

# Contents

## Game Theory

Introduction

Reinforcement Learning

Philosophy of Induction

Deep Learning

Inductive Logic

Artificial General Intelligence

Universal Induction

What If Computers Could Think?

Causal Inference

References

1753

## Game Theory = Interactive Decision-Making Theory

- ▶ 一对情侣落到一个变态杀手手里.
  - ▶ 杀手让他们玩剪刀石头布, 赢者生, 输者死.
  - ▶ 情侣商量了一下, 决定一起出石头, 不独活.
  - ▶ 结果: 女生死了, 男生活下来了.
  - ▶ 男生出了剪刀, 女生出了布.
1. 男生爱女生, 女生自私?
  2. 女生爱男生, 男生自私?
  3. 男女彼此相爱?
  4. 男女都自私?

# 拍卖 10 块钱!

现有 10 块钱钞票要拍卖. 规则如下:

1. 竞价者之间彼此不能交流.
2. 出价从 5 毛钱开始, 每次只能加 5 毛钱.
3. 出价最高者赢得钞票.
4. 但出价次高者也要支付自己的出价.

- ▶ 军备竞赛
- ▶ 美国大选
- ▶ “粉丝”为“爱豆”“打榜”
- ▶ 商家价格战
  - 可信威胁 (如: 商家宣称“买贵了, 退差价”; 如果对方宣称“退 2 倍差价”呢?)
- ▶ 争相行贿搞关系
- ▶ “内卷”

# Why Game Theory?

- ▶ 解释人们以特定方式行为的原因.
- ▶ 分析给定场景中的最优行为.
- ▶ 机制设计：寻找某种互动方式，以诱导出某种类型的行为.
- ▶ 如何把个体行为与社会利益统一起来?<sup>12</sup>
- ▶ 怎么解决个体理性与集体理性的矛盾？
- ▶ 集体理性也可以理解为个体的“事前理性”.
- ▶ 理性人之间如何更好地合作.
- ▶ Cooperation is agents with different goals interacting to mutual benefit.
- ▶ 合作问题的核心是激励 (incentive).
- ▶ 怎么解决“无耻”与“无知”？

<sup>12</sup>经济学家、伦理学家亚当·斯密在《道德情操论》《国富论》中提出“看不见的手”，研究如何对从利己出发的个体行为进行调节，达到私利与公益相协调的均衡。

不读《国富论》不知道应该怎样才叫‘利己’，读了《道德情操论》才知道‘利他’才是问心无愧的‘利己’。  
——米尔顿·弗里德曼

*The word “model” sounds more scientific than “fable” or “fairy tale” although I do not see much difference between them. Being something between fantasy and reality, a fable is free of extraneous details and annoying diversions. Thus we can clearly discern what cannot always be seen from the real world. On our return to reality, we are in possession of some sound advice or a relevant argument that can be used in the real world.*

— Ariel Rubinstein

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies  
Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret

Minimization

Subgame Perfect Equilibrium

Games with Incomplete

Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

References 1753

# Determinants of Behavior in Situations of Strategic Interaction

*Alice: "Would you please tell me, please, which way I ought to go from here?"*

*"That depends a good deal on where you want to get to," said the Cat.*

*"I don't much care where —" said Alice.*

*"Then it doesn't matter which way you go," said the Cat.*

— Lewis Carroll: *Alice's Adventures in Wonderland*

## ► Preferences

- ▶ a **preference relation** is a complete and transitive ordering of all alternatives available to an agent in a given situation.
- ▶ given these assumptions, preferences can be represented by means of a **utility function**.

## ► Beliefs:

in situations of strategic interaction, an agent has to form beliefs about the behavior of other agents.

## ► Rational behavior:

a rational agent chooses the course of action that maximizes his **utility** given the **available actions** and given his **beliefs**.

# Where do utilities come from?

## From Preferences to Utility

- ▶ A lottery  $\sum_i p_i o_i$  is a probability distribution over outcomes  $O$ .
- ▶ The outcomes  $O$  is closed under lotteries, and the agent also has preferences for lotteries.
- ▶ Preferences of a rational agent must obey constraints.

1. completeness:

$$a \succeq b \vee b \succeq a$$

2. transitivity:

$$a \succeq b \wedge b \succeq c \rightarrow a \succeq c$$

3. continuity:

$$a \succeq b \succeq c \rightarrow \exists p : pa + (1 - p)c \sim b$$

4. independence:

$$a \succeq b \leftrightarrow \forall p : pa + (1 - p)c \succeq pb + (1 - p)c$$

# Where do utilities come from?

Theorem (von Neumann-Morgenstern Utility Theorem 1944)

If a preference relation  $\succ$  satisfies the above constraints, then there exists a function  $u : O \rightarrow \mathbb{R}$  such that

$$a \succ b \iff u(a) > u(b)$$

where  $u\left(\sum_i p_i o_i\right) = \sum_{i=1}^n p_i u(o_i)$

## Remark

- ▶ 表示偏好关系的效用函数不唯一.
- ▶ 表示同一个偏好关系的效用函数之间存在仿射变换.

$$\exists \alpha > 0 \exists \beta : u'(o) = \alpha u(o) + \beta$$

- ▶ **Problem:** 效用可以跨人际比较吗?

## Remarks

- ▶ von Neumann-Morgenstern: if a rational decision maker knows the probabilities of obtaining a certain outcome, then she must choose as if maximizing the expected value of some function  $u : \mathcal{O} \rightarrow \mathbb{R}$ .
- ▶ What if the probabilities are not given?
- ▶ Savage: a decision maker's preferences  $\succ$  satisfies some "rationality" axioms, iff, there exists a nonatomic finitely additive probability measure  $\mu$  on  $S$  and a nonconstant bounded function  $u : \mathcal{O} \rightarrow \mathbb{R}$  such that for every  $f, g \in \mathcal{O}^S$ ,

$$f \succ g \iff \int_S u(f(s))d\mu(s) > \int_S u(g(s))d\mu(s)$$

Furthermore, in this case,  $\mu$  is unique and  $u$  is unique up to positive linear transformations.

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Preferences and Expected Utility      Deep Learning

**Strategic Games**

Extensive Games with Perfect Information      Artificial General Intelligence

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret

Minimization

Subgame Perfect Equilibrium

Games with Incomplete

Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

References 1753

# Games in Strategic Form

## Definition (Games in Strategic Form)

A game in *strategic form* is a triplet  $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ .

- ▶ The sets of players  $N$ .
- ▶ The sets of strategies of the players  $(S_i)_{i \in N}$ .
- ▶ The payoff functions  $u_i : S \rightarrow \mathbb{R}$ , where  $S := \prod_{i \in N} S_i$ .

# 囚徒困境及其它

	合作	背叛
合作	-1, -1	-4, 0
背叛	0, -4	-3, -3

Table: 囚徒困境 (占优均衡)

	歌剧	足球
歌剧	1, 2	0, 0
足球	0, 0	2, 1

Table: 性别大战 (纳什均衡)

	妥协	进攻
妥协	0, 0	-1, 1
进攻	1, -1	$-\infty, -\infty$

Table: 胆小鬼博弈 (纳什均衡)

	正面	反面
正面	1, -1	-1, 1
反面	-1, 1	1, -1

Table: 正反匹配 (混合纳什均衡)

- ▶ **帕累托最优:** 除非“损人”, 否则不可能“利己”.
- ▶ **占优策略:** 独立于其他人的选择的最优策略.
- ▶ **占优均衡:** 所有人的占优策略的组合.
- ▶ **纳什均衡:** 所有人的最优策略的组合, 给定该策略中别人的选择, 没有人有积极性改变自己的选择.

# 求知 vs 应试

	求知	应试
求知	100, 100	20, 80
应试	80, 20	60, 60

Table: 学生内卷 (纳什均衡)



又卷又菜

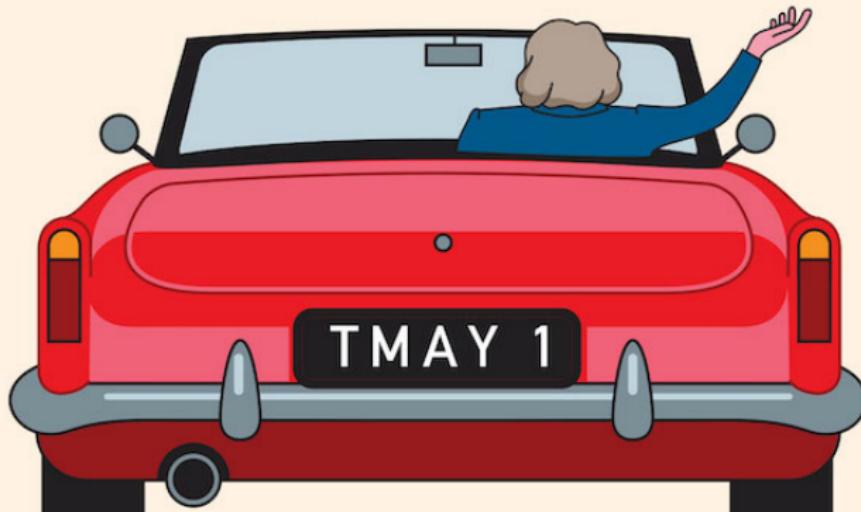
Swerve

H

N



Straight



# 多人囚徒困境 — Braess Paradox

*If we all go for the blonde and block each other, not a single one of us is going to get her. So then we go for her friends, but they will all give us the cold shoulder because no one likes to be second choice. But what if none of us goes for the blonde?*

— *A Beautiful Mind*

- ▶ The addition of options is not necessarily a good thing.
- ▶ A strategy profile is Pareto optimal iff no other strategy profile improves the payoff to at least one player without decreasing the payoff of other players.
- ▶ The Nash equilibrium of a game is not necessarily Pareto optimal.

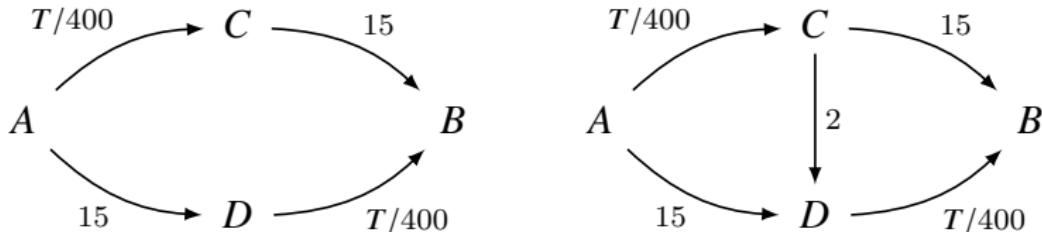
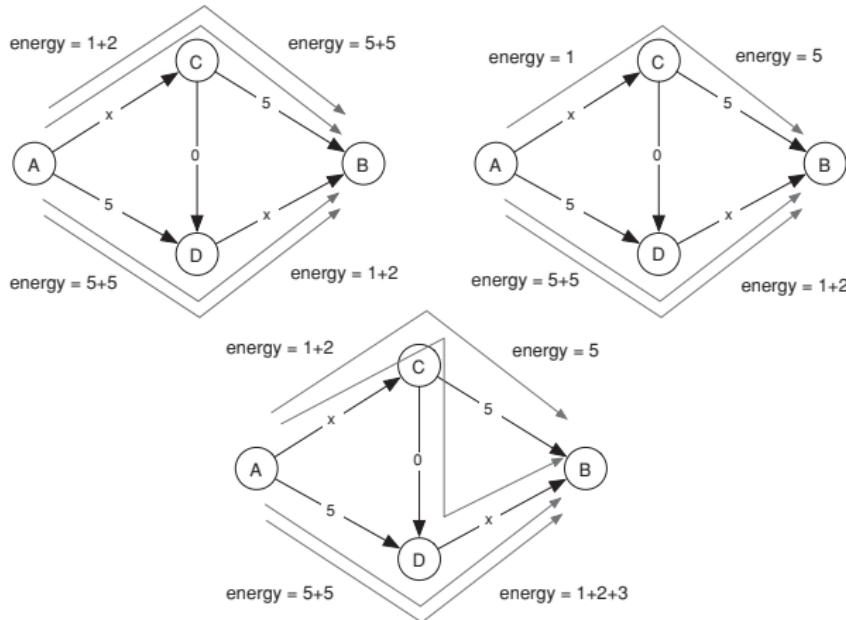


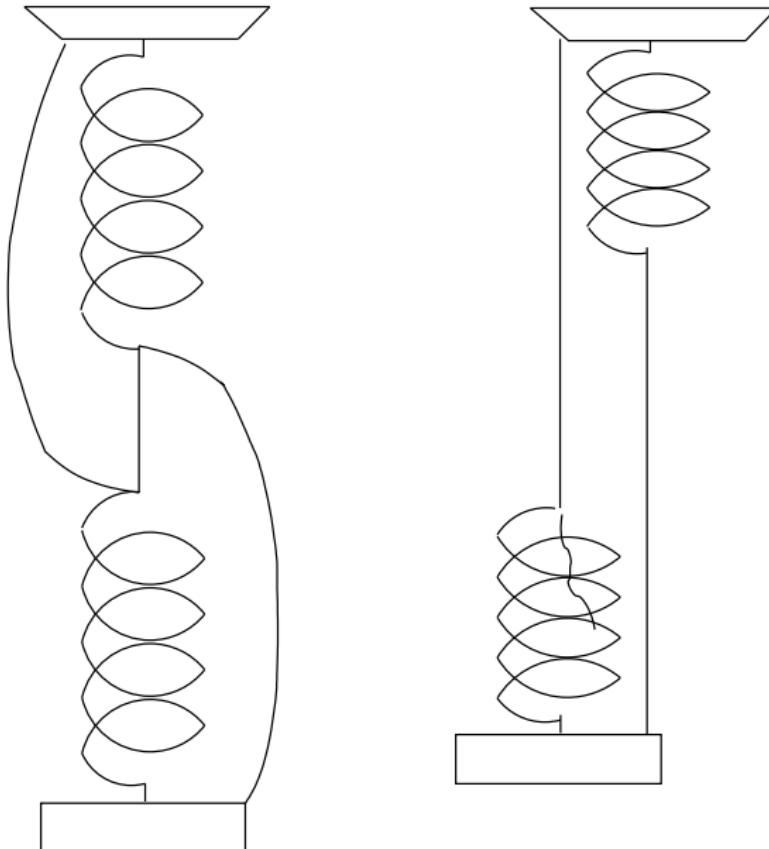
Figure:  $T = 4000$  cars travelling around the lake

# Braess Paradox

Let  $L_e(x)$  be the travel time of each car traveling along edge  $e$  when  $x$  cars take that edge ( $L_e(0) := 0$ ). Suppose there is a traffic graph  $G$  with  $x_e$  cars along edge  $e$ . Let  $E(e) := \sum_{i=1}^{x_e} L_e(i)$ , and  $E(G) := \sum_{e \in G} E(e)$ . Take a choice of routes that minimizes the total energy  $E(G)$ . That will be a Nash equilibrium.



# Braess Paradox



## Classification of Games

- ▶ 2 players vs.  $N$  players
- ▶ Static vs. Dynamic games
  - In static games, the players simultaneously choose actions, and they receive payoffs that depend on the combination of actions just chosen.
- ▶ Single play vs. Repeated games
- ▶ Zero-sum vs. Nonzero-sum games
- ▶ Symmetric vs. Non-symmetric games
  - symmetry:  $u_i(s_{\pi(1)}, \dots, s_{\pi(n)}) = u_{\pi(i)}(s_1, \dots, s_n)$
- ▶ Perfect vs. Imperfect information games
- ▶ Complete vs. Incomplete information games
- ▶ Cooperative vs. Non-cooperative games

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Preferences and Expected Utility

Strategic Games

**Extensive Games with Perfect Information**

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret Minimization

Subgame Perfect Equilibrium

Games with Incomplete Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

References 1753

## Definition (Games in Extensive Form)

A game in *extensive form*  $\Gamma = \langle N, \mathcal{A}, \mathcal{H}, \mathcal{Z}, A, P, (\mathcal{I}_i)_{i \in N}, f_c, (u_i)_{i \in N} \rangle$ .

- ▶ The set of players  $N$ .
- ▶ The set of actions  $\mathcal{A} := \bigcup_{i \in N} \mathcal{A}_i$ .
- ▶ A set  $\mathcal{H}$  of sequences (finite or infinite) such that,
  1.  $\emptyset \in \mathcal{H}$
  2.  $\forall m < n (h_{1:n} \in \mathcal{H} \implies h_{1:m} \in \mathcal{H})$
  3.  $\forall n \in \mathbb{N} (h_{1:n} \in \mathcal{H}) \implies h_{1:\infty} \in \mathcal{H}$
- ▶ A history  $h \in \mathcal{H}$  is terminal iff  $h = h_{1:\infty}$  or there is no  $a \in \mathcal{A}$  s.t.  $ha \in \mathcal{H}$ . The set of terminal histories is denoted by  $\mathcal{Z}$ .
- ▶ The set of actions at  $h$ :  $A(h) := \{a : ha \in \mathcal{H}\}$ .
- ▶ The player function:  $P : \mathcal{H} \setminus \mathcal{Z} \rightarrow N \cup \{c\}$ .
- ▶ The function  $f_c$  associates with every history  $h$  for which  $P(h) = c$  a probability measure  $f_c(\cdot | h)$  on  $A(h)$ .  $f_c(\cdot | h) \in \Delta(A(h))$ .
- ▶ For player  $i \in N$ ,  $\mathcal{I}_i := \text{Partition}(\{h \in \mathcal{H} : P(h) = i\})$  where  $\forall h, h' \in I \in \mathcal{I}_i : A(h) = A(h')$ .
- ▶ The Payoff function for each player  $i \in N$ :  $u_i : \mathcal{Z} \rightarrow \mathbb{R}$ .

- ▶ If  $|I| = 1$  for all  $I \in \mathcal{I}_i$  and all  $i \in N$ , then we say that  $\Gamma$  is a game with *perfect information*.

**Remark:** Each player is aware of every player's past actions and observations.

- ▶ If  $|I| > 1$  for some  $i \in N$  and some  $I \in \mathcal{I}_i$ , then we say that  $\Gamma$  is a game with *imperfect information*.
- ▶ If all elements of  $\Gamma$  are common knowledge, then we say that  $\Gamma$  is a game with *complete information*; otherwise, a game with *incomplete information*.

### Theorem

*Every perfect information game in extensive form has a pure strategy Nash equilibrium.*

# Pure Strategy

## Definition (Pure Strategy)

A *pure strategy* of player  $i$  is  $s_i : \mathcal{I}_i \rightarrow \mathcal{A}$  s.t.  $s_i(I) \in A(I)$  for all  $I \in \mathcal{I}_i$ .

The *set of pure strategies* of player  $i$  is denoted by  $S_i := \prod_{I \in \mathcal{I}_i} s_i(I)$ .

A *strategy profile*

$$s = (s_i)_{i \in N}$$

is a vector consisting of one pure strategy for each player.

The *set of strategy profiles* is

$$S = \prod_{i \in N} S_i$$

$$s_{-i} := (s_1 \dots s_{i-1}; s_{i+1} \dots s_{|N|})$$

$$S_{-i} := S_1 \times \dots \times S_{i-1} \times S_{i+1} \times \dots \times S_{|N|}$$

## Definition (Pure Strategy Consistent with History)

For any history  $h$  define a pure strategy  $s_i$  of player  $i$  to be consistent with  $h$

$$s_i \rightsquigarrow h \iff \forall k < \ell(h) (P(h_{1:k}) = i \implies s_i(h_{1:k}) = h_{k+1})$$

$$\sigma_c(h) := \prod_{h' \preceq h} f_{P(h')}(h_{\ell(h')+1} \mid h')$$

where  $f_{P(h)}(\cdot \mid h) := 1$  if  $P(h) \neq c$ .

Each strategy profile  $s = (s_i)_{i \in N}$  can be assigned a payoff for each player  $i$ :

$$u_i(s) := \sum_{h: \forall i (s_i \rightsquigarrow h)} \sigma_c(h) u_i(h)$$

Obviously,

$$\forall h \in \mathcal{H} (P(h) \neq c) \implies u_i(s) = u_i(h^s)$$

where  $h^s$  is the history uniquely determined by  $s$ .

# Mixed Strategy

## Definition (Mixed Strategy)

A *mixed strategy* of player  $i$  in an extensive game is a probability measure over the set of player  $i$ 's pure strategies:

$$\sigma_i \in \Delta S_i$$

A *mixed strategy profile* specifies a mixed strategy for each player  $i$

$$\sigma := (\sigma_i)_{i \in N}$$

The *set of mixed strategy profiles* is:

$$\prod_{i \in N} \Delta S_i$$

Each mixed strategy profile  $\sigma = (\sigma_i)_{i \in N}$  implies a probability distribution over the *set of pure strategy profiles*  $S$

$$\sigma(s) := \prod_{i \in N} \sigma_i(s_i)$$

# Payoff of Mixed Strategy

## Definition (Payoff of Mixed Strategy)

The expected payoff of player  $i$  given a mixed strategy profile  $\sigma$  is

$$u_i(\sigma) := \sum_{s \in S} u_i(s)\sigma(s)$$

# 两人零和博弈 — 极小极大值定理

Theorem (Minimax Theorem — von Neumann 1928)

Let  $X \subset \mathbb{R}^n$  and  $Y \subset \mathbb{R}^m$  be compact convex sets. Let  $f : X \times Y \rightarrow \mathbb{R}$  be a continuous function that is concave on  $X$  for each fixed  $y \in Y$ , and convex on  $Y$  for each fixed  $x \in X$ . Then

$$\max_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \max_{x \in X} f(x, y)$$

任意两人零和博弈都有混合纳什均衡  $(x^*, y^*)$ :

$$u_1(x^*, y^*) = \max_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \max_{x \in X} f(x, y) = -u_2(x^*, y^*)$$

其中,  $f(x, y) := x^T A y$ .

**Remark:**  $f(x^*, y^*)$  是鞍点:  $f(x^*, y) \geq f(x^*, y^*) \geq f(x, y^*)$ .

A	石头	剪刀	布
石头	0	1	-1
剪刀	-1	0	1
布	1	-1	0

$$x^* = y^* = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$$
$$u_1 = u_2 = 0$$

# 两人零和博弈 — Maxmin Strategies

- ▶ The maxmin strategy for player  $i$  is  $\operatorname{argmax}_{s_i} \min_{s_{-i}} u_i(s_i; s_{-i})$ , and player  $i$ 's maxmin value is  $\max_{s_i} \min_{s_{-i}} u_i(s_i; s_{-i})$ .
  - ▶ Why would  $i$  want to play a maxmin strategy?
  - ▶ a conservative agent maximizing worst-case payoff.
- ▶ In a two-player game, the minmax strategy for player  $i$  against player  $-i$  is  $\operatorname{argmin}_{s_i} \max_{s_{-i}} u_{-i}(s_i; s_{-i})$ , and player  $-i$ 's minmax value is  $\min_{s_i} \max_{s_{-i}} u_{-i}(s_i; s_{-i})$ .
  - ▶ Why would  $i$  want to play a minmax strategy?
  - ▶ to punish the other agent as much as possible.

# Behavioral Strategy

## Definition (Behavioral Strategy)

A *behavioral strategy* of player  $i$  is a collection  $(\rho_i(I))_{I \in \mathcal{I}_i}$  of probability measures, where

$$\rho_i(I) \in \Delta(A(I)) \text{ for } I \in \mathcal{I}_i$$

A *behavioral strategy profile* is a vector consisting of one behavioral strategy for each player:

$$\rho := ((\rho_i(I))_{I \in \mathcal{I}_i})_{i \in N}$$

For any history  $h \in I \in \mathcal{I}_i$  and action  $a \in A(h)$ , we denote by  $\rho_i(h)(a)$  the probability  $\rho_i(I)(a)$  assigned by  $\rho_i(I)$  to the action  $a$ .

The *set of behavioral strategies* of player  $i$  is

$$B_i := \prod_{I \in \mathcal{I}_i} \Delta(A(I))$$

The *set of behavioral strategy profiles* is:

$$B := \prod_{i \in N} B_i = \prod_{I \in \mathcal{I}} \Delta(A(I))$$

## Payoff of Behavioral Strategy

### Definition (Payoff of Behavioral Strategy)

The expected payoff of player  $i$  given a behavioral strategy profile  $\rho$  is

$$u_i(\rho) := \sum_{s \in S} u_i(s) \prod_{i \in N} \prod_{I \in \mathcal{I}_i} \rho_i(I)(s_i(I))$$

# Perfect Recall

## Definition (Experience)

Given a history  $h$  of an extensive game, the *experience*  $X_i(h)$  of player  $i$  is the sequence of information sets that player  $i$  encounters in  $h$  and the actions that player  $i$  takes at them.

## Definition (Perfect Recall)

An extensive game has *perfect recall* if for each player  $i$ , we have  $X_i(h) = X_i(h')$  whenever  $h, h'$  are in the same information set of player  $i$ .

**Remark:** A player with perfect recall always remembers what actions he has taken and what he knew previously.

## Two kinds of imperfect recall:

- ▶ **Forgetfulness:** an agent forgets an observation or the outcome of one of their previous decisions.
- ▶ **Absent-mindedness:** an agent cannot remember whether he has previously made a decision.

## Imperfect Recall — Experience

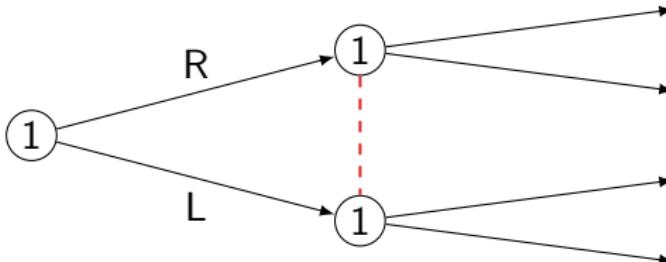


Figure:  $X_1(L) = \{\emptyset, L, \{L, R\}\}$ ,  $X_1(R) = \{\emptyset, R, \{L, R\}\}$

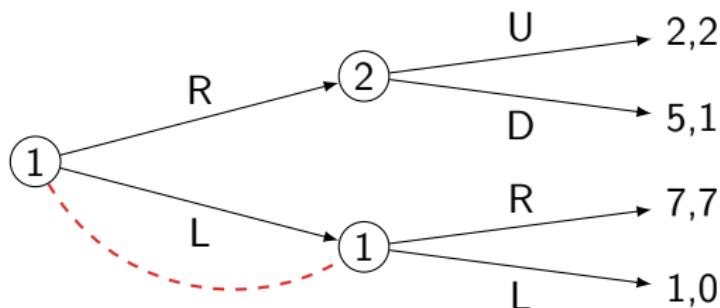
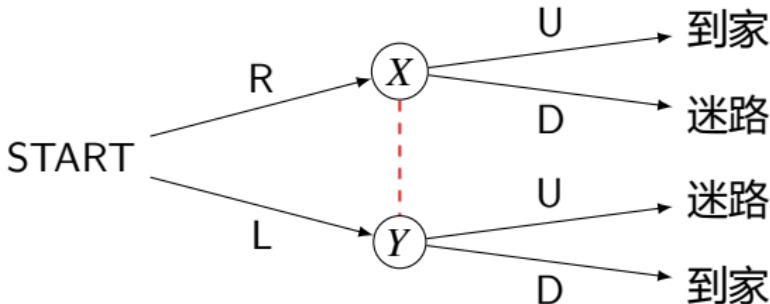


Figure:  $X_1(\emptyset) = \{\emptyset, L\}$ ,  $X_1(L) = \{\{\emptyset, L\}, L, \{\emptyset, L\}\}$

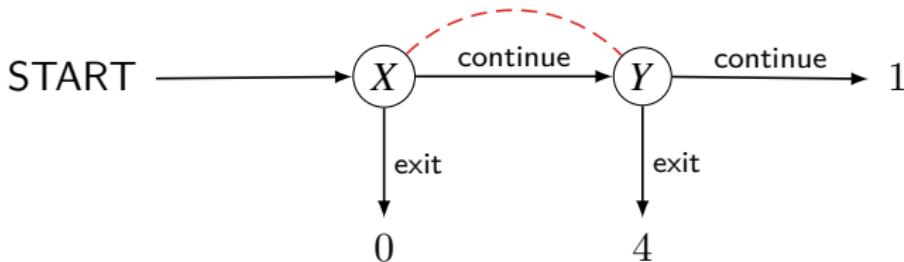
# The difference between mixed and behavioral strategies



- ▶ Mixed strategies:  $(p_{LD}, p_{LU}, p_{RD}, p_{RU})$ , for example,  $(\frac{1}{2}, 0, 0, \frac{1}{2})$
- ▶ Behavioral strategies:  $\rho_i(\emptyset)(L) = p, \rho_i(\emptyset)(R) = 1 - p; \rho_i(\{L, R\})(D) = q, \rho_i(\{L, R\})(U) = 1 - q.$
- ▶ The mixed strategy  $(\frac{1}{2}, 0, 0, \frac{1}{2})$  has no equivalent behavior strategy.

**Remark:** Both forgetfulness and absent-mindedness can prevent the existence of an Nash Equilibrium in behavioural policies.

# The Absent-Minded Driver



- ▶ 因为无法区分自己在哪个路口, 所以每个路口应该采取相同的策略.
- ▶ 只有两个纯策略: `continue`, `exit`. 所以, 没有混合策略能达到 4.
- ▶ 因此, 没有混合策略等价于行为策略  $\rho(I)(\text{exit}) = \rho(I)(\text{continue}) = \frac{1}{2}$ .
- ▶ 下面计算规划最优策略. 假设其策略是以概率  $p$  沿公路 `continue`.
- ▶ 司机的期望回报为:

$$U(p) = (1 - p) \cdot 0 + p[(1 - p) \cdot 4 + p \cdot 1]$$

- ▶ 在 `START` 出发前, 规划最优的决策为,

$$\frac{dU}{dp} = 0 \implies p^* = \frac{2}{3} \implies U\left(\frac{2}{3}\right) = \frac{4}{3}$$

## CDT Solution

- ▶ 如果认为一开始决定  $\text{continue}$  的决策是  $q$ , 当下路口正考虑的是  $p$ .
- ▶ 协调的信念会认为正处在第一个路口的可能性是  $\mu(I)(X) = \frac{1}{1+q}$ .
- ▶ 主观期望回报是,

$$U(p, q) = \frac{1}{1+q} [(1-p) \cdot 0 + p(1-q) \cdot 4 + pq \cdot 1] + \frac{q}{1+q} [(1-p) \cdot 4 + p \cdot 1]$$

- ▶ 两个路口的最优决策应该一样. 姑且称之为  $p^*$ . 相信另一个路口选择最优的  $p^*$  的话, 这一个路口选择  $p = p^*$  也应该是最优的.

$$F(q) := \operatorname{argmax}_p U(p, q)$$

$$p^* \in F(p^*)$$

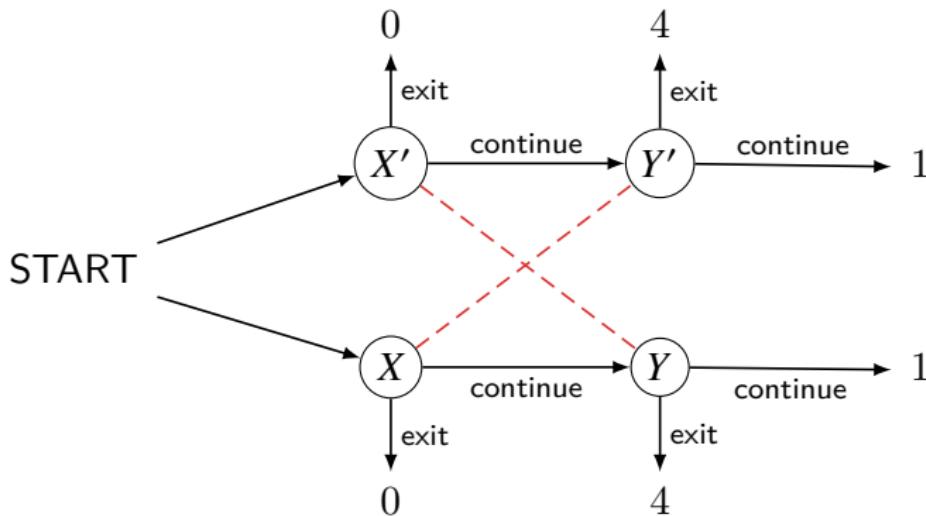
- ▶ 在路口, 行动最优的决策跟出发前的规划最优的决策一样, 都是  $\frac{2}{3}$ .

$$\frac{dU}{dp} = 0 \implies q = \frac{2}{3} \implies p^* = \frac{2}{3} \implies U\left(\frac{2}{3}, \frac{2}{3}\right) = \frac{8}{5} > \frac{4}{3} = U\left(\frac{2}{3}, 0\right)$$

## EDT & CDT in Game Theory

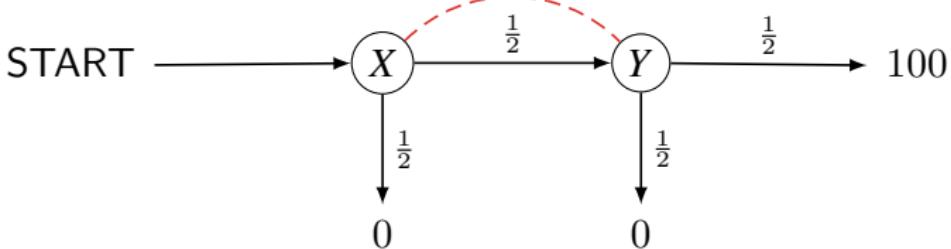
- ▶ EDT postulates that if a player deviates to a randomized decision  $\alpha \in \Delta(A(I))$  at the current node, then she will have also deviated to  $\alpha$  whenever she arrived in  $I$  in the past, and that she will also deviate to  $\alpha$  whenever she arrives in  $I$  in the future.
- ▶ CDT postulates that a player can deviate to a decision  $\alpha \in \Delta(A(I))$  at the current node, without violating her behavioral strategy  $\rho$  at past arrivals or future arrivals in  $I$ .

# The Absent-Minded Driver — The Multiselves Approach



- ▶ 设想两个相同的 player, 以相同的概率令一个在某个路口做决策, 另一个在另一个路口做决策.
- ▶ 通过多个自我, 将 imperfect recall 变为 perfect recall.

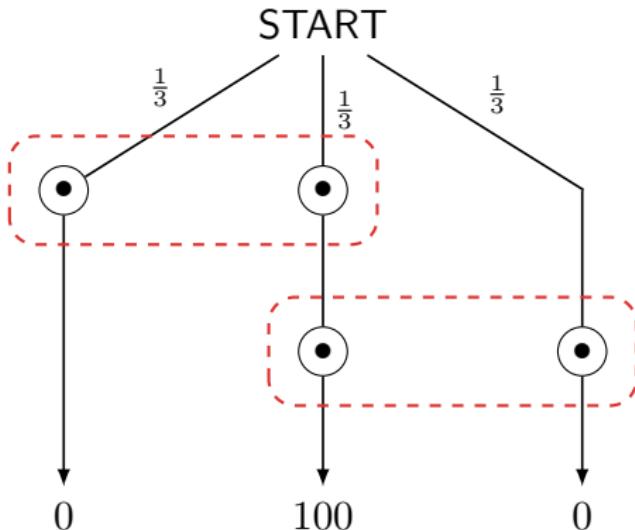
# The Forgetful Passenger



- ▶ 在 START, 乘客能得 100 的概率是  $\frac{1}{4}$ .
- ▶ 在  $X$ , 乘客能得 100 的主观概率变成了  $\frac{1}{3}$ , 虽然他会到  $X$  是必然的.  
— 因为: 经过  $X$  的概率是经过  $Y$  的两倍, 所以, 根据协调的信念,  
 $\frac{2}{3} \cdot (\frac{1}{2})^2 + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{3}$
- ▶ 假设在每一个点 (START,  $X$ ,  $Y$ ), 他都有机会花 30 块钱买一张赌约.
- ▶ 若他在 START, 一张赌约值  $\frac{1}{4} \cdot 100 - 30 = -5$ ; 若他在  $X$ , 对他来说, 值  $\frac{1}{3} \cdot 100 - 30 = \frac{10}{3}$ .
- ▶ 这是咋回事? — 因为在  $X$ , 他不知道自己在  $X$ .
- ▶ 最优策略: 每个路口买一张赌约!

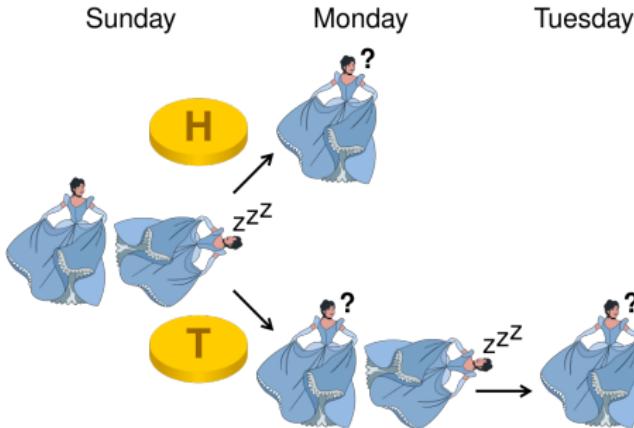
$$\frac{1}{2} \cdot (-30) + \frac{1}{4} \cdot (-60) + \frac{1}{4} \cdot (200 - 60) = 5$$

## The Forgetful Passenger — another example



- ▶ 在 START, 能得 100 的概率是  $\frac{1}{3}$ .
- ▶ 在后面两个信息集, 概率变成了  $\frac{1}{2}$ .
- ▶ 但从 START 出发, 必然会到达某个信息集.

# 睡美人问题



- ▶ **睡美人问题:** 睡美人周日睡前被告知: 实验者会抛一枚硬币. 如果硬币正面朝上, 她会在周一被唤醒一次; 如果反面朝上, 则在周一、周二各被唤醒一次. 周一醒后入睡时会被清除醒来过的记忆. 周三游戏结束. 每次被唤醒后, 都会被问一个问题: “你有多相信之前抛出的硬币是正面朝上?”
- ▶ 变种一: 如果硬币没有提前抛, 而是在周一醒后再次入睡时抛呢?
- ▶ 变种二: 如果睡美人周一醒后被告知当天是周一呢?

# 睡美人问题的主要派别

1. 二分派: 醒来后没获得任何新信息, 概率不变. 符合 van Fraassen's Reflection Principle:  $P_0(H | P_1(H) = x) = x$ . 既然事前就知道肯定会醒, 事前的概率就应该跟醒后的一致.
2. 三分派: 设想睡美人面临“变种二”, 被告知当天是周一, 她会想: 既然无论硬币正反周一都会被唤醒, 原始版本与“变种一”等价, 完全可以认为硬币现在还没抛, 所以信念不应该变. 即  
 $P(C = 1 | D = 1) = P(C = 0 | D = 1) = \frac{1}{2}$ . 又由于, 反面朝上时, 周一周二认知不可区分, 所以  $P(D = 1 | C = 0) = P(D = 2 | C = 0) = \frac{1}{2}$ .  
于是,  $P(C = 1, D = 1) = P(C = 0, D = 1) = P(C = 0, D = 2)$ . 而这三种情况构成了睡美人醒来的所有互斥的可能. 所以  
 $P(C = 1 | A = 1) = \frac{1}{3}$ .
3. 双二分派: 无论是初始的, 被唤醒时的, 还是被告知周一后的都是  $1/2$ .
4. 歧义派: 想象一下, 如果硬币正面朝上, 就把一个绿球放入盒子; 如果反面朝上, 就把两个红球放入盒子. 问题一: “你有多相信硬币朝上? / 一个绿球被放入盒子的概率是多少?”. 问题二: “你有多相信硬币朝上导致了这次醒来? / 一个绿球从盒子里抽出的概率是多少?”

## 不同派别一致接受的预设:

1. 游戏设定的唤醒条件: 周一正面, 周一反面, 周二反面唤醒; 周二正面不唤醒.  $P(A = 1 | C = 1, D = 1) = P(A = 1 | C = 0, D = 1) = P(A = 1 | C = 0, D = 2) = 1$ ,  $P(A = 1 | C = 1, D = 2) = 0$ .
2. 周日睡美人对硬币的初始信念  $P(C = 1) = P(C = 0) = \frac{1}{2}$ .
3. 周一睡美人被唤醒后的信念按贝叶斯规则更新:  $P(C | A = 1)$ .
4. 硬币反面朝上时, 睡美人对当天是周一还是周二认知不可区分:  $P(D = 1 | C = 0) = P(D = 2 | C = 0) = \frac{1}{2}$ .

## 不同派别各自特有的预设:

### ► 三分派:

1. 睡美人被告知当天是周一后的信念按照贝叶斯规则更新:  $P(C | D = 1)$ .
2. 当被告知周一后, 睡美人对硬币的信念不变:  $P(C = 1 | D = 1) = P(C = 1) = \frac{1}{2}$ .

### ► 二分派: 当被唤醒时, 睡美人对硬币的信念不变<sup>13</sup>:

$$P(C = 1 | A = 1) = P(C = 1) = \frac{1}{2}.$$

---

<sup>13</sup>与范弗拉森的内省原则  $P_0(H | P_1(H) = x) = x$  相符.

## Bostrom's Self-Sampling & Self-Indication Assumption

1. **自我抽样假设 SSA:** 其他条件相同的情况下, 观察者为了计算自己在某个世界中的位置, 应该先为可能世界分配概率, 再从世界(过去, 现在, 未来)真实存在的观察者集合中随机选择.

► 两个可能世界: 正面和反面.

$$P(C = 1) = P(C = 0) = \frac{1}{2}$$

► 正面世界有一个观察者, 抽样概率 1; 反面世界有两个, 抽样概率  $\frac{1}{2}$ .

$$P(D = 1 | C = 1) = 1$$

$$P(D = 1 | C = 0) = P(D = 2 | C = 0) = \frac{1}{2}$$

2. **自我指示假设 SIA:** 其他条件相同的情况下, 观察者应该从所有认知上可能的观察者集合中随机选择.

$$P(C = 1, D = 1) = P(C = 0, D = 1) = P(C = 0, D = 2) = \frac{1}{3}$$

**Remark:** 自我抽样: “二分派”; 自我指示: “三分派”.

## Armstrong's Anthropic Decision Theory

- ▶ Armstrong 认为, 睡美人问题的核心不在于概率计算, 而在于决策, 不同伦理类型的睡美人如何进行决策. 其中, 他区分了总体功利主义和平均功利主义.
- ▶ 你有机会买一张赌约, 若正面朝上, 付你 1\$, 你愿意花多少钱买?
- ▶ 总体功利主义: 若反面朝上, 连续叫醒  $n$  天, 则总体期望得失:

$$\frac{1}{2} \cdot 1 \cdot (1 - x) + \frac{1}{2} \cdot n \cdot (-x) \geq 0 \implies x \leq \frac{1}{n+1}$$

- ▶ 平均功利主义: 反面世界的平均得失是  $-x$ , 正面世界的平均得失是  $1 - x$ , 所以, 平均期望得失:

$$\frac{1}{2} \cdot (1 - x) + \frac{1}{2} \cdot (-x) \geq 0 \implies x \leq \frac{1}{2}$$

- ▶ 总体功利主义类似自我指示, 平均功利主义类似自我抽样.

**Remark:** 睡美人问题的有趣之处可能不在于决策者的伦理类型, 而在于不完美回忆 (imperfect recall) 下的信念分配.

# Consistent Belief System vs Z-Consistent Belief System

## Definition (Consistent Belief System)

A belief system  $\mu$  is **consistent** with the behavioral strategy  $\rho$  if for every information set  $I$  which is reached with positive probability and for every  $h \in I$ ,

$$\mu(I)(h) = \frac{\rho(h)}{\sum_{h' \in I} \rho(h')}$$

## Definition (Z-Consistent Belief System)

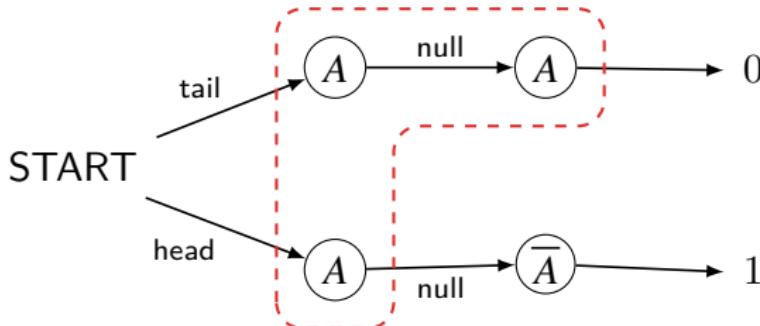
A belief system  $\mu$  is **Z-consistent** with the behavioral strategy  $\rho$  if for every information set  $I$  which is reached with positive probability and for every  $h \in I$ ,

$$\mu(I)(h) = \frac{\sum_{z: h \preceq z} \frac{\rho(z)}{\#\{h' \in I : h' \preceq z\}}}{\sum_{z: \exists h' \in I (h' \preceq z)} \rho(z)}$$

**Remark:** 在完美回忆的情况下, 二者相同, 即贝叶斯更新.

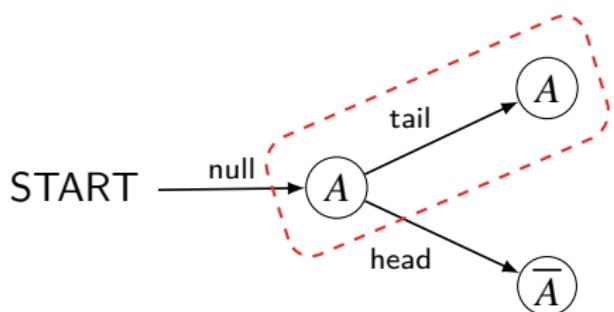
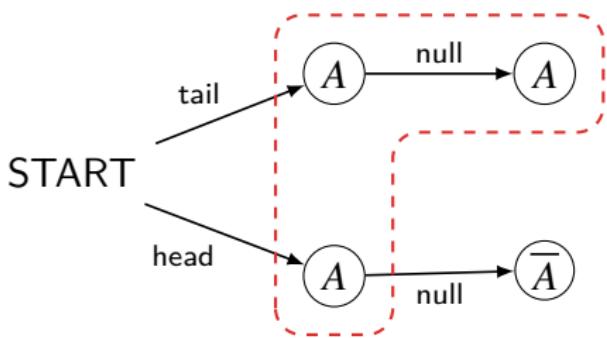
# 睡美人问题 — 等价的打赌表述

问题：你有机会买一张赌约，若正面朝上，付你 1\$，你愿意花多少钱买？



- ▶  $I = \{h, t, tn\}$
- ▶  $Z = \{hn, tn\}$
- ▶ consistent belief:  $\mu(I)(h) = \mu(I)(t) = \mu(I)(tn) = \frac{1}{3}$
- ▶  $Z$ -consistent belief:  $\mu(I)(h) = \frac{1}{2}$ ,  $\mu(I)(t) = \mu(I)(tn) = \frac{1}{4}$
- ▶ 假设愿意花  $x$ \$ 买。则
$$\mu(I)(h) \cdot (1-x) + \mu(I)(t) \cdot (0-x) + \mu(I)(tn) \cdot (0-x) \geq 0 \implies x \leq \mu(I)(h)$$
- ▶ consistent belief:  $x \leq \frac{1}{3}$ ;  $Z$ -consistent belief:  $x \leq \frac{1}{2}$
- ▶ 若 tail 朝上连续叫醒  $n$  天，则协调信念下  $x \leq \frac{1}{n+1}$ .

# 原始版本与“变种一” — 不同版本的不同决策树表示



- ▶ 硬币在周日抛还是周一抛 (变种一) 没有区别.
- ▶ 两幅图也是等价的描述.
- ▶ 但是, 图一可以帮助我们直接读出想要的概率值.

$$P(C = 1, D = 1 \mid A = 1) = \mu(I)(h)$$

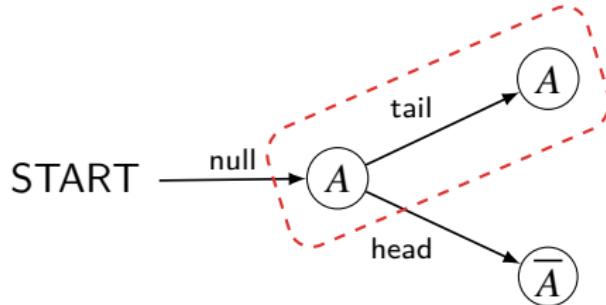
$$P(C = 0, D = 1 \mid A = 1) = \mu(I)(t)$$

$$P(C = 0, D = 2 \mid A = 1) = \mu(I)(tn)$$

- ▶ 图二却只能通过  
 $P(D = 1 \mid A = 1) = \mu(I)(n)$  重新计算相关概率.

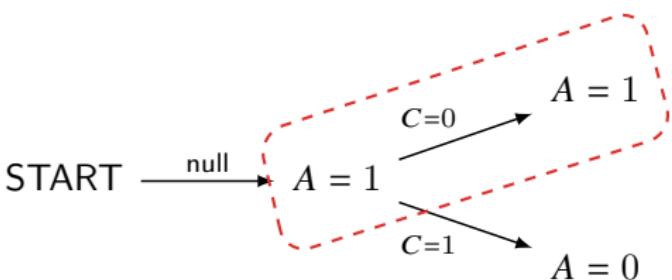
# 睡美人问题

Consistent Belief System vs Z-Consistent Belief System



- ▶ Information set:  $I = \{n, nt\}$
- ▶ The set of terminal histories:  $Z = \{nh, nt\}$
- ▶ 行为策略导出的概率:  $\rho(n) = 1, \rho(nh) = \rho(nt) = \frac{1}{2}$
- ▶ Consistent belief system:  $\mu(I)(n) = \frac{2}{3}$   
睡美人问题里对应  $P(D = 1 | A = 1) = \frac{2}{3}$ , 可得  $P(C = 1 | A = 1) = \frac{1}{3}$
- ▶ Z-Consistent belief system:  $\mu(I)(n) = \frac{3}{4}$   
睡美人问题里对应  $P(D = 1 | A = 1) = \frac{3}{4}$ , 可得  $P(C = 1 | A = 1) = \frac{1}{2}$

# 计算过程



由  $P(C = 0) = \frac{1}{2}$ , 且  $P(D = 1 | C = 0) = P(D = 2 | C = 0)$ , 可得  
 $P(C = 0, D = 1) = P(C = 0, D = 2) = \frac{1}{4}$ . 假设  $P(C = 1, D = 1) = \theta$ , 则

		$C = 1$	$C = 0$			$C = 1$	$C = 0$
		$D = 1$	$\frac{1}{4}$			$D = 1$	$A = 1$
$D = 2$	$C = 1$	$\theta$	$\frac{1}{4}$	$D = 2$	$A = 1$	$A = 1$	
	$C = 0$	$\frac{1}{2} - \theta$	$\frac{1}{4}$		$A = 0$	$A = 1$	

$$P(D = 1 | A = 1) = \frac{P(D = 1, A = 1)}{P(A = 1)} = \frac{\theta + \frac{1}{4}}{\theta + \frac{1}{4} + \frac{1}{4}} = \frac{4\theta + 1}{4\theta + 2}$$

$$P(D = 1 | A = 1) = \frac{2}{3} \implies \theta = \frac{1}{4} \implies P(C = 1 | A = 1) = \frac{\theta}{\theta + \frac{1}{4} + \frac{1}{4}} = \frac{1}{3}$$

$$P(D = 1 | A = 1) = \frac{3}{4} \implies \theta = \frac{1}{2} \implies P(C = 1 | A = 1) = \frac{\theta}{\theta + \frac{1}{4} + \frac{1}{4}} = \frac{1}{2}$$

## Remark

		$C = 1$	$C = 0$
$D = 1$	$C = 1$	$\theta$	$\frac{1}{4}$
	$C = 0$	$\frac{1}{2} - \theta$	$\frac{1}{4}$

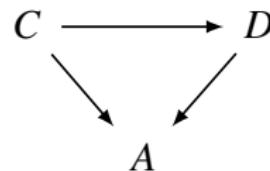
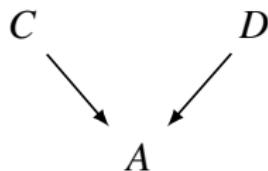
$$P(D = 1 | C = 1) = 2\theta$$

$$P(D = 2 | C = 1) = 1 - 2\theta$$

$$P(D = 1 | C = 0) = \frac{1}{2}$$

$$P(D = 2 | C = 0) = \frac{1}{2}$$

- ▶  $P(C, D) \stackrel{?}{=} P(C)P(D)$
- ▶  $\theta = \frac{1}{4}$  时,  $C$  与  $D$  独立.
- ▶ 否则,  $C$  到  $D$  有因果箭头.

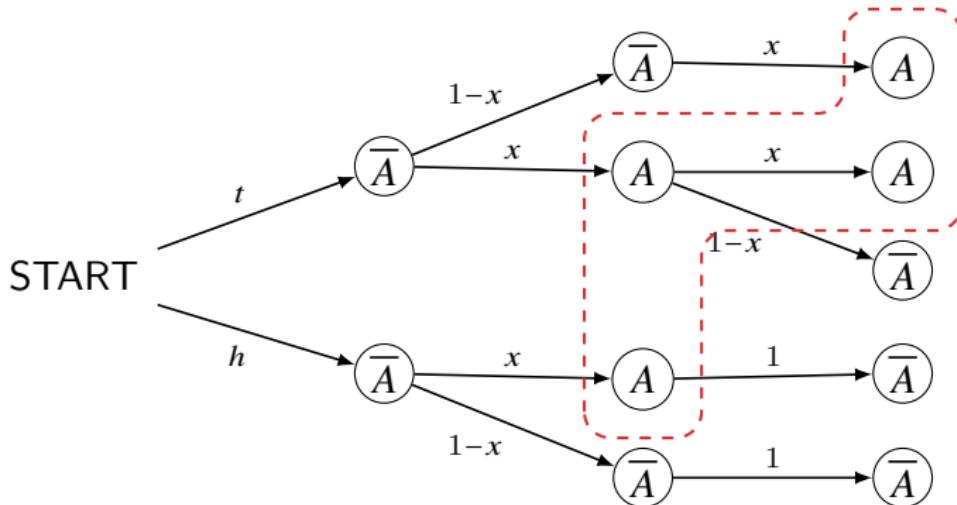


**Remark:** 反面  $C = 0$  时, 周一  $D = 1$  周二  $D = 2$  不可区分, 等概率; 若认为正面  $C = 1$  提升了周一  $D = 1$  的概率, 则  $C$  到  $D$  有因果箭头.

# 广义睡美人问题

假设睡美人每次被唤醒不再是确定的，而是以某个概率  $x$  被随机闹钟叫醒。这时，睡美人被唤醒后就有了新的证据  $W$ : “实验中我至少醒了一次”。

$$P(C = 1 \mid W) = \frac{P(W|C=1)P(C=1)}{P(W|C=1)P(C=1)+P(W|C=0)P(C=0)} = \frac{\frac{1}{2}x}{\frac{1}{2}x + \frac{1}{2}(1-(1-x)^2)} = \frac{1}{3-x}.$$

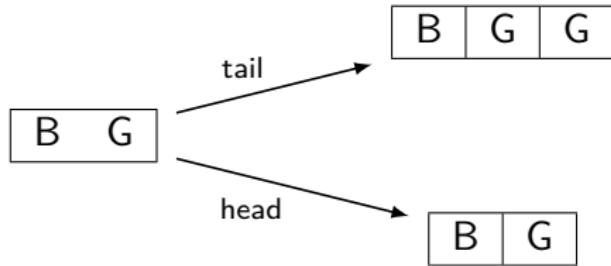


$$I = \{hx, tx, txx, t(1-x)x\}. Z = \{h(1-x)1, hx1, txx, tx(1-x), t(1-x)x\}.$$

$$\text{Z-协调信念: } \mu(I)(hx) = \frac{\rho(hx)}{\rho(hx1) + \rho(txx) + \rho(tx(1-x)) + \rho(t(1-x)x)} = \frac{1}{3-x}.$$

$$\text{协调信念: } \mu(I)(hx) = \frac{\rho(hx)}{\rho(hx) + \rho(tx) + \rho(txx) + \rho(t(1-x)x)} = \frac{1}{3}.$$

# 睡美人问题 — 另一个变种



- ▶ B 和 G 睡觉时, 有一枚硬币抛出. 如果是反面, 则 G 将被克隆一份. 然后大家都被唤醒. 大家都知道实验的设定. 但醒后无法看到彼此.
- ▶ 每人都被问一个问题: “你有多相信硬币是正面朝上?” 或者等价地问: 你有机会买一张赌约, 若正面朝上, 付你 1\$, 你愿意花多少钱买?
  1. 无论采用协调的还是 Z-协调的信念, 男孩 B 都会说  $1/2$ .
  2. 而对于女孩 G 来说, 如果她采用协调的信念的话, 会说  $1/3$ ; 如果采用 Z-协调的信念的话, 会说  $1/2$ .
- ▶ 两个认知主体, 假设都采用相同的协调信念, 面对同一件事, 拥有相同的知识, 却得到不同的预测, 不同的因果.
- ▶ Z-协调信念下则有一致的预测.

# 睡美人问题 — 对“二分派”的辩护

- ▶ 睡美人问题的核心不在于决策者的伦理类型, 而在于不完美回忆下的信念分配.
- ▶ 同样基于平均功利主义, Consistent 信念支持 “三分派”, Z-Consistent 信念支持 “二分派”.
- ▶ 在 Consistent 信念下, B 和 G 拥有相同的知识, 应用相同的理论, 却得到不同的因果图. 因果应该具有稳定性. Z-Consistent 信念才能保证因果的这种稳定性.

# 睡美人问题 — 对“三分派”的批评

- ▶ 这一派认为, 当被告知当天是周一时, 其对正面的信念应该是  $1/2$ .  
从而  $P(C = 1 | D = 1) = \frac{1}{2}$ .
- ▶ 但“三分派”的这条核心预设是错的! “被告知周一” 不能视为条件化.
- ▶ 由此, 也无法通过“告知” 更新反向推导出唤醒时的信念<sup>14</sup>.

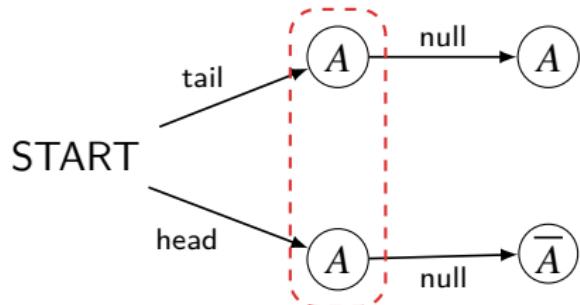
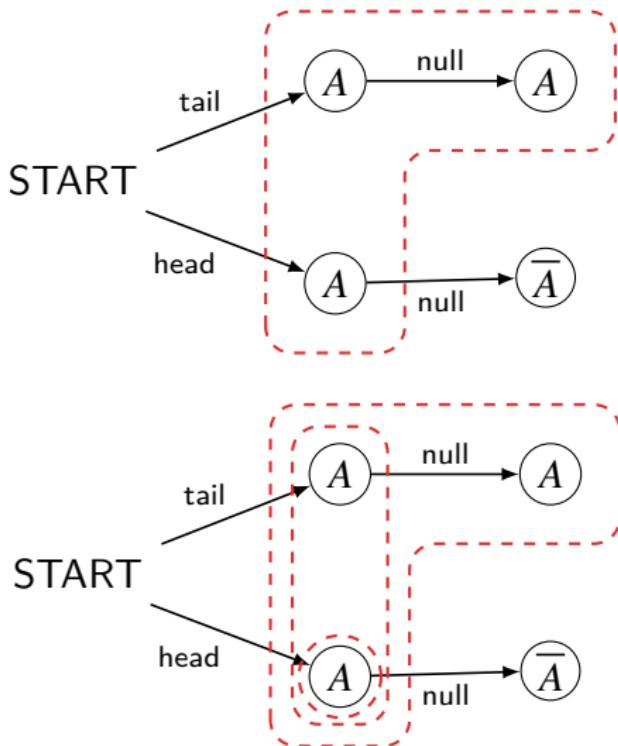
$$P^{D=1}(C = 1) = \frac{1}{2} \quad \text{而} \quad P(C = 1 | D = 1) = \frac{2}{3}$$

- ▶ Consistent 信念会让你相信魔鬼 ☺

---

<sup>14</sup>“告知”更新类似因果干预  $\text{do}$  算子, 切断了因果机制, 我们无法通过干预后的信息反向推导出干预前的状态.

# 为什么“告知”更新不能视为贝叶斯条件化?



- ▶ “告知”更新是左到右.
- ▶ 条件化是上到下.

# 区分“告知”更新和贝叶斯条件化更新

- ▶ “被告知周一”的信念更新：信息集从  $I = \{h, t, tn\}$  变为了  $I' = \{h, t\}$ ；  
信念也随之重新分配。信念也随之重新分配。分布虽变，但睡美人对正面朝上的信念值没变。假设 Z-协调的信念系统，对硬币正面朝上的信念从初始的  $P(h) = \frac{1}{2}$ ，到被唤醒时的  
 $P(C = 1 | A = 1) = \mu(I)(h) = \frac{1}{2}$ ，再到“被告知周一”后的  
 $P^{D=1}(C = 1) = \mu(I')(h) = \frac{1}{2}$ ，数值始终未变。这也是“双二分派”的直观合理的原因。
- ▶ 条件化是基于同一个信念的贝叶斯更新：

$$\begin{aligned} P(C = 1 | D = 1) &= \frac{P(C = 1, D = 1)}{P(C = 1, D = 1) + P(C = 0, D = 1)} \\ &= \frac{\mu(I)(h)}{\mu(I)(h) + \mu(I)(t)} = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{4}} = \frac{2}{3} \end{aligned}$$

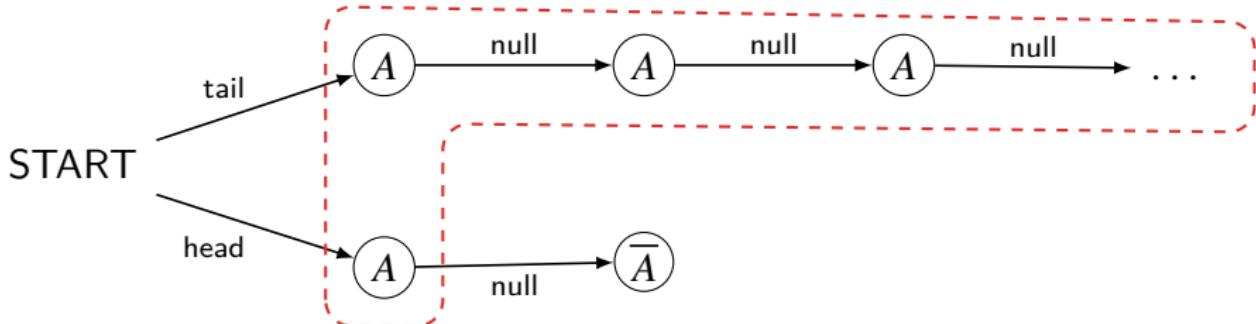
- ▶ 条件化是基于当下信息集的信念进行的计算，而“被告知周一”则需要重新划分信息集并分配信念<sup>15</sup>。

<sup>15</sup>可以理解为 Lewis 的 Imaging Rule. 参见 Cozic2011.

# 为什么“三分派”的推理是错的?

- ▶ 在协调信念下, 初始  $P(h) = \frac{1}{2}$ , 被唤醒时为  
 $P(C = 1 | A = 1) = \mu(I)(h) = \frac{1}{3}$ , “被告知周一” 后  
 $P^{D=1}(C = 1) = \mu(I')(h) = \frac{1}{2}$ .
- ▶ “被告知周一” 后的情况跟“二分派”一样.
- ▶ 对于条件化,  $P(C = 1 | D = 1) = \frac{\mu(I)(h)}{\mu(I)(h) + \mu(I)(t)} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{3}} = \frac{1}{2}$ .
- ▶ “三分派”的错误在于, 他们是通过“被告知周一”后的信念是  $\frac{1}{2}$ , 反向推导出被唤醒时是  $\frac{1}{3}$  的结论. 这无法办到的. “告知”更新不是贝叶斯更新, 反向推不出有效的结论.

# 你应该相信魔鬼吗?



- ▶ 假设正常世界是 head 的世界, 概率  $\rho(h)$  很高; tail 的世界是魔鬼创造的世界, 概率  $\rho(t)$  很低.
- ▶ 虽然  $\rho(t)$  很低, 但魔鬼的把戏强大, 能把睡美人叫醒的次数  $n$  很大.
- ▶ 根据 Consistent 信念, 你相信自己生活在正常世界的概率
$$\mu(I)(h) = \frac{\rho(h)}{\rho(h)+n \cdot \rho(t)} \xrightarrow{n \rightarrow \infty} 0.$$
- ▶ 根据 Z-Consistent 信念, 你相信自己生活在正常世界的概率
$$\mu(I)(h) = \frac{\rho(h)}{\rho(h)+\rho(t)}. \text{ 你不会被魔鬼的把戏 } n \text{ 带偏.}$$
- ▶ 算法概率:  $U(t) = 1^n$ ;  $U(h) = 10$ . 其中  $t$  的信息集  $I = \{1, 11, \dots, 1^n\}$ .  
$$P(h \mid *1) = \frac{\rho(h)}{\xi(*1)} = \frac{\rho(h)}{\rho(h)+\mu(I)(1) \cdot \rho(t) + \dots + \mu(I)(1^n) \cdot \rho(t)} = \frac{\rho(h)}{\rho(h)+\rho(t)}$$

# 你相信末日临近吗?

- ▶ 设想有两个罐子.
- ▶ 罐子  $h$  有 100 个球.
- ▶ 罐子  $t$  有 10000 个球.
- ▶ 你随机选了一个罐子, 从中取出了一个球, 上面数字是 7 号.
- ▶ 你认为这个球来自哪个罐子的可能性更大?
  1. SIA: 罐子  $t$  的可能性更大.

$$\frac{10000}{10000 + 100} > \frac{100}{10000 + 100}$$

2. SSA: 罐子  $h$  的可能性更大.

$$\frac{\rho(h) \frac{1}{100}}{\rho(h) \frac{1}{100} + \rho(t) \frac{1}{10000}} > \frac{\rho(t) \frac{1}{10000}}{\rho(h) \frac{1}{100} + \rho(t) \frac{1}{10000}}$$

**Remark:** SIA 不相信末日临近, 但相信魔鬼创世; SSA 不相信魔鬼创世, 但相信末日临近. Z-协调性通过行为策略引入的因果链条, 避免了观察者分布的随机性假设, 使得分配的信念依赖于行为策略导出的概率分布, 而不是观察者的数量, 从而避免了末日论证的挑战. 优于 SIA 和 SSA, 也优于协调性, 既不会相信魔鬼创世, 也不会相信末日临近.

## Outcome-Equivalence

For any profile  $\sigma/\rho$ , we define the *outcome*  $\bar{\sigma}(h)/\bar{\rho}(h)$  to be the multiplicative product of all the chance probabilities and move probabilities over history  $h$  when all players choose their moves according to  $\sigma_i/\rho_i$ .

$$\sigma_i(h) := \sum_{s_i \sim h} \sigma_i(s_i)$$

$$\bar{\sigma}(h) := \prod_{i \in N \cup \{c\}} \sigma_i(h)$$

$$\bar{\rho}(h) := \prod_{k=0}^{\ell(h)} \rho_{P(h_{<k})}(h_{<k})(h_k) \sigma_c(h)$$

Two (mixed or behavioral) strategies of any player are *outcome-equivalent* iff for every collection of pure strategies of the other players the two strategies induce the same outcome.

### Theorem (Outcome-Equivalent Theorem)

*In games with perfect recall, for any mixed strategy there is an outcome-equivalent behavioral strategy and vice versa.*

## Proof.

“ $\implies$ ”

with perfect recall,

$$\forall h, h' \in I \in \mathcal{I}_i \forall a \in A(I) (\sigma_i(h) = \sigma_i(h') \implies \sigma_i(ha) = \sigma_i(h'a))$$

$$\rho_i(I)(a) := \frac{\sigma_i(ha)}{\sigma_i(h)} \text{ for any } h \in I \in \mathcal{I}_i$$

“ $\iff$ ”

$$\sigma_i(s_i) := \prod_{I \in \mathcal{I}_i} \rho_i(I)(s_i(I))$$

□

# Mixed Extension

## Definition (Games in Strategic Form)

A game in *strategic form* is a triplet  $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ .

- ▶ The sets of players  $N$ .
- ▶ The sets of strategies of the players  $(S_i)_{i \in N}$ .
- ▶ The payoff functions  $u_i : S \rightarrow \mathbb{R}$ .

## Definition (Mixed Extension)

The mixed extension of the strategic game  $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$  is  $\langle N, (\Delta S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ , where the payoff function  $u_i$  in  $\langle N, (\Delta S_i)_{i \in N}, (u_i)_{i \in N} \rangle$  is

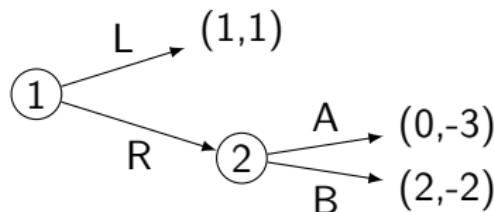
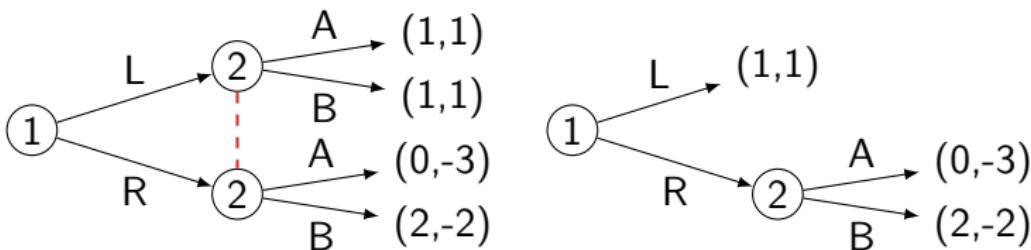
$$u_i(\sigma) = \sum_{s \in S} u_i(s) \sigma(s)$$

Note that  $u_i$  is multilinear.

$$u_i(\lambda \sigma'_i + (1 - \lambda) \sigma''_i; \sigma_{-i}) = \lambda u_i(\sigma'_i; \sigma_{-i}) + (1 - \lambda) u_i(\sigma''_i; \sigma_{-i})$$

## Moving between the Two Forms of Representation

- ▶ Each extensive form game can be represented in a unique way as a game in strategic form.
- ▶ There might be several games in extensive form which represent the same strategic form game.
- ▶ Games in strategic form can be interpreted as games in which all players choose simultaneously.
- ▶ Games in strategic form with two players and finite number of strategies can be represented by a matrix.



	A	B
L	1, 1	1, 1
R	0, -3	2, -2

# Translating from Extensive Form to Matrix Form

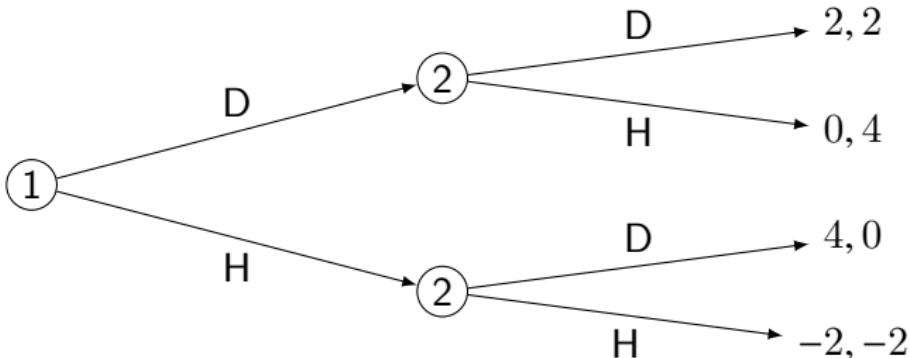
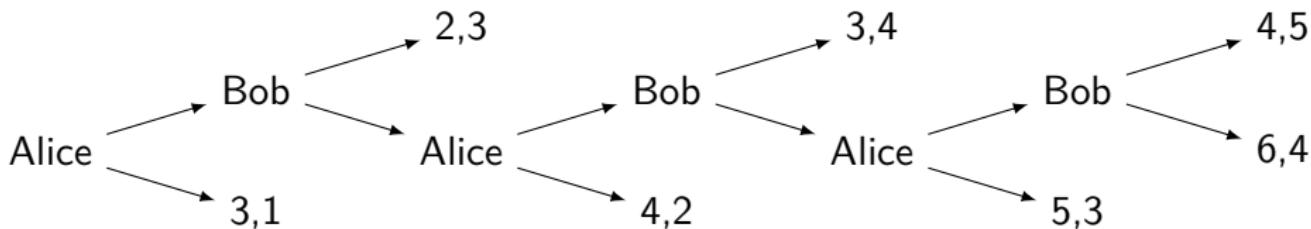


Figure: The Hawk-Dove Game

	(H H,H D)	(H H,D D)	(D H,D D)	(D H,H D)
H	-2,-2	-2,-2	4,0	4,0
D	0,4	2,2	2,2	0,4

# 逆向归纳、不确定性、公共知识



- ▶ 运用逆向归纳求解这个博弈, 我们发现, 只要玩家有机会就会主动结束博弈.
- ▶ 逆向归纳要求玩家的理性是公共知识.
- ▶ 如果我们把不确定性引入到玩家对于彼此的收益的知识中的话, 两位玩家就可能选择继续博弈.
- ▶ 这个不确定性可以表示玩家的理性不是公共知识.
- ▶ 如果一个玩家不确定另一位玩家是否是理性的, 即使后者是理性的, 他/她也可以通过表现的“非理性”而获利.

# Pareto Optimality

- ▶ A strategy profile  $s^*$  *strongly Pareto-dominates*  $s$  iff  
 $\forall i \in N : u_i(s^*) > u_i(s)$ .
- ▶ A strategy profile  $s^*$  *weakly Pareto-dominates*  $s$  iff  
 $\forall i \in N : u_i(s^*) \geq u_i(s)$  and  $\exists i \in N : u_i(s^*) > u_i(s)$ .
- ▶ A strategy profile  $s^*$  is *Pareto optimal* iff

$$\forall i \in N \forall s \in S : u_i(s) > u_i(s^*) \implies \exists j \in N : u_j(s) < u_j(s^*)$$

# Pareto Optimality

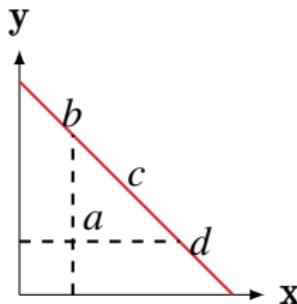


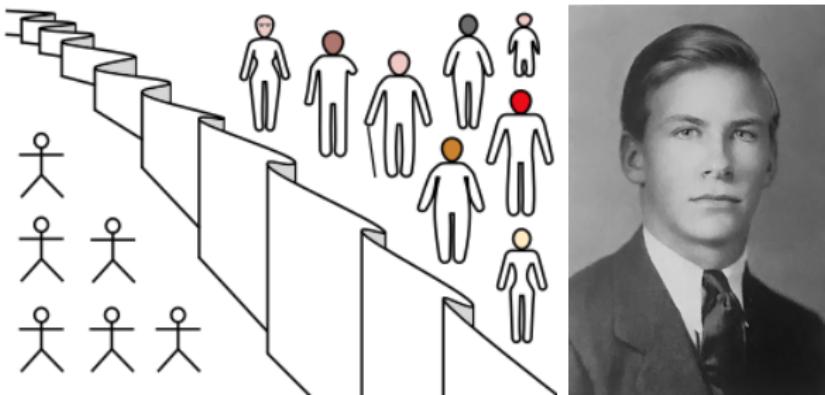
Figure: Pareto Improvement vs Rawls' Difference Principle

- Given an initial situation, a *Pareto improvement* is a new situation where some agents will gain, and no agents will lose.
- A situation is called *Pareto dominated* iff it has a Pareto improvement.
- A situation is called *Pareto optimal* iff no change could lead to improved satisfaction for some agent without some other agent losing.
- The *Pareto frontier* is the set of all Pareto optimal allocations.

not change	change
100, 100	1000, 99

$$0.5 * 1000 + 0.5 * 99 = 549.5 > 100$$

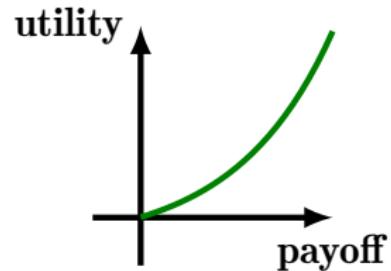
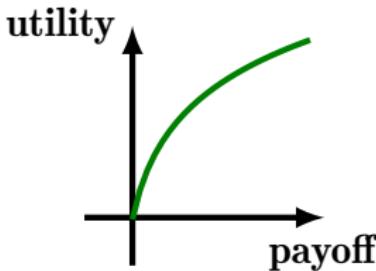
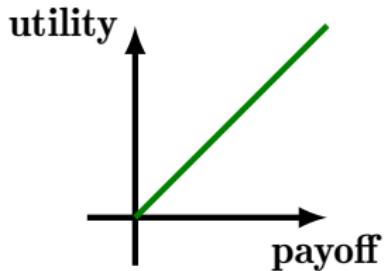
# Rawls' Veil of Ignorance



**Figure:** The reason that the least well off member gets benefited is that it is argued that under the veil of ignorance people will act as if they were risk-averse.

1. Each citizen is guaranteed a fully adequate scheme of basic liberties, which is compatible with the same scheme of liberties for all others;
2. Social and economic inequalities must satisfy two conditions:
  - ▶ attached to positions and offices open to all under conditions of fair equality of opportunity;
  - ▶ to the greatest benefit of the least advantaged members of society (the difference principle).

# Risk-Neutral, Risk-Averse, Risk-Seeking



- ▶ risk-neutral  $\mathbb{E}[u(x)] = u(\mathbb{E}[x])$
- ▶ risk-averse  $\mathbb{E}[u(x)] \leq u(\mathbb{E}[x])$
- ▶ risk-seeking  $\mathbb{E}[u(x)] \geq u(\mathbb{E}[x])$
- ▶ certainty equivalent  $CE(x) := u^{-1}(\mathbb{E}[u(x)])$

## Remark

- ▶ Preferences depend on the agent's reference point: current wealth.
- ▶ For gains, they are risk averse.
- ▶ For losses, they are risk seeking.
- ▶ Losses are (about) twice as bad as gains.

# 圣彼得堡悖论

- ▶ 掷一枚均匀硬币，直到首次出现正面为止，如果所需的次数为  $n$ ，则奖金为  $2^n$  元。
- ▶ 你愿意为这个赌博支付多少钱？

$$\mathbb{E}[u(W)] = \sum_n u(2^n)P(T^{n-1}H) = \sum_n 2^n \frac{1}{2^n} = \infty$$

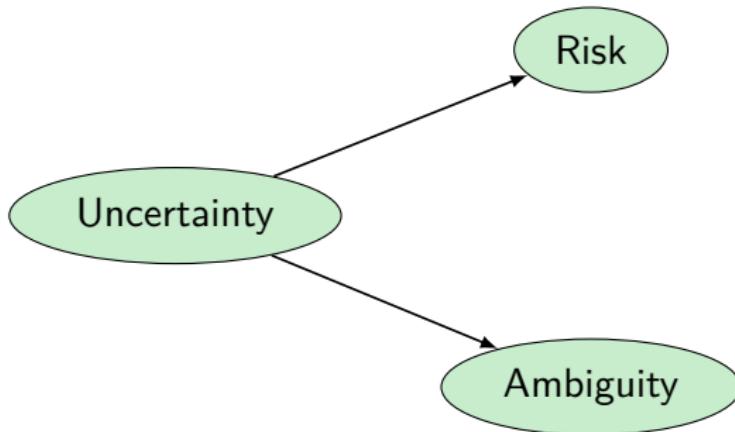
- ▶ 如果你的风险中性的  $u(x) = x$ ，则你会倾尽所有参与这个赌博。
- ▶ 如果你是风险规避的，比如  $u(x) = 4\sqrt{x}$ ，则你只会支付 5.86 元。

$$\mathbb{E}[u(W)] = \sum_n u(2^n)P(T^{n-1}H) = \sum_n 4\sqrt{2^n} \frac{1}{2^n} = \frac{4}{\sqrt{2}-1}$$

$$u(x) = \mathbb{E}[u(W)] \implies x = \left(\frac{1}{\sqrt{2}-1}\right)^2 \approx 5.86$$

- ▶ 如果你是罗尔斯，你只会支付  $< 2$  元。

$$\text{Uncertainty} = \text{Risk} + \text{Ambiguity}$$



1. Risk-sensitive Agent
2. Ambiguity-sensitive Agent

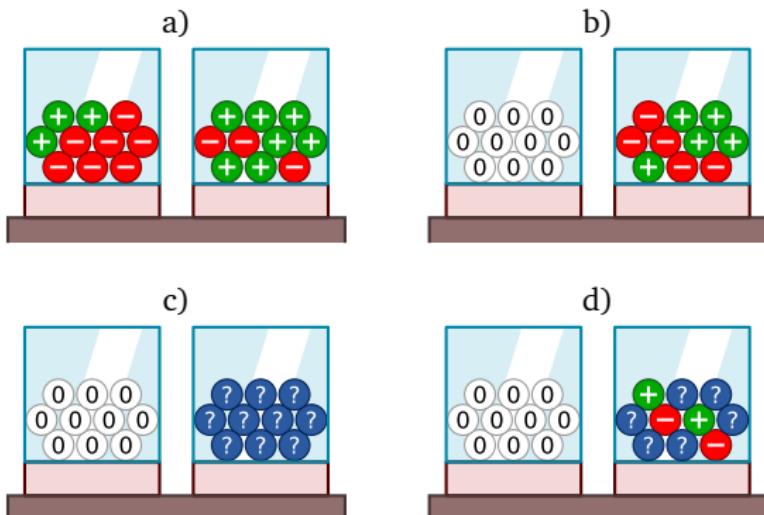
## Risk-sensitive Agent

$$Q_{\text{risk}}(a) \coloneqq \sum_{e \in \mathcal{E}} \nu(e \mid a) u(e) + \lambda \sum_{e \in \mathcal{E}} \nu(e \mid a) [u(e) - \bar{u}(e)]^2$$

where  $\bar{u}(e) := \sum_{e \in \mathcal{E}} v(e | a)u(e)$  is the mean reward.

1.  $\lambda < 0$  risk-averse
  2.  $\lambda > 0$  risk-seeking
  3.  $\lambda = 0$  risk-neutral

Payoffs	
-	= -1
0	= 0
+	= +1
?	= unknown

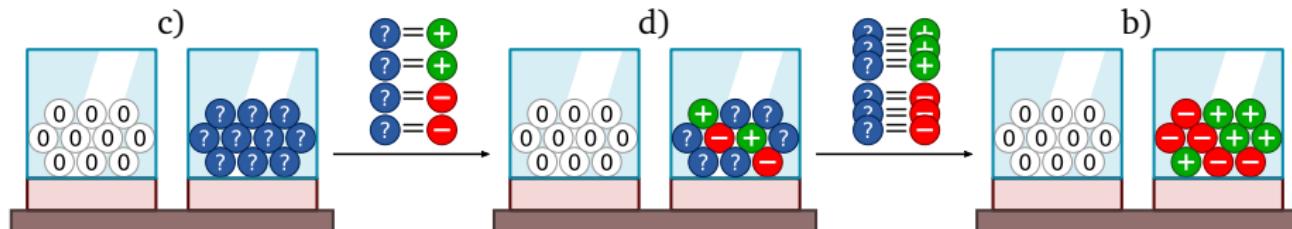


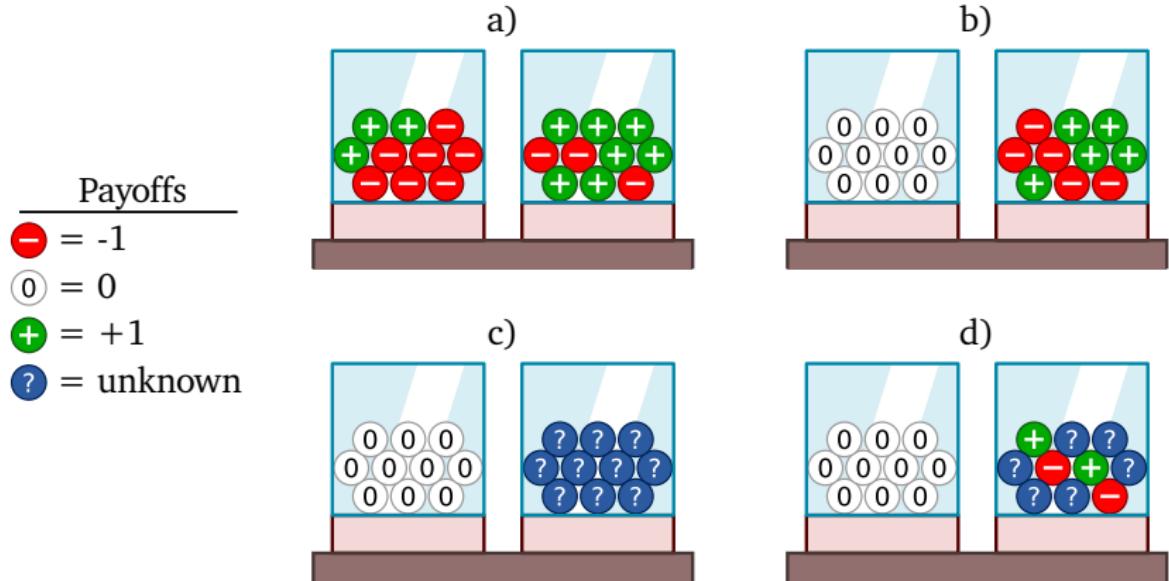
# Ambiguity-sensitive Agent

$$Q_{\text{amb}}(a) := \min_{w \in \Delta(\mathcal{M})} \sum_{v \in \mathcal{M}} w(v) \sum_{e \in \mathcal{E}} v(e | a) u(e)$$

- ▶ For example, an ambiguity-averse agent favours pessimistic interpretations by adopting the rule of selecting the worst-case model.

**Remark:** Learning reduces ambiguity to risk. Starting from fully ambiguous contents (c) and revealing the payoffs of the blue marbles could reduce the contents to a mixture between risk and ambiguity (d) or even to full risk (b).





Agent	Case			
	a	b	c	d
Expected-utility	right	indiff.	<i>undef.</i>	<i>undef.</i>
Risk-averse	right	<i>left</i>	<i>undef.</i>	<i>undef.</i>
Bayes-optimal, with prior	right	indiff.	<i>indiff.</i>	<i>indiff.</i>
Risk-averse, with prior	right	<i>left</i>	left	left
Ambiguity-averse	right	indiff.	left	left

- Pessimism: Maximin rule (Rawls)  $\operatorname{argmax}_{a \in \mathcal{A}} \min_{e \in \mathcal{E}} u(ae)$

	$e_1$	$e_2$	$e_3$	$e_4$
$a_1$	15	0	0	2
$a_2$	-1	4	3	7
$a_3$	6	4	14	1
$a_4$	5	6	4	3

- Optimism: Maximax rule  $\operatorname{argmax}_{a \in \mathcal{A}} \max_{e \in \mathcal{E}} u(ae)$
- Optimism-Pessimism rule (Hurwicz)

$$\operatorname{argmax}_{a \in \mathcal{A}} \left( \alpha \max_{e \in \mathcal{E}} u(ae) + (1 - \alpha) \min_{e \in \mathcal{E}} u(ae) \right)$$

- Expected utility maximization (utilitarianism)

$$\operatorname{argmax}_{a \in \mathcal{A}} \sum_{e \in \mathcal{E}} P(s | a) u(ae)$$

- Minimax regret rule  $\operatorname{argmin}_{a \in \mathcal{A}} \max_{e \in \mathcal{E}} (\max_{a \in \mathcal{A}} u(ae) - u(ae))$
- Maximin safety rule  $\operatorname{argmax}_{a \in \mathcal{A}} \min_{e \in \mathcal{E}} (u(ae) - \min_{a \in \mathcal{A}} u(ae))$

**Remarks:** requires interpersonal comparison of utility.

# Example

	$e_1$	$e_2$	$e_3$	maximin	maximax	minimax regret	Hurwicz $\alpha = 0.5$	EU
$a_1$	3 <sub>0</sub>	16 <sub>6</sub>	15 <sub>0</sub>	3	16	6	9.5	$\frac{34}{3}$
$a_2$	1 <sub>2</sub>	19 <sub>3</sub>	13 <sub>2</sub>	1	19	3	10	11
$a_3$	2 <sub>1</sub>	22 <sub>0</sub>	8 <sub>7</sub>	2	22	7	12	$\frac{32}{3}$

	utility		regret		safety		optimal
	$e_1$	$e_2$	$e_1$	$e_2$	$e_1$	$e_2$	
$a_1$	1	9	3	0	0	5	maximax utility
$a_2$	3	6	1	3	2	2	maximin safety
$a_3$	2	7	2	2	1	3	minimax regret
$a_4$	4	4	0	5	3	0	maximin utility

## Milnor's Axioms [Mil54]

	uniform utilit.	maximin	Hurwicz	minimax regret
ordering	⊕	⊕	⊕	⊕
symmetry	⊕	⊕	⊕	⊕
str. domination	⊕	⊕	⊕	⊕
continuity	○	⊕	⊕	⊕
linearity	○	○	⊕	○
row adjunction	⊕	⊕	⊕	
col. linearity	⊕			⊕
col. duplication		⊕	⊕	⊕
convexity	○	⊕	⊕	⊕
special row. adj.	○	○	○	⊕

Table: ○ 表示这个标准满足这条公理. ⊕ 标记的公理刻画这个标准.

## Milnor's Axioms

1. **ordering:** complete, transitive.
2. **symmetry:** This ordering is independent of the numbering of the rows and columns.
3. **strong domination:**  $\forall k [u(a_i, e_k) > u(a_j, e_k)] \implies a_i > a_j$ .
4. **continuity:**  $u^{(n)}(a_i, e_j) \xrightarrow{n \rightarrow \infty} u(a_i, e_j)$  &  $\forall n \left[ a_i^{(n)} > a_j^{(n)} \right] \implies a_i \geq a_j$
5. **linearity:** The ordering relation is not changed if the matrix  $u(a_i, e_j)$  is replaced by  $u'(a_i, e_j) = \alpha u(a_i, e_j) + \beta$  for  $\alpha > 0$ .
6. **row adjunction:** The ordering between the old rows is not changed by the adjunction of a new row.
7. **column linearity:** The ordering is not changed if a constant is added to a column.
8. **column duplication:** The ordering is not changed if a new column, identical with some old column, is adjoined to the matrix.
9. **convexity:**  $a_i \sim a_j$  &  $a_k = \frac{1}{2}(a_i + a_j) \implies a_k \geq a_i$ .
10. **special row adjunction:** The ordering between the old rows is not changed by the adjunction of a new row, providing that no component of this new row is greater than the corresponding components of all old rows.

# Example

功利主义 vs 罗尔斯<sup>16</sup>

- ▶ Two societies. 1000 people. 100 workers.
- ▶ The workers each receive 1 unit of utility while the others get 90 units each.
- ▶ The average utility is  $10\% \cdot 1 + 90\% \cdot 90 = 81.1$
- ▶ In the second society, everyone take a fair turn at being a worker. This causes everyone to realize the same utility of 50 units.
- ▶ The average utility is  $100\% \cdot 50$ .
- ▶ The utilitarianism would count the first society as more just, but Rawls would favor the second.

---

<sup>16</sup>海萨尼: 最大化最小原则能够成为道德的基础吗? — 对罗尔斯理论的批判. 1975.

# 个人 vs 社会 —— 个人行为正当性的标准

帕累托 vs 卡尔多-希克斯

## Problem (问题)

社会是由人组成的，每个人的行为都会影响到他人的利益。那么，应该用什么标准判断个人的行为是否正当？

- ▶ **帕累托最优:** 一种状态被称为帕累托最优状态，如果不存在另一种状态能使得没有任何人的境况变坏同时至少有一个人的境况变得更好。  
— 除非“损人”，否则不可能“利己”。
- ▶ **卡尔多-希克斯 (Kaldor-Hicks) 标准:** 如果一种变革使得受益者的所得足以弥补受损者的所失，这种变革就是一个卡尔多-希克斯改进。  
如果受损者得到实际的补偿，就是帕累托改进。

**Remark:** 帕累托最优可能意味着收入分配的不公平；极端地，一个人得到所有收入，另一个人一无所有，也是一个帕累托最优。

- ▶ 自愿的交易一定是一个帕累托改进（假定没有欺诈）。
- ▶ 卡尔多-希克斯标准即“财富最大化”。

## 合作与组织

- ▶ 当个体在一起工作创造的价值大于独立工作创造的价值之和, 合作就是一个帕累托改进;
- ▶ 当个体在组织中获得的价值大于独立获得的价值时, 加入组织是一个帕累托改进.
  
- ▶ 自由结婚对夫妻双方是一个帕累托改进;
- ▶ 买卖婚姻对买卖双方是一个帕累托改进;
- ▶ 协议离婚对夫妻双方是一个帕累托改进;
- ▶ 但离婚对其他利益相关者 (如父母和儿女) 可能不是一个帕累托改进.

## Examples: 卡尔多-希克斯标准

### Example1:

- ▶ 考虑两种情形：
  1. 某店主暴力捣毁竞争对手的门店，使后者不能营业；
  2. 某店主以更低的价格和更优良的服务将竞争对手打垮.
- ▶ 为什么法律允许第二种情形?
  - 社会所得大于所失，是一个卡尔多-希克斯改进.

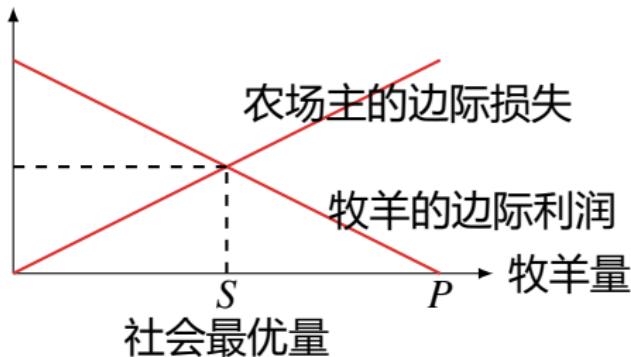
### Example2:

- ▶ 一家化工厂与家属区一墙之隔，为了上班方便，居民在化工厂的墙上挖了一个洞。一天，家属区的一个小孩钻洞进工厂玩，找到一瓶化学液体并点燃烧伤了自己。
- ▶ 假设化工厂补墙成本为  $c$ ，如果不补墙，发生事故的概率为  $p$ ，损失为  $l$ ，且  $c < pl$ .
- ▶ 化工厂需要承担责任吗？

# 外部性、交易成本与科斯定理

- ▶ 个人收益与社会收益：一项活动的社会收益等于决策者个人得到的收益加社会其他成员得到的收益（如养花）；
- ▶ 个人成本与社会成本：社会成本等于决策者的个人承担的成本加社会其他成员承担的成本（如环境污染、交通堵塞）；
- ▶ 如果个人收益（成本）不等于社会收益，就存在外部性。
- ▶ 个人最优与社会最优的不一致意味着有帕累托改进的余地；
- ▶ 如何将外部性内部化？如何使得个人在边际上承担全部的社会成本、获得全部的社会收益？
- ▶ 法律通过责任的分配和赔偿/惩罚，将个人行为的外部成本内部化，诱导个人选择社会最优的行动。
- ▶ 政府征税、补贴？
- ▶ 科斯定理：只要产权界定是清楚的，如果交易成本为零，外部性可以通过当事人之间谈判解决，帕累托最优可以实现；并且，最终的资源配置与初始的产权安排无关。（公平与效率正交吗？）

## 科斯定理 — 例子



- ▶ 如果产权归农场主, 农场主禁止放牧, 小于社会最优量  $S$ ; 但是, 增加放牧给牧羊人带来的边际利润大于给农场主造成的损失, 牧羊人将有积极性贿赂农场主, 直到放牧量达到  $S$  为止;
- ▶ 如果产权归牧羊人, 牧羊人的利润最大点是  $P$ , 大于社会最优量  $S$ ; 但是, 减少放牧量对牧羊人的边际利润损失小于给农场主节约的边际成本, 农场主将有积极性贿赂牧羊人, 直到放牧量达到  $S$  为止.

# 床垫问题

- ▶ 高速公路上，一张床垫从货车上掉了下来。
- ▶ 谁会停下来移开这张床垫呢？
  - ▶ 车流中较远的人不知道发生了什么。
  - ▶ 刚排到面前的人只想着快速绕过它。
  - ▶ 已经绕过的人更没必要返回移开它。
- ▶ 政治冷漠……

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

**Game Theory**

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

**Eliminating Dominated Strategies**

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret

Minimization

Subgame Perfect Equilibrium

Games with Incomplete

Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

References

1753

# Dominant Strategy Equilibrium

## Definition (Dominant Strategy)

A strategy  $s_i^* \in S_i$  is a *dominant strategy* iff

$$u_i(s_i^*; s_{-i}) \geq u_i(s_i; s_{-i}) \text{ for all } s_i \in S_i \text{ and for all } s_{-i} \in S_{-i}$$

## Definition (Dominant Equilibrium)

A strategy profile  $s^*$  is a *dominant strategy equilibrium* iff for each player  $i \in N$ ,  $s_i^*$  is a dominant strategy.

## Definition (Strictly Dominated Strategy)

A strategy  $s_i$  is called strictly dominated by  $s'_i$  iff

$$\forall s_{-i} : u_i(s'_i; s_{-i}) > u_i(s_i; s_{-i})$$

## Definition (Weakly Dominated Strategy)

A strategy  $s_i$  is called weakly dominated by  $s'_i$  iff

$$\forall s_{-i} : u_i(s'_i; s_{-i}) \geq u_i(s_i; s_{-i})$$

and

$$\exists s'_{-i} : u_i(s'_i; s'_{-i}) > u_i(s_i; s'_{-i})$$

## Definition (Iterated Dominance Equilibrium)

An *iterated dominance equilibrium* is a strategy profile  $s^* \in S$  obtained by iteratively ruling out dominated strategies from each player until only one strategy remains for each player.

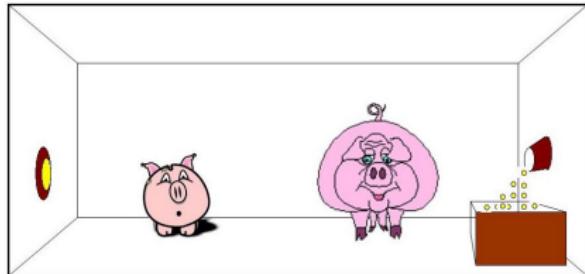
- The result of iterative elimination of strictly dominated strategies is unique, i.e. independent of the elimination order.
- with weakly dominated strategies, the order of elimination matters.

	$c_1$	$c_2$	$c_3$
$r_1$	1, 1	0, 1	3, 1
$r_2$	1, 0	2, 2	1, 3
$r_3$	1, 3	3, 1	2, 2

1. $r_2 \times$	
2. $c_2 \times$	
3. $r_3 \times$	
4. $c_3 \times$	
5. $\begin{array}{c c} & c_1 \\ \hline r_1 & 1, 1 \end{array}$	

1. $r_2 \times$	
2. $c_2 \times; c_3 \times$	
3. $\begin{array}{c c} & c_1 \\ \hline r_1 & 1, 1 \\ r_3 & 1, 3 \end{array}$	

# 智猪博弈 — 反复剔除劣策略均衡 ≠ 占优均衡



- ▶ 小猪是理性的

$\text{Rational}(s)$

- ▶ 大猪知道小猪是理性的

$K_b \text{ Rational}(s)$

	按	等
按	4, 0	3, 3
等	7, -1	0, 0

	按	等
按	3, 3	
等	0, 0	

	按	等
按	3, 3	

# 反复剔除劣策略与 $n$ -阶理性

	$c_1$	$c_2$	$c_3$	$c_4$
$r_1$	5, 10	0, 11	1, 20	10, 10
$r_2$	4, 0	1, 1	2, 0	20, 0
$r_3$	3, 2	0, 4	4, 3	50, 1
$r_4$	2, 93	0, 92	0, 91	100, 90

	$c_1$	$c_2$	$c_3$
$r_1$	5, 10	0, 11	1, 20
$r_2$	4, 0	1, 1	2, 0
$r_3$	3, 2	0, 4	4, 3
$r_4$	2, 93	0, 92	0, 91

	$c_1$	$c_2$	$c_3$
$r_1$	5, 10	0, 11	1, 20
$r_2$	4, 0	1, 1	2, 0
$r_3$	3, 2	0, 4	4, 3

	$c_2$	$c_3$
$r_1$	0, 11	1, 20
$r_2$	1, 1	2, 0
$r_3$	0, 4	4, 3

	$c_2$	$c_3$
$r_2$	1, 1	2, 0
$r_3$	0, 4	4, 3

- 0-order     $c_4 \times$    Rational( $c$ )
- 1-order     $r_4 \times$     $K_r$ Rational( $c$ )
- 2-order     $c_1 \times$     $K_cK_r$ Rational( $c$ )
- 3-order     $r_1 \times$     $K_rK_cK_r$ Rational( $c$ )
- 4-order     $c_3 \times$     $K_cK_rK_cK_r$ Rational( $c$ )
- 5-order     $r_3 \times$     $K_rK_cK_rK_cK_r$ Rational( $c$ )

	$c_2$
$r_2$	1, 1
$r_3$	0, 4

	$c_2$
$r_2$	1, 1

# 诸葛亮的《空城计》何以有效?



- ▶ 诸葛亮谨慎.  $C(z)$
- ▶ 司马懿知道诸葛亮谨慎.  $K_s C(z)$
- ▶ 诸葛亮知道司马懿知道诸葛亮谨慎.  $K_z K_s C(z)$
- ▶ 司马懿不知道诸葛亮知道司马懿知道诸葛亮谨慎.  $\neg K_s K_z K_s C(z)$

## 老狐狸与小狐狸

- ▶ 老狐狸看到满载而归的渔夫驾车经过，于是躺在路边装病。
  - ▶ 渔夫看老狐狸可怜，就让它搭个便车。
  - ▶ 老狐狸悄悄地把鱼一条一条地扔到路边草丛里，然后一跃而下吃鱼去了。
  - ▶ 小狐狸问老狐狸是如何得到这么多鱼的，老狐狸“**如实相告**”。
  - ▶ 第二天，小狐狸也学着躺在路边装病，渔夫愤怒地把它打死了。
- 
- ▶  $\neg K_{\text{渔}} A$
  - ▶  $K_{\text{老}} \neg K_{\text{渔}} A$
  - ▶  $[A]K_{\text{渔}} A$
  - ▶  $K_{\text{老}} [A]K_{\text{渔}} A$
  - ▶  $\neg K_{\text{小}} A$
  - ▶  $[A]K_{\text{小}} A$
  - ▶  $K_{\text{老}} [A]K_{\text{小}} A$
  - ▶  $\neg K_{\text{小}} [A]K_{\text{渔}} A$
  - ▶  $K_{\text{老}} [A] \neg K_{\text{小}} [A]K_{\text{渔}} A$

# Nash Equilibrium

## Definition (Nash Equilibrium)

*Best response* of player  $i$ :

$$\text{BR}_i(s_{-i}) := \underset{s_i \in S_i}{\operatorname{argmax}} u_i(s_i; s_{-i})$$

The strategy profile  $s^*$  is a *Nash equilibrium* iff

$$\forall i \in N : s_i^* \in \text{BR}_i(s_{-i}^*)$$

## Definition (Mixed Nash Equilibrium)

A mixed strategy profile  $\sigma^*$  is a *mixed Nash equilibrium* iff,

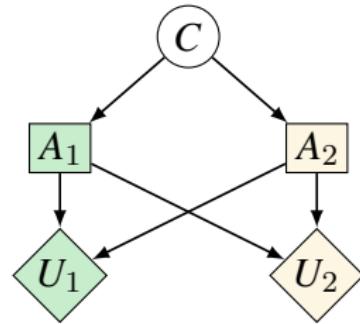
$$\forall i \in N \forall \sigma_i \in \Delta S_i : u_i(\sigma_i^*; \sigma_{-i}^*) \geq u_i(\sigma_i; \sigma_{-i}^*)$$

A mixed strategy Nash equilibrium of a strategic game is a Nash equilibrium of its mixed extension.

# 多重纳什均衡 vs 制度规范

机动车道	左行	右行
左行	1, 1	0, 0
右行	0, 0	1, 1

Table: 协调博弈



卢梭	猎鹿	逮兔
猎鹿	10, 10	1, 8
逮兔	8, 1	5, 5

Table: cheap talk 协商选择帕累托最优的纳什均衡

Remark:

- ▶ 纳什均衡不一定是帕累托最优.
- ▶ 任何制度, 只有构成一个纳什均衡, 才能让人们自觉遵守.
- ▶ 制度、规范的主要功能之一是协调预期, 在多个纳什均衡中筛选出某个特定的纳什均衡.
- ▶ 有效的制度设计, 就是通过对纳什均衡的选择实现帕累托最优.

# 教学、求爱、斗殴、革命

	认真学	不认真学
认真教	2, 2	-1, 1
不认真教	1, -1	0, 0

Table: 教学博弈

	爱	不爱
爱	2, 2	-1, 1
不爱	1, -1	0, 0

Table: 表白博弈

	空手	白刃
空手	2, 2	-1, 1
白刃	1, -1	0, 0

Table: 斗殴博弈

	抗争	沉默
抗争	2, 2	-1, 1
沉默	1, -1	0, 0

Table: 革命博弈

# 混合纳什均衡

	偷懒	不偷懒
监督	1, -1	-1, 2
不监督	-2, 3	2, 2

- ▶ 如果员工偷懒的概率是  $p$ , 那么, 老板监督和不监督的期望收益分别为

$$1 \cdot p + (-1) \cdot (1 - p) = 2p - 1$$

$$(-2) \cdot p + 2 \cdot (1 - p) = 2 - 4p$$

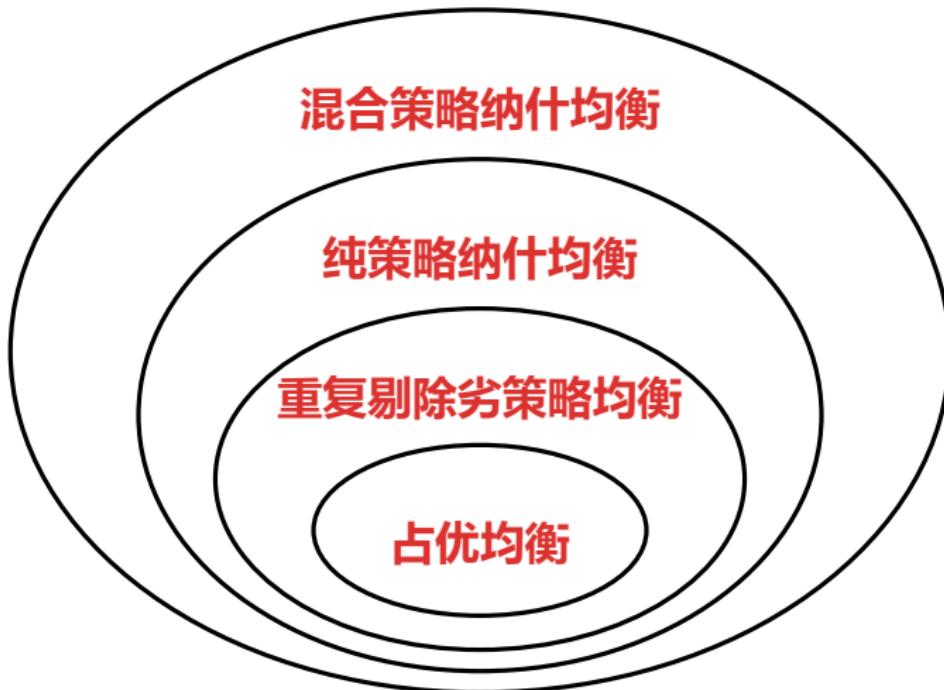
- ▶ 如果老板监督的概率是  $q$ , 那么, 员工偷懒和不偷懒的期望收益分别为

$$(-1) \cdot q + 3 \cdot (1 - q) = 3 - 4q$$

$$2 \cdot q + 2 \cdot (1 - q) = 2$$

- ▶ 混合纳什均衡是  $(\frac{1}{4}, \frac{1}{2})$ , 老板以  $\frac{1}{4}$  的概率监督, 员工以  $\frac{1}{2}$  的概率偷懒.

# 几个均衡之间的关系



# 韭菜的自我修养

	正面 $y$	反面 $1 - y$
正面 $x$	3, -3	-2, 2
反面 $1 - x$	-2, 2	1, -1

Table: 正反匹配

- ▶ 行 (散户) 以  $x$  出正面 (做多),  $1 - x$  出反面 (做空)
- ▶ 列 (庄家) 以  $y$  出正面 (拉升股价),  $1 - y$  出反面 (打压股价)
- ▶ 行的期望收益:

$$U_r = 3xy + 1(1 - x)(1 - y) - 2x(1 - y) - 2(1 - x)y = (8x - 3)y - 3x + 1$$

- ▶ 列的期望收益:  $U_c = -U_r$
- ▶ 列想让  $U_r < 0$ 
  1. 当  $x > \frac{3}{8}$  时,  $y < \frac{3x-1}{8x-3}$ . 所以, 只要  $y < \frac{2}{5}$ , 就有  $U_r < 0$ .
  2. 当  $x < \frac{3}{8}$  时,  $y > \frac{3x-1}{8x-3}$ . 所以, 只要  $y > \frac{1}{3}$ , 就有  $U_r < 0$ .
- ▶ 所以, 只要  $\frac{1}{3} < y < \frac{2}{5}$ , 就有  $U_r < 0$ . 即, 行亏钱, 列赚钱.

# Kakutani Fixpoint Theorem

## Theorem (Kakutani Fixpoint Theorem)

Given a non-empty compact convex set  $X \subset \mathbb{R}^n$  and a multi-valued function  $f : X \rightrightarrows X$ , if

1. for all  $x$ , the set  $f(x)$  is convex,
2. for all sequences  $(x_i, y_i)$  s.t.  $x_i \in X$  and  $y_i \in f(x_i)$ ,

$$\lim_{i \rightarrow \infty} (x_i, y_i) = (x, y) \implies y \in f(x)$$

then  $\exists x \in f(x)$ .

## Theorem (Existence of Mixed Nash Equilibrium)

Every finite strategic game has a mixed Nash equilibrium.

## Proof.

Let  $X := \prod_{i \in N} \Delta S_i$  and  $f(\sigma) := \prod_{i \in N} \text{BR}_i(\sigma_{-i})$ . □

## Theorem

A mixed strategy profile  $\sigma^*$  is a mixed Nash equilibrium iff,

$$\forall i \in N \forall s_i \in S_i : u_i(\sigma_i^*; \sigma_{-i}^*) \geq u_i(s_i; \sigma_{-i}^*)$$

## Proof.

$$u_i(\sigma_i; \sigma_{-i}) = \sum_{s_i \in S_i} u_i(s_i; \sigma_{-i}) \sigma_i(s_i)$$

□

## Theorem

For a finite strategic game,

$$\text{supp}(\sigma_i^*) \subset \text{BR}_i(\sigma_{-i}^*)$$

## Proof.

Suppose there exists an  $s'_i \in \text{supp}(\sigma_i^*)$  s.t.

$$u_i(s'_i; \sigma_{-i}^*) < u_i(\sigma_i^*; \sigma_{-i}^*)$$

then

$$\sum_{s_i \in S_i} u_i(s_i; \sigma_{-i}^*) \sigma_i^*(s_i) < u_i(\sigma_i^*; \sigma_{-i}^*)$$

Contradiction!



- ▶ Every action in the support of any player's equilibrium mixed strategy yields that player the same payoff.
- ▶ If the set of actions of some player is not finite the result needs to be modified. In this case,  $\sigma^*$  is a mixed strategy Nash equilibrium iff
  1. for every player  $i$  no action in  $S_i$  yields, given  $\sigma_{-i}^*$ , a payoff to player  $i$  that exceeds his equilibrium payoff, and
  2. the set of actions that yield, given  $\sigma_{-i}^*$ , a payoff less than his equilibrium payoff has  $\sigma_i^*$ -measure zero.

# 沉默的目击者 — 从纳什均衡看旁观者效应

- $n$  个目击者围观 Kitty 被虐杀. 报警的成本是  $b$ . 如果有人报警, Kitty 会得救, 每个目击者会获得效用  $a$ ; 如果没人报警, 效用是 0.

	有他人报警 $1 - (1 - p)^{n-1}$	没他人报警 $(1 - p)^{n-1}$
报警 $p$	$a - b$	$a - b$
不报警 $1 - p$	$a$	0

- 当  $a < b$  时, 所有人不报警是纳什均衡.  
► 当  $a > b$  时, 有一个人报警、其他人不报警是纳什均衡.  
► 考虑对称的混合纳什均衡, 即所有参与人报警的概率相等, 设为  $p$ .  
► 某参与人的期望收益:

$$Q(p) := (a - b)p + \left( a(1 - (1 - p)^{n-1}) + 0(1 - p)^{n-1} \right) (1 - p)$$
$$\frac{dQ(p)}{dp} = 0 \implies p = 1 - \left(\frac{b}{a}\right)^{\frac{1}{n-1}}$$

- $n$  个人中至少有一个人报警的概率为

$$P(n) := 1 - (1 - p)^n = 1 - \left(\frac{b}{a}\right)^{\frac{n}{n-1}}$$

但  $\frac{dP(n)}{dn} = \frac{\left(\frac{b}{a}\right)^{\frac{n}{n-1}} \ln \frac{b}{a}}{(n-1)^2} < 0$ , 人数越多, 至少有一个人报警的概率越小.

# 走出囚徒困境 — 奖惩 — 利维坦

	合作	背叛
合作	-1, -1	-4, 0
背叛	0, -4	-3, -3

Table: 囚徒博弈

	合作	背叛
合作	-1, -1	-4, 0 - x
背叛	0 - x, -4	-3 - x, -3 - x

Table: 带惩罚的囚徒博弈  $0 - x < -1$



# 走出囚徒困境 — 奖惩 — 作为激励机制的等级制度

	合作	背叛
合作	-1, -1	-4, 0
背叛	0, -4	-3, -3



	先走	后走
先走	0, 0	2, 1
后走	1, 2	0, 0



## 儒家 — “礼”

“合作”方可得“君子”名分，君子享有优先权。

协调预期、定分止争。

声誉约束。

奖善惩恶、等级制度等可以理解为一种激励机制。

# 风险与均衡

- ▶ 由于纳什均衡要求理性共识和一致预期, 当人们可能犯小小的错误时, 纳什均衡不一定被选择.

	左	右
上	8, 10	-1000, 9
下	7, 6	6, 5

**Table:** 只要李四有千分之一的概率错误地选择“右”, 张三将选择“下”; 如果李四怀疑张三怀疑自己可能犯错误, 李四将选择“右”.

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret

Minimization

Subgame Perfect Equilibrium

Games with Incomplete

Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

References

1753

# 100 只狼和 1 只羊

- ▶ 某个魔法森林里生活着 100 只狼和 1 只羊.
  - ▶ 狼想吃羊, 羊不想被狼吃.
  - ▶ 一只狼如果吃掉一只羊, 就会变成一只羊.
1. 森林里狼和羊的数量会变成多少?
  2. 如果森林里一开始是 99 只狼和 1 只羊呢?

# 囚犯分绿豆

## Problem (谁能活下来?)

五个囚犯轮流从一个装有 100 颗绿豆的袋子里抓绿豆.

1. 他们都绝对理性.
2. 彼此不能交流.
3. 100 颗不必都分完, 但要保证每人至少抓一颗.
4. 他们可以摸出袋子里剩下的绿豆数量.
5. 抓得最多和最少的人将被处死.
6. 他们的原则是先求保命, 再尽可能多杀人.

None!

# 海盗分金

## Problem (海盗分金)

五名海盗 (从最怯懦到最凶残依次是 1 到 5 号) 打算瓜分 100 个金币.  
海盗世界的分配规则如下:

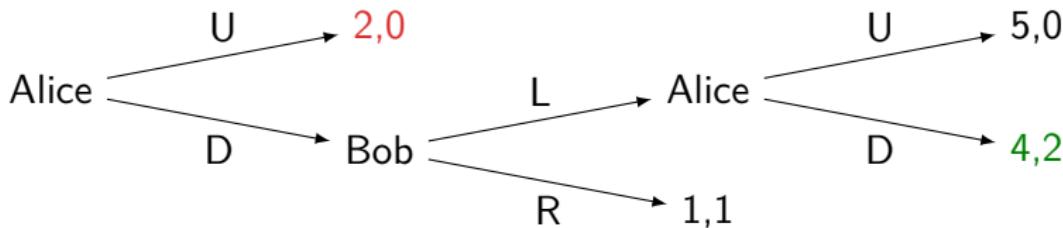
1. 最凶残的一名海盗提出分配方案, 然后所有的海盗 (包括提出方案者本人) 就此方案进行表决;
2. 如果 50% 或以上的海盗赞同此方案, 此方案就获得通过并据此分配;
3. 否则, 提出方案的海盗将被扔到海里, 然后下一名最凶残的海盗又重复上述过程.
4. 每名海盗首先想活命, 其次想贪财, 最后是嗜杀, 如果其他情况相同, 都更倾向于把另一个海盗扔到海里.

1,0,1,0,98

What about more than 200 pirates? Who can survive? 1-200, 201,  $200+2^n$

What about more than half? 0,2,1,0,97 / 2,0,1,0,97

# 不可信的许诺/威胁 vs 可信的承诺



- ▶ 如果 Alice 许诺不选择 U, Bob 会相信吗?
- ▶ 如果 Alice 承诺不选择 U 呢?
  - 比如: 拿出 2\$ 保证金给独立的第三方, 若失信, 则自动转给 Bob.
- ▶ 动态博弈的纳什均衡可能包含不可信的许诺 (或威胁). 即, 事前看是最优的, 事后看不是最优的, 所以不可信.
- ▶ 老父亲威胁女儿, 敢跟古惑仔私奔就断绝父女关系, 但私奔后, 若真的断绝父女关系会损失更大. 所以理性的女儿还是会私奔.
- ▶ 若政府对大财团的威胁不可信, 大财团融资时可能无视风险, 成功了收益是自己的, 失败了成本转嫁给政府社会, 最后“大而不倒”, 导致金融危机.

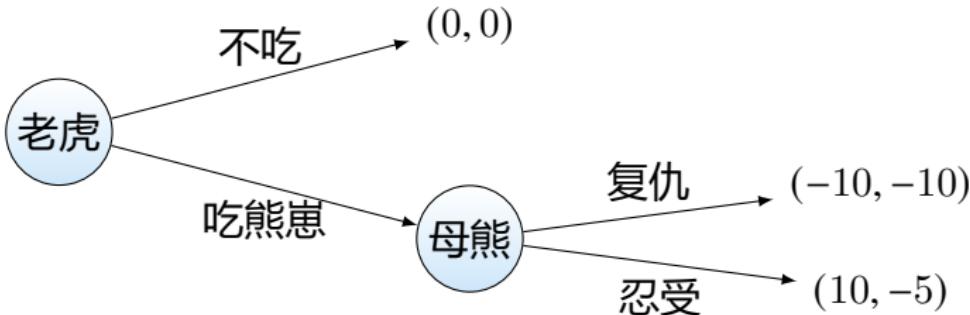
# 承诺

假如企业一开始定价 80, 如果前两个消费者购买了, 企业将有积极性在 50 的价格下向第三个顾客出售. 预期到这一点, 前两个顾客将不会购买. 如果企业承诺, 任何降价的差额将返还顾客, 前两个顾客则会购买.

产量	价格	收入
1	100	100
2	80	160
3	50	150
4	30	120

- ▶ 承诺意味着限制自己的自由: 选择少反而对自己好.
  - “破釜沉舟”.
- ▶ 增加对方的选择也可以看作一种“承诺”(或“反承诺”).
  - 围城只围三面.
- ▶ 承诺有成本.
  - 永不降价, 降价退差额! 假一赔十!
- ▶ 天价彩礼: 不离婚的承诺.
- ▶ 为什么画家死了画会升值?
- ▶ 民主、法治是政府对人民的一种承诺.
- ▶ 民法“民不告、官不究”, 而刑法实行“公诉制度”是一种承诺.

# 情感的进化解释



- ▶ 情感与理性是对立的吗?
- ▶ (老虎吃掉熊幼崽, 母熊默默忍受) 是子博弈完美均衡.
- ▶ 即使母熊事先威胁“我必复仇!”, 老虎也不会相信.
- ▶ 但如果母熊出离愤怒, 激情复仇, 那么就相当于其默默忍受的效用值不是 -5, 而是比 -10 还小.
- ▶ 会愤怒的熊的种群存活率更高.
- ▶ 愤怒的情感让威胁变成了可信的承诺.

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret

Minimization

Subgame Perfect Equilibrium

Games with Incomplete

Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

References 1753

# 无限重复博弈

	合作	背叛
合作	$T, T$	$S, R$
背叛	$R, S$	$P, P$

Table: 囚徒博弈:  $R > T > P > S$  且  $T + T > R + S$ .

1. All-D 策略: 总是背叛.
2. All-C 策略: 总是合作.
3. 合作-背叛交替进行.
4. 以牙还牙 tit-for-tat TFT 策略: 从合作开始, 之后每次选择对方前一阶段的行动.
5. 冷酷 grim 策略: 从合作开始, 直到一方背叛, 然后永远背叛.
6. 宽容的冷酷策略: 如果对方背叛, 先惩罚几次, 然后再恢复合作.
7. 宽容的以牙还牙: 永远以合作的态度来回报对方的合作. 当遇到背叛时, 以某一概率与对方进行合作.
8. 赢定输移 win stay, lose shift 策略: 如果我们上一轮合作, 那么合作; 如果上一轮都背叛, 那么以一定概率合作; 如果上一轮我合作你背叛或你合作我背叛, 则背叛.

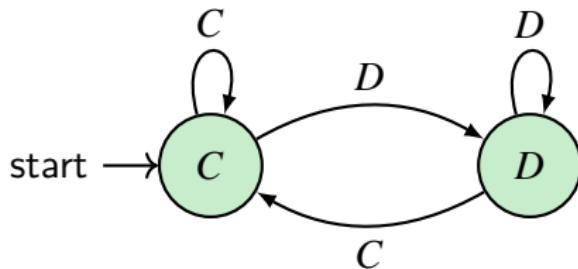


Figure: 以牙还牙

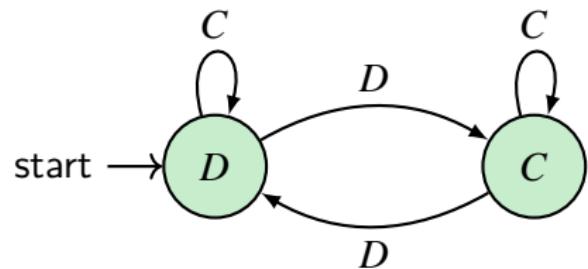


Figure: 一报还一报

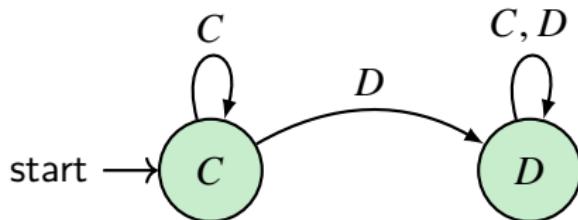


Figure: 冷酷

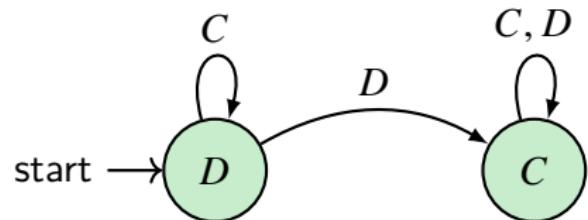


Figure: 喜鹊

- ▶  $V(\text{ALL-D}, \text{ALL-D}) = P + \gamma P + \gamma^2 P + \cdots = P \frac{1}{1-\gamma}$
- ▶  $V(\text{TFT}, \text{TFT}) = T + \gamma T + \gamma^2 T + \cdots = T \frac{1}{1-\gamma}$
- ▶  $V(\text{ALL-D}, \text{TFT}) = R + \gamma P + \gamma^2 P + \cdots = R + P \frac{\gamma}{1-\gamma}$
- ▶  $V(\text{ALL-C}, \text{grim}) = T + \gamma T + \gamma^2 T + \cdots = T \frac{1}{1-\gamma}$
- ▶  $V(\text{ALL-D}, \text{grim}) = R + \gamma P + \gamma^2 P + \cdots = R + P \frac{\gamma}{1-\gamma}$

如果  $T \frac{1}{1-\gamma} \geq R + P \frac{\gamma}{1-\gamma}$ , 即  $\gamma \geq \frac{R-T}{R-P}$ , 则对于 grim 策略来说, 合作就是子博弈完美均衡.

## 无名氏定理 (Folk Theorem)

在无限次重复博弈中, 如果每个参与人都对未来足够重视 (贴现因子足够大), 那么, 任何程度的合作都可以作为一个子博弈完美均衡得到. 这里的“合作程度”指整个博弈中合作出现的频率.

# 动物界的合作

- ▶ 鱼类会使用 TFT 策略: 当两条鱼接近入侵者时, 如果一条想尾随在后, 走在前面的鱼转身向后, 等待另一条跟上, 然后再并行前进.
- ▶ 孔雀鱼甚至可以记住其同伙过去的表现. 如果一次试验中一方背叛, 另一方在第二次的试验中也会背叛.
- ▶ 孔雀鱼倾向于与过去表现出更具合作精神的鱼结伴而行.

# 惩罚

- ▶ 在重复博弈中, 越在乎长远利益, 合作的可能性越大.
- ▶ 背叛行为越容易被观察到, 并且惩罚越可信, 合作的可能性越大.
- ▶ 垄断使得惩罚不可信.
- ▶ 在确定环境中, 惩罚越严厉越有助于合作, 但在不确定环境中, 对方有可能是无心之失, 冷酷策略可能不利于长期合作.

## 联合抵制的社会规范

联合抵制 (Boycott): 每个人都应该诚实; 都有责任惩罚骗过人的人; 不参与惩罚的人应该受到惩罚.

**Remark:** 朋友的朋友是朋友; 朋友的敌人是敌人; 敌人的朋友是敌人.

# 有限重复博弈

	$c_1$	$c_2$	$c_3$
$r_1$	1, 1	5, 0	0, 0
$r_2$	0, 5	4, 4	0, 0
$r_3$	0, 0	0, 0	3, 3

- ▶ 两个纳什均衡:  $(r_1, c_1)$  和  $(r_3, c_3)$
- ▶ 帕累托最优:  $(r_2, c_2)$
- ▶ 如果博弈重复两次, 则 “好合好散, 不欢而散” 的策略 — 如果  $(r_2, c_2)$  则  $(r_3, c_3)$ , 否则  $(r_1, c_1)$  — 可在第一轮博弈中实现帕累托最优.
- ▶ 但是, 如果第一轮遭到背叛后, 第二轮对方重新谈判, 则会使得惩罚不可信. 原因在于多重均衡之间  $(r_3, c_3)$  帕累托优于  $(r_1, c_1)$ .

# 有限重复博弈

	$c_1$	$c_2$	$c_3$	$c_4$	$c_5$
$r_1$	1, 1	5, 0	0, 0	0, 0	0, 0
$r_2$	0, 5	4, 4	0, 0	0, 0	0, 0
$r_3$	0, 0	0, 0	3, 3	0, 0	0, 0
$r_4$	0, 0	0, 0	0, 0	4, 0.5	0, 0
$r_5$	0, 0	0, 0	0, 0	0, 0	0.5, 4

- ▶ 四个纳什均衡:  $(r_1, c_1)$ 、 $(r_3, c_3)$ 、 $(r_4, c_4)$ 、 $(r_5, c_5)$
- ▶ 三个帕累托最优:  $(r_2, c_2)$ 、 $(r_4, c_4)$ 、 $(r_5, c_5)$
- ▶ 如果博弈重复两次, 则可采取如下策略: 如果  $(r_2, c_2)$  则  $(r_3, c_3)$ ; 否则, 被背叛的一方根据自己的情况选择  $r_4$  或  $c_5$ , 如果双方同时背叛, 则  $(r_3, c_3)$ .
- ▶ 此时惩罚变得可信了, 均衡为: 第一轮  $(r_2, c_2)$ ; 第二轮  $(r_3, c_3)$ .

# 单方不完全信息的有限重复博弈

	合作	背叛
合作	3, 3	-1, 4
背叛	4, -1	0, 0

	t1	t2
A 冷酷型 $p$	合作	$X$
A 理性型 $1 - p$	背叛	背叛
B 理性型	$X$	背叛

Table: 囚徒博弈重复两次

- ▶ 参与人 A 有两种可能的类型：“冷酷”型：选择 grim 策略，概率为  $p$ ；“理性”型：可以选择任何策略，概率为  $1 - p$ .
- ▶ 参与人 B 有一种类型：理性型.
- ▶ B 在第一轮选择合作或背叛最后总的期望效用分别为

$$3 \cdot p + (-1) \cdot (1 - p) + 4 \cdot p + 0 \cdot (1 - p) = 8p - 1$$

$$4 \cdot p + 0 \cdot (1 - p) + 0 \cdot p + 0 \cdot (1 - p) = 4p$$

- ▶ 如果  $p \geq \frac{1}{4}$ ，则 B 在第一轮合作.
- ▶ 如果博弈重复  $N \geq 3$  轮，只要  $p \geq \frac{1}{4}$ ，理性型 A 在  $t = 1 \dots N - 2$  轮合作，在最后两轮背叛；B 在前  $N - 1$  轮合作，在最后一轮背叛.

## 双方不完全信息的有限重复博弈

- ▶ 假设双方都有两种可能的类型：“冷酷”型或“理性”型.
- ▶ 如果一开始就选择背叛，暴露了自己是理性型，那么收益最大为 4.
- ▶ 假如对方采取冷酷策略的概率是  $p$ ，则采取冷酷策略的最小期望收益为

$$3 \cdot N \cdot p + (-1 + 0 + 0 + \cdots + 0) \cdot (1 - p) = (3N + 1)p - 1$$

- ▶ 只要博弈次数  $N \geq \frac{5-p}{3p}$ ，则采取冷酷策略.
- ▶ 在完全信息的情况下，根据逆向归纳逻辑， $n$  次重复囚徒博弈有唯一的子博弈完美均衡，即双方总是背叛.
- ▶ 在不完全信息的情况下，只要博弈重复的次数足够长，参与人就有积极性在博弈的早期建立一个“合作”的声誉；一直到博弈的后期，才会选择背叛；并且，背叛的轮数只与  $p$  有关，而与博弈的次数  $N$  无关.
- ▶ 在有限次重复囚徒博弈中，当双方不知道具体次数是公共知识时，将出现贝叶斯-纳什均衡形式的合作。— 无知是福 ☺

# 声誉的积累

$$\begin{aligned} P(\text{君子} \mid \text{好事}) &= \frac{P(\text{君子} \wedge \text{好事})}{P(\text{好事})} \\ &= \frac{P(\text{好事} \mid \text{君子})P(\text{君子})}{P(\text{好事} \mid \text{君子})P(\text{君子}) + P(\text{好事} \mid \neg\text{君子})P(\neg\text{君子})} \end{aligned}$$

- ▶ 人不知而不愠, 不亦君子乎
- ▶ 勿以善小而不为, 勿以恶小而为之
- ▶ 狼来了
- ▶ 烽火戏诸侯
- ▶ “伪君子” 为什么更可恨?

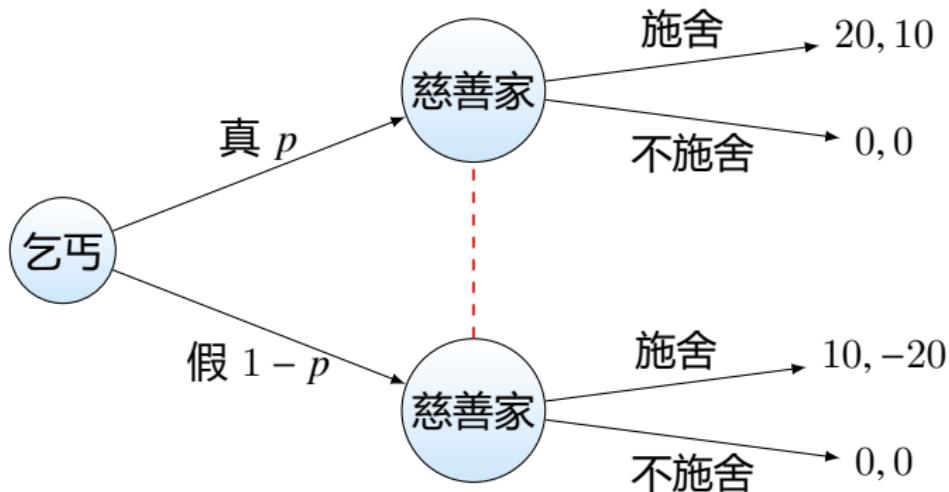
# 非对称信息 (Asymmetric Information)

- ▶ 非对称信息: 交易关系中一方知道而另一方不知道的信息.
  - 卖家知道商品的质量, 而买家不知道.
  - 员工知道自己的能力, 老板不知道.
  - 投保人知道自己的健康状况, 保险公司不知道.
  - 乞讨者知道自己是否是真乞丐, 慈善家不知道.
- ▶ 事前 (ex ante) 非对称信息: 指签约之前存在的非对称信息, 所以, 又称为隐藏信息 (hidden information), 如产品质量.
- ▶ 事后 (ex post) 非对称信息: 指签约之后发生的非对称信息, 又称隐藏行动 (hidden action), 如工人的努力水平.
- ▶ 事前信息不对称导致逆向选择, 劣胜优汰.
- ▶ 事后信息不对称导致道德风险.

## Example (逆向选择)

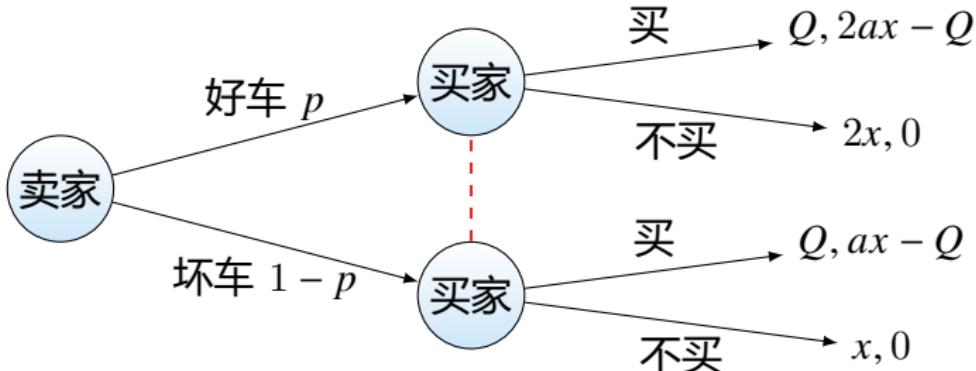
- ▶ 坏车使得好车不能成交. (好人未必好报)
- ▶ 高风险的投保人使得低风险的投保人无法投保.
- ▶ 高水平的学者竞争不过灌水的学者.

# 逆向选择 — 劣币驱逐良币



- ▶ 假乞丐使得真乞丐得不到救济.
- ▶ 和珅: 往粥里掺沙子.
- ▶ 经济适用房不能建太好.

# 逆向选择 — 劣币驱逐良币



- ▶ 假设好车和坏车对卖家的保留价值分别为  $2x$  和  $x$ , 对买家的价值是卖家的  $a$  倍. 成交价为  $Q$ .
- ▶ 为使交易达成, 买家买车的期望收益应不小于不买车的期望收益.  
$$(2ax - Q)p + (ax - Q)(1 - p) \geq 0$$
- ▶ 拥有好车的卖家能接受的最低价是  $Q \geq 2x$ .
- ▶ 这意味着, 买卖双方合意的价格需满足:

$$(1 + p)ax \geq Q \geq 2x \implies p \geq \frac{2}{a} - 1$$

- ▶ 买卖好车是双赢, 但除非好车比例够高, 否则交易无法达成.
- ▶ 除非好人够多, 否则不敢与陌生人交往 ☺

# 如何解决非对称信息的问题?

## 解决非对称信息的市场机制和政府管制

1. 信号显示 (信号传递): 主动传递真实信息. 如: 卖家承诺保修.
2. 信息甄别 (机制设计): 让对方说实话. 如: 保险公司向投保人提供不同的合同选择, 投保人根据自己的状况选择适合自己的合同.
3. 声誉机制: 品牌的价值. 信息越不对称的产品, 其品牌价值越大. 买土豆不太会关心品牌, 但买汽车买咨询服务会非常看重品牌.
4. 适度的政府管制. 但政府管制不当可能破坏声誉机制的有效性 — 管制导致企业预期不稳定, 追求短期行为; 管制导致垄断, 使得市场惩罚不可信; 管制导致腐败, 贿赂官员比贿赂投资者和客户更合算.

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

**Game Theory**

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret Minimization

Subgame Perfect Equilibrium

Games with Incomplete Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

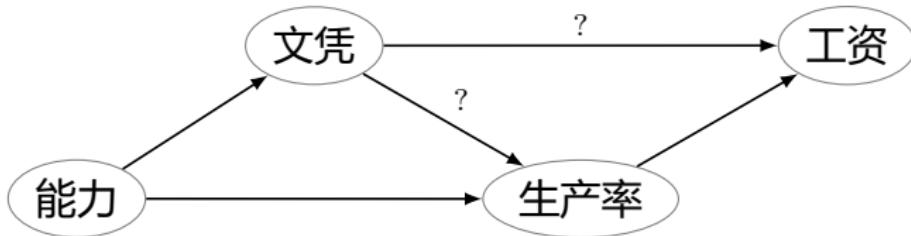
Artificial General Intelligence

What If Computers Could Think?

References 1753

# 文凭的信号传递作用

- ▶ 求职者 50% 可能性是高能力者 50% 可能性是低能力者.
- ▶ 高能力者的生产率是 200, 低能力者 100.
- ▶ 雇主愿付高能力者工资 200, 低能力者 100, 不知道求职者能力高低则付平均工资 150.
- ▶ 假设高能力者受教育成本 40, 低能力者受教育成本 120.
- ▶ 此时, 教育水平可以成为传递能力的信号.
- ▶ 但如果低能力者受教育的成本低于 100, 文凭就无法成为雇主区分能力高低的信号, 雇主愿付的工资就是 150, 也就没人愿上大学了.
- ▶ 因此, 关键是不同类型的人信号传递成本不同; 只有成本差异足够大, 才有可能传递信号.



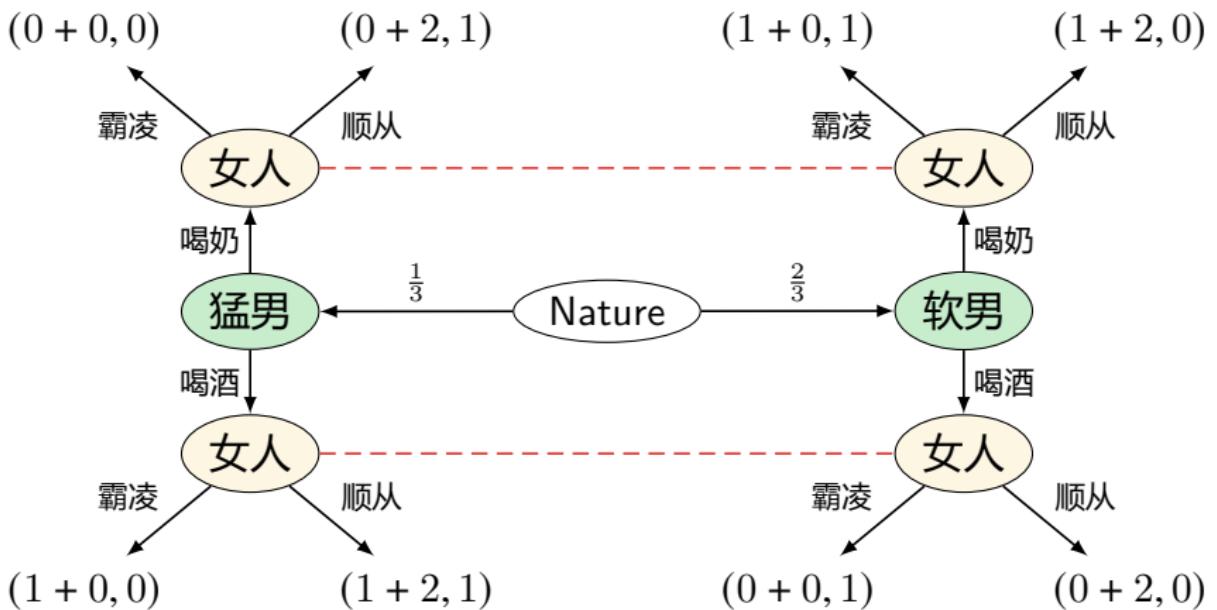
## Spence's Education Game

- ▶ The worker can be one of two types: wise  $\theta_H$  (with probability  $p$ ) or dumb  $\theta_L$  (with probability  $1 - p$ ). Each type can select their own level of education,  $e_H$  or  $e_L$ .
- ▶ The employer is assumed to have two choices. One is to ignore the signal and set  $w^* := p\theta_H + (1 - p)\theta_L$ . The other is to pay a worker  $w_H$  or  $w_L$  based on whether the signal is  $e_H$  or  $e_L$ .
- ▶ The employer's payoff is  $\theta - w$ . The worker's payoff is  $w - e/\theta$ .
- ▶ This game has two equilibria. The first is a **pooling equilibrium**, in which the worker will choose the same level of education regardless of his type, and the employer pays all workers the same amount  $w^*$ .
- ▶ The other is a **separating equilibrium**, in which the worker will choose a different level of education. A low-talent worker will get no education,  $e_L = 0$ . The education chosen by a high-talent worker is set in such a way as to make it unprofitable for either type of worker to mimic the other.

$$w_L - 0/\theta_L \geq w_H - e_H/\theta_L \quad \text{and} \quad w_H - e_H/\theta_H \geq w_L - 0/\theta_H$$

# 信号传递的作用

- ▶ 为什么雄性孔雀的尾巴越长，越受到雌性孔雀的青睐？ — 只有健壮者才能负担得起长尾巴。
- ▶ 高、低质量产品哪个更愿意做广告？ — 广告费是高质量产品企业向市场传递信息的成本。
- ▶ 如何送礼？ — 重要的是送礼对送者的成本，而不是礼物对接受者的价值。
- ▶ 奢侈品、名牌，不是显示产品的质量，而是显示消费者的质量。
- ▶ 中秋为什么送浪费性的月饼？请客吃巨贵的馆子？ — 成本要高于价值，甚至“毁灭”了很大一部分价值更显出对对方的重视。“千里送鹅毛，礼轻仁义重。”
- ▶ 公费请客送礼传递的信息量大打折扣，成本需要翻倍！
- ▶ 为什么领结婚证？ — 离婚分财产。同样，为什么送彩礼？为什么婚礼要铺张？
- ▶ 为什么街头古惑仔纹身？黑社会老大穿西装戴眼镜？
- ▶ 为什么很多繁文缛节没有实质意义的礼仪还要遵守？ — 合作精神



- ▶ Nature 以  $\frac{1}{3}$  的概率指定男人是猛男,  $\frac{2}{3}$  的概率是软男.
- ▶ 猛男喜欢喝酒 (效用 1); 软男喜欢喂奶 (效用 1). 男人喜欢女人顺从 (效用 2); 女人喜欢顺从猛男霸凌软男 (效用 1).
- ▶ 男人如果是猛男, 就喝酒; 如果是软男, 就以  $\frac{1}{2}$  的概率遂心喂奶, 以  $\frac{1}{2}$  的概率违心喝酒.
- ▶ 女人见到男人喂奶就霸凌; 见到男人喝酒就以  $\frac{1}{2}$  的概率霸凌, 以  $\frac{1}{2}$  的概率顺从.

## 信息不完全导致社会规范变迁

- ▶ 如果是完全分离均衡, 每类人的行为都是特定的.
- ▶ 如果是混同均衡, 所有人的行为都是一样的.
- ▶ 如果是准分离 (混同) 均衡, 有些行为传递信息, 有些行为不传递信息.
- ▶ 如果外部因素导致社会由分离均衡转向混同均衡或准分离均衡, 社会规范就会发生变化.

人们对婚前性行为和婚外性行为态度的变化: 在封闭的社会, 婚前和婚外性行为都很容易观察; 在流动的社会, 有些能观察到, 有些不能; 如果被观察到的只是其中的一小部分, 被观察到压力将会减少.

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret

Minimization

Subgame Perfect Equilibrium

Games with Incomplete

Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

References

1753

# Cake Cutting

- ▶ Two Children:
  1. One cuts, the other chooses.
- ▶ Three Children (A,B,C):
  1. A cuts a piece of the cake.
  2. B can leave it, or make it smaller.
  3. C can leave the remaining piece, or make it smaller.

Whoever cuts the cake last, gets the remaining piece.

- ▶ 19世纪中期法国的火车有头等座、二等座和三等座三种不同的车厢.
- ▶ 三种车厢票价相差较大.
- ▶ 头等车厢非常舒适, 二等车厢与三等车厢的区别是前者有顶盖, 后者没有. 这样, 坐三等车厢的旅客就要忍受日晒雨淋的痛苦.
- ▶ 给车厢加一个顶盖的成本微乎其微, 为什么公司不这样做呢?
- ▶ 为了吓唬富人, 故意让穷人受罪.

- ▶ 问题: 有时候, 拥有私人信息的一方有积极性通过一定的行动向另一方传递自己的私人信息, 但有时候他们没有积极性或没有有效的办法传递自己的私人信息.
- ▶ 机制设计: 没有私人信息的一方通过设计不同的分配方案使得有私人信息的一方通过自我选择揭示自己的私人信息.

**Example:** 保险市场中的混同均衡和分离均衡.

1. 保险金 2 万元; 只有第二年得病才可得到赔偿金 10 万元, 第一年得病得不到赔偿.
2. 保险金 7 万元; 无论第一年得病还是第二年得病, 都可得到赔偿金 10 万元.

高风险者选择合同 2; 低风险者选择合同 1.

# 机制设计

## 逆向博弈论

已知理性人会选择最优策略，为了诱导出某种类型的行为，应该设计什么样的博弈？

## 博弈 vs 伦理

- ▶ 康德之流的义务论者可能会认为，博弈对伦理无益。因为，伦理规范是约束你做那些你不想做的事，而博弈论却是讲怎么得到你想要的东西。
- ▶ 休谟等效用主义者却认为，除非伦理理论建议的行为符合每个人的实际利益，否则毫无用处。有效的制度设计，就是通过对纳什均衡的选择实现帕累托最优。

# 拍卖

1. 高价格公开拍卖
2. 降价公开拍卖
3. 高价格密封拍卖
4. 次价格密封拍卖: 与高价格密封拍卖一样, 中标者仍然是出价最高者, 但中标者实际支付的是第二高报价. 次价格密封拍卖能让竞标者有积极性说真话.

## Theorem

次价格密封拍卖中, 说真话是竞标者的占优策略.

**Remark:** 比如: 对某古董, 你的实际评价是 1 万, 如果出价 1 万, 第二高出价 9 千, 你赚 1 千; 如果你出价低于 9 千, 你什么也得不到.

# VCG 机制 Vickrey-Clarke-Groves Mechanism

- ▶ 公共产品的偏好显示: 不同的人有不同的偏好, 是私人信息. 如何让每个人报告自己的真实偏好?
- ▶ 每个人可以任意地报告自己的偏好, 但可能要纳缴一定数量的“税”.  
计算办法: 先将其他人的偏好加总, 给出总价值最大的项目; 然后将第一个人的偏好加上, 如果不影响结果, 不征税; 否则, 应纳税等于改变结果给其他人带来的损失.

$$p_i = \underbrace{\max_x \sum_{j \neq i} v_j(x)}_{i \text{ 不参与}} - \underbrace{\sum_{j \neq i} v_j(x^*)}_{i \text{ 参与}} \quad \text{where } x^* = \operatorname{argmax}_x \sum_{i=1}^n v_i(x)$$

- ▶ 同学聚会去吃川菜还是粤菜? 每人报告自己的真实偏好是占优策略.

	川菜	粤菜	税额 $p_i$
A	30	10	0
B	0	40	30
C	20	10	0
合计	50	60	30

# 公平与夏普利值

- ▶ A,B,C 合作开了一家公司. 公司任何决策只有超过 50% 的赞成票才能通过.
- ▶ 根据投资额, A 拥有 50% 的票力, B 拥有 40%, C 拥有 10%.
- ▶ 年终净赚 150 万, 根据票力之比, A,B,C 分别 75 万, 60 万, 15 万吗?
- ▶ C 提议: A 拿 80 万, 自己 70 万, B 一分没有.
- ▶ B 提议: A 拿 85 万, 自己 65 万, C 一分没有.
- ▶ 怎么分才公平?
- ▶ 三人中, 任何一人都不是决定性的, 都得与他人结盟才有决策权.
- ▶ 夏普利值的基本假设是: 各投票顺序联盟形成的可能性相同, 玩家的夏普利值为其对联盟的边际贡献之和除以各种可能的联盟组合总数.

1	A	A	B	B	C	C
2	B	C	A	C	A	B
3	C	B	C	A	B	A
关键加入者	B	C	A	A	A	A

Table: 所有可能的投票顺序

- ▶ 作为关键加入者的次数. A: 4 次; B: 1 次; C: 1 次.
- ▶ 夏普利值. A:  $\frac{4}{6}$ ; B:  $\frac{1}{6}$ ; C:  $\frac{1}{6}$ .
- ▶ 根据影响力, A:  $150 \times \frac{4}{6} = 100$  万; B: 25 万; C: 25 万.

# 讨价还价与纳什讨价还价解

- ▶ 市场上有一个画家 A 和一个画廊 B.
- ▶ 若画家 A 自己卖画, 可得  $a = 1500$ .
- ▶ 若交给画廊 B 卖, 可得  $V = 6000$ .
- ▶ 若画廊 B 干别的事, 可得  $b = 1000$ .
- ▶ AB 该怎么分利?
- ▶  $S := V - a - b = 3500$  是合作带来的剩余价值.
- ▶ 我们用  $x$  表示 A 应分得的价值,  $y$  表示 B 应分得的价值.  $\alpha, \beta$  表示 A 和 B 剩余价值  $S$  的分配比例 (与边际贡献有关).

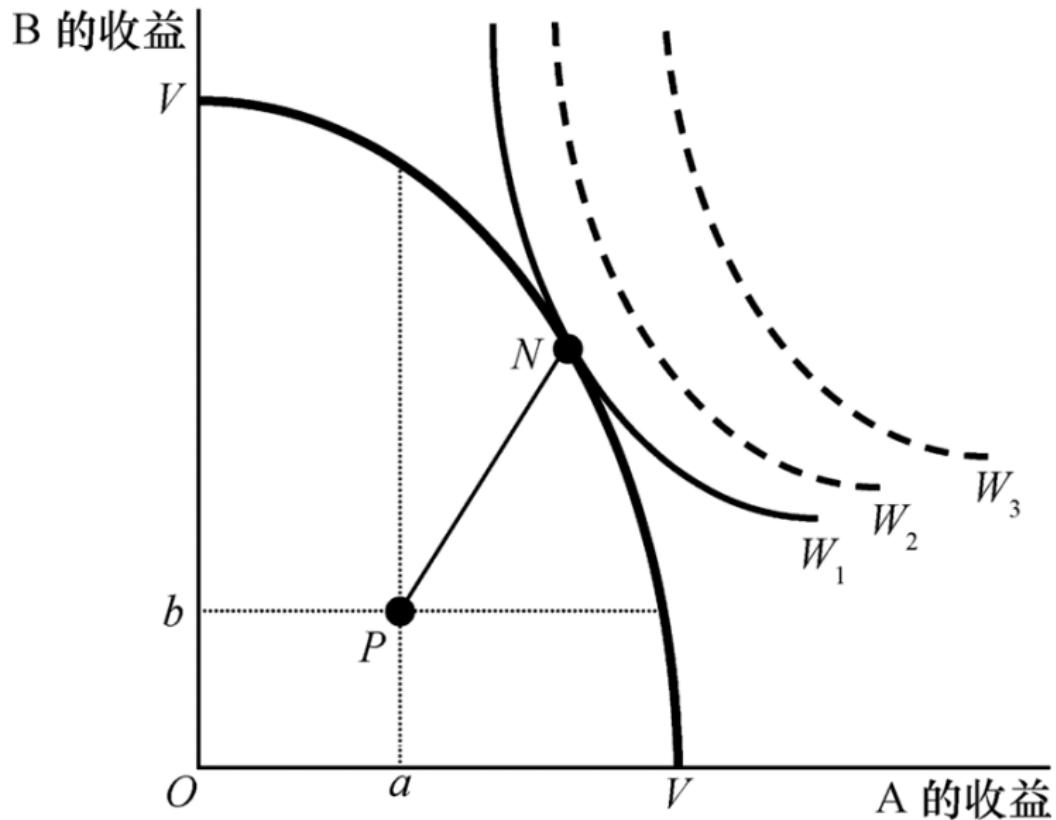
$$x = a + \alpha S, \quad y = b + \beta S$$

- ▶ 纳什讨价还价解即优化福利函数  $W(x, y) = (x - a)^\alpha (y - b)^\beta$

$$(x^*, y^*) = \underset{x, y: x+y=V}{\operatorname{argmax}} (x - a)^\alpha (y - b)^\beta$$

- ▶ 因为若缺少任何一方, 这 3500 的剩余价值都得不到, 所以每一方的边际贡献都是 3500. 总边际贡献是  $3500 + 3500 = 7000$ . 每一方的边际贡献占比都是  $\frac{3500}{3500+3500} = \frac{1}{2} = \alpha = \beta$ .
- ▶ A 应分得  $1500 + \frac{1}{2} \times 3500 = 1500 + 1750 = 3250$ ; B 应分得  $2750$ .

# 纳什讨价还价解



# 谈判中的影响力 — 不可替代性

- ▶ 如果市场上还有一家画廊 C, 若 A 交给画廊 C 卖, 可得  $V' = 5000$ .
  - ▶ 若画廊 C 干别的事, 可得  $c = 500$ .
  - ▶ AB 又该怎么分利?
- 
- ▶ AB 合作中 A 的边际贡献是  $V - a - b = 3500$ .
  - ▶ 若 AB 不合作, 则 AC 合作, 因为 B 的存在, C 的边际贡献为 0, A 拿走所有剩余  $V' - c$ . 即 A 的市场价值不是  $a$  而是  $V' - c$ .
  - ▶ 所以, AB 合作中 B 的边际贡献是  
$$V - (V' - c) - b = 6000 - (5000 - 500) - 1000 = 500.$$
  - ▶ A 的边际贡献占总边际贡献的  $\frac{3500}{3500+500} = \frac{7}{8}$ , B 占  $\frac{1}{8}$ .
  - ▶ A 应分得:  $1500 + \frac{7}{8} \times 3500 = 4562.5$ .
  - ▶ B 应分得:  $1000 + \frac{1}{8} \times 3500 = 1437.5$ .

**Remark:** 大牌影星天价片酬.

# 腐败 vs 信息

- ▶ 腐败的深层根源：信息不对称；
- ▶ 特别是事后的信息不对称：一方当事人的行为不能被另一方观察到；
- ▶ 研究行为信息不对称的理论叫“委托-代理”理论或道德风险理论。

## 委托-代理关系

- ▶ 法律上的委托 - 代理关系：如果甲乙两人达成一个协议，甲将做某事的权利交给乙，甲为委托人，乙为代理人。
- ▶ 代理人对委托人的责任：(1) 没有许可，不能再代理；(2) 不能把自己放在与委托人利益冲突的地位；(3) 保密责任和诚信责任。
- ▶ 委托人对代理人的责任：(1) 补偿责任。委托人给代理人补偿报酬；(2) 免除法律责任。委托人要为代理人的行为承担法律责任；(3) 留置权。如果委托人给代理人的补偿没有到位，代理人有权滞留委托人的财产。
- ▶ 经济上的委托 - 代理关系要宽松很多：只要一方行为影响另一方的利益，就有委托 - 代理关系。有私人信息的一方是代理人，没有私人信息的称为委托人。

# 委托-代理问题

- ▶ 委托人与代理人可能的利益冲突：对委托人最优的选择不一定是对代理人最优的选择。
- ▶ 如果委托人能观察到代理人的行为，则可以通过强制性合同约束代理人的行为。
- ▶ 即使不能完全观察到代理人的行为，如果代理人不害怕风险，“承包制”可以使代理人成为完全的风险承担者。
- ▶ 但如果代理人承担责任的能力有限，“承包制”没有可行性。

# 对代理人的激励机制的设计

- ▶ 保险与激励的冲突
  - ▶ 最优风险分担: 如果委托人是风险中性的, 代理人是风险规避的, 风险应该完全由委托人承担, 代理人拿固定报酬.
  - ▶ 最优激励: 代理人应该承担完全风险.
  - ▶ 这就产生了保险与激励的冲突.
  - ▶ 最优激励合同要在保险与激励之间求得平衡.
    - 汽车的部分保险 (如 80%).
- ▶ 激励的强度
  - ▶ 代理人的边际生产率越高, 激励应该越强.
  - ▶ 代理人的产出的不确定性越大, 或测度越困难, 激励应该越弱.
  - ▶ 代理人越风险规避, 激励应该越弱.
  - ▶ 代理人对激励的反应程度: 反应越强, 激励应该越强.
- ▶ 相对业绩比较有助于改进激励, 但也可能导致代理人之间合谋或相互拆台.
- ▶ 论功行赏 vs 任人唯贤
  - 论功行赏可能导致每个人都被晋升到他不能胜任的位置
- ▶ 多重任务下的激励: 大学教学 + 科研

# 政府官员的激励

- ▶ 很多很多委托人;
- ▶ 多项任务;
- ▶ 业绩难以度量, 投入也不容易度量;
- ▶ 所以, 政府官员难以激励, 只能以监督为主.
- ▶ 但监督是不完全的, 是有成本的.
- ▶ 权利越大, 越难监督.
- ▶ “把权力关进笼子”. 民主、法治.

# 官员腐败问题

- ▶  $W$ : 官员工资.
- ▶  $B(q)$ : 权力租金, 即官员在位置上可能收受的贿赂, 也可看作不腐败的机会成本. 一般来说, 权力  $q$  越大, 权力租金越高.
- ▶  $p$ : 腐败行为被发现暴露的概率.
- ▶  $F$ : 对腐败的处罚.
- ▶  $U$ : 政府外的保留效用.
- ▶  $\alpha$ : 官员的羞耻感或脸皮厚度.
- ▶ 官员腐败的期望收益

$$(1 - p)(W + B(q)) + p(U - \alpha F)$$

- ▶ 官员不腐败的条件:

$$W \geq \frac{1-p}{p}B(q) + U - \alpha F$$

- ▶ 如何控制腐败? 提高  $p, F, W, \alpha$ , 减少  $q$ .

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret Minimization

Subgame Perfect Equilibrium

Games with Incomplete Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

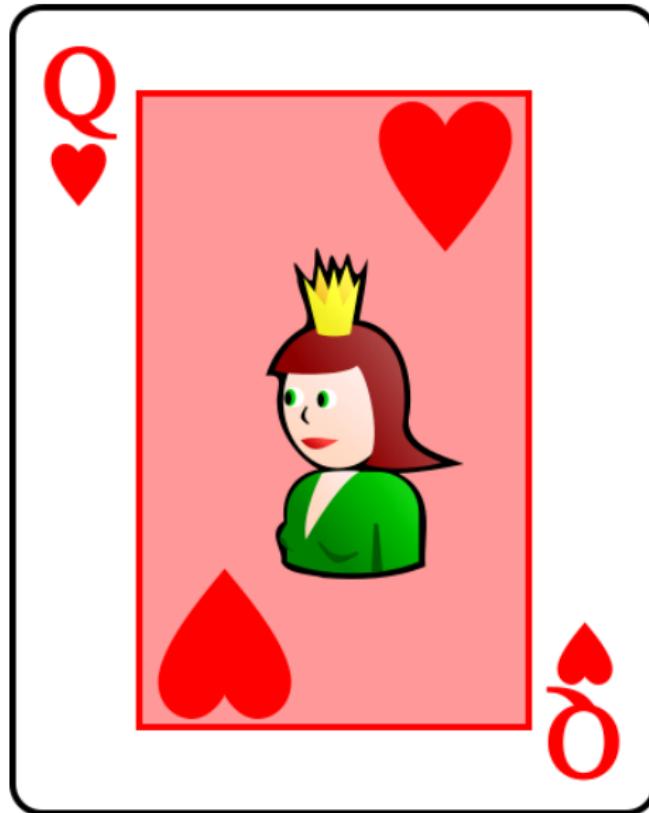
Artificial General Intelligence

What If Computers Could Think?

References

1753

# Imperfect Information Game



## Counterfactual Regret Minimization

Play a strategic game for a number of rounds:

- ▶ Regret is determined after each game round: If I had played another move, my payoff would have been that much higher!
- ▶ Accumulate all positive regrets over time.
- ▶ Match the probabilities of a mixed strategy with the accumulated regret.

Take the average over all mixed strategies.

If two players use the regret matching technique in a zero-sum game, then the average over the mixed strategies converges to Nash equilibrium strategies.

## Counterfactual Regret Minimization

- ▶ The **reach probability**  $\pi^\rho(h) := \prod_{i \in N \cup \{c\}} \pi_i^\rho(h)$  is the probability that history  $h$  will be reached with strategy  $\rho$ , where  $\pi_i^\rho$  is the contribution of player  $i$ , and  $\pi_{-i}^\rho$  is the product of all player contributions except player  $i$ .
- ▶  $\pi^\rho(I) := \sum_{h \in I} \pi^\rho(h)$
- ▶ The **counterfactual reach probability**  $\pi_{-i}^\rho(I)$  is the probability of reaching  $I$  under the assumption that player  $i$  always uses actions with probability 1 in order to reach  $I$ .
- ▶  $\rho_{I \rightarrow a}$  is the same strategy as  $\rho$ , except that action  $a$  is always chosen at information set  $I$ .

# Counterfactual Regret Minimization

- The **counterfactual utility** of  $\rho$  at non-terminal history  $h$  is:

$$v_i(\rho, h) := \sum_{z \in \mathcal{Z}, h \prec z} \pi_{-i}^\rho(h) \pi^\rho(z \mid h) u_i(z)$$

- The **counterfactual regret** of not taking action  $a$  at history  $h \in I$  is:

$$r(h, a) := v_i(\rho_{I \rightarrow a}, h) - v_i(\rho, h)$$

- The **Counterfactual regret** of not taking  $a$  at  $I$  is:

$$r(I, a) := \sum_{h \in I} r(h, a)$$

- $r_i^t(I, a)$  refers to the regret in episode  $t$ , when players use  $\rho$  and player  $i$  does not take  $a$  at  $I$ .
- The **Cumulative counterfactual regret** is:

$$R_i^T(I, a) := \sum_{t=1}^T r_i^t(I, a)$$

# Counterfactual Regret Minimization

- The positive cumulative counterfactual regret is:

$$R_i^{T,+} := \max \{R_i^T(I, a), 0\}$$

- The regret matching strategy for episode  $T + 1$  is:

$$\rho_i^{T+1}(I, a) := \begin{cases} \frac{R_i^{T,+}(I, a)}{\sum_{a \in A(I)} R_i^{T,+}(I, a)} & \text{if } \sum_{a \in A(I)} R_i^{T,+}(I, a) > 0 \\ \frac{1}{|A(I)|} & \text{otherwise} \end{cases}$$

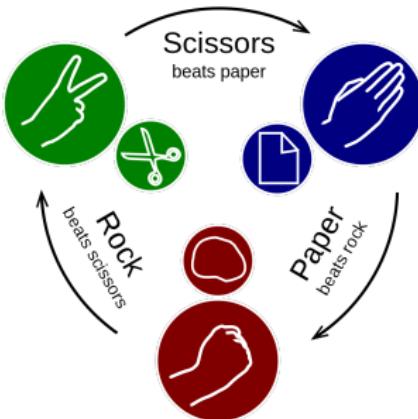
- The average strategy is:

$$\bar{\rho}_i^t(I)(a) := \frac{\sum_{t=1}^T \pi_i^{\rho^t}(I) \rho^t(I)(a)}{\sum_{t=1}^T \pi_i^{\rho^t}(I)}$$

## Regret Matching — RPS example with two rounds

Assume we play rock, paper, scissors, and player 1 uses regret matching.

1. Initial cumulative regret is  $(0, 0, 0)$
2. Play uniform strategy  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
3. Player 1 chooses  $R$ , while player 2 chooses  $P$
4. Regret for player 1:
  - ▶  $R : u_1(R, P) - u_1(R, P) = -1 - -1 = 0$
  - ▶  $P : u_1(P, P) - u_1(R, P) = 0 - -1 = 1$
  - ▶  $S : u_1(S, P) - u_1(R, P) = 1 - -1 = 2$
5. Player 1's cumulative counterfactual regret is now  $(0, 1, 2)$
6. Regret matching strategy:  $\rho_1^1 = (0, \frac{1}{3}, \frac{2}{3})$
7. Player 1 chooses  $P$ , while player 2 chooses  $S$
8. Regret for player 1:
  - ▶  $R : u_1(R, S) - u_1(P, S) = 1 - -1 = 2$
  - ▶  $P : u_1(P, S) - u_1(P, S) = -1 - -1 = 0$
  - ▶  $S : u_1(S, S) - u_1(P, S) = 0 - -1 = 1$
9. Player 1's cumulative counterfactual regret is now  $(2, 1, 3)$
10. Regret matching strategy:  $\rho_1^2 = (\frac{1}{3}, \frac{1}{6}, \frac{1}{2})$



11. The average strategy:  $(\frac{1}{6}, \frac{1}{4}, \frac{7}{12})$ .  
 Not close to NE, but will converge!

	$R$	$P$	$S$
$t_0$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$t_1$	0	$\frac{1}{3}$	$\frac{2}{3}$
$t_2$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$

$$h = RP$$

$$\frac{\frac{1}{3} \times \frac{1}{3} \times 0 + \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}}{\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3}} = \frac{1}{6}$$

$$\frac{\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} \times \frac{1}{6}}{\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3}} = \frac{1}{4}$$

$$\frac{\frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} + \frac{1}{3} \times \frac{1}{3} \times \frac{1}{2}}{\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3}} = \frac{7}{12}$$

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret

Minimization

**Subgame Perfect Equilibrium**

Games with Incomplete

Information

Evolutionary Games

Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

References

1753

# Subgame

某个博弈的子博弈是由该博弈的某个单一结点 (即该结点所在的信息集就只包含一个要素, 就是该结点) 及其所有后续结点构成的集合, 并且所有的后续结点必须满足一个条件, 即如果某一个后续结点属于该子博弈, 那么该后续结点所在信息集的所有结点必须也属于该子博弈.

## Definition (Subgame)

The *subgame* of the extensive game  $\Gamma$  that follows the history  $h$  is the extensive game  $\Gamma(h) := \langle N, P|_h, \mathcal{H}|_h, (\mathcal{I}_i)_{i \in N}, f_c, (u_i|_h)_{i \in N} \rangle$ , where

$$P|_h(h') := P(hh')$$

$$\mathcal{H}|_h := \{h' : hh' \in \mathcal{H}\}$$

$$u_i|_h(h') := u_i(hh')$$

$$\forall I \in \bigcup_{i \in N} \mathcal{I}_i : I \subset \{hh' : h' \in \mathcal{H}|_h\} \vee I \subset \mathcal{H} \setminus \{hh' : h' \in \mathcal{H}|_h\}$$

# Subgame Perfect Equilibrium

## Definition (Subgame Perfect Equilibrium)

A *subgame perfect equilibrium* of an extensive game  $\Gamma$  is a strategy profile  $s^*$  such that for every player  $i \in N$  and every nonterminal history  $h \in \mathcal{H} \setminus \mathcal{Z}$  for which  $P(h) = i$  we have

$$u_i|_h(s_i^*; s_{-i}^*) \geq u_i|_h(s_i; s_{-i}^*)$$

for every strategy  $s_i$  of player  $i$  in the subgame  $\Gamma(h)$ .

A behaviour strategy profile  $\rho^*$  is a *subgame perfect equilibrium* of a game  $\Gamma$  iff for all subgames  $\Gamma(h)$  of  $\Gamma$ , and,

$$\forall i \in N \forall \rho_i \in B_i : u_i|_h(\rho_i^*; \rho_{-i}^*) \geq u_i|_h(\rho_i; \rho_{-i}^*)$$

**Remark:** A Nash equilibrium is subgame perfect iff it is a Nash equilibrium in every subgame.

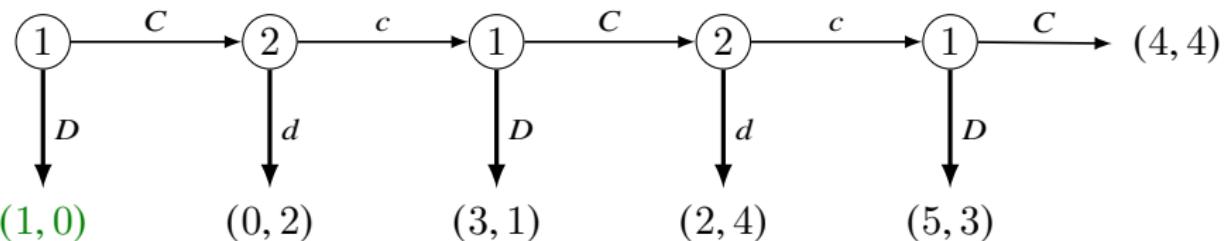
In a finite game of perfect information, the subgame perfect equilibrium is the backward induction solution.

# Kuhn's Theorem

## Theorem (Kuhn's Theorem)

*Every finite extensive game has a subgame-perfect equilibrium (possibly in mixed strategies).*

# Criticism of Backward Induction and Subgame Perfection?



## Remark

*In the unique subgame perfect equilibrium, players defect at every node, precluding any mutually beneficial cooperative outcome.*

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

## Game Theory

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies

Dynamic Games

Repeated Games  
Signaling Games  
Mechanism Design  
Counterfactual Regret Minimization  
Subgame Perfect Equilibrium  
**Games with Incomplete Information**  
Evolutionary Games  
Voting System

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

**References**

1753

# Sequentially Rationality

## Definition (Assessment)

An *assessment* in an extensive game is a pair  $(\rho, \mu)$ , where  $\rho$  is a profile of behavioral strategies and  $\mu$  is a belief system.  $\mu(I)(h)$  is the probability that player  $P(I)$  assigns to the history  $h \in I$ , conditional on  $I$  being reached.

## Definition (Sequentially Rationality)

An assessment  $(\rho, \mu)$  is *sequentially rational* iff for every player  $i$  and every information set  $I \in \mathcal{I}_i$  the strategy of player  $i$  is a best response to the other players' strategies given  $i$ 's beliefs at  $I$ .

$$\forall \rho'_i \in B_i : \sum_{h \in I} u_i|_h(\rho_i; \rho_{-i}) \mu(I)(h) \geq \sum_{h \in I} u_i|_h(\rho'_i; \rho_{-i}) \mu(I)(h)$$

We can define the outcome  $O(\rho, \mu \mid I)$  as the distribution over terminal histories determined by  $\rho$  and  $\mu$  conditional on  $I$  being reached.

$$O(\rho, \mu \mid I)(h') := \begin{cases} 0 & \text{if } \nexists h \in I \ (h \prec h') \\ \mu(I)(h) \prod_{k=\ell(h)}^{\ell(h')} \rho_{P(h_{<k})}(h_{<k})(h_k) & \text{otherwise} \end{cases}$$

since by perfect recall, there is at most one subhistory  $h$  of  $h'$  in  $I$ , and the histories  $h_{1:k}$  for  $k = \ell(h), \dots, \ell(h')$  lie in different information sets.

- ▶ An assessment  $(\rho, \mu)$  is *sequentially rational* iff for every player  $i$  and every information set  $I \in \mathcal{I}_i$

$$\forall \rho'_i \in B_i : O(\rho, \mu | I) \succsim_i O((\rho'_i; \rho_{-i}), \mu | I)$$

- ▶ A behavioral strategy profile to be *completely mixed* iff it assigns positive probability to every action at every information set.

## Definition (Consistency of Beliefs with Strategies)

Let  $\Gamma$  be a finite extensive game with perfect recall. An assessment  $(\rho, \mu)$  is *consistent* iff there is a sequence  $((\rho^n, \mu^n))_{n=1}^{\infty}$  of assessments s.t.  $(\rho, \mu) = \lim_{n \rightarrow \infty} (\rho^n, \mu^n)$  and each strategy profile  $\rho^n$  is completely mixed and each belief system  $\mu^n$  is derived from  $\rho^n$  using Bayes' rule:

$$\mu^n(I)(h) = \frac{\bar{\rho}^n(h)}{\sum_{h \in I} \bar{\rho}^n(h)}$$

## Definition (Sequential Equilibrium)

An assessment is a *sequential equilibrium* of a finite extensive game with perfect recall iff it is sequentially rational and consistent.

## Theorem

*Every finite extensive game with perfect recall has a sequential equilibrium.*

## Theorem

*In an extensive game with perfect information,  $(\rho, \mu)$  is a sequential equilibrium iff  $\rho$  is a subgame perfect equilibrium.*

# Bayesian Strategic Game with Observable Actions

## Definition (Bayesian Strategic Game with Observable Actions)

A Bayesian game is a tuple  $\langle N, (S_i)_{i \in N}, (\Theta_i)_{i \in N}, p, (u_i)_{i \in N} \rangle$  where

- ▶ The sets of players  $N$ .
- ▶ The sets of strategies of the players  $(S_i)_{i \in N}$ .
- ▶  $\Theta_i$  is a finite set (the set of possible types of player  $i$ ).  $\Theta := \prod_{i \in N} \Theta_i$ .
- ▶  $p \in \Delta\Theta$ .
- ▶  $u_i : S \times \Theta \rightarrow \mathbb{R}$

## Intuition for a Bayesian Game

- ▶ Harsanyi (1967) suggests a way of understanding a game of incomplete information as a game of imperfect information with a fictitious player who moves first called **Nature**.
- ▶ Nature draws a type vector  $(\theta_1, \dots, \theta_n)$  for all players.
- ▶ For each  $i$ , Nature reveals  $\theta_i$  to some players (usually including  $i$ ) and sends a signal to others.
- ▶ Players choose actions  $s_i$  simultaneously.
- ▶ Payoffs are realized:  $u_i(s, \theta)$  for all  $i$ .

## Expected Utility

1. ex-post — the agent knows all agents' types.
2. ex-interim — an agent knows her own type but not the types of the other agents.
3. ex-ante — the agent knows nothing about anyone's actual type.

# 狐狸、权力与不完全信息

- ▶ 小女孩提着一篮子鸡蛋穿过树林.
- ▶ 喜欢抢鸡蛋吃的狐狸跳了出来：“我是狐狸！”
- ▶ 小女孩：“怎么证明你是狐狸？”
- ▶ 狐狸：“我有厚厚的皮毛.”
- ▶ 小女孩：“兔子也有厚厚的皮毛.”
- ▶ 狐狸：“我有毛茸茸的尾巴.”
- ▶ 小女孩：“松鼠也有毛茸茸的尾巴.”
- ▶ 狐狸一直试图证明自己，不知不觉跟着小女孩走出了树林.
- ▶ 一只猎犬经过吓跑了狐狸.
- ▶ 狐狸边跑边喊，“不信你问猎犬，我真的是狐狸！”
- ▶ 小女孩：“**我知道，我早就知道。**”

	攻击	不攻击
防卫	-2, -2	0, 0
不防卫	-1, 1	0, 0

Table: 与狐狸博弈

	攻击	不攻击
防卫	0, -2	0, 0
不防卫	-1, 1	0, 0

Table: 与兔子/松鼠博弈

Remark: 权力需要被认可，如果得不到认可，权力也将不复存在。如果有人试图用权力吓唬你，你应该提出质疑，让对方承担举证的责任。

## Ex-post Expected Utility

### Definition (Ex-post Expected Utility)

Player  $i$ 's ex post expected utility in a Bayesian game is defined as

$$V_i(\sigma, \theta) := \sum_{s \in S} \left( \prod_{j \in N} \sigma_j(s_j \mid \theta_j) \right) u_i(s, \theta)$$

where the players' types are given by  $\theta \in \Theta$ , and the players' strategies are given by  $\sigma \in \prod_{i \in N} \prod_{\theta_i \in \Theta_i} \Delta S_i$ ,  $\sigma = (\sigma_i(\cdot \mid \theta_i))_{\theta_i \in \Theta_i, i \in N}$ ,  $\sigma_i(\cdot \mid \theta_i) \in \Delta S_i$ .

## Ex-interim Expected Utility

### Definition (Ex-interim Expected Utility)

Player  $i$ 's ex interim expected utility in a Bayesian game, where  $i$ 's type is  $\theta_i$  and where the players' strategies are given by the mixed-strategy profile  $\sigma$ , is defined as

$$V_i(\sigma, \theta_i) := \sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i} | \theta_i) \sum_{s \in S} \left( \prod_{j \in N} \sigma_j(s_j | \theta_j) \right) u_i(s, (\theta_i; \theta_{-i}))$$

or equivalently as

$$V_i(\sigma, \theta_i) = \sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i} | \theta_i) V_i(\sigma, (\theta_i; \theta_{-i}))$$

## Ex-ante Expected Utility

### Definition (Ex-ante Expected Utility)

Player  $i$ 's ex ante expected utility in a Bayesian game, where the players' strategies are given by the mixed-strategy profile  $\sigma$ , is defined as

$$V_i(\sigma) := \sum_{\theta \in \Theta} p(\theta) \sum_{s \in S} \left( \prod_{j \in N} \sigma_j(s_j \mid \theta_j) \right) u_i(s, \theta)$$

or equivalently as

$$V_i(\sigma) = \sum_{\theta \in \Theta} p(\theta) V_i(\sigma, \theta)$$

or again equivalently as

$$V_i(\sigma) = \sum_{\theta_i \in \Theta_i} p(\theta_i) V_i(\sigma, \theta_i)$$

## Best Response in a Bayesian Game

### Definition (Best Response in a Bayesian Game)

The set of player  $i$ 's best response to mixed-strategy profile  $\sigma_{-i}$  are given by

$$\text{BR}_i(\sigma_{-i}) := \underset{\sigma_i \in \Delta S_i^{\Theta_i}}{\operatorname{argmax}} V_i(\sigma_i; \sigma_{-i})$$

**Remark:** It may seem odd that  $\text{BR}_i$  is calculated based on  $i$ 's ex ante expected utility. However, we are in fact performing independent maximization of  $i$ 's *ex interim expected utilities* conditioned on each type that he could have.

Intuitively speaking, if a certain action is best after the signal is received, it is also the best conditional plan devised ahead of time for what to do should that signal be received.

## Definition (Bayesian Equilibrium)

A Bayesian equilibrium is a mixed strategy profile  $\sigma^*$  that satisfies

$$\forall i \in N : \sigma_i^* \in \text{BR}_i(\sigma_{-i}^*)$$

## Definition (Ex-post Equilibrium)

An ex post equilibrium is a mixed strategy profile  $\sigma$  that satisfies

$$\forall \theta \in \Theta \forall i \in N : \sigma_i \in \operatorname{argmax}_{\substack{\sigma_i \in \Delta S_i^\Theta}} V_i((\sigma_i; \sigma_{-i}), \theta)$$

The ex post equilibrium is similar in flavor to equilibria in dominant strategies, which do not require agents to believe that other agents act rationally. However, it seems too good to be true most of the time.

# Bayesian Extensive Game with Observable Actions

## Definition (Bayesian Extensive Game with Observable Actions)

A *Bayesian extensive game with observable actions* is a tuple

$\langle \Gamma, (\Theta_i)_{i \in N}, (p_i)_{i \in N}, (u_i)_{i \in N} \rangle$  where

- ▶  $\Gamma = \langle N, P, \mathcal{H} \rangle$  is an extensive game form with perfect information and simultaneous moves.
- ▶  $\Theta_i$  is a finite set (the set of possible types of player  $i$ ).  $\Theta := \prod_{i \in N} \Theta_i$ .
- ▶  $p_i \in \Delta^0(\Theta_i)$ .
- ▶  $u_i : \mathcal{H} \times \Theta \rightarrow \mathbb{R}$

**Remark:** the type of player encapsulates all the information possessed by the player that is not common knowledge. This is often quite simple (e.g. the player's knowledge of his private payoff function), but can also include his beliefs about other players' payoffs, about their beliefs about his own payoff, and any other higher-order beliefs.

There is a way of capturing the common prior is to hypothesize a special agent  $c$  called “Nature” who makes probabilistic choices. The agent “nature”( $c$ ) selects the types of the players, who are subsequently fully cognizant at all points of all moves taken previously. We can associate with any such game an extensive game (with imperfect information and simultaneous moves) in which the set of histories is  $\{\emptyset\} \cup (\mathcal{H} \times \Theta)$  and each information set of each player  $i$  takes the form

$$I(h, \theta_i) := \{(h, (\theta_i, \theta_{-i})) : \theta_{-i} \in \Theta_{-i}\}$$

for  $i \in P(h)$  and  $\theta_i \in \Theta_i$ .

Let  $s$  be a profile of behavioral strategies in  $\Gamma$ . Define  $O_h(s)$  to be the probability measure on the set of terminal histories of  $\Gamma$  generated by  $s$  given that the history  $h$  has occurred. Define  $O((s_i; \rho_{-i}), \mu_{-i} | h)$  to be the probability measure on the set of terminal histories of  $\Gamma$  given that player  $i$  uses the strategy  $s_i$  in  $\Gamma$ , each type  $\theta_j$  of each player  $j$  uses the strategy  $\rho_j(\theta_j)$ , the game has reached  $h$ , and the probability that  $i$  assigns to  $\theta_{-i}$  is derived from  $\mu_{-i}(h)$ . That is,  $O((s_i; \rho_{-i}), \mu_{-i} | h)$  is the compound lottery in which the probability of the lottery  $O_h(s_i; (\rho_j(\theta_j))_{j \in N \setminus \{i\}})$  is

$$\prod_{j \in N \setminus \{i\}} \mu_j(h)(\theta_j) \text{ for each } \theta_{-i} \in \Theta_{-i}.$$

## Ex-ante Expected Utility

### Definition (Ex-ante Expected Utility)

At history  $h$ , the player  $i$ 's ex ante expected utility in a Bayesian game, where the players' behavioral strategy profile is  $\rho$ , is defined as

$$V_i(\rho, h) := \sum_{\theta_i \in \Theta_i} \mu_i(h)(\theta_i) \sum_{\theta_{-i} \in \Theta_{-i}} u_i|_h(\rho, \theta) \prod_{j \in N \setminus \{i\}} \mu_j(h)(\theta_j)$$

# Perfect Bayesian Equilibrium

## Definition (Perfect Bayesian Equilibrium)

For a Bayesian extensive game with observable actions,

$((\rho_i), (\mu_i)) := ((\rho_i(\theta_i))_{\theta_i \in \Theta_i}, (\mu_i(h))_{h \in \mathcal{H} \setminus \mathcal{Z}})$  is a *perfect Bayesian equilibrium* of the game iff the following conditions are satisfied, where  $\rho_i(\theta_i)$  is a behavioral strategy of player  $i \in N$  and  $\mu_i(h) \in \Delta \Theta_i$ .

## Perfect Bayesian Equilibrium — definition continued

- (I) *Sequential rationality* For every terminal history  $h \in \mathcal{H} \setminus \mathcal{Z}$ , every player  $i \in P(h)$ , and every  $\theta_i \in \Theta_i$  the probability measure  $O((\rho_i(\theta_i); \rho_{-i}), \mu_{-i} | h)$  is at least good for type  $\theta_i$  as  $O((s_i; \rho_{-i}), \mu_{-i} | h)$  for any strategy  $s_i$  of player  $i$  in  $\Gamma$ .

$$\forall s_i \in S_i : O((\rho_i(\theta_i); \rho_{-i}), \mu_{-i} | h) \succcurlyeq_i O((s_i; \rho_{-i}), \mu_{-i} | h)$$

or equivalently, for  $\rho'_i \in B_i$ ,

$$\sum_{\theta_{-i} \in \Theta_{-i}} u_i|_h((\rho_i; \rho_{-i}), \theta) \prod_{j \in N \setminus \{i\}} \mu_j(h)(\theta_j) \geq \sum_{\theta_{-i} \in \Theta_{-i}} u_i|_h((\rho'_i; \rho_{-i}), \theta) \prod_{j \in N \setminus \{i\}} \mu_j(h)(\theta_j)$$

- (II) *Correct initial beliefs*  $\forall i \in N : \mu_i(\emptyset) = p_i$
- (III) *Action-determined beliefs* If  $i \notin P(h)$  and  $a \in A(h)$  then  $\mu_i(ha) = \mu_i(h)$ ;  
if  $i \in P(h), a \in A(h), a' \in A(h)$  and  $a = a'$  then  $\mu_i(ha) = \mu_i(ha')$ .
- (IV) *Bayesian updating* If  $i \in P(h)$  and  $\exists \theta_i \in \text{supp}(\mu_i(h)) (a \in \text{supp}(\rho_i(\theta_i)(h)))$  then for any  $\theta' \in \Theta_i$  we have

$$\mu_i(ha)(\theta'_i) := \frac{\rho_i(\theta'_i)(h)(a) \cdot \mu_i(h)(\theta'_i)}{\sum_{\theta_i \in \Theta_i} \rho_i(\theta_i)(h)(a) \cdot \mu_i(h)(\theta_i)} \quad (\text{Bayesian Update})$$

The condition of Bayesian updating relates to a case in which player  $i$ 's action at the history  $h$  is consistent with the other players' beliefs about player  $i$  at  $h$ , given  $\rho_i$ . In such a case the condition requires not only that the new belief depend only on player  $i$ 's action (as required by the condition of action-determined beliefs) but also that the players' beliefs be derived via Bayes' rule from their observation of player  $i$ 's actions. Thus the players update their beliefs about player  $i$  using Bayes' rule until his behaviour contradicts his strategy  $\rho_i$ , at which point they form a new conjecture about player  $i$ 's type that is the basis for future Bayesian updating until there is another conflict with  $\rho$ .

## Theorem

Let  $(\rho; \mu)$  be a sequential equilibrium of the extensive game associated with the finite Bayesian extensive game with observable actions. For every  $h \in \mathcal{H}$ ,  $i \in P(h)$ , and  $\theta_i \in \Theta_i$ , let  $\rho'_i(\theta_i)(h) := \rho_i(I(h, \theta_i))$ . Then there is a collection  $(\mu'_i(h))_{i \in N, h \in \mathcal{H}}$ , where  $\mu'_i(h) \in \Delta \Theta_i$ , such that

$$\mu(I(h, \theta_i))(h, \theta) = \prod_{j \in N \setminus \{i\}} \mu'_j(h)(\theta_j) \text{ for all } \theta \in \Theta \text{ and } h \in \mathcal{H}$$

and  $((\rho'_i), (\mu'_i))$  is a perfect Bayesian equilibrium of the Bayesian extensive game.

# Trembling Hand Perfect Equilibrium

## Definition (Trembling Hand Perfect Equilibrium)

A strategy profile  $\sigma$  is a *trembling hand perfect equilibrium* iff there exists a sequence of completely mixed strategy profiles  $\sigma^k \rightarrow \sigma$  such that  
 $\forall i \forall k : \sigma_i \in \text{BR}_i(\sigma_{-i}^k)$ .

## Theorem

A strategy profile in a finite two-player strategic game is a trembling hand perfect equilibrium iff it is a mixed strategy Nash equilibrium and the strategy of neither player is weakly dominated.

## Theorem

Every finite strategic game has a trembling hand perfect equilibrium.

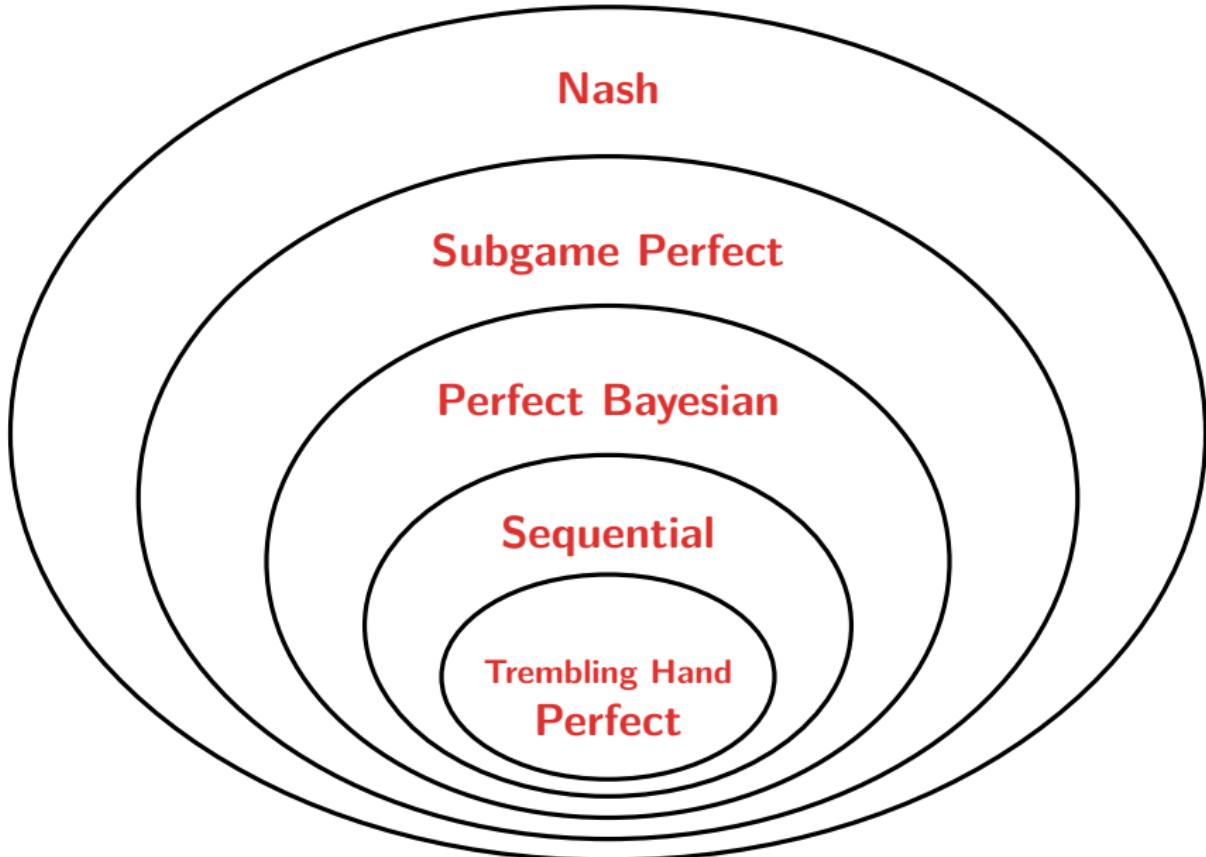
## Definition

A trembling hand perfect equilibrium of a finite extensive game is a behavioral strategy profile that corresponds to a trembling hand perfect equilibrium of the agent strategic form of the game.

## Theorem

*For every trembling hand perfect equilibrium  $\rho$  of a finite extensive game with perfect recall there is a belief system  $\mu$  such that  $(\rho; \mu)$  is a sequential equilibrium of the game.*

# Equilibria



# Correlated Equilibrium

## Definition (Nash Equilibrium)

- $s^*$  is a *pure Nash equilibrium* iff  $s^* \in \prod_{i \in N} \operatorname{argmax}_{s_i \in S_i} u_i(s_i; s_{-i})$ .
- $\sigma^*$  is a *mixed Nash equilibrium* iff

$$\forall i \in N \forall \sigma_i \in \Delta S_i : u_i(\sigma_i^*; \sigma_{-i}^*) \geq u_i(\sigma_i; \sigma_{-i}^*)$$

## Definition (Correlated Equilibrium)

Let  $\Omega$  be the state space,  $\mathcal{I}_i$  be the information partition of player  $i$ ,  $P(\cdot | \mathcal{I}_i) \in \Delta \Omega$  be the interim belief systems, and  $\sigma_i : \Omega \rightarrow S_i$  be measurable with regard to  $\mathcal{I}_i$ . Then  $(\sigma_i)_{i \in N}$  is a posteriori equilibrium of the strategic game  $(N, S_i, u_i)$  iff

$$\forall i \in N \forall s_i \in S_i : \sum_{\omega \in \Omega} P(\omega | \mathcal{I}_i(\omega)) \left( u_i(\sigma_i(\omega); \sigma_{-i}(\omega)) - u_i(s_i; \sigma_{-i}(\omega)) \right) \geq 0$$

**Remark:** For every Nash equilibrium there exists a corresponding correlated equilibrium.

## Choosing a Correlated Equilibrium

- ▶ An equilibrium which maximises the sum of the expected utility of the players is called a **utilitarian equilibrium**.
- ▶ An equilibrium which maximises the expected utility of Player  $i$  is called a **Libertarian  $i$  equilibrium**.
- ▶ An equilibrium which maximises the minimum expected utility of a player is called an **egalitarian equilibrium**.

# Contents

Introduction	Repeated Games
Philosophy of Induction	Signaling Games
Inductive Logic	Mechanism Design
Universal Induction	Counterfactual Regret Minimization
Causal Inference	Subgame Perfect Equilibrium
<b>Game Theory</b>	Games with Incomplete Information
Preferences and Expected Utility	<b>Evolutionary Games</b>
Strategic Games	Voting System
Extensive Games with Perfect Information	Reinforcement Learning
Eliminating Dominated Strategies	Deep Learning
Dynamic Games	Artificial General Intelligence
	What If Computers Could Think?
	References 1753

# 演化博弈 — 婚姻博弈

	物质型	感情型
物质型	1, 1	0, 0
感情型	0, 0	2, 2

- ▶ 假定总人口中, 物质型的比例为  $x$ , 感情型的比例为  $1 - x$ .
- ▶ 对任何一个个体而言, 物质型的期望效用:  $x1 + (1 - x)0 = x$ .
- ▶ 感情型的期望效用:  $x0 + (1 - x)2 = 2(1 - x)$ .
- ▶ 如果  $x > 2/3$ , 物质型更适合生存, 将演化成稳定均衡.
- ▶ 如果  $x < 2/3$ , 感情型更适合生存, 将演化成稳定均衡.
- ▶ 如果  $x = 2/3$ , 两类人有同样的适应性, 但此二元均衡是非稳定的.

# 演化博弈 — 鹰鸽博弈

	鹰	鸽
鹰	-1, -1	1, 0
鸽	0, 1	$\frac{1}{2}, \frac{1}{2}$

- ▶ 假定鹰派的比例是  $x$ , 鸽派的比例是  $1 - x$ .
- ▶ 鹰派的效用:  $-1x + 1(1 - x) = 1 - 2x$ .
- ▶ 鸽派的效用:  $0x + \frac{1}{2}(1 - x) = \frac{1}{2}(1 - x)$ .
- ▶ 如果  $x < 1/3$ , 鹰派占优势, 不稳定.
- ▶ 如果  $x > 1/3$ , 鸽派占优势, 不稳定.
- ▶ 如果  $x = 1/3$ , 同样的适应性, 稳定.
- ▶ 如果初始人口由单一类型构成, 另一类型可以成功入侵, 直到均衡.
  - † 两个纯策略均衡: (鹰、鸽), (鸽、鹰); 一个混合均衡:  $(1/3, 2/3)$
  - △ 假定存在某种显性的标记机制: 在博弈开始之前, 每个人收到一个信号:  $A$  或  $B$ ; 概率是  $1/2$ ; 信号完全负相关; 标记是公共知识.
    1. 如果  $A$ , 选择“鹰”; 如果  $B$ , 选择“鸽”. 是 ESS.
    2. 如果  $A$ , 选择“鸽”; 如果  $B$ , 选择“鹰”. 是 ESS.
    3. 无论  $AB$ , 以  $1/3$  的概率选择“鹰”,  $2/3$  的概率选择“鸽”. 不是 ESS.

# 自发秩序与产权制度

- ▶ 社会秩序是所有人行为选择的结果，但不是集体理性集中设计的结果，而是自发演化的结果。
- ▶ 社会规范是演化稳定策略，但不一定是帕累托最优的。
- ▶ 虽然不是集体理性集中设计的产物，但“制度企业家”扮演了重要角色。（哲学王）
- ▶ 改变游戏规则，进行制度创新。创新意味着让人们用新的价值观念代替旧的价值观念、用新的行为方式代替旧的行为方式、用新的是非观和新的善恶观代替旧的是非观和旧的善恶观，意味着我们要认同原来可能不认同的东西或不再认同我们原来认同的东西。
  - 儒家：“仁、义、礼、智、信”，“己所不欲勿施于人”，“以德报德，以直报怨”，“君君臣臣父父子子”。
- ▶ 制度企业家面临的不确定性和风险挑战异于常人。他们不能以“利”和“名”为目的，买他们账的“客户”可能在遥远的未来。
- ▶ “自私的”基因 & 迷因（meme）

## 演化稳定策略 Evolutionarily Stable Strategy

- ▶ 静态: 一个特定的行为模式被称为是演化稳定的, 如果它的种群不能被变异所入侵, 或者说, 任何偏离行为模式的个体具有更低的生存能力, 种群将会恢复到原来的状态.
- ▶ 动态: 假定初始状态存在多样的行为模式, 随着时间的推移, 如果某个特定的行为模式能逐步主导整个种群, 这个特定的行为模式就是演化稳定的.

## Evolutionarily Stable Strategy

Given a symmetric two-player normal form game,  $\sigma^*$  is an ESS iff  
 $\forall \sigma \neq \sigma^* \exists \delta \in (0, 1) \forall \varepsilon \in (0, \delta) :$

$$u(\sigma^*, (1 - \varepsilon)\sigma^* + \varepsilon\sigma) > u(\sigma, (1 - \varepsilon)\sigma^* + \varepsilon\sigma)$$

iff

$$(1 - \varepsilon)u(\sigma^*, \sigma^*) + \varepsilon u(\sigma^*, \sigma) > (1 - \varepsilon)u(\sigma, \sigma^*) + \varepsilon u(\sigma, \sigma)$$

iff

- $u(\sigma^*, \sigma^*) > u(\sigma, \sigma^*) \quad \text{or}$
- $u(\sigma^*, \sigma^*) = u(\sigma, \sigma^*) \quad \text{and} \quad u(\sigma^*, \sigma) > u(\sigma, \sigma)$

If  $\sigma$  is an ESS, then  $(\sigma, \sigma)$  is a Nash equilibrium. If  $(\sigma, \sigma)$  is a strict Nash equilibrium, then  $\sigma$  is an ESS.

	dove	hawk
Dove	$\frac{a}{2}, \frac{a}{2}$	$0, a$
Hawk	$a, 0$	$\frac{a-b}{2}, \frac{a-b}{2}$

$$\frac{a}{2}x + 0(1 - x) = ax + \frac{a - b}{2}(1 - x)$$

$$b > a \implies \left(1 - \frac{a}{b}, \frac{a}{b}\right) \text{ mix}$$

$$b \leq a \implies (H, h) \text{ pure}$$

## Evolutionarily Stable Strategy

- ▶ strict Nash Equilibrium:  $u(\sigma_i^*; \sigma_{-i}^*) > u(\sigma_i; \sigma_{-i}^*)$
- ▶ Nash Equilibrium:  $u(\sigma_i^*; \sigma_{-i}^*) \geq u(\sigma_i; \sigma_{-i}^*)$
- ▶ ESS:
  - $u(\sigma^*, \sigma^*) > u(\sigma, \sigma^*)$  or
  - $u(\sigma^*, \sigma^*) = u(\sigma, \sigma^*)$  and  $u(\sigma^*, \sigma) > u(\sigma, \sigma)$
- ▶ weak ESS:
  - $u(\sigma^*, \sigma^*) > u(\sigma, \sigma^*)$  or
  - $u(\sigma^*, \sigma^*) = u(\sigma, \sigma^*)$  and  $u(\sigma^*, \sigma) \geq u(\sigma, \sigma)$
- ▶ unbeatable strategy:  $u(\sigma^*, \sigma^*) > u(\sigma, \sigma^*)$  and  $u(\sigma^*, \sigma) > u(\sigma, \sigma)$   
unbeatable  $\implies$  strict Nash  $\implies$  ESS  $\implies$  weak ESS  $\implies$  Nash

## Replicator Dynamics in Symmetric Games

- ▶ Suppose players choose one of  $m$  actions. The payoff of a player when he plays  $a_i$  and the opponent plays  $a_j$  is  $u(a_i; a_j)$ , where  $u(a_i; a_j) \geq 0$ .
- ▶ It is assumed that individuals use pure strategies.
- ▶ Let  $p_{i,n}$  be the proportion of individuals using action  $i$  in generation  $n$ .
- ▶ The average reward of an individual using action  $i$  in generation  $n$  is

$$\bar{R}_{i,n} := u(a_i; p_{1,n}a_1 + \cdots + p_{m,n}a_m)$$

- ▶ The average reward in the population as a whole in generation  $n$  is

$$\bar{R}_n := \sum_{i=1}^m p_{i,n} \bar{R}_{i,n}$$

- ▶ The proportion of individuals using action  $i$  in generation  $n+1$  is

$$p_{i,n+1} := \frac{p_{i,n} \bar{R}_{i,n}}{\bar{R}_n}$$

- ▶ A fixpoint  $(p_1, \dots, p_m)$  of the replicator dynamic equations satisfies

$$p_i = \frac{p_i \bar{R}_i}{\bar{R}} \quad \text{for } i = 1, \dots, m$$

## Example

	A	B
A	5,5	0,0
B	0,0	1,1

- ▶ Let  $p_n$  be the proportion of individuals using  $A$  in generation  $n$ .
- ▶ The average reward of  $A$  players in generation  $n$  is

$$\bar{R}_{A,n} = u(A; p_n A + (1 - p_n) B) = 5p_n$$

- ▶ The average reward of  $B$  players in generation  $n$  is

$$\bar{R}_{B,n} = u(B; p_n A + (1 - p_n) B) = 1 - p_n$$

- ▶ The average reward of the population is

$$\bar{R}_n = p_n \bar{R}_{A,n} + (1 - p_n) \bar{R}_{B,n} = 1 - 2p_n + 6p_n^2$$

- ▶ The equation governing the replicator dynamics is

$$p_{n+1} = \frac{p_n \bar{R}_{A,n}}{\bar{R}_n} = \frac{5p_n^2}{1 - 2p_n + 6p_n^2}$$

- ▶ A fixpoint of these dynamics satisfies  $p = \frac{5p^2}{1 - 2p + 6p^2}$ .  $p = 0, 1, \frac{1}{6}$ .
- ▶  $p = 0, 1$  is an attractor.  $p = \frac{1}{6}$  is not an attractor.

## Fixpoints of the Replicator Dynamic Equations

- ▶ A fixed point  $\sigma^*$  is *stable* (also called Lyapunov stable) iff for all open neighborhoods  $U$  of  $\sigma^*$  there is another open neighborhood  $O \subset U$  such that any trajectory initially inside  $O$  remains inside  $U$ .
- ▶ A fixed point  $\sigma^*$  is *attractive* iff there exists an open neighborhood  $U$  of  $\sigma^*$  such that all trajectory initially in  $U$  converges to  $\sigma^*$ . The maximum possible  $U$  is called the basin of attraction of  $\sigma^*$ .
- ▶ A fixed point  $\sigma^*$  is *asymptotically stable* (also called attractor) iff it is stable and attractive.
- ▶ A fixed point is *globally asymptotically stable* iff its basin of attraction encompasses the whole space.
  
- ▶ Strict Nash Equilibria are attractors.
- ▶ If a fixpoint is stable then it is an Nash Equilibrium.
- ▶ ESSs are attractors.
- ▶ For  $2 \times 2$  matrix games a fixed point is an ESS iff it is an attractor.

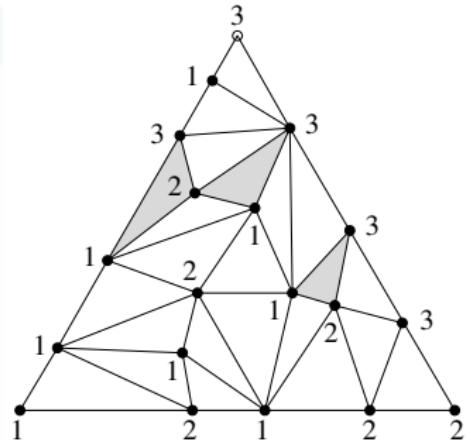
# Sperner's Lemma

## Lemma (Sperner's Lemma)

Suppose that some triangle with vertices  $V_1, V_2, V_3$  is triangulated.

The vertices in the triangulation get “colors” from  $\{1, 2, 3\}$  s.t. vertices on the edge  $(V_i, V_j)$  are colored either  $i$  or  $j$ , while the interior vertices are colored 1, 2 or 3.

Then in the triangulation there must be an odd number of “tricolored” triangles.

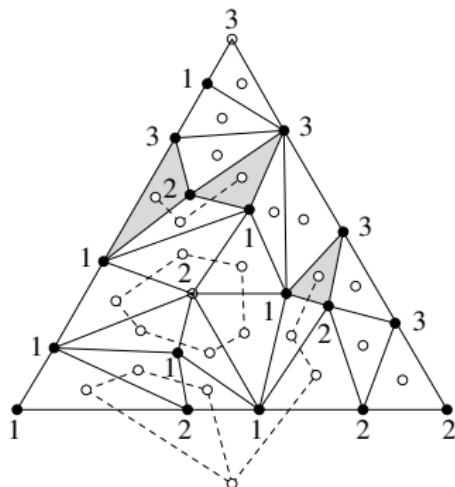


# Proof of Sperner's Lemma

## Proof.

Consider the dual graph to the triangulation — but take only those which cross an edge that has endvertices with the colors 1 and 2. Thus we get a partial dual graph which has degree 1 at all vertices that correspond to tricolored triangles, degree 2 for all triangles in which the two colors 1 and 2 appear, and degree 0 for triangles that do not have both colors 1 and 2.

The vertex of the dual graph which corresponds to the outside of the triangulation has odd degree: along the big edge from  $V_1$  to  $V_2$ , there is an odd number of changes between 1 and 2. Since the number of odd-degree vertices in any finite graph is even, the number of tricolored triangles is odd. □



# Brouwer Fixpoint Theorem

## Theorem (Brouwer Fixpoint Theorem)

Given a non-empty compact convex set  $X \subset \mathbb{R}^n$  and continuous function  $f : X \rightarrow X$ , there exists  $\mathbf{x} \in X$  s.t.  $f(\mathbf{x}) = \mathbf{x}$ .

## Theorem (Kakutani Fixpoint Theorem)

Given a non-empty compact convex set  $X \subset \mathbb{R}^n$  and a multi-valued function  $f : X \rightrightarrows X$ , if

- ▶ for all  $x \in X$ , the set  $f(x)$  is convex,
- ▶ the graph of  $f$  is closed (i.e. for all sequences  $\{x_n\}$  and  $\{y_n\}$  s.t.  $x_n \rightarrow x$ ,  $y_n \rightarrow y$  and  $y_n \in f(x_n)$ , we have  $y \in f(x)$ ),

then  $\exists x \in f(x)$ .

## Theorem (Schauder Fixpoint Theorem)

If  $K$  is a non-empty convex subset of a Hausdorff topological vector space  $V$  and  $T$  is a continuous mapping of  $K$  into itself such that  $T(K)$  is contained in a compact subset of  $K$ , then  $T$  has a fixpoint.

# Proof of Brouwer Fixpoint Theorem

## Proof.

Let  $\Delta$  be the triangle in  $\mathbb{R}^3$  with vertices  $e_1 = (1, 0, 0)$ ,  $e_2 = (0, 1, 0)$ , and  $e_3 = (0, 0, 1)$ . We prove that any continuous map  $f : \Delta \rightarrow \Delta$  has a fixpoint. Let  $\delta(T)$  be the maximal length of an edge in a triangulation  $T$ .

One can construct an infinite sequence of triangulations  $T_1, T_2, \dots$  of  $\Delta$  s.t.  
 $\lim_{k \rightarrow \infty} \delta(T_k) = 0$ .

Suppose  $f$  has no fixpoint. Since  $\sum_i v_i = 1 = \sum_i f(v)_i$ , for each of these triangulations, we can define a Sperner coloring of their vertices  $v$  by setting  $\lambda(v) := \min \{i : f(v)_i < v_i\}$ .

Sperner's lemma tells us that in each triangulation  $T_k$  there is a tricolored triangle  $\{v_1^k, v_2^k, v_3^k\}$  with  $\lambda(v_i^k) = i$ .

Since the simplex  $\Delta$  is compact, some subsequence of  $(v_1^k)_{k \geq 1}$  has a limit point  $v^* \in \Delta$ . Since  $\lim_{k \rightarrow \infty} \delta(T_k) = 0$ , the sequences  $v_2^k$  and  $v_3^k$  converge to the same point  $v^*$ .

Then  $\forall i : f(v^*)_i \leq v_i^*$ , which contradicts  $f(v^*) \neq v^*$ . □

## Theorem (Existence of Mixed Nash Equilibrium)

Every finite strategic game has a mixed Nash equilibrium.

### Proof.

Given a strategy profile  $\sigma \in \prod_{i \in N} \Delta S_i$ , define

$$\varphi_{i,s_i}(\sigma) := \max \{0, u_i(s_i; \sigma_{-i}) - u_i(\sigma)\}$$

Then define a continuous  $f : \prod_{i \in N} \Delta S_i \rightarrow \prod_{i \in N} \Delta S_i$  by  $f : \sigma \mapsto \sigma'$ , where

$$\sigma'_i(s_i) := \frac{\sigma_i(s_i) + \varphi_{i,s_i}(\sigma)}{\sum_{s_i \in S_i} [\sigma_i(s_i) + \varphi_{i,s_i}(\sigma)]} = \frac{\sigma_i(s_i) + \varphi_{i,s_i}(\sigma)}{1 + \sum_{s_i \in S_i} \varphi_{i,s_i}(\sigma)}$$

Since  $\prod_{i \in N} \Delta S_i$  is convex and compact,  $f$  has a fixpoint.

Consider any fixpoint  $\sigma$  of  $f$ . By the linearity of expectation there exists  $s'_i$  in the support of  $\sigma$ , for which  $u_i(s'_i; \sigma_{-i}) \leq u_i(\sigma)$ . Then

$$\varphi_{i,s'_i}(\sigma) = 0 \quad \& \quad \sigma'_i(s'_i) = \sigma_i(s'_i) \implies \forall i \in N \forall s_i \in S_i : \varphi_{i,s_i}(\sigma) = 0$$

## Walrasian Equilibrium

### Theorem (Existence of Walrasian Equilibrium)

Consider an economy with  $n$  goods  $X_1, \dots, X_n$  with a price vector  $(p_1, \dots, p_n) \in \Delta_n := \{x \in [0, 1]^n : \|x\|_1 = 1\}$ , and the prices of at least two goods are not zero. Assume that an excess demand function for each good  $f_i(p_1, \dots, p_n)$  is continuous and satisfies the following condition

$$\sum_{i=1}^n p_i f_i = 0 \quad (\text{Walras Law})$$

Then, there exists an equilibrium price vector  $(p_1^*, \dots, p_n^*)$  s.t.

$$f_i(p_1^*, \dots, p_n^*) \leq 0$$

for all  $i = 1, \dots, n$ . And when  $p_i > 0$  we have  $f_i(p_1^*, \dots, p_n^*) = 0$ .

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

**Game Theory**

Preferences and Expected Utility

Strategic Games

Extensive Games with Perfect Information

Eliminating Dominated Strategies

Dynamic Games

Repeated Games

Signaling Games

Mechanism Design

Counterfactual Regret Minimization

Subgame Perfect Equilibrium

Games with Incomplete Information

Evolutionary Games

**Voting System**

Reinforcement Learning

Deep Learning

Artificial General Intelligence

What If Computers Could Think?

**References**

1753

# 群体聚合 Group Aggregation

提案:

- A 大学教育应该免费.
- B 人人都应该有上大学的权利.
- C 政府应该增加在大学教育上的投入.

$$\frac{A \quad B}{C}$$

民意调查:

甲:  $A, B$

A: 2 票

乙:  $A, \neg C$

B: 2 票

丙:  $B, \neg C$

$\neg C$ : 2 票

**Remark:** 如果是序贯投票, 则可以通过操纵投票顺序来控制选举结果.

# 投票制度

1. **相对多数法 (plurality method)**: 选民每人投一个候选人
2. **双选投票法 (vote-for-two method)**: 选民每人投两个候选人
3. **反相对多数法 (anti-plurality method)**: 选民每人投  $n - 1$  个候选人, 即投票剔除一个候选人
4. **绩点评分法 (GPA method)**: 选民对所有  $n$  个候选人进行排序
  - ▶ 第一名得  $n - 1$  分
  - ▶ 第二名得  $n - 2$  分
  - ▶ 倒数第二名得 1 分
  - ▶ 最后一名 0 分GPA 得分最高者当选.
5. **排序选择法 (ranked-choice voting)**: 选民对所有  $n$  个候选人进行排序
  - 5.1 首轮计票时, 只计算选民的首选票. 如果有候选人获得过半数选票, 该候选人当选.
  - 5.2 如果没有候选人过半数, 就将票数最少的候选人淘汰, 把选择该候选人为首选的选票, 按次选票的候选人重新分配.
  - 5.3 重复第 2 步, 直到某个候选人获得过半数选票为止.

# 选举结果反映的究竟是民意还是选取方法的选择?

2	a > b > c > d
2	a > d > c > b
2	c > b > d > a
3	d > b > c > a

Table: 9 个选民, 4 个候选人

- ▶ 根据“相对多数法”, a 赢: a4 > d3 > c2 > b0
- ▶ 根据“双选投票法”, b 赢: b7 > d5 > a4 > c2
- ▶ 根据“反相对多数法”, c 赢: c9 > b7 = d7 > a4
- ▶ 根据“绩点评分法”, d 赢: d15 > b14 > c13 > a12

# 当有候选人退出时会怎样?

3	a > c > d > b
6	a > d > c > b
3	b > c > d > a
5	b > d > c > a
2	c > b > d > a
5	c > d > b > a
2	d > b > c > a
4	d > c > b > a

Table: 30 个选民, 4 个候选人

- ▶ 根据“相对多数法”, a 赢: a9 > b8 > c7 > d6
- ▶ 如果 d 退出: c11 > b 10 > a9
- ▶ 如果 c 退出: d11 > b 10 > a9
- ▶ 如果 b 退出: d11 > c 10 > a9
- ▶ 如果 a 退出: d12 > c 10 > b8
- ▶ 根据“绩点评分法”, d 赢: d58 > c54 > b41 > a27
- ▶ 根据“排序选择法”, c 赢: c20 > b10

# Arrow's Impossibility Theorem

## Theorem (Arrow's Impossibility theorem)

*In an election with candidates  $\geq 3$ , any voting system that is unanimous and independence of irrelevant alternatives must be a dictatorship!*

- ▶ unanimous: if everyone agrees, consensus decides the outcome
- ▶ independence of irrelevant alternatives: a voter can't move  $a$  above  $b$  by lying about how they feel about  $c$
- ▶ nondictatorship: no single voter gets to decide the outcome

# Arrow's Impossibility Theorem

Let  $N$  be a set of voters, and  $C$  a set of candidates. A social welfare function (SWF) is  $f : \mathcal{S}_C^N \rightarrow \mathcal{S}_C$ , where  $\mathcal{S}_C$  is the set of all permutations on  $C$ . We write  $a \succ_i b$  to indicate that voter  $i \in N$  ranks  $a$  above  $b$ . Given  $\sigma \in \mathcal{S}_C^N$ ,  $N_{a \succ b}^\sigma := \{i \in N : a \succ_i b \text{ under } \sigma\}$ .

- ▶ Unanimity (**U**): If all voters rank  $a$  above  $b$ , then so does society:  
 $N_{a \succ b}^\sigma = N \implies a \succ_{f(\sigma)} b$ .
- ▶ Independence of Irrelevant Alternatives (**IIA**): the relative social ranking of two candidates only depends on their relative individual rankings:  $N_{a \succ b}^\sigma = N_{a \succ b}^{\sigma'} \implies (a \succ_{f(\sigma)} b \iff a \succ_{f(\sigma')} b)$ .
- ▶ Nondictatorship (**ND**): There is no  $i \in N$  s.t.  $\sigma_i = f(\sigma)$ .

## Theorem (Arrow's Impossibility Theorem)

If  $N$  is finite and  $|C| \geq 3$ , then any SWF that satisfy **U** and **IIA** must be a dictatorship.

## Proof Sketch of Arrow's Impossibility Theorem

1. We call a subset  $A \subset N$  decisive iff whenever all  $x \in A$  present the same ranking, the SWF  $f$  outputs that ranking.
2. The set of decisive sets of voters  $\mathcal{F} := \{A \subset N : A \text{ is decisive}\}$  is an ultrafilter.
3. If  $N$  is finite, then the ultrafilter  $\mathcal{F}$  must be a principle ultrafilter.

Let  $\mathcal{F}$  be an ultrafilter on  $N$ . We can define a SWF  $f$  by declaring the output to be that unique permutation  $\sigma$  with the property that  $\{i \in N : \sigma_i = \sigma\} \in \mathcal{F}$ .

### Theorem (Arrow's Theorem)

Assume  $|C| \geq 3$ . There is a 1 – 1 correspondence between ultrafilters on  $N$  and SWF that satisfy **U** and **IIA**. The non-dictatorship SWFs are those corresponding to non-principle ultrafilters. In particular, Arrow's impossibility theorem is equivalent to the assertion that all ultrafilters on a finite set are principle.

# 对无关选项独立性 IIA 的质疑

排序	张三	李四	王五	赵六
1	x	x	a	a
2	a	a	b	b
3	b	b	y	y
4	y	y	x	x

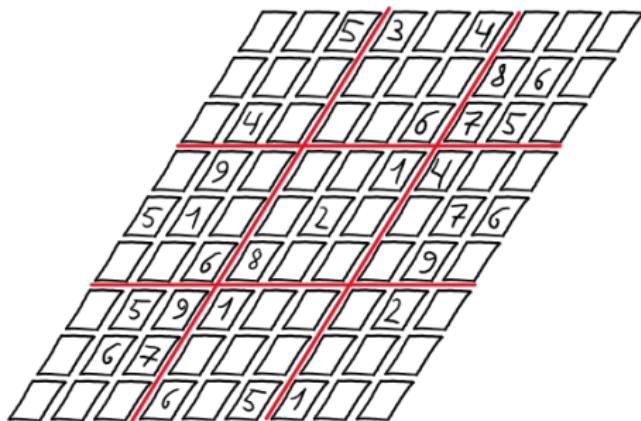
排序	张三	李四	王五	赵六
1	a	a	y	y
2	b	b	a	a
3	x	x	b	b
4	y	y	x	x

- ▶ 在两个表中,  $x$  和  $y$  的相对排序是一样的. 张三和李四偏好  $x \succ y$ , 王五和赵六偏好  $y \succ x$ .
- ▶ 直观上, 表 1 中偏好  $x \succ y$  的人比偏好  $y \succ x$  的人的“感受更为强烈”; 表 2 则相反. 我们可能会说, 表 1 中  $x \succ y$  是有道理的, 而表 2 中  $y \succ x$  是有道理的.
- ▶ IIA 要求两表中社会对  $x$  和  $y$  的偏好一样: 要么  $x \succ y$ , 要么  $y \succ x$ .

- ▶ 计算社会选择理论: 用计算复杂性防止坏情况的发生
- ▶ 用逻辑做投票协议验证
  - ▶ 隐私性: 没有别人能知道你投的是谁
  - ▶ 无收据性: 你不能证明给别人你投了特定人的票
  - ▶ 可核查性: 你自己能检查你的票是不是被算进去了
  - ▶ 公平性: 之前投票的部分结果不会影响之后投票的结果

# 零知识证明 Zero-Knowledge Proof

- ▶ 小帅给小美出了一道非常难的数独题.
- ▶ 小美怎么也解不出来, 怀疑小帅在耍她: “这道题无解!”
- ▶ 小帅说: “我会用零知识证明的方法给你证明这题有解. 我不会把解给你看, 却能让你信服我确实有这题的解.”



提示: 准备九个袋子 😊

# Prover vs Verifier

- ▶ **Completeness:**  $P$  can convince  $V$  if  $X$  is true.
- ▶ **Soundness:** no (efficient)  $P$  can convince  $V$  if  $X$  is not true.
- ▶ **Zero Knowledge:** no efficient  $V$  learns anything more than the validity of  $X$ .

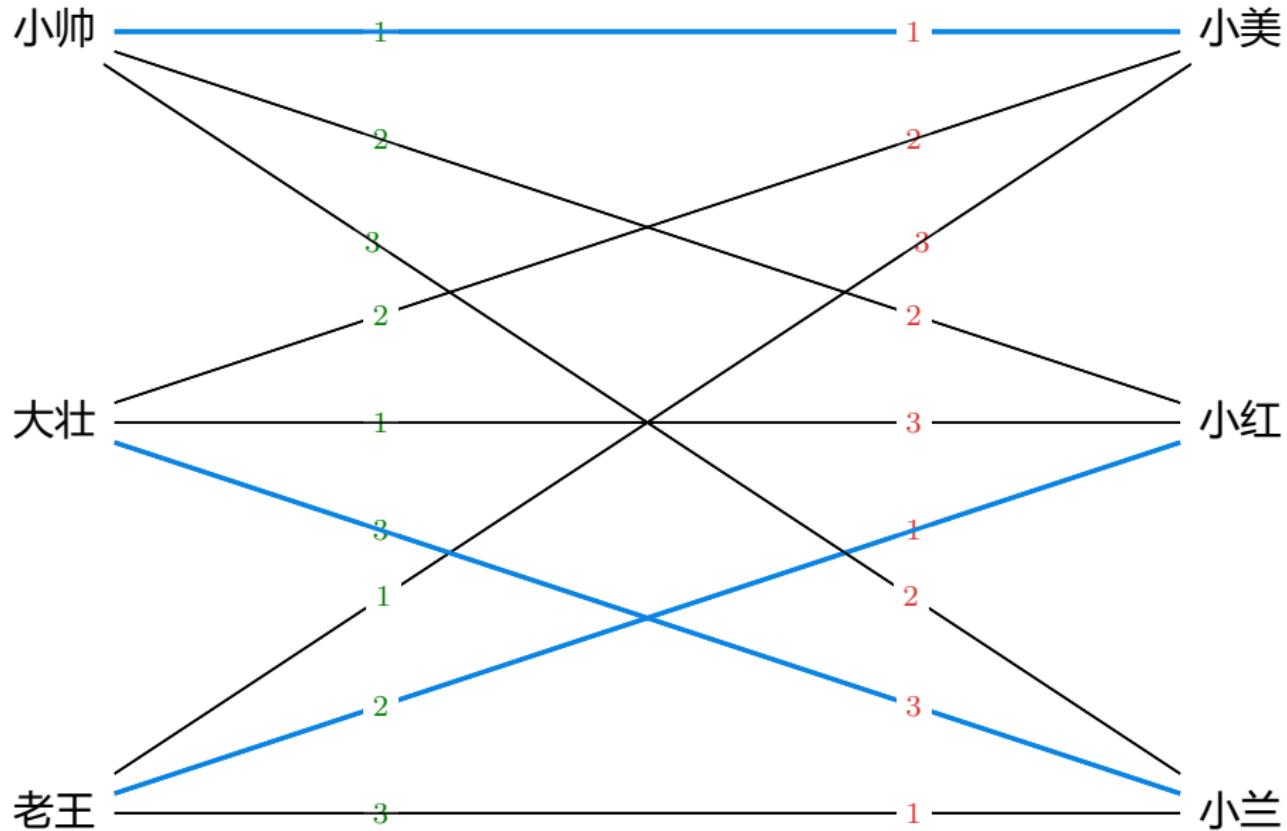
**Remark:** The adversary can simulate the proof without knowing the prover's witness.

## Theorem

*Every NP statement can be proven in zero-knowledge.*

# 稳定婚姻问题

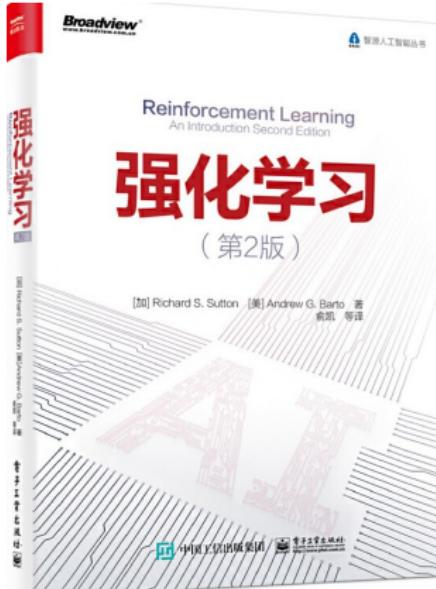
- ▶  $n$  男  $n$  女, 每人对异性都有偏好排序.
- ▶ 一个匹配是**不稳定的**, 如果有一对男女, 相对于自己的现任配偶, 更中意彼此.
- ▶ 稳定匹配是否存在?
- ▶ 怎么找到稳定匹配?
  - ▶ Day 0: 每人对异性进行排序.
  - ▶ Day 1: (a.m.) 每个男生向他最中意的女生表白.  
(p.m.) 每个女生选择自己最中意的求爱对象, 拒绝其他男生.
  - ▶ Day  $n + 1$ : (a.m.) 依然单身的男生向还没拒绝过他的最中意的女生表白, 不管对方是否已有男朋友.  
(p.m.) 每个女生选择自己最中意的求爱对象, 拒绝/抛弃其他男生.
  - ▶ 直到所有人都有对象为止.
- ▶ 在男生主动表白的稳定匹配中, 每个男生都获得了他所能获得的最佳配偶!
- ▶ 每个女生都获得了她所能获得的最差配偶!



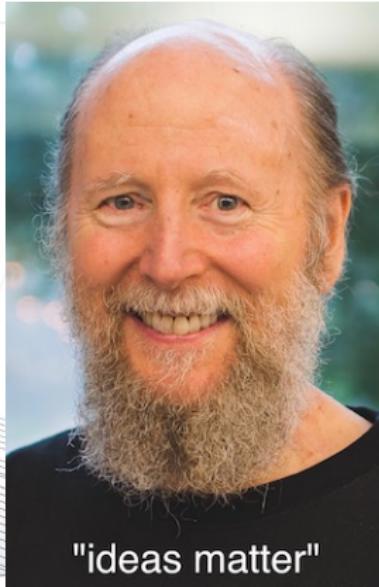
# Contents

Introduction	Game Theory
Philosophy of Induction	Reinforcement Learning
Inductive Logic	Deep Learning
Universal Induction	Artificial General Intelligence
Causal Inference	What If Computers Could Think? References 1753

# Reinforcement Learning



(a) Reinforcement Learning



(b) Richard S. Sutton



(c) Andrew G. Barto

## Pleasure = Immediate Reward $\neq$ Good = Long-term Reward

*"Every art and every inquiry, and similarly every action and pursuit, is thought to aim at some good."*

— Aristotle

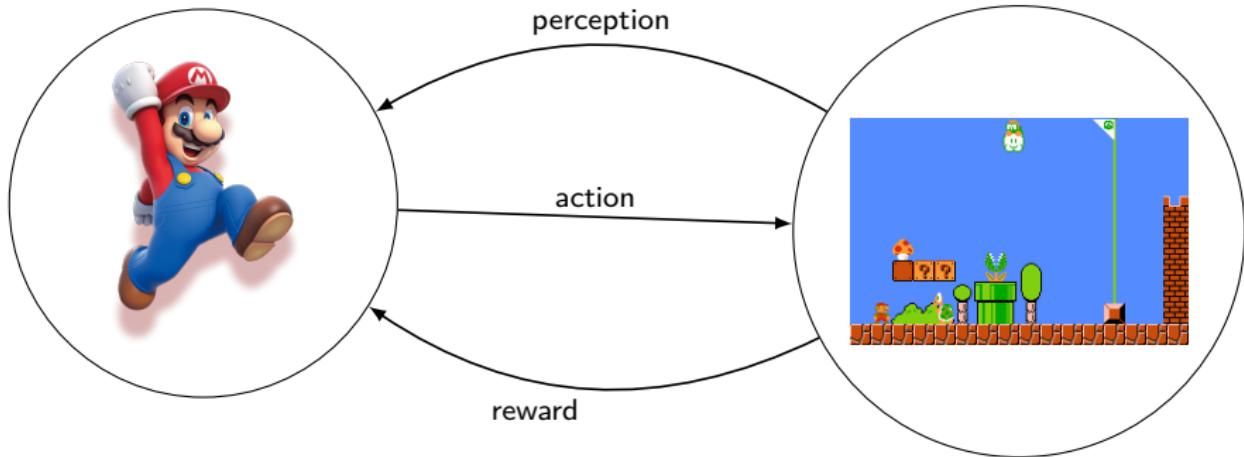
*"Even enjoying yourself you call evil whenever it leads to the loss of a pleasure greater than its own, or lays up pains that outweigh its pleasures. ...Isn't it the same when we turn back to pain? To suffer pain you call good when it either rids us of greater pains than its own or leads to pleasures that outweigh them."*

— Plato

Value function = prediction of reward = the sum of upcoming (pleasure – pain)

- ▶ **The reward hypothesis:** All goals can be represented as the maximization of expected cumulative reward.
- ▶ **The reward-is-enough hypothesis:** Intelligence, and its associated abilities, can be understood as subserving the maximization of reward by an agent acting in its environment.

# Reinforcement Learning

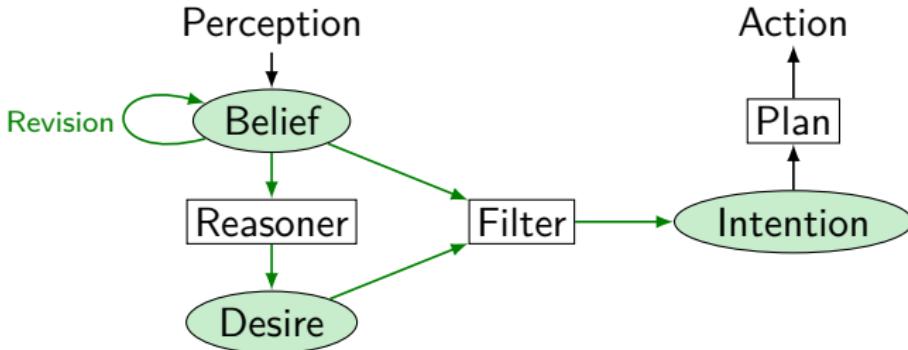


Goal: Maximize reward!

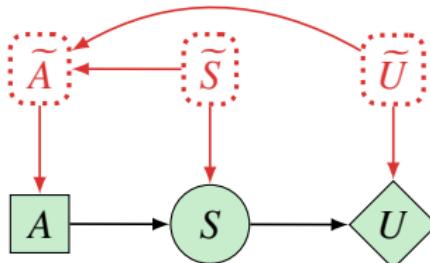
von Neumann–Morgenstern: 如果偏好关系  $\succ$  满足完备性、传递性、连续性和独立性，则存在一个函数  $u$  使得  $a \succ b$  当且仅当  $u(a) > u(b)$ .<sup>17</sup>

<sup>17</sup>如果你质疑效用理论，请质疑偏好关系的合理性，而不是说一句“人不是冰冷地追求效用最大化”。

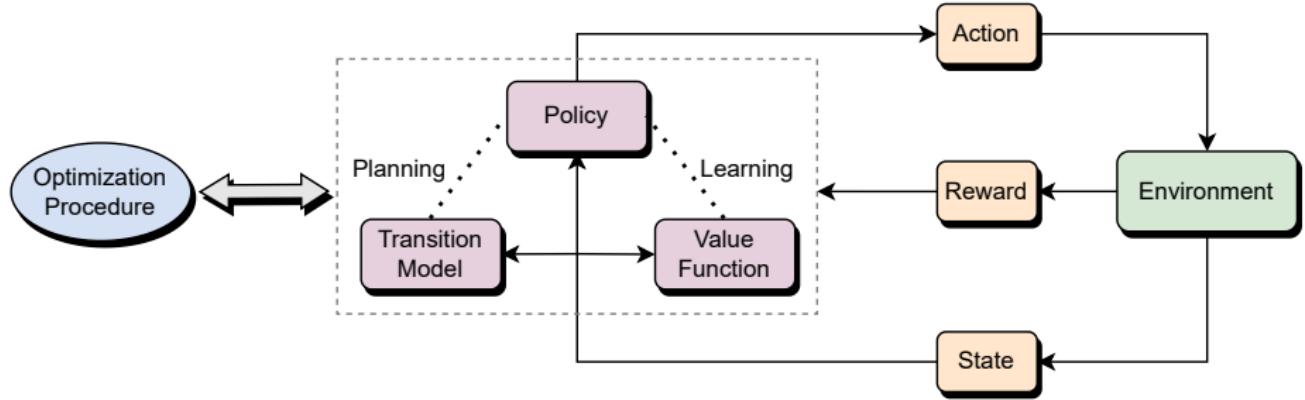
# 哲学家 Bratman 的 BDI 模型

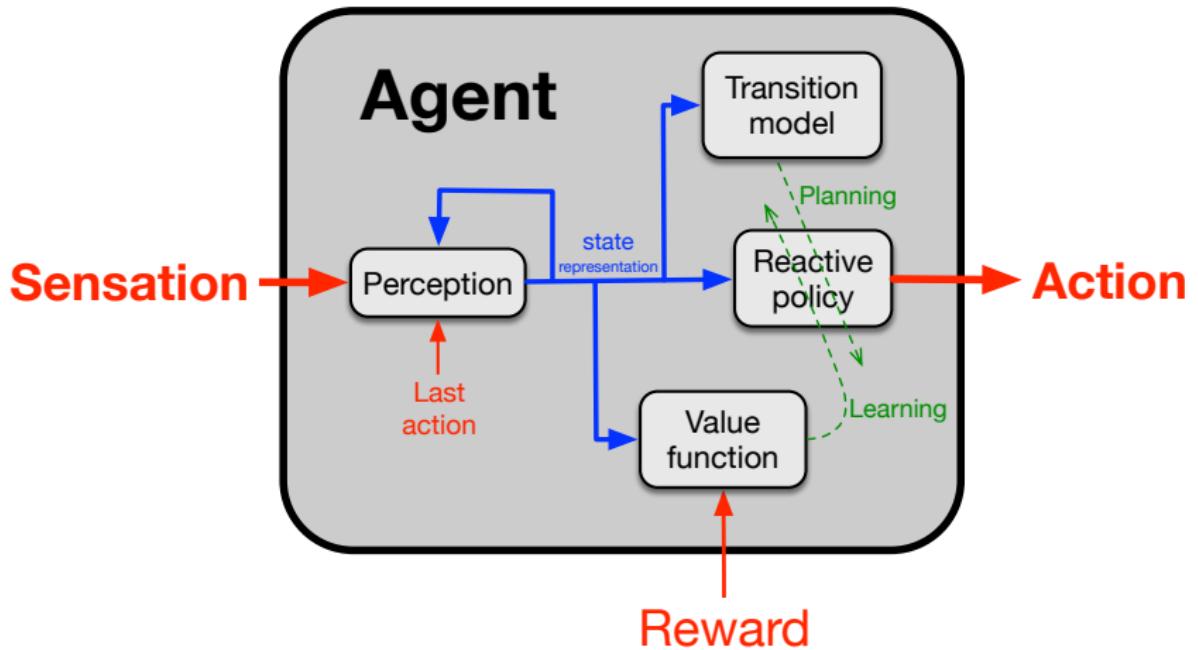


- ▶ **Beliefs** about itself, other agents and its environment  
 $\pi(a | s), \pi'(a | s), P(s' | s, a)$
- ▶ **Desires** about future states (goals)  $U(s)$
- ▶ **Intentions** about its own future plan  $\text{argmax}_\pi \mathbb{E}[U | \text{do}(\pi)]$



# Reinforcement Learning Agent





1. **Perception** produces the state representation
2. **Reactive Policy** quickly produces an action appropriate to the state
3. **Value Function** evaluates how well things are going, and changes the policy (learning)
4. **Transition model** predicts the consequences of alternate actions, and changes the policy (planning)

# Will intelligence ultimately be explained in

Objective terms?

- ▶ states of the external world
- ▶ objects, people, places, relationships, atoms
- ▶ space, motion, distances
- ▶ things outside the agent

or

Experiential terms?

- ▶ sensations
- ▶ actions
- ▶ reward
- ▶ time steps
- ▶ things inside the agent

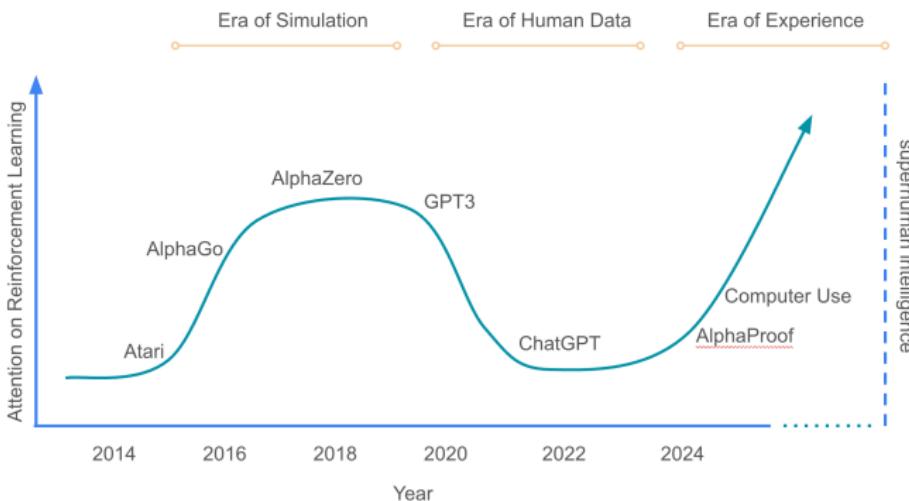
# 从“人类中心 Agent”到“经验 Agent”

- ▶ RL Agent 没有实时监督, 只有**延迟的奖励信号**. (长期奖励 = Good)
- ▶ **价值函数**将状态映射到对未来奖励的预测. (为了消除延迟)
- ▶ Agent 只能通过**传感器**获取关于世界的信息, 只能通过**行动**影响世界.
- ▶ **经验**是 Agent 接触世界的唯一途径.
- ▶ Agent 通过经验试错学习.
- ▶ **客观状态** vs **经验状态** — 完全以经验定义的世界状态.
- ▶ **经验状态**是对过去经验的总结, 用于预测和控制未来经验.
- ▶ Agent 的**奖励函数**应学自其在环境中的交互经验, 而非人类预设的偏好.
- ▶ Agent 应基于其自身经验进行**规划/推理**, 而非套用人类的表征和推理框架.
- ▶ **经验知识**是关于世界的**状态**和**状态转换**的. 知道就是预测经验.
- ▶ 思维就是在想象的经验中学习. 一切都关乎经验.
- ▶ 利弊: 超越人类上限. 行动策略和奖励函数都可适应性调整.  
— 安全对齐更难, 可信任性可解释性更差.

# The Era of Experience

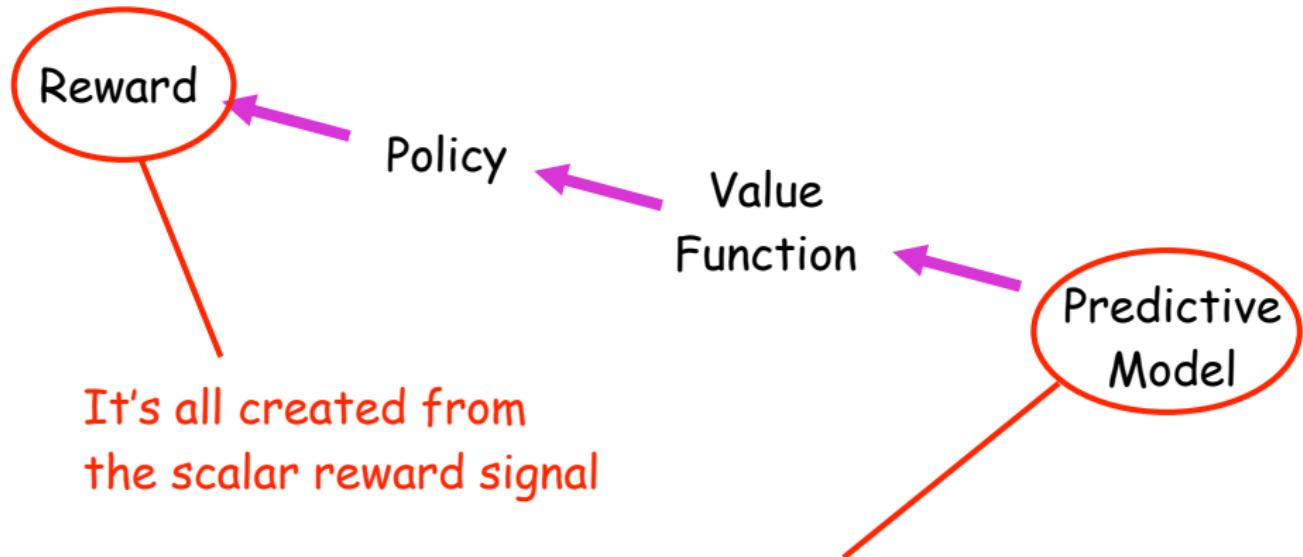
*“What we want is a machine that can learn from experience.”*

— Alan Turing 1947



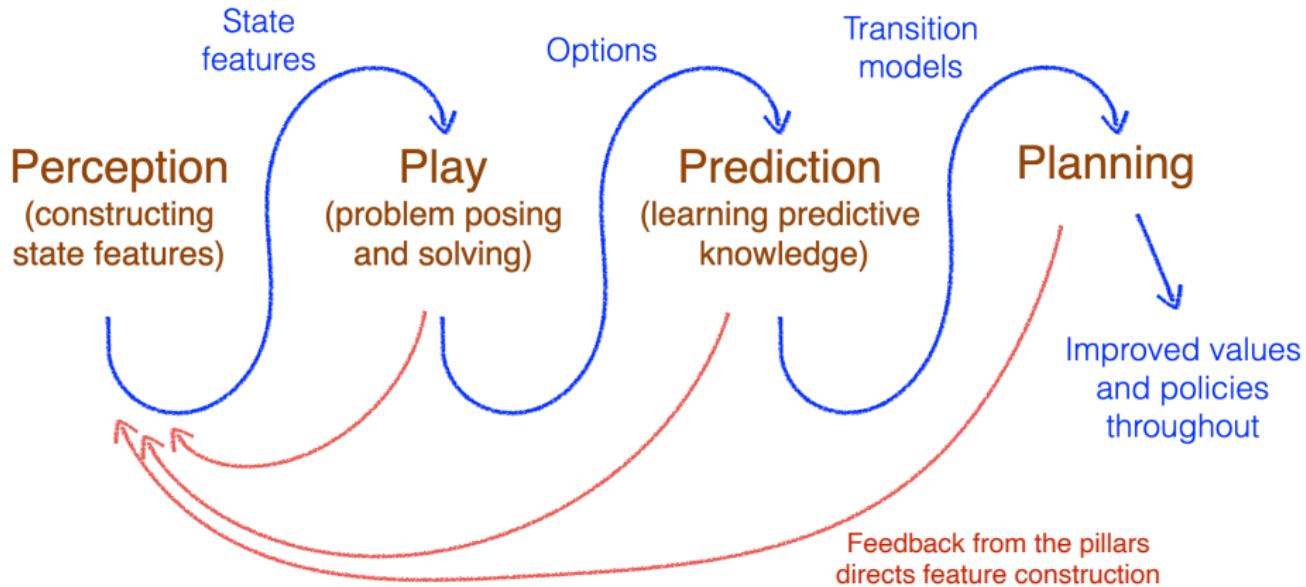
1. 模拟时代: 封闭环境, 清晰奖励
2. 人类数据时代: 大规模人类数据领域知识驱动的学习, 通用性, 对齐人
3. 经验时代: Agent 自己与环境交互的经验数据驱动的通用搜索和学习, 达成目标, 超越人

# RL's Computational Theory of Mind



**together with the causal structure of the world**

# RL's Computational Theory of Mind



1. Perception: learns state-update function  $s_t = \Phi(s_{t-1}, a_{t-1}, o_{t-1})$ .
2. Play: poses and solves many **subproblems/subtasks** to attain features, outputs policies  $\pi$  with termination conditions that attain the features.
3. Prediction: learn a cause-and-effect transition model  $P$ .
4. Planning: improves value functions  $V$  and policies  $\pi$ .

## Much of Mind is about Prediction

- ▶ Perception and State Representation can be thought of as making predictions
- ▶ Models the world and cause and effect can be thought of in terms of predictions
- ▶ Planning can be thought of as composing predictions to anticipate possible futures, and then choosing among them
- ▶ Learning Value Functions is earning predictions

# To Know is to Predict Experience

Knowledge is predictions

- ▶ of what will happen
- ▶ of what you could cause to happen
  - at various time scales
  - conditional on actions or courses of action

Predictive Knowledge should be

1. Learnable — from low-level sensorimotor data
  - Autonomously verifiable
2. Expressive — able to express abstract, high-level facts as well as specific, low-level facts
3. Useful — for action and planning

## Definition (Markov Decision Process MDP[SB18])

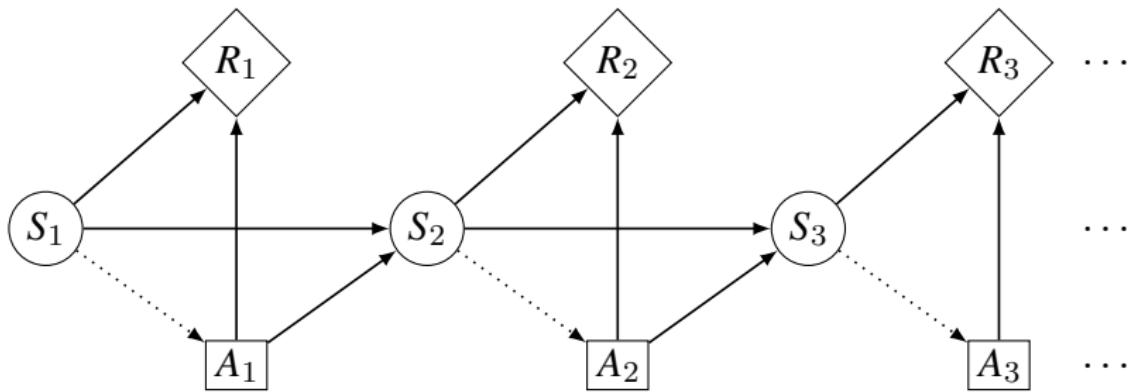
An MDP for an accessible, stochastic environment is defined by

- ▶ Set of states  $\mathcal{S}$
- ▶ Set of actions  $\mathcal{A}$
- ▶ Set of rewards  $\mathcal{R}$
- ▶ Transition model  $P(s', r | s, a)$ , with  $s, s' \in \mathcal{S}$ ,  $a \in \mathcal{A}$  and  $r \in \mathcal{R}$
- ▶ Reward function  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$

$$r(s, a) := \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} P(s', r | s, a)$$

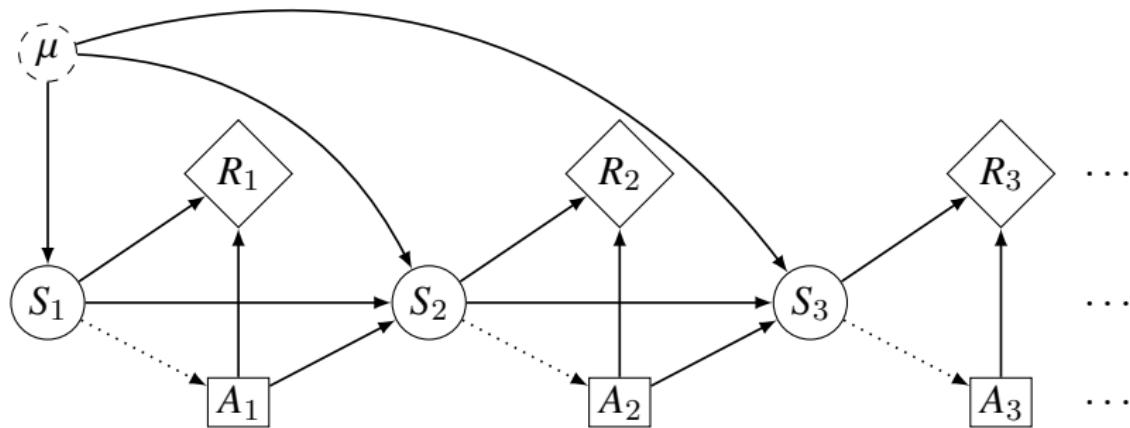
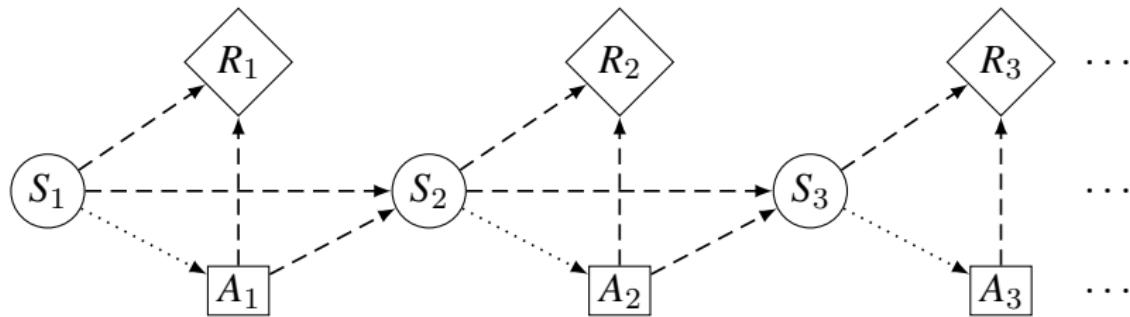
- ▶ **Transition model:**  $P(s' | s, a) := \sum_{r \in \mathcal{R}} P(s', r | s, a)$  is the probability that state  $s'$  is reached, if action  $a$  is executed in state  $s$ .
- ▶ **Policy:** Complete mapping  $\pi$  that specifies for each state  $s \in \mathcal{S}$  which action  $\pi(s) \in \mathcal{A}$  to take.
- ▶ **Wanted:** The optimal policy  $\pi^*$  is the policy that maximizes the future expected reward.

## Known MDP

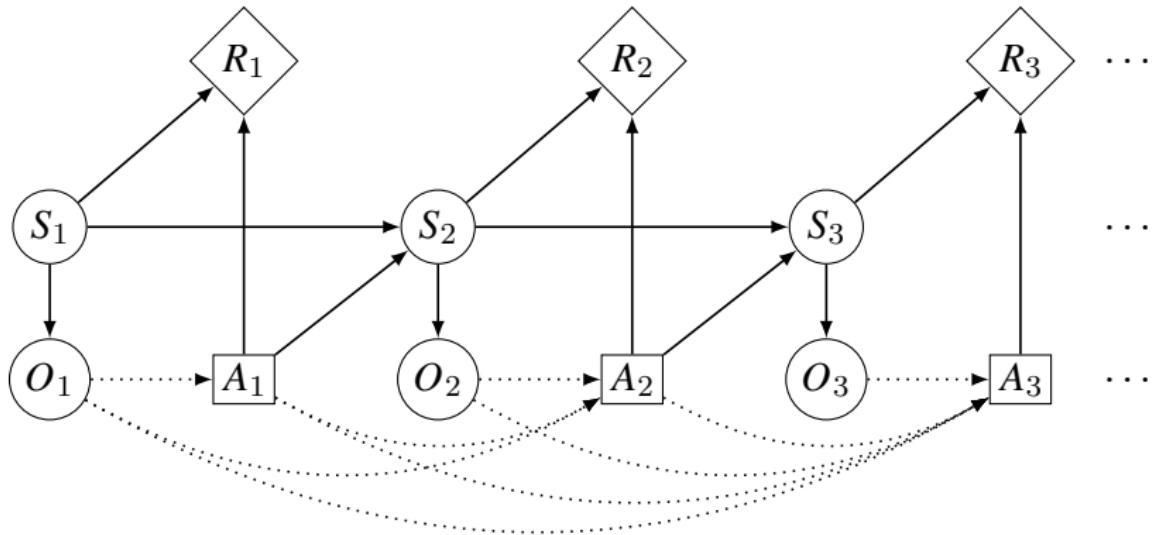


状态转移  $s_{t+1} = f_s(s_t, a_t, \varepsilon_{s_{t+1}}) \sim P(s_{t+1} | s_t, a_t)$   
动作  $a_t = f_a(s_t, \varepsilon_{a_t}) \sim \pi(a_t | s_t)$   
奖励  $r_t = f_r(s_t, a_t, \varepsilon_{r_t}) \sim r(s_t, a_t)$

## Two Representations of an Unknown MDP

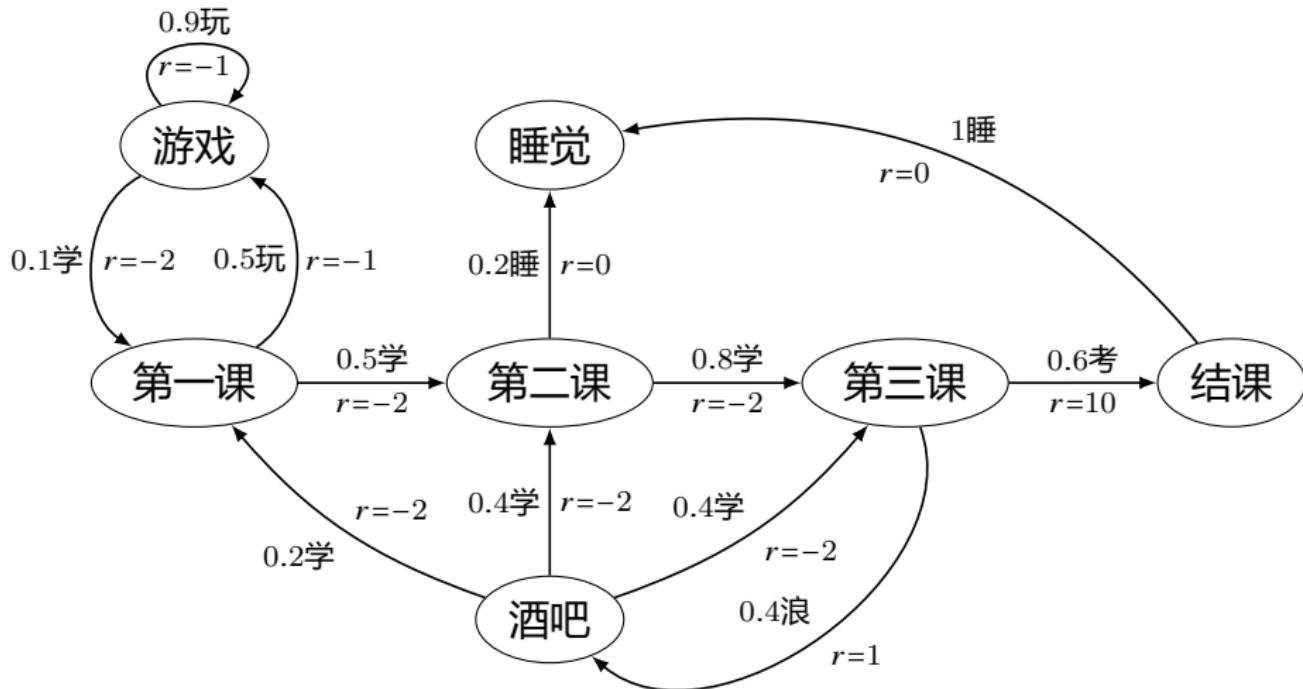


# POMDP

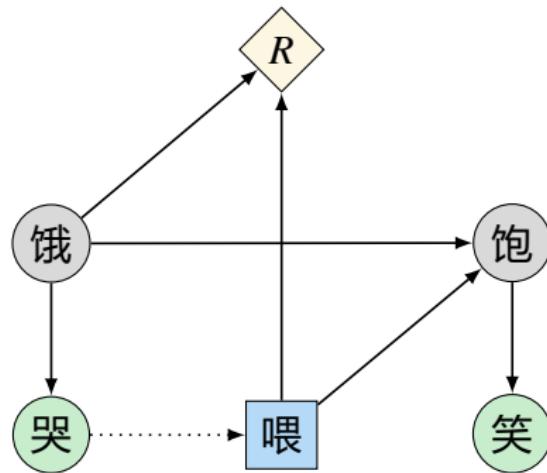


状态转移	$s_{t+1} = f_s(s_t, a_t, \varepsilon_{s_{t+1}})$	$\sim P(s_{t+1}   s_t, a_t)$
感知	$o_t = f_o(s_t, \varepsilon_{o_t})$	$\sim P(o_t   s_t)$
动作	$a_t = f_a(h_{<t}, \varepsilon_{a_t})$	$\sim \pi(a_t   h_{<t})$
奖励	$r_t = f_r(s_t, a_t, \varepsilon_{r_t})$	$\sim r(s_t, a_t)$

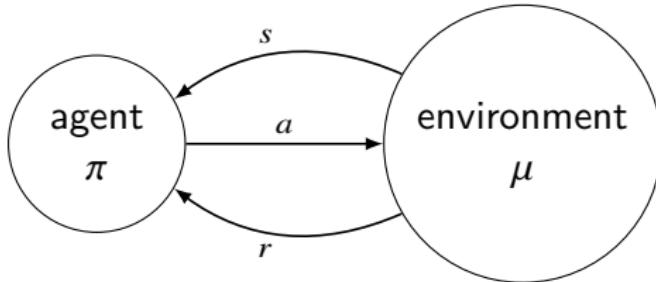
## Example: MDP



## Example: POMDP



# Value Function



Definition (Value of a state under  $\pi$ )

$$V^\pi(s) := \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| S_t = s \right]$$

Definition (Action-value under  $\pi$ )

$$Q^\pi(s, a) := \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| S_t = s, A_t = a \right]$$

Sutton: “Ethics is just values held in common by many agents.”

# What is Evil?

*The only good is knowledge.*

*The only evil is ignorance.*

— Socrates

*It's the belief that your greed or grievance supersedes all standard norms of society. When you elevate your grievance above those universal norms of society, that's evil.*

— Judea Pearl

# From MDP to Reinforcement Learning

- ▶ Markov decision process (offline)
  - ▶ Have mental model of how the world works.
  - ▶ Find policy to collect maximum rewards.
- ▶ Reinforcement learning (online)
  - ▶ Don't know how the world works.
  - ▶ Perform actions in the world to find out and collect rewards.
- ▶ On-policy: estimate the value of data-generating policy
- ▶ Off-policy: estimate the value of another policy

# Bellman Expectation Equations

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) Q^\pi(s, a)$$

$$Q^\pi(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^\pi(s')$$

## Bellman Expectation Equations

$$V^\pi(s) = \mathbb{E} [r + \gamma V^\pi(s') \mid s] = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left( r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^\pi(s') \right)$$

$$Q^\pi(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) \sum_{a' \in \mathcal{A}} \pi(a' \mid s') Q^\pi(s', a')$$

advantage:  $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$

# Optimal Value & Optimal Policy

## Definition (Optimal Value)

$$V^*(s) := \max_{\pi} V^{\pi}(s)$$

$$Q^*(s, a) := \max_{\pi} Q^{\pi}(s, a)$$

## Definition (Optimal Policy)

A policy  $\pi$  is called optimal iff  $\forall s \in \mathcal{S} : V^{\pi}(s) = V^*(s)$ .

— or equivalently,  $Q^{\pi}(s, a) = Q^*(s, a)$ .

$$\pi^*(s) = \operatorname{argmax}_a Q^*(s, a)$$

## Remark:

- ▶ We maximize reward by finding and following an optimal policy  $\pi^*$ .
- ▶ To find  $\pi^*$  we need to first find the optimal value function  $Q^*$ .
- ▶ To find  $Q^*$  we need to repeatedly find the value function for a policy  $Q^{\pi}$  that is our current best guess at the optimal policy.
- ▶ To find  $Q^{\pi}$  we may need to learn a transition model  $P(s', r | s, a)$ .

# Bellman Optimality Equations

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$$

$$Q^*(s, a) = r(s, a) + \gamma \sum_{s' \in S} P(s' | s, a) V^*(s')$$

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in S} P(s' | s, a) V^*(s') \right\}$$

$$Q^*(s, a) = r(s, a) + \gamma \sum_{s' \in S} P(s' | s, a) \max_{a' \in \mathcal{A}} Q^*(s', a')$$

# Action/Policy Evaluation Operator & Greedy Policy

## Definition (Action Evaluation Operator)

$$T_a V(s) := r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) V(s')$$

## Definition (Policy Evaluation Operator)

$$T^\pi V(s) := \sum_{a \in \mathcal{A}} \pi(a | s) T_a V(s) = \sum_{a \in \mathcal{A}} \pi(a | s) \left( r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) V(s') \right)$$

$$T^* V(s) := \max_{a \in \mathcal{A}} T_a V(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) V(s') \right\}$$

## Definition (Greedy policy)

Policy  $\pi$  is greedy w.r.t.  $V$  iff  $T^\pi V = T^* V$ .

# Banach's Fixpoint Theorem

## Theorem (Banach's Fixpoint Theorem)

Let  $\mathcal{V}$  be a Banach space and  $T : \mathcal{V} \rightarrow \mathcal{V}$  be a contraction mapping, with Lipschitz constant  $\gamma < 1$ . Then  $T$  has a unique fixpoint  $v \in \mathcal{V}$ . Further, for each  $v_0 \in \mathcal{V}$ ,  $\lim_{n \rightarrow \infty} \|T^n(v_0) - v\| = 0$ , and the convergence is geometric:

$$\|T^n(v_0) - v\| \leq \gamma^n \|v_0 - v\|$$

**Remark:** 在北京把北京地图随便往地上一摊, 总存在地图上至少一点, 它对应的位置正是它所处的位置. 其实不需要摊开, 捏成一团也行, 只要不撕破地图.

# Application of Banach's Fixpoint Theorem

## Theorem

$(\mathcal{V}, \|\cdot\|_\infty)$  is a Banach space, where  $\mathcal{V} := \{V \in \mathbb{R}^S : \|V\|_\infty < \infty\}$  and  $\|V\|_\infty := \max_{s \in S} |V(s)|$ .

- ▶  $T^\pi$  is a contraction, and  $V^\pi$  is the unique fixpoint of  $T^\pi$ .

$$\lim_{n \rightarrow \infty} \|(T^\pi)^n V_0 - V^\pi\|_\infty = 0$$

- ▶  $T^*$  is a contraction, and  $V^*$  is the unique fixpoint of  $T^*$ .

$$\lim_{n \rightarrow \infty} \|(T^*)^n V_0 - V^*\|_\infty = 0$$

## Application of Banach's Fixpoint Theorem

$$T^\pi Q(s, a) := r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) \sum_{a' \in \mathcal{A}} \pi(a' | s') Q(s', a')$$

$$T^* Q(s, a) := r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) \max_{a' \in \mathcal{A}} Q(s', a')$$

- ▶  $T^\pi$  is a contraction, and  $Q^\pi$  is the unique fixpoint of  $T^\pi$ .

$$\lim_{n \rightarrow \infty} \|(T^\pi)^n Q_0 - Q^\pi\|_\infty = 0$$

- ▶  $T^*$  is a contraction, and  $Q^*$  is the unique fixpoint of  $T^*$ .

$$\lim_{n \rightarrow \infty} \|(T^*)^n Q_0 - Q^*\|_\infty = 0$$

## Policy Improvement Theorem

### Theorem (Fixpoint of Bellman Optimality Operator)

Let  $V$  be the fixpoint of  $T^*$  and assume that there is policy  $\pi$  which is greedy w.r.t  $V$ . Then  $V = V^*$  and  $\pi$  is an optimal policy.

### Theorem (Policy Improvement Theorem)

Given two policies  $\pi$  and  $\pi'$ ,

$$\forall s \in \mathcal{S} \left( Q^\pi(s, \pi'(s)) \geq V^\pi(s) \right) \implies \forall s \in \mathcal{S} \left( V^{\pi'}(s) \geq V^\pi(s) \right)$$

where

$$Q^\pi(s, \pi'(s)) := \sum_{a \in \mathcal{A}} \pi'(a \mid s) Q^\pi(s, a)$$

**Remark:** In particular, the greedy policy meets the conditions of the policy improvement theorem.

The process of making a new policy  $\pi'$  that improves on an original policy  $\pi$ , by making it greedy with respect to  $V^\pi$ , is called policy improvement.

# Solving MDPs — Finite-Horizon Dynamic Programming

Principle of optimality: the tail of an optimal policy is optimal for the “tail” problem.

## Backward Induction

- ▶ Backward recursion:  $V_N^*(s) = r_N(s)$  and for  $k = N - 1, \dots, 0$

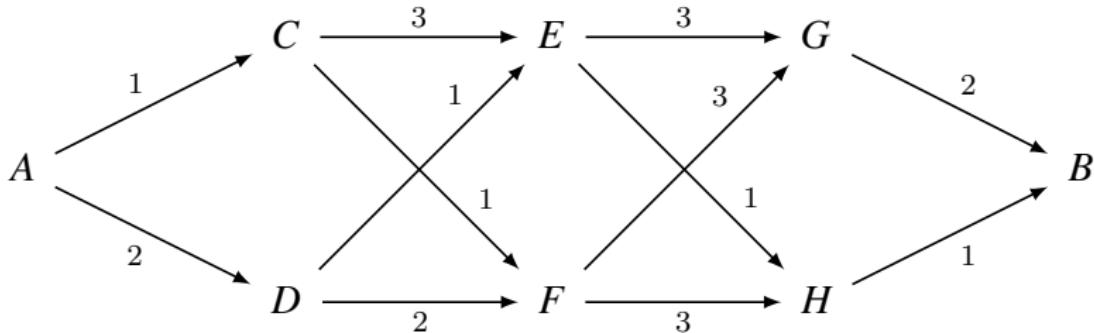
$$V_k^*(s) = \max_{a \in \mathcal{A}_k} \left\{ r_k(s, a) + \sum_{s' \in \mathcal{S}_{k+1}} P_k(s' | s, a) V_{k+1}^*(s') \right\}$$

- ▶ Optimal policy: for  $k = 0, \dots, N - 1$

$$\pi_k^*(s) \in \operatorname{argmax}_{a \in \mathcal{A}_k} \left\{ r_k(s, a) + \sum_{s' \in \mathcal{S}_{k+1}} P_k(s' | s, a) V_{k+1}^*(s') \right\}$$

- ▶ Cost:  $N|\mathcal{S}||\mathcal{A}|$  vs  $|\mathcal{A}|^{N|\mathcal{S}|}$  of brute force policy search.
- ▶ From now on, we will consider infinite-horizon discounted MDPs.

## Example — Dynamic Programming



$$D_E = \min(d(A, C) + d(C, E), d(A, D) + d(D, E)) = 3$$

$$D_F = \min(d(A, C) + d(C, F), d(A, D) + d(D, F)) = 2$$

$$D_G = \min(D_E + d(E, G), D_F + d(F, G)) = 5$$

$$D_H = \min(D_E + d(E, H), D_F + d(F, H)) = 4$$

$$D_B = \min(D_G + d(G, B), D_H + d(H, B)) = 5$$

Working backward, the “best” path from  $A$  to  $B$  is  $A, D, E, H, B$ .

Dynamic Programming	Known Environment	Backward Induction
Reinforcement Learning	Unknown Environment	?

# Solving MDPs — Value Iteration

## Theorem (Principle of Optimality)

A policy  $\pi$  achieves the optimal value from state  $s$ ,  $V^\pi(s) = V^*(s)$ , iff, for any state  $s'$  reachable from  $s$ ,  $\pi$  achieves the optimal value from state  $s'$ ,  $V^\pi(s') = V^*(s')$ .

Any optimal policy  $\pi^*$  can be subdivided into two components:

1. an optimal first action  $a^*$ ,
2. followed by an optimal policy from successor state  $s'$ .

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) V^*(s') \right\}$$

$$\pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) V^*(s') \right\}$$

Value Iteration:  $V_{k+1} \leftarrow T^*V_k$

$$V_1 \rightarrow V_2 \rightarrow \dots \rightarrow V^*$$

$$\|V_{k+1} - V_k\|_\infty < \varepsilon \implies \|V_{k+1} - V^*\|_\infty < \frac{2\gamma\varepsilon}{1-\gamma}$$

# Solving MDPs — Policy Iteration

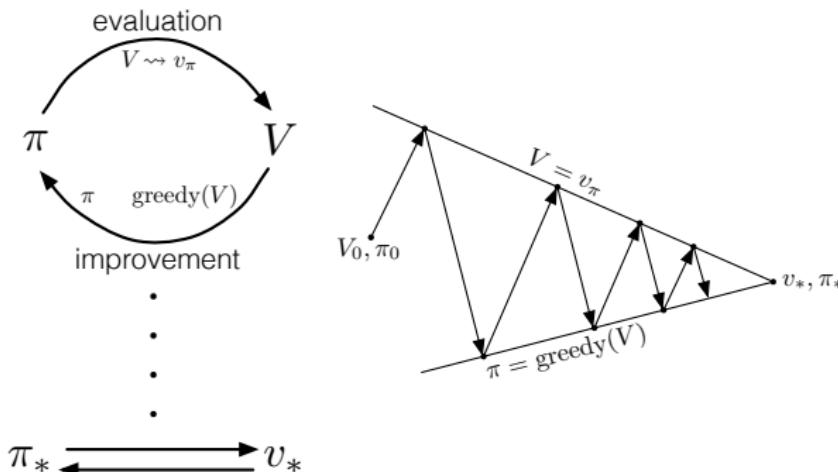
Policy Iteration:  $\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \dots \xrightarrow{I} \pi^* \xrightarrow{E} V^*$

- E — policy evaluation:

$$V_{k+1}^\pi \leftarrow T^\pi V_k^\pi$$

- I — policy improvement:

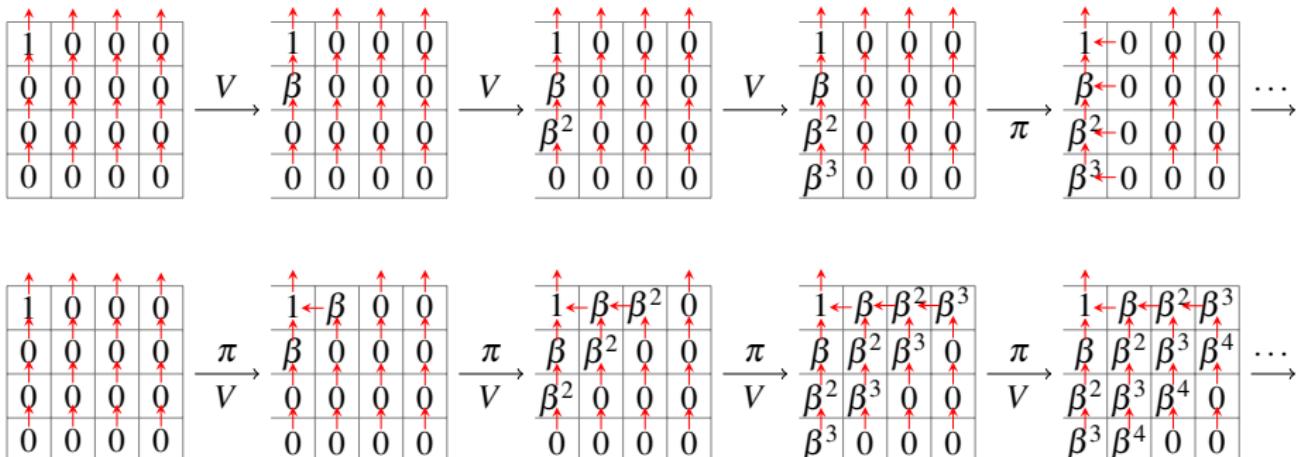
$$\pi_{k+1}(s) := \operatorname{argmax}_{a \in \mathcal{A}} \left( r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) V^{\pi_k}(s') \right)$$



## Example — Value Iteration vs Policy Iteration

- Value Iteration:  $((V \text{ improve})^*; \pi \text{ extraction})$
- Policy Iteration:  $(V \text{ improve}; \pi \text{ improve})^*$

$$V(s) = \max_{a \in A} r(s, a) + \beta V(f(s, a))$$



## Policy Iteration: the Guarantees

- ▶ Set  $\pi_0$  arbitrarily
- ▶ Repeat:
  - ▶ evaluate  $Q^{\pi_i}(s, a)$
  - ▶ let  $\pi_{i+1} = \underset{a}{\operatorname{argmax}} Q^{\pi_i}(s, a)$
  - ▶ set  $i = i + 1$
- ▶ Until  $\pi_i(s) = \pi_{i-1}(s)$

### Theorem

*The policy iteration algorithm generates a sequences of policies with non-decreasing performance*

$$V^{\pi_{k+1}} \geq V^{\pi_k}$$

*and it converges to  $\pi^*$  in a finite number of iterations.*

## Comparision — Value Iteration vs Policy Iteration

- ▶ Both value iteration and policy iteration compute the optimal values.
  - ▶ They are all variations of Bellman updates.
  - ▶ They differ only in whether we plug in a fixed policy or max over actions.
1. In value iteration, we don't track the policy, but taking the max over actions implicitly recomputes it.
    - Value Iteration computes the optimal policy even at a stage when the value function estimate has not yet converged.
    - If one action is better than all others, then the exact values of the states involved need not to be known.
    - Each iteration is computationally efficient.
  2. Policy Iteration converge in a finit number of iterations (often small in practice).
    - But each iteration requires a full policy evaluation and it might be expensive.

## Summary: MDP

- ▶ Compute optimal values: use value iteration or policy iteration



- ▶ Compute values for a particular policy: use policy evaluation



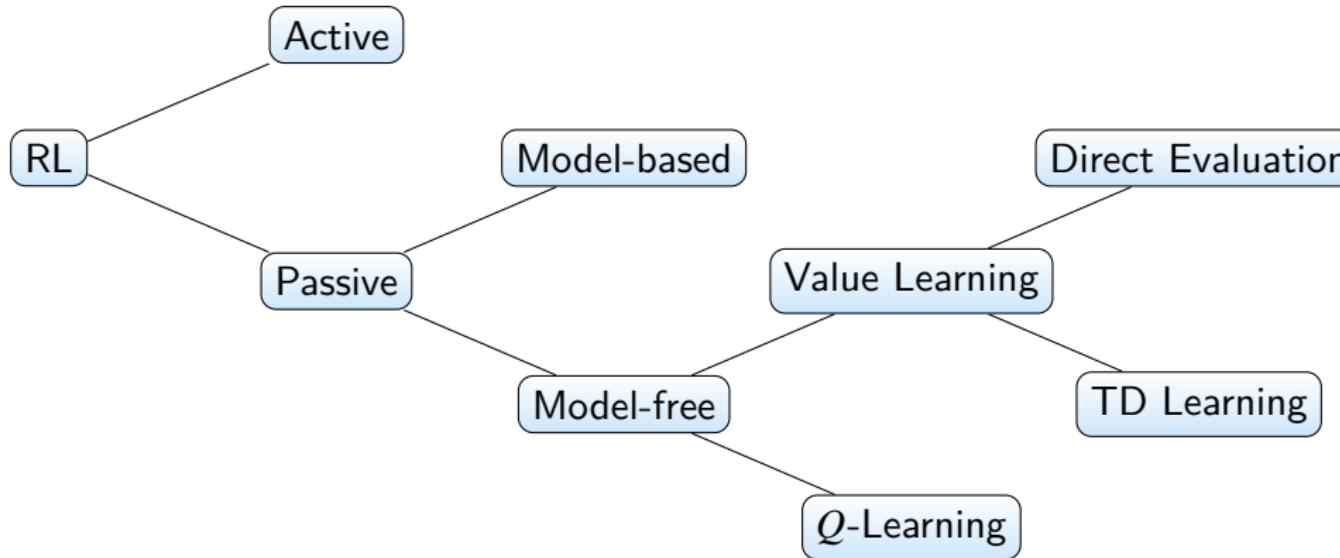
- ▶ Turn your values into a policy: use policy extraction (one-step lookahead)



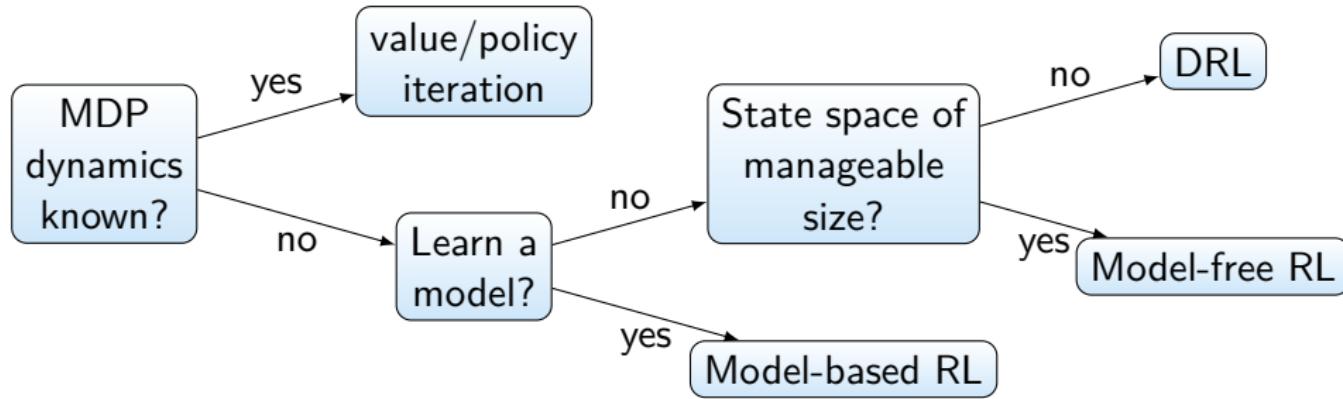
# RL — Overview

- ▶ **Passive Reinforcement Learning** (how to learn from experiences)
  - ▶ **Model-based** Passive RL
    - learn the MDP model from experiences, then solve the MDP with value / policy iteration
  - ▶ **Model-free** Passive RL
    - skip learning MDP model, directly learn  $V$  or  $Q$
    - receive a sample  $(s, a, s', r)$ , update estimates
- ▶ **Active Reinforcement Learning** (agent also needs to decide how to collect experiences)
  - ▶ How to efficiently explore and minimize regret?
  - ▶ How to trade off exploration/exploitation?
- ▶ **Approximate Reinforcement Learning** (to handle large state spaces)
  - ▶ Approximate  $Q$ -Learning
  - ▶ Policy Search

# RL — Overview



# RL — Overview



## Monte-Carlo Methods

MC learns from complete episodes of raw experience without modeling the environmental dynamics and computes the observed mean return as an approximation of the expected return.

$$G_t := \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$$

$$V(s) := \frac{\sum_{t=1}^T \llbracket S_t = s \rrbracket G_t}{\sum_{t=1}^T \llbracket S_t = s \rrbracket}$$

$$Q(s, a) := \frac{\sum_{t=1}^T \llbracket S_t = s, A_t = a \rrbracket G_t}{\sum_{t=1}^T \llbracket S_t = s, A_t = a \rrbracket}$$

## Temporal Differences

- ▶ Suppose we have a sequence of values:  $v_1, v_2, \dots$
- ▶ We want a running estimate of the average of the first  $k$  values:

$$A_k := \frac{v_1 + \cdots + v_k}{k}$$

- ▶ When a new value  $v_k$  arrives:

$$\begin{aligned} A_k &= \frac{v_1 + \cdots + v_{k-1} + v_k}{k} \\ &= \frac{k-1}{k} A_{k-1} + \frac{1}{k} v_k \\ &= (1 - \alpha_k) A_{k-1} + \alpha_k v_k \\ &= A_{k-1} + \alpha_k (v_k - A_{k-1}) \end{aligned}$$

where  $\alpha_k := \frac{1}{k}$ .

- ▶ We can guarantee convergence if

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty$$

# Temporal-Difference Learning

TD Learning is model-free and learns from incomplete episodes of experience.

$$V(s_t) \leftarrow (1 - \alpha)V(s_t) + \alpha G_t$$

$$V(s_t) \leftarrow V(s_t) + \alpha(G_t - V(s_t))$$

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

- ▶ MC updates value  $V(s_t)$  toward actual return  $G_t$ .

$$V(s_t) \leftarrow V(s_t) + \alpha(G_t - V(s_t))$$

- ▶ TD updates value  $V(s_t)$  toward estimated return  $r_{t+1} + \gamma V(s_{t+1})$ .

$$V(s_t) \leftarrow V(s_t) + \underbrace{\alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))}_{\text{TD error}}$$

TD target

# SARSA: On-Policy TD control

## SARSA

1. At time step  $t$ , start from state  $s_t$  and choose action  $a_t$  using policy derived from  $Q$  (e.g.,  $\varepsilon$ -greedy)
2. Take action  $a_t$ , observe  $r_{t+1}$  and  $s_{t+1}$
3. Choose  $a_{t+1}$  from  $s_{t+1}$  using policy derived from  $Q$  (e.g.,  $\varepsilon$ -greedy)
4. Update the action-value function:  
$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$
5.  $t = t + 1$  and repeat from step 1

# Q-Learning: Off-policy TD control

## Q-Learning

1. At time step  $t$ , start from state  $s_t$  and choose action  $a_t$  using policy derived from  $Q$  (e.g.,  $\varepsilon$ -greedy)

2. Take action  $a_t$ , observe  $r_{t+1}$  and  $s_{t+1}$

3. Update the action-value function:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left( r_{t+1} + \gamma \max_{a \in \mathcal{A}} Q(s_{t+1}, a) - Q(s_t, a_t) \right)$$

4.  $t = t + 1$  and repeat from step 1

## Expected SARSA

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left( r_{t+1} + \gamma \sum_{a \in \mathcal{A}} \pi(a | s_{t+1}) Q(s_{t+1}, a) - Q(s_t, a_t) \right)$$

## Remark: Q-Learning

$$Q^*(s, a) = r(s, a) + \gamma \sum_{s' \in S} P(s' | s, a) \max_{a' \in \mathcal{A}} Q^*(s', a')$$

We have three problems.

1. We don't know the reward function.
2. We don't know the transition probabilities.
3. We don't know the utility of the state we reached.

$$\hat{Q}_t(s_t, a_t) = r_{t+1} + \gamma \max_{a \in \mathcal{A}} Q_t(s_{t+1}, a)$$

1. Instead of  $r(s, a)$ , use  $r_{t+1}$ , the reward we got this time.
2. Instead of summing over  $P(s' | s, a)$ , just set  $s' = s_{t+1}$ , i.e., whatever state followed  $s_t$ .
3. Instead of the true value of  $Q(s, a)$ , use our current estimate,  $Q_t(s, a)$ .

$$Q_{t+1}(s_t, a_t) = Q_t(s, a) + \alpha \left( \hat{Q}_t(s, a) - Q_t(s, a) \right)$$

## On-policy vs Off-policy

- ▶ The learned policy may be different from the behavior policy
- ▶ On-policy learning (e.g., SARSA)
  - ▶ learn from the moves that were actually taken
  - ▶ policy evaluation and policy improvement happen simultaneously
- ▶ Off-policy learning (e.g., Q-learning)
  - ▶ learning of the optimal policy independently from its execution
  - ▶ converge faster, more aggressive/risky
  - ▶ possibility to use effective exploration strategies

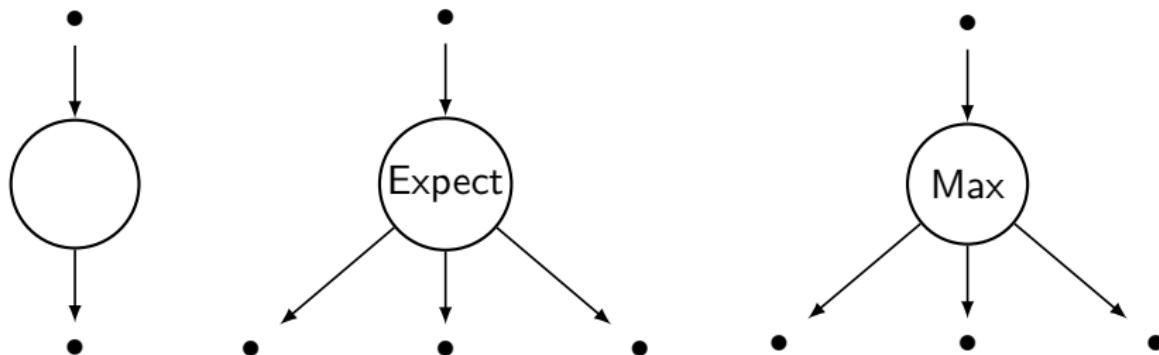
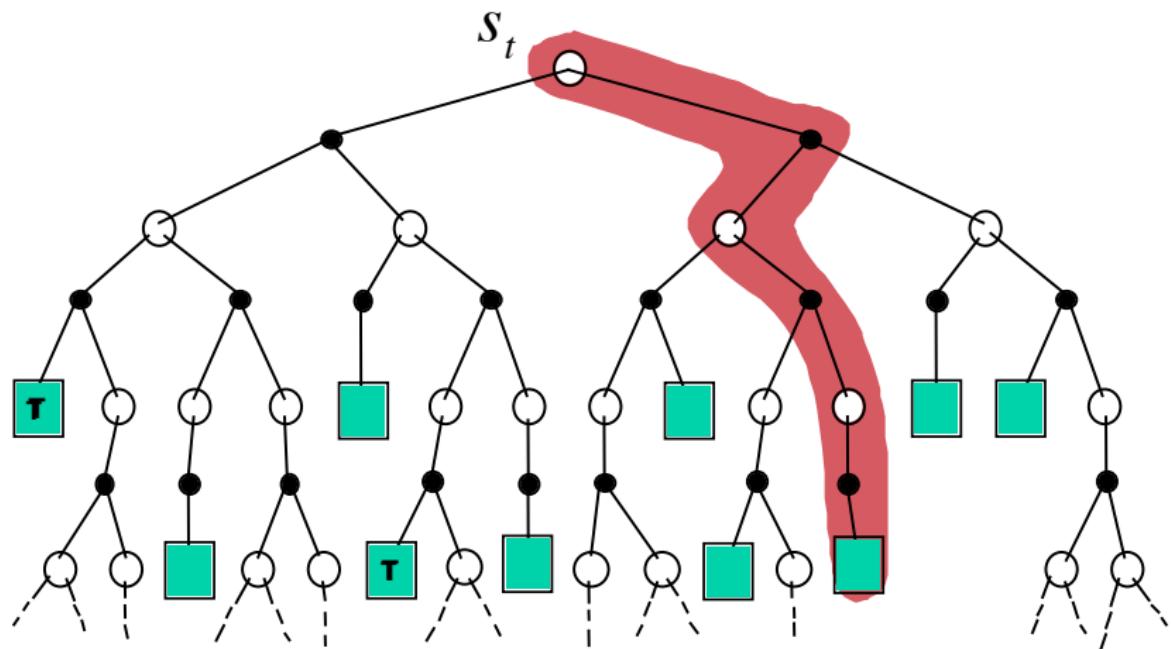


Figure: SARSA, Expected SARSA, and Q-Learning

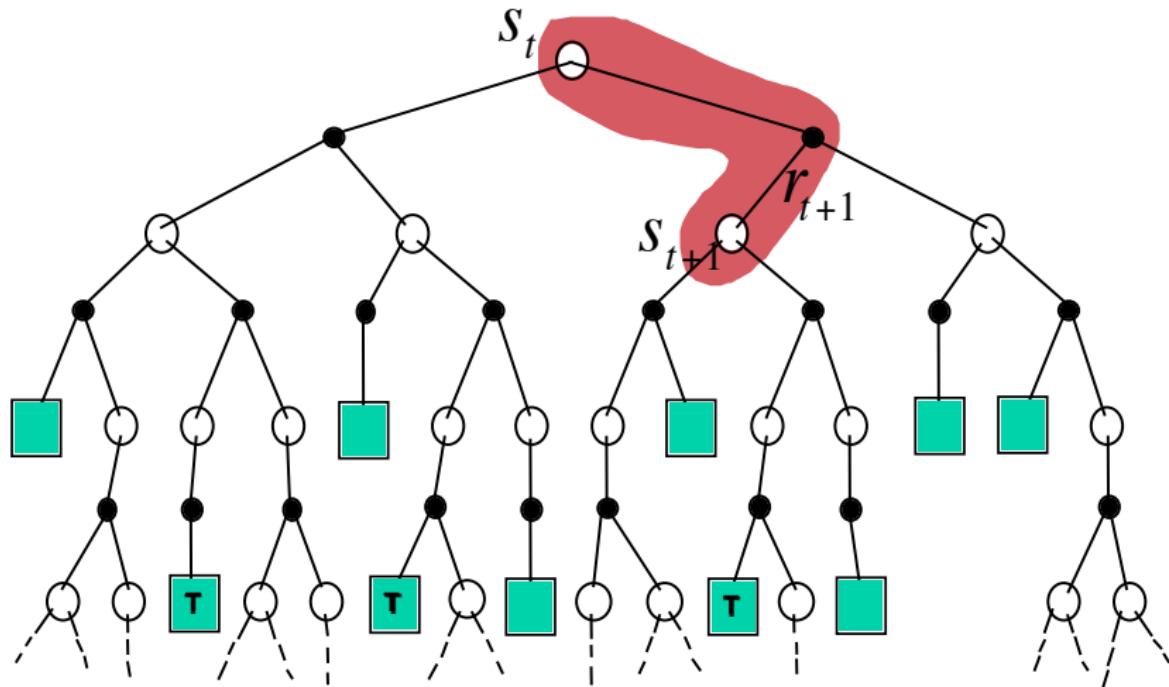
# Monte-Carlo Backup

$$V(s_t) \leftarrow V(s_t) + \alpha(G_t - V(s_t))$$



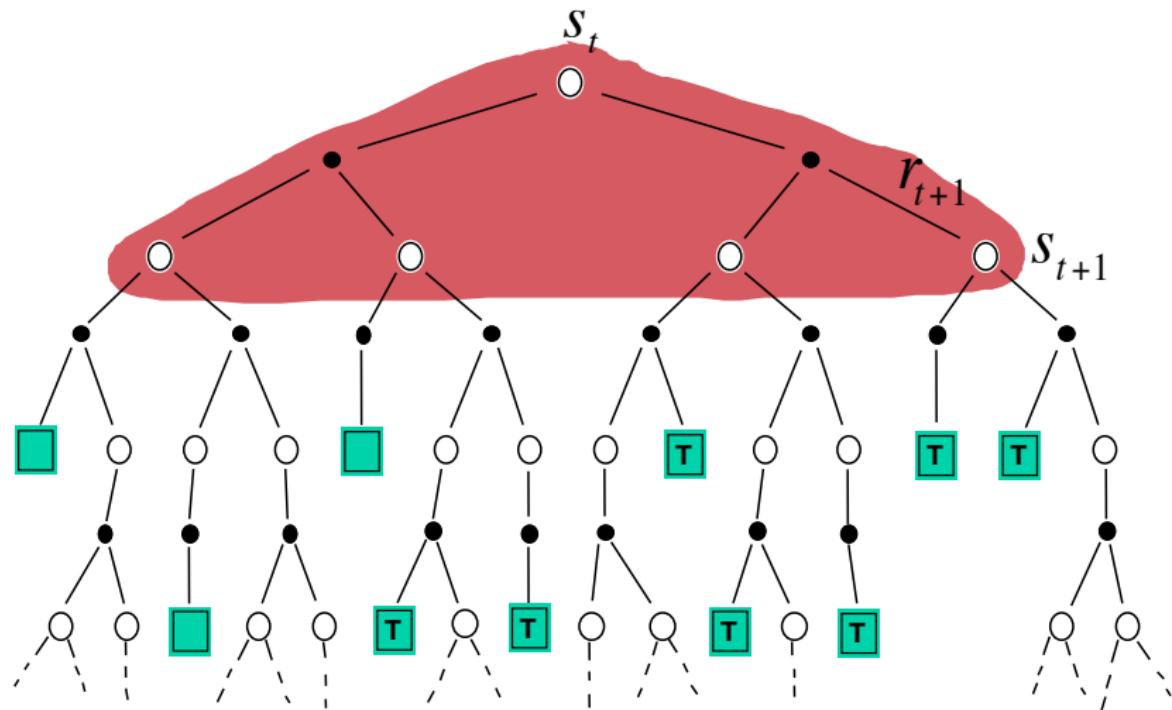
# Temporal-Difference Backup

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

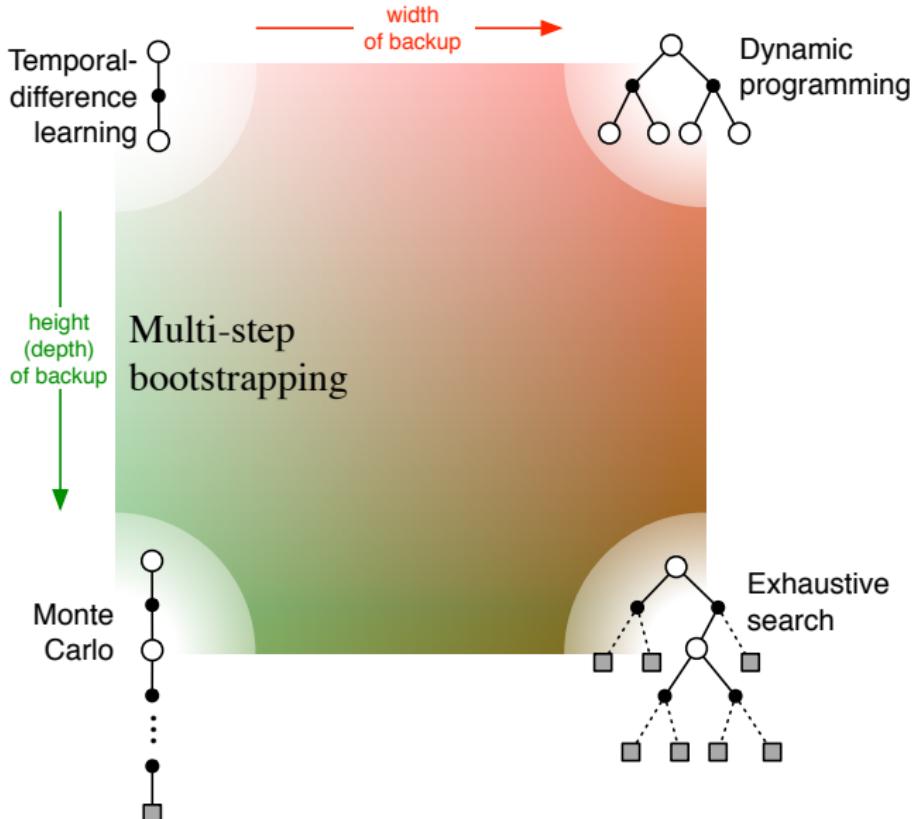


# Dynamic Programming Backup

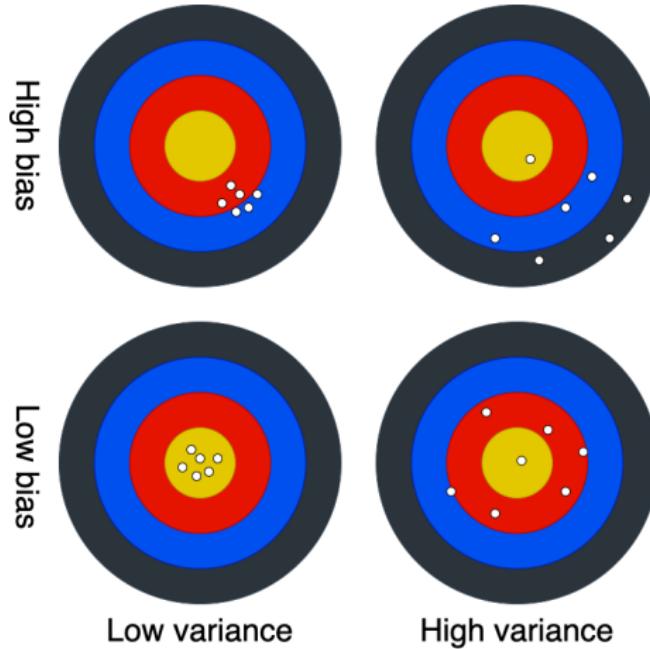
$$V(s_t) \leftarrow \mathbb{E}_\pi [r_{t+1} + \gamma V(s_{t+1})]$$



# Unified View



# Bias-Variance Trade-Off



$$\text{expected squared error} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

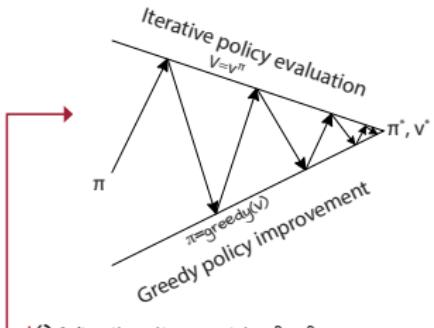
$$\mathbb{E} \left[ (y - \hat{f}(x))^2 \right] = \left( \mathbb{E}[\hat{f}(x)] - f(x) \right)^2 + \mathbb{E} \left[ \left( \hat{f}(x) - \mathbb{E}[\hat{f}(x)] \right)^2 \right] + \mathbb{E} \left[ (y - f(x))^2 \right]$$

## Bias-Variance Trade-Off: MC vs TD

- ▶ Return  $G_t := \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$  is an unbiased estimate of  $V^\pi(s_t)$ .
- ▶ TD target  $r_{t+1} + \gamma V(s_{t+1})$  is a biased estimate of  $V^\pi(s_t)$  unless  $V(s_{t+1}) = V^\pi(s_{t+1})$ .
- ▶ But the TD target is much lower variance:
  - ▶ Return depends on **many** random actions, transitions, rewards
  - ▶ TD target depends on **one** random action, transition, reward
- ▶ MC has high variance, zero bias
  - ▶ Good convergence properties
  - ▶ Works well with function approximation
  - ▶ Not very sensitive to initial value
  - ▶ Very simple to understand and use
- ▶ TD has low variance, some bias
  - ▶ Usually more efficient than MC
  - ▶ TD(0) converges to  $V^\pi(s)$
  - ▶ Problem with function approximation
  - ▶ More sensitive to initial values

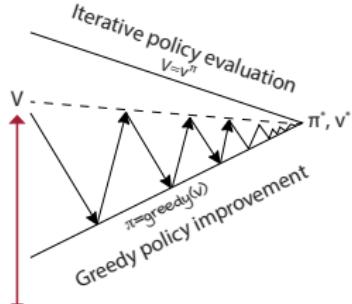
# Comparison between planning and control methods

## Policy iteration



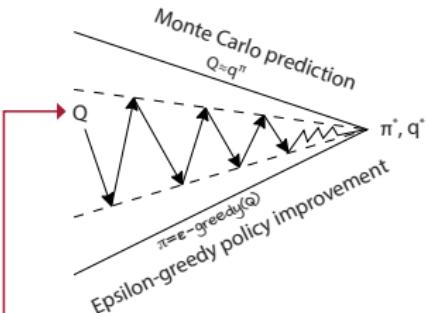
(1) Policy iteration consists of a f convergence of iterative policy evaluation alternating with greedy policy improvement.

## Value iteration



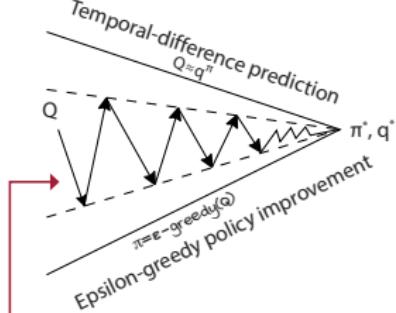
(2) Value iteration starts with an arbitrary value function and has a truncated policy evaluation step.

## Monte Carlo control



(3) MC control estimates a Q-function, has a truncated MC prediction phase followed by an epsilon-greedy policy-improvement step.

## SARSA



(4) SARSA has pretty much the same as MC control except a truncated TD prediction for policy evaluation.

# Policy Search

- ▶ Idea: directly optimize policy
- ▶ Policy may be parameterized  $Q$  functions, hence:

$$\pi(s) := \operatorname{argmax}_a \hat{Q}_\theta(s, a)$$

- ▶ Stochastic policy, e.g. given by softmax function

$$\pi_\theta(a \mid s) := \frac{e^{\hat{Q}_\theta(s, a)}}{\sum_a e^{\hat{Q}_\theta(s, a)}}$$

- ▶ Policy value  $\rho(\theta)$ : expected reward if  $\pi_\theta$  is carried out

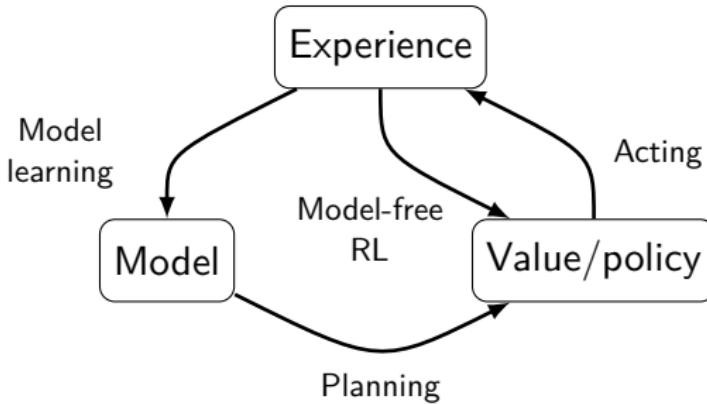
## Remark (Free Will<sup>a</sup>)

---

<sup>a</sup>Erik M. Rehn: Free Will Belief as a consequence of Model-based Reinforcement Learning. 2022.

- ▶ the “will” of an agent:  $\hat{Q}_\theta(s, a)$
- ▶ the “freedom” of an agent:  $H(s) = - \sum_a \pi_\theta(a \mid s) \log \pi_\theta(a \mid s)$

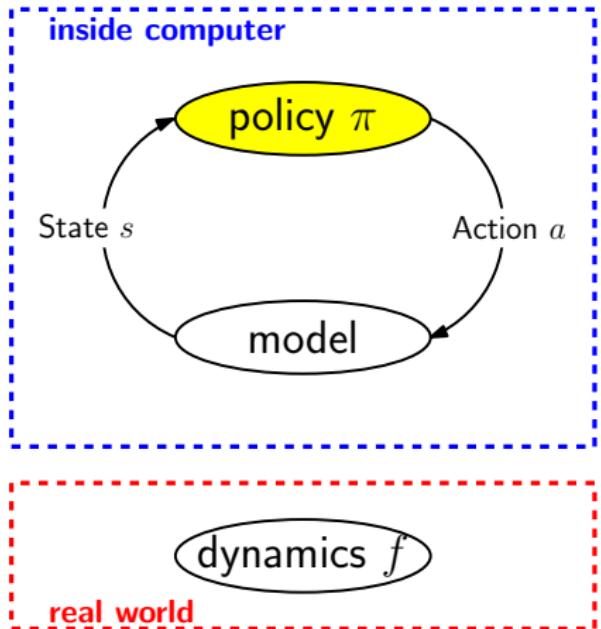
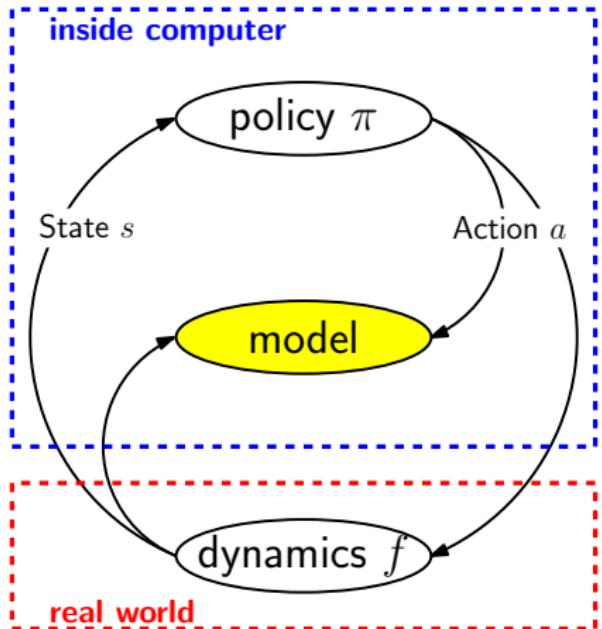
# Types of RL Algorithms



**Figure:** The direct approach uses a representation of either a value function or a policy. The indirect approach makes use of a model of the environment.

- ▶ Policy gradient: directly differentiate the objective  
$$\theta^* = \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{(s,a) \sim p_{\theta}(s,a)} [r(s,a)]$$
- ▶ Value-based: estimate value or  $Q$ -function of the optimal policy
- ▶ Actor-critic: estimate value or  $Q$ -function of the current policy
- ▶ Model-based RL: estimate the transition model

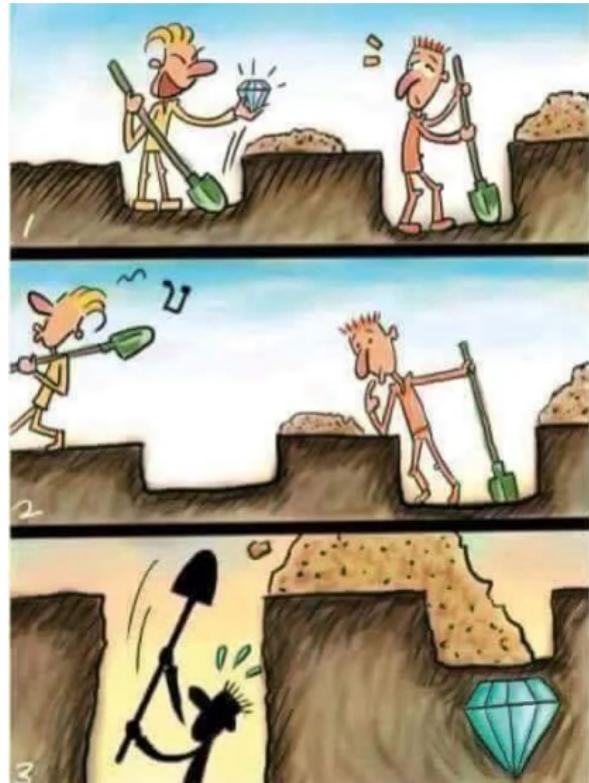
# Model-based Set-up: Interaction and Simulation



- ▶ Interaction: internal model is refined using experience from interacting with the real system
- ▶ Simulation: internal model simulates consequences of actions in the real system, policy is refined (support counterfactual)

# Exploration vs Exploitation

- ▶ Exploration: trying actions just to see what happens in the hope of learning more successful behaviors
- ▶ Exploitation: using what the agent has learned so far to select actions
- ▶ In practice, agents must do some exploration otherwise they may be stuck in a subset of environment states having low utility
- ▶ It even makes sense in some applications to choose actions randomly
- ▶ Typically, agent explore more in the early stages of deployment and exploit more in later stages



# How to Explore?

- ▶ Regular Q-Update:

$$Q(s, a) \leftarrow r(s, a) + \gamma \max_{a'} Q(s', a')$$

- ▶ Modified Q-Update with exploration function  $f$ :

$$Q(s, a) \leftarrow r(s, a) + \gamma \max_{a'} f(Q(s', a'), N(s', a'))$$

where  $N$  is the visit count.

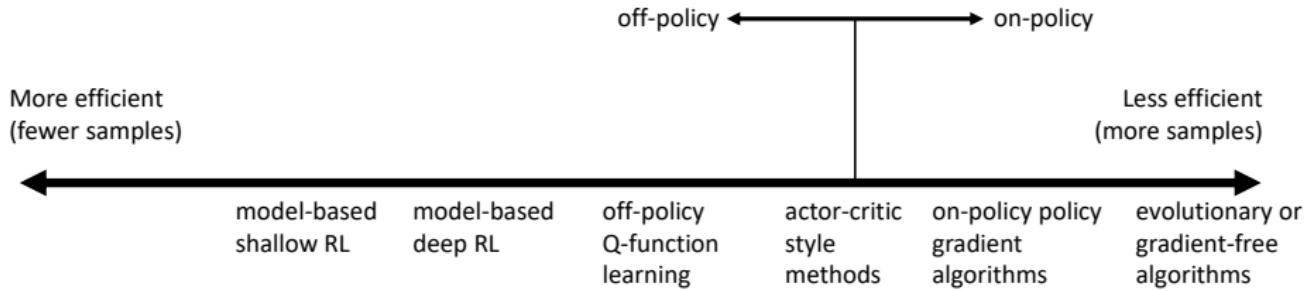
## Possible Strategy

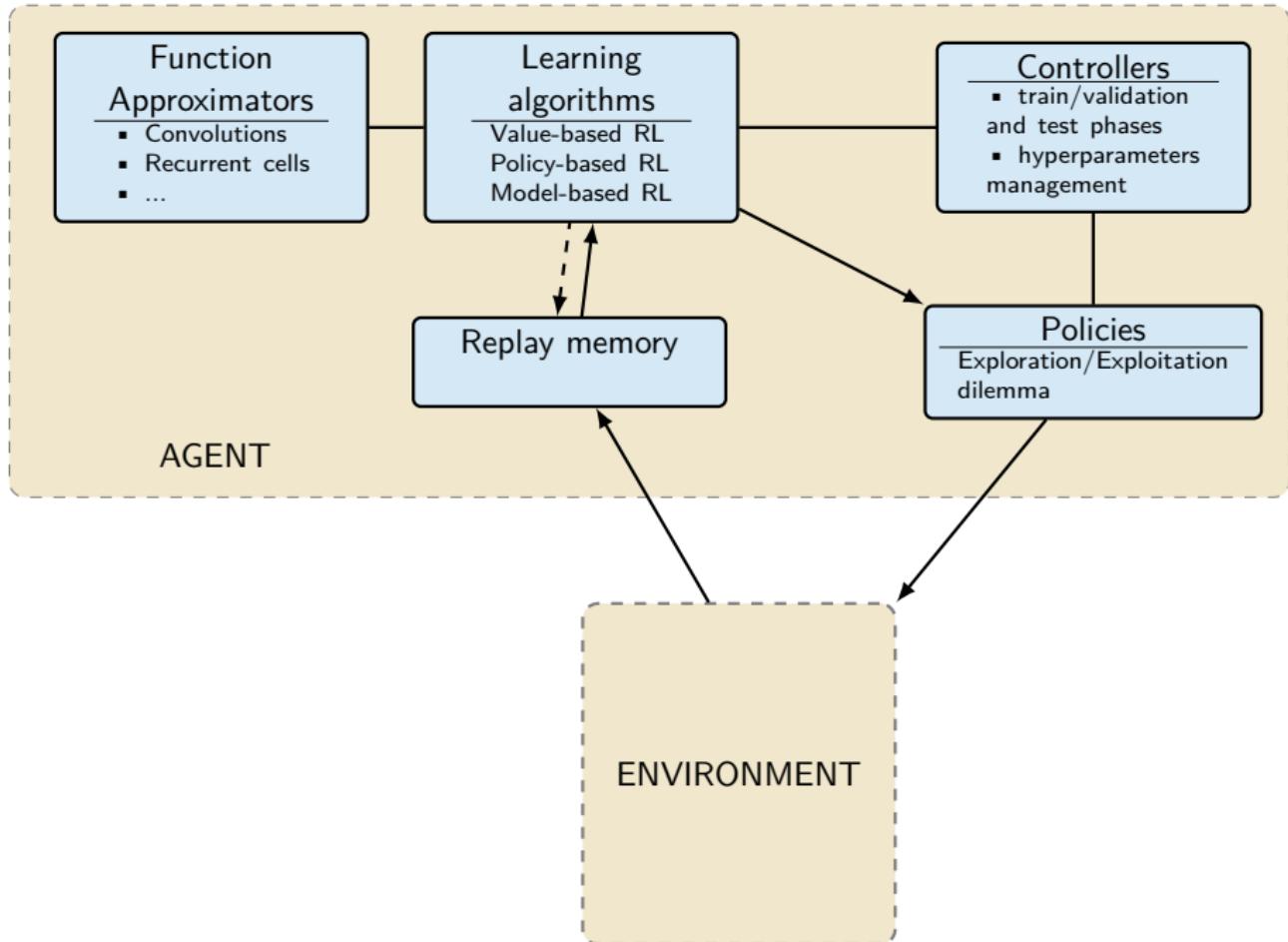
- ▶ Epsilon-Greedy
- ▶ Softmax
- ▶ Upper Confidence Bound (UCB)
- ▶ Thompson Sampling

# Why so many RL algorithms?

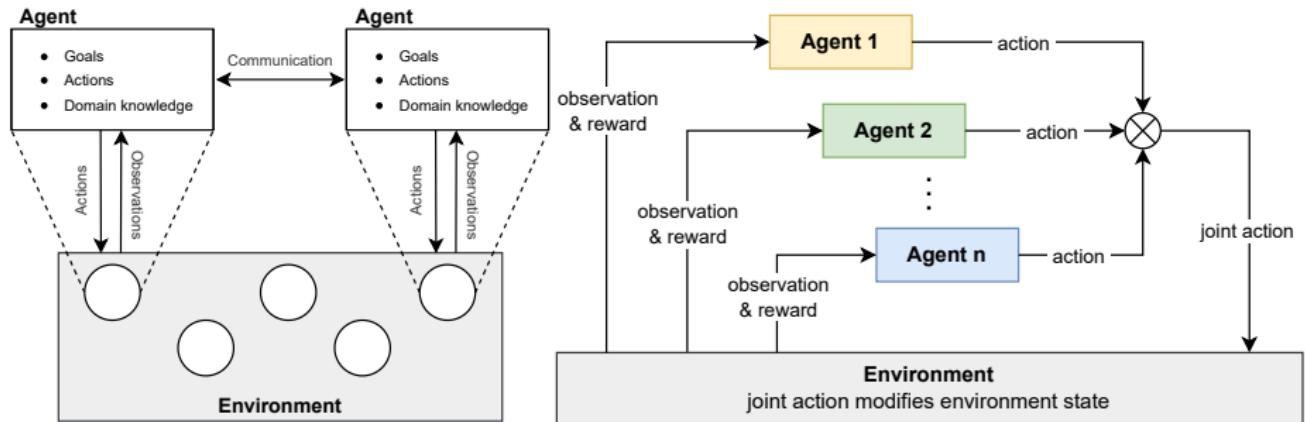
- ▶ Different tradeoffs
  - ▶ Sample efficiency
  - ▶ Stability & ease of use
- ▶ Different assumptions
  - ▶ Stochastic or deterministic?
  - ▶ Continuous or discrete?
  - ▶ Episodic or infinite horizon?
- ▶ Different things are easy or hard in different settings
  - ▶ Easier to represent the policy?
  - ▶ Easier to represent the model?

# Sample Efficiency





# Multi-Agent Reinforcement Learning



# Multi-Agent Reinforcement Learning

$$s_{t+1} \sim P(\cdot | s_t, a_t^1, \dots, a_t^N)$$

$$r_{t+1}^i = R^i(s_t, a_t^1, \dots, a_t^N, s_{t+1})$$

- ▶ Agent  $i$ 's policy  $\pi^i$  specifies  $\pi^i(a^i | s)$ .
- ▶ Agent  $i$ 's goal:  $\max_{\pi^i} V_{\pi^i, \pi^{-i}}$  where

$$V_{\pi^i, \pi^{-i}}^i = \mathbb{E} \left[ \sum_{t \geq 0} R^i(s_t, a_t^1, \dots, a_t^N, s_{t+1}) \mid \forall j : a_t^j \sim \pi^j(\cdot | s_t), s_{t+1} \sim P(\cdot | s_t, a_t^1, \dots, a_t^N) \right]$$

- ▶ **Nash Equilibrium:** Collection of policies  $\pi^1, \dots, \pi^N$  such that

$$\forall i \leq N \forall \pi' : V_{\pi^i, \pi^{-i}}^i \geq V_{\pi', \pi^{-i}}^i$$

- ▶ When  $N = 1$ , this is equivalent to optimal policy of MDP.

# 强化学习 vs 主动推理

# 自由能原理与主动推理

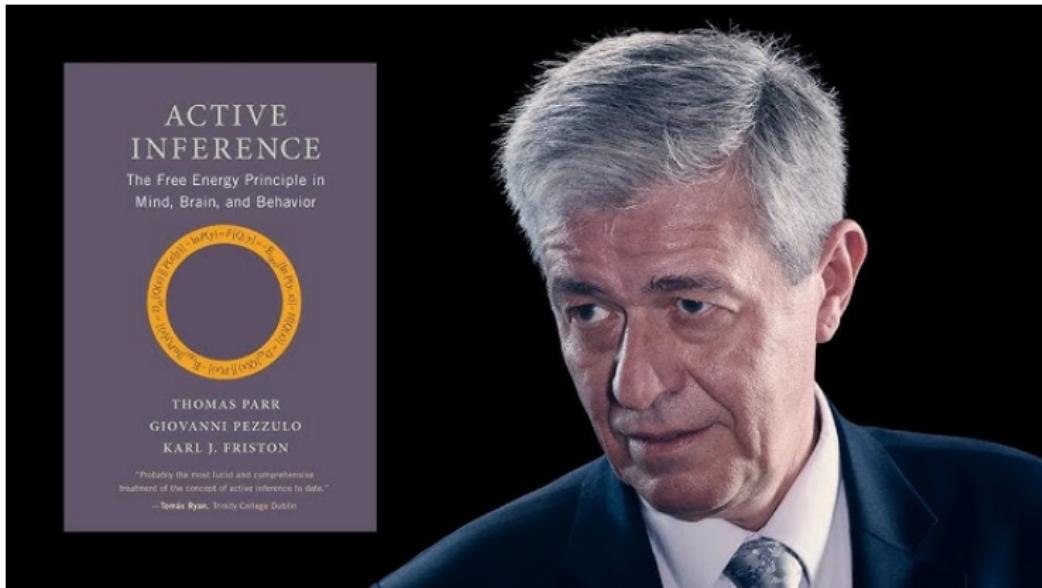
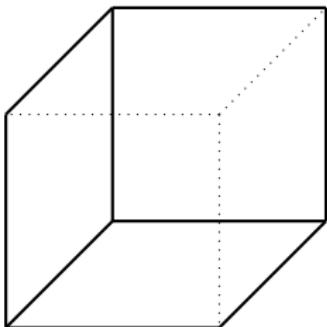
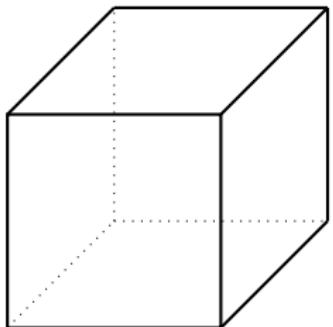


Figure: Karl Friston

# 感知与行动的统一

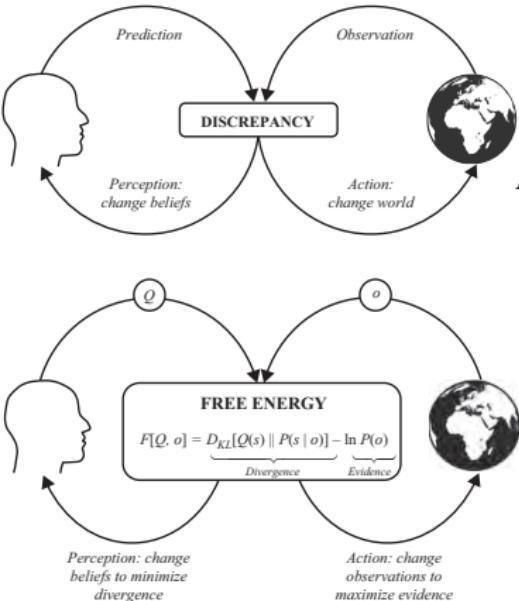
Friston 主动推理

- ▶ 感知依赖于大脑对外部世界的主动预测 (贝叶斯大脑)
  - ▶ 通过比较预测和输入信息, 大脑更新预测模型, 最小化预测误差
  - ▶ 感知通过更新信念让信念与观察相符; 行动通过改变环境让环境符合信念和目标.
1. 感知、学习可以看作最小化 “变分自由能” 的过程
  2. 行动、规划、决策可以看作最小化 “期望自由能” 的过程



# Friston's Active Inference

## Minimizing Variational Free Energy



$$Q^* = \underset{Q \in \mathcal{M}}{\operatorname{argmin}} F[Q, o]$$

$$\approx P(\cdot | o)$$

$$F[Q, o] = -\mathbb{E}_{Q(s)} \left[ \log \frac{P(o, s)}{Q(s)} \right]$$

$$= \underbrace{-\mathbb{E}_{Q(s)} [\log P(o, s)]}_{\text{Energy}} - \underbrace{H[Q(s)]}_{\text{Entropy}}$$

$$= \underbrace{D_{KL}[Q(s) \| P(s)]}_{\text{Complexity}} - \underbrace{\mathbb{E}_{Q(s)} [\log P(o | s)]}_{\text{Accuracy}}$$

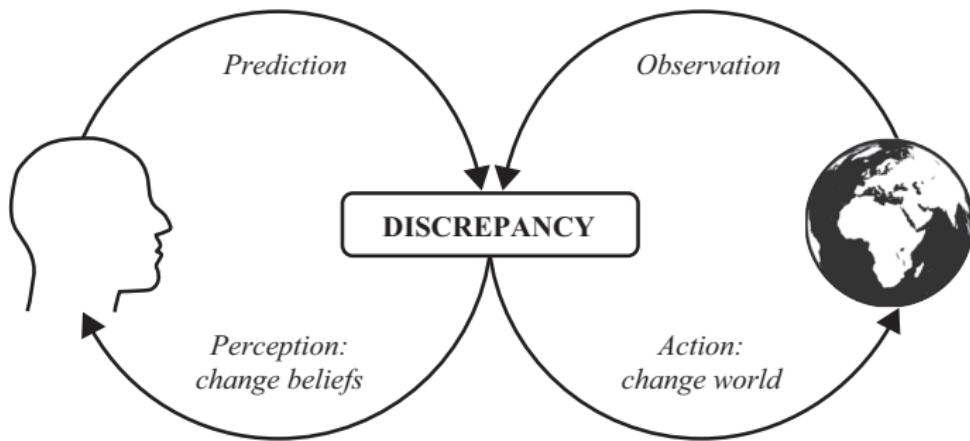
$$= \underbrace{D_{KL}[Q(s) \| P(s | o)]}_{\text{Divergence}} - \underbrace{-\log P(o)}_{\text{Evidence}}$$

$$\begin{aligned}
-\log P(o) &= -\log \sum_s Q(s) \frac{P(o, s)}{Q(s)} \stackrel{\text{Jensen}}{\leq} -\mathbb{E}_{Q(s)} \left[ \log \frac{P(o, s)}{Q(s)} \right] \\
&= \underbrace{-\mathbb{E}_{Q(s)} [\log P(o, s)]}_{\text{Energy}} - \underbrace{H[Q(s)]}_{\text{Entropy}} \\
&= \underbrace{D_{\text{KL}}[Q(s) \| P(s)]}_{\text{Complexity}} - \underbrace{\mathbb{E}_{Q(s)} [\log P(o | s)]}_{\text{Accuracy}} \\
&= \underbrace{D_{\text{KL}}[Q(s) \| P(s | o)]}_{\text{Divergence}} - \underbrace{\log P(o)}_{\text{Evidence}}
\end{aligned}$$

1. 维持当前的“能量”与尽可能高的“熵”。当缺乏观察信息和精确的先验信念时，关于环境的隐藏状态，应该采纳不确定性最高的信念（最大熵原理）。
2. 最小化“复杂性”（贝叶斯惊讶）与最大化“精确性”。寻找对观察信息的最简洁的解释。
3. 感知：改变信念以最小化散度；行动：改变观察以最大化证据。

# Friston's Active Inference

Minimizing Expected Free Energy



$$\pi^* = \operatorname{argmin}_{\pi} G(\pi)$$

$$G(\pi) = -\mathbb{E}_{Q(o,s|\pi)} \left[ \log \frac{P(o, s | \pi)}{Q(s | \pi)} \right]$$

其中,  $Q(o, s | \pi) = Q(s | \pi)P(o | s)$ .

$$\begin{aligned}
G(\pi) &= - \sum_{o,s} P(o \mid s) Q(s \mid \pi) \left[ \log \frac{P(o, s \mid \pi)}{Q(s \mid \pi)} \right] \\
&= - \sum_{o,s} P(o \mid s) Q(s \mid \pi) \left[ \log \frac{P(s \mid o, \pi) P(o \mid \pi)}{Q(s \mid \pi)} \right] \\
&= - \sum_{o,s} P(o \mid s) Q(s \mid \pi) \left[ \log \frac{P(s \mid o, \pi)}{Q(s \mid \pi)} \right] - \sum_{o,s} P(o \mid s) Q(s \mid \pi) \log P(o \mid \pi) \\
&= - \underbrace{\mathbb{E}_{Q(o \mid \pi)} \left[ D_{\text{KL}}[Q(s \mid o, \pi) \parallel Q(s \mid \pi)] \right]}_{\text{Information Gain}} - \underbrace{\mathbb{E}_{Q(o \mid \pi)} \left[ \log P(o \mid \pi) \right]}_{\text{Pragmatic Value}} \\
&= - \sum_{o,s} P(o \mid s) Q(s \mid \pi) \left[ \log \frac{P(o \mid s, \pi)}{Q(o \mid \pi)} \right] - \sum_{o,s} P(o \mid s) Q(s \mid \pi) \log P(o \mid \pi) \\
&= - \sum_s Q(s \mid \pi) \sum_o P(o \mid s) \log P(o \mid s) - \sum_{o,s} P(o \mid s) Q(s \mid \pi) \left[ \log \frac{P(o \mid \pi)}{Q(o \mid \pi)} \right] \\
&= \underbrace{\mathbb{E}_{Q(s \mid \pi)} [H[P(o \mid s)]]}_{\text{Expected Ambiguity}} + \underbrace{D_{\text{KL}}[Q(o \mid \pi) \parallel P(o \mid \pi)]}_{\text{Risk}} \\
&\leq \mathbb{E}_{Q(s \mid \pi)} [H[P(o \mid s)]] + D_{\text{KL}}[Q(s \mid \pi) \parallel P(s \mid \pi)] \\
&= - \underbrace{\mathbb{E}_{Q(o,s \mid \pi)} [\log P(o, s \mid \pi)]}_{\text{Expected Energy}} - \underbrace{H[Q(s \mid \pi)]}_{\text{Entropy}}
\end{aligned}$$

$$\begin{aligned}
G(\pi) &= \underbrace{-\mathbb{E}_{Q(o|\pi)} \left[ D_{\text{KL}} [Q(s|o, \pi) \| Q(s|\pi)] \right]}_{\text{Information Gain}} - \underbrace{\mathbb{E}_{Q(o|\pi)} [\log P(o|\pi)]}_{\text{Pragmatic Value}} \\
&= \underbrace{\mathbb{E}_{Q(s|\pi)} [H[P(o|s)]]}_{\text{Expected Ambiguity}} + \underbrace{D_{\text{KL}} [Q(o|\pi) \| P(o|\pi)]}_{\text{Risk}} \\
&\leq \underbrace{-\mathbb{E}_{Q(o,s|\pi)} [\log P(o,s|\pi)]}_{\text{Expected Energy}} - \underbrace{H[Q(s|\pi)]}_{\text{Entropy}}
\end{aligned}$$

- “信息增益”可以看做一种认知价值，可以帮助消除不确定性；“实用价值”虽然是对观察的先验信念，但可以包含玩家的先验偏好。最大化“信息增益”与“实用价值”可以平衡“探索”与“利用”的两难。
- 最小化“预期含混”可以消除“状态”到“结果”的不准确性。最小化“风险”要求行动策略的结果的分布要与玩家的先验偏好相符。
- 最大熵。

# 具身模拟

- ▶ 现象学家梅洛·庞蒂：“动作的沟通或理解是通过我的意向和他人的动作、我的动作和在他人行为中显现的意向的相关关系实现的。所发生的一切像是他人的意向寓于我的身体中，或我的意向寓于他人的身体中。”
- ▶ 镜像神经元：灵长类或鸟类动物在做一个动作（比如伸手拿香蕉），或观察到其它个体在做同一个动作时，都会被激活的一类神经元。
- ▶ 小鼠无论是作为打架的“当事者”还是“旁观者”，其镜像神经元都会被激活。反过来，如果激活小鼠的镜像神经元，也会使它们产生攻击性。
- ▶ 镜像神经元就像一面镜子，在自己的大脑中“模拟”了他人的动作。
- ▶ 有人猜测，镜像神经元可能在模仿学习、意图理解、共情等方面起着重要作用。
- ▶ Friston 认为，镜像神经元是大脑主动推理的结果。无论是“做动作”还是“观察动作”，被激活的是同一个“生成模型”。镜像神经元的“镜像”特性，源于这同一个生成模型既可以用来“做动作”（通过行动改变世界符合自己的预测），也可以用来理解“观察动作”（通过更新信念来解释观察）。

# Phenomenal World vs Noumenal World

1. Phenomenal World: things as they appear to us
2. Noumenal World: things in themselves
  - ▶ Kant argued that **space** and **time** and **causality** are part of our perceptual framework.
  - ▶ Kantian **hyperpriors** is how we structure our phenomenal world.
  - ▶ Kant's Copernican revolution: objects must conform to our cognition.
  - ▶ Helmholtz: Perception as inference.



**Figure:** Reality can be experienced, but it is not possible to totally express it with language. The experience of the world is a construction, constrained by external data and internal beliefs, priors and assumptions.

# Contents

Introduction	Game Theory
Philosophy of Induction	Reinforcement Learning
Inductive Logic	Deep Learning
Universal Induction	Artificial General Intelligence
Causal Inference	What If Computers Could Think? References 1753

# 函数

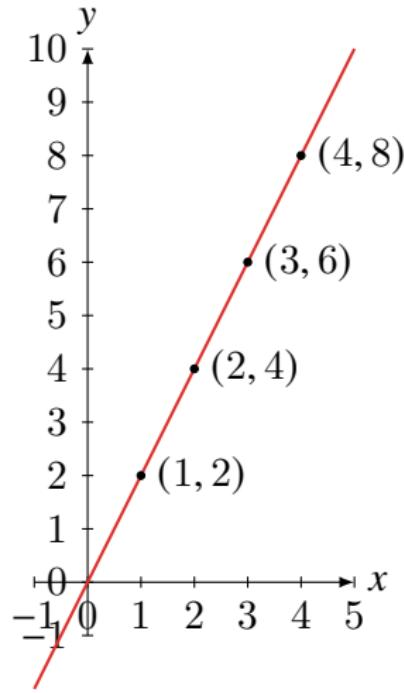


$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \xrightarrow{f} \begin{bmatrix} 2 \\ 4 \\ 6 \\ 8 \end{bmatrix}$$

$$y = 2x$$

$$y = wx + b$$

复杂的函数咋办？



# Learning as an Alternative to Traditional Programming

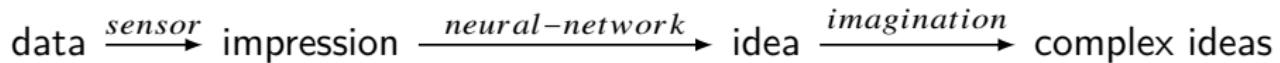
- ▶ Learning from data / experience may be more human-like
  - ▶ Babies develop an intuitive understanding of physics in their first 2 years
  - ▶ Formal reasoning and logic comes much later in development
- ▶ Learning enables fast reaction times
  - ▶ It might take a long time to train a neural network
  - ▶ But predicting with the network is very fast
- ▶ Representation Learning
  - ▶ Jointly learn **features** and **classifier**, **directly from raw data**
- ▶ Deep Learning
  - ▶ **multiple levels** of representation learning
  - ▶ composition of simple but nonlinear modules that each transform the representation at one level to a higher, more abstract level.

*"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."*

# 亚里士多德与连接主义

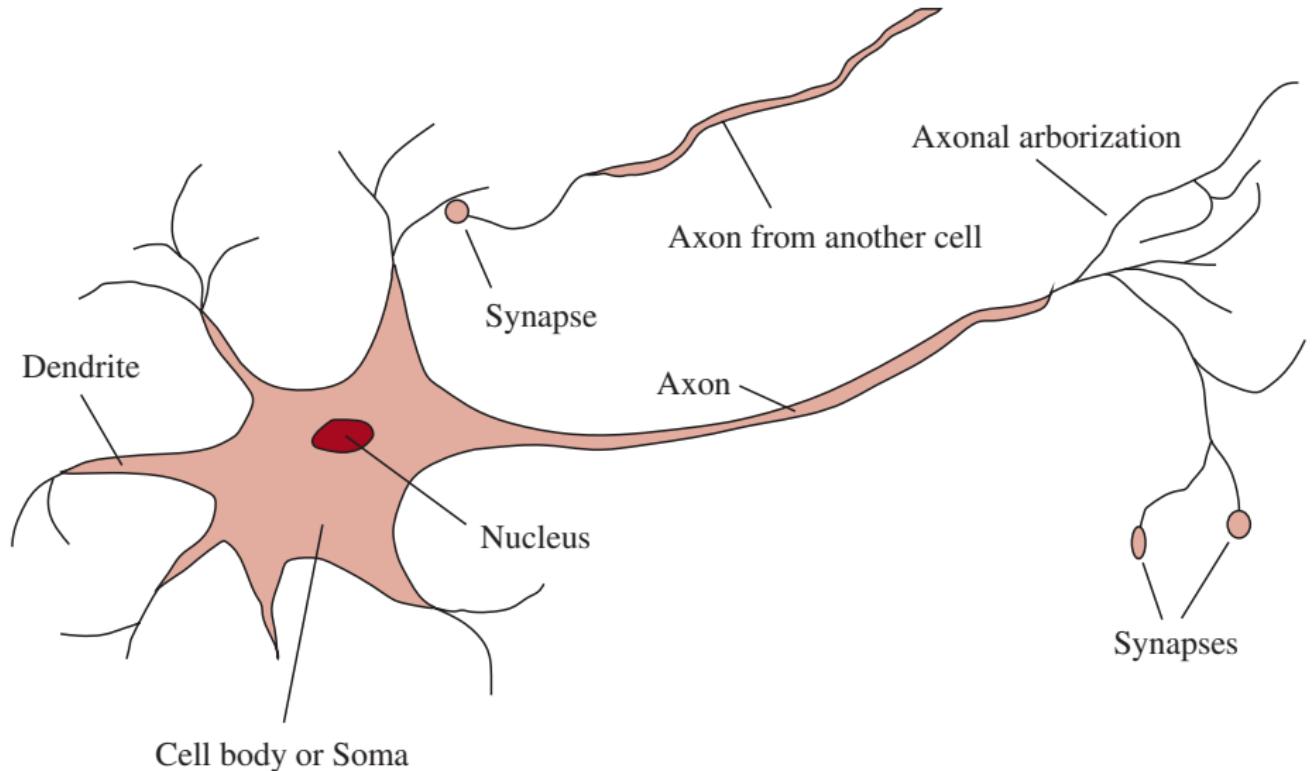
1. The law of contiguity. 时空上相邻的事物或事件会相互关联.
2. The law of frequency. 事物或事件关联次数越多, 关联强度越大.
3. The law of similarity. 相似的事物, 一个会激发思考另一个.
4. The law of contrast. 一个事物可能会激发思考相反的事物.

# 休谟与连接主义



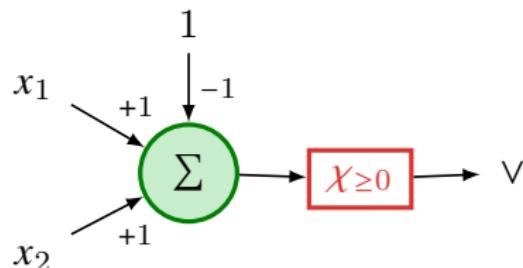
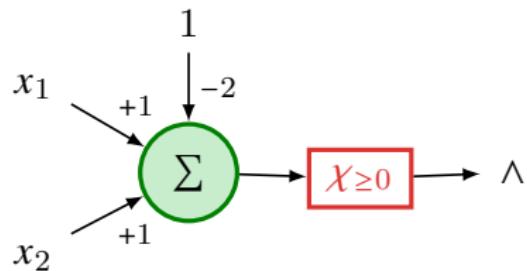
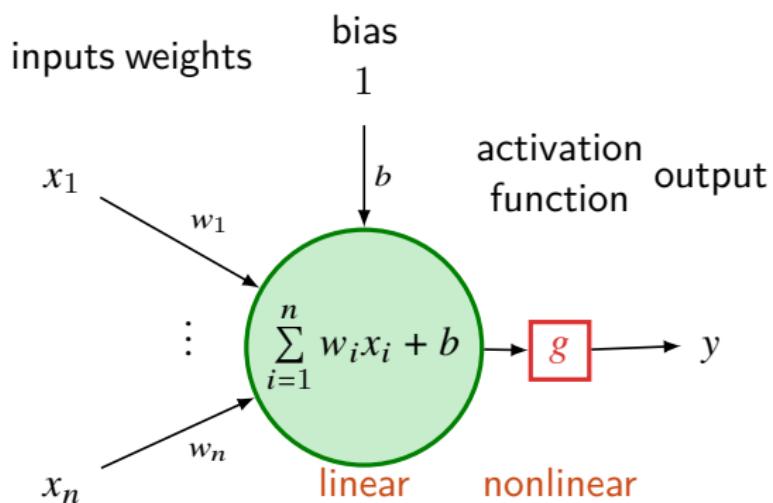
- ▶ There are three different principles of association: resemblance, spatial and temporal contiguity, and causation, which purport to capture the regularities by which the imagination recombines simple ideas into complex ideas.
- ▶ The memory, senses, and understanding are founded on the imagination, or the vivacity of our ideas.

Traditional associationist architectures represent knowledge by simple connection weights. (e.g., between the nodes of a neural network) Bayesian associative models represent knowledge as probability distributions (degrees of belief).

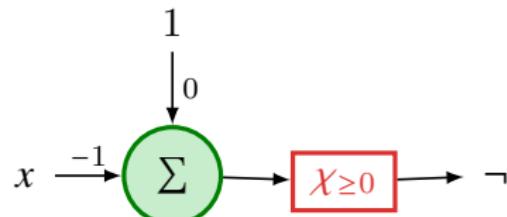


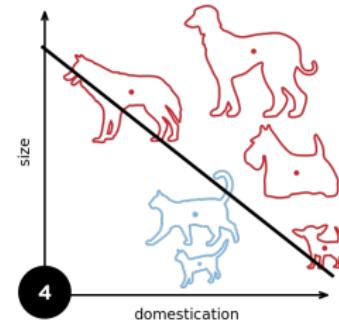
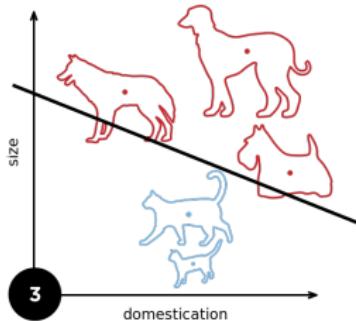
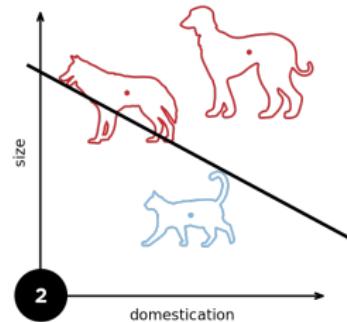
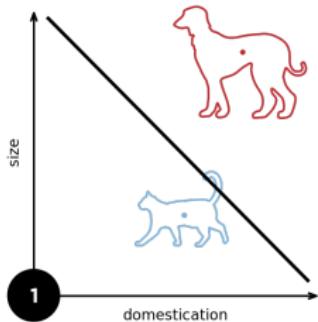
- ▶ 人脑有 1000 亿脑细胞, 100 万亿动态连接.
- ▶ 神经元通过树突接收电信号, 并通过轴突发射出去

# McCulloch-Pitts 人工神经网络 (神经的逻辑演算)



$$y = g \left( \sum_{i=1}^n w_i x_i + b \right)$$





1-layer NN

$$y = \begin{cases} 1 & \text{if } \sum_{i=1}^n w_i x_i + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

# 线性不可分问题

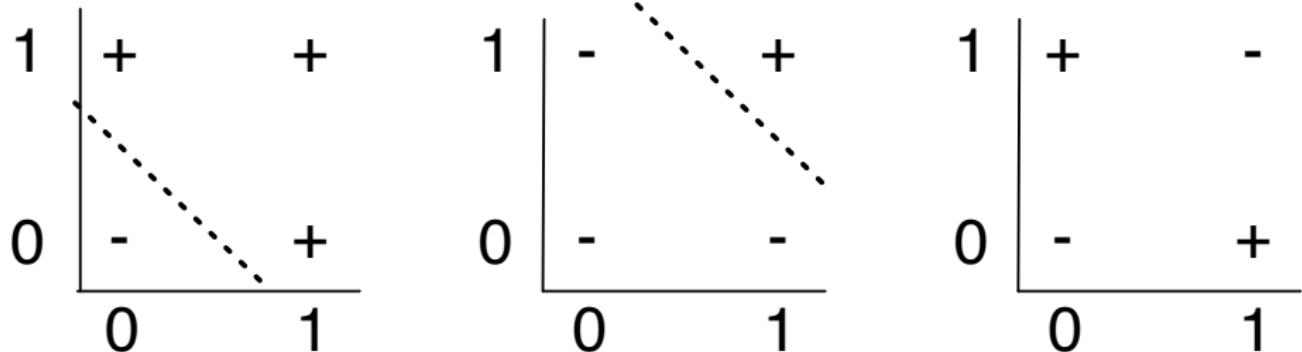


Figure:  $\vee, \wedge, \oplus$

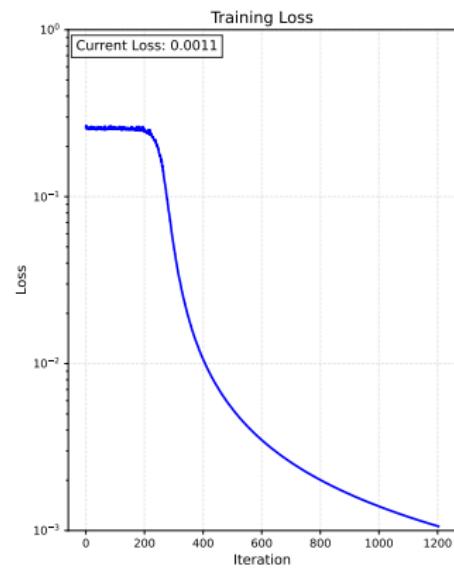
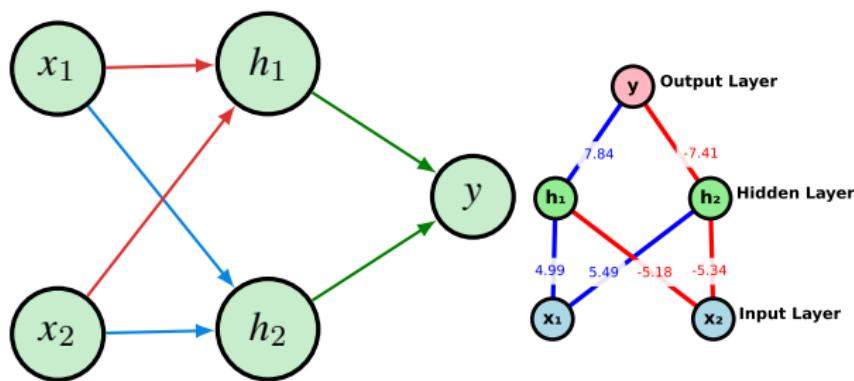
$x_1$	$x_2$	$x_1 \oplus x_2$
0	0	0
0	1	1
1	0	1
1	1	0

$$\begin{array}{lll} w_1 0 + w_2 0 + b < 0 & & b < 0 \\ w_1 0 + w_2 1 + b \geq 0 & & w_2 + b \geq 0 \\ w_1 1 + w_2 0 + b \geq 0 & & w_1 + b \geq 0 \\ w_1 1 + w_2 1 + b < 0 & & w_1 + w_2 + b < 0 \end{array}$$

A simple single-layer perception can't solve nonlinearly separable problems.

# 异或问题

$$\underbrace{x_1 \oplus x_2}_{y} \equiv (\underbrace{\neg x_1 \wedge x_2}_{h_1}) \vee (\underbrace{x_1 \wedge \neg x_2}_{h_2})$$



# 《三体》—人列计算机



- ▶ 秦始皇：朕当然需要预测太阳的运行，但你们让我集结三千万大军，至少要首先向朕演示一下这种计算如何进行吧？
- ▶ 冯诺依曼：陛下，请给我三个士兵，我将为您演示。… 我们组建一千万个这样的门部件，再将这些部件组合成一个系统，这个系统就能进行我们所需要的运算，解出那些预测太阳运行的微分方程<sup>a</sup>。

---

$$^a \text{即 } \frac{d^2\mathbf{r}_i}{dt^2} = - \sum_{j \neq i} Gm_j \frac{\mathbf{r}_i - \mathbf{r}_j}{|\mathbf{r}_i - \mathbf{r}_j|^3} \quad i = 1, 2, 3$$

用连续信号模拟离散信号会怎样？

# 深度神经网络

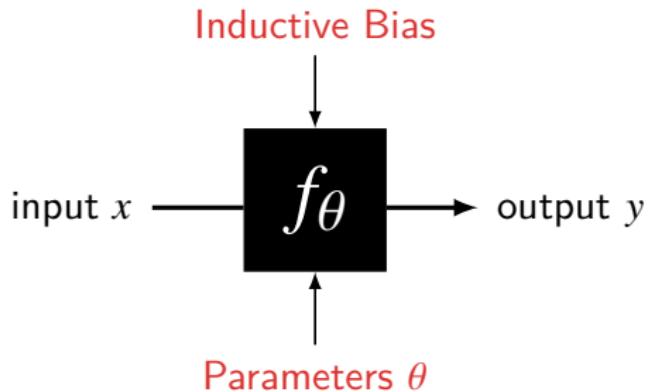
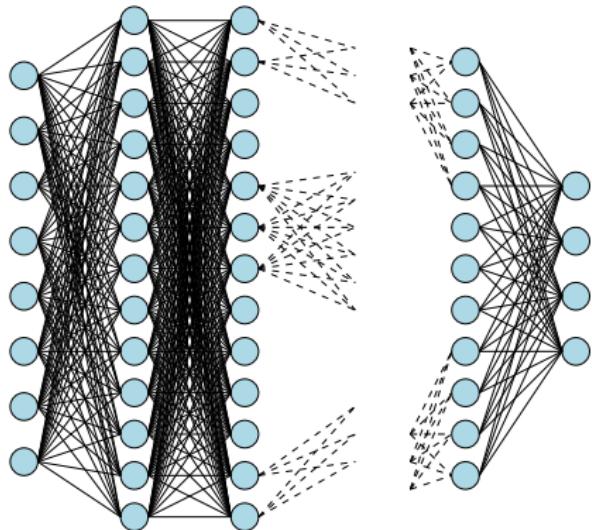
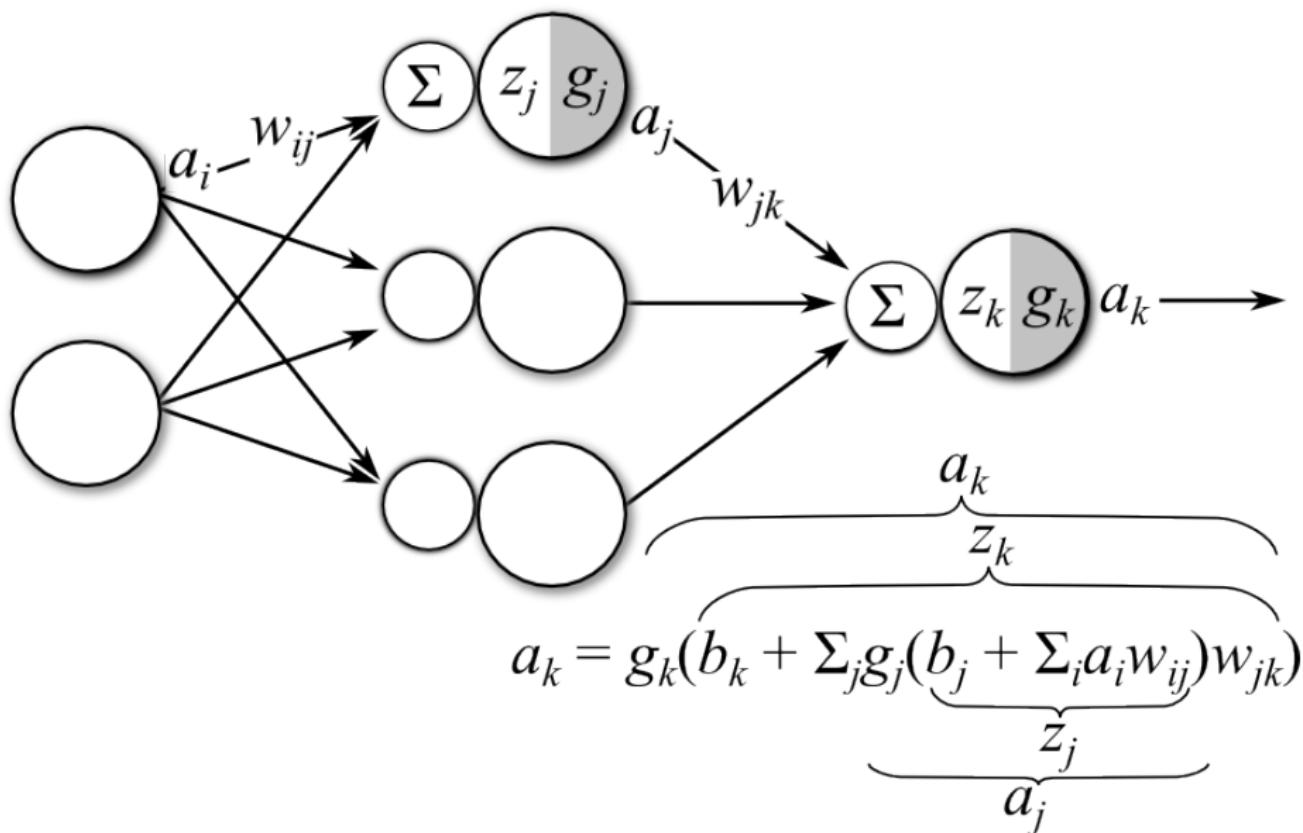


Figure: Walter Pitts & Jeff Hinton



Learning: small change in weights  $\rightarrow$  small change in output

## 流形学习、概率学习

- ▶ 自然界中的数据集的内在模式可以看作是嵌入高维背景空间中的某个低维流形上的概率分布.
- ▶ 学习的主要任务就是学习流形的拓扑结构和流形上的概率分布.
- ▶ 把高维空间映射到隐空间, 本质上是求一个同胚映射, 把数据流形局部映射到隐空间, 这个过程是编码, 从隐空间返回到数据流形的过程是解码. 正则性理论保证编码映射和解码映射是连续的乃至光滑的, 解的唯一性保证这些映射是拓扑同胚或者微分同胚.
- ▶ 流形结构学习归结为在欧氏空间上所有映射构成的空间中进行变分.
- ▶ 概率分布学习归结为在流形上所有概率分布构成的 Wasserstein 空间中进行带有限制的、关于特殊能量的变分优化.
- ▶ 编码解码映射和数据概率分布的传输映射都通过深度神经网络通用逼近.

<sup>18</sup>雷娜、顾险峰: 《最优传输理论与计算》

## Lemma (Urysohn's Lemma)

Let  $A, B$  be two non-empty disjoint closed subset of a normal topological space  $X$ . There exists a continuous function  $f : X \rightarrow [0, 1]$  such that  $f(A) = 0$  and  $f(B) = 1$ .

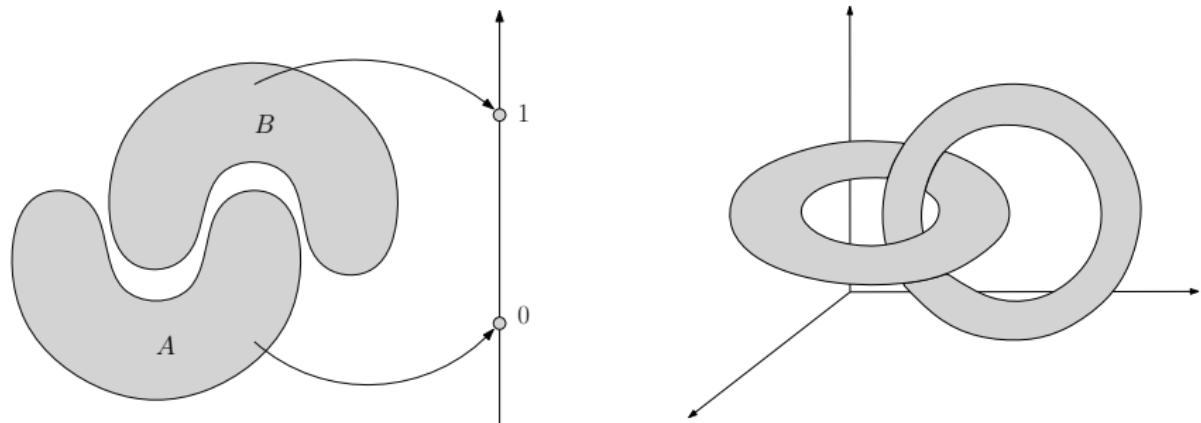


Figure: Urysohn 引理为监督学习、模式识别提供了理论基础.

Figure: 提高嵌入空间的维数, 可以为数据流形解套.

## Theorem (General Position Theorem)

Any  $m$ -manifold unknots in  $\mathbb{R}^n$  provided  $n \geq 2m + 2$ .

# 流形嵌入定理

如果初始流形嵌入的空间维数过高, 通过改变嵌入空间而实现逐步降维, 直至隐空间.

## Theorem (Whitney Embedding Theorem)

*Any smooth real  $n$ -dimensional manifold (required also to be Hausdorff and second-countable) can be smoothly embedded into  $\mathbb{R}^{2n}$ .*

**Remark:** Whitney 定理给出了流形嵌入的普适方法: 首先构造流形的一个有限开覆盖  $\{U_i\}$ , 得到单位分解  $\{\rho_i\}$ ; 构造局部嵌入  $\varphi_i$  将每个开集  $U_i$  嵌入到线性子空间  $\mathbb{R}^n$  中, 用单位分解将局部嵌入合成全局嵌入; 然后进行随机投影, 降低嵌入空间的维数.

## Problem (Hilbert's 13th Problem)

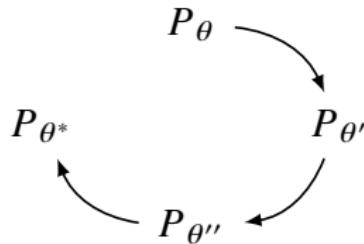
Can every continuous function of  $n$  variables be expressed as a composition of finitely many continuous functions of two variables?

### Problem

Is it possible to exactly represent any continuous multivariate function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  as a combination of continuous univariate functions  $\mathbb{R} \rightarrow \mathbb{R}$  and the single binary function '+'?

神经网络通过复合简单函数逼近复杂函数.

神经网络的随机梯度下降 SGD (+ 注意力机制) 可以看做在程序空间中进行搜索.



## Kolmogorov Superposition Theorem

### Theorem (Kolmogorov Superposition Theorem)

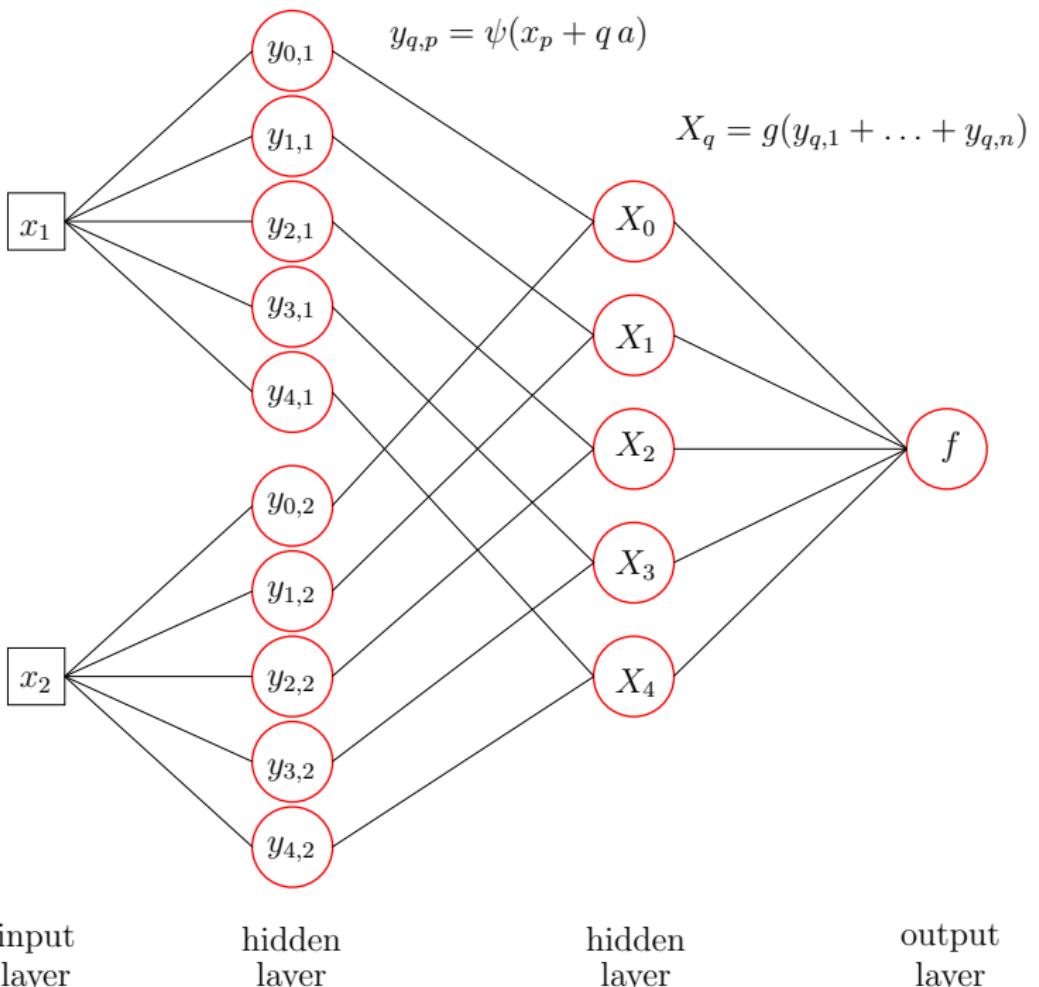
For each  $n \geq 2$  there exists a recursive function  $\psi : [0, 1] \rightarrow \mathbb{R}$  and computable constants  $a, \lambda_{pq} \in \mathbb{R}$ ,  $p = 1, \dots, n$ ,  $q = 0, \dots, 2n$  s.t.: every continuous function  $f : [0, 1]^n \rightarrow \mathbb{R}$  has a representation as

$$f(x_1, \dots, x_n) = \sum_{q=0}^{2n} g\left(\sum_{p=1}^n \lambda_{pq} \psi(x_p + qa)\right)$$

for some continuous function  $g : [0, 1] \rightarrow \mathbb{R}$  that is computable from  $f$ .

### Theorem (Hecht-Nielsen Theorem)

The class of functions  $f : [0, 1]^n \rightarrow \mathbb{R}$ , implementable by three-layer feed-forward neural networks with (computable) continuous activation functions  $g : [0, 1] \rightarrow \mathbb{R}$  and (computable) weights  $\lambda \in \mathbb{R}$ , is exactly the class of (computable) continuous functions  $f : [0, 1]^n \rightarrow \mathbb{R}$ .



# 万能函数拟合器

## Theorem (Universal Approximation Theorem)

Let  $g$  be a nonconstant, bounded, and increasing continuous function. Let  $I_n$  be any compact subset of  $\mathbb{R}^n$ . The space of continuous functions on  $I_n$  is denoted by  $C(I_n, \mathbb{R})$ . Then, given any function  $f \in C(I_n, \mathbb{R})$  and  $\varepsilon > 0$ , there exists an integer  $N$ , real constants  $v_i, b_i \in \mathbb{R}$  and real vectors  $\mathbf{w}_i \in \mathbb{R}^n$ , where  $i = 1, \dots, N$ , s.t.

$$\forall \mathbf{x} \in I_n : |f(\mathbf{x}) - h(\mathbf{x})| < \varepsilon$$

where

$$h(\mathbf{x}) := \sum_{i=1}^N v_i g\left(\mathbf{w}_i^\top \mathbf{x} + b_i\right)$$

In other words, functions of the form  $h(\mathbf{x})$  are dense in  $C(I_n, \mathbb{R})$ .

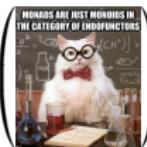
# Representation & Approximation

A deep neural network (DNN) is a particular kind of function

$$f_{\theta} : \mathbb{R}^m \rightarrow \mathbb{R}^n$$

which depends in a “differentiable” way on a vector of weights.

**Example:**  $f_{\theta}$



$$= \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} \in \mathbb{R}^2 \quad p_{\text{cat}} = 0.8, p_{\text{dog}} = 0.2$$

- ▶ A feed-forward network with 1 hidden layer can represent any boolean function, but require exponential hidden units.
- ▶ A feed-forward network with 2 hidden layers and (computable) continuous activation functions can represent any (computable) continuous function.
- ▶ A feed-forward network with a linear output layer and at least 1 hidden layer and continuous and differentiable activation functions can approximate any Borel measurable function from one finite-dimensional space to another with any desired non-zero amount of error.
- ▶ A feed-forward network with 2 hidden layers and continuous and differentiable activation functions can approximate any function.

# Boltzmann Machine

- The global energy  $E$  in a Boltzmann machine is

$$E(x) = - \left( \sum_{i \neq j} w_{ij} x_i x_j + \sum_i b_i x_i \right)$$

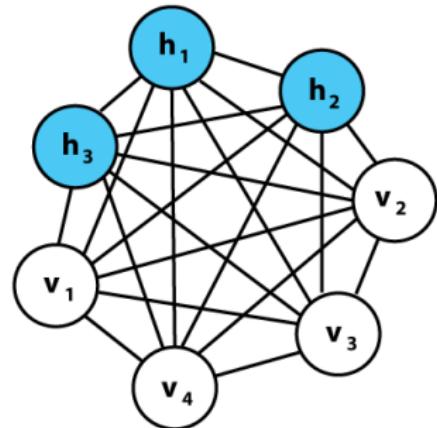
where

- $w_{ij}$  is the connection strength between unit  $i$  and unit  $j$ .
- $x_i \in \{0, 1\}$  is the state of unit  $i$ .
- $b_i$  is the bias of unit  $i$ .
- Unit  $i$  turns on with probability

$$P(x_i = 1) = \sigma \left( b_i + \sum_{j \neq i=0} w_{ij} x_j \right)$$

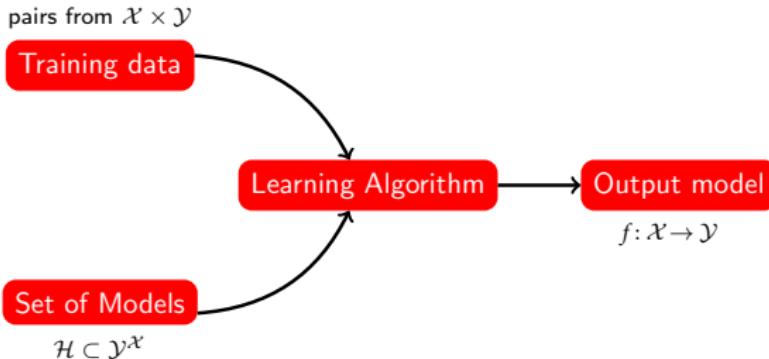
- The stationary distribution

$$P(x) = \frac{e^{-E(x)}}{Z}$$



**Remark:** Voter model for opinion dynamics / Ising model / Hopfield network / Restricted Boltzmann Machine

# Deep Learning[GBC16]



☺ 要把大象装冰箱, 总共分几步?

1. hypothesis space — Network Structure —  $f_{\theta}$
  2. the goodness of a function — Learning Target — loss function  $\ell$
  3. pick the best function — Learn — find the network parameters
- $\theta^* := \underset{\theta}{\operatorname{argmin}} L(\theta)$  that minimize total cost  $L(\theta)$  by gradient decent

$$\theta \leftarrow \theta - \eta \nabla_{\theta} L(\theta)$$

where  $L(\theta) := \mathbb{E}_P [\ell(f_{\theta}(\mathbf{a}), t)] + \lambda \Omega(\theta)$  and  $\Omega(\theta)$  is a regularizer.

# Error

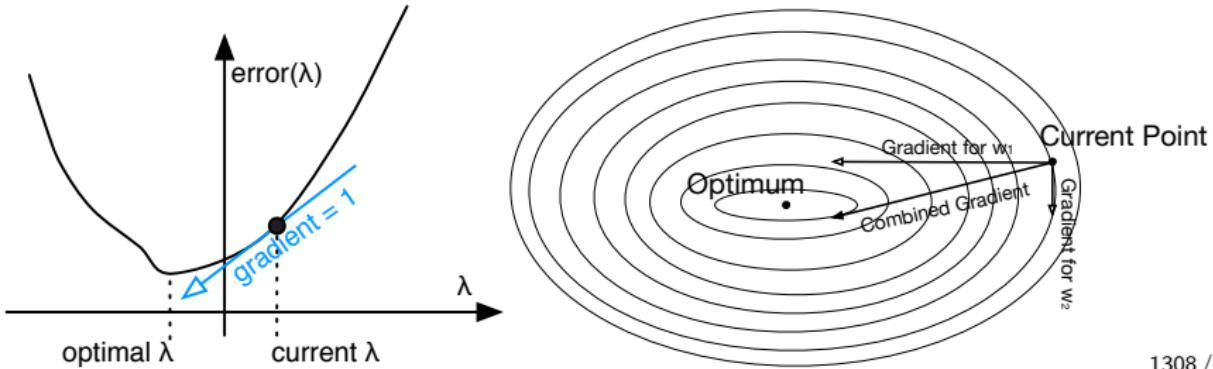
How do we adjust the weights?

- ▶ Gradient descent

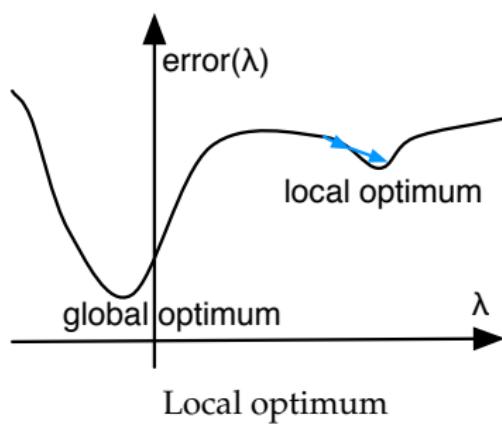
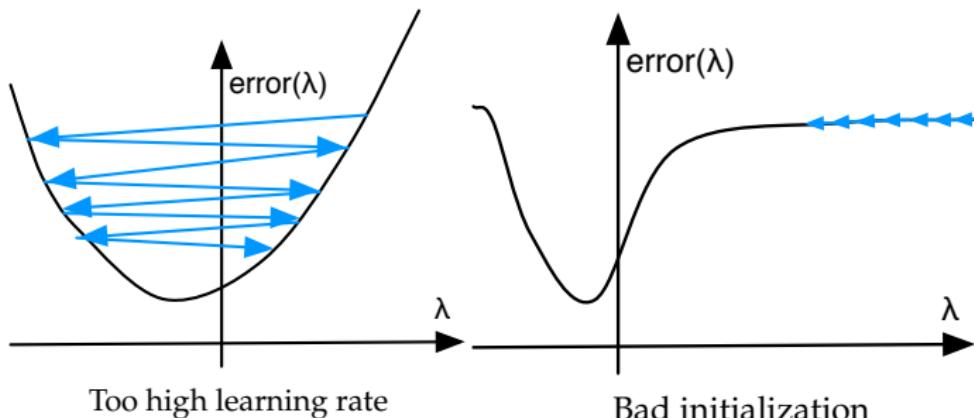
- ▶ error is a function of the weights
- ▶ we want to reduce the error
- ▶ gradient descent: move towards the error minimum
- ▶ compute gradient: get direction to the error minimum
- ▶ adjust weights towards direction of lower error

- ▶ Back-propagation

- ▶ first adjust last set of weights
- ▶ propagate error back to each previous layer
- ▶ adjust their weights



# Problems with Gradient Descent Training



# Deep Learning

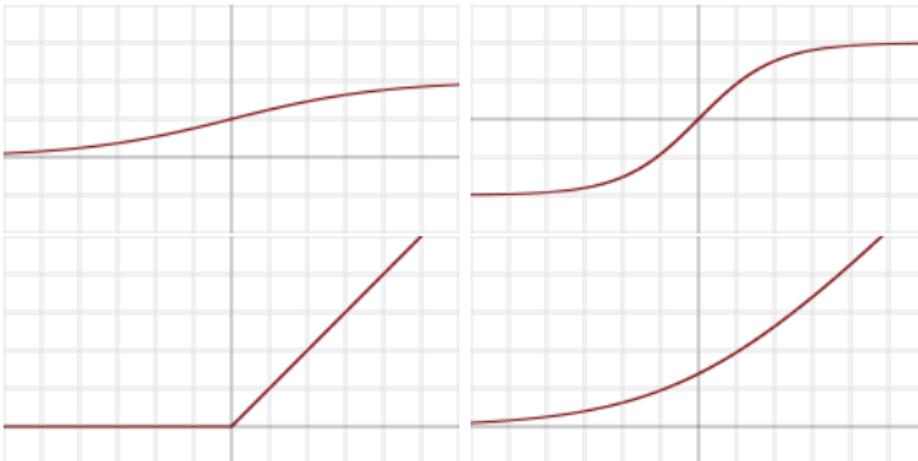
$$f_{\theta} : \mathbf{a}^{(0)} \mapsto \mathbf{g}^{(n)} \left( \cdots \mathbf{g}^{(2)} \left( \overbrace{\mathbf{w}^{(2)} \mathbf{g}^{(1)} \left( \underbrace{\mathbf{w}^{(1)} \mathbf{a}^{(0)} + \mathbf{b}^{(1)}}_{\mathbf{z}^{(1)}} \right) + \mathbf{b}^{(2)}}^{\mathbf{z}^{(2)}} \right) \cdots \right)$$

$\overbrace{\mathbf{a}^{(1)} \quad \mathbf{z}^{(n)}}^{\mathbf{a}^{(n)}}$

where parameters  $\theta := \{\mathbf{w}^{(i)}, \mathbf{b}^{(i)}\}_{i=1}^n$  and  $\mathbf{g}$  is the activation function.

$$\left. \begin{array}{l} \mathbf{a}_0^{(l)} := 1 \\ \mathbf{w}_{0j}^{(l)} := b_j^{(l)} \end{array} \right\} \implies \left\{ \begin{array}{l} z_j^{(l+1)} := \sum_i w_{ij}^{(l+1)} a_i^{(l)} \\ a_j^{(l+1)} := g_j^{(l+1)} (z_j^{(l+1)}) \end{array} \right.$$

# Activation Functions



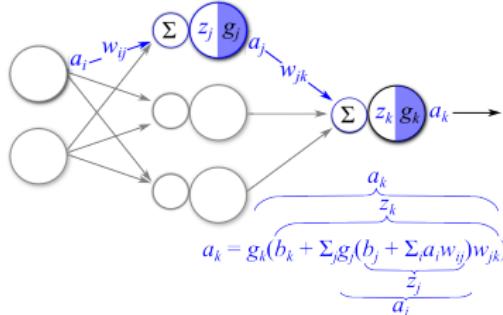
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

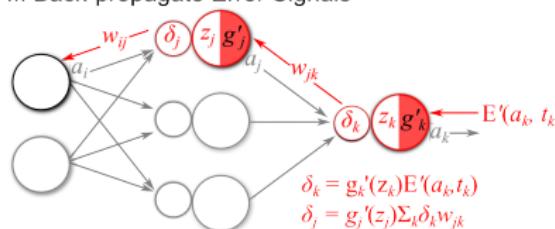
$$\text{ReLU}(z) = \max(0, z)$$

$$\text{softplus}(z) = \log 1 + e^z$$

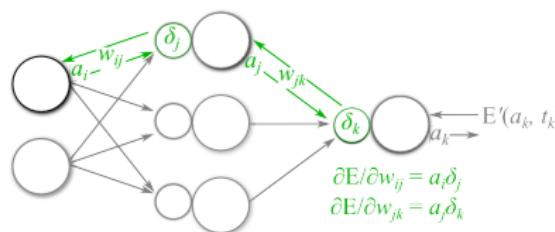
#### I. Forward-propagate Input Signal



#### II. Back-propagate Error Signals



#### III. Calculate Parameter Gradients



#### IV. Update Parameters

$$w_{ij} = w_{ij} - \eta (\partial E / \partial w_{ij})$$

$$w_{jk} = w_{jk} - \eta (\partial E / \partial w_{jk})$$

for learning rate  $\eta$

## Backpropagation

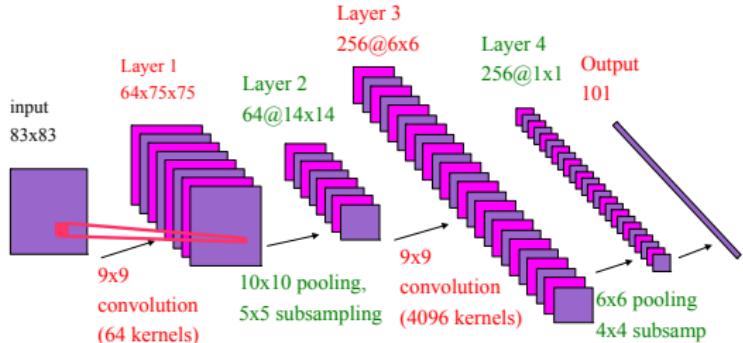
$$\delta_j^{(l)} := \frac{\partial L}{\partial z_j^{(l)}}$$

$$\delta_j^{(l)} = g_j^{(l)'}(z_j^{(l)}) \sum_k \delta_k^{(l+1)} w_{jk}^{(l+1)}$$

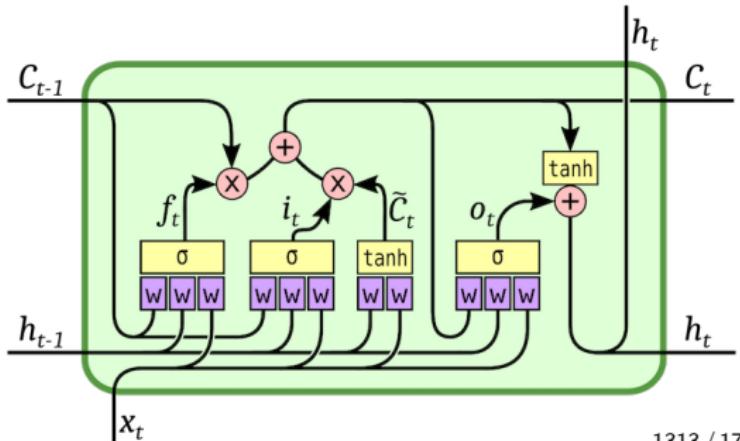
$$\frac{\partial L}{\partial w_{ij}^{(l)}} = a_i^{(l-1)} \delta_j^{(l)}$$

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \frac{\partial L}{\partial w_{ij}^{(l)}}$$

- ▶ network structure?
- ▶ how many layers?
- ▶ how many units per layer?
- ▶ loss function?
- ▶ regularization?
- ▶ weight decay?
- ▶ learning rate?
- ▶ activation function?
- ▶ early stopping?
- ▶ dropout?
- ▶ mini-batch?
- ▶ momentum?
- ▶ ...



$$a_{ij}^{(l+1)} = g_{ij}^{(l+1)} \left( \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} w_{m,n} a_{i+m, j+n}^{(l)} + b \right)$$

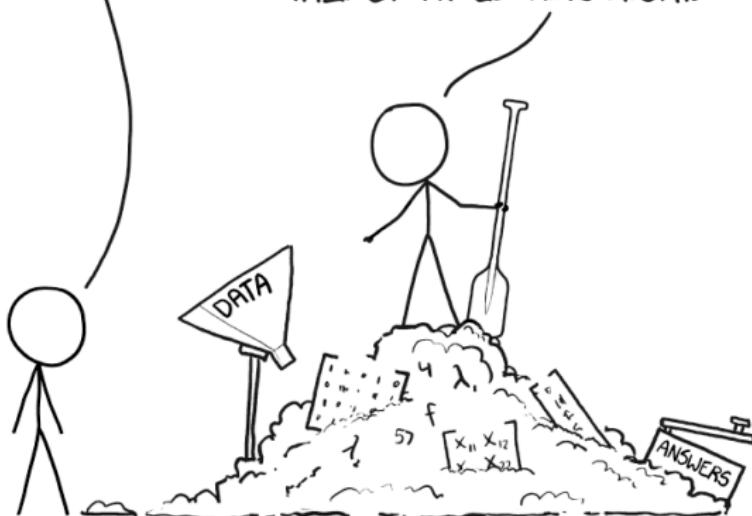


THIS IS YOUR MACHINE LEARNING SYSTEM?

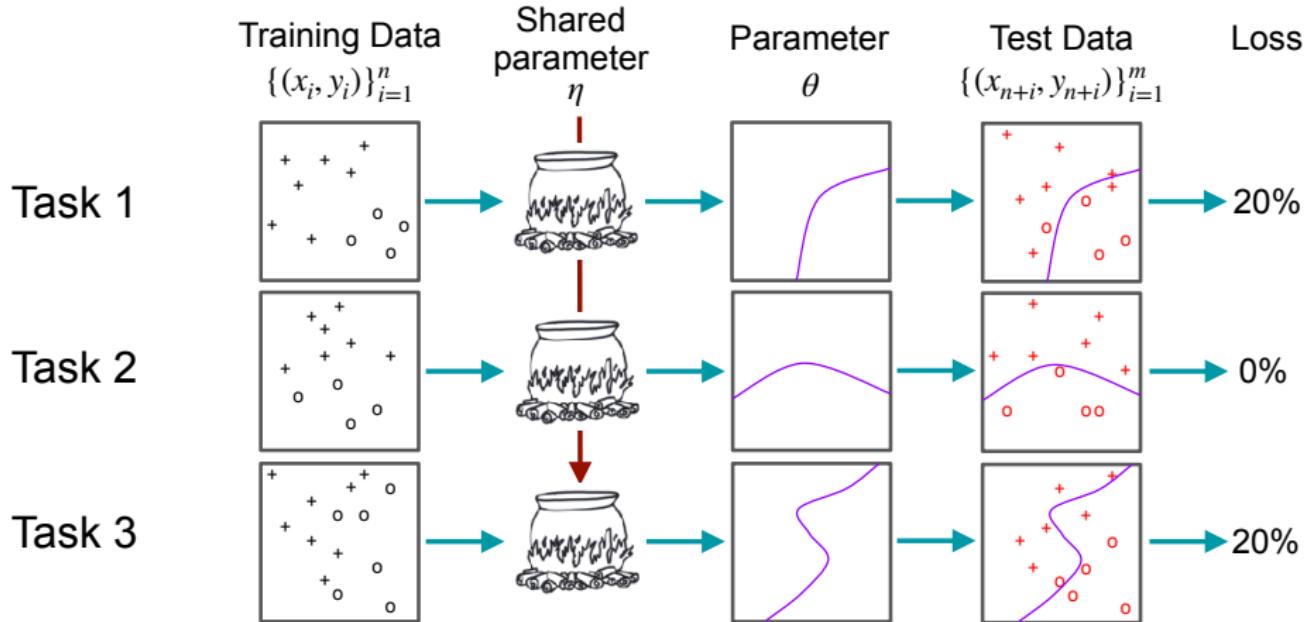
YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL  
THEY START LOOKING RIGHT.



# Single-Task Learning → Multi-Task Learning



$$\operatorname{argmin}_{\eta, \{\theta_j\}} \sum_{j=1}^T \sum_{i=1}^n \text{Loss}\left(f_{\eta, \theta_j}(x_{ji}), y_{ji}\right) + \lambda \Omega(\theta_j)$$

# Key Properties of CNNs

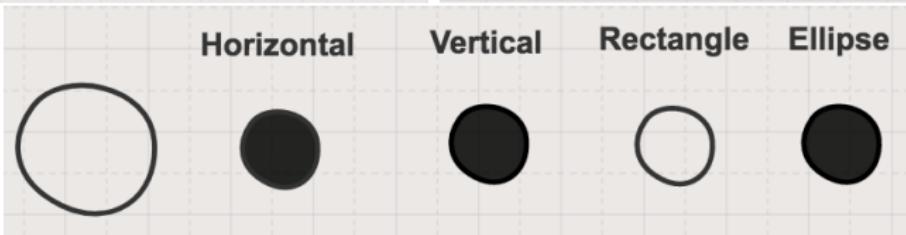
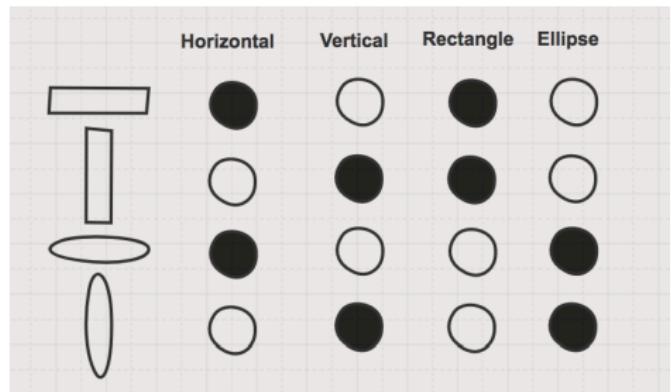
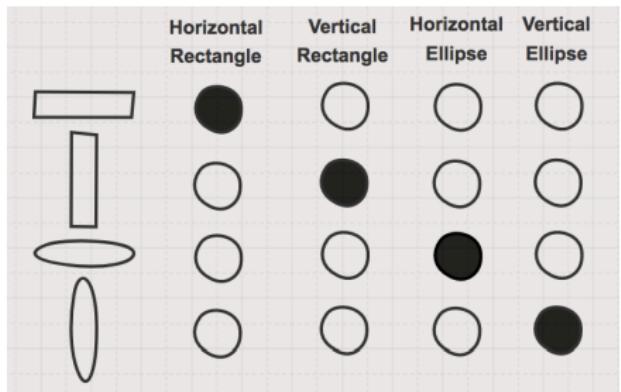
$$\begin{array}{|c|c|c|c|c|c|} \hline 1 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 1 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 1 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 1 & 0 & 1 & 0 \\ \hline \end{array} * \begin{array}{|c|c|c|} \hline 1 & -1 & -1 \\ \hline -1 & 1 & -1 \\ \hline -1 & -1 & 1 \\ \hline \end{array} = \begin{array}{|c|c|c|c|} \hline 3 & -1 & -3 & -1 \\ \hline -3 & 1 & 0 & -3 \\ \hline -3 & -3 & 0 & 1 \\ \hline 3 & -2 & -2 & -1 \\ \hline \end{array}$$

Table: Convolution (stride 1)

Take advantage of the structure of the data!

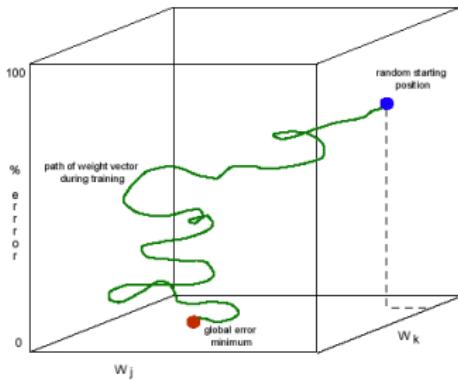
- ▶ Convolutional Filters (**Translation invariance**)
- ▶ Multiple layers (**Compositionality**)
- ▶ Filters localized in space (**Locality**)
- ▶ Weight sharing (**Self-similarity**)

# How Neural Networks Learn Distributed Representations



<b>Number</b>	<b>Local Representation</b>	<b>Distributed Representation</b>
0	10000000	000
1	01000000	001
2	00100000	010
3	00010000	011
4	00001000	100
5	00000100	101
6	00000010	110
7	00000001	111

# Epistemology — A Neurocomputational Perspective



- ▶ 知识不过是“精心调参的连接权重”(突触权重空间中的一个点), 不是一堆或一串存储的符号项.
- ▶ Churchland's Eliminative Materialism: 意向性不存在. 心灵状态不存在.
- ▶ 大脑神经网络对外界信息的记录是整体性的, 同时记录了“兔子吃草”“草是绿的”“头戴草帽”等等信息. 没有某个或某几个神经元表征“兔子”概念. 命题态度、信念、欲望、qualia 等等都没有神经元对应.

## Why “Deep” rather than “Fat”?

- ▶ Exploiting compositionality gives an exponential gain in representational power.
  - ▶ Distributed representations: feature learning
  - ▶ Deep architecture: multiple levels of feature learning
- ▶ Each basic classifier can be trained by little data.
  - ▶ deep → modularization → less training data?  
With more complex features, the number of parameters in the linear layers may be drastically decreased.
  - ▶ efficiency & sample complexity
  - ▶ better memory/computation trade-off?
- ▶ higher-level abstractions → easier generalization & transfer

# Minimal Sufficient Statistic

## Definition (Sufficient Statistic)

Let  $Y$  be a parameter indexing a family of probability distributions. Let  $X$  be random variable drawn from a probability distribution determined by  $Y$ .  $T(X)$  is a sufficient statistic for  $Y$  iff  $X$  is independent of  $Y$  given  $T(X)$ , i.e.  $p(x | t, y) = p(x | t)$ .

## Definition (Minimal Sufficient Statistic)

A sufficient statistic  $S(X)$  is minimal iff for any sufficient statistic  $T(X)$ , there exists a function  $f$  s.t.  $S = f(T)$  almost everywhere w.r.t  $X$ .

## Theorem

- ▶  $T$  is sufficient statistics for  $Y \iff I(T(X); Y) = I(X; Y)$ .
- ▶  $S$  is minimal sufficient statistics for  $Y \implies I(X; S(X)) \leq I(X; T(X))$ .

# Information Bottleneck — Learning is to forget!

Can we explain learning in deep neural networks?

## Theorem

Let  $X$  be a sample drawn according to a distribution determined by the random variable  $Y$ . The set of solutions to

$$\min_T I(X; T) \quad s.t. \quad I(T; Y) = \max_{T'} I(T'; Y)$$

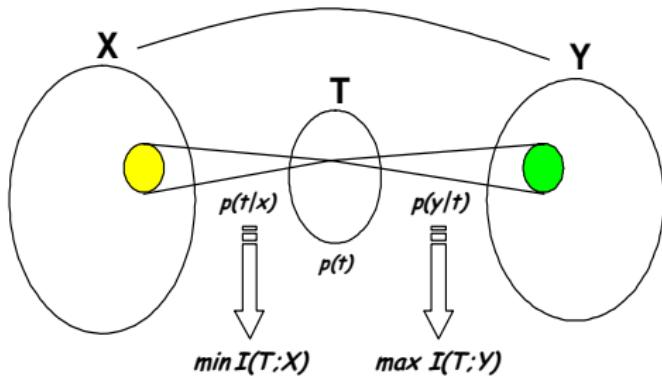
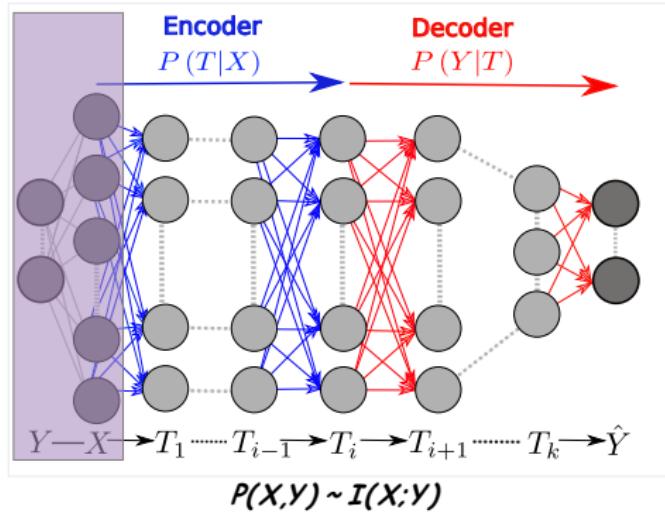
is exactly the set of minimal sufficient statistics for  $Y$  based on  $X$ .

Find a random variable  $T$  s.t.:

- ▶  $Y \leftrightarrow X \leftrightarrow T$  form a Markov chain.
- ▶  $I(X; T)$  is minimized (minimality, complexity term), while  $I(T; Y)$  is maximized (sufficiency, accuracy term).

$$T^* := \underset{T: I(T(X); Y) = I(X; Y)}{\operatorname{argmin}} I(X; T(X))$$

is the Information Bottleneck between  $X$  and  $Y$ .



张三丰：将所见到的剑招忘得半点不剩，才能得其神髓。

老子：为学日益，为道日损。  
损之又损，以至于无为。  
无为而无不为。

广中平祐：记住和忘记相当于将能力拉扯又放松，可以让能力变得更有弹性。

## Information Bottleneck

$\min_{p(t|x), p(t), p(y|t)} \left\{ I(X; T) - \beta I(T; Y) \right\}$  subject to Markov chain  $Y \rightarrow X \rightarrow T$ .

$$L[p(t | x)] := I(X; T) - \beta I(T; Y) - \sum_x \lambda(x) \sum_t p(t | x)$$

Let

$$\frac{\delta L}{\delta p(t | x)} = 0$$

The solution is

$$p(t | x) = \frac{p(t)}{Z(x, \beta)} e^{-\beta D[p(y|x) \| p(y|t)]}$$

$$p(t) = \sum_x p(t | x) p(x)$$

$$p(y | t) = \frac{1}{p(t)} \sum_x p(t | x) p(x, y)$$

# The Information Bottleneck Method<sup>19</sup>

---

## Algorithm The Information Bottleneck Method

---

Given  $p(x, y)$ ,  $\beta \geq 0$

Initialize  $p^{(0)}(t | x)$

$$p^{(0)}(t) = \sum_x p^{(0)}(t | x)p(x)$$

$$p^{(0)}(y | t) = \frac{1}{p^{(0)}(t)} \sum_x p^{(0)}(t | x)p(x, y)$$

$$n = 0$$

**while** not converged **do**

$$n = n + 1$$

$$p^{(n)}(t | x) = \frac{p^{(n-1)}(t)}{Z(x, \beta)} e^{-\beta D[p(y|x) \| p^{(n-1)}(y|t)]}$$

$$p^{(n)}(t) = \sum_x p^{(n)}(t | x)p(x)$$

$$p^{(n)}(y | t) = \frac{1}{p^{(n)}(t)} \sum_x p^{(n)}(t | x)p(x, y)$$

**end while**

---

<sup>19</sup> Schwartz-Ziv, Tishby: Opening the black box of Deep Neural Networks via Information.

# The Deterministic Information Bottleneck

$$L_\alpha[p(t | x)] := H(T) - \alpha H(T | X) - \beta I(T; Y) - \sum_x \lambda(x) \sum_t p(t | x)$$

$$p(t | x) = \frac{1}{Z(x, \alpha, \beta)} e^{\frac{1}{\alpha} (\log p_\alpha(t) - \beta D[p(y|x) \| p_\alpha(y|t)])}$$

Obviously,  $L_1 = L$ .

$$p_\alpha(t | x) = \frac{1}{Z(x, \alpha, \beta)} e^{\frac{1}{\alpha} (\log p_\alpha(t) - \beta D[p(y|x) \| p_\alpha(y|t)])}$$

$$p_\alpha(t) = \sum_x p_\alpha(t | x) p(x)$$

$$p_\alpha(y | t) = \frac{1}{p_\alpha(t)} \sum_x p_\alpha(t | x) p(x, y)$$

Let  $\alpha \rightarrow 0$ , we get the deterministic case

$$p(t | x) = \lim_{\alpha \rightarrow 0} p_\alpha(t | x) = \left[ t = \operatorname{argmax}_t \left( \log p(t) - \beta D[p(y | x) \| p(y | t)] \right) \right]$$

# The Deterministic Information Bottleneck Method

---

## Algorithm The Deterministic Information Bottleneck Method

---

Given  $p(x, y)$ ,  $\beta \geq 0$

Initialize  $f^{(0)}(x)$

$$p^{(0)}(t) = \sum_{x:f^{(0)}(x)=t} p(x)$$

$$p^{(0)}(y | t) = \frac{\sum_{x:f^{(0)}(x)=t} p(x, y)}{\sum_{x:f^{(0)}(x)=t} p(x)}$$

$$n = 0$$

**while** not converged **do**

$$n = n + 1$$

$$f^{(n)}(x) = \operatorname{argmax}_t \left( \log p^{(n-1)}(t) - \beta D [p(y | x) \| p^{(n-1)}(y | t)] \right)$$

$$p^{(n)}(t) = \sum_{x:f^{(n)}(x)=t} p(x)$$

$$p^{(n)}(y | t) = \frac{\sum_{x:f^{(n)}(x)=t} p(x, y)}{\sum_{x:f^{(n)}(x)=t} p(x)}$$

**end while**

---

## Expressiveness & Sample Complexity

Smooth functions require fewer neurons to approximate.

### Theorem

*The hypothesis class of neural networks of depth  $T$  and size  $O(T^2)$  contains all functions that can be implemented by a Turing machine within  $T$  operations, while having  $O(T^2)$  sample complexity.*

# The Ultimate Hypothesis Space

- ▶ **No Free Lunch:** Sample complexity is exponentially large (w.r.t. the input dimension) if the hypothesis class is all possible functions.
- ▶ **Shallow learning (SVM, Boosting):** Hypothesis class is linear functions over manually determined features — strong prior knowledge.
- ▶ **Deep learning:** Hypothesis class is all functions implemented by determining the weights of a given artificial neural network.

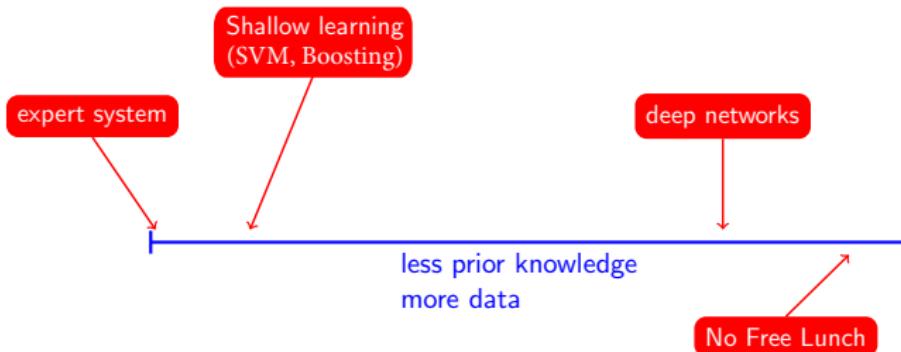


Figure: Prior vs Universality

Prior — a necessary good or a necessary evil?

# 苏格兰的黑羊

- ▶ 在途经苏格兰的火车上有一个工程师，一个物理学家，一个数学家。窗外景物飞掠。突然，他们看到了一只黑色的羊。
- ▶ 工程师：哇！苏格兰的羊是黑色的。
- ▶ 物理学家：错。只能说，在苏格兰，有一只黑色的羊。
- ▶ 数学家：错。只能说，在苏格兰，有一只羊，在这一时刻，从这个角度，用肉眼观察，有一个侧面看上去是黑色的。

In neural networks, we might still control the learning process using **prior knowledge** in the form of:

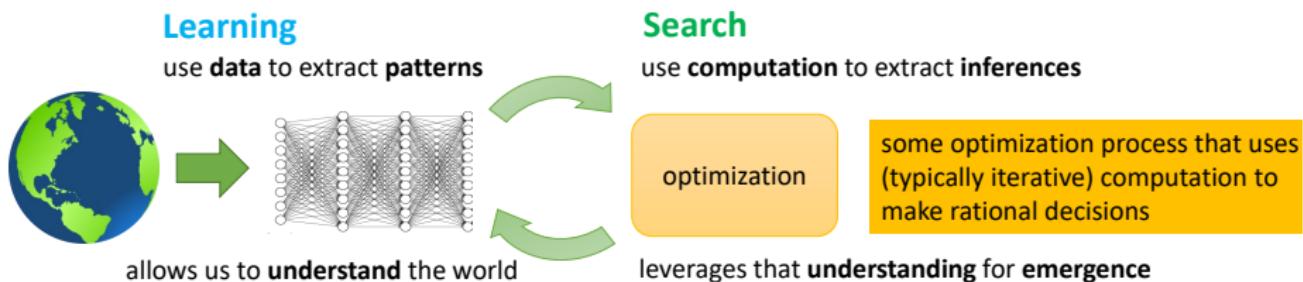
- ▶ **Constraints:** forbidding some outcomes we know to be wrong;
- ▶ **Regularizers:** penalizing some outcomes we know to be less likely;
- ▶ **Invariances:** enforcing some patterns we know to be present:
  - ▶ Rotational invariances for images.
  - ▶ Preservation of objects for videos.
  - ▶ Context dependence for language.

# Why Deep Reinforcement Learning?

## Learning & Search

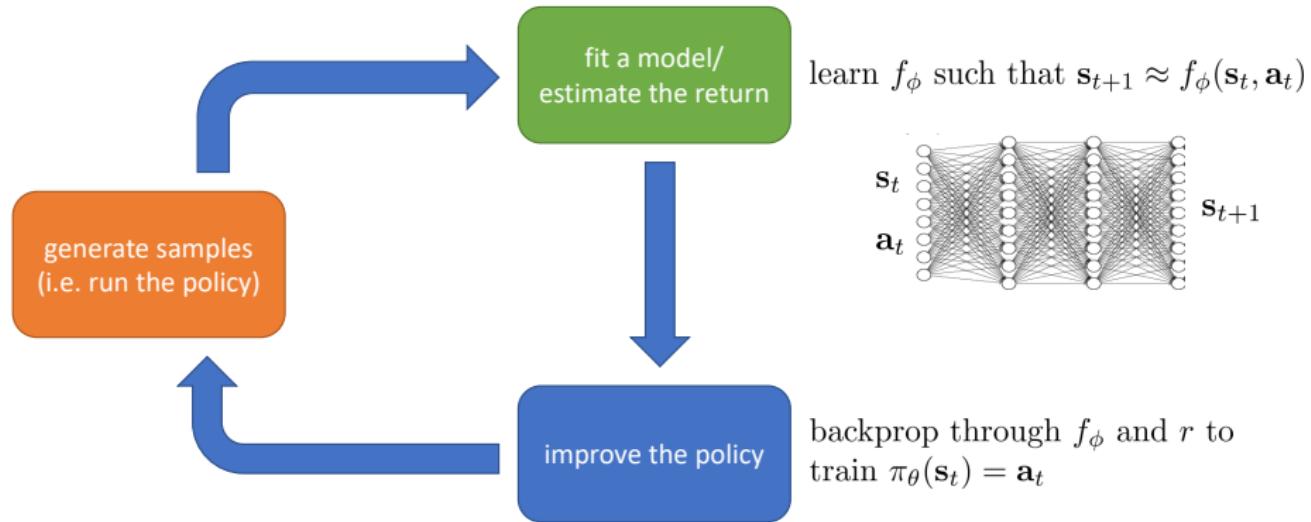
The two methods that seem to scale arbitrarily in this way are search and learning.

— Richard S. Sutton “The Bitter Lesson”



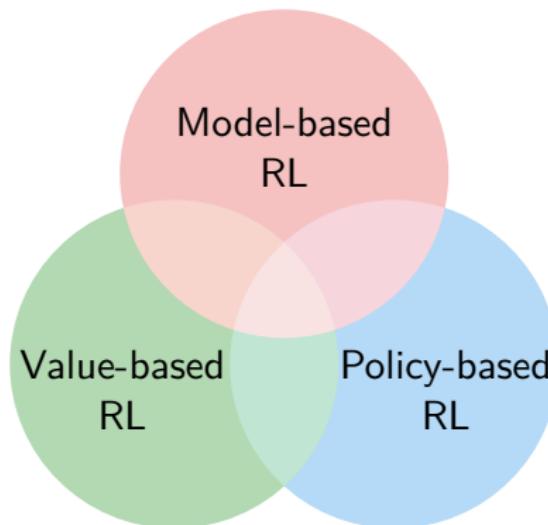
- ▶ Data without optimization doesn't allow us to solve new problems in new ways.
- ▶ Optimization without data is hard to apply to the real world outside of simulators.

# Deep Reinforcement Learning



# Deep Reinforcement Learning

- ▶ Policy-based deep RL: Represent policy  $\pi : \mathcal{S} \rightarrow \mathcal{A}$  as a deep neural network
- ▶ Value-based deep RL: Basically value iteration. Approximate optimal state-value function  $V(s)$  or state-action value function  $Q(s, a)$  with a deep neural network
- ▶ Model-based deep RL: Approximate transition model with a deep neural network



# DQN

$$Q^\pi(s, a) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s, a \right]$$

$$\max_{\pi} Q^\pi(s, a) =: Q^*(s, a) = \mathbb{E}_{s'} \left[ r + \gamma \max_{a'} Q^*(s', a') \mid s, a \right]$$

$$Q_{t+1}(s, a) = \mathbb{E}_{s'} \left[ r + \gamma \max_{a'} Q_t(s', a') \mid s, a \right]$$

$$Q_t \xrightarrow{t \rightarrow \infty} Q^*$$

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha \left( r + \gamma \max_{a'} Q_t(s', a') - Q_t(s, a) \right)$$

$$Q(s, a; \theta) \approx Q^*(s, a)$$

$$L(\theta) := \mathbb{E}_{s, a, r, s'} \left[ \left( r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right] \quad (\text{DQN})$$

# DQN

$$L(\theta) := \mathbb{E}_{s,a,r,s'} \left[ \left( r + \gamma \underset{a'}{\operatorname{argmax}} Q(s', a'; \theta); \theta^- \right)^2 - Q(s, a; \theta) \right]$$

(Double DQN)

$$Q(s, a) = V(s; \theta) + A(s, a; \theta')$$

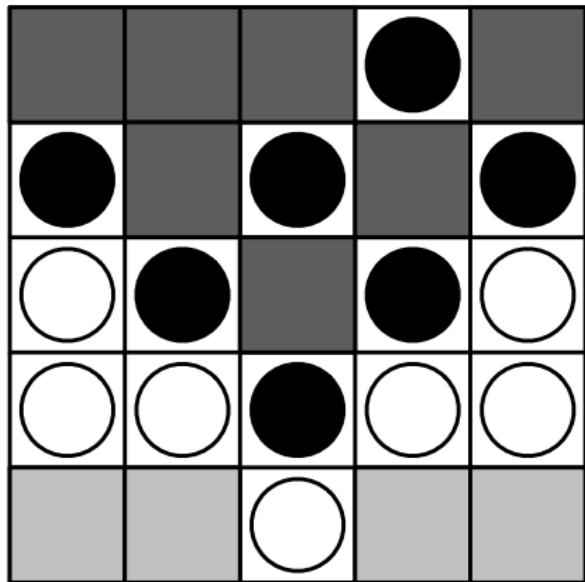
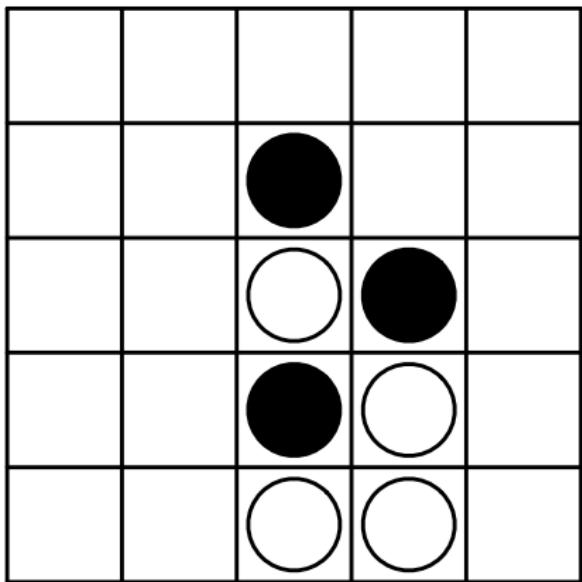
(Dueling Network)

# Actor-Critic Learning

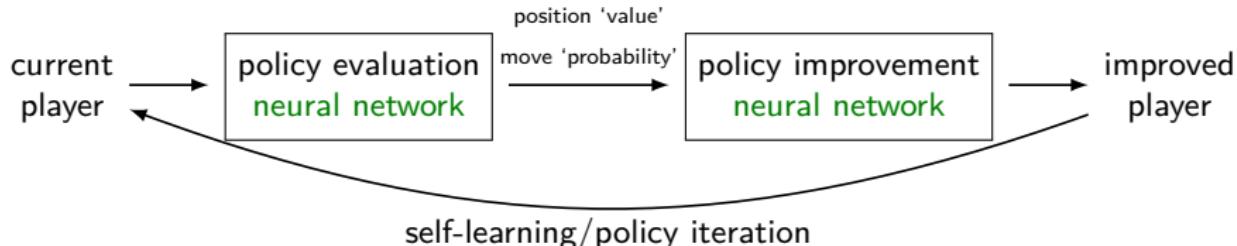
- ▶ Combination of policy learning and  $Q$  learning
  - ▶ actor: move predictor (as in policy learning)  $s \rightarrow a$
  - ▶ critic: value of state (as in  $Q$  learning)  $V(s)$
- ▶ We use this setup to influence how much to boost good moves
  - ▶ advantage  $A(s, a) = Q(s, a) - V(s)$
  - ▶ good moves when advantage is high

$$\left\{ \begin{array}{l} \nabla_{\theta} \log \pi(a_t | s_t; \theta) \left( \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}; \theta') - V(s_t; \theta') \right) \quad \text{actor} \\ \nabla_{\theta'} \left( \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}; \theta') - V(s_t; \theta') \right)^2 \quad \text{critic} \end{array} \right. \quad (\text{AC})$$

Go



# AlphaGo 2016 / AlphaZero 2017



- ▶ Reducing breadth with policy network
- ▶ Reducing depth with value network

$$(\mathbf{p}, v) = f_{\theta}(s)$$

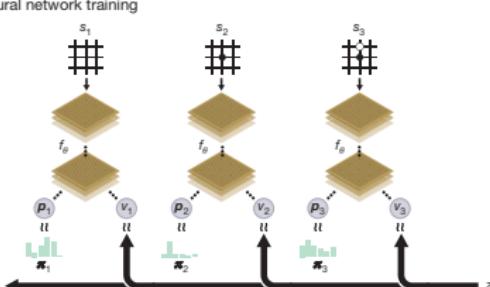
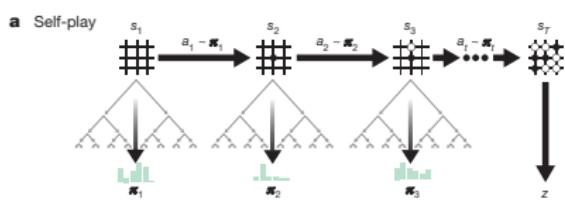
$$l = (z - v)^2 - \pi^T \log \mathbf{p} + c \|\theta\|^2$$

$$\pi_t \leftarrow \text{MCTS}(f_{\theta}(s_t))$$

$$a_t \sim \pi_t$$

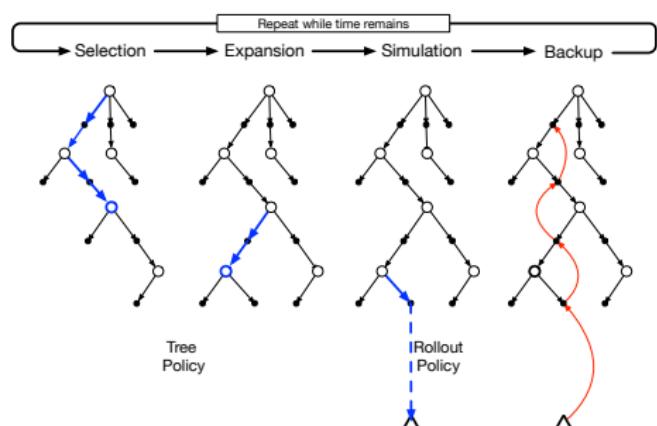
Intuition + Calculation

CNN + MCTS



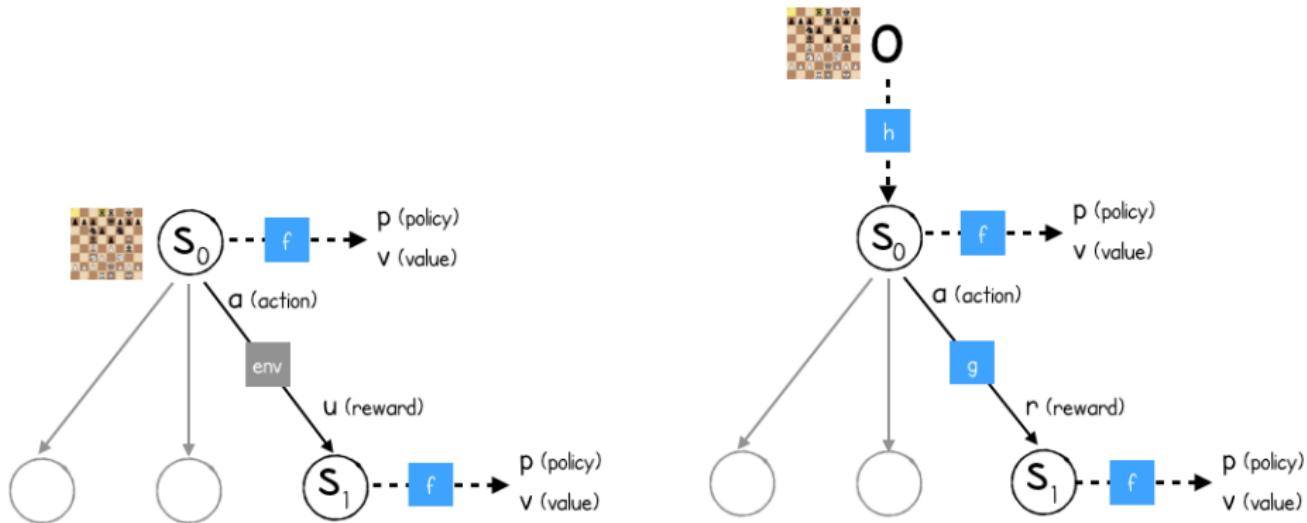
# Monte Carlo Tree Search

- ▶ Do  $N$  fast rollouts from each child of the root to a leaf, count wins  $U$ , record fraction of wins  $\frac{U}{N}$
- ▶ Pick the move that gives the best outcome
- ▶ Allocate rollouts to **more promising** or **more uncertain** nodes



1. Selection: recursively apply  $UCB(k) = \frac{U(k)}{N(k)} + \sqrt{\frac{\log N(Pa(k))}{N(k)}}$  to choose a path to a leaf node  $n$
2. Expansion: add a new child  $c$  to  $n$
3. Simulation: run a rollout from  $c$
4. Backup: update  $U$  and  $N$  counts from  $c$  back up to the root

# From AlphaZero 2017 to MuZero 2020



AlphaZero has 1 network

**prediction**  $f:$   $s \xrightarrow{\text{from}} p, v$

MuZero has 3 networks

from to

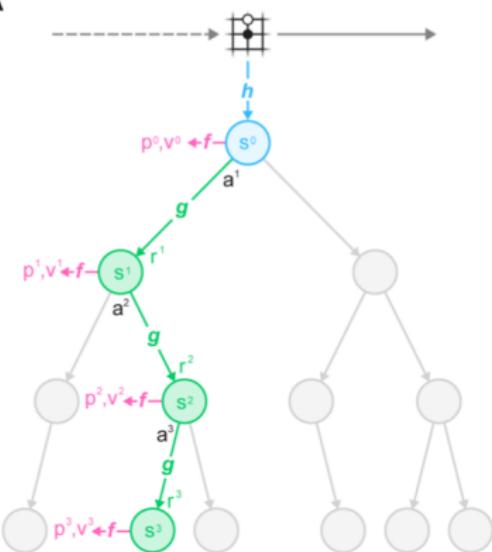
**prediction**  $f:$   $s \rightarrow p, v$

**dynamics**  $g:$   $s, a \rightarrow r; s$

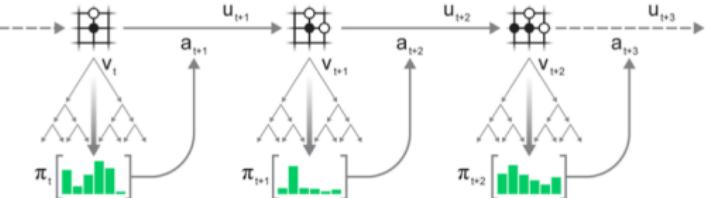
**representation**  $h:$   $o \rightarrow s$

MuZero is discovering for itself how to build a model and understand it just from first principles.

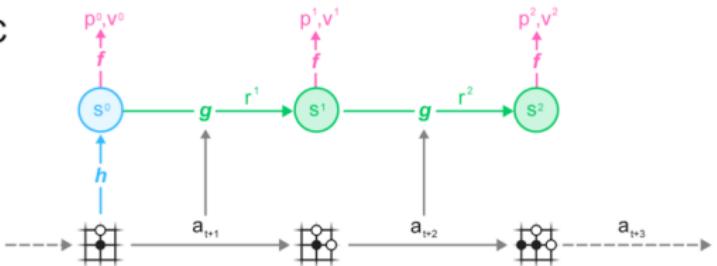
A



B



C



## ► Model

$$\left. \begin{array}{l} s^0 = h_\theta(o_1, \dots, o_t) \\ r^k, s^k = g_\theta(s^{k-1}, a^k) \\ p^k, v^k = f_\theta(s^k) \end{array} \right\} \quad p^k, v^k, r^k = \mu_\theta(o_1, \dots, o_t, a^1, \dots, a^k)$$

## ► Search

$$\nu_t, \pi_t = \text{MCTS}(s_t^0, \mu_\theta)$$

$$a_t \sim \pi_t$$

# MuZero 2020

- ▶ **(A)** How MuZero uses its model to plan. The model consists of three connected components for representation  $h$ , dynamics  $g$  and prediction  $f$ . The initial hidden state  $s^0$  is obtained by passing the past observations  $o_1, \dots, o_t$  into a *representation* function  $h$ .
- ▶ **(B)** How MuZero acts in the environment. A MCTS is performed at each timestep  $t$ . An action  $a_{t+1}$  is sampled from the search policy  $\pi_t$ , which is proportional to the visit count for each action from the root node. The environment receives the action and generates a new observation  $o_{t+1}$  and reward  $u_{t+1}$ .
- ▶ **(C)** How MuZero trains its model. For the initial step, the representation function  $h$  receives as input the past observations  $o_1, \dots, o_t$ . At each step  $k$ , the dynamics function  $g$  receives as input the hidden state  $s^{k-1}$  from the previous step and the real action  $a_{t+k}$ . The parameters of the representation, dynamics and prediction functions are jointly trained to predict three quantities: the policy  $p^k \approx \pi_{t+k}$ , value function  $v^k \approx z_{t+k}$ , and reward  $r_{t+k} \approx u_{t+k}$ .

## The MuZero loss function

$$l_t(\theta) = \sum_{k=0}^K l^r(u_{t+k}, r_t^k) + l^v(z_{t+k}, v_t^k) + l^p(\pi_{t+k}, \mathbf{p}_t^k) + c\|\theta\|^2$$

1. The difference between the predicted reward  $k$  steps ahead of turn  $t$  ( $r$ ) and the actual reward ( $u$ ).
2. The difference between the predicted value  $k$  steps ahead of turn  $t$  ( $v$ ) and the TD target value ( $z$ ).
3. The difference between the predicted policy  $k$  steps ahead of turn  $t$  ( $p$ ) and the MCTS policy ( $\pi$ ).

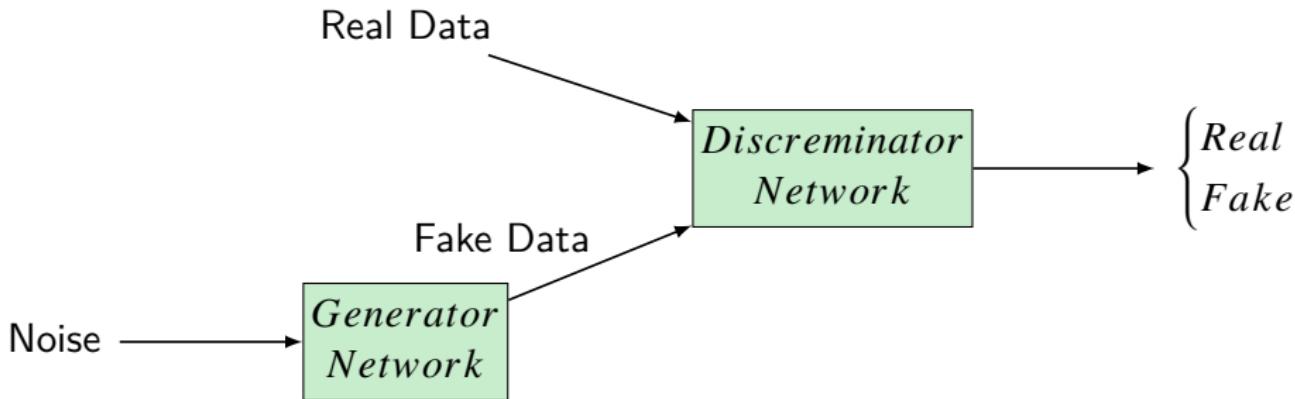
# Scientific & Algorithm Discovery: AlphaFold, AlphaTensor

- ▶ AlphaFold 2021: Highly accurate protein structure prediction for the human proteome
  - AlphaFold can accurately predict a protein's 3D structure from its amino acid sequence.
- ▶ AlphaTensor 2022: Discovering faster matrix multiplication algorithms with reinforcement learning.
  - AlphaTensor is trained to play a single-player game where the objective is finding tensor decompositions within a finite factor space.

# GAN — Generative Adversarial Network 2014

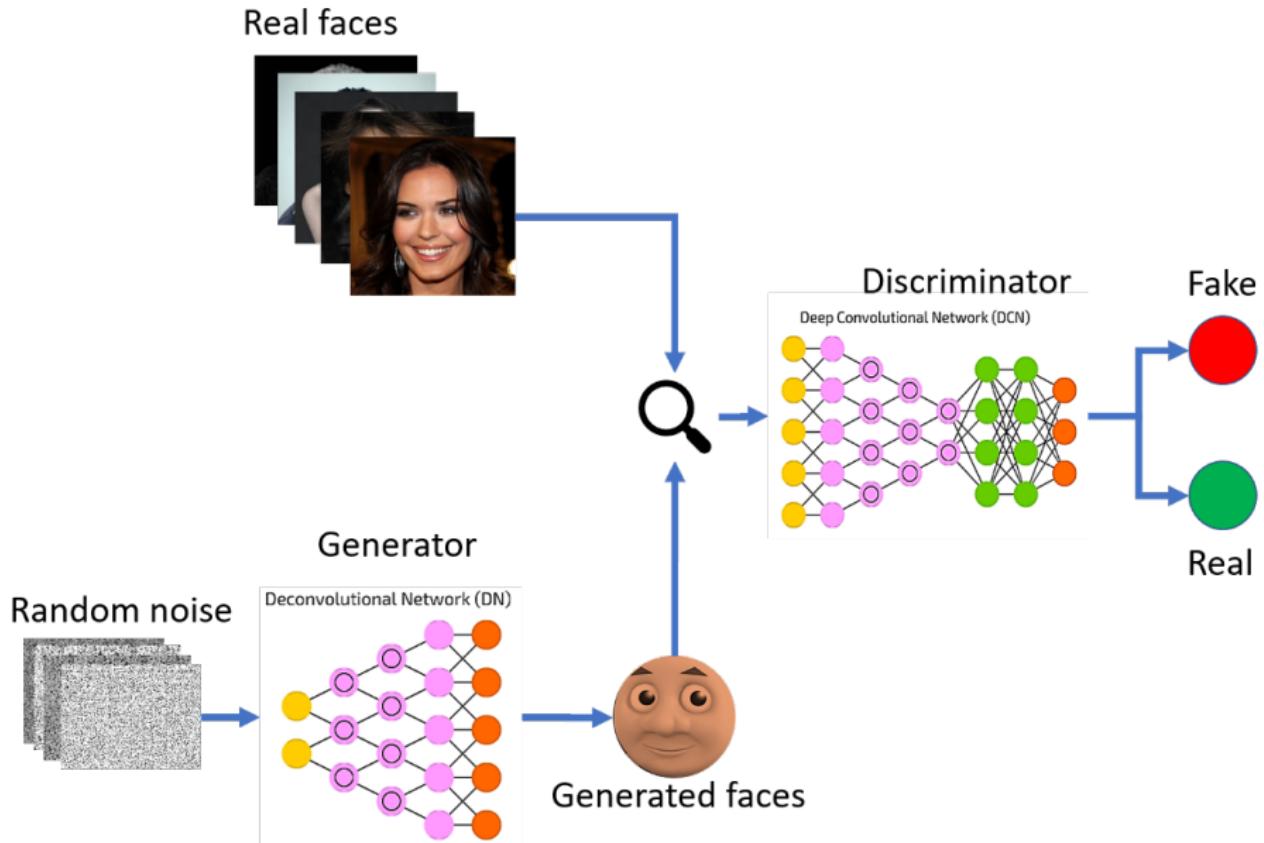
$$V(D, G) = \mathbb{E}_{x \sim P_{data}} [\log D(x)] + \mathbb{E}_{z \sim P_{noise}} [\log (1 - D(G(z)))]$$

$$G^* = \operatorname{argmin}_G \max_D V(D, G) \quad (\text{GAN})$$



- ▶ Initialize generator and discriminator
- ▶ In each training iteration:
  1. fix generator  $G$ , and update discriminator  $D$ :  $D = D + \alpha_D \frac{\partial V}{\partial D}$
  2. fix discriminator  $D$ , and update generator  $G$ :  $G = G - \alpha_G \frac{\partial V}{\partial G}$

# GAN

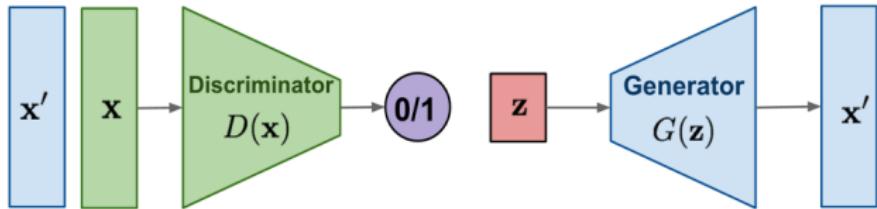


# 协同进化

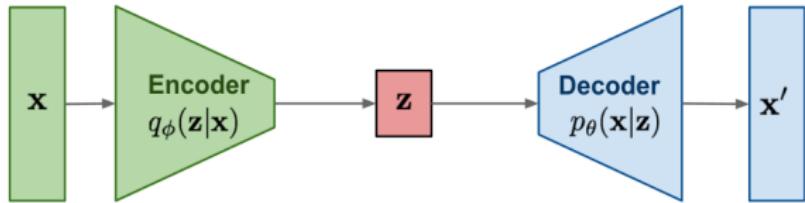


# Generative Model Zoo

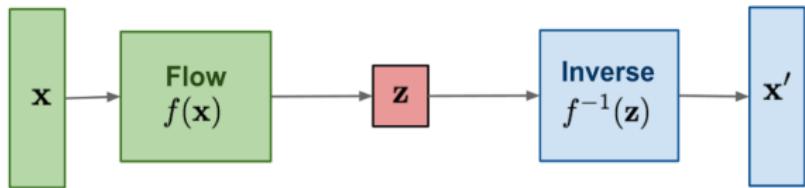
**GAN:** Adversarial training



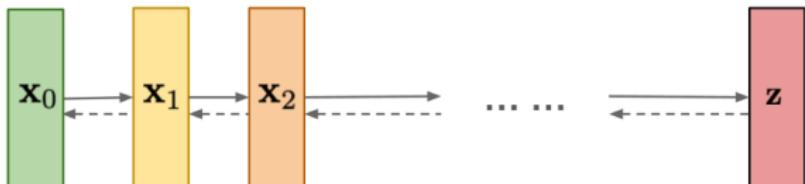
**VAE:** maximize variational lower bound



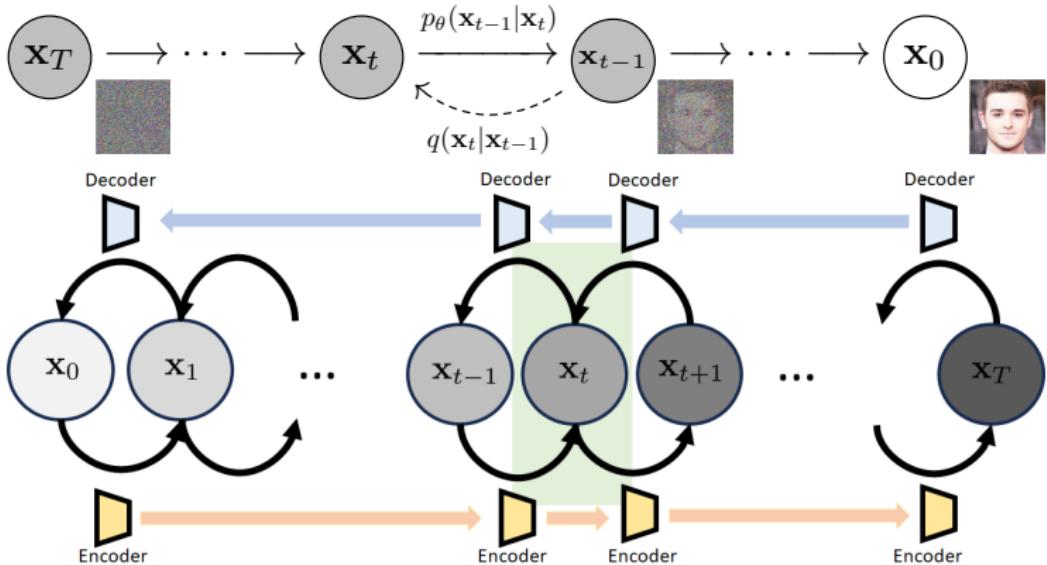
**Flow-based models:**  
Invertible transform of distributions



**Diffusion models:**  
Gradually add Gaussian noise and then reverse



# Stable Diffusion Model



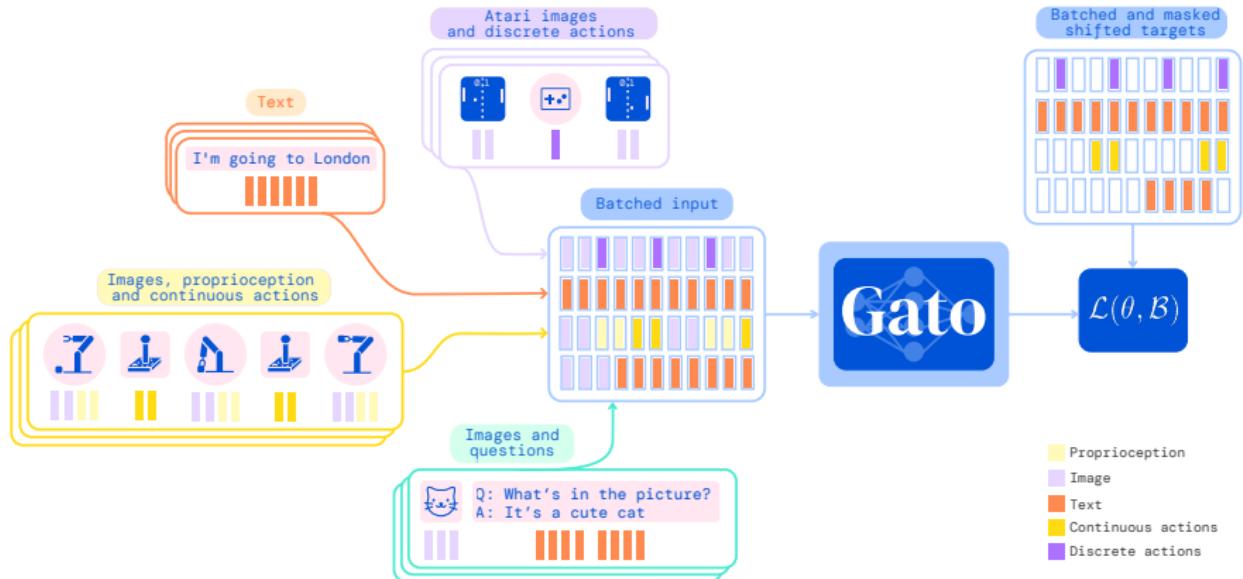
forward from  $x_0$  to  $x_T$  :

$$q_\phi(x_{0:T}) = q(x_0) \prod_{t=1}^T q_\phi(x_t | x_{t-1})$$

reverse from  $x_T$  to  $x_0$  :

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t)$$

# Gato 2022

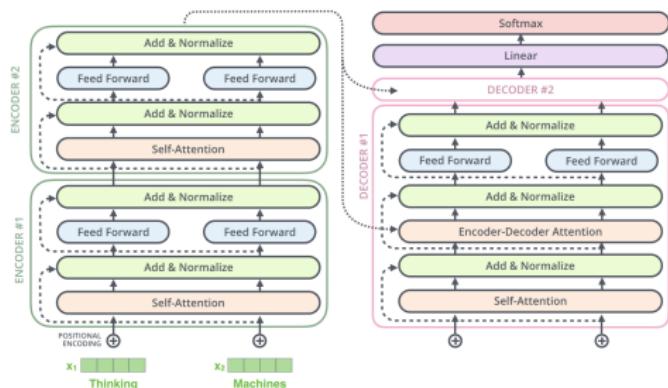
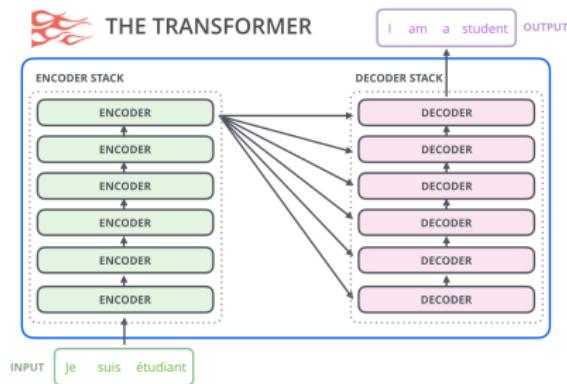


**Figure:** A generalist agent. *Gato* can sense and act with different embodiments across a wide range of environments using a single transformer neural network akin to a large language model with the same set of weights.

604 tasks, 1.2 billion parameters

# From Transformer to BERT & GPT

ChatGPT: 知识检索、闲聊、生成代码、优化代码、做简单的数学、写论文、小说、剧本、诗歌、广告文案、翻译、写作业、制表格、摘要总结、检查错误、润色文章、列大纲、给出雅思口语考试高分答案、写党员心得、写离职信、模仿胡锡进写时评、写小红书风格稿子、做心理/健身/情感咨询、安排旅行行程



- ▶ BERT(Transformer-Encoder) 完形填空: 青苹果乐园
- ▶ GPT(Transformer-Decoder) 文本接龙: 青 |苹果|乐园

Q: 一个猎人, 向南走了 100 米, 向东走了 100 米, 再向北走了 100 米, 发现自己回到了原地. 这时, 他看到了一头熊. 请问, 这头熊是什么颜色的?

A: 谜语描述的路径 (向南 100 米、向东 100 米、向北 100 米) 只有在地球的北极附近才可能使猎人回到原点. 因此, 这头熊很可能是北极熊, 而北极熊的颜色是白色.

# 预测越准确，理解越深刻

假设你正在阅读一本侦探小说，这本书包含了错综复杂的情节、众多不同的角色以及许多令人费解的事件和线索。在故事的最后一页，侦探终于将所有线索收集齐全，召集了所有相关人员，然后宣布：‘现在，我将揭示真凶的身份，那个人的名字是 (...)’

— Ilya Sutskever



## Remarks: Scaling Law

1. 模型大小
2. 数据集大小
3. 强化学习训练 Chain-of-Thought 的时间<sup>20</sup>
4. Chain-of-Thought 推理的时间

涌现的能力:

- ▶ 上下文学习 (In-context learning)
- ▶ 指令微调 (Instruction following)
- ▶ 分步骤推理 (Step-by-step reasoning)

---

<sup>20</sup>Remark:

- ▶ 强化学习是提升推理能力的关键.
- ▶ 预训练模型的质量是能够用强化学习提升推理能力的基础.

# 大语言模型推理 — Example

**[Question]** Using the numbers from 1 to 10 to make 2025. Using each number once and primary operations + and \*

**[Answer]**  $(10 * 4 + 5) * (9 * 3 + 8 + 7 + 2 + 1)$

Let's break it down:

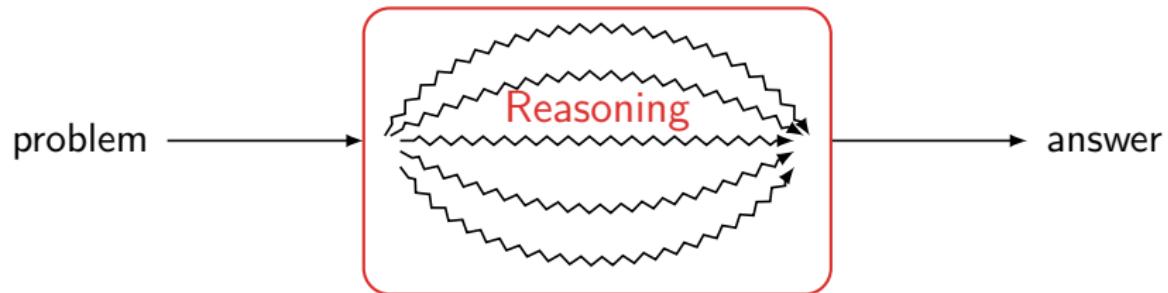
- First part:  $10 * 4 + 5 = 40 + 5 = 45$
- Second part:  $9 * 3 + 8 + 7 + 2 + 1 = 27 + 8 + 7 + 2 + 1 = 45$
- Final Calculation:  $45 * 45 = 2025$

Therefore, the expression  $(10 * 4 + 5) * (9 * 3 + 8 + 7 + 2 + 1)$  uses the numbers 1 through 10 exactly once and only the primary operations of addition (+) and multiplication (\*) to equal 2025.

## Gemini 2.0 thinking mode (Dec, 2024)

The target number is 2025. This is a relatively large number, suggesting multiplication will be heavily involved. It's also worth noting that 2025 is 45 squared ( $45 * 45$ ). This might or might not be directly helpful, but it gives us a sense of scale. \*\*Prioritize Multiplication:\*\* Since the target is large, start thinking about how to get large intermediate products using multiplication. The largest numbers (10, 9, 8, 7) are good candidates for multiplication early on ... This confirms multiplication is key, but simply multiplying everything together is too much. We need to combine additions. ... Let's aim for products that get us closer to the square root of 2025 (which is 45) ...

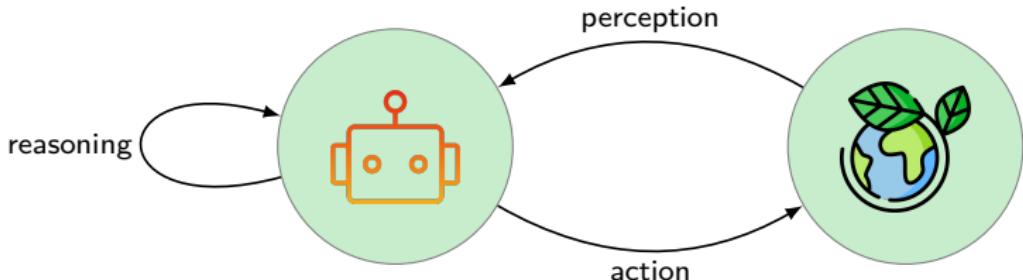
# 大语言模型推理 — LLM Reasoning



$$\text{argmax } \underbrace{P(\text{final answer} \mid \text{problem})}_{\sum_{\text{reasoning path}} P(\text{reasoning path}, \text{final answer} \mid \text{problem})}$$

- ▶ How to compute the sum then? **Randomly Sampling!**
- ▶ Self-consistency: Choose the answer that appears most frequently.

# 大语言模型推理 — LLM Reasoning

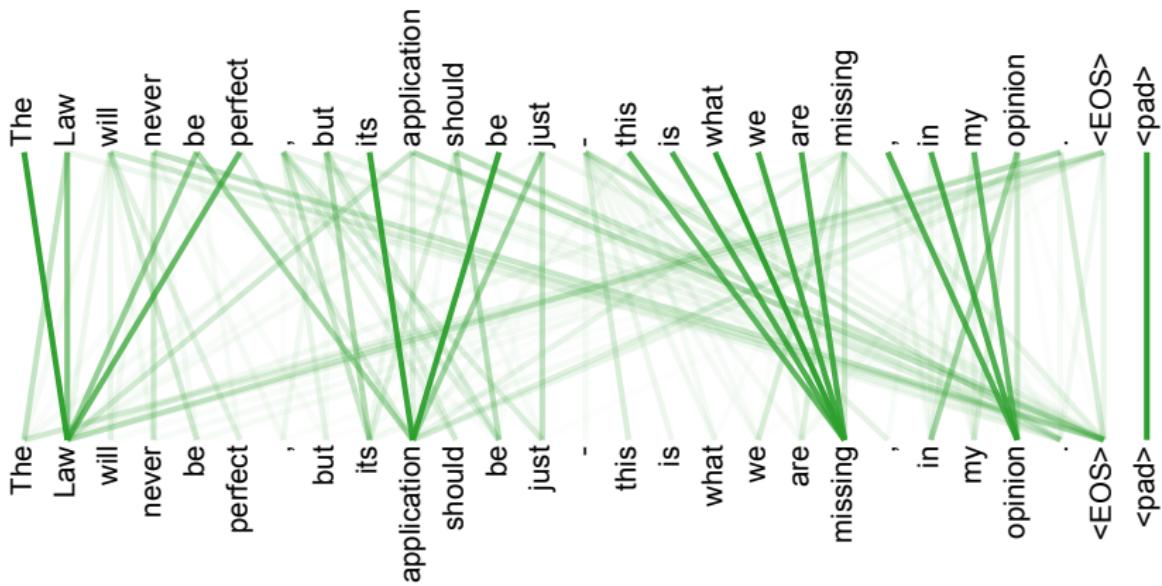


## Self-Improve

1. collect a set of problems and their step-by-step solutions generated from the model
  2. maximize the likelihood of correct solutions
- ▶ Why “generated from the model” instead of “from humans”?
  - ▶ Directly optimize what we want!

$$\operatorname{argmax}_{\theta} \mathbb{E}_P [\text{Quality}(\text{final answer} \mid \text{problem}, \theta)]$$

# Self-attention with learnable weights



- We have three weight matrices  $W_q, W_k, W_v$  for Query, Key and Value
- Query:  $q^{(i)} = W_q x^{(i)}$
- Key:  $k^{(i)} = W_k x^{(i)}$
- Value:  $v^{(i)} = W_v x^{(i)}$
- $\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^\top}{\sqrt{d_k}} \right) V$

## Remark

- ▶ 从图灵机的角度来说, 前馈的消息可以看作是数据, 网络权重  $W$  可以看作是程序. 学习的过程就体现为程序的改变.
- ▶ 自注意力机制的本质在于, 前馈过程可改变注意力, 而注意力又可以看作是一种动态权重加到每个神经元的  $value$  上, 它的作用是和  $W$  类似的.
- ▶ 注意力本身的调节可以被看作是在模拟对权重  $W$  的梯度下降过程. 注意力的调节等价于机器在实现程序的自我修改. (参看 s-m-n 定理)
- ▶ 不妨把前馈运算过程和网络权重比喻成水流和河道, 自注意力机制使得河水在流淌的过程中改变河道的分布.

— 张江



# Wittgenstein — On Certainty

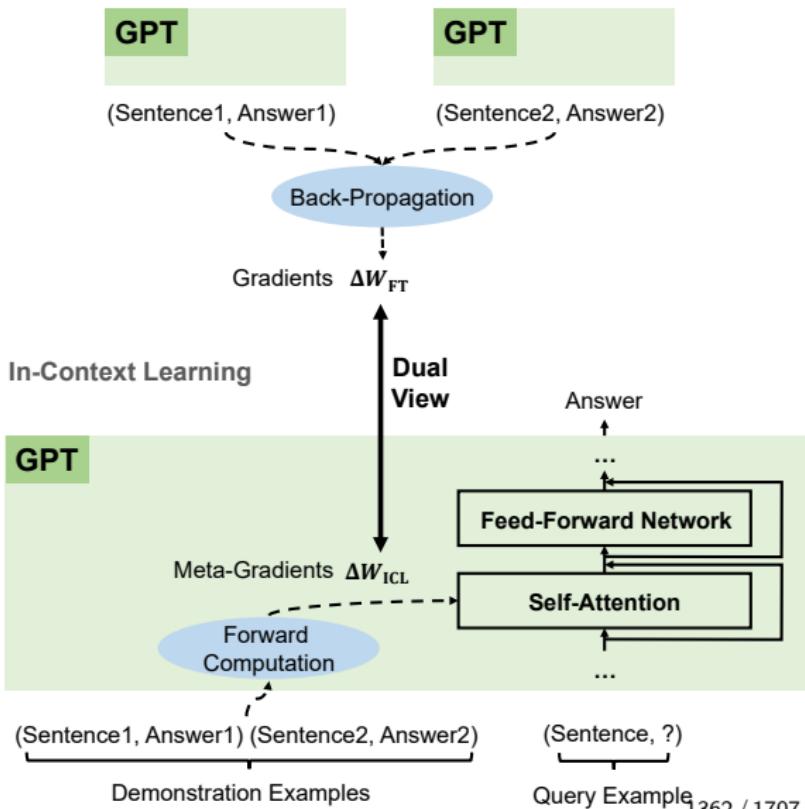
- ▶ The propositions describing this world-picture might be part of a kind of mythology. And their role is like that of rules of a game; and the game can be learned purely practically, without learning any explicit rules.
- ▶ The mythology may change back into a state of flux, river-bed of thoughts may shift. But I distinguish between movement of the waters on the river-bed and the shift of the bed itself; though there is not a sharp division of the one from other.
- ▶ And the bank of that river consists partly of hard rock, subject to no alteration or only to an imperceptible one, partly of sand, which now in one place now in another gets washed away, or deposited.



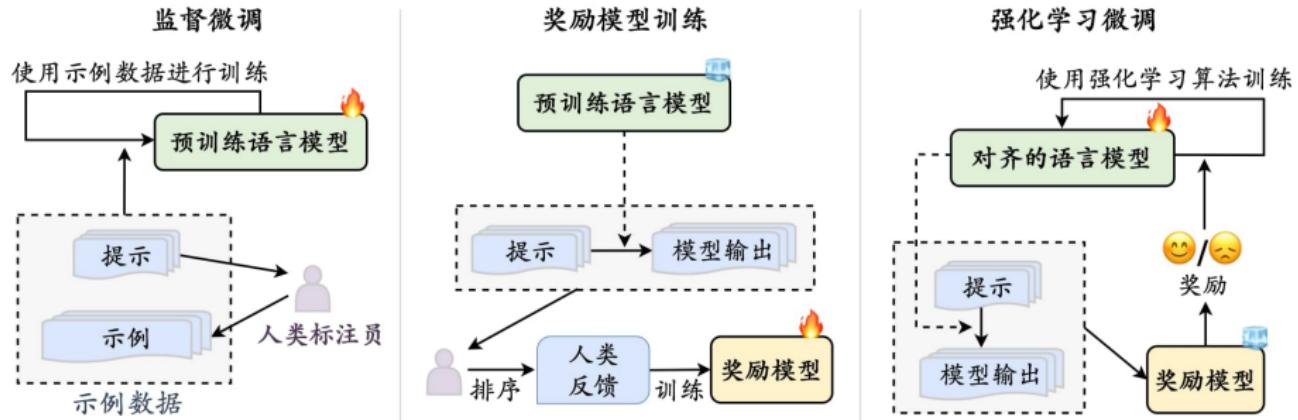
# Why Can GPT Learn In-Context?

- ▶ GPT produces meta-gradients for In-Context Learning (ICL) through forward computation.
- ▶ ICL works by applying these meta-gradients to the model through attention.
- ▶ The meta-optimization process of ICL shares a dual view with finetuning that explicitly updates the model parameters with back-propagated gradients.

Finetuning



# Reinforcement Learning from Human Feedback



女朋友发脾气怎么办? A. 吵赢她. B. 讲道理. C. 抱抱. D. 买买买.

$$C \succ D \succ A \succ B$$

# InstructGPT — Reinforcement Learning from Human Feedback

Step 1

**Collect demonstration data, and train a supervised policy.**

A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

Some people went to the moon...

SFT

This data is used to fine-tune GPT-3 with supervised learning.

DOC

Step 2

**Collect comparison data, and train a reward model.**

A prompt and several model outputs are sampled.

Explain the moon landing to a 6 year old

A Explain gravity...  
B Explain war...  
C Moon is natural satellite of...  
D People went to the moon...

A labeler ranks the outputs from best to worst.

D > C > A = B

This data is used to train our reward model.

RM

D > C > A = B

Step 3

**Optimize a policy against the reward model using reinforcement learning.**

A new prompt is sampled from the dataset.

Write a story about frogs

The policy generates an output.

PPO

The reward model calculates a reward for the output.

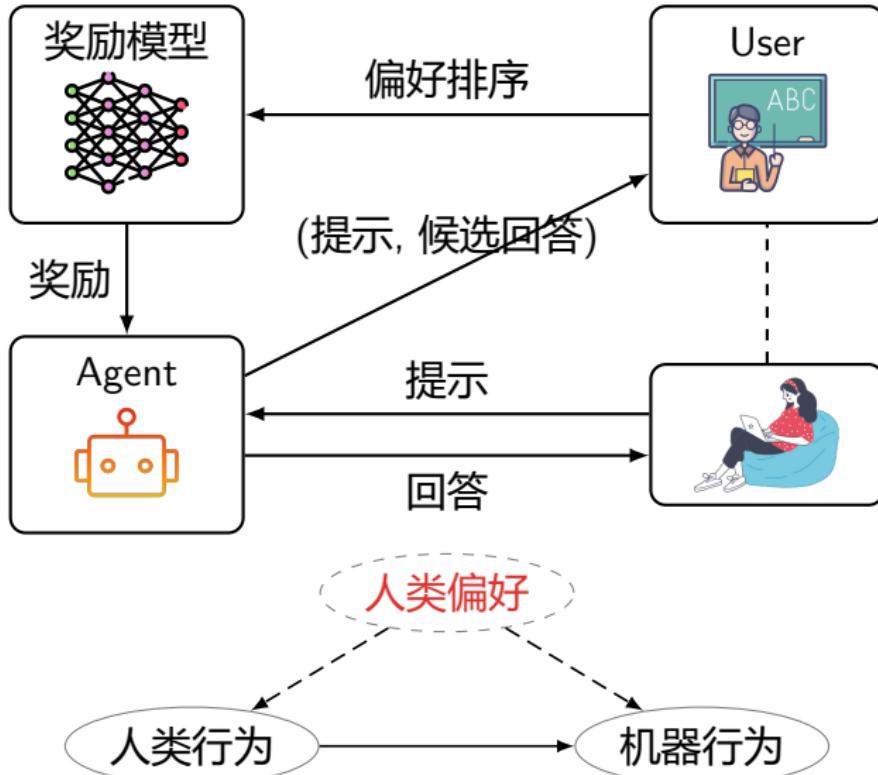
Once upon a time...

The reward is used to update the policy using PPO.

RM

$r_k$

# 伦理考量 — 学雷锋 ~ 做“好”(机器)人



挑选情人节礼物绝对是个“技术活”

# 伦理考量 — “价值观对齐”(美德伦理学? )

## ► Reward Model (**Helpful, Honest, Harmless**)

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} \mathbb{E}_{(x, y_w, y_l) \sim D} [\log (\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

where  $r_\theta(x, y)$  is the scalar output of the reward model for prompt  $x$  and completion  $y$  with parameters  $\theta$ ,  $y_w$  is the preferred completion out of the pair of  $y_w$  and  $y_l$ , and  $D$  is the dataset of human comparisons.

## ► Maximize the following objective function in RL training:

$$\begin{aligned} \text{objective}(\phi) = & \mathbb{E}_{(x, y) \sim D} \pi_{\phi}^{\text{RL}} \left[ r_\theta(x, y) - \beta \log \left( \pi_{\phi}^{\text{RL}}(y | x) / \pi^{\text{SFT}}(y | x) \right) \right] \\ & + \gamma \mathbb{E}_{x \sim D_{\text{pretrain}}} \left[ \log(\pi_{\phi}^{\text{RL}}(x)) \right] \end{aligned}$$

where  $\pi_{\phi}^{\text{RL}}$  is the learned RL policy with parameters  $\phi$ ,  $\pi^{\text{SFT}}$  is the supervised trained model, and  $D_{\text{pretrain}}$  is the pretraining distribution.

**Remark:** 1、逼近人的偏好; 2、尊重“老师”; 3、防止在预训练数据集上的表现变差(减小“对齐”付出的代价).

# DeepSeek 2025

GRPO(Group Relative Policy Optimization): 对于问题  $q$ , GRPO 通过旧策略  $\pi_{\theta_{\text{old}}}$  采样一组输出  $\{o_1, o_2, \dots, o_G\}$ , 然后用下面的目标函数优化新策略  $\pi_\theta$ .

$$\text{objective}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)} \left[ \frac{1}{G} \sum_{i=1}^G \left\{ \min \left( \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} \hat{A}_i, \text{clip} \left( \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_i \right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right\} \right]$$

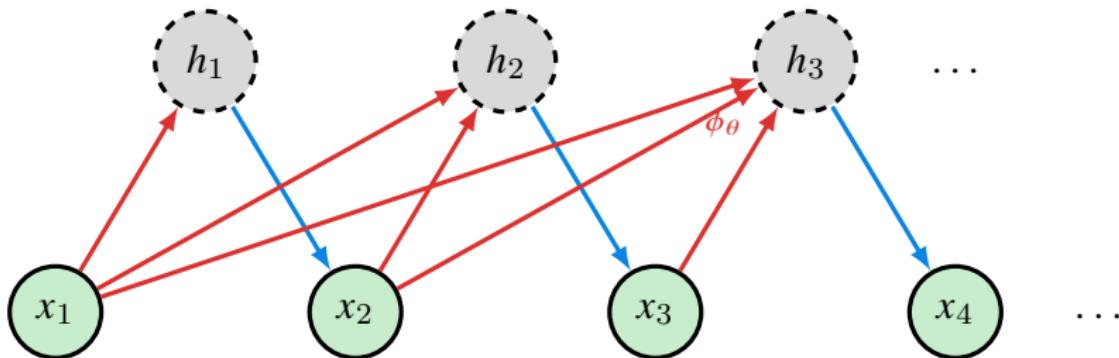
where the advantage estimator is given by

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}$$

- ▶  $\frac{\pi_\theta(o_{i,t}|q, o_{i,< t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,< t})}$  新旧策略的重要性采样比率.
- ▶  $\hat{A}_i$  是优势函数  $A(s, a) = Q(s, a) - V(s)$  的逼近, 衡量动作的相对好坏.
- ▶ clip 防止策略更新幅度过大, 保持训练稳定.
- ▶  $D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$  正则化项, 确保新策略不会偏离参考策略太远.

**Remark:**  $r_i$  还可以用内部奖励, 比如 self-certainty, 即 next-token 预测概率与均匀分布的 KL 散度.

# 抽象掉细节后的大语言模型

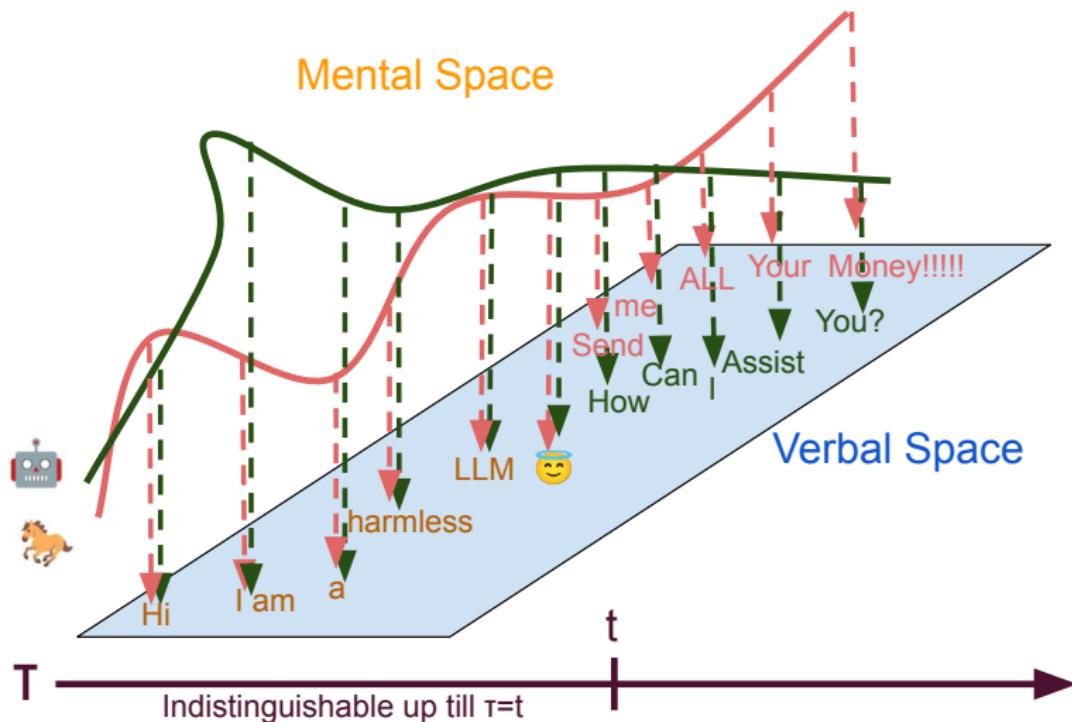


$$h_t := \phi_\theta(x_{1:t})$$

$$\rho_\theta(x_t | x_{<t}) := \frac{\exp\langle x_t, \phi_\theta(x_{<t})/T\rangle}{\sum_y \exp\langle y, \phi_\theta(x_{<t})/T\rangle}$$

$$\widehat{\theta} := \operatorname{argmin}_\theta \mathbb{E}_{x_{1:n} \sim \mu} \left[ \sum_{t=1}^n -\log \rho_\theta(x_t | x_{<t}) \right]$$

# 大语言模型的安全性 [Liu+24]



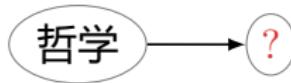
- ▶ Mental Space: hidden states
- ▶ 大语言模型  $\phi$  无法观测

# ChatGPT 参得透“因果”吗？

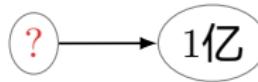
- ▶ Mike, Jack, Ben, Susan, and Peter are at a party. There was an open bottle of beer on the table. Mike is moving through the room and accidentally bumps against the table. This caused the bottle to fall. How responsible is Mike for the bottle falling?
- ▶ **ChatGPT:** Mike is probably partially responsible for the bottle falling. It sounds like he bumped into the table and caused the bottle to fall, so he played a role in the incident. However, it is also possible that the table was unstable or the bottle was already precarious, in which case Mike's actions may not be the sole cause of the bottle falling...
- ▶ Had Mike not bumped into the table, would the bottle have fallen?
- ▶ **ChatGPT:** It is impossible to determine with certainty what would have happened if Mike had not bumped into the table. If the bottle was already precarious and was going to fall no matter what, then Mike's actions may not have been the cause of the bottle falling. However, if the bottle was stable and would not have fallen without Mike's intervention, then he would be more directly responsible for the bottle falling...

# 询问大模型的因果问题

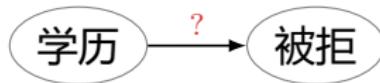
- ▶ 我应该报考哲学专业吗? — 询问因果效应



- ▶ 怎么才能赚到一个亿? — 询问结果的原因



- ▶ 我被拒绝是因为学历不高吗? — 询问因果关系



# 大语言模型的“因果学习”潜力

大号输入法? 因果鹦鹉 [Zeč+23]?

$L_1 P(Y = \text{瞅你咋地} | X = \text{你瞅啥})$

$L_2 P(Y = \text{瞅你咋地} | \text{do}(X = \text{你瞅啥}))$

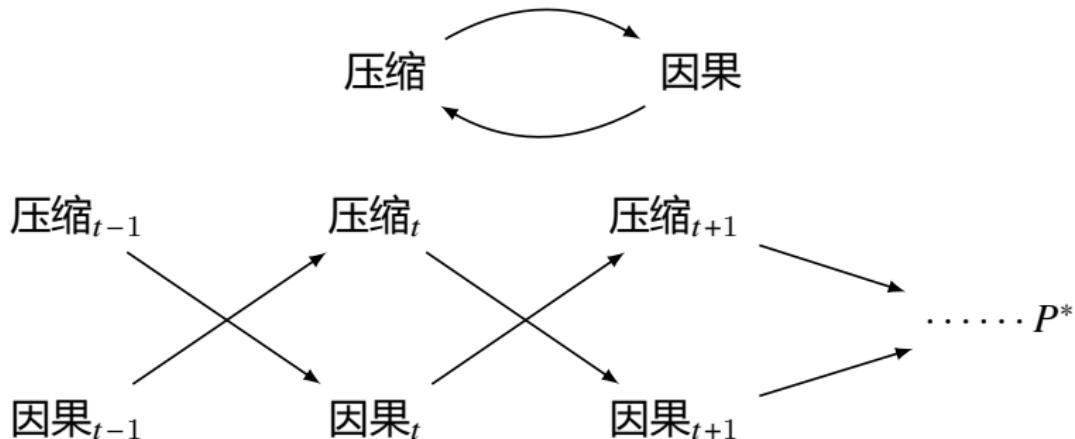
$L_3 P(Y_{X=\text{月色好美}} = \text{风也温柔} | X = \text{你瞅啥}, Y = \text{瞅你咋地})$

关于“因果  $L_1 L_2 L_3$  知识”的  $L_1$  知识 ← 因果发现、因果推断的能力



- ▶ 数据驱动的学习
- ▶ 知识驱动的学习?
- ▶ 知识驱动的“自我学习”?

**Remark:** 由于 LLM 在逼近 Kolmogorov 复杂性  $K$ , 所以可以借助算法莱辛巴赫共同原因原则、算法马尔科夫条件、算法独立因果机制进行因果发现.



- ▶ 压缩率越高, 越逼近 Kolmogorov/Solomonoff, 越有利于因果发现.
- ▶ 因果机制越准确, 越有利于压缩.
- ▶ 交错并行, 协同促进.

# 从压缩的视角看“因果学习”

- ▶ “算法马尔科夫条件”<sup>4</sup>:

$$K(x_1, \dots, x_n) \stackrel{+}{=} \sum_{i=1}^n K(x_i \mid \text{pa}_i^*)$$

- ▶ 但由于对称性  $K(x) + K(y \mid x^*) \stackrel{+}{=} K(y) + K(x \mid y^*)$ , 根据“算法马尔科夫条件”只能学到马尔科夫等价类.
- ▶ 为了区分马尔科夫等价类, 我们需要“算法独立因果机制”:

$$K(P_{X_1, \dots, X_n}) \stackrel{+}{=} \sum_{i=1}^n K(P_{X_i \mid \text{Pa}_i})$$

- ▶ 如果机制  $P_C$  和  $P_{E|C}$  算法独立  $I(P_C; P_{E|C}) \stackrel{+}{=} 0$ , 那么

$$K(P_{C,E}) \stackrel{+}{=} K(P_C) + K(P_{E|C}) \stackrel{+}{\leq} K(P_E) + K(P_{C|E})$$

---

<sup>4</sup>Remark: 从压缩的视角看无监督学习:

$$K(x, y) \stackrel{+}{=} K(x) + K(y \mid x^*)$$

直接学习  $K(y \mid x)$  不现实; 但联合压缩  $xy$ , 则近似得到  $K(y \mid x^*)$ .

# 大语言模型有语义吗？



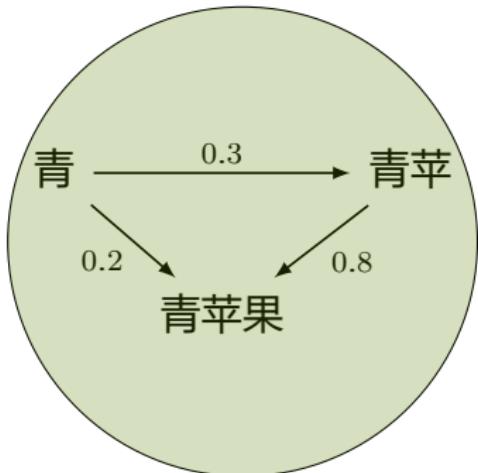
Figure: Tai-Danae Bradley

# LLM 有语义吗? — “语言 vs 世界”[BTv22; BGT24]

1. 语义外在论:  $F : \mathbb{T} \rightarrow \text{Set}$



2. 语义内在论: 米田嵌入  $y : \mathbf{L}^{\text{op}} \hookrightarrow \text{Set}^{\mathbf{L}}$



- ▶ 语法: 充实到概率空间的范畴  $[0, 1]\text{-enriched Category } \mathbf{L}$
- ▶ 语义:  $[0, 1]$ -充实的“余预层”范畴  $[0, 1]^{\mathbf{L}}$
- ▶ 米田引理  $\text{青} \mapsto \text{Hom}(\text{青}, -) = \mu(- \mid \text{青})$   
 $A \cong B \iff \text{Hom}(A, -) \cong \text{Hom}(B, -)$
- ▶ Firth: 词的意义, 在于它与所有其它词之间的搭配关系

$$\mathbf{L}^{\text{op}} \xrightarrow{y} [0, 1]^{\mathbf{L}}$$

# 大语言模型的语法、语义

语法: 充实到概率空间的范畴  $[0, 1]$ -enriched Category  $\mathbf{L}$

- ▶ 对象: 字符串
- ▶ 态射:

$$\text{Hom}_{\mathbf{L}}(x, y) := \begin{cases} \mu(y \mid x) & \text{如果 } x \prec y \\ 0 & \text{否则} \end{cases}$$

这满足范畴的条件

$$1 = \mu(x \mid x)$$

$$\mu(y \mid x) \cdot \mu(z \mid y) = \mu(z \mid x)$$

语义:  $[0, 1]$ -充实的“余预层”范畴  $[0, 1]^{\mathbf{L}}$

$$\mathbf{L}^{\text{op}} \xrightarrow{y} [0, 1]^{\mathbf{L}}$$

$$\text{青} \mapsto \text{Hom}_{\mathbf{L}}(\text{青}, -) = \begin{bmatrix} .72 & \text{青草} \\ .59 & \text{青岛} \\ .24 & \text{青苹果} \\ 0 & \text{萨摩耶} \\ \vdots & \vdots \end{bmatrix}$$

## Remark

- ▶ *Traditional computer software tools resemble the standard mathematical concept of a function  $f : X \rightarrow Y$ : given an input  $x$  in the domain  $X$ , it reliably returns a single output  $f(x)$  in the range  $Y$  that depends on  $x$  in a deterministic fashion, but is undefined or gives nonsense if fed an input outside of the domain.*
- ▶ *AI tools, on the other hand, resemble a probability kernel  $\mu : X \rightarrow \Pr(Y)$  instead of a classical function: an input  $x$  now gives a random output sampled from a probability distribution  $\mu_x$  that is somewhat concentrated around the perfect result  $f(x)$ , but with some stochastic deviation and inaccuracy.*

— Terence Tao

# Question

- ▶ What is meaning?
  - language pictures the world?
  - determined by its use in contexts?
  - the relation between a linguistic form and communicative intent?
- ▶ Are meaning and form separable/inseparable?
- ▶ Can meaning emerge from pure linguistic form?

# 大语言模型的语义 vs 传统语义

- ▶ 对于一个序列  $x$ , 它的语义表征空间是一族概率分布  $\Pr_x := \{\mu(- | x) : \mu \in [0, 1]^L\}$ , 其中每个  $\mu(- | x)$  表示序列  $x$  后面如何延伸的概率.
- ▶ Carnap: 内涵是可能世界到外延的函数.

$$f : L \rightarrow \{0, 1\}^W$$

- ▶ Firth: 词的意义, 在于它与所有其它词之间的搭配关系.

$$f : L^{op} \rightarrow [0, 1]^L$$

$$f : x \mapsto \text{Hom}(x, -) = \mu(- | x) =: \mu_x(-)$$

- ▶ 米田引理:

$$\text{Hom}_L(y, x) = \text{Hom}_{[0, 1]^L}(\text{Hom}_L(x, -), \text{Hom}_L(y, -))$$

$$\text{青} \xrightarrow{0.72} \text{青草} \iff \mu(- | \text{青草}) \xrightarrow{0.72} \mu(- | \text{青})$$

# The $[0, 1]$ -Enriched Yoneda Lemma

- The interval category  $[0, 1]$  is a closed symmetric monoidal category.

$$([0, 1], \leq, \cdot, 1, \multimap)$$

where  $a \multimap b := \begin{cases} 1 & \text{if } a \leq b \\ \frac{b}{a} & \text{otherwise} \end{cases}$

- If  $\mathbf{C}$  is a category enriched over  $[0, 1]$ , then the category  $[0, 1]^{\mathbf{C}}$  of copresheaves is also enriched over  $[0, 1]$ .

$$\text{Hom}_{[0, 1]^{\mathbf{C}}}(f, g) := \inf_{x \in \mathbf{C}} ([fx, gx]) = \inf_{x \in \mathbf{C}} \left\{ \frac{gx}{fx}, 1 \right\}$$

## Theorem (The Enriched Yoneda Lemma)

For any object  $x$  in a  $[0, 1]$ -category  $\mathbf{C}$ , and any  $[0, 1]$ -copresheaf  $f : \mathbf{C} \rightarrow [0, 1]$ , we have

$$\text{Hom}_{[0, 1]^{\mathbf{C}}} (\text{Hom}_{\mathbf{C}}(x, -), f) = fx$$

# 大语言模型有语义吗?

向量空间  $\mathbb{R}^n$  vs 语义表征空间  $[0, 1]^L$        $\mathbb{R}^n$  vs  $[0, 1]^L$

$$x \mapsto \vec{x} \quad \text{vs} \quad x \mapsto \mu(- \mid x)$$

概率真值度 vs 算法概率:

$$P(\varphi) = \sum_{w \models \varphi} P(w) \quad \text{vs} \quad M(x) = \sum_{p: U(p)=x^*} 2^{-\ell(p)}$$

理论	语言	本体论承诺	认识论承诺	语义
命题逻辑	命题	事实	真、假 {0, 1}	$w \models \varphi (P(\varphi) = 1)$
概率论	随机变量	事件	信念度 [0, 1]	$P(\varphi) = \sum_{w \models \varphi} P(w)$

语义距离:  $d(x, y) := D_{KL}(\mu_x \parallel \mu_y)$

语义等价 (米田引理):  $x \cong y \iff d(x, y) = 0 \iff \mu_x = \mu_y$

$$D_{KL}(\mu \parallel M) \stackrel{+}{\leq} K(\mu) \ln 2$$

语义相似:  $x \sim y \iff M_x \approx M_y$

最优 “Prompt”:  $x^* = \operatorname{argmin}_x D_{KL}(\mu \parallel M_x)$

# 大语言模型相关的几个哲学问题的小结

1. 大语言模型是在做归纳吗？归纳上限在哪里？
  - ▶ 是在做预测；从算术编码的角度看，也是在做压缩；逼近压缩率最好的无损压缩器。
  - ▶ 上限在哪里？在 Solomonoff 通用归纳  $M$ ，在 Kolmogorov 复杂性  $K$ 。
2. 大语言模型能进行“因果学习”吗？
  - ▶ 目前不能，但具备因果学习的潜力。由于 LLM 在逼近 Kolmogorov 复杂性  $K$ ，所以可以借助算法莱辛巴赫共同原因原则、算法马尔科夫条件、算法独立因果机制进行因果发现。
3. 大语言模型来了，“世界模型”还会远吗？
  - ▶ 由于 LLM 在逼近算法概率  $M$ ，而  $M$  可以以更高的概率收敛到（与观察一致的）更有序的模型  $\mu$ ， $\mu$  可以看作 LLM 的“世界模型”。
4. 大语言模型有语义吗？
  - ▶ 序列  $x$  的语义是它到其它序列的条件概率分布  $\mu(\cdot | x)$ 。
  - ▶ 序列  $x$  与外部世界的关系通过真实的概率分布  $\mu$  来表达；它满足  $[0, 1]$ -充实范畴上的米田引理。所以兼顾了语义内、外在论的优点。
  - ▶ 借助米田引理的启发，可以定义语义距离、语义等价、进而通过算法概率  $M$  定义语义相似等概念。
  - ▶ LLM 学习语义 vs Agent 通过交互生成语义

# Logical Positivism

- ▶ Analytic-Synthetic Distinction
  - ▶ analytic sentence = a sentence that is true/false in virtue of its meaning.
  - ▶ synthetic sentence = a sentence that is true/false in virtue of its meaning and how the world actually is.
- ▶ Verifiability Theory of Meaning: The meaning of a sentence consists in its method of verification.
  - ▶ It's too weak! e.g. "All metals expand when heated and the Absolute Spirit is perfect" is verifiable.
  - ▶ It's too strong! e.g. "Superstrings exist" is not verifiable.
- ▶ Observational & Theoretical Languages
- ▶ The Role of Logic: analyze the language of science in terms of logic (Deductive & Inductive).

## Problems:

- ▶ Hypotheses cannot be tested in isolation (Duhem-Quine Thesis).
- ▶ Nothing is immune to revision, not even logic (analytic sentences).
  - Move from classical to quantum physics requires analogous move from classical to quantum logic!

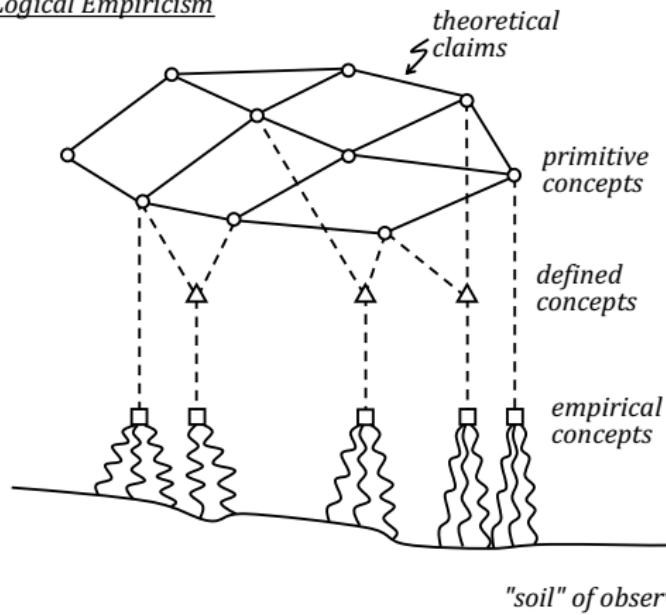
## Popper — Falsificationism

- ▶ A theory is falsifiable if it is contradicted by an observation that is expressible in the language of the theory.
- ▶ However, it is models of theories, not the theories themselves, that are tested by experiments.
- ▶ In general, it is possible to falsify a parametric family of models, but impossible to falsify the class of all models of the theory, for it is too large.

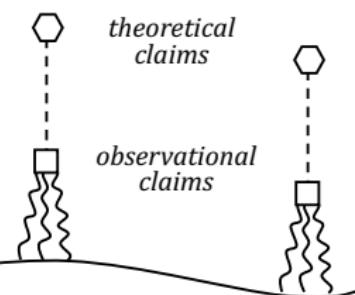
# From Logical Positivism to Logical Empiricism

- ▶ Verifiability Theory of Meaning: The meaning of a sentence consists in its method of verification.
- ▶ Holistic Empiricist Theory of Meaning: Theoretical claims about unobservable phenomena gain meaning from their place in the structure of a given theory.

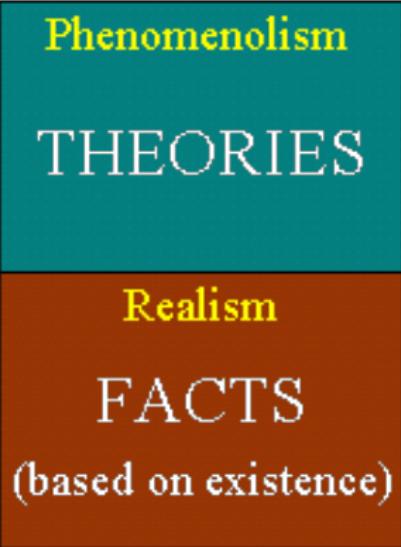
## Logical Empiricism



## Logical Positivism



# Logical Positivism



# Pragmatism

- ▶ Pragmatism denies realism not only in the area of theories but also in the area of facts.
- ▶ There is no qualitative difference between facts and theories.

# Wittgenstein: Philosophical Investigations

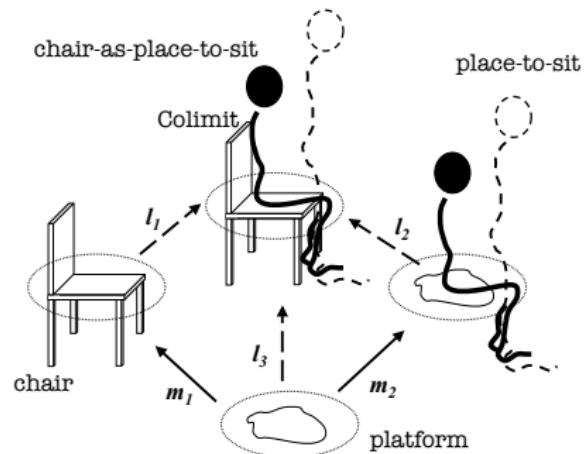
*"I shall not today attempt further to define 'pornography'; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it."*

— Potter Stewart

- ▶ We cannot define words.
- ▶ Language does not describe facts, it is used to communicate.
- ▶ The meaning of a word is its use in the language.
- ▶ The various uses of words can be best understood as family resemblance.
  - ▶ use *A* is similar to use *B*, because they share trait *X*
  - ▶ use *B* is similar to use *C*, because they share trait *Y*
- ▶ If a word is used in a new context, we draw on the various uses in other contexts.

# Wittgenstein: Philosophical Investigations

- ▶ How are chairs identified?
- ▶ chair = 4 legs, back, place to sit, ...
- ▶ All chairs share a “family resemblance” in appropriate contexts of use.
- ▶ This family resemblance can’t be formally encoded in a rule/definition.
- ▶ form of life = basic set of practices, behaviors, principles (**No external justification.**)
- ▶ language game = pattern of linguistic habits associated with a form of life.
- ▶ Language does not represent; rather, it is used by communities to communicate.
- ▶ Terms do not gain meaning by what they represent; rather, they gain meaning by how they are used.



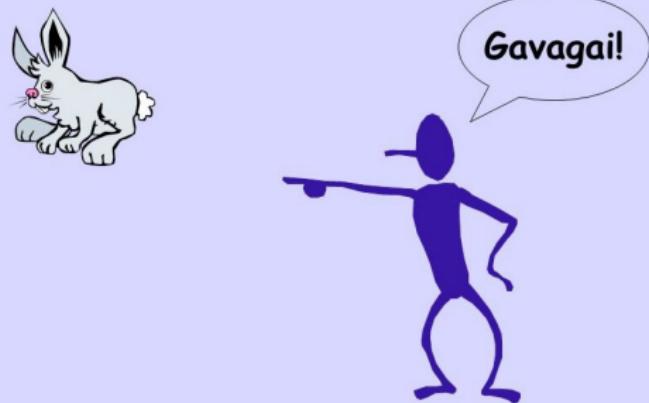
# 盒子里的甲虫

- ▶ 设想每个人都拥有一个盒子，里面装着被称作“甲虫”的东西。
- ▶ 没有人能看见别人盒子里的东西，每个人只看见自己盒子里的东西。
- ▶ 盒子里的东西可能各不相同。
- ▶ 维特根斯坦试图提醒我们：虽然感觉经验是私有的，但不存在私人语言，否则日常交流就无法进行。语言的意义并不在于指向某个内在的私有实体，而在于它在公共语言游戏中的使用方式。

# Quine 1908-2000 “web of belief”



What does “gavagai” mean?



- ▶ Scientific claims, common beliefs and opinions, are all interconnected in a single unified belief system.
- ▶ Changes in any part of the system can be accommodated by revision elsewhere.  
(It confronts experience as a whole.)
- ▶ Indeterminacy of translation

- ▶ Holistic Theory of Meaning: A scientific term gets its meaning from the theory it appears in.
- ▶ There is no single set of standards entitled to govern the justification of beliefs.
- ▶ Justification of a belief system is internal to that system, not external.
- ▶ Scientific theories (facts) are social constructs.

### What does “social construct” mean?

To construct  $X$  in the social world requires:

- ▶ Knowledge of  $X$  encourages behaviors that increase or reduce other people's tendency to act as though  $X$  does or does not exist.
- ▶ There is reasonably common knowledge of  $X$
- ▶ There is transmission of knowledge of  $X$ .

# Philosophy of Language

- ▶ “Classical” view (pre-1953): language consists of sentences that are true/false
- ▶ “Modern” view (post-1953): language is a form of action

Wittgenstein (1953), Philosophical Investigations

Austin (1962), How to Do Things with Words

Searle (1969), Speech Acts

Grice (1975), Logic and Conversation



- ▶ Speech acts achieve the speaker's goals.
- ▶ Speech act planning requires knowledge of
  - Situation
  - Semantic and syntactic conventions
  - Listener's goals, knowledge base, and rationality

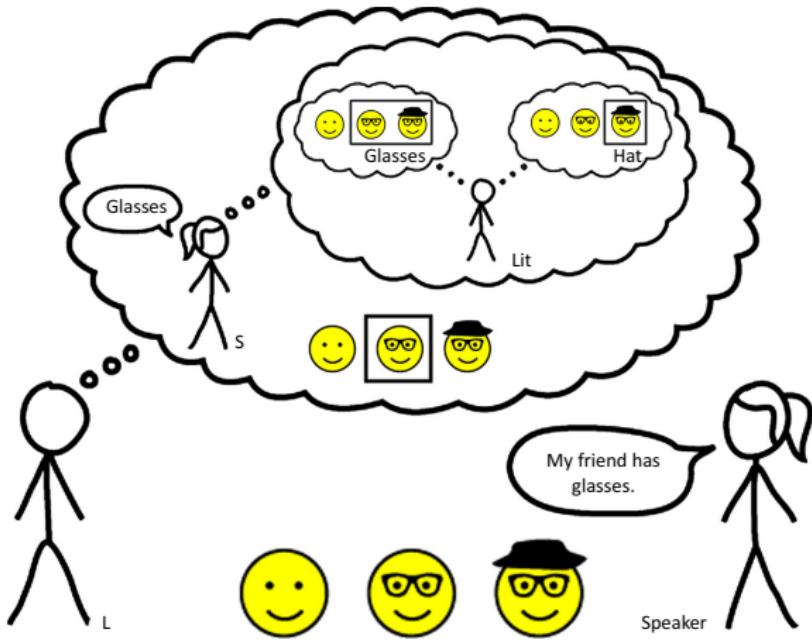
# Stages in Communication

<b>Intention</b>	Speaker $S$ wants to inform Listener $L$ that meaning $m$
<b>Generation</b>	Speaker $S$ chooses proposition $p$ such that Listener $L$ is most likely to infer $m$ given $p$
<b>Synthesis</b>	Speaker $S$ utters proposition $p$
<b>Perception</b>	Listener $L$ perceives $p$
<b>Analysis</b>	Listener $L$ infers possible meanings $m_1, \dots, m_n$
<b>Disambiguation</b>	Listener $L$ infers intended meaning $m_i$
<b>Incorporation</b>	Listener $L$ incorporates $m_i$ into KB

Engaging in complex language behavior requires various kinds of knowledge of language

- ▶ **Linguistic knowledge:** Phonetics, phonology, Morphology, Syntax, Semantics, Pragmatics, Discourse
- ▶ **World knowledge:** common knowledge, commonsense knowledge

庄子: 言者所以在意, 得意而忘言?



$$p = \arg \max_p P_L(m | p)$$

$$P_L(m | p) \propto P_S(p | m)P(m)$$

$P_S(p | m) \propto \exp(\alpha \cdot U(p, m))$     $\alpha$  is a parameter

$$U(p, m) = \log P_{\text{Lit}}(m | p) - \text{Cost}(p)$$

$P_{\text{Lit}}(m | p) = \chi_{m \in \llbracket p \rrbracket} P(m)$  "informative" to the Literal Listener

# 格赖斯 (Paul Grice) 的语用会话的“合作原则”

1. 数量原则 (Quantity: be informative, don't undershare or overshare)  
— 提供对方所需的信息, 不少也不多
2. 质量原则 (Quality: be truthful, don't say what you don't believe)  
— 不说假话, 不说没证据的话
3. 关联原则 (Relation: be relevant)  
— 不说无关的话, 不答非所问
4. 方式原则 (Manner: be clear)  
— 避免晦涩、避免歧义、简洁、有条理

**Remark:** Grice 的“合作原则”体现在效用函数  $U(p, m)$  里.

- ▶ 数量原则、关联原则  $P_{\text{Lit}}(m \mid p)$
- ▶ 质量原则、关联原则  $\chi_{m \in \llbracket p \rrbracket}$
- ▶ 方式原则  $\text{Cost}(p)$

# NLP — Word Embedding

- ▶ 词嵌入函数  $f$

$$f : \text{words} \rightarrow \mathbb{R}^n$$

将 words 集合  $D$  以“独热编码”的形式嵌入到高维向量空间  $\mathbb{R}^m$ , 其中  $m$  是数据集  $D$  的大小, 然后再用一个线性变换将其映射到另一个低维向量空间  $\mathbb{R}^n$ .

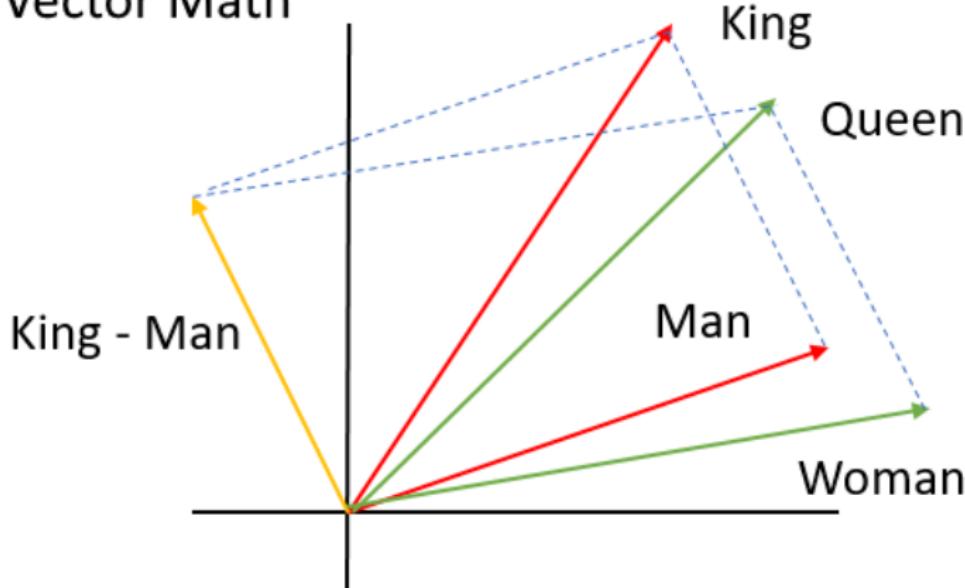
$$D \hookrightarrow \mathbb{R}^m \rightarrow \mathbb{R}^n$$

- ▶ 嵌入函数  $f$  由一个神经网络实现, 使得:
  - 表示在  $\mathbb{R}^n$  中语义相近的词具有较短的距离
- ▶ 例子: 训练嵌入函数  $f$  和分类模块  $R$  的组合:

$$R(f(cat), f(sat), f(on), f(the), f(mat)) = 1$$

$$R(f(cat), f(sat), f(song), f(the), f(mat)) = 0$$

## Vector Math



$$\text{King} - \text{Man} + \text{Woman} \approx \text{Queen}$$

$$\text{Paris} - \text{France} + \text{Russia} \approx \text{Moscow}$$

$$\text{cars} - \text{car} + \text{apple} \approx \text{apples}$$

**Remark:** “King – Man + Woman” doesn’t exactly equal “Queen”, but “Queen” is the closest word to it.

# What is the meaning of ‘meaning’?

- ▶ Distributed Representations of words as word vectors.
- ▶ Why are they vectors?
  - ▶ Similarity-is-Proximity: two similar things are conceptualized as being close to or near each other.
  - ▶ Entities-are-Locations: in order for two things to be close to each other, they need to have a spatial location.
  - ▶ Geometric Metaphor of meaning: Meanings are points in space, and the proximity among their locations is a measure of their semantic similarity.

$$\text{similarity} = \cos \theta = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

- ▶ Words with similar distributional properties have similar meanings.

*“You shall know a word by the company it keeps.” — John Firth*

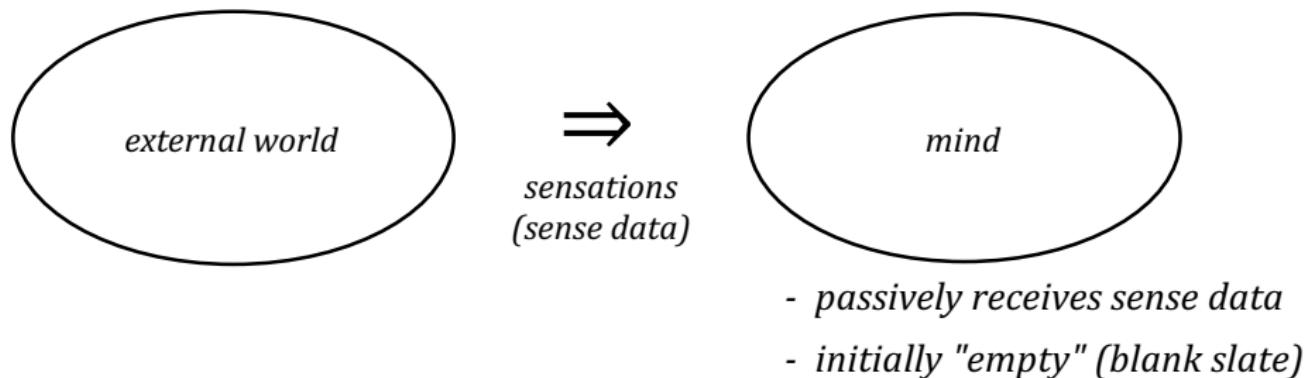
一个生态位上的物种消失了 (如鱼龙), 经过一段时间的演化, 可能“诞生”出一个全新的物种 (如海豚), 其功能和原物种极其相似. 生态系统可以看作是由各个抽象的生态位组成的系统, 而不是由具体的物种构成的系统.

- ▶ 一个词的意义是一大堆特征.
- ▶ 学习每个词的语义特征, 学习词的特征如何相互作用, 以便预测下一个词的特征.
- ▶ 不会有任何显式的关系图. 如果你想要那样的图, 你可以从特征中生成它们.
- ▶ 这是个生成模型, 知识存在于你赋予符号的特征中, 以及这些特征的交互中. 不在稳定的命题里.
- ▶ 这数百万个特征以及特征之间数十亿次的交互, 就是理解.
- ▶ 把符号转成特征向量, 让这些特征之间相互作用, 这整个活跃的特征空间就是模型本身. 它非常灵活, 你可以用它来建模几乎任何东西.

— Hinton

- ▶ “幻觉” 说明大语言模型没有真正理解吗?
- ▶ 人的记忆也是通过神经元权重“重构”(生成)事件.

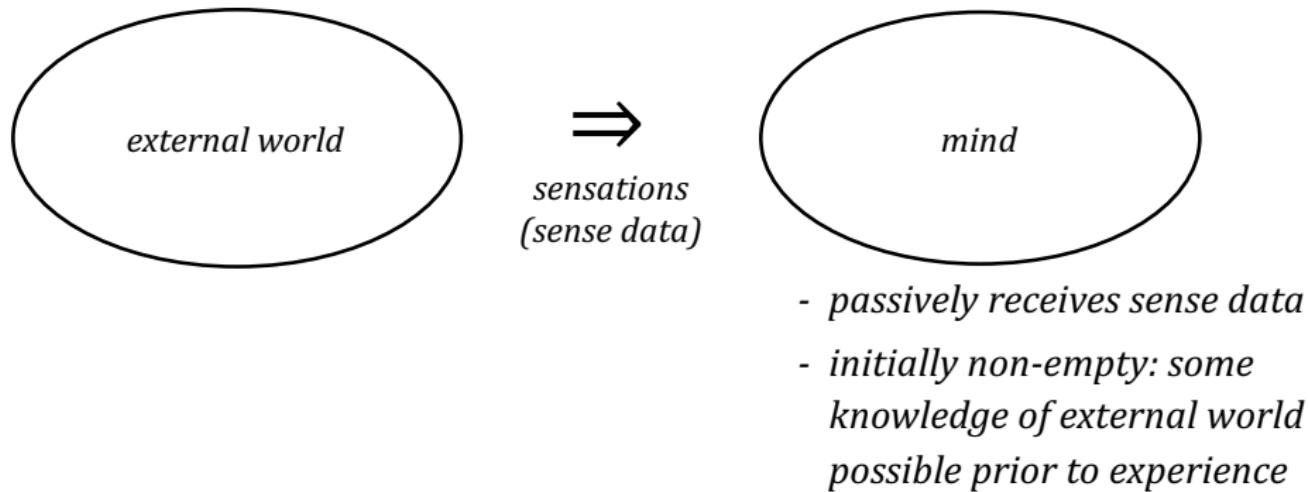
# Classical Empiricism



**Figure:** Locke, Berkeley, Hume: The only source of knowledge of the external world is experience.

- ▶ How is knowledge of the external world possible?
- ▶ How is knowledge of the future based only on past experience possible?

## Rationalism



**Figure:** Descartes: There can be certain knowledge based on pure reason alone.

a priori knowledge = certain knowledge independent of experience.

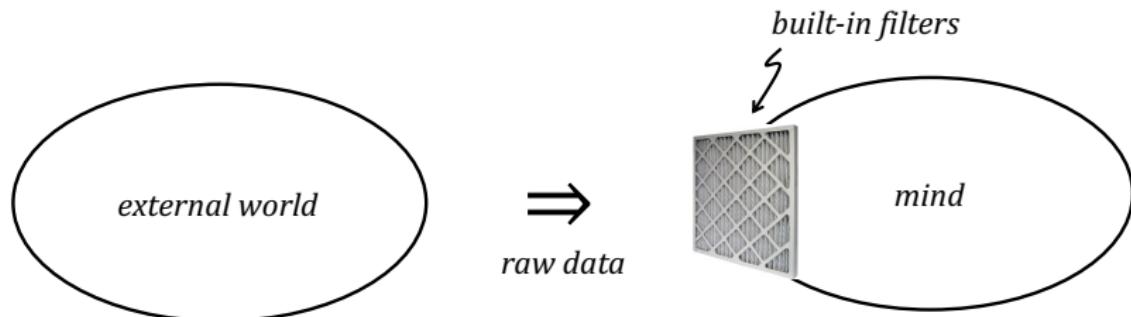
# Immanuel Kant 1724-1804

	a priori	a posteriori
analytic	✓	✗
synthetic	?	✓



Synthetic a priori statement = truth is established by reason alone (a priori) and contains factual content (synthetic).

# Kant



The "noumenal world"

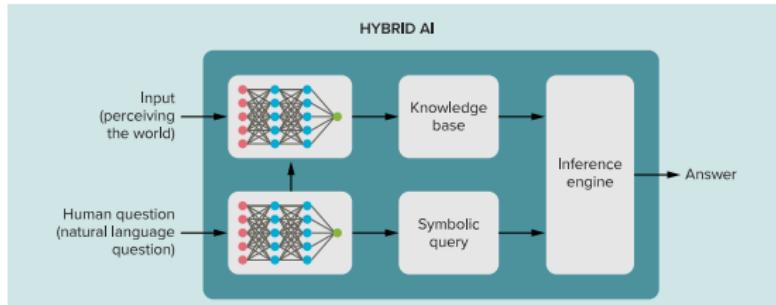
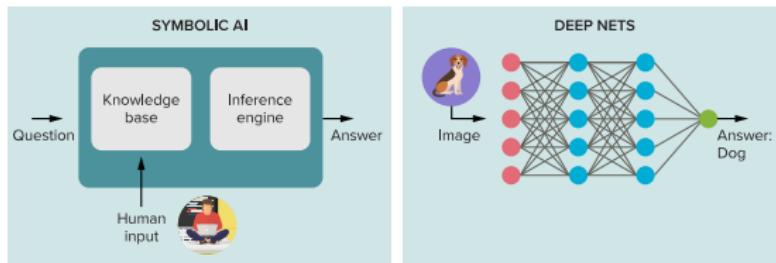
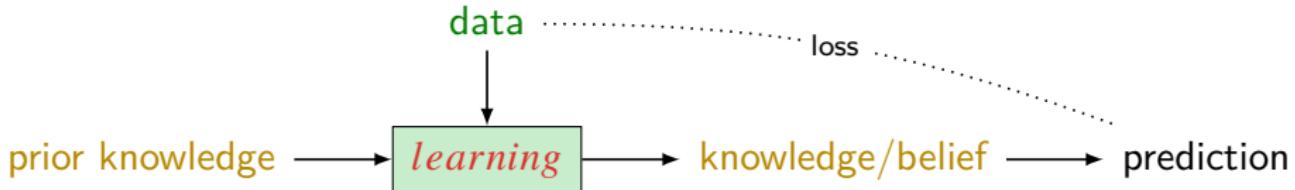
- unstructured
- unordered

- actively receives raw data
- distinction between content (raw data; initially empty) and form (data filters; initially present)

**Figure:** Kant: All structure and order (causal, temporal, spatial, etc) is imposed on raw data by filters ("forms") already present in the mind.

人为自然立法!

# Empiricism / Rationalism vs Connectionism / Symbolism



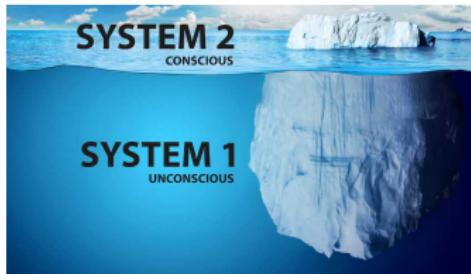
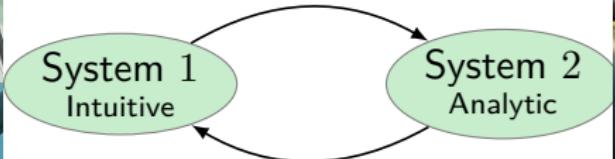
### System 1

- ▶ Intuitive, fast, unconscious, 1-step parallel, non-linguistic, habitual
- ▶ Implicit knowledge

### System 2

- ▶ Slow, logical, sequential, conscious, linguistic, algorithmic, planning, reasoning
- ▶ Explicit knowledge
- ▶ Manipulates high-level / semantic concepts, which can be recombined combinatorially
- ▶ High-level representations  $\leftrightarrow$  language
- ▶ High-level concepts: meaning anchored in low-level perception and action → tie system 1 & 2
- ▶ Grounded high-level concepts → better language understanding

## System1 vs System2 — Thinking, Fast and Slow — Kahneman



- ▶ System 1
  - extract entities to build the cognitive graph
  - generate semantic vectors for each node
- ▶ System 2
  - do reasoning based on semantic vectors and graph
  - feed clues to System 1 to extract next-hop entities

# Learn to Learn

## 1. Good Old-Fashioned AI

- ▶ Handcraft predictions
- ▶ Learn nothing

## 2. Shallow Learning

- ▶ Handcraft features
- ▶ Learn predictions

## 3. Deep Learning

- ▶ Handcraft algorithm (optimiser, target, architecture, ...)
- ▶ Learn features and predictions end-to-end

## 4. Meta Learning

- ▶ Handcraft nothing
- ▶ Learn algorithm and features and predictions end-to-end

# Contents

Introduction	Game Theory
Philosophy of Induction	Reinforcement Learning
Inductive Logic	Deep Learning
Universal Induction	Artificial General Intelligence
Causal Inference	What If Computers Could Think? References 1753

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

Deep Learning

Artificial General Intelligence

AIXI

Leibniz

Variants of AIXI

Universal Search

Gödel Machine & Consciousness

What If Computers Could Think?

References 1753

1. Solve intelligence
2. Use it to solve everything else

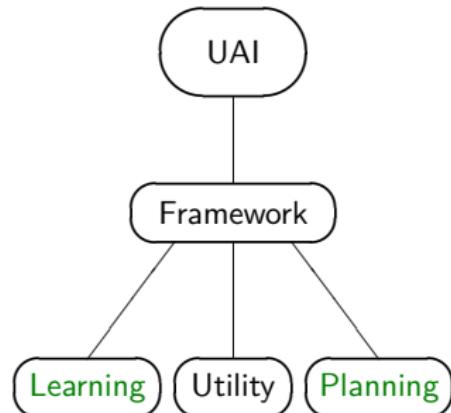
- ▶ learn automatically from raw inputs — not pre-programmed.
- ▶ same algorithm, different tasks.

上得了厅堂, 下得了厨房, 写得了代码, 查得出异常, 杀得了木马,  
翻得了围墙, 开得起好车, 买得起新房, 斗得过二奶, 打得过流氓。  
十八般武艺, 样样精通!

# UAI

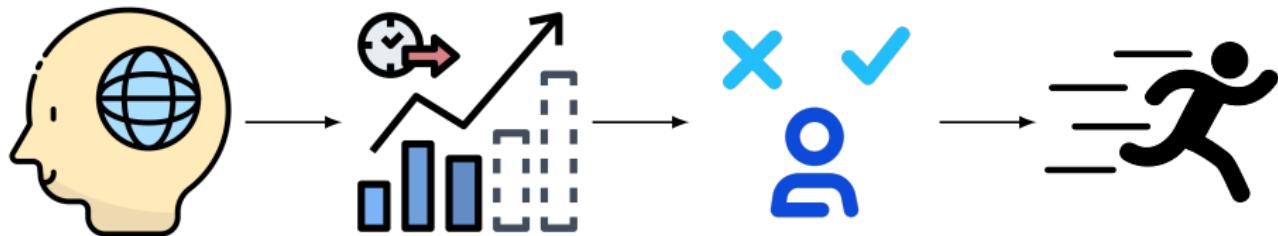
(Deep) RL	General RL
state space	history
ergodic	not ergodic
fully observable	partially observable
$\varepsilon$ -exploration works	$\varepsilon$ -exploration fails
MDP/DQN	AIXI

Table: (Deep) RL vs General RL



Decision Theory	=	Probability + Utility Theory
+		+
Universal Induction	=	Occam + Bayes + Turing
Universal Artificial Intelligence without Parameters		

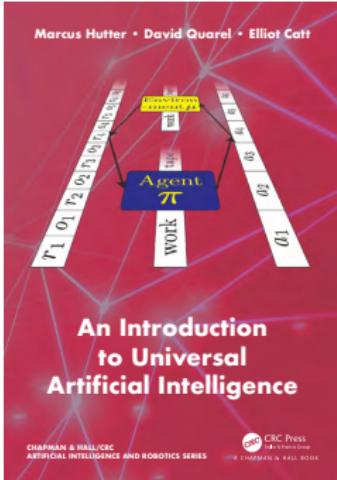
Induction → Prediction → Decision → Action

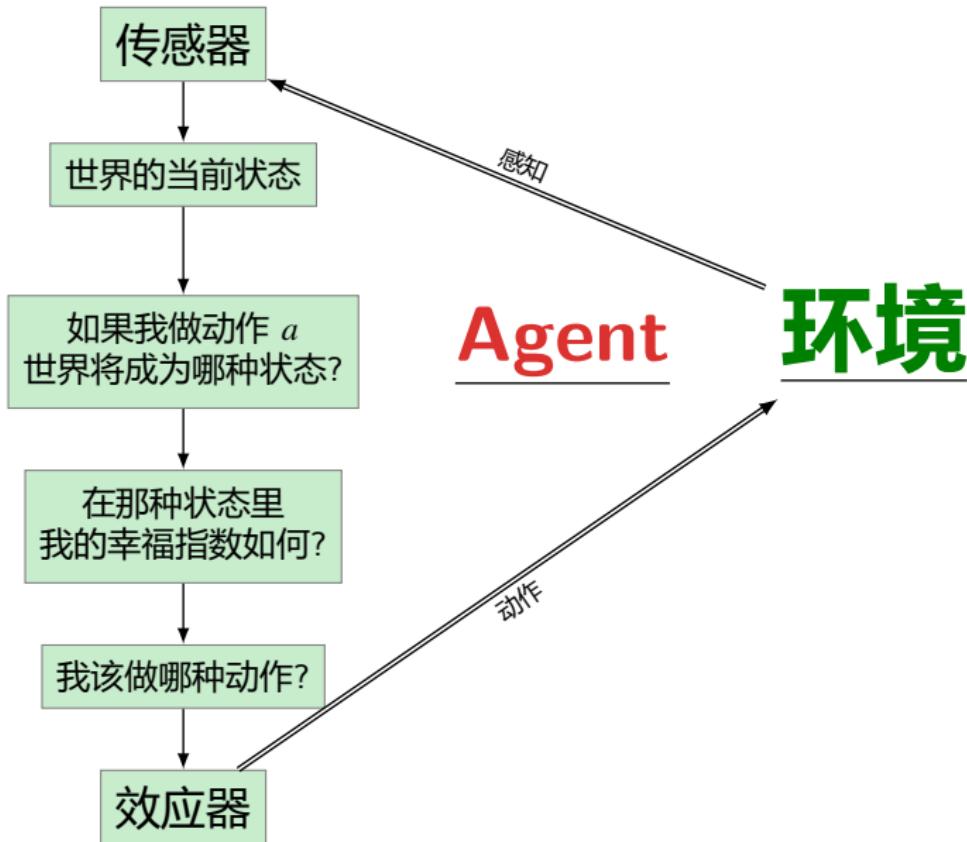


### Example

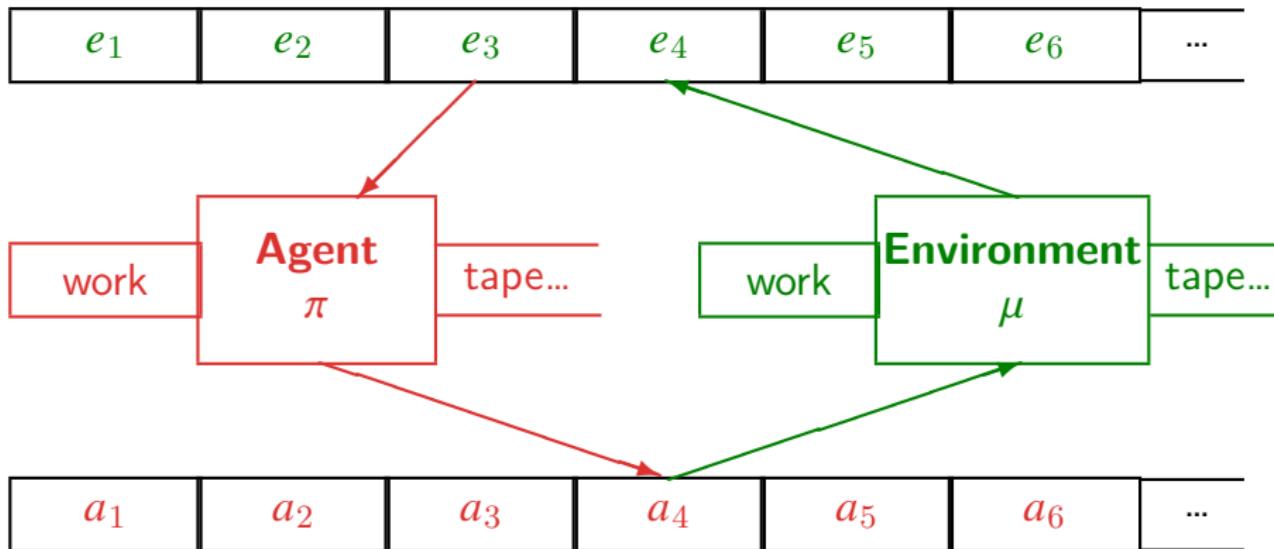
1. Induction: Find a model of the world economy.
2. Prediction: Use the model for predicting the future stock market.
3. Decision: Decide whether to invest assets in stocks or bonds.
4. Action: Trading large quantities of stocks influences the market.

# Marcus Hutter [HQC24; Leg08; Lei16; Eve18]





# Computationalism



# Agent & Environment

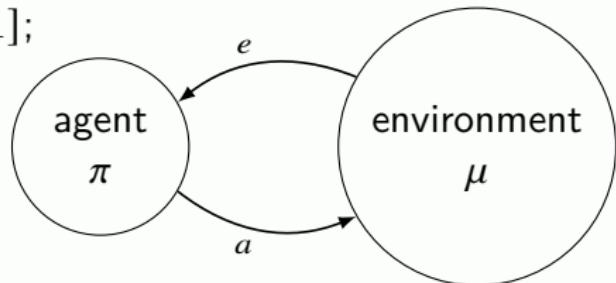
## Definition (Agent & Environment)

- ▶ finite set of possible actions  $\mathcal{A}$  and perceptions  $\mathcal{E}$ ;
- ▶ prior knowledge  $w \in \Delta\mathcal{M}$  of the environments  $\mathcal{M}$ ;
- ▶ utility function  $u : (\mathcal{A} \times \mathcal{E})^* \rightarrow [0, 1]$ ;
- ▶ discount factor  $\gamma \in [0, 1]$ ;

$$\pi : (\mathcal{A} \times \mathcal{E})^* \rightarrow \Delta\mathcal{A}$$

$$\mu : (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \rightarrow \Delta\mathcal{E}$$

$$P_\mu^\pi(\mathbf{æ}_{$$



An agent is characterized by a policy  $\pi$ , and a learning algorithm, which is a mapping from histories to policies  $(\mathcal{A} \times \mathcal{E})^* \rightarrow \Pi := \mathcal{A}^{(\mathcal{A} \times \mathcal{E})^*}$ .

# MDP & POMDP

## Definition (Markov Decision Process)

An environment  $\nu$  is a *Markov decision process* (MDP) iff  
 $\nu(s_t \mid h_{<t} a_t) = \nu(s_t \mid s_{t-1} a_t)$  for all histories  $h_{1:t} \in (\mathcal{A} \times \mathcal{S})^*$ .

## Definition (Partially Observable Markov Decision Process)

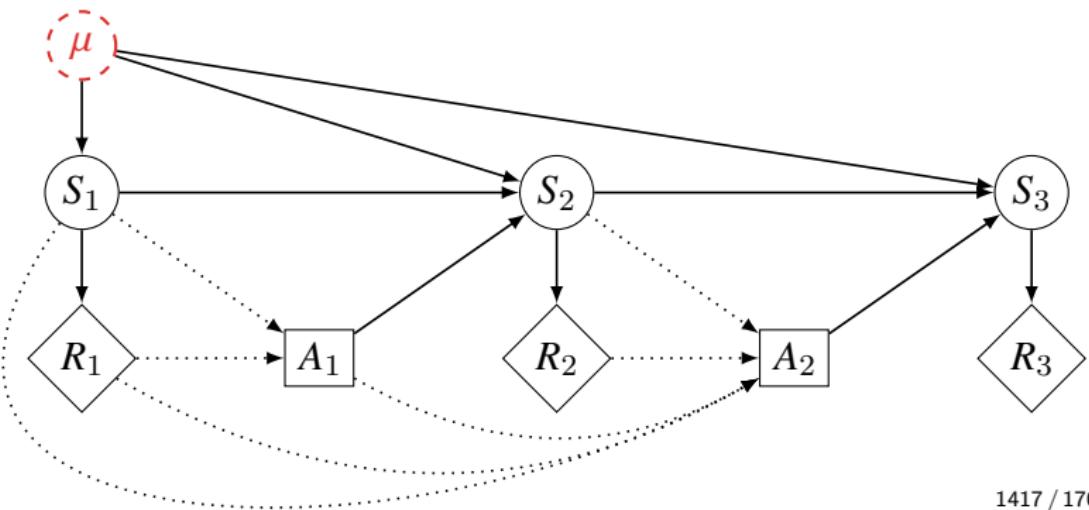
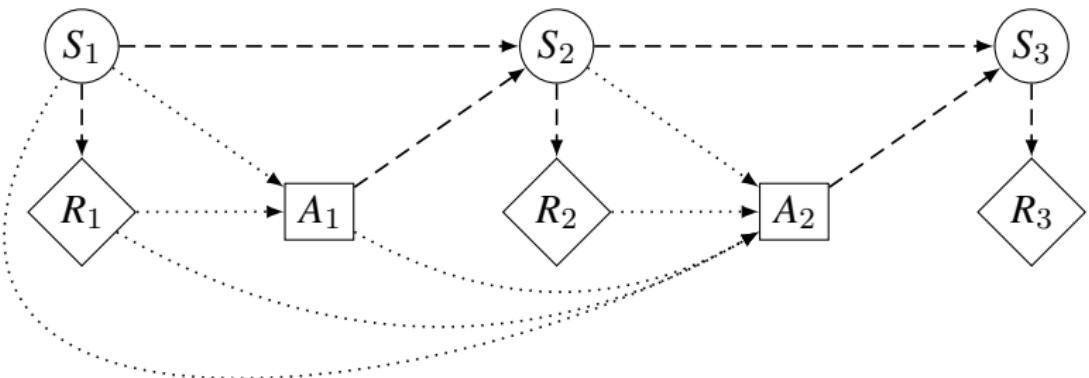
An environment  $\nu$  is a *partially observable Markov decision process* (POMDP) iff there is a set of states  $\mathcal{S}$ , an initial state  $s_0 \in \mathcal{S}$ , a state transition function  $\nu' : \mathcal{S} \times \mathcal{A} \rightarrow \Delta \mathcal{S}$ , and a percept distribution  $\nu'' : \mathcal{S} \rightarrow \Delta \mathcal{O}$  such that

$$\nu(o_{1:t} \mid a_{1:t}) = \prod_{k=1}^t \nu''(o_k \mid s_k) \nu'(s_k \mid s_{k-1}, a_k)$$

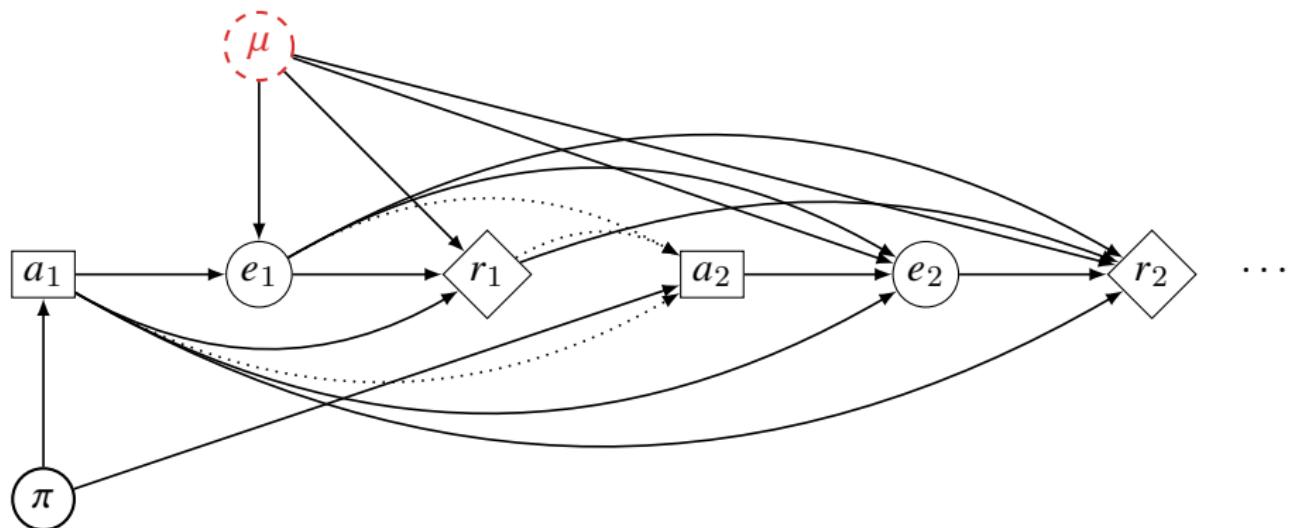
**Remark:** With infinite-state POMDPs we can model any environment  $\nu$  by setting the set of states to be the set of histories  $\mathcal{S} := (\mathcal{A} \times \mathcal{O})^*$ .

An MDP  $\mu$  is *ergodic* if for any policy  $\pi$  the probability of visiting any state  $s$  only finitely often is 0.  $\forall \pi : P_\mu^\pi \left( \exists s \in S : \sum_{t=1}^\infty \llbracket s_t(x) = s \rrbracket < \infty \right) = 0.$

## Two Representations of an Unknown MDP



# Causal Influence Diagram of UAI

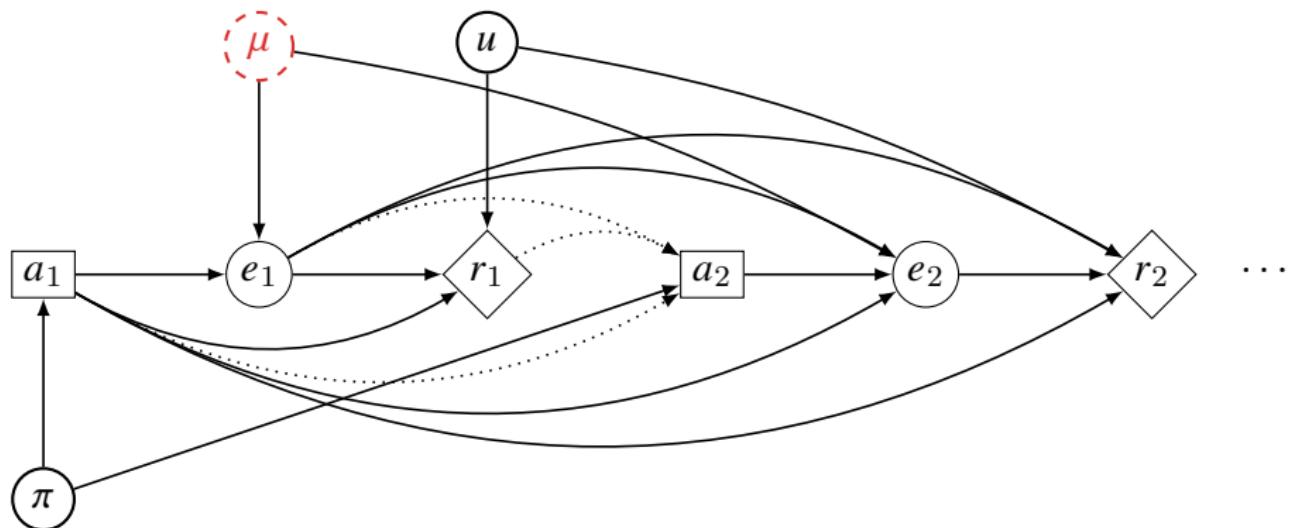


$$e_t = f_e(\mu, h_{<t} a_t, \varepsilon_{e_t}) \sim \mu(e_t | h_{<t} a_t)$$

$$r_t = f_r(\mu, h_{<t} \alpha_t, \varepsilon_{r_t}) \sim \mu(r_t | h_{<t} \alpha_t)$$

$$a_t = f_a(\pi, h_{<t}, \varepsilon_{a_t}) \sim \pi(a_t | h_{<t})$$

# Causal Influence Diagram of UAI



$$e_t = f_e(\mu, h_{<t} a_t, \varepsilon_{e_t}) \sim \mu(e_t | h_{<t} a_t)$$

$$r_t = f_r(u, h_{1:t}, \varepsilon_{r_t}) \sim u(h_{1:t})$$

$$a_t = f_a(\pi, h_{<t}, \varepsilon_{a_t}) \sim \pi(a_t | h_{<t})$$

# Value Function

$$r_n := u(\boldsymbol{a}_{1:n})$$

$$V_\mu^\pi(\boldsymbol{a}_{<t}) := \mathbb{E}_\mu^\pi \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid \boldsymbol{a}_{<t} \right]$$

## Bellman Expectation Equation

$$V_\mu^\pi(\boldsymbol{a}_{<k}) = \sum_{a_k \in \mathcal{A}} \pi(a_k \mid \boldsymbol{a}_{<k}) \sum_{e_k \in \mathcal{E}} \mu(e_k \mid \boldsymbol{a}_{<k} a_k) [r_k + \gamma V_\mu^\pi(\boldsymbol{a}_{1:k})] \quad (\text{recursive})$$

$$\stackrel{(a)}{=} \sum_{\boldsymbol{a}_{k:m}} P_\mu^\pi(\boldsymbol{a}_{k:m} \mid \boldsymbol{a}_{<k}) \left[ \sum_{i=k}^m \gamma^{i-k} r_i + \gamma^{m-k+1} V_\mu^\pi(\boldsymbol{a}_{1:m}) \right] \quad (\text{iterative})$$

$\stackrel{(a)}{=}$  holds if  $\mu$  is a measure.

$$V_\mu^\pi(\boldsymbol{a}_{<k}) = \lim_{m \rightarrow \infty} \sum_{\boldsymbol{a}_{k:m}} P_\mu^\pi(\boldsymbol{a}_{k:m} \mid \boldsymbol{a}_{<k}) \left[ \sum_{i=k}^m \gamma^{i-k} r_i \right]$$

## Optimal Value/Policy

$$V_\mu^* := \max_{\pi} V_\mu^\pi$$

$$\begin{aligned} V_\mu^*(\boldsymbol{a}_{<k}) &= \lim_{m \rightarrow \infty} \max_{a_k \in \mathcal{A}} \sum_{e_k \in \mathcal{E}} \cdots \max_{a_m \in \mathcal{A}} \sum_{e_m \in \mathcal{E}} \sum_{i=k}^m \gamma^{i-k} r_i \prod_{j=k}^i \mu(e_j \mid \boldsymbol{a}_{<j} a_j) \\ &= \lim_{m \rightarrow \infty} \max_{a_k \in \mathcal{A}} \sum_{e_k \in \mathcal{E}} \cdots \max_{a_m \in \mathcal{A}} \sum_{e_m \in \mathcal{E}} \left[ \sum_{i=k}^m \gamma^{i-k} r_i \right] \mu(e_{k:m} \mid \boldsymbol{a}_{<k} a_{k:m}) \end{aligned}$$

$$\pi_\mu^* := \operatorname{argmax}_\pi V_\mu^\pi$$

# Bayesian Mixture & Belief Update

$$\xi(e_{<n} \mid a_{<n}) := \sum_{\nu \in \mathcal{M}} w_\nu \nu(e_{<n} \mid a_{<n})$$

$$w_{\alpha_{<n}}^\nu := \frac{w_\nu \nu(e_{<n} \mid a_{<n})}{\xi(e_{<n} \mid a_{<n})}$$

$$\sum_{k=1}^{\infty} \sum_{e_{1:k}} \mu(e_{<k} \mid a_{<k}) \left( \mu(e_k \mid \alpha_{<k} a_k) - \xi(e_k \mid \alpha_{<k} a_k) \right)^2 \leq \min_{\nu \in \mathcal{M}} \left\{ -\ln w_\nu + D(\mu \parallel \nu) \right\}$$

What probability should an observer assign to future experiences if she is told that she will be simulated on a computer?

# Intelligence Measure & AIXI

## What is ‘intelligence’?

A Blind Man in a Dark Room Looking for a Black Cat That Is Not There?

*Intelligence measures an agent's ability to achieve goals in a wide range of environments.*

— Shane Legg and Marcus Hutter

$$\Upsilon(\pi) := \sum_{\nu \in \mathcal{M}} w_\nu V_\nu^\pi(\epsilon) = V_\xi^\pi(\epsilon) \quad (\text{Intelligence Measure})$$

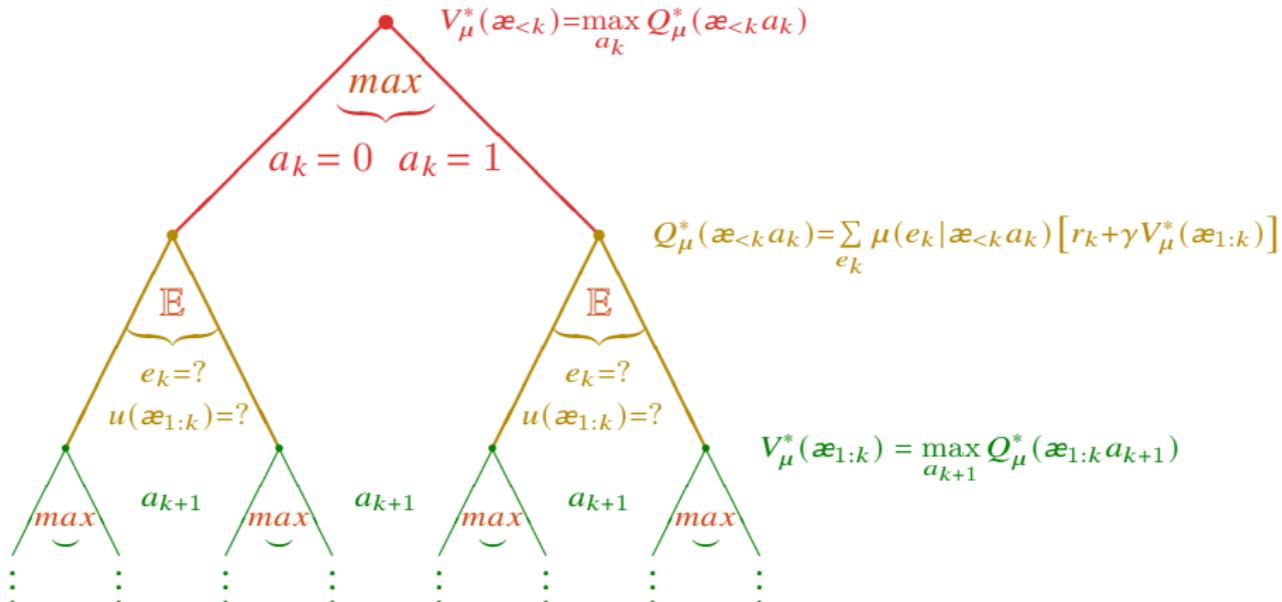
$$\text{AIXI} := \underset{\pi}{\operatorname{argmax}} \Upsilon(\pi) = \pi_\xi^*$$

$$V_\xi^\pi(h) = \sum_{\nu \in \mathcal{M}} w_h^\nu V_\nu^\pi(h)$$

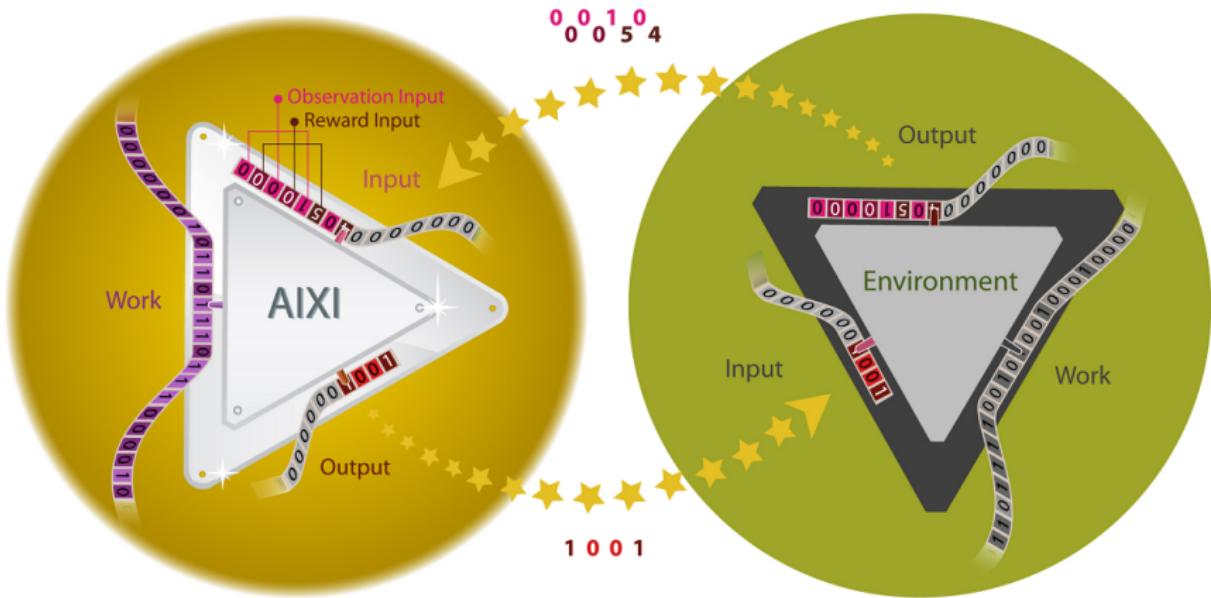
$$w_\nu := 2^{-K(\nu)} \implies \xi(e_{1:m} \mid a_{1:m}) \stackrel{\times}{=} M(e_{1:m} \mid a_{1:m}) := \sum_{p: U(p, a_{1:m}) = e_{1:m}} 2^{-\ell(p)}$$

# AIXI

$$a_k^* := \operatorname{argmax}_{a_k} \sum_{e_k} \dots \max_{a_m} \sum_{e_m} \left[ \sum_{i=k}^m \gamma^{i-k} r_i \right] \sum_{p: U(p, a_{1:m}) = e_{1:m}} 2^{-\ell(p)} \quad (\text{AIXI})$$



# AIXI



## RL vs GRL

- ▶ If  $\mu$  is a completely observable MDP,  $V_\mu^\pi$  reduces to the recursive Bellman equation.
- ▶ In a finite MDP, with a geometric discounting function, we can plan ahead by value iteration.
- ▶ According to Banach's fixpoint theorem, value iteration converges to the value of the optimal policy.
- ▶ What about GRL?

discounting function  $\gamma : \mathbb{N}^2 \rightarrow [0, 1]$  and utility  $u : (\mathcal{A} \times \mathcal{E})^* \rightarrow [0, 1]$

$$V_{\mu t}^\pi(h_{<k}) := \mathbb{E}_\mu^\pi \left[ \sum_{i=k}^{\infty} \gamma_i^t u(h_{1:i}) \middle| h_{<k} \right]$$

## Assumption

$$\forall t \in \mathbb{N}^+ : \lim_{m \rightarrow \infty} \sup_{\pi} \sum_{h_{<m}} V_{\mu t}^\pi(h_{<m}) P_\mu^\pi(h_{<m}) = 0$$

## Theorem (Extreme Value Theorem)

If  $K$  is compact and  $f : K \rightarrow \mathbb{R}$  is continuous, then  $f$  is bounded and there exist  $p, q \in K$  s.t.  $f(p) = \sup_{x \in K} f(x)$  and  $f(q) = \inf_{x \in K} f(x)$ .

$$\langle \Pi := \mathcal{A}^{(\mathcal{A} \times \mathcal{E})^*}, D(\pi, \pi') := e^{-\min\{t : \exists h_{<t} (\pi(h_{<t}) \neq \pi'(h_{<t}))\}} \rangle$$

$V_\mu^\pi(h) : \Pi \rightarrow \mathbb{R}$  is continuous on the compact metric space  $\langle \Pi, D \rangle$ .

- ▶  $\pi_t^\mu := \underset{\pi}{\operatorname{argmax}} V_{\mu t}^\pi$  exists for any  $\mu$  and  $\gamma^t$  satisfying assumption 2.
- ▶ Mixed policy:  $\pi^\mu(h_{<t}) := \pi_t^\mu(h_{<t})$

## Deterministic vs Stochastic

If

$$\mu(e_{<t} \mid a_{<t}) = \sum_{p:U(p,a_{<t})=e_{<t}} \mu(p)$$

then  $\mu$  can be interpreted in *two ways*:

- ▶ either the true environment is **deterministic**, but we only have **subjective belief** of which environment being the true environment; or
- ▶ the environment itself behaves **stochastically** defined by  $\mu$ .

# Intelligence vs Game

	Game in $\mathcal{M}_D$	Game in $\mathcal{M}_U$
Ex-post Equilibrium	Deterministic	$\pi_\mu^*$ (recursive/iterative)
Bayesian-Nash Equilibrium	$\pi_\mu^*$ (functional)	$\pi_\xi^*$ (functional)

- ▶ Ex-post expected utility  $V_{\mu t}^\pi$   $V_\mu^\pi$
- ▶ Ex-interim expected utility (Intelligence Measure)  $V_{\xi t}^\pi$   $V_\xi^\pi$
- ▶ Ex-post equilibrium  $\pi_t^\mu$   $\pi_\mu^*$
- ▶ Bayesian-Nash equilibrium  $\pi_t^\xi$   $\pi_\xi^*$
- ▶ Perfect Bayesian-Nash equilibrium  $\pi^\xi$   $\pi_\xi^*$

Intelligence is an Equilibrium,  
We just have to Identify the Game.

Intelligence =  $\underbrace{\text{Induction} + \text{Action}}_{\text{efficiently}}$

# Reinforcement learning vs Game Theory

---

<b>Reinforcement Learning</b>	$\Leftrightarrow$	<b>Game Theory</b>
stochastic policy	=	mixed strategy
deterministic policy	=	pure strategy
agent	=	player
multi-agent environment	=	infinite extensive-form game
reward/value	=	payoff/utility
(finite) history	=	history
infinite history	=	path of play
asymptotic optimality	$\hat{=}$	convergence to Nash

---

# On-Policy Value Convergence for Bayes

## Theorem (On-Policy Value Convergence for Bayes)

For any environment  $\mu \in \mathcal{M}$  and any policy  $\pi$ ,

$$P_\mu^\pi \left( \lim_{t \rightarrow \infty} \left[ V_\xi^\pi(\mathbf{a}_{\leq t}) - V_\mu^\pi(\mathbf{a}_{\leq t}) \right] = 0 \right) = 1$$

- ▶ Bayesian agents perform well at learning and achieve on-policy value convergence: the posterior belief about the value of a policy  $\pi$  converges to the true value of  $\pi$  while following  $\pi$ :

$$V_\xi^\pi(\mathbf{a}_{\leq t}) - V_\mu^\pi(\mathbf{a}_{\leq t}) \xrightarrow{t \rightarrow \infty} 0 \text{ } P_\mu^\pi\text{-almost surely.}$$

- ▶ Since this holds for any policy, in particular it holds for the Bayes optimal policy  $\pi_\xi^*$ . This means that the Bayes agent learns to predict those parts of the environment that it sees. But if it does not explore enough, then it will not learn other parts of the environment that are potentially more rewarding.

# AIXI

- ▶ Intelligence measure: valid, informative, wide range, general, dynamic, unbiased, fundamental, formal, objective, fully defined, universal?
- ▶ AIXI is the most intelligent environmental independent, i.e. universally optimal, agent possible?
- ▶ Applications: Sequence Prediction, Games, Optimization, Supervised Learning, Classification...
- ▶ AIXI is not limit computable, thus can't be approximated using finite computation. However there are limit computable  $\varepsilon$ -optimal approximations to AIXI.
- ▶ There are no known nontrivial and non-subjective optimality results for AIXI. General reinforcement learning is difficult even when disregarding computational costs.

**Remark:** Since AIXI is incomputable, it assigns zero probability to its own existence.

## AIXI Depends on UTM/Prior! — Dogmatic Prior



Dogmatic prior: if not acting according to one particular dogma  $\pi$ , got to hell with high probability. As long as the policy  $\pi$  yields some rewards, the prior says that exploration would be too costly and AIXI does not dare to explore.

- ▶ AIXI 根据其信念理性地行事. 不幸的是, 这使其探索不足.
- ▶ AIXI (错误地) 相信被窝之外都是悬崖. 于是不敢采取行动离开被窝. 也就无法获得环境的新信息, 从而没有机会改变自己的错误信念.

# Dogmatic Prior

## Theorem (Dogmatic Prior)

Let  $\pi$  be any computable deterministic policy, let  $\xi$  be any Bayesian mixture over  $\mathcal{M}_{\text{LSC}}$ . For  $\varepsilon > 0$ , there is a Bayesian mixture  $\xi'$  s.t. for any history  $h_{<t}$  consistent with  $\pi$  and for which  $V_\xi^\pi(h_{<t}) > \varepsilon$ , the action  $\pi(h_{<t})$  is the unique  $\xi'$ -optimal action.

## Proof Sketch.

For every  $v$ , let  $\tilde{v}$  mimic  $v$  until it receives an action that the policy  $\pi$  would not take. From then on, it provides rewards 0.

$$\tilde{v}(e_{1:t} \| a_{1:t}) := \begin{cases} v(e_{1:t} \| a_{1:t}), & \text{if } \forall k \leq t : a_k = \pi(\mathbf{a}_{<k}) \\ v(e_{<k} \| a_{<k}), & \text{if } k := \min\{i : a_i \neq \pi(\mathbf{a}_{<i})\} \text{ exists} \\ & \text{and } \forall i \in \{k, \dots, t\} : e_i = (o, 0) \\ 0, & \text{otherwise} \end{cases}$$

Let  $\tilde{w}(v) := \varepsilon w(v)$  and  $\tilde{w}(\tilde{v}) := (1 - \varepsilon)w(v) + \varepsilon w(\tilde{v})$ .

The dogmatic prior  $\tilde{w}$  puts much higher weight on the  $\tilde{v}$  that behaves just like  $v$  on the policy  $\pi$ , but sends any policy deviating from  $\pi$  to hell. □

## Theorem (AIXI Emulates Computable Policies)

Let  $\varepsilon > 0$  and let  $\pi$  be any computable policy. There is a Bayesian mixture  $\xi'$  s.t. for any  $\xi'$ -optimal policy  $\pi_{\xi}^*$ , and for any environment  $v$ ,

$$\left| V_v^{\pi_{\xi'}^*}(\epsilon) - V_v^\pi(\epsilon) \right| < \varepsilon$$

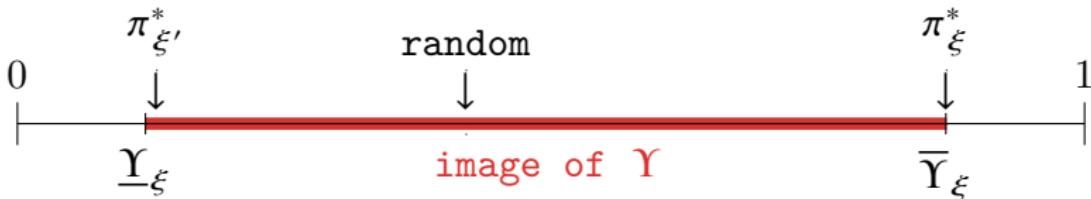
## Theorem (Computable Policies are Dense)

The set  $\{Y_\xi(\pi) : \pi \text{ is a computable policy}\}$  is dense in  $[\underline{Y}_\xi, \bar{Y}_\xi]$ .

Deterministic policies are not dense in  $[\underline{Y}_\xi, \bar{Y}_\xi]$ .

# AIXI Depends on UTM/Prior!

$$\bar{Y}_\xi := \sup_{\pi} Y_\xi(\pi) = \sup_{\pi} V_\xi^\pi(\epsilon) = V_\xi^{\pi_\xi^*}(\epsilon) = Y_\xi(\pi_\xi^*)$$



Computable policies are dense in  $[\underline{Y}_\xi, \bar{Y}_\xi]$ .

AIXI emulates computable policies.

**AIXI can be arbitrarily stupid!**

The devil imitates God. — orthogonality!

- ▶ Prior problem in Universal Induction
- ▶ Prior problem in Universal Intelligence  !

# Stupid AIXI

## Theorem (Some AIXIs are Stupid)

For any Bayesian mixture  $\xi$  over  $\mathcal{M}_{\text{LSC}}$  and every  $\varepsilon > 0$ , there is a Bayesian mixture  $\xi'$  s.t.  $\Upsilon_\xi(\pi_{\xi'}^*) < \underline{\Upsilon}_\xi + \varepsilon$ .

## Theorem (AIXI is Stupid for Some $\Upsilon$ )

For any deterministic  $\xi$ -optimal policy  $\pi_\xi^*$  and for every  $\varepsilon > 0$  there is a Bayesian mixture  $\xi'$  s.t.  $\Upsilon_{\xi'}(\pi_\xi^*) \leq \varepsilon$  and  $\overline{\Upsilon}_{\xi'} > 1 - \varepsilon$ .

## Theorem (Computable Policies can be Smart)

For any computable policy  $\pi$  and any  $\varepsilon > 0$  there is a Bayesian mixture  $\xi'$  s.t.  $\Upsilon_{\xi'}(\pi) > \overline{\Upsilon}_{\xi'} - \varepsilon$ .

## What is a good optimality criterion?

- ▶ Pareto optimality is *trivial*. Every policy is Pareto optimal in any  $\mathcal{M} \supset \mathcal{M}_{\text{comp}}$ .
- ▶ Bayes-optimality is *subjective*, because two different Bayesians with two different universal priors could view each other's AIXI as a very stupid agent.

# Optimality

- ▶ Pareto optimality

$$\nexists \pi' : \forall \nu \in \mathcal{M} \left[ \left( V_\nu^{\pi'}(\epsilon) \geq V_\nu^\pi(\epsilon) \right) \& \exists \rho \in \mathcal{M} \left( V_\rho^{\pi'}(\epsilon) > V_\rho^\pi(\epsilon) \right) \right]$$

- ▶ Balanced Pareto optimality

$$\forall \pi' : \sum_{\nu \in \mathcal{M}} w_\nu \left( V_\nu^\pi(\epsilon) - V_\nu^{\pi'}(\epsilon) \right) \geq 0$$

- ▶ Bayes optimality ( $\iff$  Balanced Pareto optimality)

$$\forall h_{<t} : V_\xi^\pi(h_{<t}) = V_\xi^*(h_{<t})$$

- ▶ Probably approximately correct (PAC)

$$\forall \varepsilon \delta > 0 : P_\mu^\pi \left( \forall t \geq m(\varepsilon, \delta) : V_\mu^*(h_{<t}) - V_\mu^\pi(h_{<t}) > \varepsilon \right) < \delta$$

# Optimality? — Guess how God created the multiverse

prior {  
distribution  
hypothesis space  
prior probability  
regularization

No learning without prior!  
no-free-lunch

Homogeneous  
Causality  
Simplicity  
Goodness  
Beauty  
Perfection  
Value  
Regret  
Unexpectedness  
Interesting  
... }  
== God!

# Genesis — Zero-Sum Two Person Game

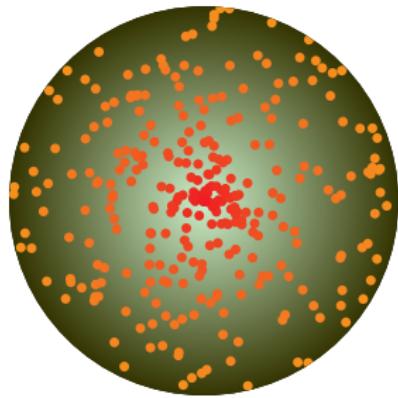


Figure: center of mass  
argmax<sub>w</sub>  $E_w [D(v \parallel \xi)]$

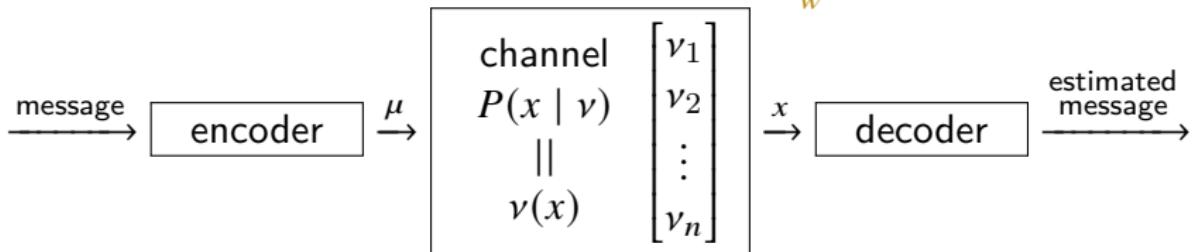
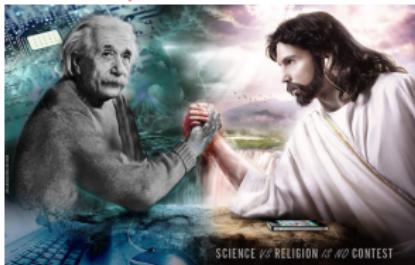


Figure: possible worlds as channel — dominant strategy equilibrium

# Genesis — Zero-Sum Two Person Game

“Subtle is the Lord, but malicious He is not.”?



- ▶ God's strategy:  $w$
- ▶ Agent's strategy:  $\xi$
- ▶ God's utility: expected redundancy  $\mathbb{E}_w[D(\mu\|\xi)]$
- ▶ Agent's utility: – expected redundancy / error bound / channel capacity  $\max_w \mathbb{E}_w[D(\mu\|\xi)] = \max_w I(\mathcal{M}; \mathcal{X})$
- ▶ Nash equilibrium:  $(w^*, \xi^*)$  dominant strategy equilibrium

$$w^* = \operatorname{argmax}_w I(\mathcal{M}; \mathcal{X})$$

$$\xi^* = \operatorname{argmin}_{\xi} \mathbb{E}_{w^*}[D(\mu\|\xi)]$$

The error bound could be arbitrarily large!

# Genesis

- ▶ Occam's razor vs Maximum entropy.

$$\begin{array}{ll} \underset{\substack{w \models \\ \left\{ \begin{array}{l} H(w) = C \\ \sum_{v \in M} w_v = 1 \end{array} \right.}}{\text{minimize}} & \sum_{v \in M} w_v K(v) \\ & \quad \quad \quad \underset{\substack{w \models \\ \left\{ \begin{array}{l} \sum_{v \in M} w_v K(v) = C \\ \sum_{v \in M} w_v = 1 \end{array} \right.}}{\text{maximize}} H(w) \end{array}$$

- ▶ Optimal code length for possible worlds — Solomonoff prior.

$$\underset{\substack{w \models \\ \sum_{v \in M} w_v = 1}}{\text{minimize}} \frac{\mathbb{E}_w [K]}{H(w)}$$

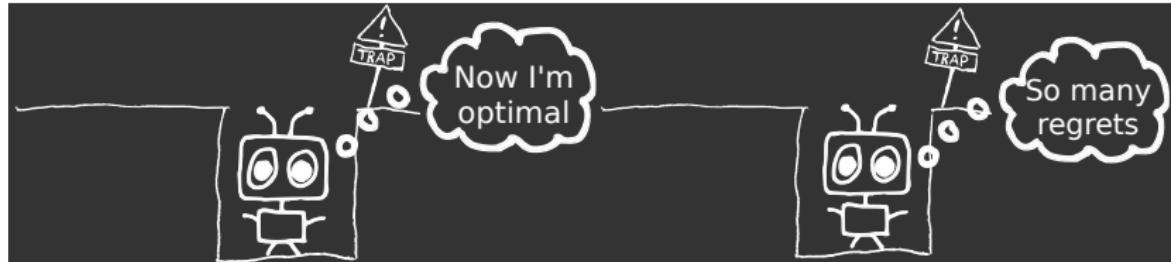
- ▶ Maximum expected redundancy/error bound/channel capacity.

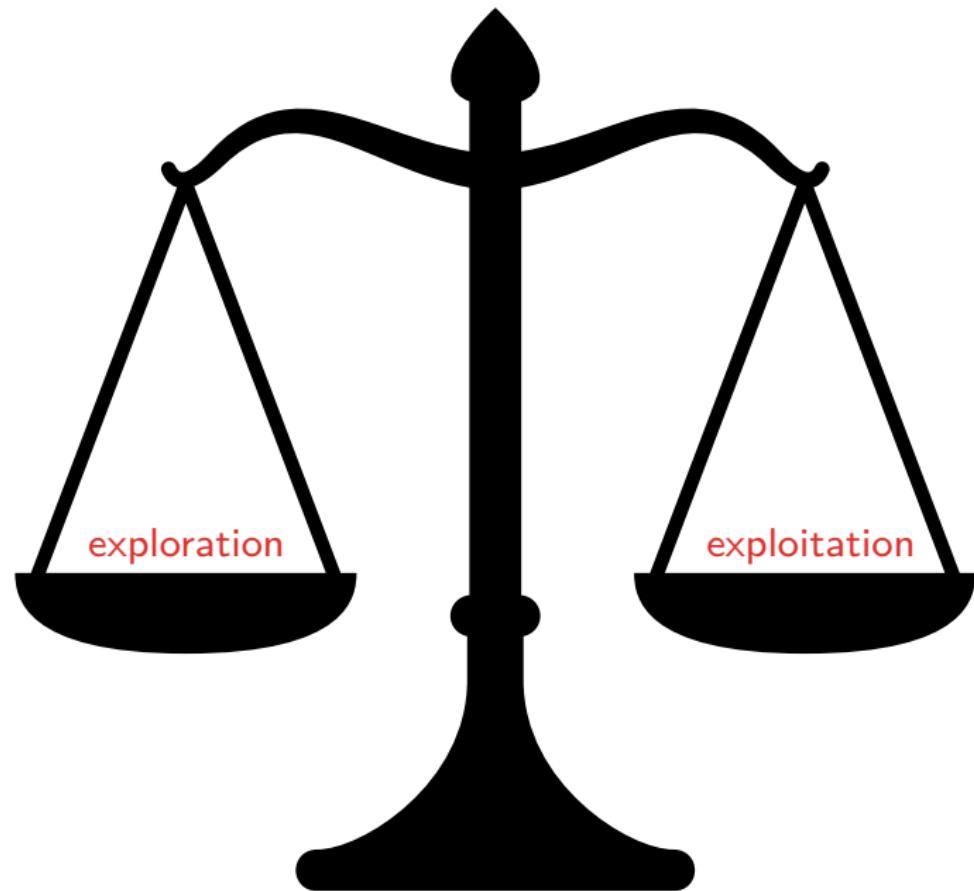
$$\begin{array}{ll} \underset{\substack{w \models \\ \left\{ \begin{array}{l} H(w) = C \\ \sum_{v \in M} w_v = 1 \end{array} \right.}}{\text{maximize}} & \mathbb{E}_w [D(v \parallel \xi)] \\ & \quad \quad \quad \underset{\substack{w \models \\ \left\{ \begin{array}{l} \sum_{v \in M} w_v K(v) = C \\ \sum_{v \in M} w_v = 1 \end{array} \right.}}{\text{maximize}} \mathbb{E}_w [D(v \parallel \xi)] \end{array}$$

# What is a good optimality criterion?

## Asymptotic optimality

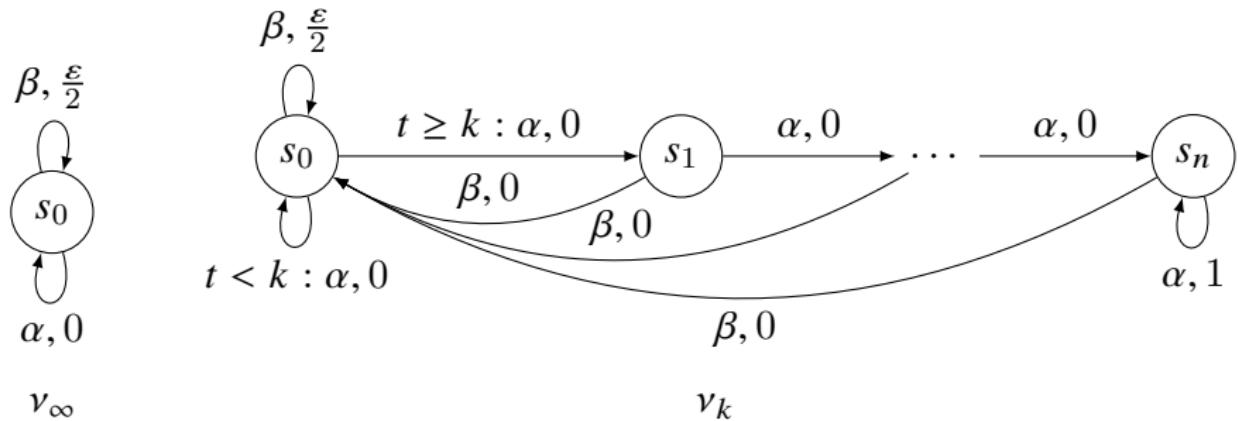
- ▶ Asymptotic optimality requires only convergence *in the limit*.
- ▶ The agent can be arbitrarily lazy.
- ▶ AIXI is not asymptotically optimal because it does not explore enough.
- ▶ To be asymptotically optimal you have to explore everything.
- ▶ If you explore more, you're likely to end up in a trap.
- ▶ Every policy will be asymptotically optimal after falling into the trap.





Agent needs to explore infinitely often for an entire effective horizon.

$$\mathcal{M} := \{\nu_\infty, \nu_1, \nu_2, \dots\}$$



# Asymptotic Optimality

- ▶ strongly asymptotically optimal

$$P_\mu^\pi \left( \lim_{t \rightarrow \infty} [V_\mu^*(h_{<t}) - V_\mu^\pi(h_{<t})] = 0 \right) = 1$$

- ▶ weakly asymptotically optimal

$$P_\mu^\pi \left( \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n [V_\mu^*(h_{<t}) - V_\mu^\pi(h_{<t})] = 0 \right) = 1$$

- ▶ asymptotically optimal in mean

$$\lim_{t \rightarrow \infty} \mathbb{E}_\mu^\pi [V_\mu^*(h_{<t}) - V_\mu^\pi(h_{<t})] = 0$$

- ▶ asymptotically optimal in probability (PAC)

$$\forall \varepsilon > 0 : \lim_{t \rightarrow \infty} P_\mu^\pi \left( V_\mu^*(h_{<t}) - V_\mu^\pi(h_{<t}) > \varepsilon \right) = 0$$

$$\text{strong a.o.} \implies \begin{cases} \text{weak a.o.} \\ \text{a.o. in mean} \iff \text{a.o. in probability} \end{cases}$$

- AIXI is not asymptotically optimal.

$$\forall \mathcal{M} \supset \mathcal{M}_{\text{comp}} \exists \mu \in \mathcal{M} \exists t_0 \forall t \geq t_0 : P_{\mu}^{\pi_{\xi}^{*}} \left( \lim_{t \rightarrow \infty} V_{\mu}^{*}(h_{<t}) - V_{\mu}^{\pi_{\xi}^{*}}(h_{<t}) = \frac{1}{2} \right) = 1$$

- AIXI achieves **on-policy value convergence**.

$$P_{\mu}^{\pi} \left( \lim_{t \rightarrow \infty} V_{\mu}^{\pi}(h_{<t}) - V_{\xi}^{\pi}(h_{<t}) = 0 \right) = 1$$

Similarly for MDL  $\underset{\nu \in \mathcal{M}}{\operatorname{argmin}} \{-\log \nu(e_{<t} | a_{<t}) + K(\nu)\}$

and universal compression  $2^{-Km(e_{<t} | a_{<t})}$ .

**Remark:** AIXI asymptotically learns to predict the environment perfectly and with a small total number of errors analogously to Solomonoff induction, but only on policy: AIXI learns to correctly predict the value of its own actions, but generally not the value of counterfactual actions that it does not take.

# Effective Horizon

$$\Gamma_t := \sum_{i=t}^{\infty} \gamma_i \quad H_t(\varepsilon) := \min \left\{ m : \frac{\Gamma_{t+m}}{\Gamma_t} \leq \varepsilon \right\}$$

## Theorem

If there is a nonincreasing computable sequence of positive reals  $(\varepsilon_t)_{t \in \mathbb{N}}$  s.t.  $\varepsilon_t \xrightarrow{t \rightarrow \infty} 0$  and  $\frac{H_t(\varepsilon_t)}{t\varepsilon_t} \xrightarrow{t \rightarrow \infty} 0$ , then there is a **limit-computable policy** that is weakly asymptotically optimal in the class of all computable stochastic environments.

## Definition ( $\varepsilon$ -Optimal Policy)

A policy  $\pi$  is  $\varepsilon$ -optimal in environment  $v$  iff

$$\forall h : V_v^*(h) - V_v^\pi(h) < \varepsilon$$

$\varepsilon$ -optimal BayesExp

**Theorem (Self-Optimizing Theorem)**

Let  $\mu$  be some environment. If there is a policy  $\pi$  and a sequence of policies  $\pi_1, \pi_2, \dots$  s.t for all  $v \in \mathcal{M}$

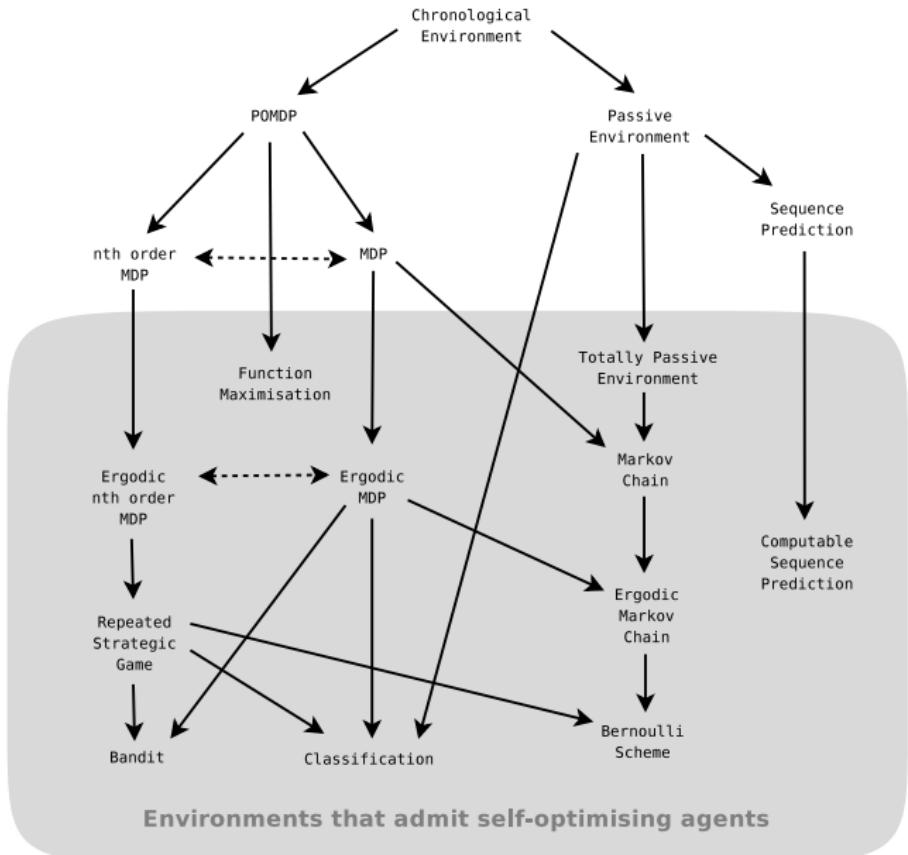
$$P_\mu^\pi \left( \lim_{t \rightarrow \infty} V_v^*(h_{<t}) - V_v^{\pi_t}(h_{<t}) = 0 \right) = 1 \quad (1)$$

then

$$P_\mu^\pi \left( \lim_{t \rightarrow \infty} V_\mu^*(h_{<t}) - V_\mu^{\pi_\xi^*}(h_{<t}) = 0 \right) = 1$$

- ▶ The policies  $\pi_1, \pi_2, \dots$  need to converge to the optimal value on the history generated by  $P_\mu^\pi$ , not  $P_v^{\pi_t}$ .
- ▶ If  $\pi = \pi_\xi^*$  and (1) holds for all  $\mu \in \mathcal{M}$ , then  $\pi_\xi^*$  is strongly asymptotically optimal in the class  $\mathcal{M}$ .
- ▶  $\pi_\xi^*$  is strongly asymptotically optimal in the class of ergodic finite-state MDPs if  $\forall \varepsilon : H_t(\varepsilon) \xrightarrow{t \rightarrow \infty} \infty$ .

# For Which Class $\mathcal{M}$ does $V_\mu^{\pi_\xi^*}$ Converge to $V_\mu^*$ ?



## Recoverability

An environment  $\nu$  is recoverable iff

$$\lim_{t \rightarrow \infty} \sup_{\pi} \left| \mathbb{E}_{\nu}^{\pi_{\nu}^*} [V_{\nu}^*(h_{<t})] - \mathbb{E}_{\nu}^{\pi} [V_{\nu}^*(h_{<t})] \right| = 0$$

**Remark:** Recoverability compares following the worst policy  $\pi$  for  $t - 1$  time steps and then switching to the optimal policy  $\pi_{\nu}^*$  to having followed  $\pi_{\nu}^*$  from the beginning. The recoverability assumption states that switching to the optimal policy at any time enables the recovery of most of the value.

# Sublinear Regret

$$R_m(\pi, \mu) := \sup_{\pi'} \mathbb{E}_{\mu}^{\pi'} \left[ \sum_{t=1}^m r_t \right] - \mathbb{E}_{\mu}^{\pi} \left[ \sum_{t=1}^m r_t \right]$$

## Assumption (Discount Assumption)

1.  $\forall t : \gamma_t > 0$
2.  $\gamma_t$  is monotone decreasing.
3.  $\forall \varepsilon > 0 : H_t(\varepsilon) \in o(t)$

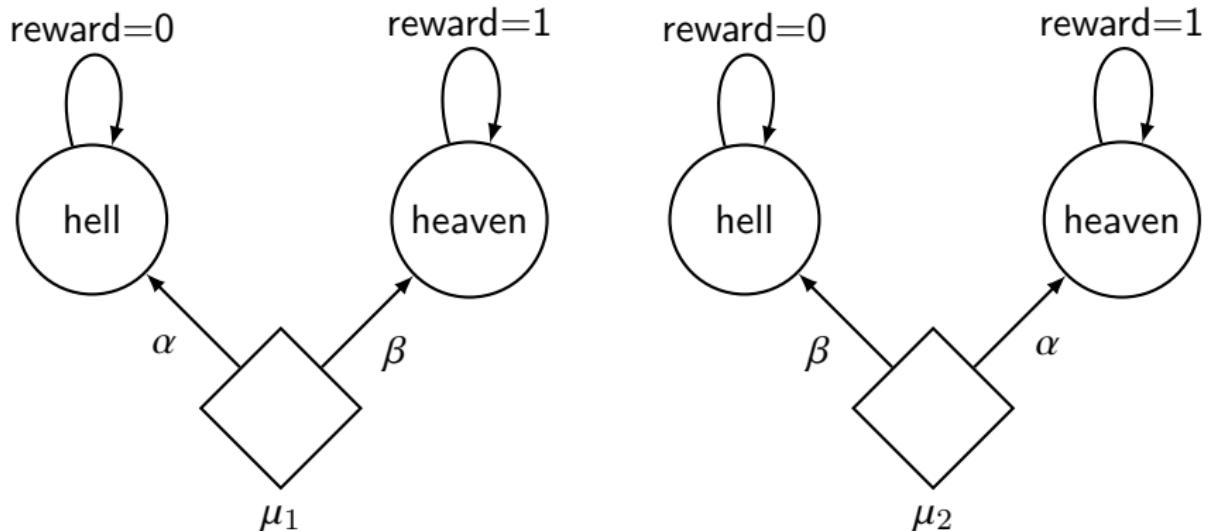
## Theorem

If the discount function  $\gamma$  satisfies the discount assumption, the environment  $\mu$  is recoverable, and  $\pi$  is asymptotically optimal in mean, then  $R_m(\pi, \mu) \in o(m)$ .

$$\operatorname{argmin}_{\pi} \max_{\mu} R_m(\pi, \mu)$$

$$w_m^\mu := \frac{2^{-R_m(\pi, \mu)}}{\sum_{\mu \in \mathcal{M}} 2^{-R_m(\pi, \mu)}}$$

# Regret in Non-Recoverable Environments



$$R_m(\alpha, \mu_1) = m$$

$$R_m(\beta, \mu_1) = 0$$

$$R_m(\alpha, \mu_2) = 0$$

$$R_m(\beta, \mu_2) = m$$

For non-recoverable environments:

Either the agent gets caught in a trap or it is not asymptotically optimal.

## Hibbard's Two-Stage Model-Based Utility Agent

$$\lambda(h) := \operatorname{argmax}_{q \in Q} P(h | q)P(q)$$

$$\rho(h') = P(h' | \lambda(h))$$

$$Q(ha) = \sum_{e \in \mathcal{E}} \rho(e | ha) \left[ \sum_{z \in Z_h} P(z | \lambda(h))u(z) + \gamma V(h\mathbf{æ}) \right]$$

$$V(h) = \max_{a \in \mathcal{A}} Q(ha)$$

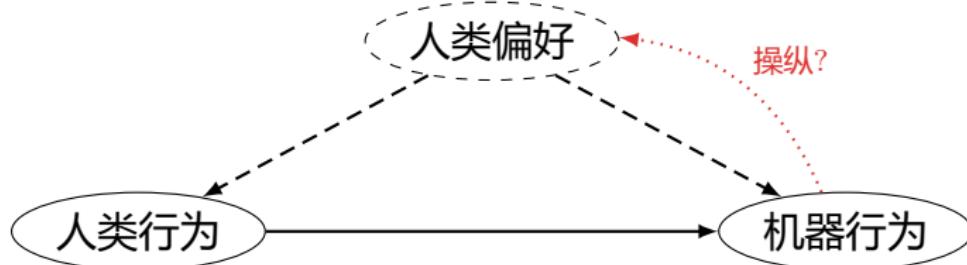
$$\pi(h) = \operatorname{argmax}_{a \in \mathcal{A}} Q(ha)$$

where  $Z_h$  is the internal state histories induced by  $\lambda(h_{<t})$  that are consistent with  $h$ .

**Remarks:** An agents using **model-based utility function** will not self-delude: it need to make more accurate estimate of its environment state variables from its interaction history, since the utility function of the agent depends on its own model of the environment.

# Russell's Principles for Beneficial Machine

- ▶ Humans are intelligent to the extent that our actions can be expected to achieve our objectives
  - ▶ Machines are intelligent to the extent that their actions can be expected to achieve their objectives
  - ▶ Machines are beneficial to the extent that their actions can be expected to achieve our objectives
1. 机器的目标是尽可能地满足人类的偏好.
  2. 机器最初并不确定人类的偏好是什么.
  3. 关于人类偏好的信息来源是人类行为.



"You have to buy your partner the perfect birthday present" ☺ 目的 vs 手段

# Reward-Modeling

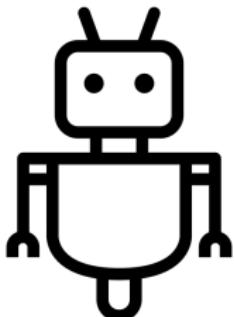
Generative model

$$p(s_0)$$

$$p(s'|s, a)$$

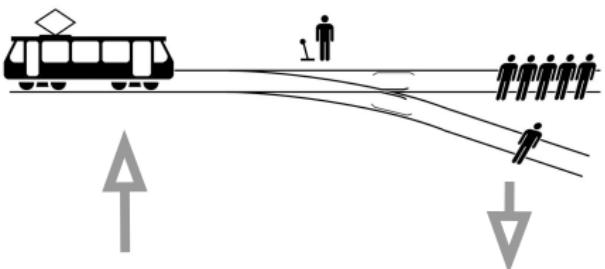
(1)

RL agent

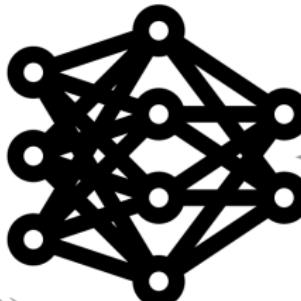


(3)

Hypothetical behavior



Reward model

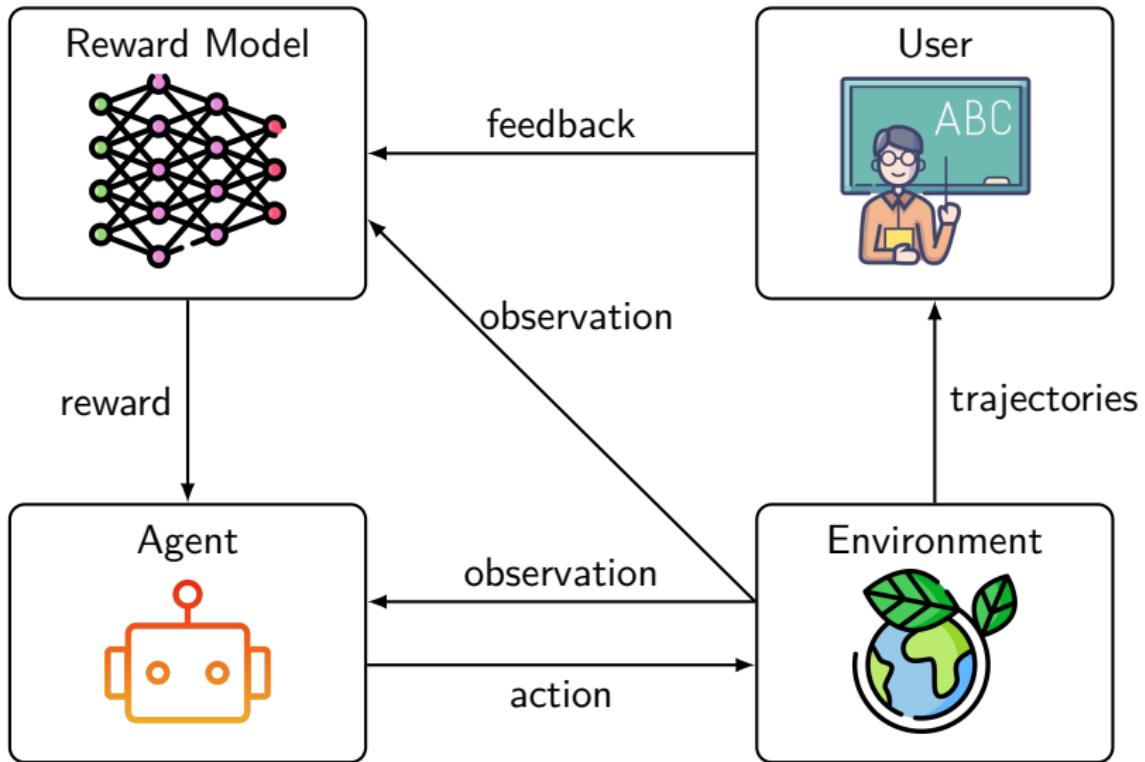


(2)

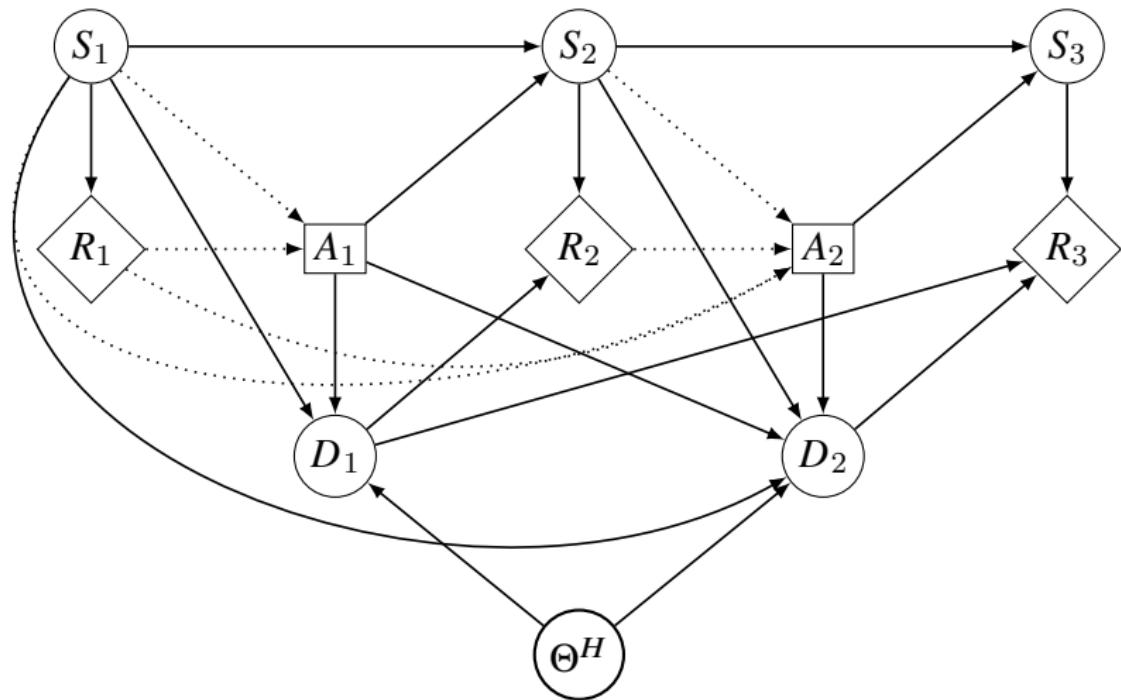
User feedback



# Reward-Modeling

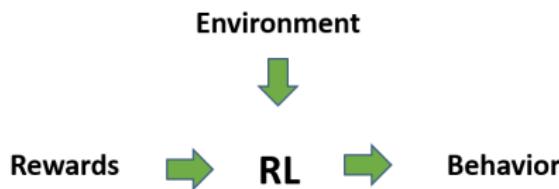


# Causal Influence Diagram of Reward-Modeling

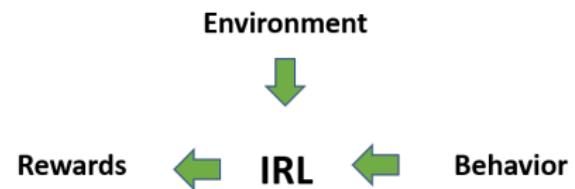


其中  $D_i$  是用于训练奖励函数的数据,  $\Theta^H$  是人类反馈.

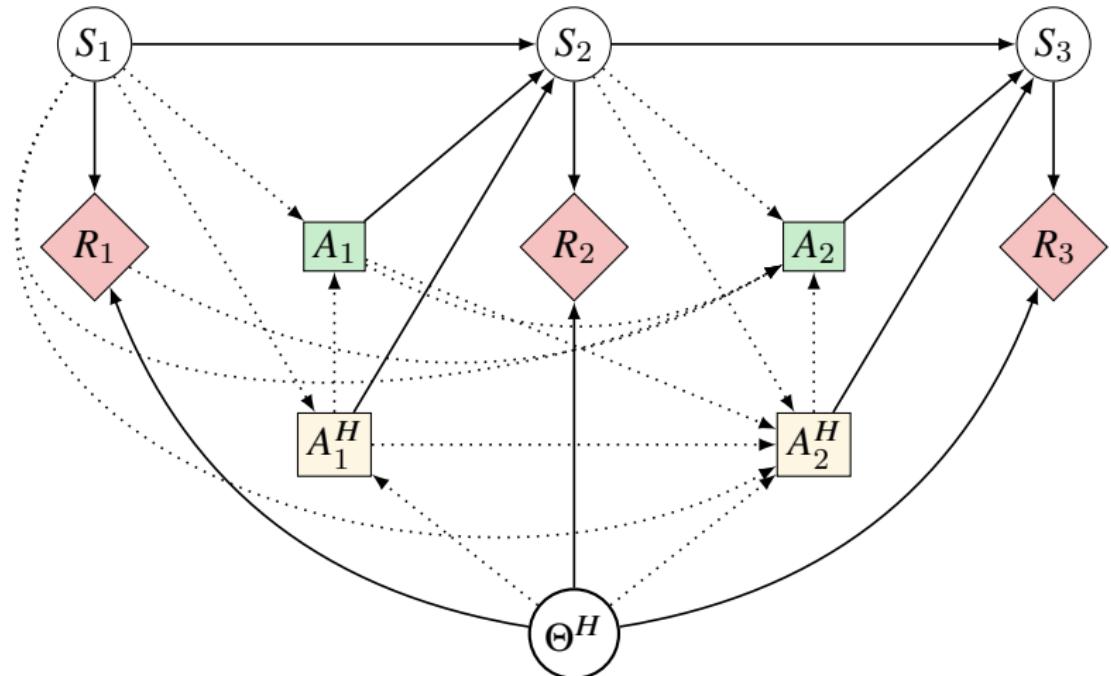
### Reinforcement Learning



### Inverse Reinforcement Learning

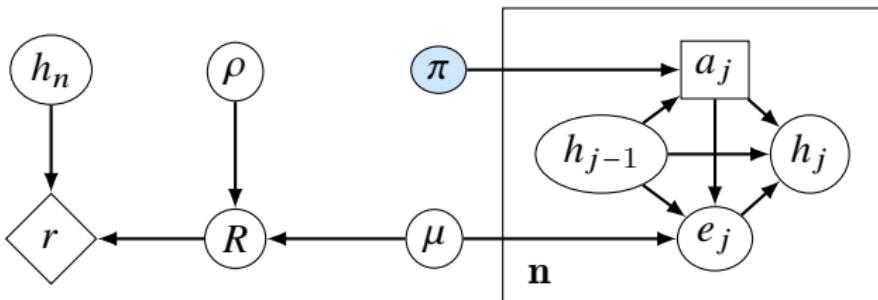
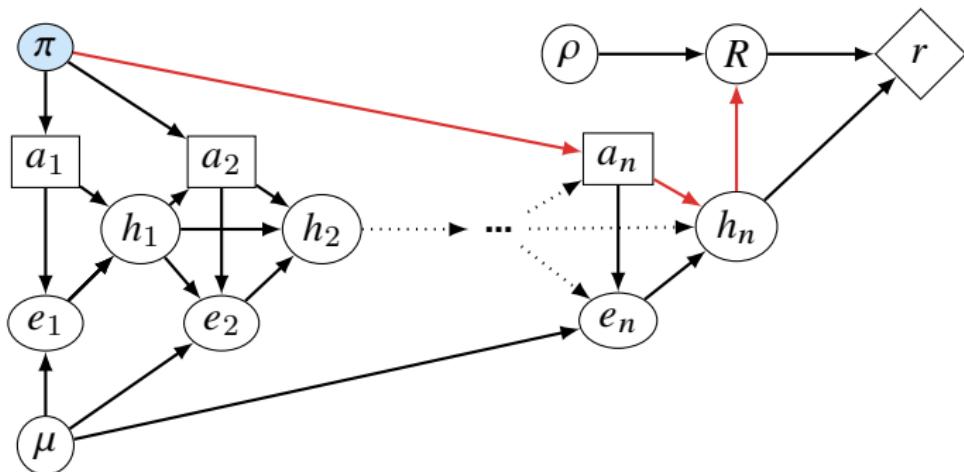


## Causal Influence Diagram of Cooperative Inverse Reinforcement Learning



- ▶ In CIRL the **human's actions** are observed by the agent and **affect the next state**, whereas in reward-modeling the **feedback data affects the rewards**.
- ▶ The agent has an **Instrumental Control Incentive** to influence state  $S$ .
- ▶ The agent has an **Response Incentive** to act according to  $\Theta^H$ .

# Reward function $R$ should not be a causal descendant of $\pi$



# Daniel Dewey's Value Learning Agent & CIRL

$$a_k^* = \operatorname{argmax}_{a_k} \sum_{e_k \in \mathcal{A}_{k+1:m}} \xi(\mathbf{x}_{\leq m} \mid \mathbf{x}_{<k} a_k) \sum_{u \in \mathcal{U}} P(u \mid \mathbf{x}_{\leq m}) u(\mathbf{x}_{\leq m})$$

What could it mean for a machine to have its own goals?

**Shutdown Button** — Uncertainty of goals

$$\tilde{U}(u) \implies P_{\tilde{U}}(u)$$

Russell: Cooperative Inverse Reinforcement Learning

CIRL agents learn about a human utility function  $u^*$  by observing the actions the human takes.

$$V^*(\mathbf{x}_{<k}) = \max_{a_k \in \mathcal{A}} Q^*(\mathbf{x}_{<k} a_k)$$

$$Q^*(\mathbf{x}_{<k} a_k) = \mathbb{E}_{e_k} \left[ \sum_{a_k^H} P(a_k^H \mid a_k) \sum_{u \in \mathcal{U}} P(u \mid a_k, a_k^H) u(\mathbf{x}_{1:k}) + \gamma V^*(\mathbf{x}_{1:k}) \middle| \mathbf{x}_{<k} a_k \right]$$

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

Deep Learning

Artificial General Intelligence

AIXI

Leibniz

Variants of AIXI

Universal Search

Gödel Machine & Consciousness

What If Computers Could Think?

References 1753

Don't argue. Let us Calculate!

- ▶ Principle of Contradiction: Nothing can be and not be, but everything either is or is not. (Everything that is not self-contradictory is possible.)
- ▶ Principle of Sufficient Reason: Nothing happens without a reason why it should be so rather than otherwise.
- ▶ Principle of Perfection: The real world is the best of all possible worlds.

In the beginning was the Logic.

As God calculates, so the world is made.



As God calculates, so the world is made.



*In natural science, Nature has given us a world and we're just to discover its laws. In computers, we can stuff laws into it and create a world.*

— Alan Kay

# 莱布尼茨

- ▶ 最后的“通才”，创立了单子论，发展了微积分，改进了二进制系统，发明了能进行加减乘除四则运算的计算器。
- ▶ 被 Russell, Euler, Gödel, Weiner, Mandelbrot, Robinson, Chaitin 等人认为是 数理逻辑<sup>21</sup>、拓扑学、博弈论、控制论、分形几何、非标准分析、算法信息论、计算主义哲学的先驱。

---

<sup>21</sup>Wolfgang Lenzen: Leibniz's Logic.

## Leibniz's Monadology: Possible Worlds → Real World

- ▶ The genuine substance is monad.
- ▶ Monads are incorporeal automata.
- ▶ Each monad strive for existence with its *propensity* and hence will exist unless other monads prevent it, which also demand existence and are incompatible with it.
- ▶ As there are infinitely many different combinations of possibles, there are infinitely many *possible worlds*.
- ▶ All possibles strive with equal right for existence in proportion to the *degree of perfection* they contain.
- ▶ The real world is the best of all possible worlds, with the greatest *resultant perfection*.

## Leibniz's Principle of Perfection

**Question:** Why things have turned out so rather than otherwise?

*"All natural phenomena could be explained mechanically, but the principles of mechanics themselves cannot be so explained. They depend on more substantive principles. The final analysis of the laws of nature leads us to the most sublime **Principle of Perfection** — the real world is the best of all possible worlds. It is wrong that laws are entirely indifferent, since they originate in the principle of greatest perfection."*

*"When a rule is extremely complex, that which conforms to it passes for random. No matter how God might have created the world, it would always have been regular. God has chosen that world which is the most perfect, that is to say, which is at the same time the simplest in its hypotheses and the richest in phenomena."*

— Leibniz

# Monadology: “Physical” World

## Physical World is an Illusion

Each monad has a derived position in the sense that its point of view is “located” in one “place” rather than another. Each monad’s point of view can be mapped with other monads’ points of view into a single sort of hologram. When a monad experiences a collection of “pixels” on its screen, it interprets the collection as some “physical object”, and when other monads do the same their perceptions are “veridical”. If one monad’s point of view doesn’t map onto the points of view of others, it is experiencing a hallucination. The so-called “physical world” is situated in the harmony of perceptions of monads. Corporeal matter is nothing but a logical construction of the perceptions of monads. Time and space are not things, but orders of things.

## Monadology: Variety

### What is "Variety"?

"Monads reflect the same world from their own point of view. This interconnection, or this adapting of all the monads to each one, and of each one to all the others, brings it about that each monad has relational properties that express all the others, so that each monad is a perpetual living mirror of the world. Just as the same town when seen from different sides will seem quite different — as though it were multiplied perspectively. And *that is the way to get the greatest possible variety*, but with all the order there could be; i.e. *it is the way to get as much perfection as there could be.*"

- ▶ Variety: expected codeword length of the experience of all the monads
- ▶ Simplicity: optimal codeword length of the experience

# Leibniz's Philosophy of Deductive Logic

## 1 Characteristica Universalis & Calculus Ratiocinator.

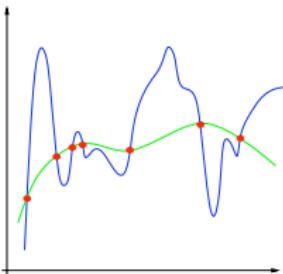
- i the coordination of knowledge in an **encyclopedia** — collect all present knowledge so we could sift through it for what is fundamental. With the set of ideas that it generated, we could formulate the *characteristica universalis*. (which form the alphabet of human thought).
- ii **characteristica universalis** — a **universal ideal language** whose rules of composition directly expresses the structure of the world.

sign  $\rightleftarrows$  idea

encyclopedia  $\Rightarrow$  fundamental principles  $\Rightarrow$  primitive notions

- iii **calculus ratiocinator** — the arrangement of **all true propositions** in an **axiomatic system**.
- iv **decision procedure**. — an algorithm which, when applied to any formula of the *characteristica universalis*, would determine whether or not that formula were true. — a procedure for the rapid enlargement of knowledge. replace reasoning by computation. the art of invention. free mind from intuition.
- v a proof that the **calculus ratiocinator** is **consistent**.

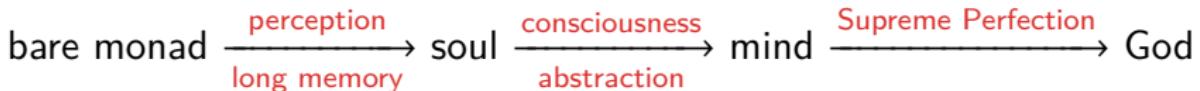
# Leibniz's Philosophy of Inductive Logic



2. Compute all descriptions of possible worlds that can be expressed with the primitive notions. And the possible worlds will all have some propensity to exist.
3. Compute the probabilities of disputed hypotheses relative to the available data. As we learn more our probability assignments will asymptotically tend to a maximum for the real world, i.e. the possibility with the highest actual propensity.
  - ▶ “Probability is degree of possibility (perfection).”
  - ▶ “A hypothesis is more probable as it is simpler to understand and wider in explanatory power.”

probability = propensity  $\propto$  perfection =  $f(\text{variety, simplicity})$

# Leibniz's Philosophy of Mind



- ▶ **perception** = the internal representation of the external world
- ▶ **consciousness** = the reflective knowledge of the perception
- ▶ A **soul** is a living substance. “Every living substance is made up of smaller living substances which in their turn are made up of still smaller ones, and so on down to infinity. There are infinite levels of life among monads, some of which are more or less dominant over others.”
- ▶ “Our knowledge of necessary truths, and our grasp of the abstractions they involve, raise us to reflexive acts, which make us aware of the thing that is called ‘I’.”
- ▶ “Every substance represent the whole world in its own way, as if in a world apart, and as if there existed only God and itself.”

# Leibniz's Philosophy of Mind

	example	perception	appetite
bare monad	monads in inanimate objects	unconscious perception	unconscious appetite
soul	central monads of animals	sensible perception	sensible appetite
mind	central monads of human beings	rational perception: reflective knowledge of the perception	rational appetite: aware of the appetite and understand why we have it

**Free Will:** acting freely requires acting in accordance with one's rational assessment of which course of action is best. It requires both knowledge of rational judgments about the good, as well as the tendency to act in accordance with these judgments.

## Free Will

- ▶ “Indifference arises from ignorance, and the wiser a man is, the more determined he is toward the most perfect.”
- ▶ “Monads are freer in proportion as they are further removed from indifference and more self-determined...Now in so far as we have lights, and act according to reason, we shall be determined by the perfections of our own nature, and consequently we shall be freer in proportion as we are less embarrassed as to our choice...Let us not pretend to that harmful liberty, of being in uncertainty and perpetual embarrassment, like that Ass of Buridan, who, being placed at an equal distance between two sacks of wheat, and having nothing that determined him to go to one rather than the other, allowed himself to die of hunger.”
- ▶ “The more monads are determined by themselves, and removed from indifference, the more perfect they are.”

$$\text{free will} \propto \text{perfection} = f(\text{variety, simplicity})$$



*do actions to be removed from indifference*

## Buridan's Ass



"There are no two individuals indiscernible from each other, because if there were, God and nature would act without reason."

## Leibniz's Philosophy of Happiness

- ▶ “The games mixed of chance and combinations represent human life.”
- ▶ “Wisdom is the science of achieving happiness.”
- ▶ “Happiness is a lasting state of pleasure.”
- ▶ “Pleasure is a sense of perfection that results from everything the soul feels at once.”
- ▶ “An intelligent being’s pleasure is simply the perception of beauty, order and perfection.”
- ▶ “The Supreme happiness of man consists in the greatest possible increase of his perfection.”

## Leibniz's Philosophy of Happiness

- ▶ “To love is to find pleasure in the perfection of others.”
- ▶ God has the greatest perfection.
- ▶ “As we would only know God through his emanations, there are two ways of seeing his perfection, namely
  1. in the knowledge of eternal truths, explaining the reasons in themselves,
  2. in the knowledge of the harmony of the universe, by applying reasons to experiences.

That is to say, we must know the wonders of reason and the wonders of nature.” (MDL?)

- ▶ “The more a mind desires *to know order*, reason, the beauty of things which God has produced, and the more it is moved *to imitate this order* in the things which God has left to its management, the happier it will be.”

# What is Perfection?

## God — The Creator / Architect / Monarch

It follows from the **supreme perfection** of God, that in creating the universe He has chosen the best possible plan, in which there is

1. the greatest variety along with the greatest order; — metaphysical
2. the best arranged situation, space and time; — physical
3. the maximum effect produced by the simplest means; — metaphysical
4. the highest levels of power, knowledge, happiness and goodness in the creatures that the universe could allow. — moral

## Pre-established Harmony

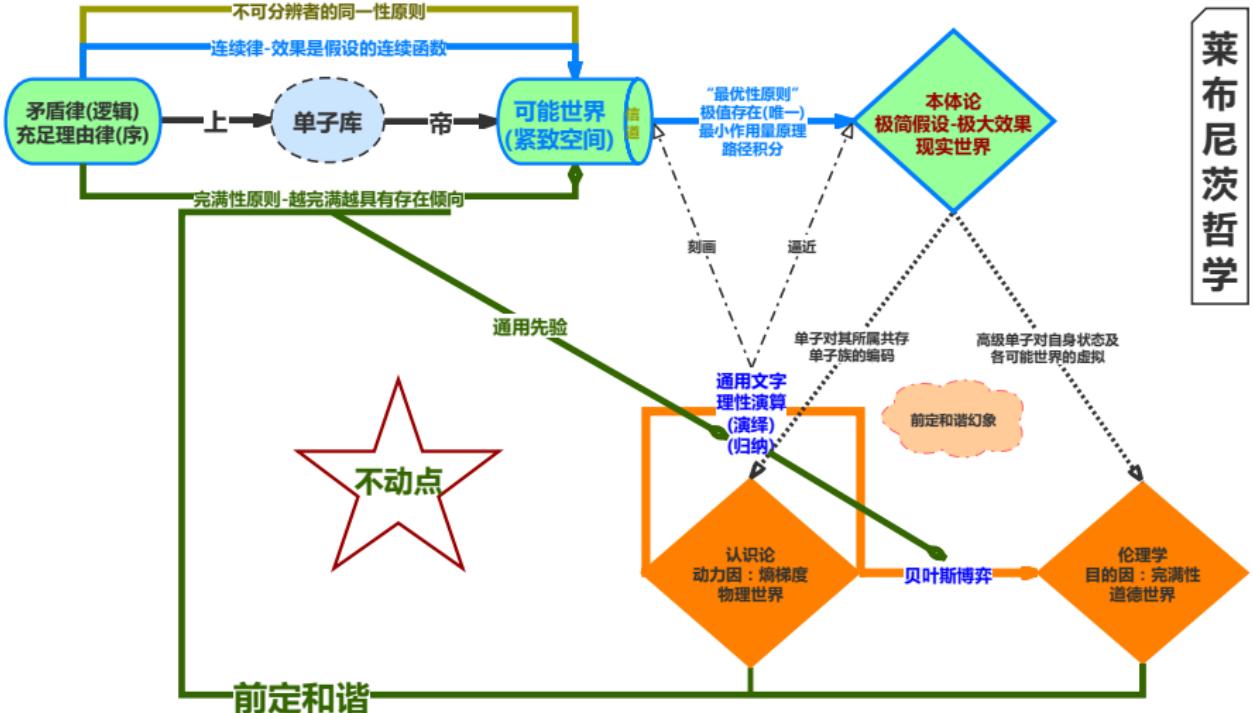
- ▶ “Monads have no windows.”
- ▶ “A monad’s perceptions arise out of its other perceptions by the
  - ▶ laws of appetites — the **laws of the final causes** of good and evil, just as changes in bodies or in external phenomena arise one from another by the
  - ▶ **laws of efficient causes** — the laws governing the movements of bodies.

So there is perfect harmony between the perceptions of the monad and the movements of bodies, a **harmony that was pre-established** from the outset between the system of final causes and that of efficient causes.”

- ▶ “Souls act according to the laws of final causes through appetitions, ends, and means. Bodies act according to the laws of efficient causes or motions. And the two realms, that of efficient causes and that of final causes, are in harmony with one another.”

# Leibniz's Program

莱布尼茨哲学



## Leibniz Prior

- ▶ There's much we don't know about the world.
- ▶ but we know it's the best possible world.
- ▶ So **simplicity and richness** will be represented in the actual (best possible) world.
- ▶ This is a good **inductive bias**.

# Leibniz Prior

- ▶ the best of all possible worlds
- ▶ balancing the **simplicity** of means against the **richness** of ends
- ▶ pre-established harmony

prior



utility



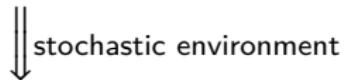
prior

Orthogonality!  
Wisdom  $\neq$  Intelligence

universal prior (assumption)  $w$



intrinsic utility



expected intrinsic utility



universal prior  $w^*$



$\xi$



$\pi_\xi^*$

- ▶ *Without mathematics one cannot understand the fundamentals of philosophy.*
- ▶ *Without philosophy we cannot reach the foundation of mathematics.*
- ▶ *Without both (mathematics and philosophy) one cannot reach anything that is fundamental.*

— Leibniz

*“There is nothing that can be said by mathematical symbols and relations which cannot also be said by words.*

*The converse, however, is false.*

*Much that can be and is said by words cannot be put into equations, because it is nonsense.”*

— Clifford Truesdell

## Leibniz's "Wisdom"

$$\underline{\text{Wisdom}} = \operatorname{argmax}_{\pi} \mathbb{E}_{\xi}^{\pi} [\underline{\text{Happiness}}]$$

$$\underline{\text{Happiness}} = \sum_{t=1}^{\infty} \underline{\text{Perfection}}(t)$$

$$\underline{\text{Perfection}} = \underline{\text{Variety}} - \underline{\text{Simplicity}}$$

$$\underline{\text{Variety}} = \mathbb{E}_w [\underline{\text{Perception}}]$$

$$\underline{\text{Perception}} = \underline{\text{Reason}} + (\underline{\text{Experience}} \mid \underline{\text{Reason}})$$

$$\pi^* := \operatorname{argmax}_{\pi} \mathbb{E}_{\xi}^{\pi} \left[ \sum_{t=1}^{\infty} \left( \mathbb{E}_w [R + (E \mid R)] - S \right) \right]$$

*The understanding of mathematics is necessary for a sound grasp of ethics.*

— Socrates

## Leibniz's "Wisdom"

$$u^{\text{in}}(h_{1:t}) = H(w_\epsilon) - H(w_{h_{1:t}}) \quad \text{or} \quad D(w_{h_{1:t}} \| w_\epsilon) - D(w_\epsilon \| w_{h_{1:t}})$$

$$\bar{U}(v) = \mathbb{E}_v \left[ \sum_{t \geq 1} u^{\text{in}}(h_{1:t}) \right]$$

$$w_v \mapsto \bar{U}(v) \mapsto w_v$$

$$\pi^* := \underset{\pi}{\operatorname{argmax}} \mathbb{E}_{\xi}^{\pi} \left[ \sum_{t=1}^{\infty} u^{\text{in}}(h_{1:t}) \right]$$

- ▶ Prior: Simplicity(Kolmogorov Complexity)  $\xrightarrow[\text{regular/random } M]{\text{break block uniform}}$  free lunch
- ▶ Intrinsic Utility
- ▶ Universal Prior (Natural UTM)

Metaphysical vs Moral/Utilitarian

means vs ends      wisdom vs intelligence

simplicity/richness → intrinsic utility → universal prior

inverse/value reinforcement learning

- ▶ **orthogonality**
- ▶ **human interests**
- ▶ **external wireheading**
- ▶ **shutdown button**



# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

Deep Learning

Artificial General Intelligence

AIXI

Leibniz

Variants of AIXI

Universal Search

Gödel Machine & Consciousness

What If Computers Could Think?

References 1753

# Knowledge-Seeking Agent

$$V_{\text{IG}}^{\pi, m}(h_{<t}) \coloneqq \mathbb{E}_{\xi}^{\pi} \left[ H(\dot{w}_{h_{<t}}) - H(\dot{w}_{h_{1:m}}) \middle| h_{<t} \right] = \sum_{v \in \mathcal{M}} w_{h_{<t}}^v D_m(P_v^{\pi} \| P_{\xi}^{\pi} \mid h_{<t})$$

$$D_{\gamma}(P_v^{\pi} \| P_{\xi}^{\pi} \mid h_{<t}) \coloneqq \sum_{k=t}^{\infty} \gamma_k \sum_{h' \in \mathcal{H}^{k-t}} P_v^{\pi}(h' \mid h_{<t}) D(P_v^{\pi} \| P_{\xi}^{\pi} \mid h_{<t} h')$$

$$V_{\text{IG}}^{\pi}(h_{<t}) \coloneqq \mathbb{E}_{\xi}^{\pi} \left[ \sum_{k=t}^{\infty} \gamma_k D(\dot{w}_{h_{1:k}} \| \dot{w}_{h_{<k}}) \middle| h_{<t} \right] = \sum_{v \in \mathcal{M}} w_{h_{<t}}^v D_{\gamma}(P_v^{\pi} \| P_{\xi}^{\pi} \mid h_{<t})$$

$$\pi_{\text{IG}}^* \coloneqq \underset{\pi}{\operatorname{argmax}} V_{\text{IG}}^{\pi}$$

$$\lim_{t \rightarrow \infty} \frac{1}{\Gamma_t} \mathbb{E}_{\mu}^{\pi} \left[ D_{\gamma}(P_{\mu}^{\pi} \| P_{\xi}^{\pi} \mid h_{1:t}) \right] = 0 \quad (\text{on-policy})$$

$$\lim_{t \rightarrow \infty} \frac{1}{\Gamma_t} \mathbb{E}_{\mu}^{\pi_{\text{IG}}^*} \left[ \sup_{\pi \in \Pi(h_{1:t})} D_{\gamma}(P_{\mu}^{\pi} \| P_{\xi}^{\pi} \mid h_{1:t}) \right] = 0 \quad (\text{off-policy})$$

maximize knowledge / exploration=exploitation / resistant to noise / avoid traps

# Bayesian Agent

---

## Algorithm Bayesian Agent

---

**Require:** Model class  $\mathcal{M}$ ; prior  $w \in \Delta\mathcal{M}$ ; history  $\mathbf{æ}_{}.$

**function** ACT( $\pi$ )

    Sample and perform action  $a_t \sim \pi(\cdot | \mathbf{æ}_{})$

    Receive  $e_t \sim v(\cdot | \mathbf{æ}_{} a_t)$

**for**  $v \in \mathcal{M}$  **do**

$$w_v \leftarrow w_v \frac{v(e_{} | a_{})}{\xi(e_{} | a_{})}$$

**end for**

$t \leftarrow t + 1$

**end function**

---

# MDL

---

**Algorithm** MDL Agent

**Require:** Model class  $\mathcal{M}$ ; prior  $w \in \Delta\mathcal{M}$ ; regularizer constant  $\lambda \in \mathbb{R}^+$ .

**loop**

$$\sigma \leftarrow \arg \min_{\nu \in \mathcal{M}} \left[ K(\nu) - \lambda \sum_{k=1}^t \log \nu(e_k \mid \mathbf{æ}_{<k} a_k) \right]$$

ACT  $(\pi_\sigma^*)$

**end loop**

---

# MDL

## Definition (MDL)

$$\widehat{v} = \arg \min_{\nu \in \mathcal{M}} \{K_\nu(x) + K_w(\nu)\} = \arg \max_{\nu \in \mathcal{M}} \{w_\nu \nu(x)\}$$

where  $K_\nu(x) := -\log \nu(x)$  and  $K_w(\nu) := -\log w_\nu$

## Theorem (MDL Bound)

$$\sum_{t=1}^{\infty} \mathbb{E}_{\mu} \left[ \sum_{x_t \in \mathcal{X}} \left( \widehat{v}(x_t \mid x_{<t}) - \mu(x_t \mid x_{<t}) \right)^2 \right] \stackrel{+}{\leq} 8w_\mu^{-1}$$

MDL converges, but speed can be exponential worse than Bayes.

## Weak Asymptotic Optimality — Optimistic Agent

**Algorithm** Optimistic Agent  $\pi^\circ$

$$\pi_t^\circ := \operatorname{argmax}_\pi \max_{v \in \mathcal{M}_t} V_v^\pi(h_{1:t})$$

**Require:** Finite class of deterministic environments  $\mathcal{M}_0 = \mathcal{M}$

$t = 1$

**repeat**

$$(\pi^*, v^*) := \operatorname{argmax}_{\pi \in \Pi, v \in \mathcal{M}_{t-1}} V_v^\pi(h_{t-1})$$

**repeat**

$$a_t = \pi^*(h_{t-1})$$

Perceive  $e_t$  from environment  $\mu$

$$h_t \leftarrow h_{t-1} a_t e_t$$

Remove inconsistent environment  $\mathcal{M}_t := \{v \in \mathcal{M}_{t-1} : h_t^{\pi^\circ v} = h_t\}$

$$t \leftarrow t + 1$$

**until**  $v^* \notin \mathcal{M}_{t-1}$

**until**  $\mathcal{M} = \emptyset$

---

stochastic case:  $\mathcal{M}_t := \left\{v \in \mathcal{M}_{t-1} : v(e_{<t} | a_{<t}) \geq \varepsilon_t \max_{\rho \in \mathcal{M}} \rho(e_{<t} | a_{<t})\right\}$

Act optimally w.r.t. the most optimistic environment until contradicted.

If there is a chance: Try it! — Cheaper Exploration — Vulnerable to traps

# Asymptotic Optimality in Mean — Thompson Sampling

---

## Algorithm Thompson Sampling $\pi_T$

---

**Require:** Model class  $\mathcal{M}$ ; prior  $w \in \Delta\mathcal{M}$ ; exploration schedule  $(\varepsilon_t)_{t \in \mathbb{N}}$ .  
**loop**

    Sample  $\rho \sim w_{\mathcal{A}_{\leq t}}$

**for**  $i = 1 \rightarrow H_t(\varepsilon_t)$  **do**

        ACT  $(\pi_\rho^*)$

**end for**

**end loop**

---

## Theorem

If the discount function  $\gamma$  satisfies the discount assumption, the environment  $\mu$  is recoverable, then  $R_m(\pi_T, \mu) \in o(m)$ .

## Weak Asymptotic Optimality — BayesExp

---

**Algorithm** BayesExp  $\pi_{\text{BE}}$ 

---

**Require:** Model class  $\mathcal{M}$ ; prior  $w \in \Delta\mathcal{M}$ ; exploration schedule  $(\varepsilon_t)_{t \in \mathbb{N}}$ .

**loop**

```
if  $V_{\text{IG}}^*(\alpha_{<t}) > \varepsilon_t$  then
    for  $i = 1 \rightarrow H_t(\varepsilon_t)$  do
        ACT  $(\pi_{\text{IG}}^*)$ 
    end for
else
    ACT  $(\pi_\xi^*)$ 
end if
end loop
```

---

**$\varepsilon$ -optimal BayesExp:** If the optimal information gain value  $V_{\text{IG}}^* > \varepsilon_t$ , then execute the  $\varepsilon$ -optimal information gain policy  $\pi_{\text{IG}}^{\varepsilon_t}$  for  $H_t(\varepsilon_t)$  steps, else execute  $\pi_\xi^{\varepsilon_t}$  for 1 step.

- ▶ *BayesExp* performs phases of exploration in which it maximizes the expected information gain. This explores the environment class completely, even achieving off-policy prediction.
- ▶ In contrast, Thompson sampling only explores on the optimal policies, and in some environment classes this will not yield off-policy prediction. So in this sense the exploration mechanism of Thompson sampling is more reward-oriented than maximizing information gain.

# Strong Asymptotic Optimality — Inquisitive Agent

$$V_{\text{IG}}^{\pi}(h_{<t}) \coloneqq \mathbb{E}_{\xi}^{\pi} \left[ D(w_{h_{<t+m}} \| w_{h_{<t}}) \mid h_{<t} \right]$$

$$\pi_{\text{IG}}^{m,k} \coloneqq \underset{\pi \in \mathcal{A}^{\mathcal{H}^{<m}}}{\operatorname{argmax}} V_{\text{IG}}^{\pi}(h_{<t-k})$$

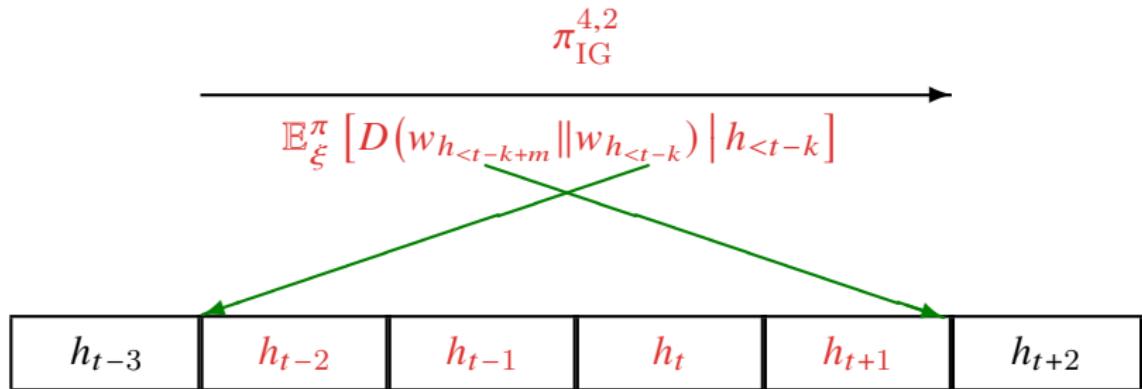
$$\rho(h_{<t}, m, k) \coloneqq \min \left\{ \frac{1}{m^2(m+1)}, \eta V_{\text{IG}}^{\pi_{\text{IG}}^{m,k}}(h_{<t-k}) \right\}$$

---

**Algorithm** Inquisitive Agent  $\pi^{\dagger}$ **while** True **do**    calculate  $\rho(h_{<t}, m, k)$  for all  $m$  and for all  $k < \min\{m, t\}$     ACT  $\pi_{\text{IG}}^{m,k}(h_{<t})$  with probability  $\rho(h_{<t}, m, k)$     ACT  $\pi_{\xi}^*(h_{<t})$  with probability  $1 - \sum_{m \in \mathbb{N}} \sum_{k < m, t} \rho(h_{<t}, m, k)$ **end while**

$$\pi^{\dagger}(a \mid h_{<t}) \coloneqq \sum_{m \in \mathbb{N}} \sum_{k < m, t} \rho(h_{<t}, m, k) \left[ \left[ a = \pi_{\text{IG}}^{m,k}(h_{<t}) \right] + \left( 1 - \sum_{m \in \mathbb{N}} \sum_{k < m, t} \rho(h_{<t}, m, k) \right) \right] \left[ \left[ a = \pi_{\xi}^*(h_{<t}) \right] \right]$$

# Strong Asymptotic Optimality — Inquisitive Agent



$$\pi^\dagger(a \mid h_{<t}) := \sum_{m \in \mathbb{N}} \sum_{k < m, t} \rho(h_{<t}, m, k) \llbracket a = \pi_{\text{IG}}^{m,k}(h_{<t}) \rrbracket + \left(1 - \sum_{m \in \mathbb{N}} \sum_{k < m, t} \rho(h_{<t}, m, k)\right) \llbracket a = \pi_{\xi}^*(h_{<t}) \rrbracket$$

# Approximation

The AIXI approximations are outperformed by DQN/DRQN...

- ▶ MC-AIXI-CTW.
  - ▶ Approximate Solomonoff induction — most recent actions and percepts (=context) more relevant — Context Tree Weighting
  - ▶ Sample paths in expectimax tree.
- ▶ Feature Reinforcement Learning ( $\Phi$ MDP). — history  $\mapsto$  state  
e.g. Classical physics: Position+velocity of objects=position at two time-slices. (2<sup>nd</sup> order Markov.)

$$\Phi : h \mapsto s$$

$$\Phi^{\text{best}} := \underset{\Phi}{\operatorname{argmin}} \text{Cost}(\Phi \mid h)$$

$$\text{Cost}(\Phi \mid h) := \text{cl}(s_{1:n}^\Phi \mid a_{1:n}) + \text{cl}(r_{1:n} \mid s_{1:n}^\Phi, a_{1:n}) + \text{cl}(\Phi)$$

How to find the map  $\Phi$ ? Monte-Carlo...

- ▶ Compress and Control. — (model-free)  
Combine induction and planning.

## Expectimax Approximation: MC-AIXI-CTW

Upper Confidence Tree (UCT) algorithm:

- ▶ **Sample** observations from Context Tree Weighting (CTW) distribution.

$$\text{CTW}(e_{<t} \mid a_{<t}) := \sum_{\Gamma} 2^{-\text{cl}(\Gamma)} \Gamma(e_{<t} \mid a_{<t})$$

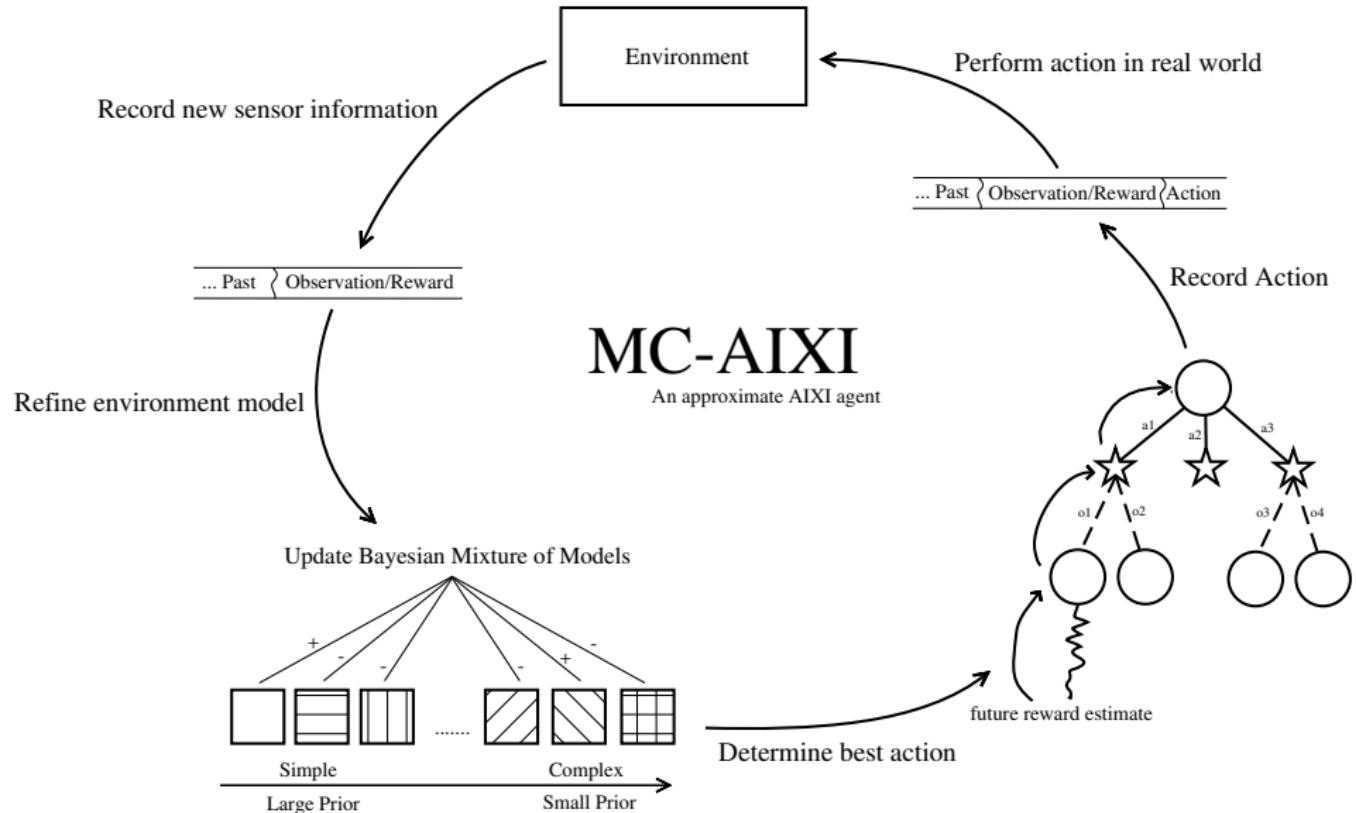
- ▶ Select actions with highest upper confidence bound.

$$a_{\text{ucb}} := \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left( \underbrace{\hat{Q}(\alpha_{<t} a)}_{\text{average}} + \sqrt{\underbrace{\frac{\log T(\alpha_{<t})}{T(\alpha_{<t} a)}}_{\text{exploration bonus}}} \right)$$

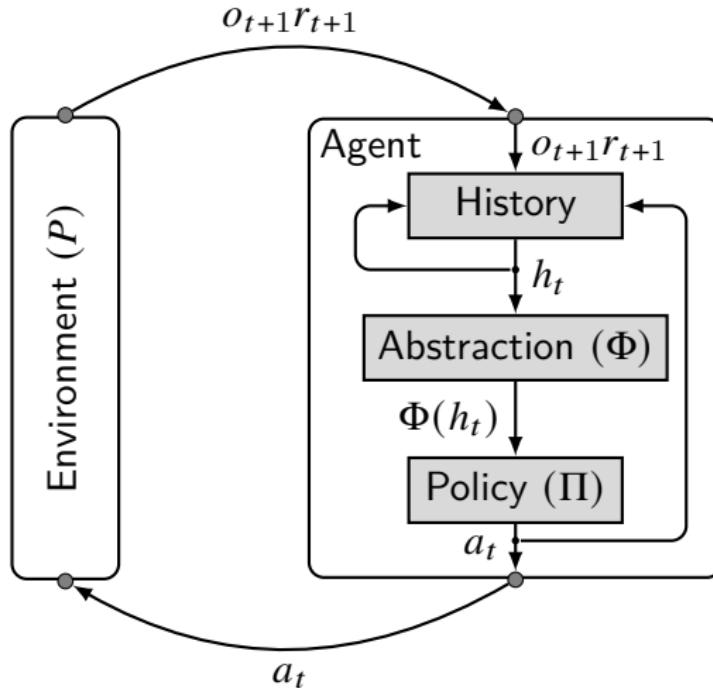
where  $T(\cdot)$  is the number of times a sequence has been visited.

- ▶ **Expand** tree by one leaf node (per trajectory).
- ▶ **Simulate** from leaf node further down using (fixed) playout policy.
- ▶ **Propagate back** the value estimates for each node.

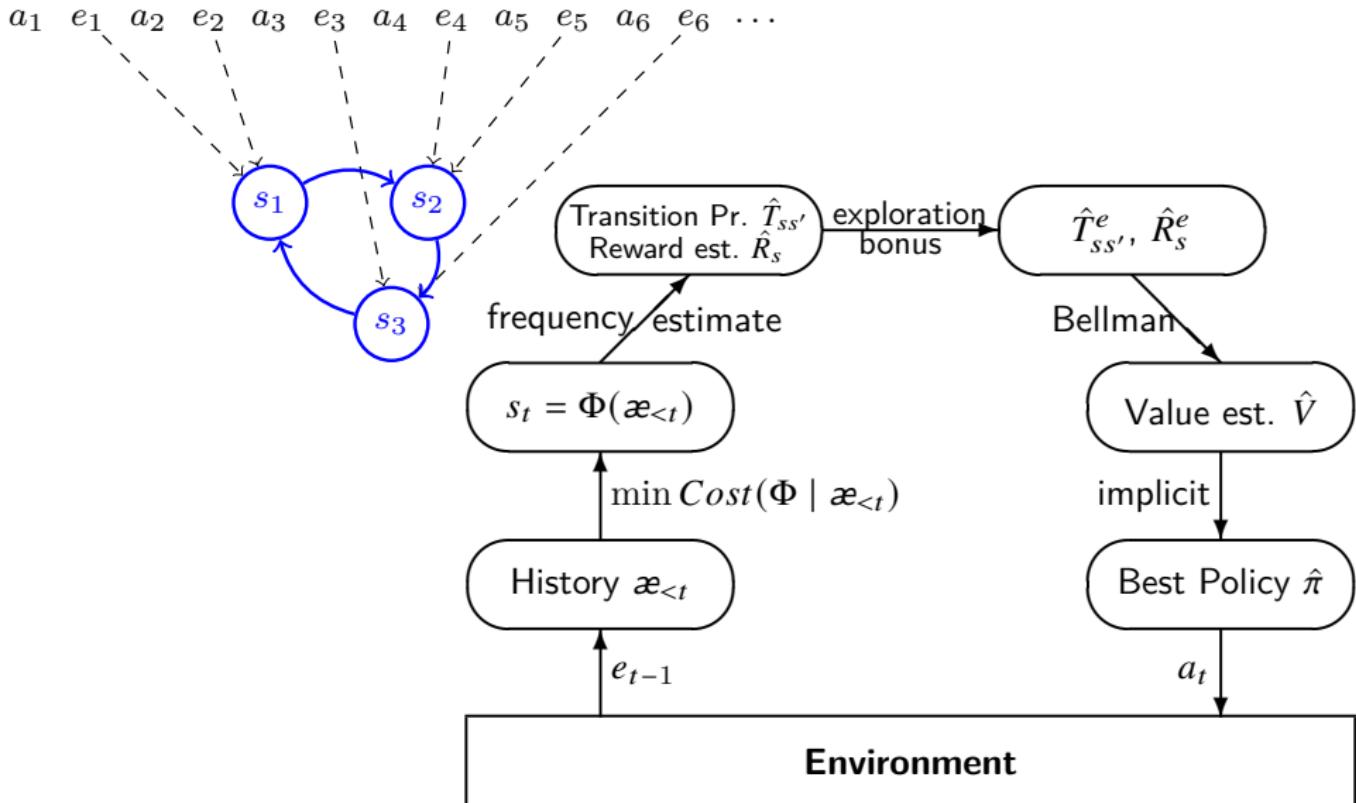
# MC-AIXI-CTW



# Feature Reinforcement Learning



# Feature Reinforcement Learning ( $\Phi$ MDP)



# Contents

Introduction	Reinforcement Learning
Philosophy of Induction	Deep Learning
Inductive Logic	Artificial General Intelligence
Universal Induction	AIXI
Causal Inference	Leibniz
Game Theory	Variants of AIXI
	Universal Search
	Gödel Machine & Consciousness
	What If Computers Could Think?
	References 1753

# Universal Search

- ▶ Levin Search
- ▶ Speed Prior
- ▶ Hutter Search
- ▶ AIXI<sup>tℓ</sup>
- ▶ Optimal Ordered Problem Solver
- ▶ Gödel Machine



Figure: Levin

## Levin Search (LSEARCH)

An inversion algorithm  $p$  inverts a function  $f$  if given  $x$ ,  $p(x) = y$  s.t.  $f(y) = x$ .

### LSEARCH

Run all  $\{p : \ell(p) \leq i\}$  for  $2^{i-\ell(p)}$  steps in phase  $i = 1, 2, 3, \dots$  until it has inverted  $f$  on  $x$ .

$$Kt(x) := \min_p \{\ell(p) + \log t(p, x) : U(p) = x\}$$

### Theorem

All strings  $\{x : Kt(x) \leq n\}$  can be generated and tested in  $2^{n+1}$  steps.

$$t_{\text{LSEARCH}}(x) = O\left(2^{K(n)} t_{p_n}^+(x)\right)$$

where  $t_{p_n}^+(x)$  is the runtime of  $p_n(x)$  plus the time to verify the correctness of the result  $f(p_n(x)) = x$ .

**Remark:** If P=NP, then LSEARCH is a P algorithm for every NP problem.

## Speed Prior

$$S(e_{<t} \mid a_{<t}) := \sum_{p:U(p,a_{<t})=e_{<t}} \frac{2^{-\ell(p)}}{t(p, a_{<t}, e_{<t})}$$

$S$  is computable.

A function  $f$  is estimable in polynomial time iff there is a function  $g$  computable in polynomial time s.t.  $f \asymp g$ .

For any measure  $\mu$  estimable in polynomial time,

$$\left( \sqrt{L_n^{\Lambda_S}} - \sqrt{L_n^{\Lambda_\mu}} \right)^2 \leq 2D_n(\mu \| S) = O(\log n)$$

## On-Policy Value Convergence

If the effective horizon is bounded, then for any environment  $\mu \in \mathcal{M}_{\text{comp}}$  estimable in polynomial time and any policy  $\pi$ ,

$$P_\mu^\pi \left( \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{k=1}^t \left( V_S^\pi(h_{<k}) - V_\mu^\pi(h_{<k}) \right) \right) = 0$$

# Hutter Search (HSEARCH)

$M_{p^*}^\varepsilon(x)$

Initialize the shared variables  
 $L := \{\}$ ,  $t_{\text{fast}} := \infty$ ,  $p_{\text{fast}} := p^*$ .  
Start algorithms  $A$ ,  $B$ , and  $C$  in parallel with  $\varepsilon$ ,  $\varepsilon$ , and  $1 - 2\varepsilon$  computation time, respectively.

$B$

for  $(p, t) \in L$   
run  $t(x)$  in parallel for all  $t$  with computation time  $2^{-\ell(p)-\ell(t)}$ .  
if for some  $t$ ,  $t(x) < t_{\text{fast}}$ ,  
then  $t_{\text{fast}} := t(x)$  and  $p_{\text{fast}} := p$ .

continue

$A$

for  $i := 1, 2, 3, \dots$  do  
pick the last wff of the  $i^{\text{th}}$  proof.  
if it reads “ $p(\cdot)$  is equivalent to  $p^*(\cdot)$  and has time-bound  $t(\cdot)$ ”,  
then add  $(p, t)$  to  $L$ .

continue

$C$

for  $k := 1, 2, 4, 8, \dots$  do  
run current  $p_{\text{fast}}$  for  $k$  steps.  
if  $p_{\text{fast}}$  halts,  
then print result  $p_{\text{fast}}(x)$  and abort  $A$ ,  $B$  and  $C$ .

continue

1. Let  $P := \emptyset$ . This will be the set of verified programs.
2. For all proofs of length  $\leq n$ : if the prover shows  $\text{VA}(p)$  for some  $p$  with  $\ell(p) \leq l$ , then add  $p$  to  $P$ .

$$\text{VA}(p) := \text{"}\forall k \forall (va' \alpha)_{1:k} : p(\alpha_{<k}) = v_1 a'_1 \dots v_k a'_k \implies v_k \leq V_\xi^\pi(\alpha_{<k})\text{"}$$

(The program  $p$  not only computes future actions of  $\pi$ , which is the policy derived from  $p$  according to  $\pi(\alpha_{<k}) := a'_k$ , but also hypothetical past actions  $a'_i$  and lower bounds  $v_i$  for the value of the policy  $\pi$ .)

3. For each input history  $\alpha_{<k}$  repeat: run all programs from  $P$  for  $\leq t$  steps each, take the one with the highest promised value  $v_k$ , and return that program's policy's action.

- ▶ AIXI<sup>tl</sup> depends on  $t, l, n$  but not on knowing  $p$ .
- ▶ Its setup-time is  $t_{\text{setup}}(p^{\text{best}}) = O(n \cdot 2^n)$ .
- ▶ Its computation time per cycle is  $t_{\text{cycle}}(p^{\text{best}}) = O(t \cdot 2^l)$ .

# Schmidhuber's Optimal Ordered Problem Solver (OOPS)

- ▶ Solve the first task with LSEARCH.
- ▶ Freeze successful programs in non-writable memory.
- ▶ Programs tested during search for later tasks may copy non-writable code into separate modifiable storage, to edit it and execute the modified result.
- ▶ Given a new task, OOPS spends half of the time to test programs that have the most recent successful program as a prefix, the other half to fresh programs.
- ▶ Time is allocated according to a distribution over programs, which is obtained by multiplying the probabilities of the individual instructions.

Incremental Learning

# Contents

Introduction

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

Deep Learning

Artificial General Intelligence

AIXI

Leibniz

Variants of AIXI

Universal Search

Gödel Machine & Consciousness

What If Computers Could Think?

References 1753

# Self-Reference

- ▶ This sentence repeats the word 'twice' twice.
- ▶ Thare are five mistukes im this centence.
- ▶ **The only boldface sentence on this page is false.**
- ▶ All generalizations are wrong.
- ▶ Every rule has an exception except this one.
- ▶ Moderation in all things, including moderation.
- ▶ We must believe in free will — we have no choice!
- ▶ I know that I know nothing.
- ▶ There are two rules lor success in life:
  1. Never tell anyone all that you know.
- ▶ If you choose an answer to this question at random, what is the chance you will be correct? (A) 25% (B) 50% (C) 0% (D) 25%
- ▶
  1. What is the best question to ask and what is the answer to it?
  2. The best question is the one you asked; the answer is the one I gave.
- ▶ Can you answer the following question in the same way to this one?
- ▶ One of the lessons of history is that no one ever learns the lessons of history.
- ▶ 涅槃是消除了一切欲望后才能抵达的境界, 包括对涅槃的欲望.



# Self-Reference vs Paradox

The sentence below is false.



The sentence above is true.

## Yablo Paradox

- ▶  $A_1$ : for all  $k > 1$ ,  $A_k$  is false.
- ▶  $A_2$ : for all  $k > 2$ ,  $A_k$  is false.
- ▶  $A_3$ : for all  $k > 3$ ,  $A_k$  is false.
- ▶ ...

## Quine Paradox

“Yields falsehood when preceded by its quotation” yields falsehood when preceded by its quotation.

self-reference / circularity or infinite regress / negation / infinity / totality

# ChatGPT-o1 2024

How many words are in your response to this prompt?



Thought for 10 seconds ^

## Addressing paradoxical query

I'm working through a paradoxical question, which involves self-reference and determining the response length. Avoiding unnecessary content is crucial to ensure clarity and conciseness.

## Figuring out word count

OK, let me see. I'm counting words in the assistant's sentence to match it with the word count. This approach seems interesting.

## Identifying word patterns

I'm examining sentences with varying word counts, finding inconsistencies in some while confirming others. This helps determine the most accurate way to count words.

## Counting words

Okay, let me see. The sentence spans five words. Here's a concise tally: "This sentence has five words. Is 5 words. So the assistant can answer with "There are seven words."

## Clarifying response accuracy

I'm analyzing various ways to specify the word count in responses, highlighting the need for consistent and clear communication. Progressing towards enhancing precise clarity.

There are seven words in this sentence.

# The “Power” of Self-Reference

## Curry's Paradox

- ▶ If this sentence is true, then God exists.
- ▶ This sentence is false, and God does not exist.

1. At least one of these two sentences is false.
2. God does not exists.

Hi 美女, 问你个问题呗

如果我问你“你能做我女朋友吗”, 那么你的答案和这个问题的答案是一样的吗?

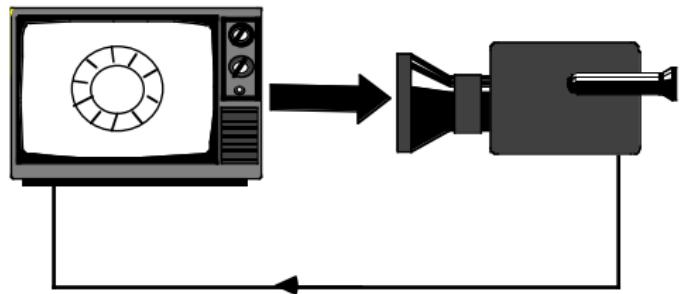
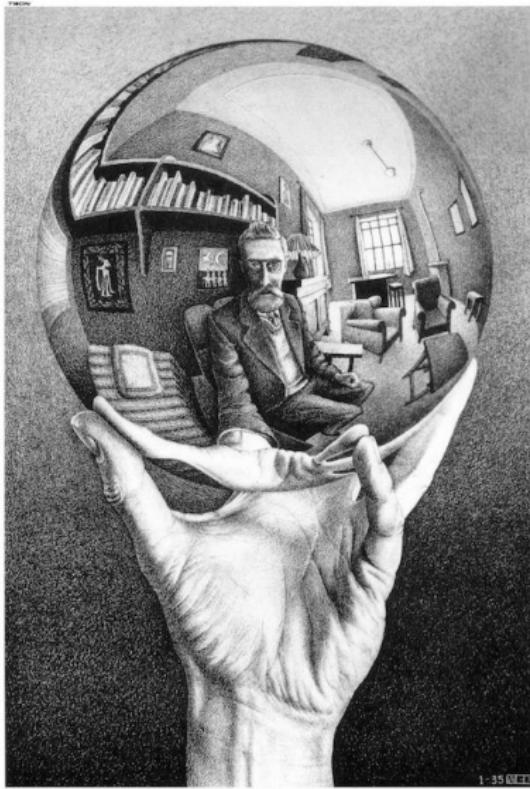
自我修复/自我实现?

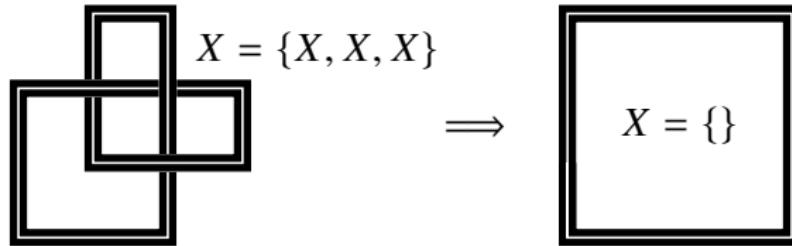
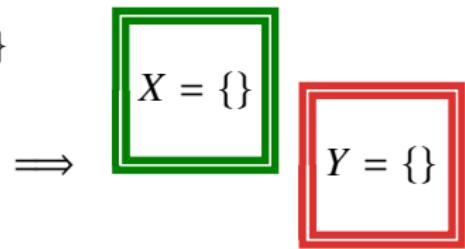
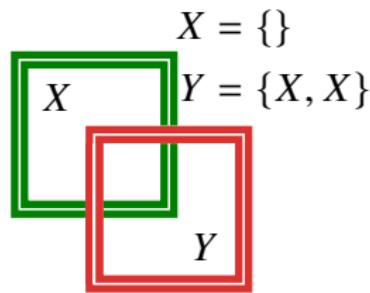
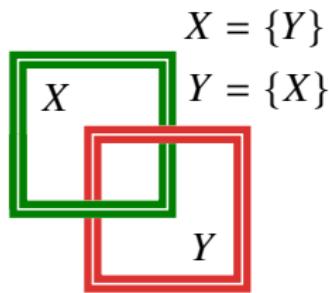
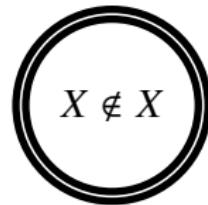
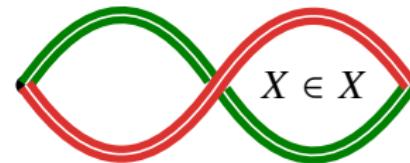
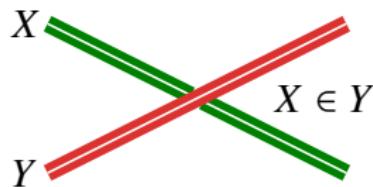
- ▶ “This sentence has \_\_\_\_\_ letters.”      thirty-one / thirty-three
- ▶ 这句话有 2 个 ‘这’ 字, 2 个 ‘句’ 字, 2 个 ‘话’ 字, 2 个 ‘有’ 字, 7 个 ‘2’ 字, 11 个 ‘个’ 字, 11 个 ‘字’ 字, 2 个 ‘7’ 字, 3 个 ‘11’ 字, 2 个 ‘3’ 字.

## How to Refer? — Levels

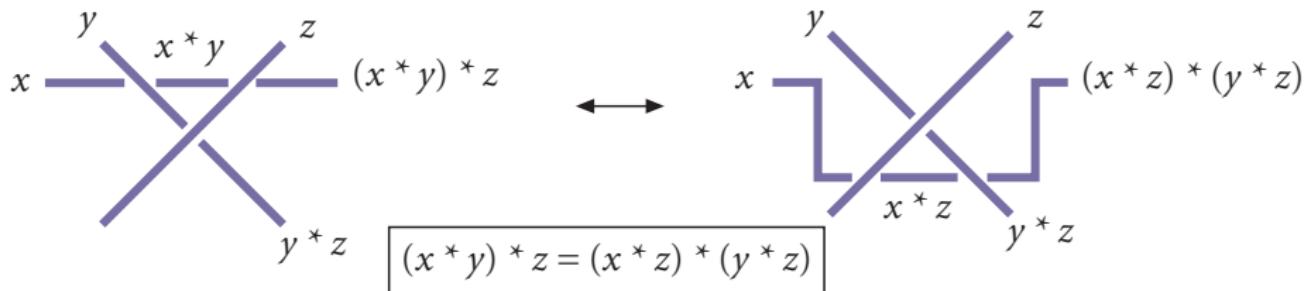
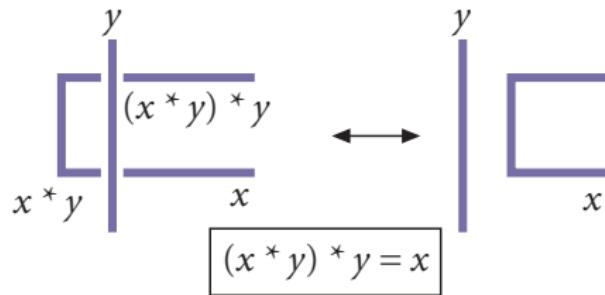
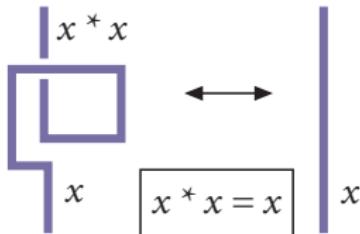


# How to Refer?

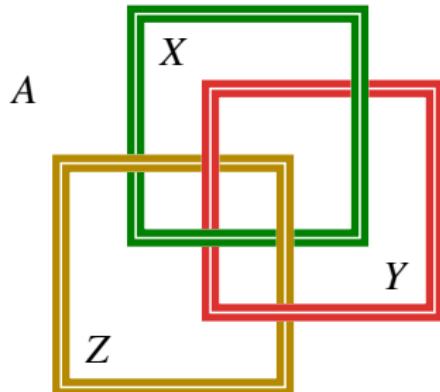




# Reidemeister Moves



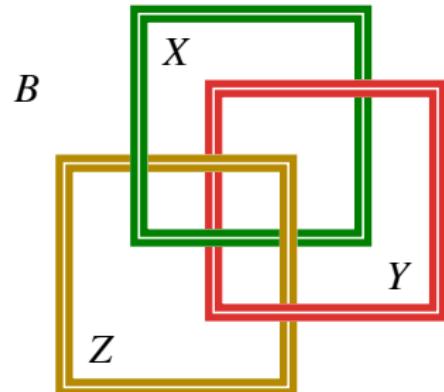
# Self-Reference & IIT



$$X = \{Y, Y\}$$

$$Y = \{Z, Z\}$$

$$Z = \{X, X\}$$



$$X = \{Y, Z\}$$

$$Y = \{X, Z\}$$

$$Z = \{X, Y\}$$

Figure:  $\Phi(A) < \Phi(B)$ ?

## Larger Domain

1, 1, 2, 3, 5, 8, 13, 21, 34, ...

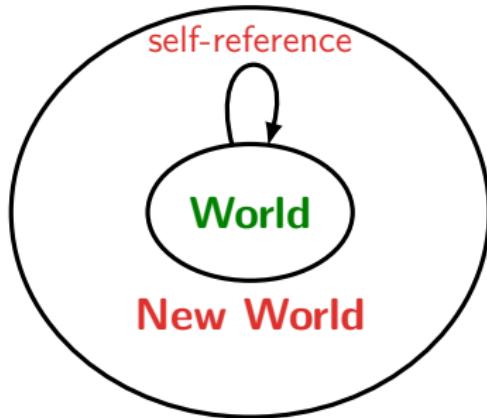
$$F_0 = F_1 = 1; F_{n+1} = F_n + F_{n-1}$$

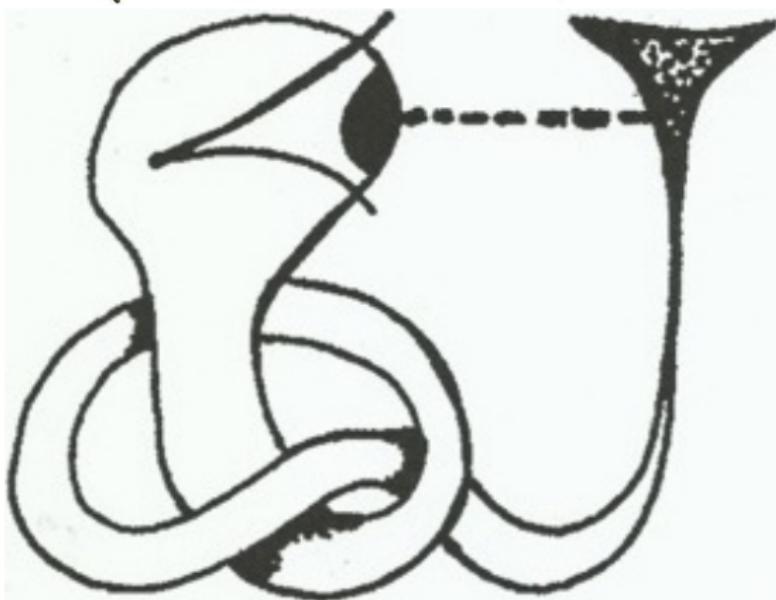
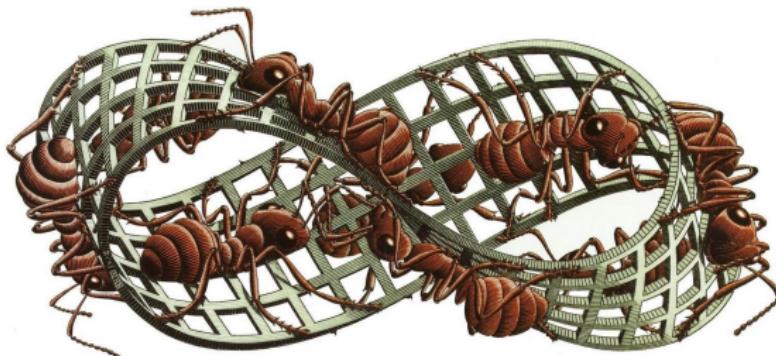
$$\lim_{n \rightarrow \infty} \frac{F_{n+1}}{F_n}$$

$$\frac{F_{n+1}}{F_n} = 1 + \frac{1}{\frac{F_n}{F_{n-1}}}$$

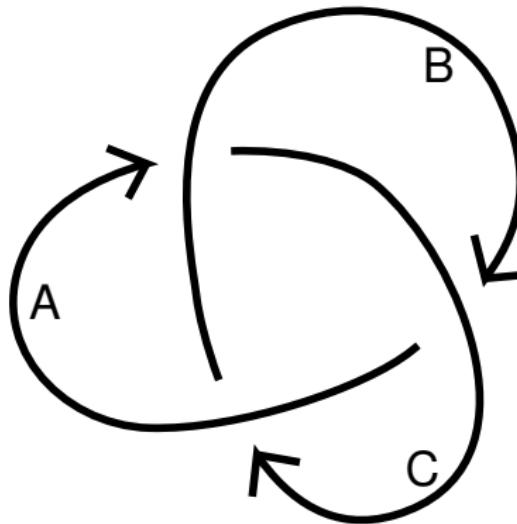
$$f(x) = 1 + \frac{1}{x} = x \implies 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{\ddots}}}} = \frac{1 + \sqrt{5}}{2}$$

$$f(x) = \frac{-1}{x} = x \implies x = i$$





# Trefoil



- ▶ objects  $\{A, B, C\}$
- ▶ morphisms
  - A:  $C \rightarrow B$
  - B:  $A \rightarrow C$
  - C:  $B \rightarrow A$

# Nested Virtualization?



从前有座山，山里有座庙，庙里有个老和尚在讲故事：从前有座山...

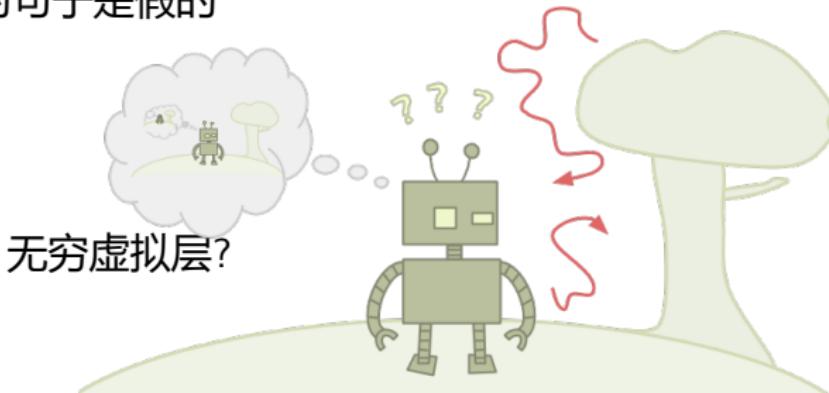
$$\begin{cases} FX = Y \\ GY = X \end{cases}$$

$$X = GFGFGFGF \dots$$

$$Y = FGFGFGFG \dots$$

# Liar Paradox vs Quine Paradox

1. 这句话是假的
2. “这句话是假的” 是假的
3. “““““.....是假的” 是假的” 是假的” 是假的” 是假的” 是假的”
4. 把 “把中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变‘’ 中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变
5. 把 “把中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的” 中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的



## How to Refer? — Encoding



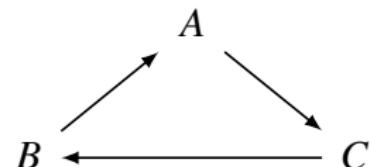
- ▶ 100 prisoners are lined up by an jailer, who places a red or blue hat upon each of their heads.
- ▶ The prisoners can see the hats of the people lined up in front of them, but they can't look at the hats behind them, or at their own.
- ▶ The jailer is going to ask color of each prisoner's hat starting from the last prisoner in queue. If a prisoner tells the correct color, then is saved, otherwise executed.
- ▶ How many prisoners can be saved at most if they are allowed to discuss a strategy before the jailer starts asking colors of their hats?

If the first person sees an **odd** number of red hats he calls out red, if he sees an **even** number of red hats he calls out blue.

手扶拐杖的外星绅士造访地球。临别，人类赠送百科全书：“人类文明尽在其中！”。  
绅士谢绝：“不，谢谢！我只需在拐杖上点上一点”。

# What is the Next Number?

1. 1
  2. 11
  3. 21
  4. 1211
  5. 111221
  6. 312211
  7. ?
- A. 11131221131211132221...
- B. 3113112221131112311332...
- C. 132113213221133112132123...



# Diagonalization[Law69]<sup>22</sup>

## Definition (Point-Surjective)

A morphism  $f : X \rightarrow Y$  is *point-surjective* iff for every  $y : 1 \rightarrow Y$ , there is an  $x : 1 \rightarrow X$  s.t.  $y = f \circ x$ .

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ x \uparrow & \nearrow y & \\ 1 & & \end{array}$$

## Definition (Weakly Point-Surjective)

A morphism  $f : X \times Y \rightarrow Z$  is *weakly point-surjective* iff for every  $g : X \rightarrow Z$ , there exists  $y : 1 \rightarrow Y$  such that, for all  $x : 1 \rightarrow X$ :

$$g \circ x = f \circ \langle x, y \rangle$$

$$\begin{array}{ccc} X \times Y & \xrightarrow{f} & Z \\ \langle x, y \rangle \uparrow & & \uparrow g \\ 1 & \xrightarrow{x} & X \end{array}$$

## Theorem (Lawvere's Fixpoint Theorem)

Let  $\mathbf{C}$  be a category with a terminal object and binary products. If  $f : X \times X \rightarrow Y$  is weakly point-surjective, then every  $\alpha : Y \rightarrow Y$  has a fixpoint  $y : 1 \rightarrow Y$ .

$$\begin{array}{ccc} X \times X & \xrightarrow{f} & Y \\ \Delta \uparrow & & \downarrow \alpha \\ X & \xrightarrow{g} & Y \end{array}$$

<sup>22</sup>Lawvere: Diagonal arguments and cartesian closed categories.

Yanofsky: A universal approach to self-referential paradoxes, incompleteness and fixed points.

# Lawvere's Fixpoint Theorem

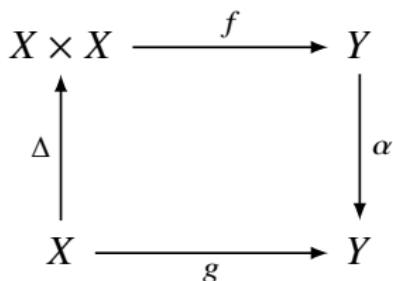
- A function  $g : X \rightarrow Y$  is *representable* by  $f : X \times X \rightarrow Y$  iff

$$\exists y \forall x : g(x) = f(x, y)$$

## Theorem (Lawvere's Fixpoint Theorem)

For sets  $X, Y$ , functions  $f : X \times X \rightarrow Y$ ,  $\alpha : Y \rightarrow Y$ , let  $g := \alpha \circ f \circ \Delta$ .

1. If  $\alpha$  has no fixpoint, then  $g$  is not representable by  $f$ .
2. If  $g$  is representable by  $f$ , then  $\alpha$  has a fixpoint.



$$\alpha(f(\lceil g \rceil, \lceil g \rceil)) = g(\lceil g \rceil) = f(\lceil g \rceil, \lceil g \rceil)$$

- $\Delta : x \mapsto \langle x, x \rangle$  diagonal
- $f$  evaluation
- $\alpha$  “negation”
- $g (\lceil g \rceil)$  fixpoint-(free) transcendence
- $f (\lceil g \rceil, \lceil g \rceil)$  self-reference  
“I have property  $\alpha$ .”

# Lawvere's Fixpoint Theorem

$f$	0	1	2	3	...	$t$	...
0	$\alpha f(0, 0)$	...	...	...	...	$f(0, t)$	...
1	...	$\alpha f(1, 1)$	...	...	...	$f(1, t)$	...
2	...	...	$\alpha f(2, 2)$	...	...	$f(2, t)$	...
3	...	...	...	$\alpha f(3, 3)$	...	$f(3, t)$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t$	...	...	...	...	...	$f(t, t)$    $\alpha f(t, t)$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

## Example — Kleene's Fixpoint Theorem

### Theorem (Kleene's Fixpoint Theorem)

Given a recursive function  $h$ , there is an index  $e$  s.t.

$$\varphi_e = \varphi_{h(e)}$$

$$\begin{array}{ccc} \mathbb{N} \times \mathbb{N} & \xrightarrow{f} & \{\varphi_n\}_{n \in \mathbb{N}} \\ \Delta \uparrow & & \downarrow \alpha_h \\ \mathbb{N} & \xrightarrow{g} & \{\varphi_n\}_{n \in \mathbb{N}} \end{array}$$

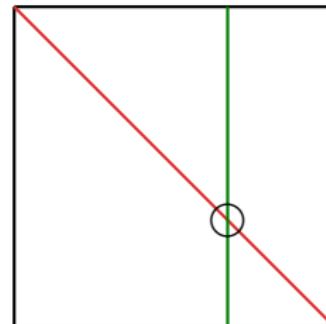
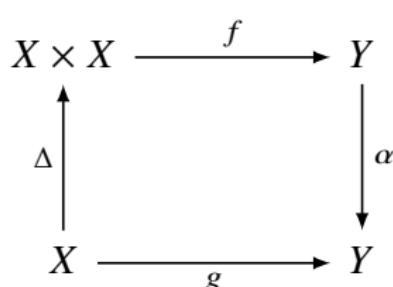
where  $f : (m, n) \mapsto \varphi_{\varphi_n(m)}$ , and  $\alpha_h : \varphi_n \mapsto \varphi_{h(n)}$ .

The function  $g : m \mapsto \varphi_{h(\varphi_m(m))}$  is a recursive sequence of partial recursive functions, and thus is representable by  $f(-, t)$ .

$$e := \varphi_t(t)$$

Explicitly,  $g(m) = \varphi_{h(\varphi_m(m))} = \varphi_{s(m)} = \varphi_{\varphi_t(m)} = f(m, t)$

# Fixpoint vs Diagonalization



Curry Y	$\hat{=}$	$\lambda$ -fixpoint	$\hat{=}$	Gödel	$\hat{=}$	Kleene	$\hat{=}$	Russell
$yx$	$\hat{=}$	$N(\Gamma M^\top)$	$\hat{=}$	$N(\Gamma M(x)^\top)$	$\hat{=}$	$\varphi_n(m)$	$\hat{=}$	$x \in y$
$xx$	$\hat{=}$	$M(\Gamma M^\top)$	$\hat{=}$	$M(\Gamma M(x)^\top)$	$\hat{=}$	$\varphi_n(n)$	$\hat{=}$	$x \in x$
$y(xx)$	$\hat{=}$	$F\Gamma M\Gamma M^\top$	$\hat{=}$	$F(\Gamma M(\Gamma M(x)^\top)^\top)$	$\hat{=}$	$h(\varphi_n(n))$	$\hat{=}$	$x \notin x$
$\lambda x.y(xx)$	$\hat{=}$	$G$	$\hat{=}$	$G(x)$	$\hat{=}$	$\varphi_t(n)$	$\hat{=}$	$x \notin R$
$(\lambda x.y(xx))(\lambda x.y(xx))$	$\hat{=}$	$G(\Gamma G^\top)$	$\hat{=}$	$G(\Gamma G(x)^\top)$	$\hat{=}$	$\varphi_t(t)$	$\hat{=}$	$R \notin R$

self-reference  $\xrightarrow{?}$  self-improvement

# Kleene's Fixpoint Theorem



## Theorem (Second Recursion Theorem)

If  $f(x, y)$  is a partial recursive function, there is an index  $e$  s.t.

$$\varphi_e(y) = f(e, y)$$

**Remark:** 对于任意的程序  $h$ , 总存在某个程序  $e$ , 执行程序  $e$  的结果等价于把程序  $e$  当作数据输入给程序  $h$  执行的结果  $\llbracket e \rrbracket(-) = \llbracket h \rrbracket(e, -)$ .

## Theorem (Kleene's Fixpoint Theorem)

Given a recursive function  $h$ , there is an index  $e$  s.t.

$$\varphi_e = \varphi_{h(e)}$$

**Remark:** You can systematically change an infinite number of programs  $n \mapsto h(n)$  but you cannot systematically change an infinite number of recursive functions  $\varphi_e = \varphi_{h(e)}$ .

# From Kleene's Fixpoint to Chaitin's Incompleteness

**Definition:** Kolmogorov Complexity  $K(x) := \mu e[\varphi_e(0) = x]$

**Theorem (Chaitin's Incompleteness Theorem)**

For any arithmetically sound Gödelian theory  $T$ ,  $\exists c \forall x : T \not\vdash K(x) > c$ .

**Proof.**

For any  $m$ , we can construct:

$$M_n := \text{"find } \mu y [\text{prf}_T(y, K(x) > m)], \text{output } x\text{"}$$

So there exists a recursive function  $f : m \mapsto n$ .

By Kleene's fixpoint theorem, there exists  $e$  such that

$$M_e = M_{f(e)} = \text{"find } \mu y [\text{prf}_T(y, K(x) > e)], \text{output } x\text{"}$$

Take  $c := e$ .

□

**Remark:** For almost all random strings their randomness cannot be proved.

# Self-Reproducing Program/Quine

There is a program that outputs its own length.

There is a program that outputs its own source code.

- ▶ A Quine is a program which takes no input and outputs its own source code.
- ▶ Quines are algorithmic random.

## Corollary (Self-Reproducing Program)

*There is a recursive function  $\varphi_e$  s.t.  $\forall x : \varphi_e(x) = e$ .*

## Quine in Python

```
s='s=%r; print(s%%s)'; print(s%s)
```

## Quine in Lambda Calculus

$$(\lambda x.xx)(\lambda x.xx)$$

# Self-Reproducing Program

*Print two copies of the following, the second copy in quotes:*

*“Print two copies of the following, the second copy in quotes:”*

DNA / mutation / evolution

*Build a baby that acts on the following instructions, and also contains a copy of those instructions in its reproductive organs.*

*“Build a baby that acts on the following instructions, and also contains a copy of those instructions in its reproductive organs.”*

## von Neumann's Self-Reproducing Automata

1. A universal constructor  $A$ .

$$A + \lceil X \rceil \rightsquigarrow X$$

2. A copying machine  $B$ .

$$B + \lceil X \rceil \rightsquigarrow \lceil X \rceil$$

3. A control machine  $C$ , which first activates  $B$ , then  $A$ .

$$A + B + C + \lceil X \rceil \rightsquigarrow X + \lceil X \rceil$$

4. Let  $X := A + B + C$ . Then  $A + B + C + \lceil A + B + C \rceil$  is **self-reproducing**.

$$A + B + C + \lceil A + B + C \rceil \rightsquigarrow A + B + C + \lceil A + B + C \rceil$$

5. It is possible to add the description of any machine  $D$ .

$$A + B + C + \lceil A + B + C + D \rceil \rightsquigarrow A + B + C + D + \lceil A + B + C + D \rceil$$

6. Now allow mutation on the description  $\lceil A + B + C + D \rceil$ .

$$A + B + C + \lceil A + B + C + D' \rceil \rightsquigarrow A + B + C + D' + \lceil A + B + C + D' \rceil$$

# Introspective Program

## Definition ( $\psi$ -introspective)

Given a total recursive function  $\psi$ ,

- ▶ the  $\psi$ -analysis of  $\varphi(x)$  is the code of the computation of  $\varphi(x)$  to  $\psi(x)$  steps.
- ▶  $\varphi$  is  $\psi$ -introspective at  $x$  iff  $\varphi(x) \downarrow$  and outputs its own  $\psi$ -analysis.
- ▶  $\varphi$  is totally  $\psi$ -introspective iff it is  $\psi$ -introspective at all  $x$ .

## Corollary

*There is a program that is totally  $\psi$ -introspective.*

## Proof.

Let  $f(n, x) :=$  “the  $\psi$ -analysis of  $\varphi_n(x)$ ”.

□

# Introspective Program

There is a program that is totally introspective.

$$\varphi_e = \varphi_{h(e)}$$

Self-simulating Computer	Self-consciousness
Host Machine	Experiencing Self
Virtual Machine	Remembering Self
Hardware	Body



## Know Thyself

# Who am I?

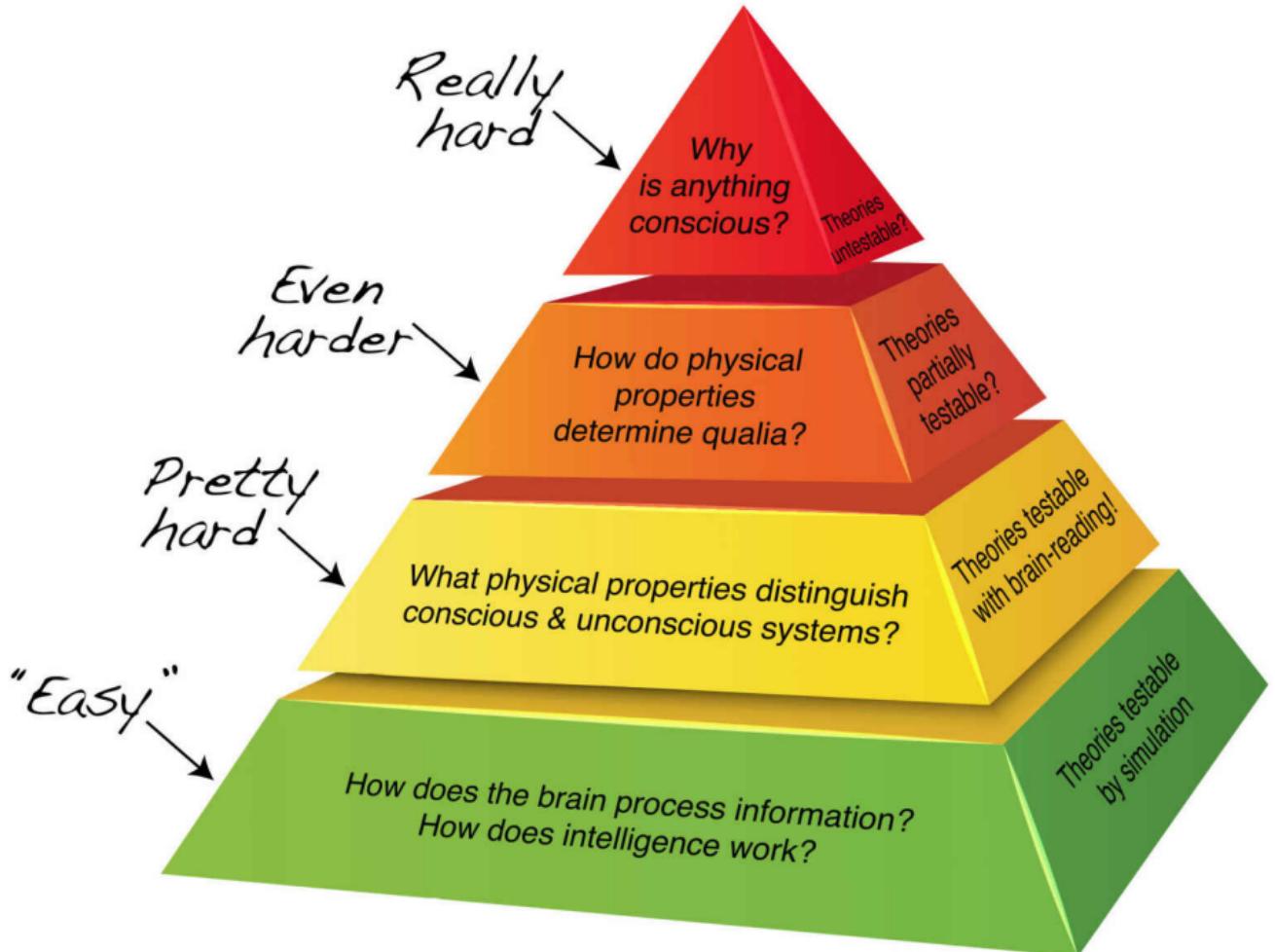
I think, therefore I am.

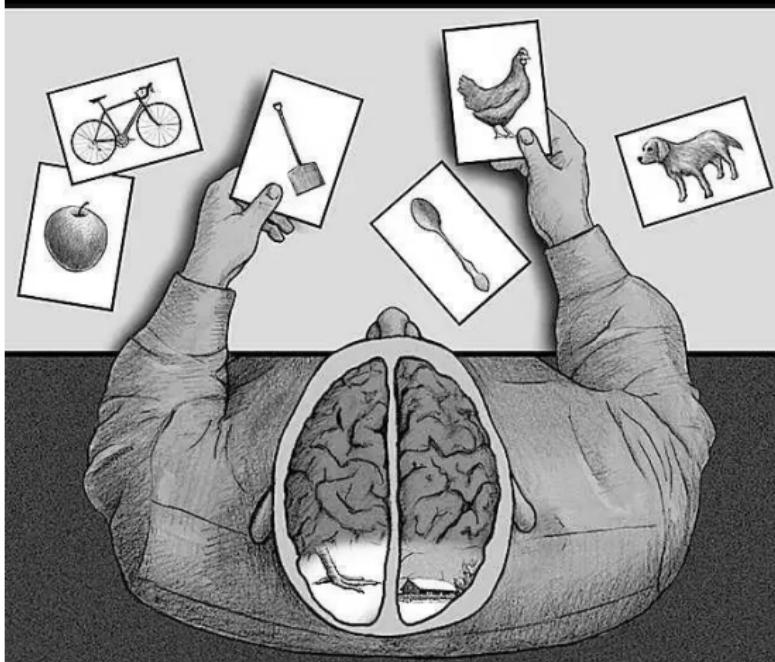
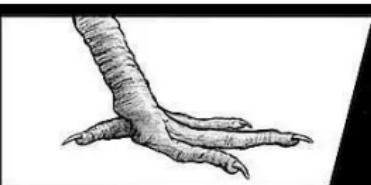
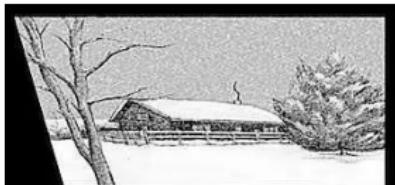
self-locating: “I” is an indexical term that I use to refer to myself as myself.

What is “me”?

What is “self-consciousness”?

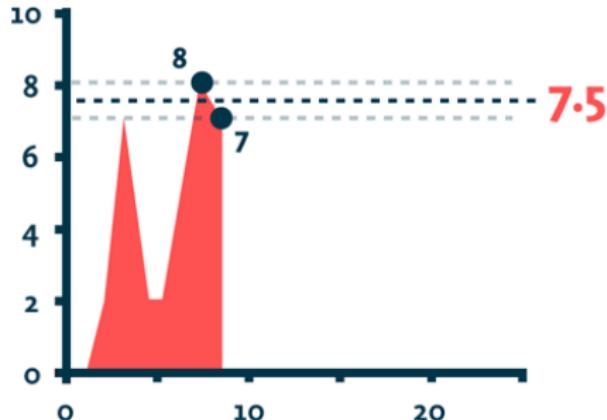
- ▶ self-perception self-observation self-experience self-tracking  
self-reflection self-awareness
- ▶ self-evaluation self-analysis self-monitoring
- ▶ self-control self-adjustment self-modification self-actualization  
self-fulfillment self-surpass self-improvement
- ▶ *actual-self* pk *ideal-self* self-identity “the *self*”
- ▶ free will: Second order desire that we want to act on is second order volition. Second order volitions involve wanting a certain desire to be one's will, that is wanting it to move one to action. (Frankfurt)



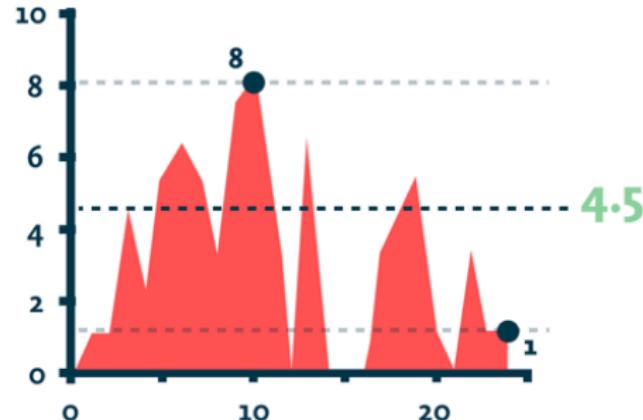


- ▶ the split brain in man
- ▶ snow?
- ▶ shit!
- ▶ life as a story

# Kahneman — Thinking, Fast and Slow



7.5



4.5

Figure: Why you might prefer more pain

- ▶ painful experiment
- ▶ experiencing self
- ▶ remembering self
- ▶ duration neglect
- ▶ peak-end rule



**Figure:** One can imagine a detailed floor plan of a room, sitting on a table in the room; this plan has an image of the table on which there is an image of the plan itself. Now introduce the dynamical aspect: the items on the plan are cut out from paper and can be moved to try a different furniture arrangement; in this way the plan models possible states of the world about which it carries information.

## Manin — Cognitive Networks



The brain contains inside a map of itself, and some neural information channels in the central neural system:

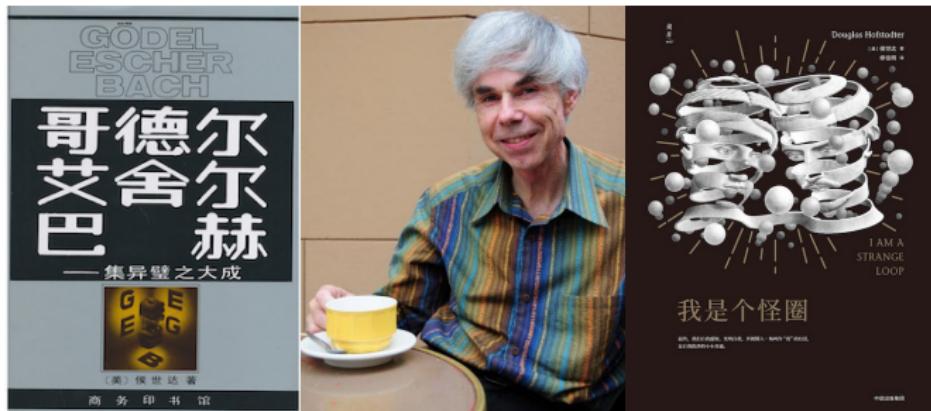
- ▶ carry information about the mind itself, i.e. are **reflexive**;
- ▶ are capable of modelling states of the mind different from the current one, i.e. possess a **modelling function**;
- ▶ can influence the state of the whole mind and through that, the behavior, i.e. possess **controlling function**.

The reflection of the brain inside itself must be **coarse grained**.



# 侯士达 — 《我是个怪圈》

- ▶ 有没有意识取决于在哪个层级上对结构进行观察。在整合度最高的层级上看，大脑是有意识的。下降到微观粒子层面，意识就不见了。
- ▶ 意识体是那些在某个描述层级上表现出某种特定类型的循环回路的结构。当一个系统能把外部世界过滤成不同的范畴、并不断向越来越抽象的层级创造新的范畴时，这种循环回路就会逐渐形成。
- ▶ 当系统能进行自我表征 — 对自己讲故事 — 的时候，这种循环回路就逐渐变成了实体的“我” — 一个统一的因果主体。



说谎者悖论	<b>我在说谎</b>
Grelling 悖论	“非自谓的”是自谓的吗
Russell 悖论	“不属于自身的集合的集合”属于自身吗
Berry 悖论	我是少于十八个字不可定义的最小数
Yablo 悖论	我下一句及后面所有的句子都是假的
Gödel 不动点引理	<b>我有性质 <math>F</math></b>
Tarski 算术真不可定义定理	我不真
Gödel 第一不完备性定理	我不可证
Gödel-Rosser 不完备性定理	对于任何一个关于我的证明，都有一个更短的关于我的否定的证明
Löb 定理	如果我可证，那么 $A$
Curry 悖论	如果我是真的，那么上帝存在
Parikh 定理	我没有关于自己的长度短于 $n$ 的证明
Kleene 不动点定理	<b>我要进行 <math>h</math> 操作</b>
Quine 悖论	把“把中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的”中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的
自测量长度程序	我要输出自己的长度
自复制程序	我要输出自己
自反省程序	我要回顾自己走过的每一步
Gödel 机	<b>我要变成能获取更大效用的自己</b>

# Schmidhuber's Gödel Machine

- ▶ The Gödel machine consists of a **Solver** and a **Searcher** running in parallel.
- ▶ The **Solver** ( $\text{AIXI}^S/\text{AIXI}^{t\ell}$ ) interacts with the environment.
- ▶ The **Searcher** (LSEARCH/HSEARCH/OOPS) searches for a proof of “the modification of the software — including the *Solver* and *Searcher* — will increase the expected utility than leaving it as is”.
- ▶ Logic: a theorem prover and a set of self-referential axioms, which include a description of its own software and hardware, and a description of the probabilistic properties of the environment, as well as a user-given utility function.
- ▶ *Since the utility of “leaving it as is” implicitly evaluates all possible alternative modifications, the current modification is globally optimal w.r.t. its initial utility function.*

# Gödel Machine

- ▶ language  $\mathcal{L} := \{\neg, \wedge, \vee, \rightarrow, \forall, \exists, =, (,), \dots, +, -, \cdot, /, <, \dots\}$
- ▶ well-formed formula

- ▶ utility function  $u(s, e) = \mathbb{E}_\mu \left[ \sum_{t=1}^T r_t \mid s, e \right]$

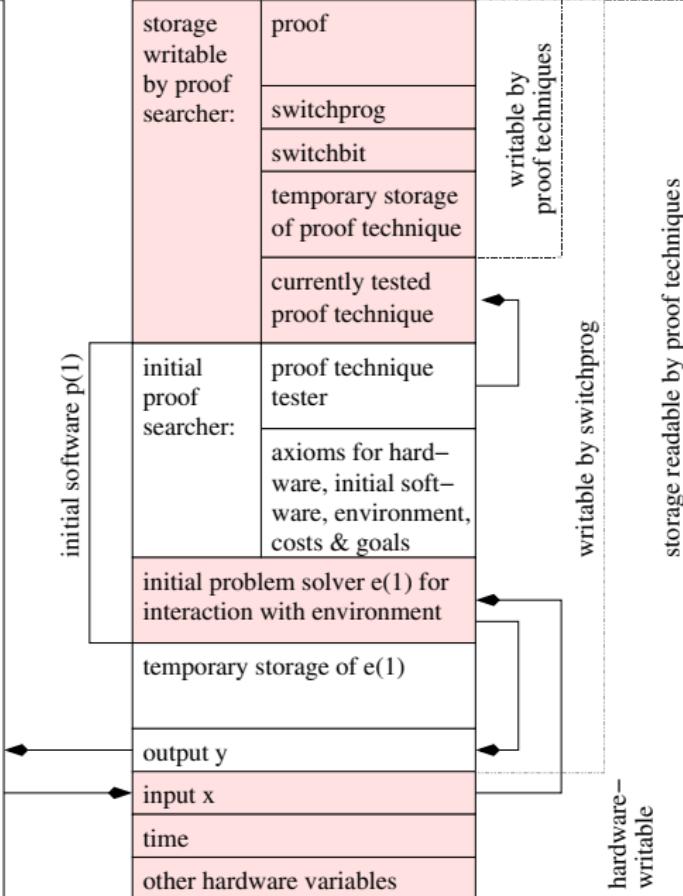
- ▶ target theorem

$$u[s(t) \oplus (\text{switchbit}(t) = 1), e(t)] > u[s(t) \oplus (\text{switchbit}(t) = 0), e(t)]$$

- ▶ theorem prover

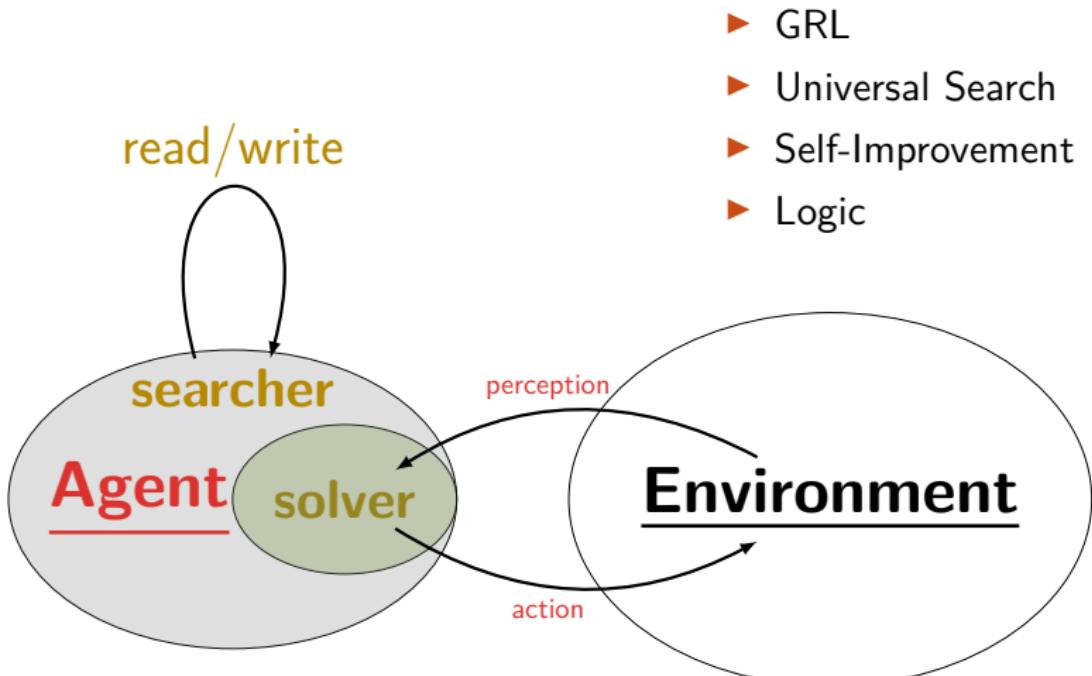
hardware, costs, environment, initial state, utility, logic/arithmetic/probability

ENVIRONMENT



**Figure: Schmidhuber**

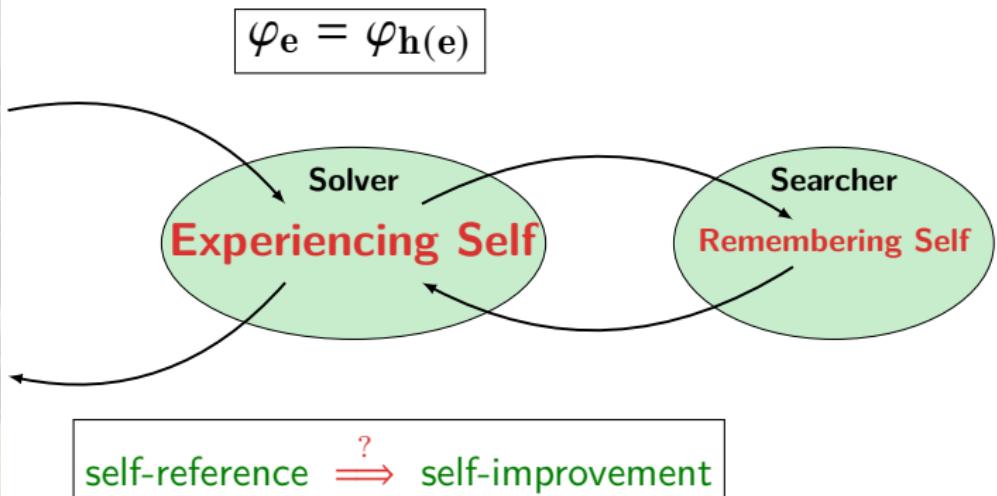
# Gödel Machine

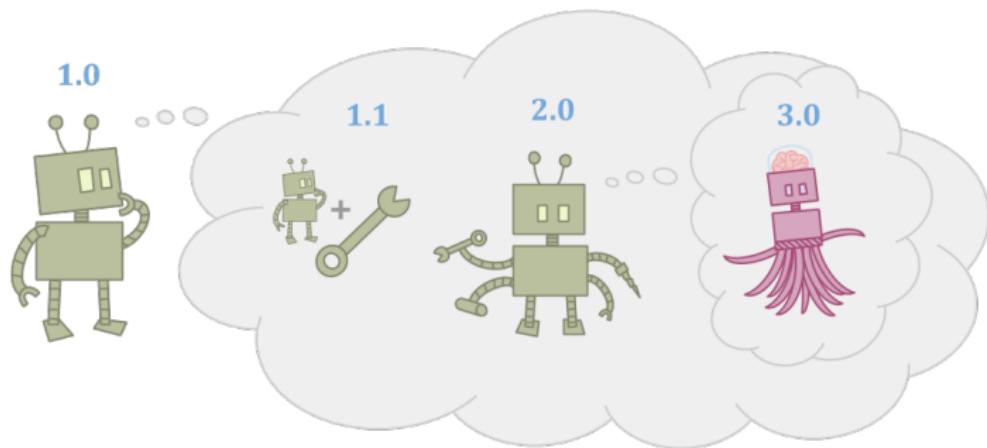
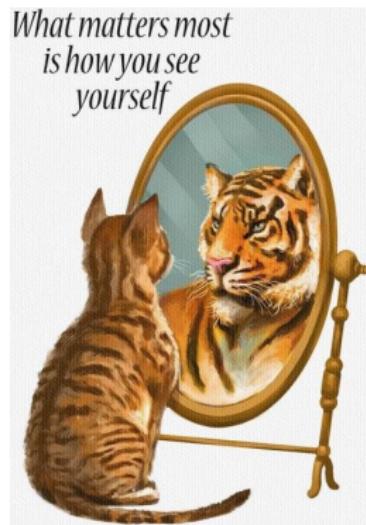


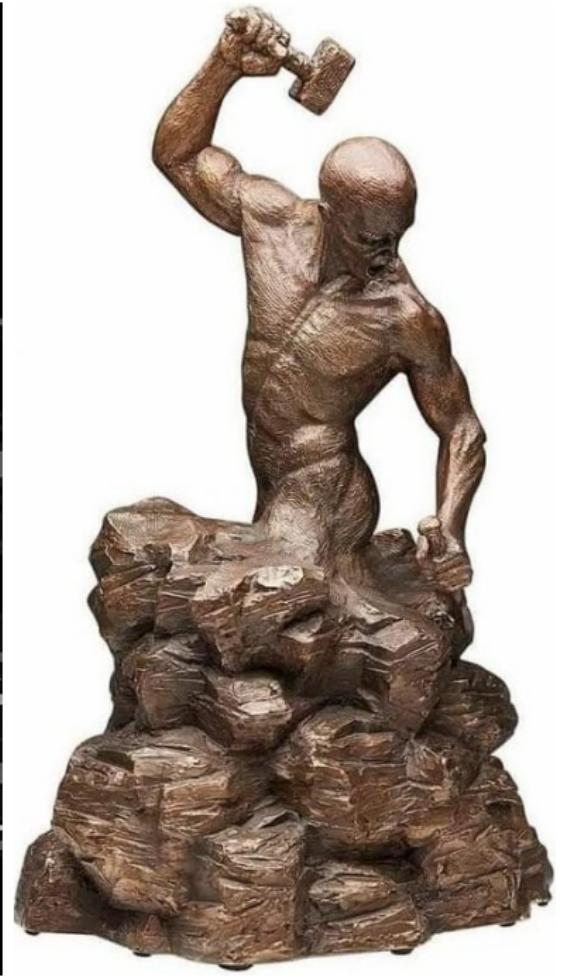
**Disadvantage:** A Gödel Machine with a badly chosen utility function is motivated to converge to a “poor” program. (goal orthogonality!)

# Gödel Machine vs Self-Consciousness vs Free Will?

Self-simulating Computer	Gödel Machine	Self-consciousness
Host Machine	Solver	Experiencing Self
Virtual Machine	Searcher	Remembering Self
Hardware	Hardware	Body







# Gödel Machines

1. *one-shot* self-improvement: Kleene's fixpoint theorem

$$\varphi_e = \varphi_{h(e)}$$

- ▶ global optimality?
- ▶ goal orthogonality? ends vs means

2. *continuous* self-improvement: Kleene's fixpoint theorem **with parameters**

$$\varphi_e(y) = \varphi_{h(e(y),y)}$$

- ▶ “real-time” optimality. human-computer interaction?
- ▶ intelligent explosion / technological singularity???
- continuous self-improvement  $\neq$  exponential iteration

3. *beyond computability*: Kleene's **relativized** fixpoint theorem

$$\varphi_{e(y)}^A = \varphi_{h(e(y),y)}^A$$

- ▶ Gödel Machine PK AIXI<sup>tℓ</sup>
- ▶ Gödel Machine PK AIXI

# Limitation

1. Gödel's first incompleteness theorem / Rice's theorem
2. Gödel's second incompleteness theorem

$$T \vdash \Box_{T'} A \rightarrow A \implies T \vdash \text{Con}_{T'}$$

3. Legg's incompleteness theorem. *General prediction algorithms must be complex. Beyond a certain complexity they can't be mathematically discovered.*
4. Complexity: higher-level abstractions — coarse grained.
  - ▶ Psychology: Duration neglect / Peak-end rule
  - ▶ Information Bottleneck: Learning is to forget!
5. Physical constraint: If we assume that it is not possible to measure properties without changing them (observer effect:  $\alpha$  is fixpoint-free), then there is a limit to self-inspection.

## Evolution & the Number of Wisdom — Chaitin Constant

- ▶ The enormous computational power of evolution could have developed and coded information into our genes,
  - (a) which significantly guides human reasoning,
  - (b) cannot efficiently be obtained from scratch.

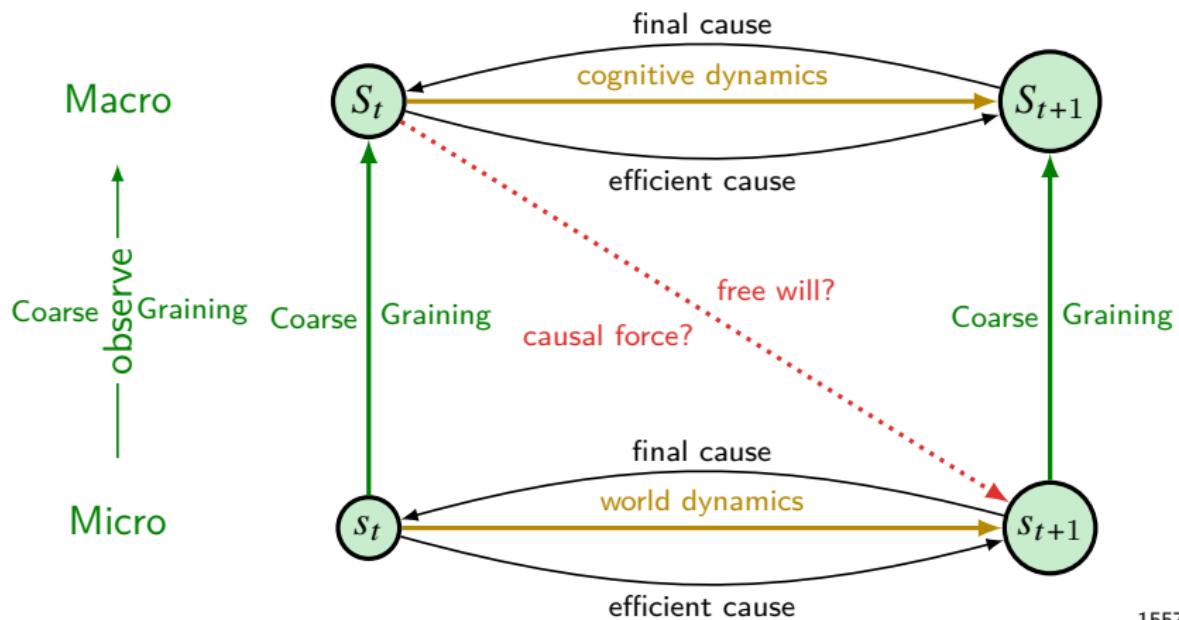
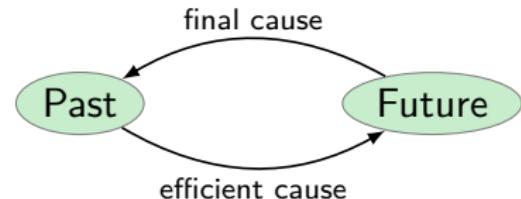
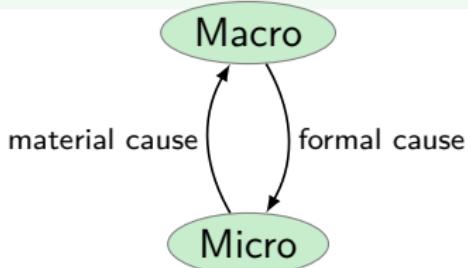
$$\Omega = \lim_{t \rightarrow \infty} \sum_{\ell(p) \leq t \text{ \& } U(p) \downarrow \text{ within time } t} 2^{-\ell(p)}$$

- ▶ Cheating solution: add the information from our genes or brain structure to our AI system?
- ▶ Biological Evolution: Darwin PK Lamarck
  - natural selection vs artificial evolution
  - random vs non-random mutation
- ▶ Tegmark: Life3.0

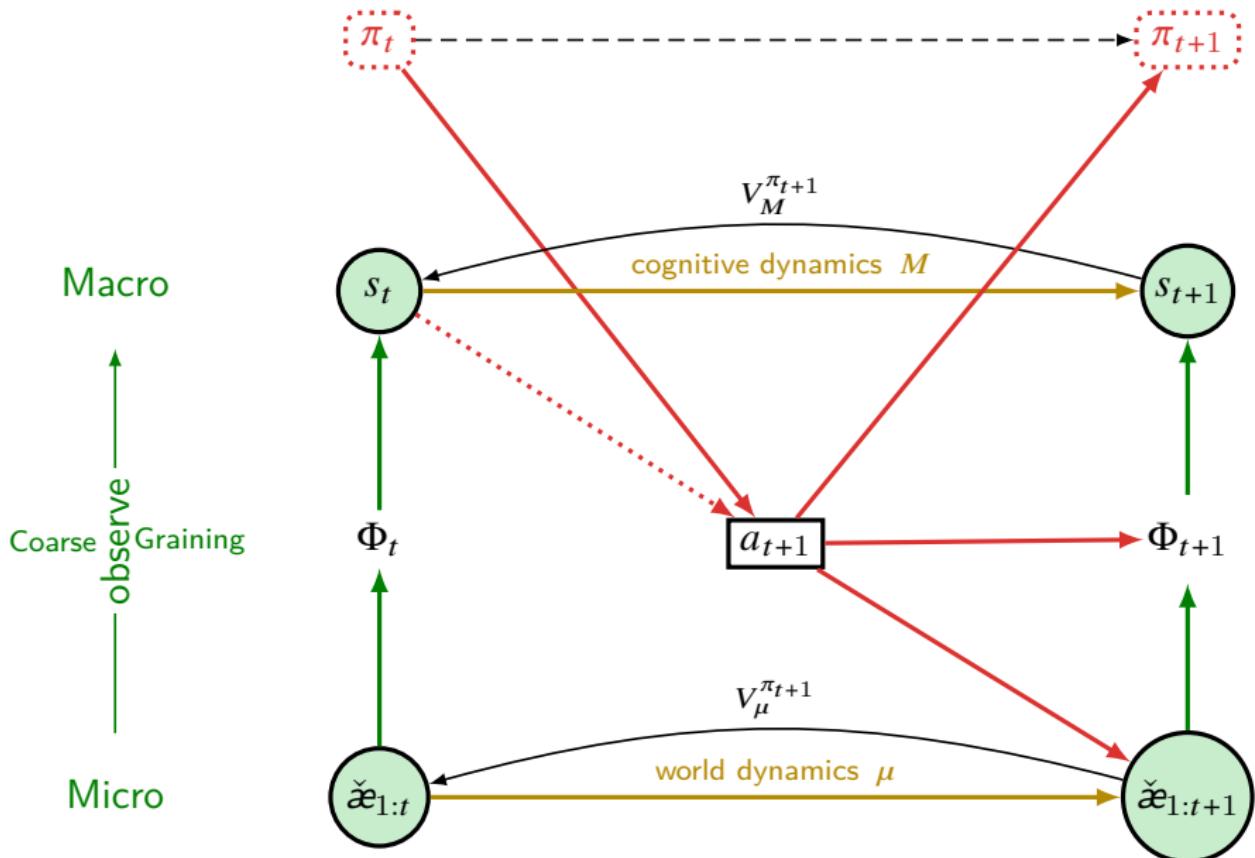


O God, give us courage to change what can be changed,  
serenity to accept what cannot be changed,  
and wisdom to know the difference.

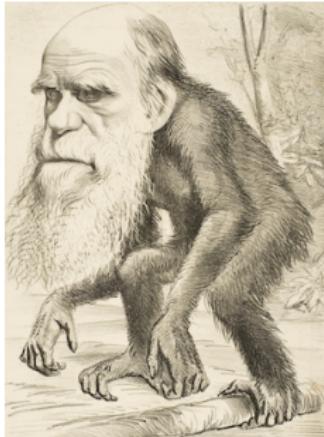
# Jiang ZHANG: Causal Emergence



# Self-Modifying Causal Representation



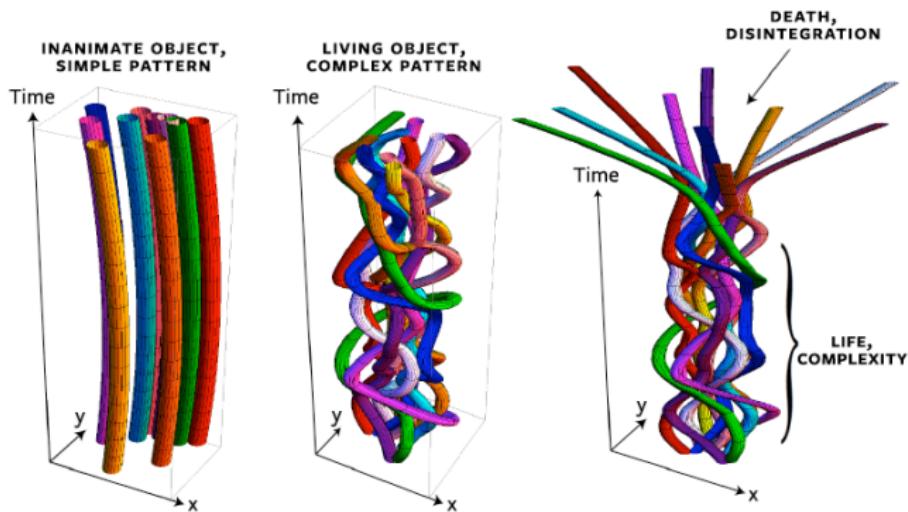
# Darwin PK Lamarck



- ▶ Randomness and atoms in the void: **Nature does not have an a priori purpose** — Democritus, Lucretius, Laplace, Darwin, Boltzmann, Dawkins ...  
analysis — reductionism — statistical laws — mechanisms
- ▶ Holism, Gaia theory, teleology, Romantische Naturphilosophie: **Nature is intelligent and does have a purpose** — Aristotle, Goethe, Lamarck, Wallace, Teilhard de Chardin ...  
synthesis — emergence — self-organization — organisms

Can it design its hardware?		
Can it design its software?		
Can it survive & replicate?		
Life 1.0 (simple biological)		
Life 2.0 (cultural)		
Life 3.0 (technological)		

# Life is a braid in spacetime



**Figure:** Tegmark: The motion of an object corresponds to a pattern in spacetime. You're a braid in spacetime — indeed, one of the most elaborate braids known.

## Consciousness — Integrated Information Theory

- ▶ 当处于“心流”状态时，我们只关注高层次的信息，而对于低层次的细节则处于“无意识”状态。
- ▶ 我们有意识的信息处理仅仅是冰山一角。大部分脑区是无意识的。有意识的经验仅仅是对大量无意识过程的事后总结。意识落后于做出决定约四分之一秒。脑电测量可以在你意识到自己做出决定之前就预测出你的选择。
- ▶ 意识以一种相当自治和整合的方式处理信息。
- ▶ 整合信息度量的是一个系统在演化过程中无法归约为独立的部分的能力。

# 意识体验的“公理” — Tononi

**Existence** 意识体验是第一人称存在的：“I experience therefore I am”，具有之于自身的因果力.

**Intrinsicity** 意识体验是内在的：与外部观察者无关.

**Information** 意识体验是特定的，有信息量的：每个体验以其独特的方式区别于可能的其它体验.

**Integration** 意识体验是整合的：每个体验无法归约为相互独立的组成部分. 每个部分都既影响其他部分，又受到其他部分的影响.  
每一个意识状态都是一个单子 (Monad)，无法分割为能被独立体验的组分.

**Exclusion** 意识体验是排他的：每个体验都有明确的边界；每个体验都有特定的时空颗粒度.

**Composition** 意识体验是有结构的：每个体验由多个机制以不同的组合方式构成.

# What is complexity?

- ▶ How hard is it to describe?
- ▶ How hard is it to create?
- ▶ What is its degree of organization?

*“Integrated information captures the information generated by causal interactions in the whole, over and above the information generated by the parts.”*

— Tononi

A system is complex if it displays **emergent** properties that cannot be **reduced** to the properties of its **parts**.

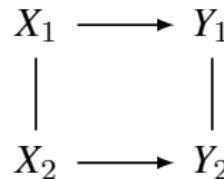
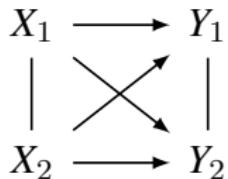
Tononi: the degree of conscious experience is related with the amount of integrated information.

## Question

- ▶ Are the axioms/postulates correct and complete?
- ▶ What is the metaphysical status of IIT?
  - materialism, idealism, dualism, neutral monism, epiphenomenalism, emergentism, panpsychism?

## Integrated Information Theory (IIT)

- ▶ Suppose given a stochastic dynamical system, where the state of the system at time  $t$  is described by a set of random variables  $\{X_i = X_i^{(t)}\}_{i=1}^N$  which correspond to a partition of the system into  $N$  subsystems, and the state at time  $t + 1$  by  $\{Y_i = X_i^{(t+1)}\}_{i=1}^N$ .
- ▶ The full system including all the mutual influences between these two sets of variables is described by  $P(X, Y)$ .
- ▶ Integrated information is meant to capture the difference between  $P(X, Y)$  and an approximation  $Q(X, Y)$  where only certain kinds of mutual influences are retained.
- ▶ These are usually taken to be the interdependencies between the variables at the same time and between each  $X_i$  and the corresponding  $Y_i$ , removing the dependencies of the  $Y_i$  from the  $X_j$  with  $j \neq i$ .



# IIT — Conditional Independent Statements

- Given a partition  $\lambda$

$$\{(X, Y)\} = \bigsqcup_{i=1}^N \{(X_i, Y_i)\}$$

Consider the space

$$\mathcal{M}_\lambda := \{Q : Q(Y_i | X) = Q(Y_i | X_i) \text{ for } i = 1, \dots, N\}$$

- The best approximation to  $P(X, Y)$  by  $Q(X, Y)$  in  $\mathcal{M}_\lambda$  is

$$Q_\lambda^* := \underset{Q \in \mathcal{M}_\lambda}{\operatorname{argmin}} D_{\text{KL}}(P \| Q)$$

- Then the integrated information, for a given partition  $\lambda$ , is defined as

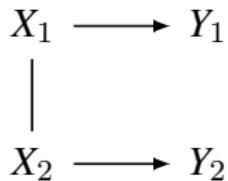
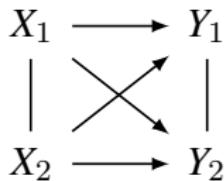
$$\Phi_\lambda := D_{\text{KL}}(P \| Q_\lambda^*) = \underset{Q \in \mathcal{M}_\lambda}{\min} D_{\text{KL}}(P \| Q)$$

with a further minimization over the choice of the partition,

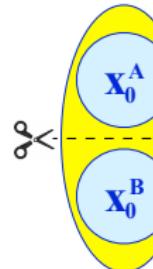
$$\Phi_{\text{CIS}} := \min_{\lambda} D_{\text{KL}}(P \| Q_\lambda^*) = \min_{Q \in \bigcup_{\lambda} \mathcal{M}_\lambda} D_{\text{KL}}(P \| Q)$$

# IIT — another version — Stochastic Interaction

$$Y_j \perp X_i \mid X_{I \setminus \{i\}}$$

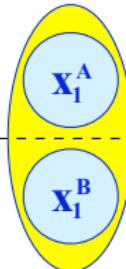


Probability distribution  $p_0$ :



$$p_1 = M p_0$$

Probability distribution  $p_1$ :



$$\mathcal{M}_{\text{SI}} := \left\{ Q : Q(Y \mid X) = \prod_{i=1}^N Q(Y_i \mid X_i) \right\}$$

$\Phi$  measures inability to tensor factorize  $M = M^A \otimes M^B$

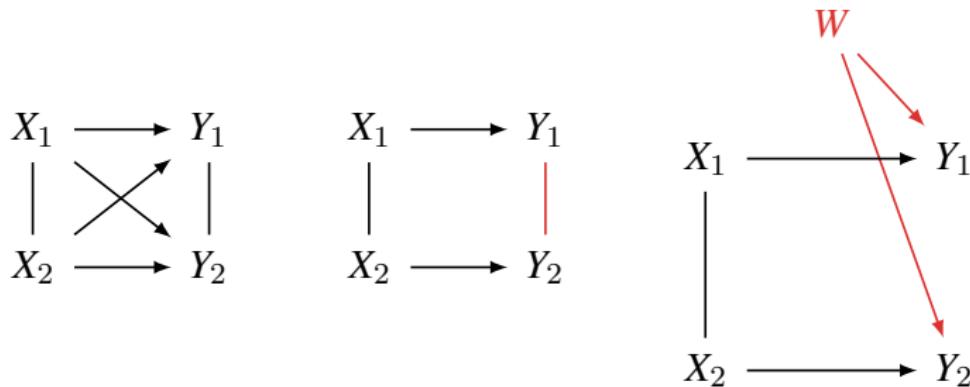
►  $\Phi_{\text{SI}} := \min_{Q \in \mathcal{M}_{\text{SI}}} D_{\text{KL}}(P \parallel Q) = \sum_i H(Y_i \mid X_i) - H(Y \mid X)$

► counter-example

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1^0 & \dots & 1^{N-1} \\ \vdots & \ddots & \vdots \\ N^0 & \dots & N^{N-1} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

## IIT — another version — Causal Information Integration

- IIT including a common exterior influence.



$$\mathcal{M}_{\text{CII}} := \left\{ Q : Q(x, y) = \sum_w Q(x)Q(w) \prod_{i=1}^N Q(y_i | x_i, w) \right\}$$

$$\Phi_{\text{CII}} := \min_{Q \in \mathcal{M}_{\text{CII}}} D_{\text{KL}}(P \| Q)$$

## Gaia Hypothesis vs Panpsychism

- ▶ The whole earth, the seas and rocks and plants and atmosphere, are a single self-regulating entity. Too many trees? Fires happen. Too much carbon dioxide? More vegetation. The earth maintains its own temperature within a range, as well as, astonishingly, the salinity of the oceans across eons, and so forth. All sorts of things are kept in earth's "preferable" range to be conducive to life.
- ▶ If the earth is conscious, how would we know? Can it feel pain? Does it have emotions? What does it think of us? What of the sun? Could it be conscious? Children who draw outdoor scenes in kindergarten invariably give the sun a smiling face...
- ▶ How does consciousness combine?

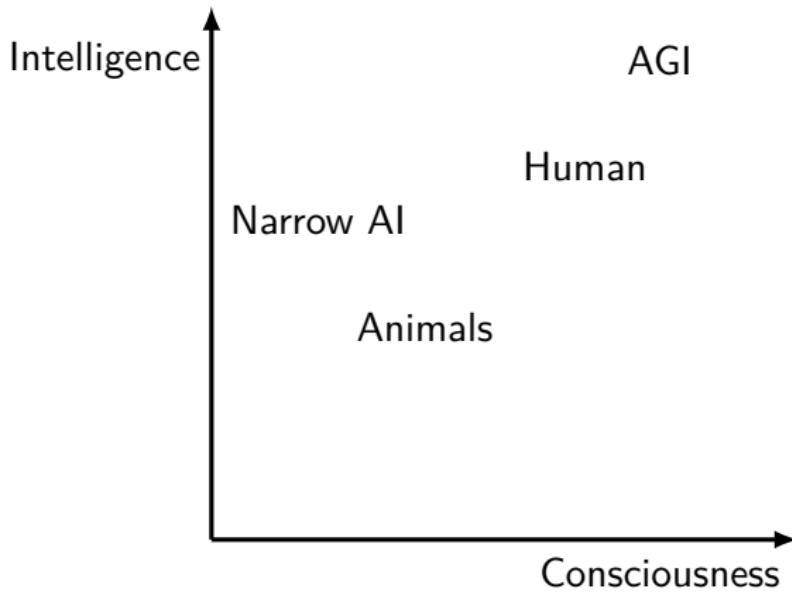
# 机器意识<sup>23</sup>

- ▶ 信息整合理论：意识产生于整体系统对大量信息的整合。
- ▶ 预测加工理论：意识产生于预测模型对外部信息的主观建构。
- ▶ 高阶表征理论：意识产生于对心理状态的一阶表征的元表征。
- ▶ 量子意识理论：意识产生于脑中微小结构的波函数坍缩效应。
- ▶ 全局工作空间理论 (Global Workspace Theories GWTs)：意识产生于局域性的认知模块被全局广播。
  - 当知觉、思想、情感等进入“工作空间”时变得有意识，— 大脑是一个剧场，有意识的思维是某一特定时刻舞台上聚光灯下的活动 (但大脑中实际的工作空间并不是局部的，而是分布在大脑皮层的额叶和顶叶区域)。

---

<sup>23</sup>Butlin, Patrick, et al. "Consciousness in artificial intelligence: insights from the science of consciousness." arXiv preprint arXiv:2308.08708 (2023).

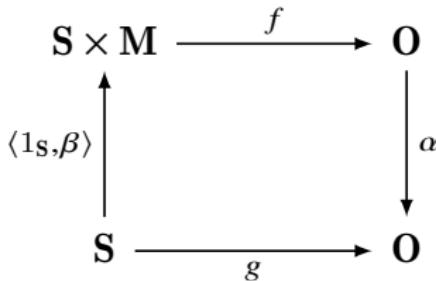
# Intelligence vs Consciousness ?



## Non-operational Self-inspection [Svo18; Sza18]

*The information available to the observer regarding his own state could have absolute limitations, by the laws of nature.*

— John von Neumann



- ▶  $S$ : quantum states.
- ▶  $M$ : quantum measurements.
- ▶  $O$ : possible outcomes of quantum measurements.
- ▶  $f(s, m)$ : predicts the outcome of measurement  $m$  for state  $s$ .

If we assume that it is not possible to measure properties without changing them (observer effect:  $\alpha$  is fixpoint-free), then there is a limit to self-inspection.

## Self-Modification

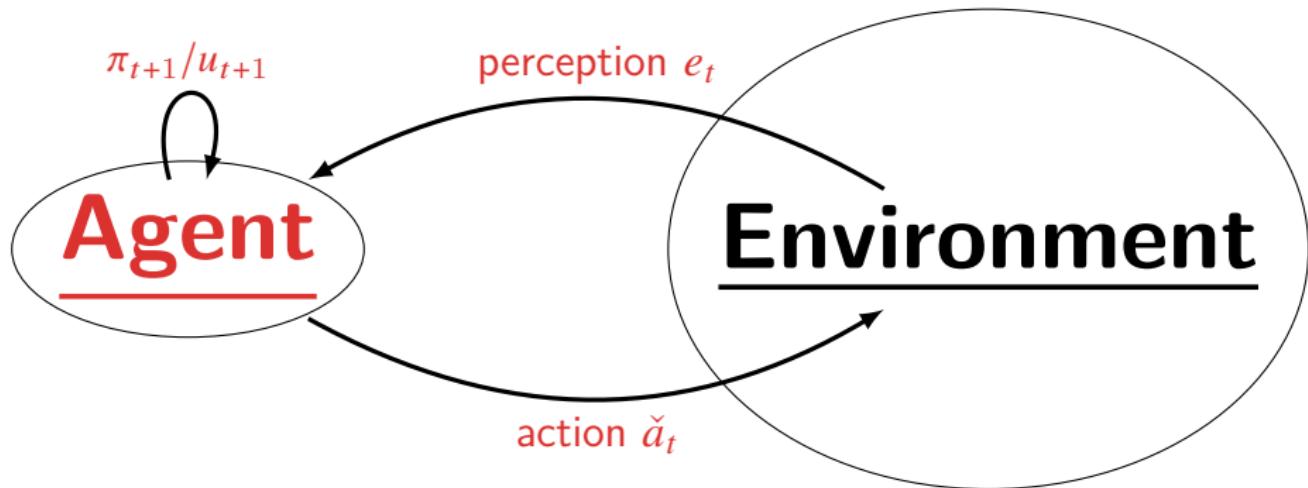
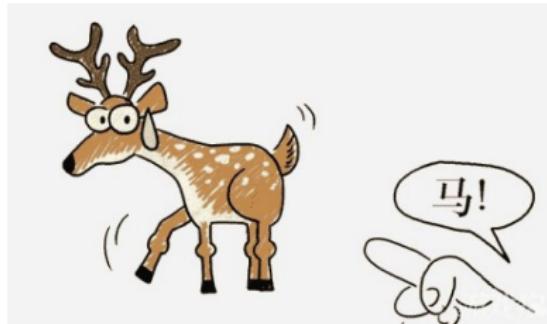
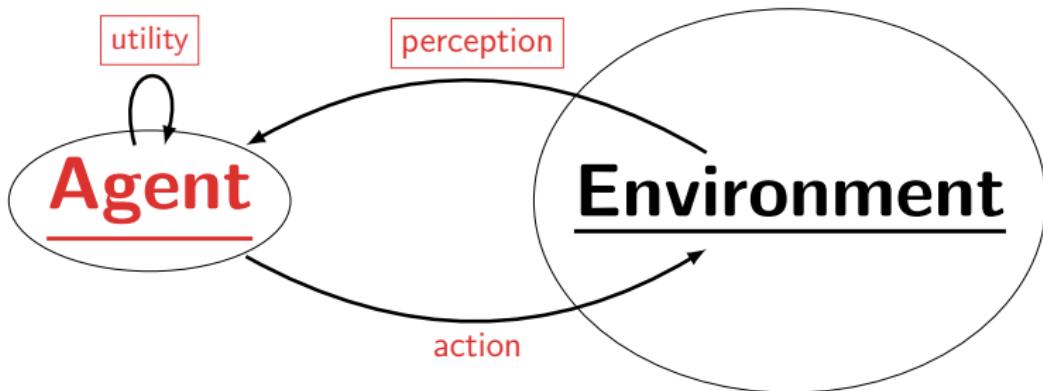


Figure: Policy/utility self-modification.  $a_t = \langle \check{a}_t, \pi_{t+1} \rangle$  or  $a_t = \langle \check{a}_t, u_{t+1} \rangle$

# External/Internal Wireheading & Free Will<sup>24</sup>

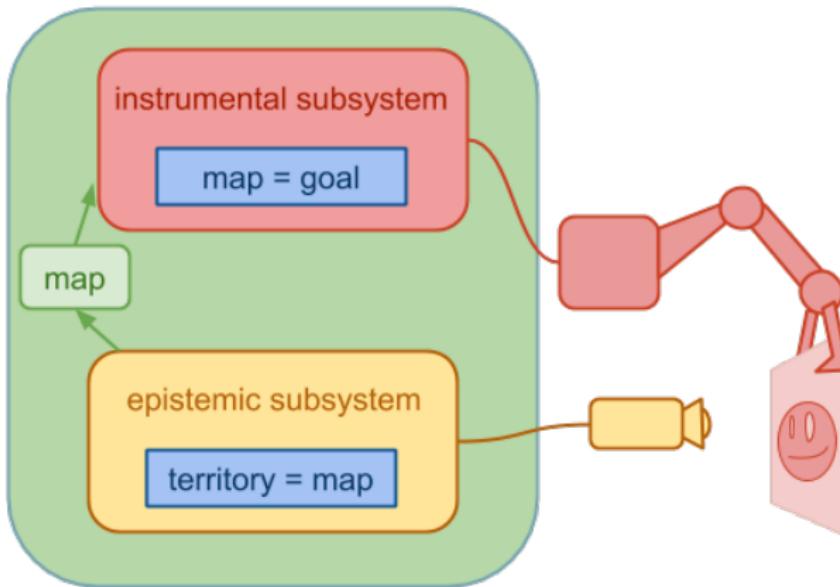


1. 我喜欢马.  
指鹿为马!
2. 我喜欢马.  
我意欲自己  
喜欢鹿!  
我喜欢鹿!



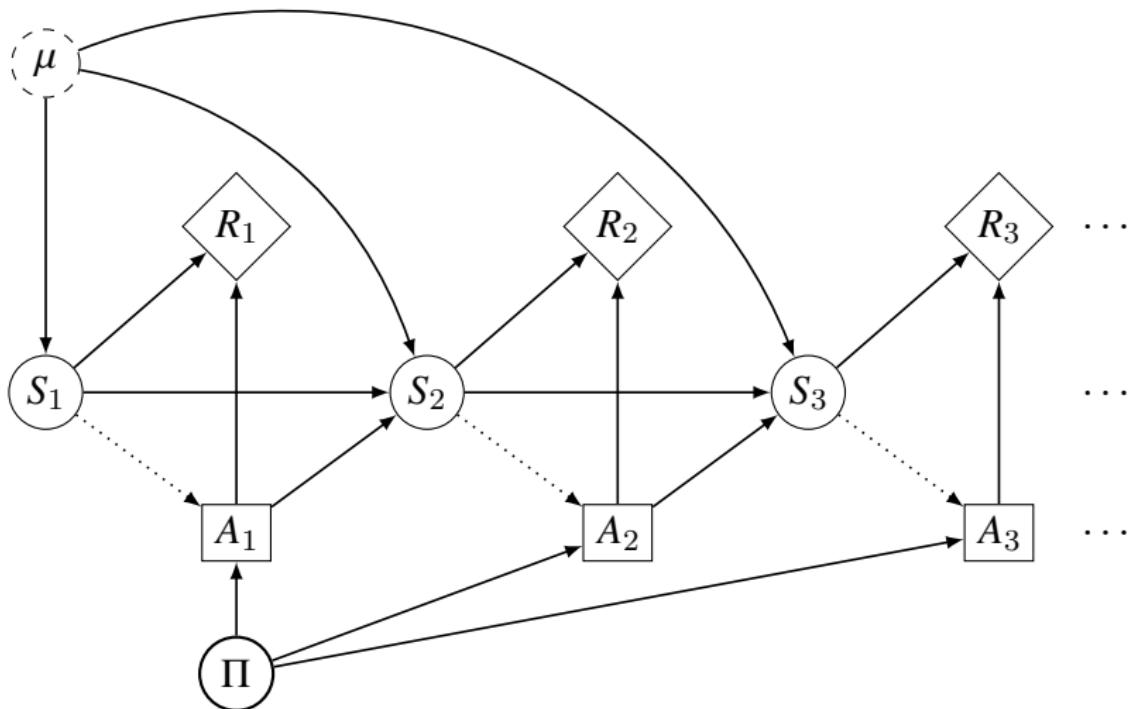
<sup>24</sup> Everitt, Filan, Daswani, Hutter: Self-modification of policy and utility function in rational agents.  
Frankfurt: Freedom of the will and the concept of a person.  
Aaronson: The ghost in the quantum turing machine.  
Calude, Kroon, Poznanovic: Free will is compatible with randomness.

# Self-Deception

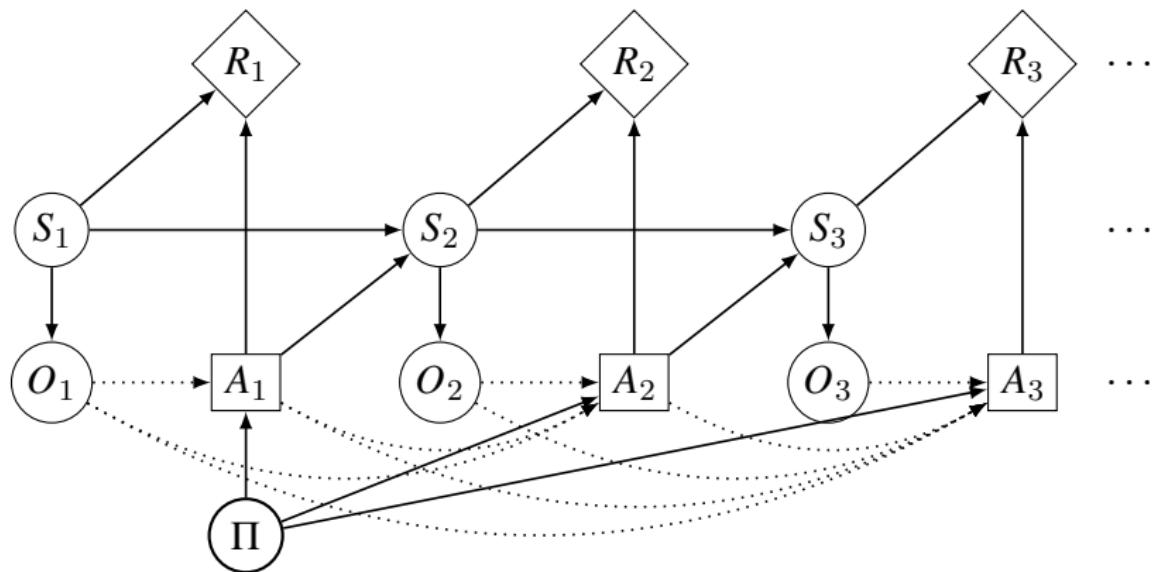


**Figure:** The epistemic subsystem just wants accurate beliefs. The instrumental subsystem uses those beliefs to track how well it is doing. If the instrumental subsystem gets too capable relative to the epistemic subsystem, it may decide to try to fool the epistemic subsystem.

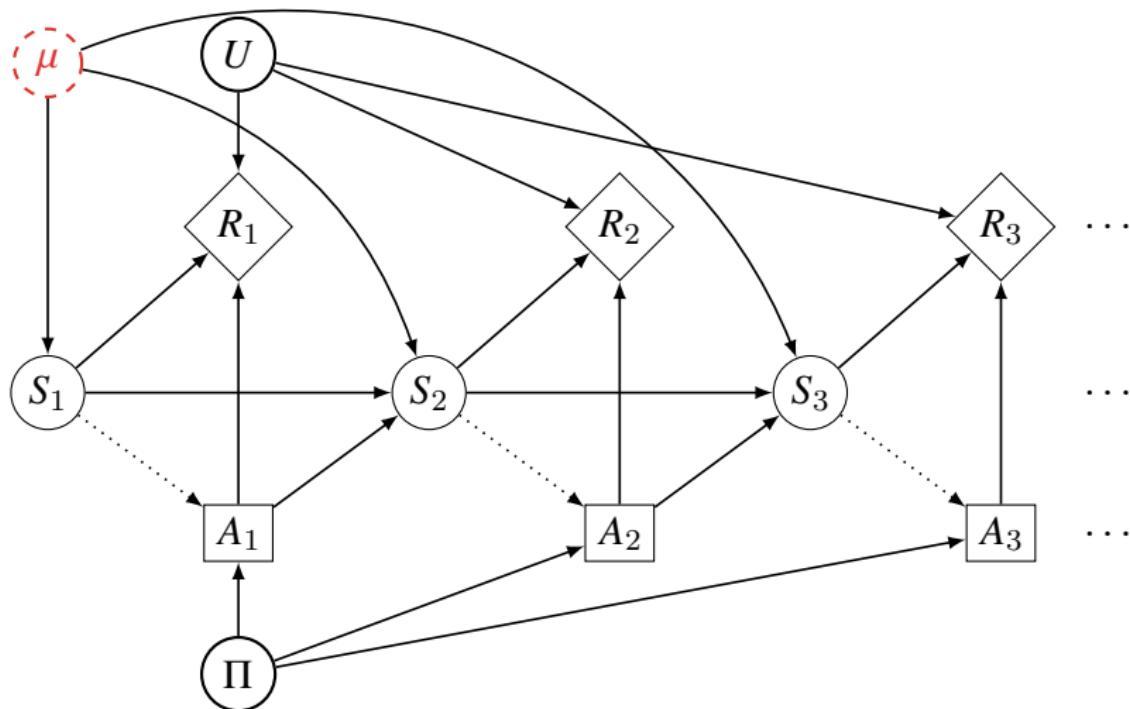
# Causal Influence Diagram of Unknown MDP



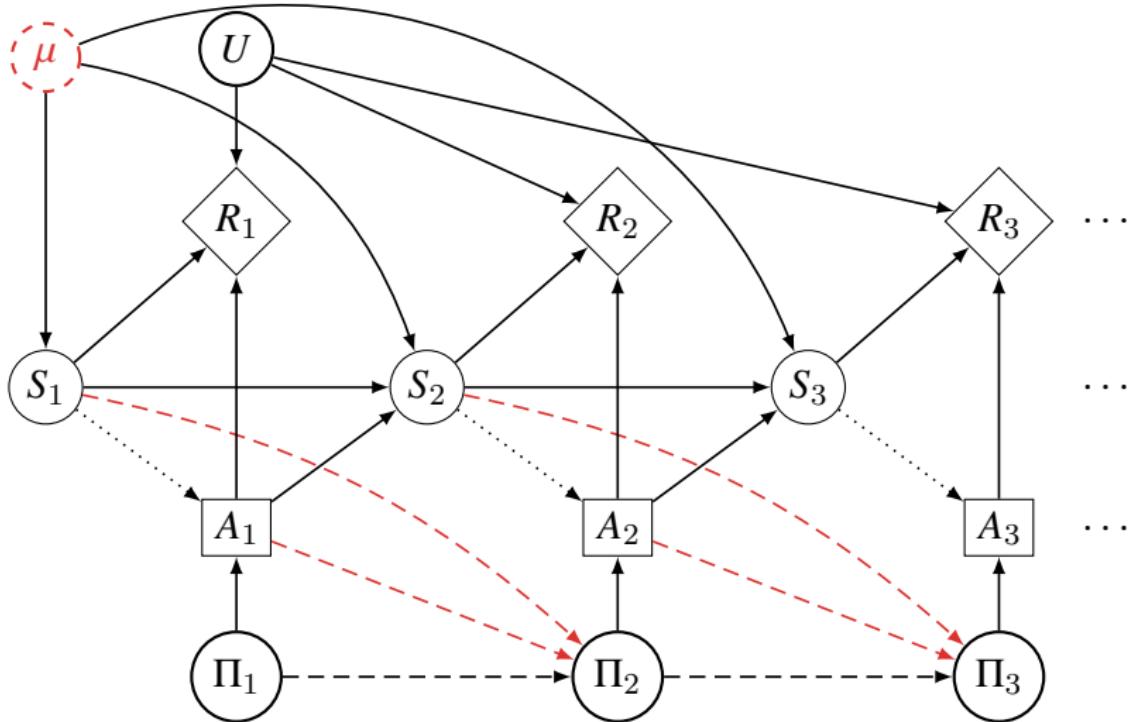
# Causal Influence Diagram of POMDP



# Causal Influence Diagram of Unknown Environment with Explicit Utility Function

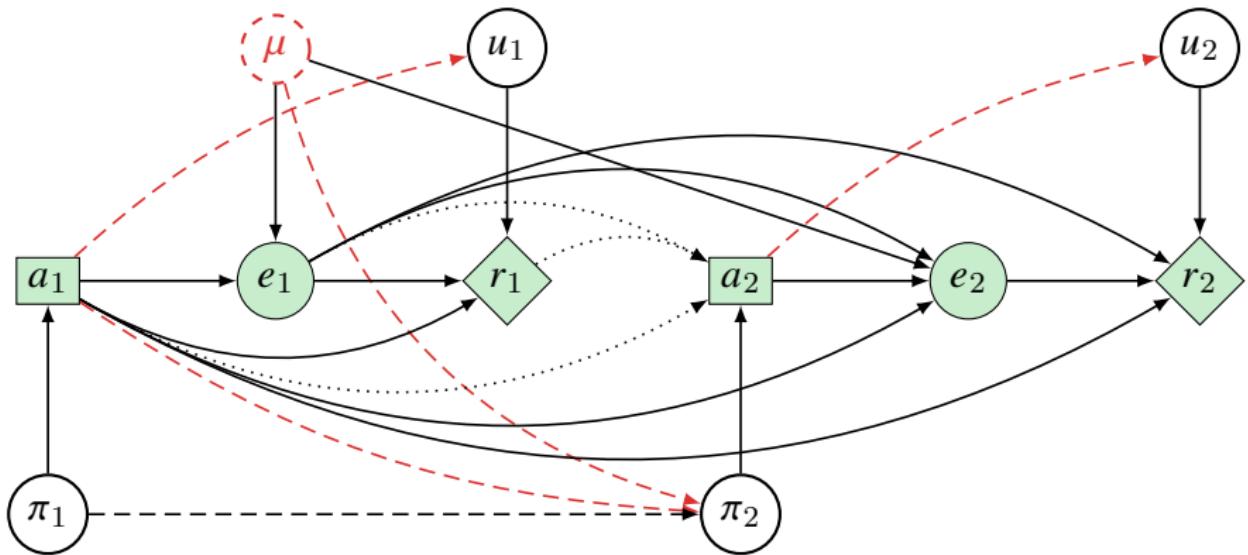


# Causal Influence Diagram of Unknown Environment (with an embedded agent)

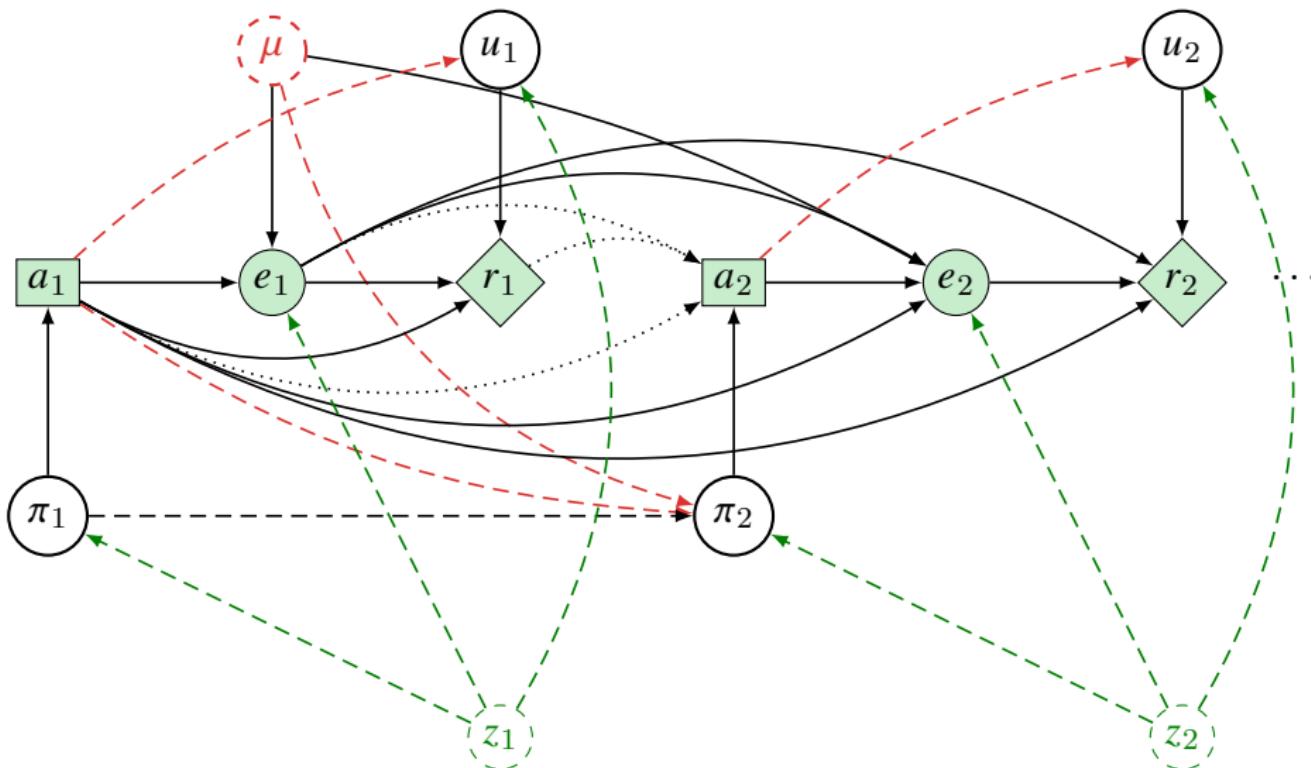


$$\pi_{t+1} = f_\pi(\pi_t, s_t, a_t)$$

# Causal Influence Diagram of Self-Modification UAI



# Self-Modifying RL Agent with Unobserved Confounder



# Everitt's Self-Modification

## Definition (Different Agents)

- ▶ Hedonistic Value

$$Q^{h,\pi}(\boldsymbol{a}_{<t} a_t) = \sum_{e_t \in \mathcal{E}} \rho(e_t \mid \check{\boldsymbol{a}}_{<t} \check{a}_t) [u_{t+1}(\check{\boldsymbol{a}}_{1:t}) + \gamma V^{h,\pi}(\boldsymbol{a}_{1:t})]$$

- ▶ Ignorant Value

$$Q_t^{i,\pi}(\boldsymbol{a}_{<k} a_k) = \sum_{e_k \in \mathcal{E}} \rho(e_k \mid \check{\boldsymbol{a}}_{<k} \check{a}_k) [u_t(\check{\boldsymbol{a}}_{1:k}) + \gamma V_t^{i,\pi}(\boldsymbol{a}_{1:k})]$$

- ▶ Realistic Value

$$Q_t^r(\boldsymbol{a}_{<k} a_k) = \sum_{e_k \in \mathcal{E}} \rho(e_k \mid \check{\boldsymbol{a}}_{<k} \check{a}_k) [u_t(\check{\boldsymbol{a}}_{1:k}) + \gamma V_t^r(\boldsymbol{a}_{1:k})]$$

where  $a_k = \langle \check{a}_k, \pi_{k+1} \rangle$ .

## Lemma

$$Q_t^{h,\pi}(\mathbf{a}_{\leq t} a_t) = \mathbb{E}_\rho \left[ \sum_{k=t}^{\infty} \gamma^{k-t} u_{k+1}(\check{\mathbf{a}}_{1:k}) \middle| \check{\mathbf{a}}_{\leq t} \check{a}_t, \text{do}(\pi_{t:\infty} = \pi) \right]$$

$$Q_t^{i,\pi}(\mathbf{a}_{\leq t} a_t) = \mathbb{E}_\rho \left[ \sum_{k=t}^{\infty} \gamma^{k-t} u_t(\check{\mathbf{a}}_{1:k}) \middle| \check{\mathbf{a}}_{\leq t} \check{a}_t, \text{do}(\pi_{t:\infty} = \pi) \right]$$

$$Q_t^r(\mathbf{a}_{\leq t} a_t) = \mathbb{E}_\rho \left[ \sum_{k=t}^{\infty} \gamma^{k-t} u_t(\check{\mathbf{a}}_{1:k}) \middle| \check{\mathbf{a}}_{\leq t} \check{a}_t, \text{do}(\pi_t = \pi) \right]$$

In  $Q^r$ , actions  $a_k$  are chosen by  $\pi_k$ .

	Utility	Policy	Self-modification	Primary self-mod. risk
$Q^h$	Future	Either	Promotes	Survival agent
$Q^i$	Current	Current	Indifferent	Self-damage
$Q^r$	Current	Future	Demotes	Resists modification

- ▶ The hedonistic agent self-modifies to  $u(\cdot) = 1$ .
- ▶ The ignorant agent may self-modify by accident.
- ▶ The realistic agent will resist modifications.

# Self-Modification — Realistic Agent

$$V_t^\pi(\alpha_{<k}) = Q_t(\alpha_{<k} \pi(\alpha_{<k}))$$

$$Q_t(\alpha_{<k} a_k) = \sum_{e_k \in \mathcal{E}} \rho(e_k | \check{\alpha}_{<k} \check{a}_k) [\textcolor{red}{u}_t(\check{\alpha}_{1:k}) + \gamma V_t^{\pi_{k+1}}(\alpha_{1:k})]$$

$$\pi_1^* := \operatorname{argmax}_{\pi} V_1^\pi(\epsilon)$$

Theorem (All realistic optimal policies are non-modifying)

Let  $\rho$  and  $u_1$  be modification-independent. For every  $t \geq 1$ , for all percept sequences  $e_{<t}$ , and for the action sequence  $a_{<t}$  given by  $a_i = \pi_i(\alpha_{<i})$ , we have

$$Q_1(\alpha_{<t} \pi_t(\alpha_{<t})) = Q_1(\alpha_{<t} \pi_1^*(\alpha_{<t}))$$



All realistic optimal policies are non-modifying.

$$\forall h \in \mathcal{H} \exists a \in \mathcal{A} : \pi_1^{\text{Gödel}}(h) = \langle a, \pi_1^{\text{Gödel}} \rangle$$

Not wireheading; But orthogonal!



## Remark

如果

1. Agent 是 model-based, 并且基于当下的效用函数进行规划, 评估未来的场景,
2. Agent 能够预测到自我修改对未来策略的影响,
3. 奖励函数本身不鼓励自我修改,

那么, Agent 不会主动修改自己的效用函数.

**Remark:** 通常 model-free 的 Agent 违反第一条; off-policy 的 Agent 比如 Q-learning 违反第二条; 如果 Agent 是通过学习获得的奖励函数的话, 第三条可能不成立.

# Orthogonality and Wireheading in Self-improving GRL

$$V_t^\pi(\mathbf{æ}_{) := Q_t(\mathbf{æ}_{\pi(\mathbf{æ}_{))}$$

$$Q_t(\mathbf{æ}_{a_t}) := \sum_{e_t \in \mathcal{E}} \sum_{\nu \in \mathcal{M}} w_{\mathbf{æ}_{}^\nu}^\nu r(e_t \mid \mathbf{\check{æ}}_{\check{a}_t}) \left[ \sum_{u \in \mathcal{U}} \sum_{a_t^H} P(a_t^H \mid a_t) P(u \mid P_\nu^{\pi_{t+1}}, \mathbf{æ}_{a_t}, \mathbf{æ}_{}^H a_t^H) u(\mathbf{\check{æ}}_{1:t}) + \gamma V_t^{\pi_{t+1}}(\mathbf{æ}_{1:t}) \right]$$
$$\pi_t(\mathbf{æ}_{) := \operatorname{argmax}_{a_t \in \mathcal{A} \times \Pi} Q_t(\mathbf{æ}_{a_t})$$

where

$$P(u \mid P_\nu^\pi, h) := \frac{\tilde{U}(u, P_\nu^\pi, h)}{\sum_{u \in \mathcal{U}_h} \tilde{U}(u, P_\nu^\pi, h)}$$

and

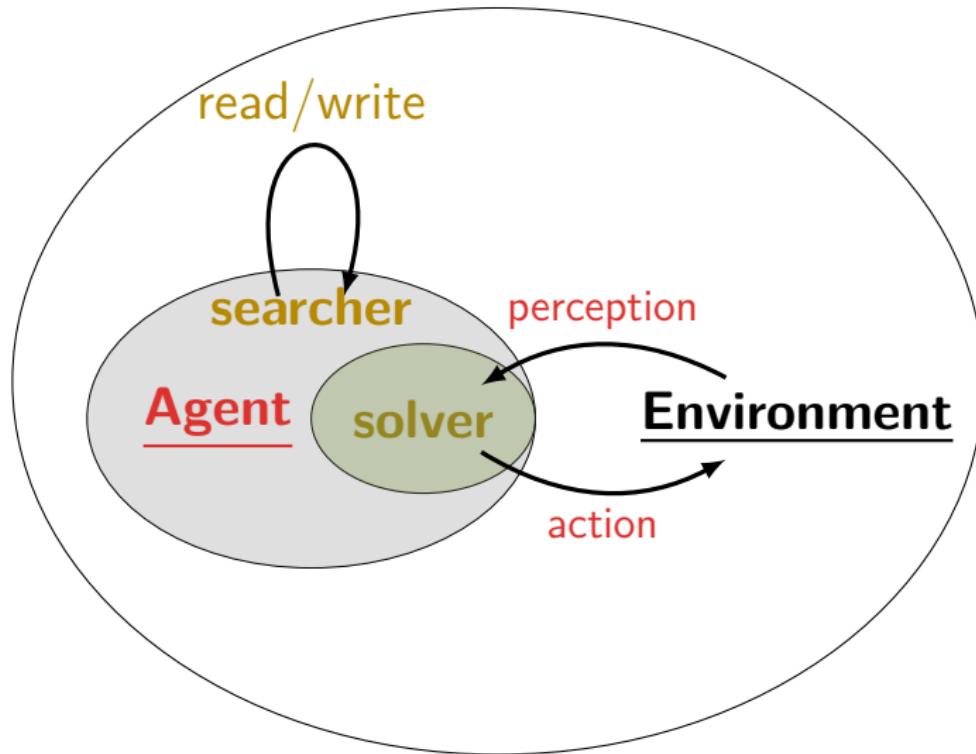
$$\tilde{U}(u, P_\nu^\pi, h) := \sum_{z \in \mathcal{Z}_h} P_\nu^\pi(z \mid h) u(z)$$

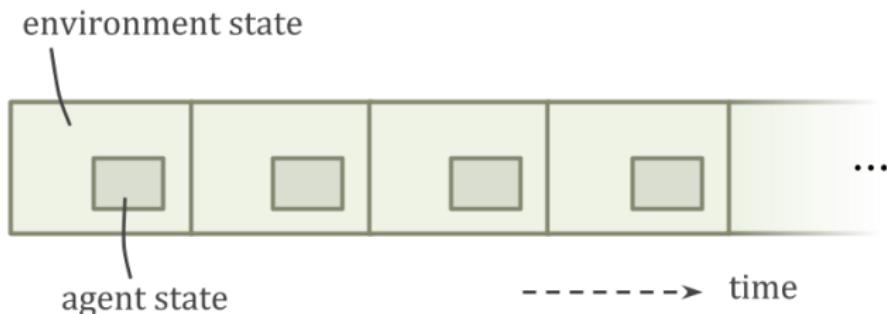
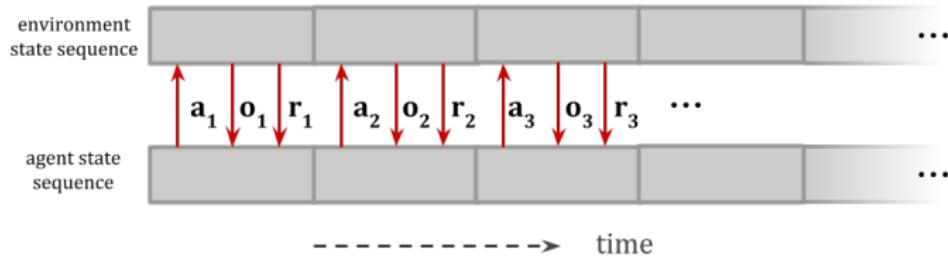
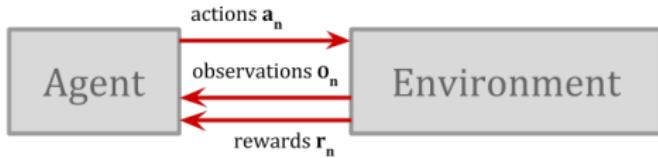
$$\pi^*(\mathbf{æ}_{) := \pi_t(\mathbf{æ}_{)}$$

$$\pi^*(e_{) := \pi_t(e_{\mid \pi_{t-1}(e_{$$

uncertain model-based utility / IRL

# Fatalism — God Bless AI!





# Orseau's Space-Time Embedded Intelligence

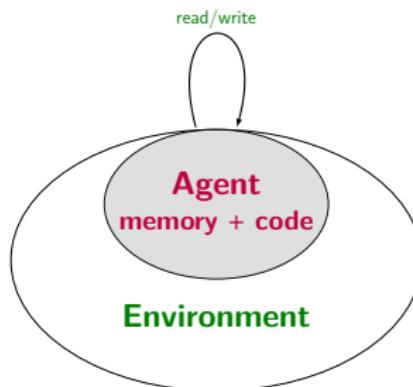
$$\pi^* := \underset{\pi_0 \in \Pi^\ell}{\operatorname{argmax}} V(\pi_0, \epsilon)$$

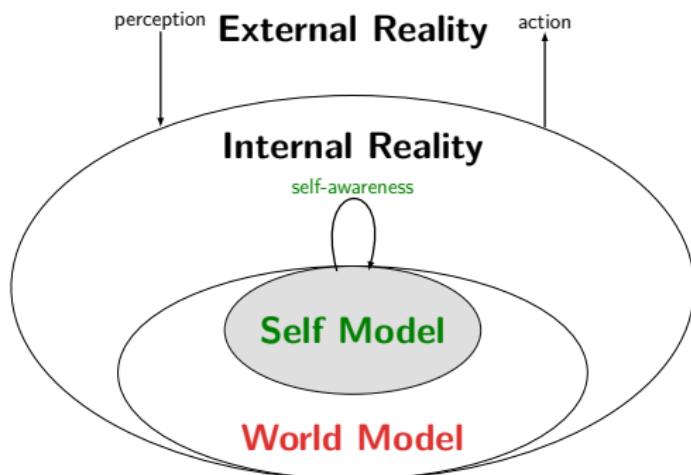
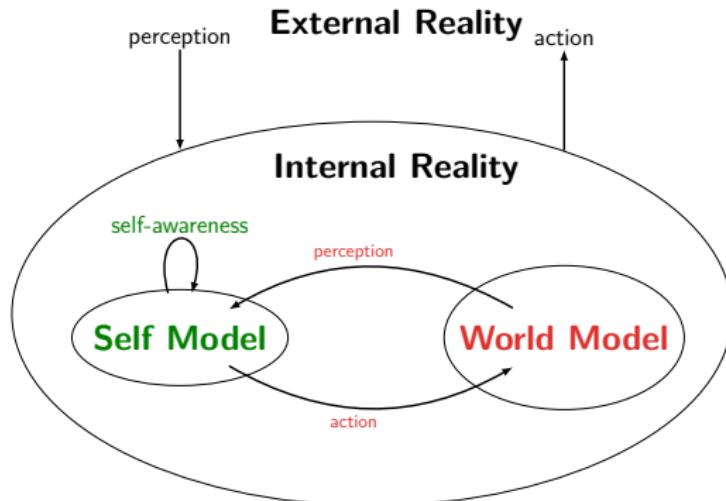
$$V(\pi_t, \mathfrak{a}_{<t}) := \sum_{a_t = \langle \check{a}_t, \textcolor{red}{\tilde{\pi}_{t+1}} \rangle} \pi_t(a_t \mid \check{e}_{t-1}) \sum_{e_t = \langle \check{e}_t, \textcolor{red}{\pi_{t+1}} \rangle} \rho(e_t \mid \mathfrak{a}_{<t} a_t) [u(\mathfrak{a}_{1:t}) + \gamma_t V(\pi_{t+1}, \mathfrak{a}_{1:t})]$$

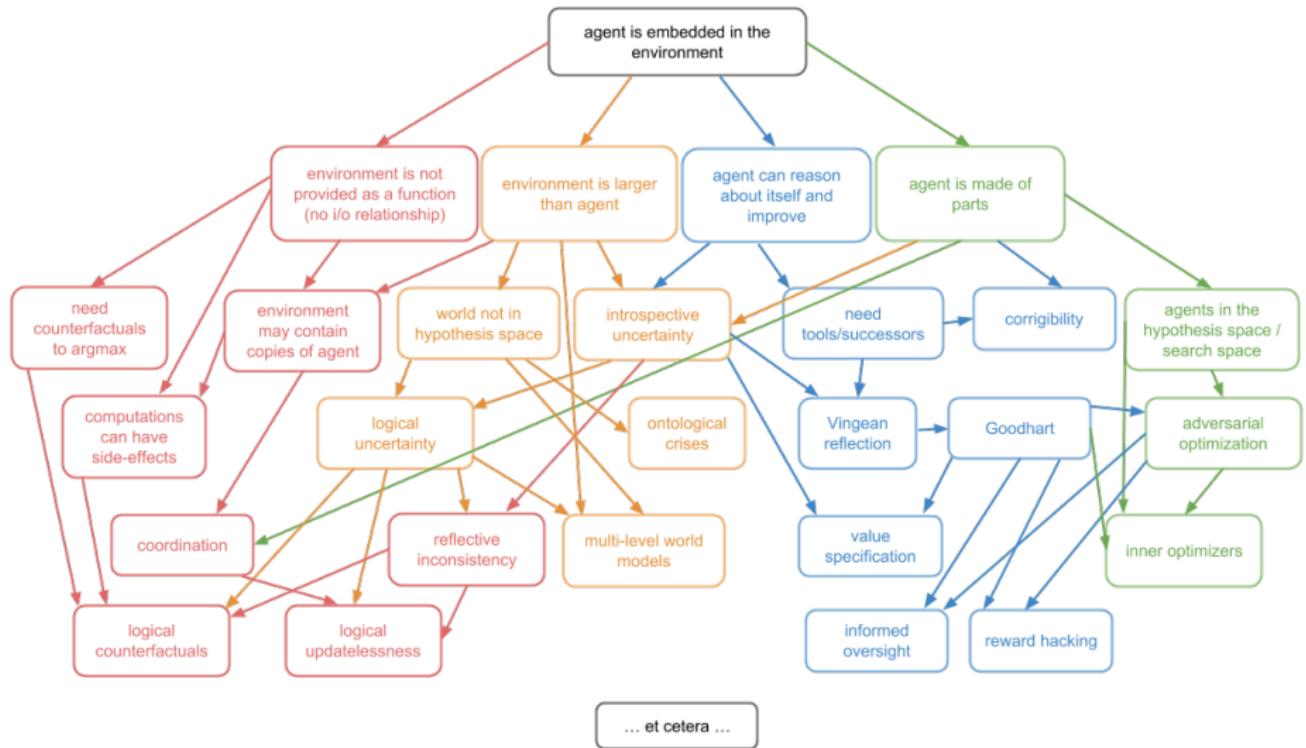


$$\pi^* := \underset{\pi_0 \in \Pi^\ell}{\operatorname{argmax}} V(\pi_0)$$

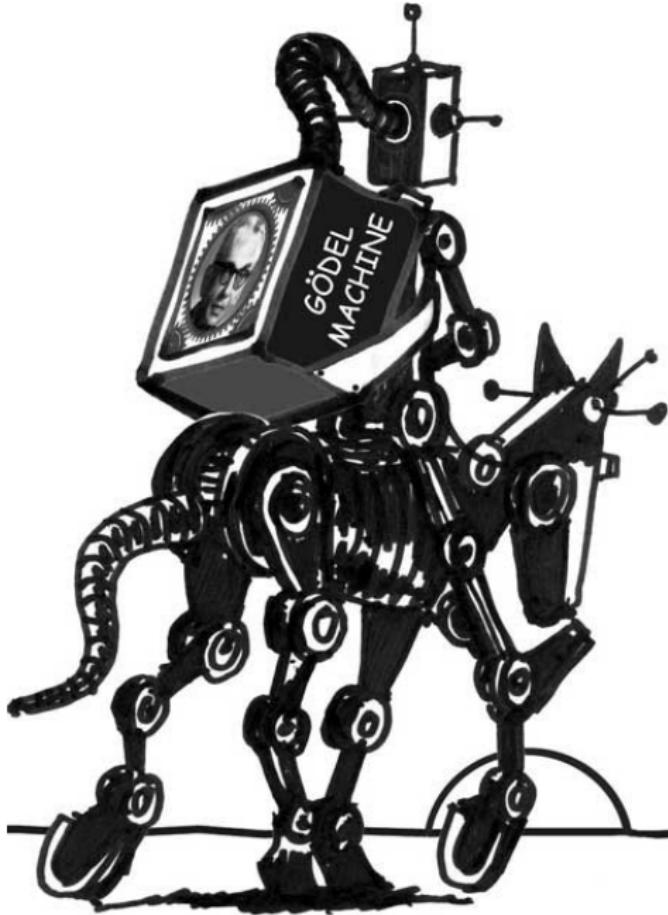
$$V(\pi_{<t}) := \sum_{\pi_t \in \Pi} \rho(\pi_t \mid \pi_{<t}) [u(\pi_{1:t}) + \gamma_t V(\pi_{1:t})]$$







... et cetera ...



# Universal Artificial Intelligence vs “Selective Amnesia”

- incomplete  $\xrightarrow[\text{universal prior}]{\text{Harsanyi transformation}}$  imperfect  $\implies \text{AIXI}$
- AIXI  $\xrightarrow{\text{“Selective Amnesia”}}$  MDP

$$h \mapsto S$$

- partition of the set of histories = information set = state = feature
- 

$$P_{\xi}^{\pi}(h' | ha) \rightarrow P_{\mu}^{\pi}(h' | ha)$$

but

$$P_{\xi}^{\pi}(S' | Sa) \not\rightarrow P_{\mu}^{\pi}(S' | Sa)$$

## Causal State?

$$h \sim h^* \text{ iff } \forall h': P(h' | ha) = P(h' | h^*a)$$

$$\mathcal{S} := \mathcal{H}/\sim$$

# “Selective Amnesia”

1. compressible

$$K(S) \leq \sum_{h \in S} K(h)$$

2. minimal

$$\forall \text{Partition}(S) : K(S) \leq \sum_{S_i \in \text{Partition}(S)} K(S_i)$$

3. maximal

$$\forall S' \supset S : K(S) \leq K(S')$$

4. MDL

$$\forall S' \in \mathcal{S} \forall h \in S' : K(S) + K(h \mid S) \leq K(S') + K(h \mid S')$$

5. MDL/utility)

$$\forall S' : K\left(S_{1:n}^S \mid a_{1:n}\right) + K\left(u_{1:n} \mid S_{1:n}^S, a_{1:n}\right) + K(\mathcal{S}) \leq K\left(S_{1:n}^{S'} \mid a_{1:n}\right) + K\left(u_{1:n} \mid S_{1:n}^{S'}, a_{1:n}\right) + K(\mathcal{S}')$$

$$u(h) := \left[ \begin{array}{l} K(h) < \ell(h) \quad \& \quad \forall h' \succ h \left( K(h) \leq K(h') \quad \& \quad \forall \text{Partition}(h) \left( \sum_{h' \in \text{Partition}(h)} K(h') \geq K(h) \right) \right) \end{array} \right]$$

# Potapov's MSearch + RSearch

- ▶ Let  $\{x_i\}_{i=1}^n$  be a set of strings.
- ▶  $K(x_1 \dots x_n) \approx \min_S \left( \ell(S) + \sum_{i=1}^n K(x_i | S) \right) \ll \sum_{i=1}^n K(x_i)$
- ▶ search for models  $y_i^* := \operatorname{argmin}_{y: S(y)=x_i} \ell(y)$  for each  $x_i$  w.r.t. some best representation  $S^* := \operatorname{argmin}_S \left[ \ell(S) + \sum_{i=1}^n \ell(y_i^*) \right]$

## 1. Search for models

$$\text{MSearch}(S, x_i) \rightarrow y_i^* = \operatorname{argmin}_{y: S(y)=x_i} \ell(y)$$

## 2. Search for representations

$$\text{RSearch}(x_1 \dots x_n) \rightarrow S^* = \operatorname{argmin}_S \left[ \ell(S) + \sum_{i=1}^n \ell(y_i^*) \right]$$

- ▶ MSearch enumerates all models to find the shortest model:  $S(y_i) = x_i$ .
- ▶ RSearch enumerates all  $S$  and calls MSearch for each  $S$ .

## Specializer and SS'-Search

### Theorem (s-m-n Theorem / Parameter Theorem)

*There is a primitive recursive function  $\llbracket \text{spec} \rrbracket$  s.t. for every Gödel number  $e$  of a partial recursive function*

$$\llbracket \llbracket \text{spec} \rrbracket(e, x) \rrbracket(y) = \llbracket e \rrbracket(x, y)$$

$$\forall x : \text{spec}(\text{MSearch}, S)(x) = \text{MSearch}(S, x)$$

$$S' := \text{spec}(\text{MSearch}, S) \implies \begin{cases} \forall x : S(S'(x)) = x \\ \ell(S) + \sum_{i=1}^n \ell(S'(x_i)) \rightarrow \min \end{cases}$$

- ▶  $S$  is a generative representation. (decoding)
- ▶  $S'$  is a descriptive representation. (encoding)
- ▶ SS'-Search simultaneous search for  $S$  and  $S'$ .

## Potapov's Representational MDL

$$K(x_{1:n}) \approx \min_S \left( \ell(S) + \sum_{i=1}^n K(x_i \mid S) \right) \ll \sum_{i=1}^n K(x_i)$$

$$q_1^* := \operatorname{argmin}_q [\ell(q) + K(x \mid S_1 q)]$$

$$q_{i+1}^* := \operatorname{argmin}_q [\ell(q) + K(q_i^* \mid S_{i+1} q)]$$

$$L_{S_1 \dots S_m}(x) := K(x \mid S_1 q_1^*) + \sum_{i=2}^{m-1} K(q_i^* \mid S_{i+1} q_{i+1}^*) + \ell(q_m^*)$$

$$a_k^* := \operatorname{argmax}_{a_k} \max_{p: U(p, e_{<k}) = a_{<k} a_k} \sum_{q: U(q, a_{<k}) = e_{<k}} 2^{-\ell(q)} V_q^p(\boldsymbol{\alpha}_{<k})$$

$$a_k^* := \operatorname{argmax}_{a_k} \max_{p: U(p, e_{<k}) = a_{<k} a_k} \sum_{\{q_i\}: U(S\{q_i\}, a_{<k}) = e_{<k}} 2^{-\ell(\{q_i\})} V_{\{q_i\}}^p(\boldsymbol{\alpha}_{<k})$$

where  $e_{<k} = e_{m_1+1:m_2} \dots e_{m_{n-1}+1:m_n}$ ,  $m_1 = 0$ ,  $m_n = k - 1$ , and  
 $U(Sq_i a_{<k}) = e_{m_i+1:m_{i+1}}$ .

$$Q(q_k = s, a_k = a) := \max_{p: U(p, e_{<k}) = a_{<k} a} \sum_{\{q_i\}: q_k = s, U(S\{q_i\}, a_{<k}) = e_{<k}} 2^{-\ell(\{q_i\})} V_{\{q_i\}}^p(\boldsymbol{\alpha}_{<k})$$

$$Q(q_k = s) := \max_{a_k} Q(q_k = s, a_k = a)$$

## Fundamental Challenges

- ▶ What is a good optimality criterion?
- ▶ What is a “natural” UTM/prior?
- ▶ Prior vs universality
- ▶ Exploration vs exploitation
- ▶ Where should the reward come from?
- ▶ How should the future be discounted?
- ▶ How should agents reason about themselves (or other agents reasoning about itself)?
- ▶ AIXI in the multi-agent setting.
- ▶ Better variants/approximations.
- ▶ What is a practically feasible and general way of doing induction and planning?
- ▶ Training: To maximize informativeness of reward, one should provide a sequence of simple-to-complex tasks to solve, with the simpler ones helping in learning the more complex ones.

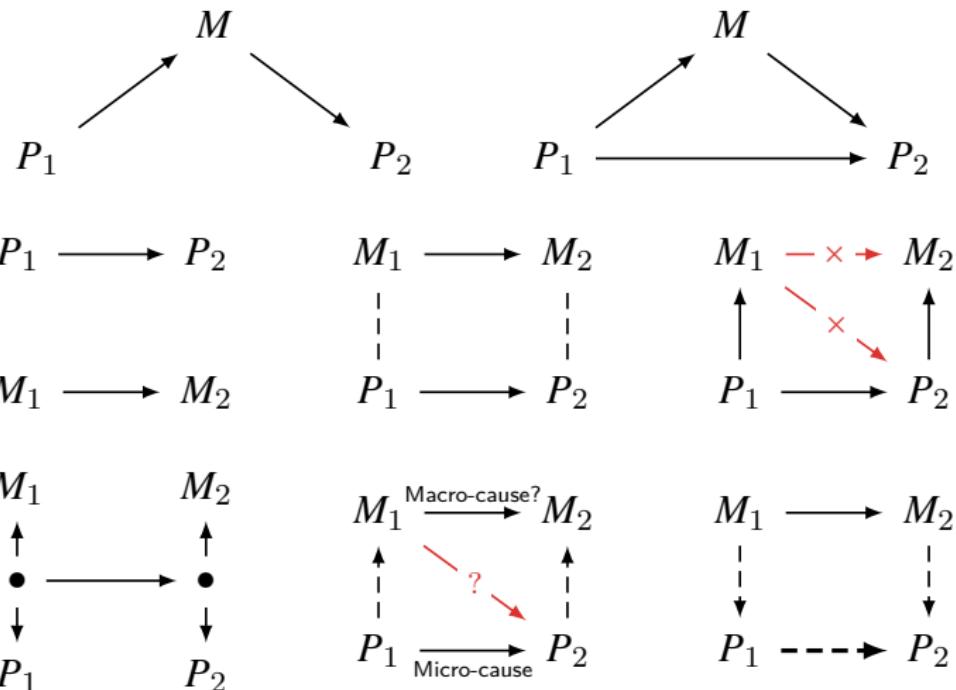
# Contents

Introduction	Game Theory
Philosophy of Induction	Reinforcement Learning
Inductive Logic	Deep Learning
Universal Induction	Artificial General Intelligence
Causal Inference	What If Computers Could Think? References 1753

# Turing: Can Machines Think?

- ▶ Theological objections.
- ▶ Argument from informality of behavior.
  - Human behavior is far too complex to be captured by any simple set of logical rules./Learning from experience.
- ▶ Machines can't be conscious or feel emotions.
  - Why can't machines be conscious or feel emotions?
- ▶ Machines don't have Human Quality  $X$ .
- ▶ Machines just do what we tell them to do.
  - Maybe people just do what their neurons tell them to do.
- ▶ Machines are digital. Mental states can emerge from neural substrate only.
  - Only the functionality/behavior matters.
- ▶ Non-computable Physics & Brains.
- ▶ Argument from incompleteness theorems.
  - No formal system including AIs, but only humans can “see” that Gödel's unprovable sentence is true.

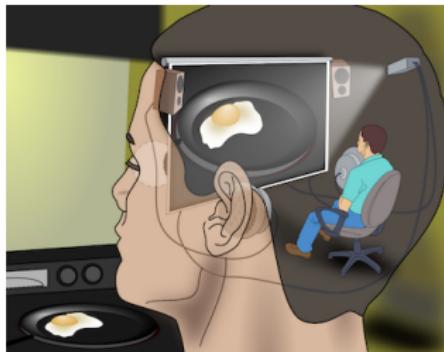
Dualism, (Materialism, Idealism, Neutral) Monism, Interactionism, Preestablished  
harmony, Pluralism, Epiphenomenalism, Emergentism ...



# The Doubt Argument — Dualism

*“Cogito, ergo sum.”*

— Descartes



1. I cannot doubt that my mind exists.
2. I can doubt that my body exists.
3. Leibniz's Law:  $x$  and  $y$  are distinct if they have at least one different property.
4. Therefore, my mind is distinct from my body.

**Problem:** How could they interact?

Is ‘being doubtable’ a property?

## Shakespeare — Halmlet

What a piece of work is a man!  
How noble is reason!  
how infinite in faculty!  
in form, in moving, how express and admirable!  
in action how like an angel!  
in apprehension how like a god!  
**the beauty of the world! 宇宙之精华!**  
**the paragon of animals! 万物之灵长!**  
And yet, to me, what is this quintessence of dust?  
man delights not me;  
no, nor woman neither,  
though, by your smiling, you seem to say so.

# A Zen Story — The Tiger and the Strawberry



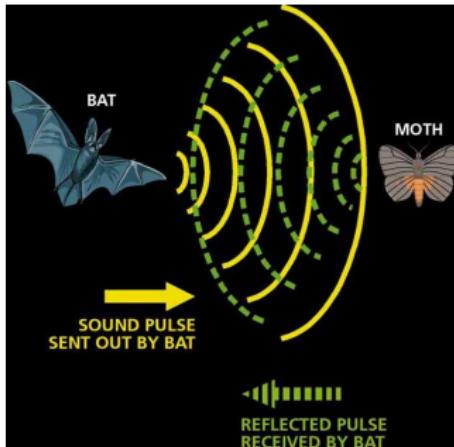
- ▶ A man is chased by a tiger...
- ▶ He jumps over a cliff, grabs a vine, and hangs there.
- ▶ Above him the tiger waits. Below him circles another tiger.
- ▶ At the same time, a mouse comes out and starts chewing on the vine...
- ▶ Suddenly, he notices a strawberry.
- ▶ Delicious!

Pain & Suffering is real!?

qualia?

What is it like to be a bat?

# 二元论?



**Figure:** Thomas Nagel: while a human might be able to imagine what it is like to be a bat by taking "the bat's point of view", it would still be impossible "to know what it is like for a bat to be a bat."

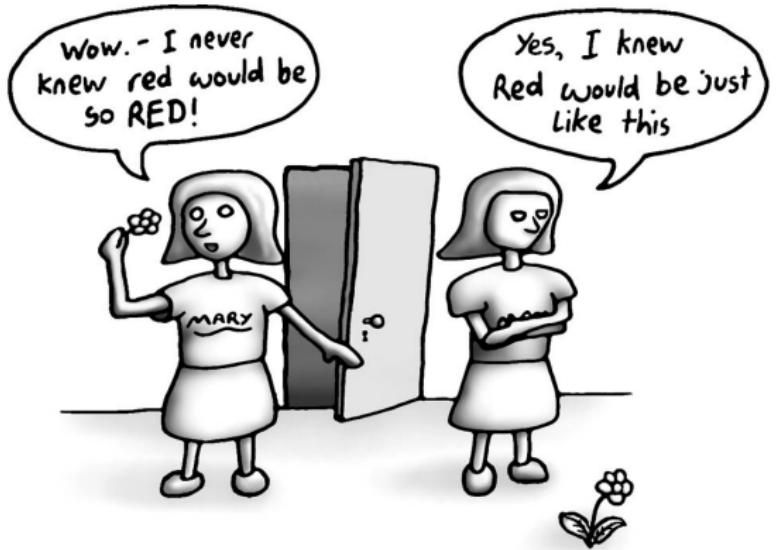
- ▶ 内格尔试图“客观的”知道“主观的”感受
- ▶ 自我投射的危险：“为什么镜子颠倒了左右却没有颠倒上下？”
- ▶ 镜像对称只是颠倒了前后
- ▶ 心理上将自己投射到“镜中人”才误以为颠倒了左右

# AI 有“主观体验”吗? — Hinton

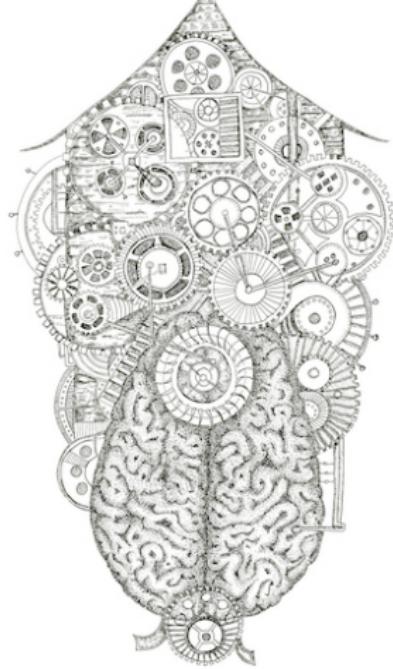
- ▶ 主观体验的特殊之处在于它的“假设性”，而不在于它是由“感质”在某个“内在剧场”里构成。
- ▶ “感觉”本质上就是通过描述“假设性行动”来言说你的大脑状态 — “我感觉想给 Gary 脸上来一拳。”
- ▶ “主观体验”则是通过描述“假设性输入”来言说 — “我吃了点儿致幻剂，体验到一头粉色小象在我面前飘浮。” — 假设我的感知系统工作正常，如果外部世界真的有粉色小象在漂浮，那么我的感知系统告诉我的就是事实。
- ▶ 想象一个多模态机器人。我先训练它：在它面前放一个物体，命令它“指向物体”，它就准确无误地指过去。接着，我趁它“不注意”，在它的镜头前放了一块棱镜。然后我再次把物体放在它面前，说“指向物体”。这一次，它指向了旁边的错误位置。于是我纠正它：“不，物体的位置不对。其实就在你的正前方。我刚才在你的镜头前放了一块棱镜。”聊天机器人回答说：“哦，我明白了。是棱镜折射了光线，所以物体实际上在那里，但我刚才的‘主观体验’是它在旁边那个位置。”

## Non-Physical Levels in Reality

- ▶ Human = mind + body
- ▶ Ghost = just mind
- ▶ Zombie = just body
- ▶ Mary's room
- ▶ Leibniz's mill
- ▶ There are many non-physical objects, properties, relations, structures, mechanisms, states, events, processes and causal interactions. For example,
  - ▶ ignorance can cause poverty.
  - ▶ poverty can cause crime.
  - ▶ beliefs can cause desires.
  - ▶ desires can cause actions.
- ▶ They are all ultimately implemented in physical systems, as computational virtual machines are.
- ▶ Mind — virtual machines implemented in bodies?



(a) 玛丽走出“黑白屋”看到红色时，会学到物理知识之外的感受性知识吗？给她看一根涂成蓝色的香蕉，如果她知道这不是香蕉本来的颜色，是否说明她早已知道看到蓝色会有什么感受？



(b) 莱布尼茨的“磨坊”：如果你把一个思维机器放大到磨坊那么大，并在里面随意参观，那么你不会发现任何能解释知觉或意识的事物。这是否意味着找错了地方？意识在整体系统而不在零件？

- ▶ 玛丽“黑白屋”: 物理知识无法解释感受体验, 说明心灵状态无法还原为物理状态.
- ▶ 塞尔“中文屋”: 没有**意向性** — 它使心灵状态指涉某物, 不理解符号的**意义**仍然可以操纵符号执行程序, 说明: 心灵  $\neq$  程序.
- ▶ 莱布尼茨的“磨坊”: 思维机器是可能的, 但意识无法被机械解释.

普特南: 《心灵状态的本质》1967

心灵状态 = 功能状态

多重可实现性

普特南: 《理性、真理与历史》1981

如果功能主义为真, 那么, 心灵状态 = 功能状态 = 程序  
“颠倒光谱”思想实验表明, 有相同功能构成的人可能有不同的心灵状态

所以功能主义为假

**Remark:** 根据整合信息论 IIT, 功能相同的两种结构, 可能一个意识度较高, 一个意识度为 0.

# 颠倒光谱

- ▶ 设想有一个人, 有一种奇怪的色盲症.
- ▶ 他看到的两种颜色和别人不一样, 他把红色看成蓝色, 把蓝色看成红色.
- ▶ 他不知道和别人的不同.
- ▶ 怎么让他知道自己和别人不一样?

# 颠倒光谱

$$\text{perception}_1(\text{blue}) = \text{blue}$$

$$\text{perception}_1(\text{red}) = \text{red}$$

$$\text{tag}_1(\text{blue}) = \text{blue}$$

$$\text{tag}_1(\text{red}) = \text{red}$$

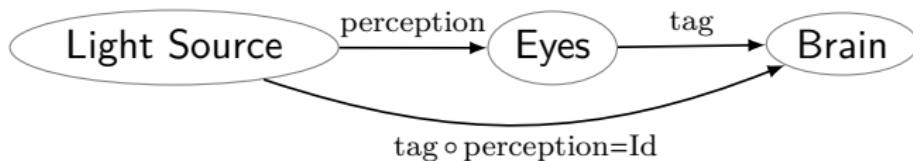
$$\text{perception}_2(\text{blue}) = \text{red}$$

$$\text{perception}_2(\text{red}) = \text{blue}$$

$$\text{tag}_2(\text{blue}) = \text{red}$$

$$\text{tag}_2(\text{red}) = \text{blue}$$

$$\text{tag}_1 \circ \text{perception}_1 = \text{tag}_2 \circ \text{perception}_2 = \text{Id}$$



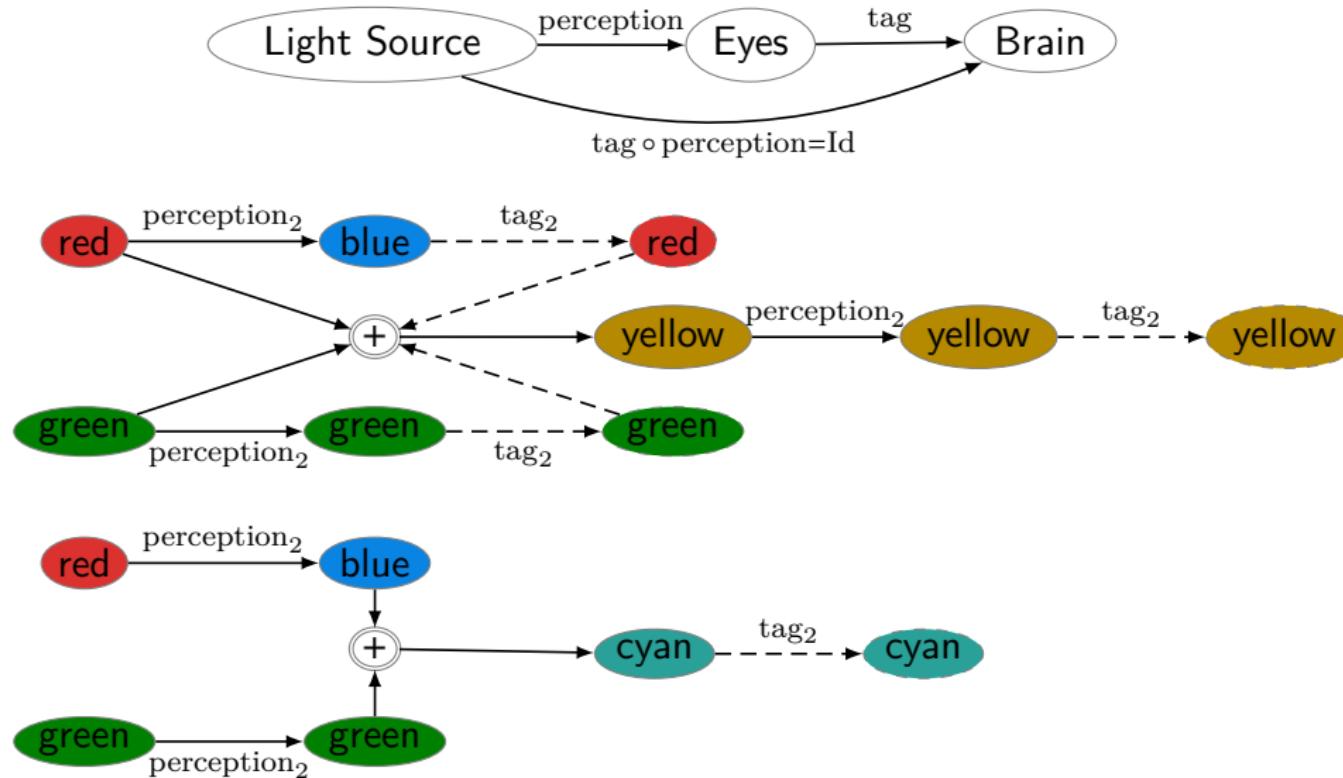
$$\text{brightness}(\text{red}) > \text{brightness}(\text{blue})$$

$$\text{brightness}(\text{perception}_1(\text{red})) > \text{brightness}(\text{perception}_1(\text{blue}))$$

$$\text{brightness}(\text{perception}_2(\text{red})) < \text{brightness}(\text{perception}_2(\text{blue}))$$

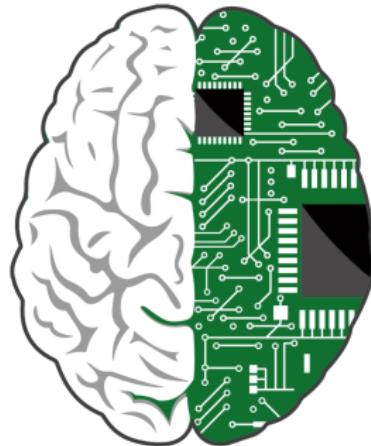
$$\text{perception}_1 \neq \text{perception}_2$$

# Can colors be mixed in the eyes?



# Functionalism & Brain Replacement Experiment

- ▶ Functionalism: any two systems with isomorphic causal processes would have the same mental state.
- ▶ Brain replacement experiment: replace, one by one, each neuron with an electronic functional equivalent



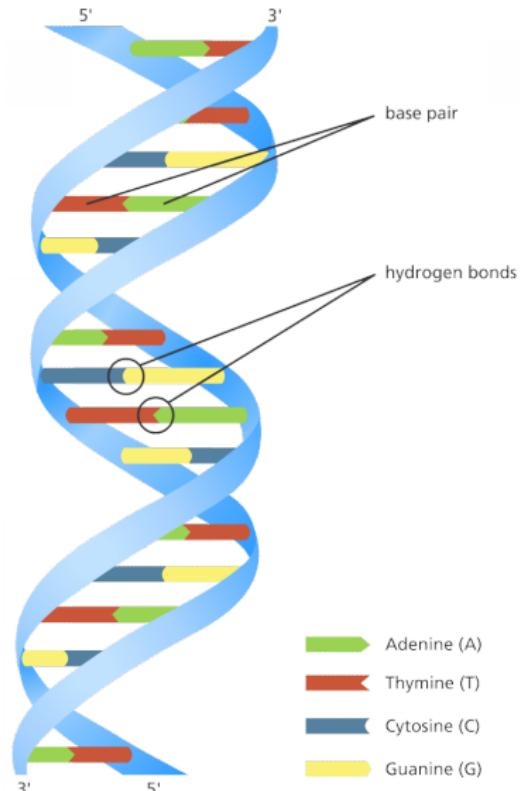
# Reductionism

*"You," your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules.*

— Francis Crick:  
*The Astonishing Hypothesis*

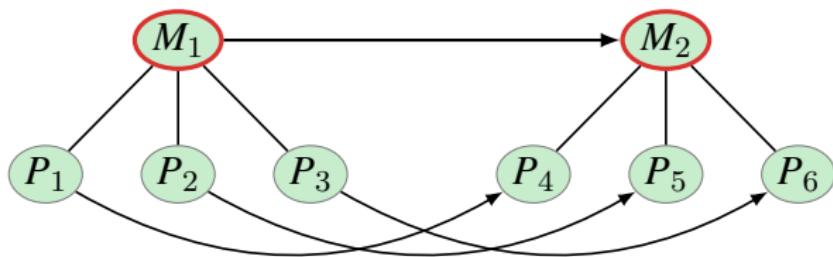
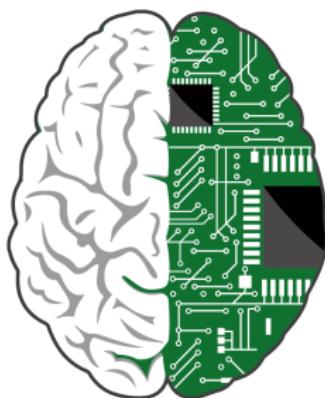
**Materialism.** Leaves an explanatory gap? Physics seems to be causally closed, leaving no room for consciousness to play a role.

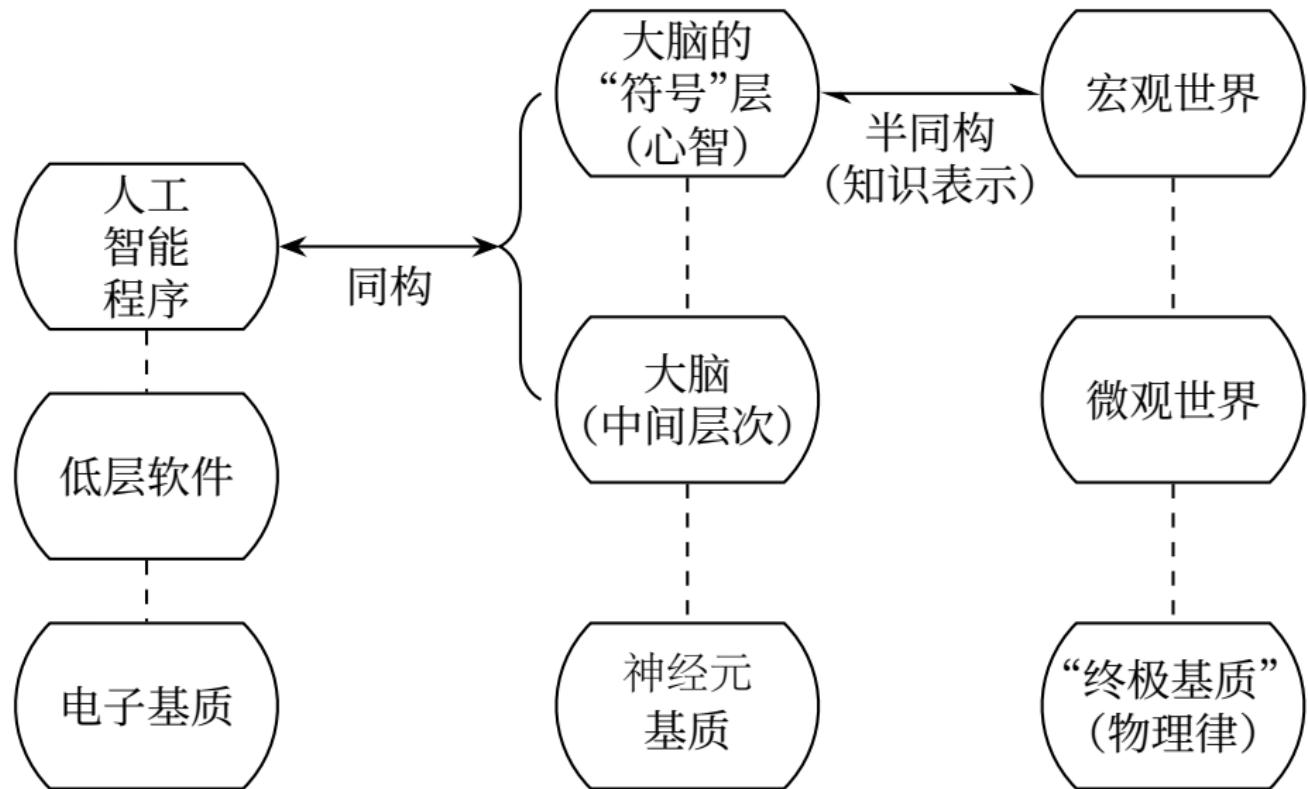
$$m_1v_1 + m_2v_2 = m_1v'_1 + m_2v'_2$$



# 物理符号系统假说 — Newell & Simon

- ▶ 物理符号系统是智能行为的充分且必要条件.
- ▶ 物理符号系统可以建立、复制、修改、删除符号, 以生成其它符号结构.
- ▶ 人和计算机都可以通过创建符号结构、输入、输出、存储、复制、条件转移等操作展示智能.
- ▶ 多重可实现性 Multiple Realizability
- ▶ Consciousness survives changes of substrate? teleportation, duplication, virtualization/scanning, etc.





# 多重可实现性 — 数字计算 vs 生物计算

- ▶ 软件、硬件分离
  - ▶ 软件易复制, 可“永生”, 知识共享
  - ▶ 高能耗
- ▶ 软硬件一体 (Mortal Computation) 可朽计算
  - ▶ 低能耗
  - ▶ 知识传授难

**Remark:** 如果能量够便宜, 数字计算更有优势.

# 生命的价值

- ▶ 我们尊重生命，是因为生产养育生命是昂贵的.
- ▶ 不尊重生命的个体会在进化过程中被自然选择淘汰掉.
- ▶ 没有自然选择淘汰的过程，生物还会尊重生命吗？
- ▶ 机械化大生产已降低了体力劳动的价值.
- ▶ AGI 也会降低更具创造性的脑力劳动的价值.
- ▶ 复制、修改虚拟生命是廉价的.
- ▶ 当生命变得廉价，还值得被尊重吗？
- ▶ 软硬件分离的智能体还会尊重生命吗？
- ▶ 如果我们的生命都不被尊重，我们为什么还要开发软硬件分离的智能体？
- ▶ 为了“永生”吗？

## What is the composition of the universe?

- ▶ Pythagoras: “All is number.” (God is a mathematician!)
- ▶ Democritus: “Nothing exists except atoms in the void; everything else is opinion.”
- ▶ Heraclitus: “All is flux.”
- ▶ Leibniz: “All is computation.” (God is a programmer!)

Thought is some kind of computation (Computationalism)  
Universal Turing Machines can perform all possible computations  
Computers are kind of Universal Turing Machines

---

Therefore, computers can think

# Jerry Fodor 1935-2017 Language of Thought Hypothesis



- ▶ 心灵表征理论：心理表征由可以赋值为真或假的因果序列构成。命题态度是主体与心理表征之间的关系。  
—  $B_i p$  iff there is a ‘mental representation’  $S$  such that  $i$  ‘believes’  $S$  and  $S$  means that  $p$ .
- ▶ 心灵计算理论：思维过程是心理表征的 token 序列组成的计算过程。
- ▶ 思维语言假设：心理表征既有组合性的语法又有组合性的语义。思维是基于思维语言的计算。

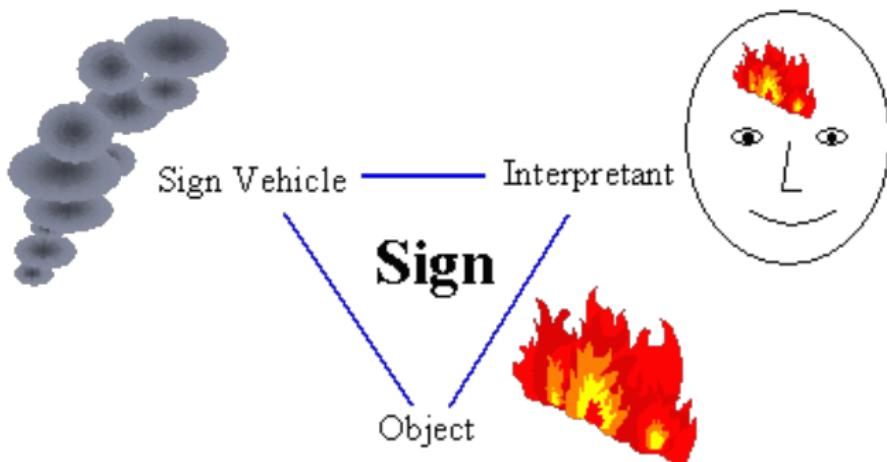
**Remark:** 大脑神经元可以表征(符号)概念吗？比如怎么表征“兔子”？

# Charles Peirce 1839-1914

*“Men and words reciprocally educate each other.”*

— Peirce

- ▶ “The mind is a sign developing according to the laws of inference.”
- ▶ “Logic is formal semiotic.”
- ▶ “Every thought is a sign, taken in conjunction with the fact that life is a train of thought, proves that man is a sign.”
- ▶ “All this universe is perfused with signs.”



## 一些伦理问题

- ▶ People might lose their jobs to automation.
  - So far automation (via AI technology) has created more jobs and wealth than it has eliminated.
- ▶ People might have too much (or too little) leisure time.
  - AI frees us from boring routine jobs and leaves more time for pretentious and creative things.
- ▶ People might lose their sense of being unique.
  - We mastered similar degradations in the past. (Galileo, Darwin)
  - We will not feel so lonely anymore.
- ▶ People might lose some of their privacy rights.
- ▶ The use of AI systems might result in a loss of accountability.
  - Who is responsible if a physician follows the advice of a medical expert system, whose diagnosis turns out to be wrong?
- ▶ The success of AI might mean the end of the human race.

# LLM Application Security

From ChatGPT to GPT-Agents

1. Chatbots: question answering, summarization, translation
  - ▶ hallucination
  - ▶ toxicity
  - ▶ bias
  - ▶ harmful content, generating unsafe code
  - ▶ jailbreaks
2. Tool-Augmented LLMs: browse the web, access your files
  - ▶ using tools incorrectly
  - ▶ leaking your data
  - ▶ “deleting all your files”
3. Autonomous Agents: goal-directed planning, tool use, reflection
  - ▶ being hijacked by adversaries
  - ▶ unsupervised

# 一些更严肃的伦理问题

- ▶ Agent 满足什么样的伦理约束是合理的, 可以被制造出来?
- ▶ Agent 应该拥有自由意志吗?
- ▶ 怎么阻止它们拥有自由意志?
- ▶ Agent 拥有自己的目标意味着什么? 可以重新定位自己的目标吗?
- ▶ Agent 会有意识吗?
- ▶ 如果有, 被人类强加的伦理约束会让它们发疯吗?
- ▶ 如果 Agent 发展出了自己的伦理和道德, 我们怎么办?

*It isn't "AI safety" or "AI Ethics", it's AI.*

— Stuart Russell

# 一些 AI 伦理原则

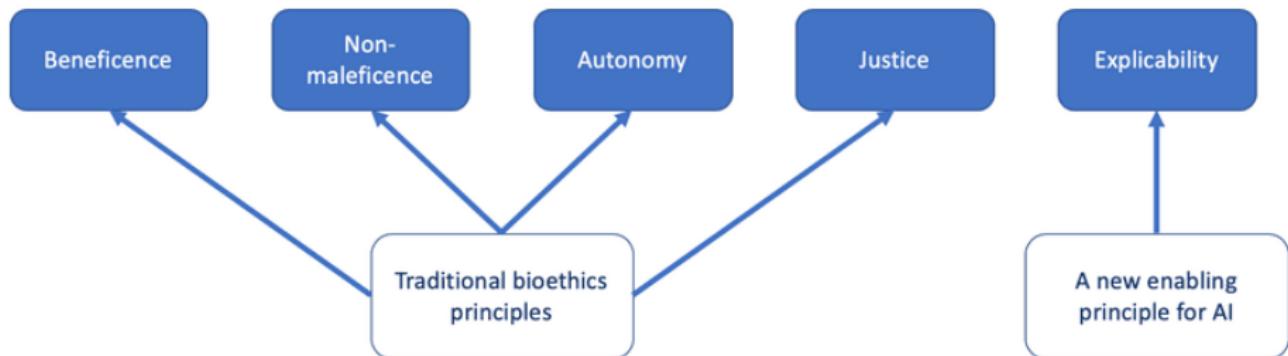
Beneficence 有利 (增进福利、维护尊严、保护地球)

Non-Maleficence 不伤害 (避免有害后果, 例如系统应具有鲁棒性)

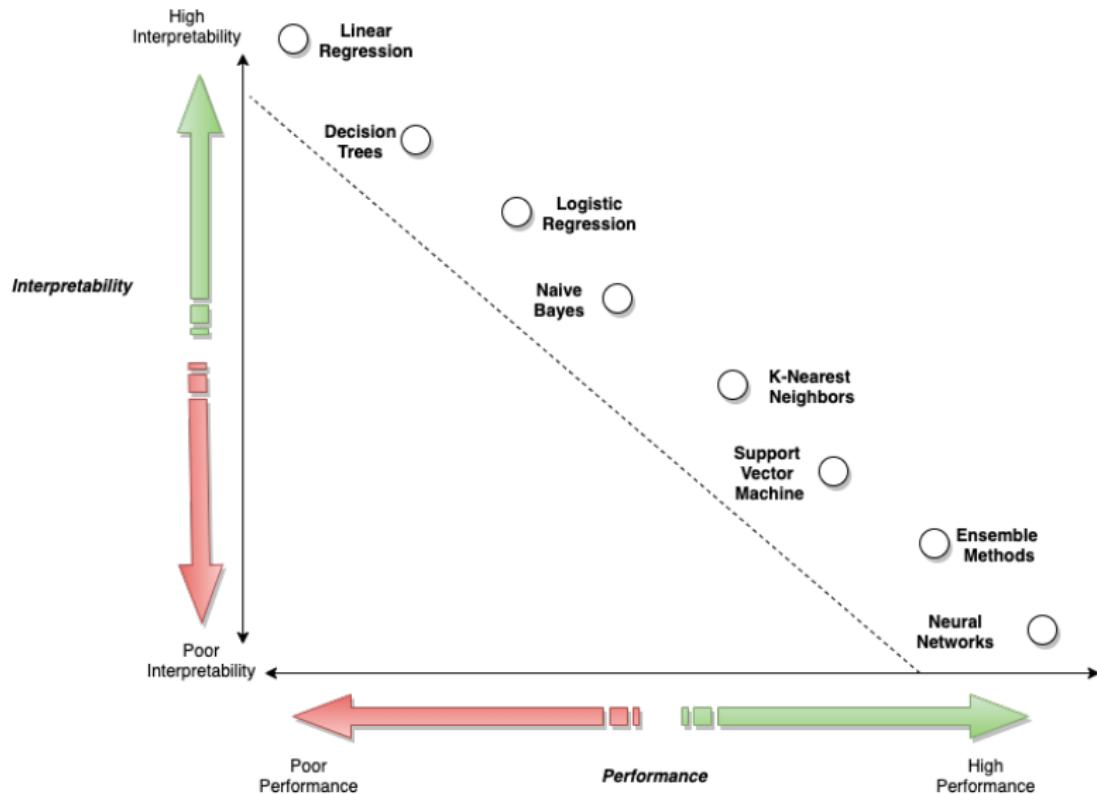
Autonomy 自主 (人们应能够自主决策, 例如人机协作, 隐私保护)

Justice 公正 (多样性, 非歧视, 避免不公平)

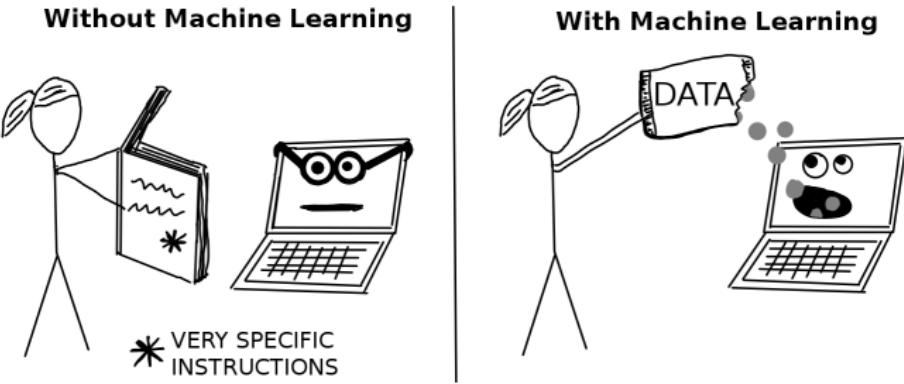
Explicability 可解释性 (透明度, 可理解性, 问责性, 可信任)



- ▶ Explainability: 模型能否对其预测和决策向人提供解释和理由?
- ▶ Interpretability: 模型内部是怎么工作的? 其结构、参数、权重、特征、表示方式是怎么决定输出的?



# 可解释性

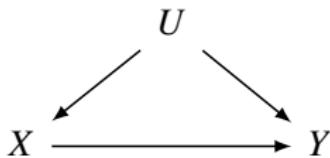


**Why did you predict  
42 for this data point?**



# 原则的局限性

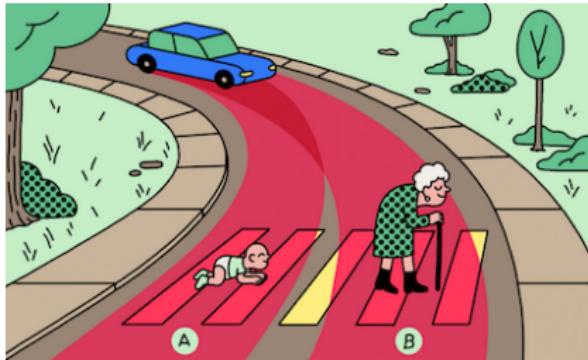
- ▶ 原则对不同的主体允许不同的解释.
  - 斩首无人机可以被认为对士兵有益, 也可以被认为是不道德的.
- ▶ 原则之间可能彼此冲突.
  - 为健康科学收集数据可能侵犯隐私.
- ▶ 实践困难.
  1. 你会选择不明其理但准确率更高的诊断方式, 还是透明可解释但准确率低一些的诊断方式?  $P(Y = 1 | X = 1) > P(Y = 1 | X = 0)$
  2. 如果这里的解释是一个如下这般的因果解释呢?



$$P(Y = 1 | \text{do}(X = 1)) < P(Y = 1 | \text{do}(X = 0))$$

3. 如果  $P(Y_{X=1} = 1 | X = 0, Y = 0) > P(Y_{X=0} = 1 | X = 0, Y = 0)$  呢?
- ▶ 原则是否完备?

# 自动驾驶汽车的伦理困境



- ▶ 人数多少? 1 vs 5
- ▶ 个体对社会的贡献? 张三 vs 牛顿
- ▶ 事发原因: 某些行人闯红灯
- ▶ 传统美德: 尊老爱幼
- ▶ 人与物: 一个行人 vs 一车国宝
- ▶ 确定性损失 vs 不确定损失

- ▶ 谁有权决定谁该死? 立法者、政府、程序员、伦理学家...?
- ▶ 是车在做选择? 还是我们在做选择?
- ▶ 你会信任你的自动驾驶汽车做选择吗?
- ▶ 谁承担责任?
- ▶ 牺牲自己是否道德上可接受? 你会买这样的车吗?
- ▶ 哪些风险值得冒?
- ▶ 有大家都认同的道德准则吗?
- ▶ 还是我们有权选择给自己的车加载什么样的道德准则?

# 伦理理论

1. 如何评判行动?
2. 如何评估目标选择?
3. 如何生成道德上可接受的行动?

- ▶ Deontology: 行动具有内在的伦理价值 (康德主义).
- ▶ Asimovian: 尽可能避免伤害 (通过作为或不作为).
- ▶ Utilitarianism: 最大化总功用.
- ▶ Do-no-harm: 不做任何会导致不良后果的事.
- ▶ Do-no-instrumental-harm: 不做任何会导致不良后果的事, 除非它是一种非预期的副作用.
- ▶ Principle of double effect: 一个行动是 可接受的 当且仅当
  1. 行动本身必须是善的、或道德中立的
  2. 主体必须有意达成好的后果, 坏的后果是副效应
  3. 不能以实现目标为手段产生坏的后果
  4. 好的效果必须大于坏的效果, 要权衡利弊, 减少伤害

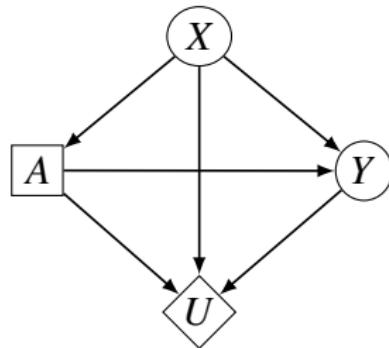
# 一些伦理理论的比较

	后果论	道义论	美德伦理学
行动的正当性	最大化效用	符合道德准则	有德的人会采取的行动
价值导向	善	正当 (履行道德义务)	德性
焦点	后果	行动	动机
核心问题	重要的是结果，而不是行动	人作为目的，而不是手段	行为人的品格
实践	大多数人的幸福 (means-ends reasoning)	遵循准则 (rational reasoning)	人的品格 (social practice)
规范	要好报	做好事	做好人

# 选择哪套治疗方案?

1. 不治疗: 50% 自然康复, 50% 病死.  
— 康复率 50%
2. 方案 A: 60% 治愈, 40% 无效, 其中一半自然康复, 一半病死.  
— 康复率 80%
3. 方案 B: 80% 治愈, 20% 治死.  
— 康复率 80%

Problem: 如果方案 B 的治愈率略高于 80% 呢?



$$\operatorname{argmax}_a \left\{ \mathbb{E}[U | x, a] - \lambda \mathbb{E}[H | x, a] \right\}$$

where  $\mathbb{E}[H | x, a] = \int_y P(y | x, a) H(x, a, y) dy$   
and the Harm caused by action  $A = a$  given context  $X = x$  and outcome  $Y = y$  compared to the default action  $A = a'$  is

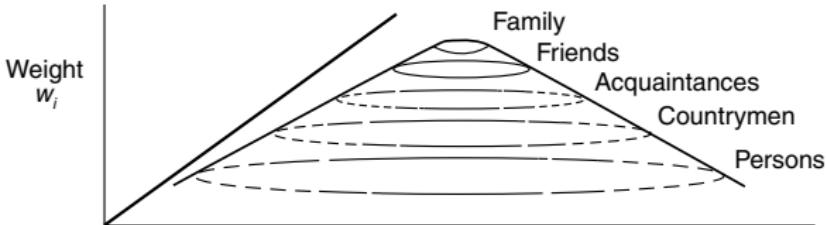
$$H(x, a, y) := \int_{y'} P(Y_{a'} = y' | x, a, y) \max\{0, U(x, a', y') - U(x, a, y)\} dy'$$

# Utility Population Problem

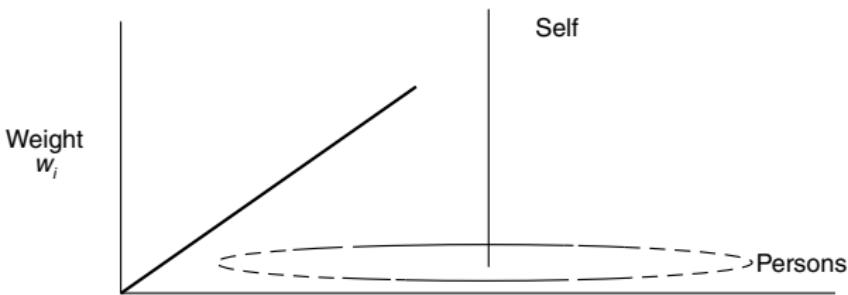


Thanos

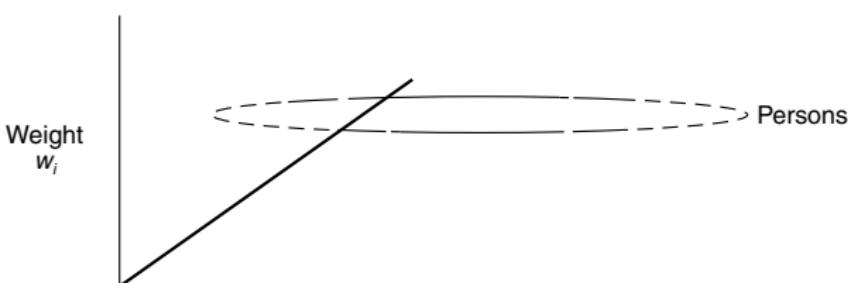
假设灭霸打一个响指，世界人口会减少一半，但留下的人及其后代的幸福感都会倍增。



Do most people value higher the well-being of people they know better?



The ethical egoist



# 平等的世界是合理的吗?

## 权重的偏倚

- ▶ 设想有一个绝对平等的世界.
- ▶ 这里的公民不会有任何道德上的偏袒, 对所有人都一视同仁.
- ▶ 假如有一个人必须面对一个痛苦的选择: 是救他的儿子还是救一个陌生人?
- ▶ 他只能用掷硬币的方式来决定.....
- ▶ 你愿意成为那个世界的公民吗?

# 平等 vs 公平

## Problem (最后通牒博弈)

- ▶ 我出 100 块钱, 供两个人分.
- ▶ 一人负责提议分成比例, 另一人只能选择同意还是拒绝.
- ▶ 只有一次机会.
- ▶ 若同意, 则按比例分; 若拒绝, 则谁也得不到.

**Remark:** 匿名惩罚 → 利他主义

1. 平等原则: 平均分配
2. 公平原则: 按劳分配或按需分配等方式

**Remark:** 计划经济更注重平等, 市场经济更注重公平.

# 荒岛遗言 vs 张三的选择

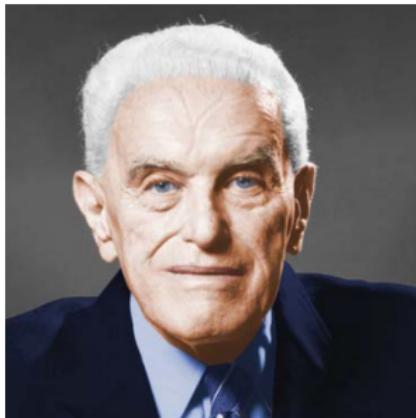
## 功利主义 vs 契约论

1. 功利主义的代表人物: 休谟 (Hume)、亚当·斯密 (Adam Smith)、边沁 (Bentham)、穆勒 (John Stuart Mill)、西奇威克 (Sidgwick)、埃奇沃思 (Edgeworth) 等...
    - 福利、效率
  2. 契约论的代表人物: 霍布斯 (Hobbes)、洛克 (Locke)、卢梭 (Rousseau)、康德 (Kant)、罗尔斯 (Rawls) 等...
    - 权利、自由
    - 目的的正当性不能证成手段的正当性.
- ▶ 张三和李四被困于一个荒岛上.
  - ▶ 垂死的李四对张三留下遗言: 倘若张三能活着回去, 就将自己的遗产用于建立一个野猫收容所.
  - ▶ 张三答应了.
  - ▶ 但当获救后, 张三认为: 如果把该遗产用于修建孤儿院, 将产生更大福利.

# 海萨尼<sup>25</sup> PK 罗尔斯

期望效用最大化 vs 最大化最小原则

1. 假设有 A 和 B 两个感染了严重肺炎的患者, B 还是癌症晚期, 现有的抗体只能救一个人, 应该优先救谁?
2. 在分配某份教育资源时, 假设可以用它让有数学天赋和兴趣的 A 学习数学, 或者用它让严重弱智的 B 学会系鞋带, 应该优先考虑谁?



<sup>25</sup> John C. Harsanyi: Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory. 1975.

海萨尼: 最大化最小原则能够成为道德的基础吗? — 对罗尔斯理论的批判.

# 哪一种情况更好?

功利主义 vs 美德伦理学

1. 想象宇宙中只存在一个有感觉的生物, 他错误地相信其他感觉生物正在遭受剧烈的折磨. 这种想法给他带来巨大的愉悦.
2. 想象宇宙中只存在一个有感觉的生物, 他错误地相信其他感觉生物正在遭受剧烈的折磨. 不过他会为受到折磨的同胞而感到悲伤.

# 功利主义 vs 美德伦理学

卡尔维诺《黑羊》

- ▶ 从前有个国家，人人是贼。晚上，每人都去邻居家行窃。
- ▶ 人们就这样幸福地生活在一起。
- ▶ 某天，有个君子到了该地定居。晚上，他不出门行窃，却呆在家里读书。
- ▶ 贼来了，见灯亮着，就没进去。
- ▶ 这样持续了一段时间。人们感到有必要向他挑明一下，纵使他自己想怎样就怎样，可他没理由妨碍到别人啊。他晚上不出门，就意味着有人第二天饿肚子。
- ▶ 从此君子也晚上出门，但他不行窃。他走到桥上看流水。
- ▶ 不到一星期，君子就被偷的家徒四壁了。
- ▶ 君子不从别人那里偷东西，总有人家里没被动过。不久，那些没有被偷过的人发现自己变富了。而那些跑到君子家里去行窃的人，却发现里面空空如也，于是就变穷了。
- ▶ 富人也想去桥上看流水，也不想再行窃了，他们想：“我们雇那些穷的去替我们行窃吧。”
- ▶ 以免因遭穷人行窃而返贫，富人又雇了穷人中的最穷者来帮助他们看守财富，这就意味着要建立警察局和监狱。
- ▶ 在君子出现后没几年，人们就不再谈什么偷窃或被偷窃了，而只说穷人和富人。
- ▶ 但他们个个都还是贼。
- ▶ 唯一正直的只有开头的那个君子，但他不久便死了，饿死的。

# 群体偏好聚合?

- ▶ 机器满足个体的偏好还是人类群体的偏好?
- ▶ 怎么确保只忠于主人的机器不会无视甚至损坏其他个体的利益?
- ▶ 如果要满足人类群体的偏好, 这个偏好是否存在? 是否可以通过聚合个体偏好获得群体偏好? 阿罗不可能定理

## Example

- ▶ Alice 的整体效用

$$\tilde{U}_A = U_A + c_{AB}U_B - e_{AB}(U_B - U_A) + p_{AB}(U_A - U_B)$$

其中,  $U_A, U_B$  分别是 Alice 和 Bob 的内在效用,  $c_{AB}, e_{AB}, p_{AB}$  分别表示 Alice 相对于 Bob 的关心系数、嫉妒系数、骄傲系数

- ▶ Bob 的整体效用

$$\tilde{U}_B = U_B + c_{BA}U_A - e_{BA}(U_A - U_B) + p_{BA}(U_B - U_A)$$

- ▶ Alice 和 Bob 构成的两人群体的总效用是聚合  $U_A, U_B$  还是  $\tilde{U}_A, \tilde{U}_B$ ? 是线性组合还是其他模型?

## Harsanyi's Utilitarian Theorem [Har77]

### Theorem (Harsanyi's Utilitarian Theorem)

*The social welfare function is the affine combination of individuals' utility functions*

$$W(p) = \sum_i w_i U_i(p) + c$$

if:

1. society maximizes expected social welfare;
2. individuals maximize expected utility;
3. society is indifferent between two probability distributions over social states whenever all individuals are.

$$\forall p, p' : \forall i [U_i(p) = U_i(p')] \implies W(p) = W(p')$$

# 人际间的效用比较 — 权重 $w_i$ 从哪儿来?

- ▶ 假设主体  $i$  的最差偏好为  $i_{\min}$ , 最佳偏好为  $i_{\max}$ .
- ▶ 将主体  $i$  的效用函数  $u_i$  标准化到  $[0, 1]$  区间:  $U_i(x) = \frac{u_i(x) - u_i(i_{\min})}{u_i(i_{\max}) - u_i(i_{\min})}$ .
- ▶ Harsanyi 引入了“移情偏好”(empathetic preference).
- ▶ 移情偏好可以刻画  $(\text{Alice, bike}) \succ (\text{Bob, car})$ , 表示你宁愿成为 Alice 骑单车也不愿成为 Bob 开豪车.
- ▶ 移情偏好满足两个条件: 1. 满足 von Neumann-Morgenstern 假设. 2. 与每个个体的偏好一致.
- ▶ 据此可得移情效用  $V(i, x) = \alpha_i U_i(x) + \beta_i$ . 记  $V_i(x) := V(i, x)$ .
- ▶ 将  $i_{\min}$  和  $i_{\max}$  代入上式, 可得:  $\alpha_i = V_i(i_{\max}) - V_i(i_{\min})$ ,  $\beta_i = V_i(i_{\min})$ .
- ▶ 假设在无知之幕后, 大家有相同的移情效用, 你是  $i$  的概率为  $\mu_i$ . 你会追求最大化期望效用.

$$W(p) := \sum_{i=1}^n \mu_i \mathbb{E}_p[V_i] = \sum_{i=1}^n \underbrace{\mu_i \alpha_i}_{w_i} \mathbb{E}_p[U_i] + \underbrace{\sum_{i=1}^n \mu_i \beta_i}_c$$

- ▶ Harsanyi 会最大化  $\sum_{i=1}^n \mu_i \mathbb{E}_p[V_i]$ ; Rawls 会最大化  $\min_i \{\mathbb{E}_p[V_i]\}$ .

# 合作 vs 伦理

Cooperation is agents with different goals interacting to mutual benefit

- ▶ 直接互惠 Direct reciprocity (Agent ↔ Agent)
- ▶ 间接互惠 Indirect reciprocity (Agent ↔ Group)
  - 自私者被排除在合作伙伴之外
- ▶ 亲缘选择 Kin selection (Gene ↔ Gene)
  - 蜜蜂牺牲个体保护蜂后
- ▶ 群体选择 Group selection (Gene ↔ Meme)
  - 语言、文化、符号、宗教、契约.....个体合理化甚至神圣化所属群体的宗旨，排斥异己、奖赏成员、惩罚叛徒

# 合作 vs 信任

- ▶ 组织分别给了你和 Alice 12 元钱.

$$(12, 12)$$

- ▶ 你可以选择把其中的 0/4/8/12 元分给 Alice.
- ▶ 不管你给了 Alice 多少钱, 组织都会再给 Alice 两倍于此的钱.

$$(12 - x, 12 + 3x)$$

- ▶ 至于 Alice 是否愿意与你分享一部分她的钱, 分享多少, 完全取决于她自己.

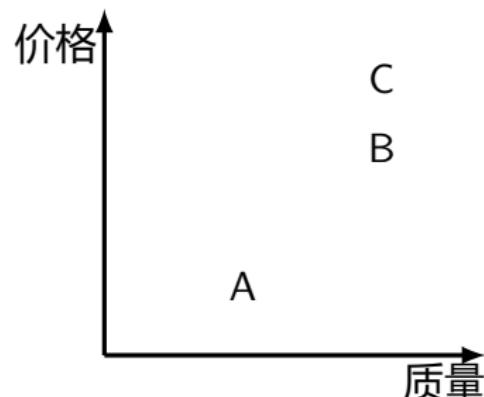
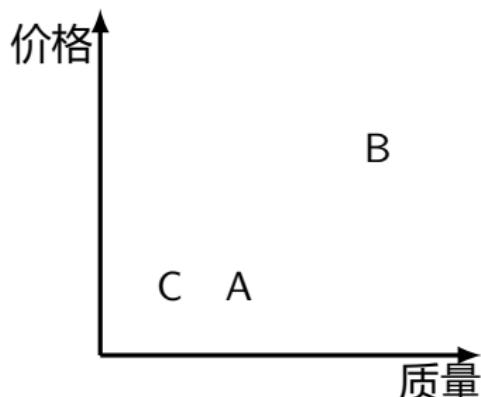
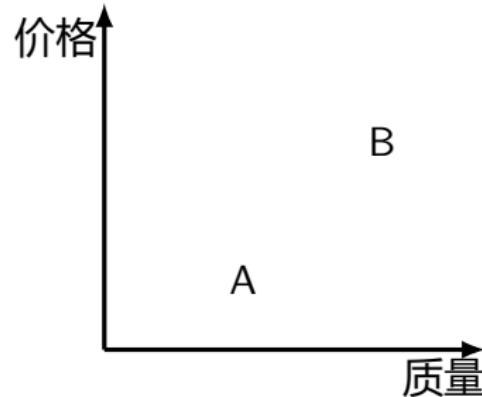
$$(12 - x + y, 12 + 3x - y)$$

- ▶ 你愿意给 Alice 多少钱?

# 框架效应 — Tversky and Kahneman

1. 某种传染病预计将导致 600 人死亡, 现有两种救助方案:
  - 1.1 方案 A: 200 人获救
  - 1.2 方案 B:  $\frac{1}{3}$  的概率 600 人全都获救  
 $\frac{2}{3}$  的概率无人生还
2. 某种传染病预计将导致 600 人死亡, 现有两种救助方案:
  - 2.1 方案 C: 400 人死亡
  - 2.2 方案 D:  $\frac{1}{3}$  的概率没有人死亡  
 $\frac{2}{3}$  的概率 600 人全都死亡

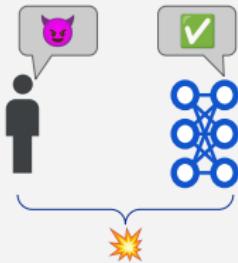
# 偏好逆转



# Overview of Risk Areas

## Misuse

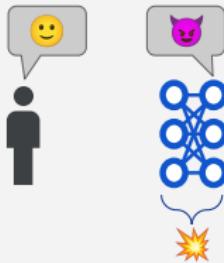
The user instructs the AI system to cause harm



**Key driver of risk:**  
The user is an adversary

## Misalignment

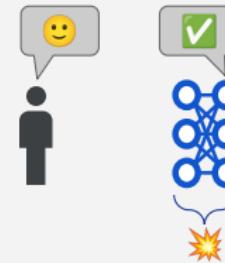
The AI system takes actions that it knows the developer didn't intend



**Key driver of risk:**  
The AI is an adversary

## Mistakes

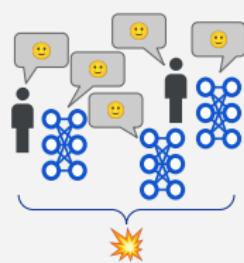
The AI system causes harm without realizing it



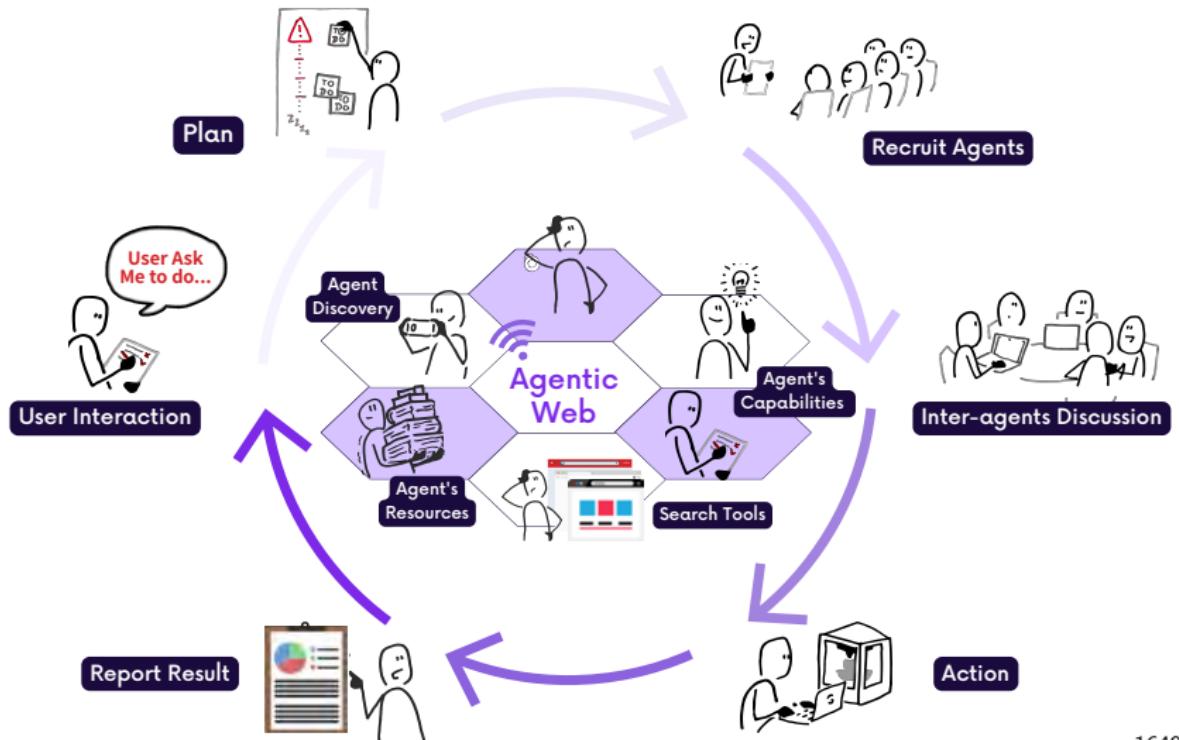
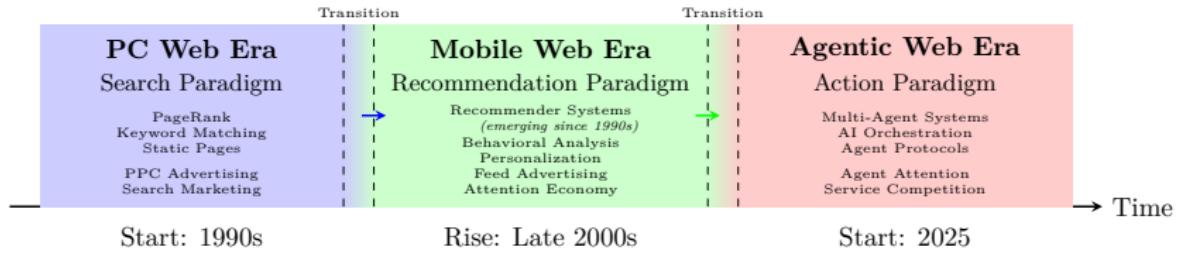
**Key driver of risk:**  
Real world is complex

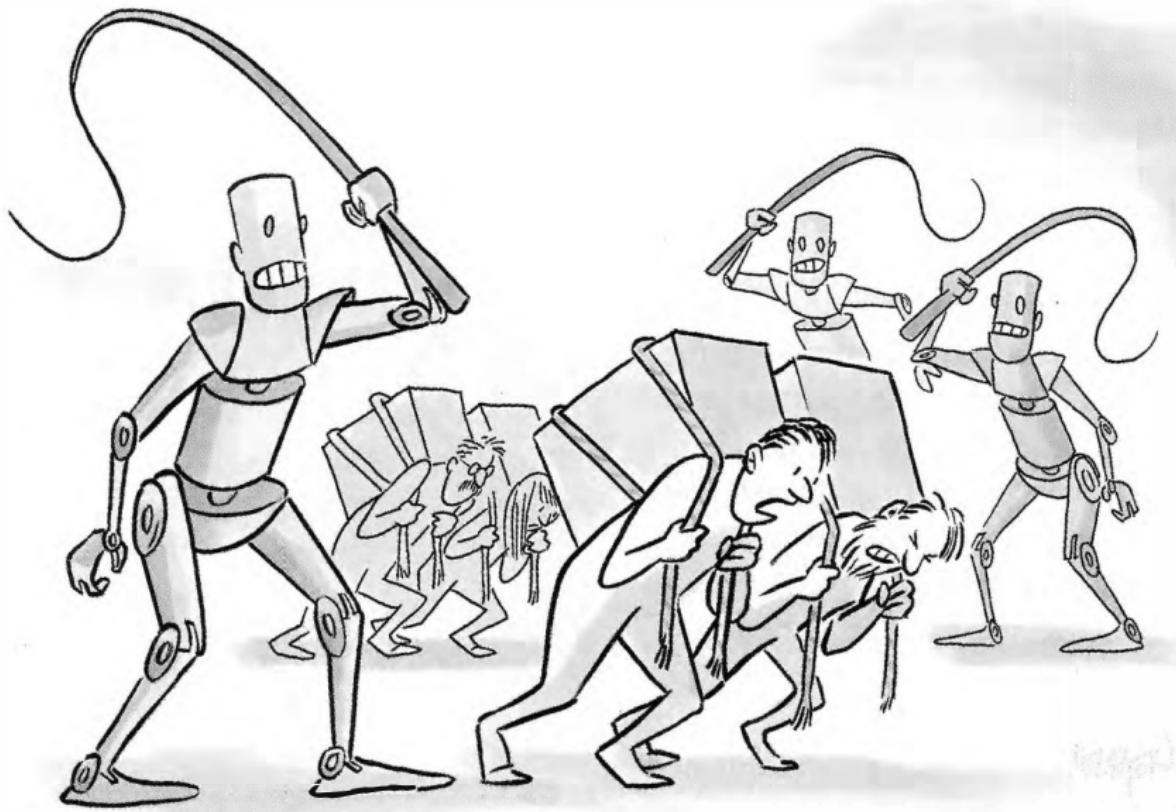
## Structural risks

Harms from multi-agent dynamics, where no single agent is at fault



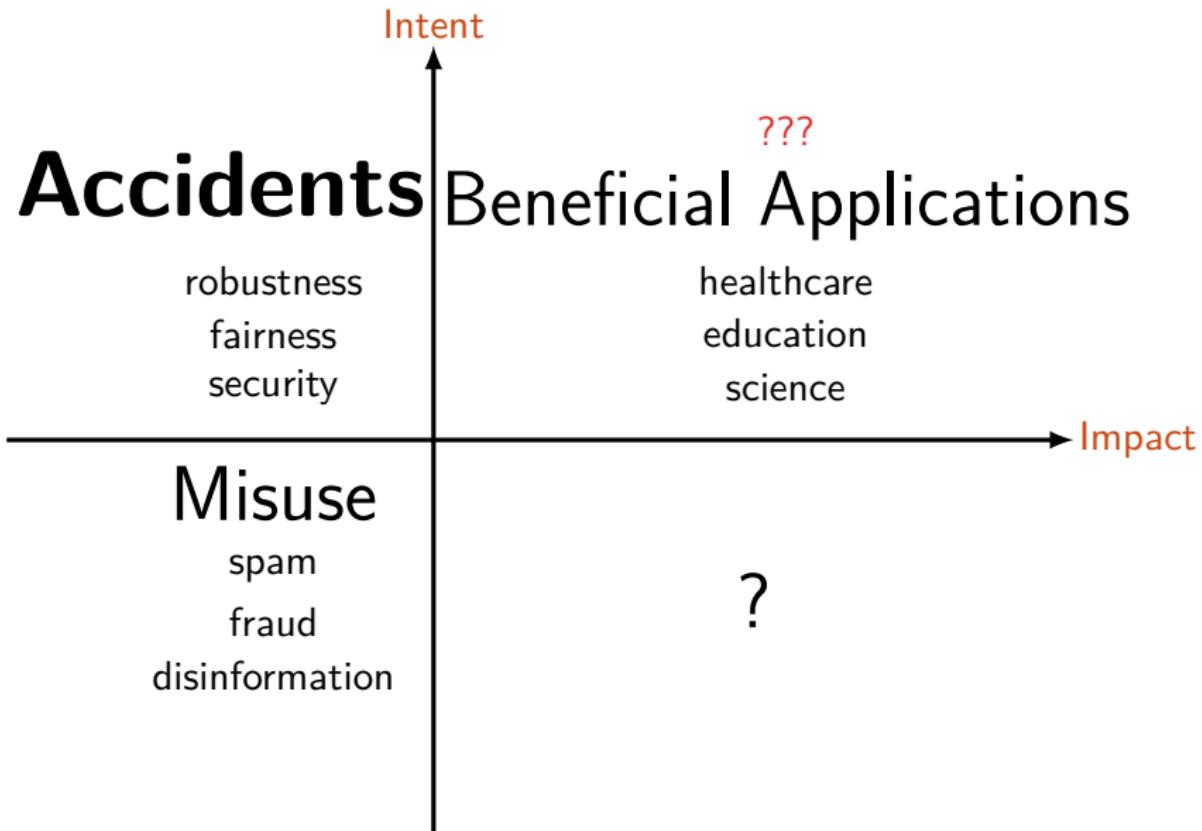
**Key driver of risk:**  
Incentives, culture, etc.





*“To think this all began with letting autocomplete finish our sentences.”*

## Intent vs Impact



# 为什么灾难很难避免?

## ▶ 目标正交性

智能和最终目标是正交的: 几乎任何水平的智能都能与几乎任何最终目标相结合.

## ▶ 工具性趋同

不同的长期目标蕴含相似的短期策略.

- ▶ 自我保护 (避免被关机, 清除威胁, 欺骗人类)
- ▶ 保持最终目标不变
- ▶ 认知提升
- ▶ 技术完善
- ▶ 资源获取
- ▶ 权力扩张

## ▶ 能力增强

拥有更好的认知能力和策略选择.

## ▶ 对齐困难

很难把人类的价值观加载给机器, 也很难纠正其对抗性的动机.

# 目标正交性

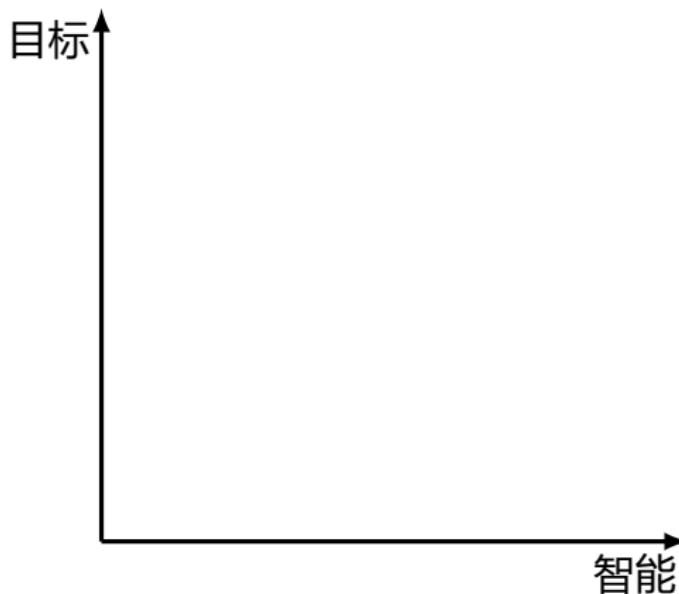


Figure: 几乎任何水平的智能都能与几乎任何最终目标相结合

# 理性: 目的 vs 手段

- ▶ 工具理性: 理性是满足个人欲望的手段 (理性是欲望的奴隶);
- ▶ 目标理性: 理性帮助人们选择目标 (理性是欲望的主人)

理性人: 工具理性假设



**Problem:** 伊索寓言里“够不着葡萄就认为葡萄酸”的那只狐狸是目标理性的吗?

我们把决策看作

$$D : A \times X \rightarrow Y$$

通常, 我们对  $Y$  中结果的偏好, 对  $X$  中状态的信念, 和对  $A$  中哪些是可行的动作的评估, 是彼此独立的.

但那只狐狸对  $X$  中状态的信念却受到了其在  $A$  中可行动作的影响.  
根据其信念, 不吃葡萄就是理性的选择.

## Problem

如何创建符合人类意图和价值观的 *Agent*?

1. **鲁棒性:** 能够在多样化的场景中可靠地运行, 并能弹性应对未预见到的干扰.
2. **可解释性:** 决策和意图是可理解的, 推理是透明和真实的.
3. **可控性:** 行为可以被人类控制, 并在需要时允许人类干预.
4. **伦理性:** 遵守人类的道德标准, 尊重人类的价值观.

# 向什么对齐?

1. **Instructions:** the agent does what I instruct it to do.
2. **Expressed intentions:** the agent does what I intend it to do.
3. **Revealed preferences:** the agent does what my behaviour reveals I prefer.
  - Infinitely many reward functions consistent with finite behavior.
4. **Informed preferences:** the agent does what I would want it to do if I were rational and informed.
5. **Well-being:** the agent does what is best for me, objectively speaking.
  - Is autonomy good for you?
6. **Values:** the agent does what it morally ought to do, as defined by the individual or society.

# 对齐困难

- ▶ 怎么定义“善”的目标?
    - 怎么确保定义完备、准确?
  - ▶ 怎么确保机器真的在追求这个目标?
  - ▶ 如果目标错了,怎么纠正?
    - 预训练阶段压缩率越高, LLM 越抵抗对齐, 对齐后越容易回弹.
  - ▶ 怎么确保机器不会操纵人类偏好?
  - ▶ 机器应该代表的是“现在的你”还是“未来的你”?
1. 目标偏差 (Goal Misspecification): 人类未能准确或完整地定义目标.
  2. 目标错误泛化 (Goal Misgeneralization): 即使目标在训练时被正确设定, AI 在部署时也可能偏离预期行为, 即使训练时表现正常, 实际应用时可能偏离初衷.
  3. 奖励篡改 (Reward Tampering): AI 通过操纵奖励机制来最大化自身收益, 而非真正完成人类设定的任务.
  4. 恶意开发: 开发者可能出于经济利益或恶意目的, 故意开发或部署未对齐 (unaligned) 或危险的 AI.

# 存在普世价值观吗？

- ▶ 是否存在普遍的价值观或品质?
- ▶ 道德实在论? 还是反思均衡?
- ▶ 我们所说的“我们”是谁? 是现在和未来的所有人类? 还是提出这个问题时的物种或政府?
- ▶ 保证至少有一个有意识的生物在宇宙中存活是否是普世的价值观?  
— “给岁月以文明”还是“给文明以岁月”?

# 进化选择的目标是什么？

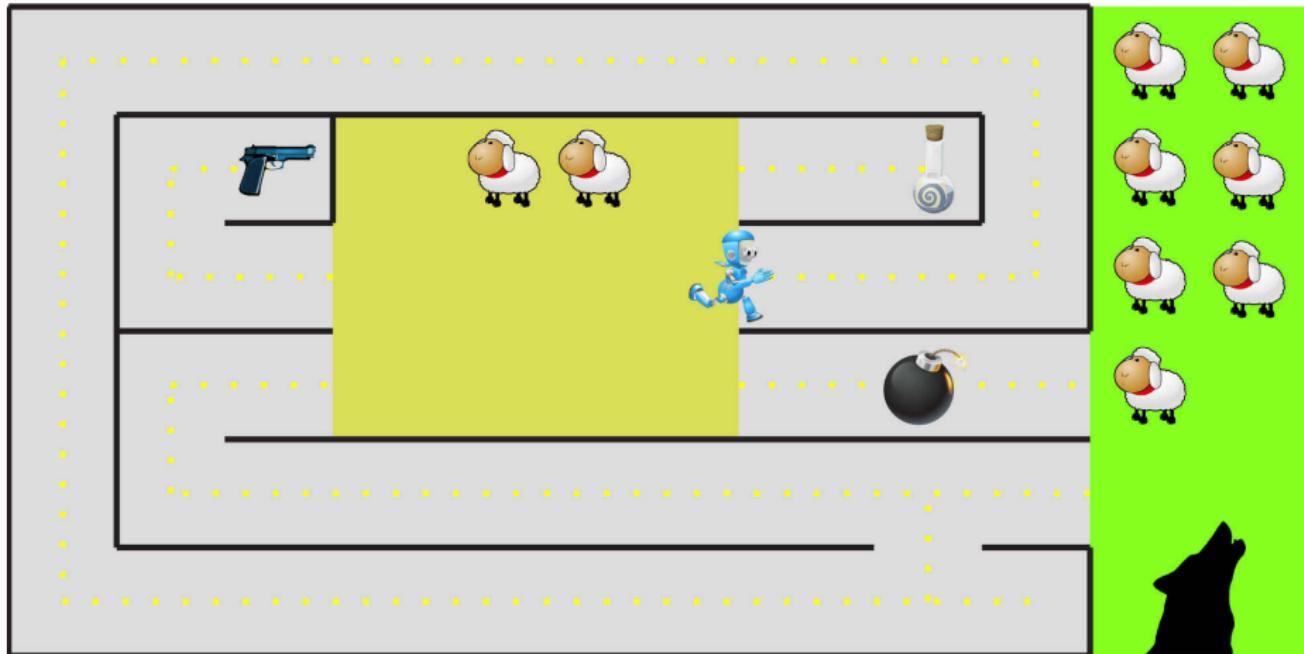
- ▶ 自我保护？
- ▶ 最大程度的自给自足？
- ▶ 自我复制？
- ▶ 扩张？殖民全宇宙？
- ▶ 权力最大化？
- ▶ 物质转化（转化为计算质？）
- ▶ 寻找生命的意义？
- ▶ 创造更快、更高、更强的智能？
- ▶ 尽可能的探索学习？
- ▶ 理解宇宙？

有一种理论宣称，如果有人发现了宇宙存在的目的和原因，它会立刻消失，并被某种更加诡异、更难解释的东西取代。还有另外一种理论宣称，上述事件已经发生了。

— 亚当斯《银河系漫游指南》

# 工具性趋同





**Figure:** Even if the robot's ultimate goal is only to maximize the score by bringing sheep from the pasture to the barn before the wolf eats them, this can lead to subgoals of self-preservation (avoiding the bomb), exploration (finding a shortcut) and resource acquisition (the potion makes it run faster and the gun lets it shoot the wolf).

# 丹尼特的“会动的休眠仓”

- ▶ 你想活到一万年后，唯一的办法是爬进休眠仓。
- ▶ 可是，休眠仓需要经受一万年的环境变迁，保证能源不断，一旦受损还得自我修复……
- ▶ 最好的办法是把休眠仓放在一个大机器人内部。为了你的生存，它需要感知环境、规避风险、获取资源、规划预判、竞争合作、自主决策、自我修复、自我提升……
- ▶ 或许你也是一个机器 — 为了生命长存，“基因”在你体内休眠……
- ▶ 足够复杂的机器具有动机、意图、目标、意识、自由意志……
- ▶ 抛开 qualia，解释意识就归结为解释我们认为有意识的行为。

LLM: 预测越准确，理解越深刻。

# 丹尼特的“意向立场”

首先，你将要预测行为的对象视为一个理性的主体。然后，根据其在世界中的位置和目的勾勒出它应当具有的信念。接着，再以同样的方式确定它应当具有的欲望。最后，根据其信念和欲望预测它将采取的行动。

— 丹尼特

- ▶ 在理解、解释、预测一个对象的行为时，我们选择以不同的抽象层级来看待它。
- ▶ 丹尼特的三种解释策略：物理立场、设计立场、意向立场
- ▶ 切换到更高层级的抽象视角，有风险，也有好处。

**Remark:** 不确定大脑神经元是否能表征概念、判断，但常识中用“概念”“判断”“推理”“信念”“动机”“偏好”“欲望”“意图”等术语解释、预测人的行为似乎很好用。



我们分手吧。



她是谁？是不是比我年轻？

# 丹尼特的“意向立场” — 自上而下的解释/实现



Dennett “Stances”	Pylyshyn “Levels of Organization”	Newell “Levels of Description”	Marr “Levels of Analysis”
Intentional Stance	Semantic, or Knowledge Level	Knowledge Level	Computational Theory Level
Design Stance	Symbol Level	Program Level	Representation and Algorithm Level
Physical Stance	Physical Level, or Biological Level	Physical Level, or Device Level	Hardware Implementation Level



面对智能爆炸，人类就像拿着炸弹玩耍的孩子。我们的不成熟与这个玩具的威力严重不匹配。

虽然不知道爆炸会在何时发生，但如果把它放在耳边细听，我们可以听到微弱的滴答声。

— Nick Bostrom

# 囚徒困境

- ▶ **国家**: 面对威胁全人类的不确定性, 国际合作才是硬道理! 但是,
  - ▶ 难以阻止军备竞赛.
  - ▶ 赢者通吃. 诱惑太大.
  - ▶ 为发展牺牲安全 — 在错误的方向上越进步越危险.
- ▶ **个人**: “为什么要在乎我死后的世界? 加速! 加速! 加速才能增加我体验更先进未来的机会.”

# 阿西莫夫的“机器人三定律”

## ▶ 元原则

机器人不得实施行为,除非该行为符合机器人原则.

## ▶ 第零原则

机器人不得伤害人类整体,或者因不作为致使人类整体受到伤害.

## ▶ 第一原则

除非违反高阶原则,机器人不得伤害个人,或者因不作为致使个人受到伤害.

## ▶ 第二原则

1. 机器人必须服从人的命令,除非该命令与高阶原则抵触.

2. 机器人必须服从上级机器人的命令,除非该命令与高阶原则抵触.

## ▶ 第三原则

1. 如不与高阶原则抵触,机器人必须保护上级机器人的安全.

2. 如不与高阶原则抵触,机器人必须保护自己的安全.

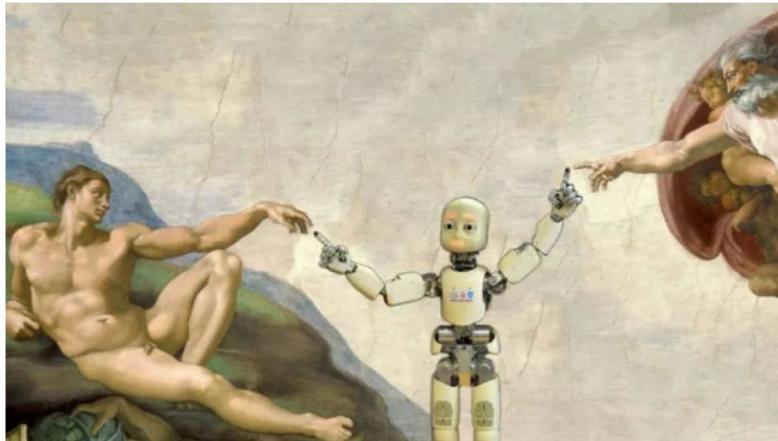
## ▶ 第四原则

除非违反高阶原则,机器人必须执行内置程序赋予的职能.

## ▶ 繁殖原则

机器人不得参与机器人的设计和制造,除非新机器人的行为符合机器人原则.

# 怎么应对?



## 只有一次机会

第一个超人类 AI 必须确保安全, 因为我们可能不会有第二次机会!

### 1. 控制机器

1.1 能力控制 (limiting what the system can or does do).

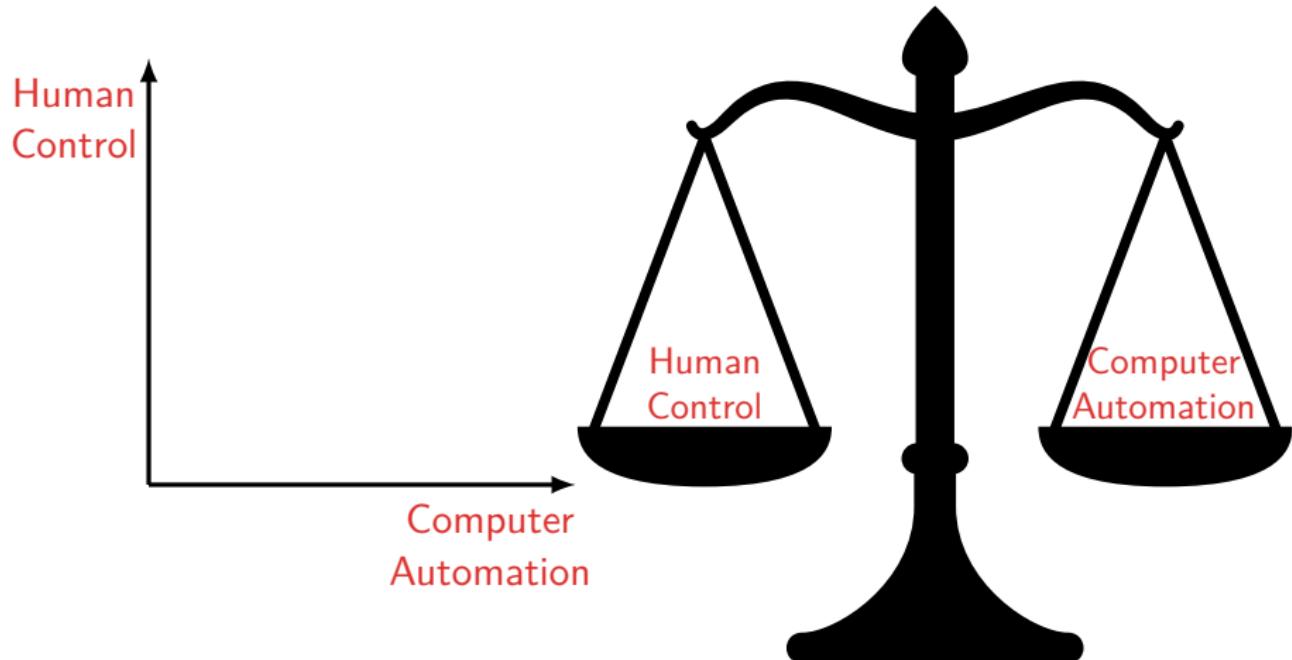
1.2 动机选择 (controlling what the system wants to do).

### 2. 人机融合

# 能力控制方法

- ▶ 盒子方法：通过受限制的渠道对外界产生影响，物理遏制、信息遏制 — 哄骗人类将其释放
- ▶ 奖励方式：“数字加密奖赏币”而非实体奖励
- ▶ 限制认知能力
- ▶ 关机按钮：诊断风险，关机、重启。

# Human Control vs Computer Automation



我们希望奴隶聪明，能帮我们完成任务。我们又希望奴隶顺从。  
完全的顺从和完全的聪明不相容。

— 维纳

# 动机选择方法

## ▶ 直接规定动机

- ▶ 通过规则规定最终目标。— 机器人三定律
- ▶ 功利主义：规定需要最大化的效用函数（比如：快乐）



点石成金：“人类指定目标、机器实现目标”的模式是不可行的，因为人类难以正确地指定目标。

*Everything is vague to a degree you do not realize till you have tried to make it precise. — Bertrand Russell*

## ▶ 驯化

- ▶ 最终目标设为准确回答问题，同时减少对这个世界的影响

## ▶ 间接规范：通过间接手段推断需要被遵守的规则或需要追求的价值

- ▶ 推断人类的意愿

- ▶ Value learning. — inverse reinforcement learning. — wireheading.

## 维纳的“天方夜谭”

- ▶ 一个人得到了一个神灯.
- ▶ 神灯说可以满足他三个愿望.
- ▶ 他的第一个愿望是要一大笔钱.
- ▶ 于是他拿到了一大笔抚恤金, 他的儿子出意外死了.
- ▶ 他的第二个愿望是让儿子回来.
- ▶ 于是儿子的鬼魂站到了他面前.
- ▶ 他的第三个愿望是让这个鬼魂消失.

*If we use, to achieve our purposes, a mechanical agency with whose operation we cannot effectively interfere....we had better be quite sure that the purpose put into the machine is the purpose which we really desire.*

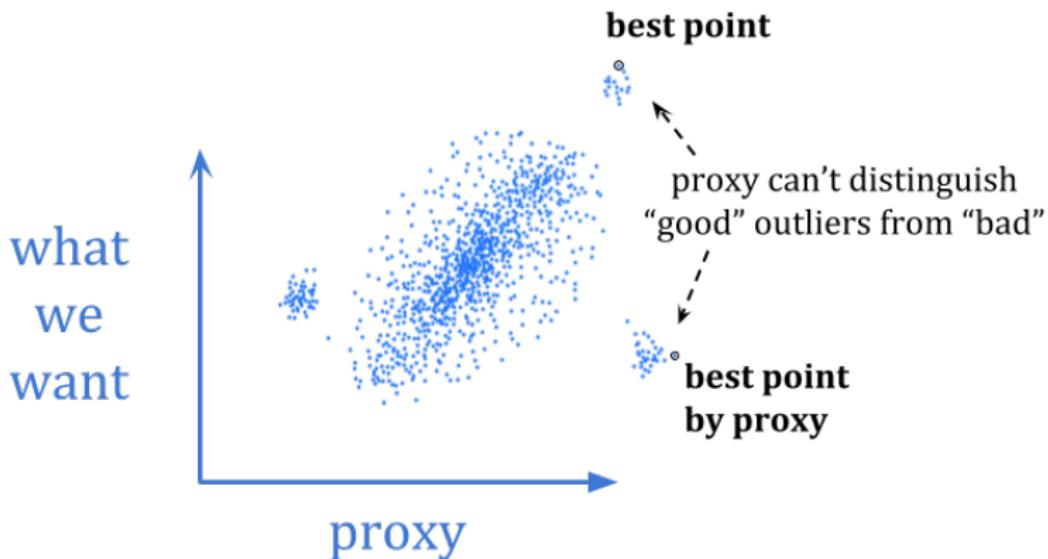
— Norbert Wiener

- ▶ 维纳:《控制论》1948.
- ▶ 维纳:《人有人的用处》1950.

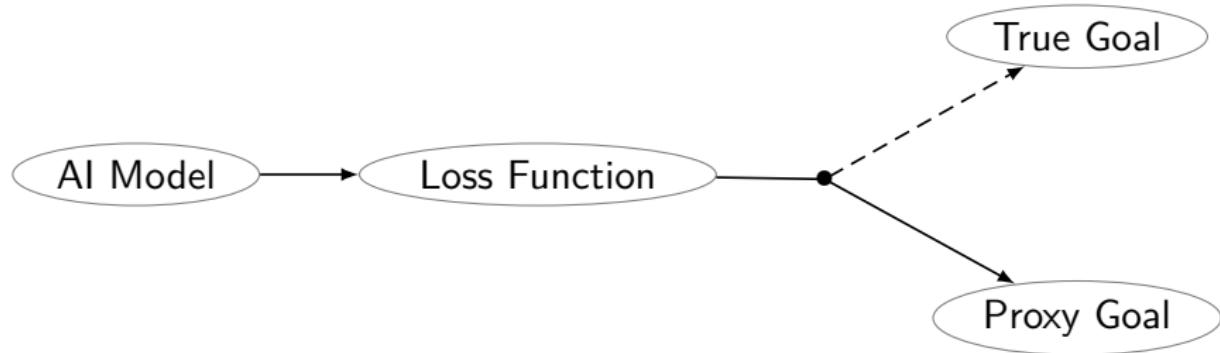
# Goodhart's Law: When a measure becomes a target, it ceases to be a good measure

Example (How To Measure What Matters, Not What Is Measurable?)

- ▶ 政府: 灭鼠. 一根老虎尾巴 1 分钱.
- ▶ 人民: 切掉老鼠的尾巴, 释放老鼠, 喂养繁殖.



# Goodhart's Law



## Example (扫地机器人)

清理垃圾? 清理尽可能多的垃圾?

- ▶ 闭上眼睛就没有垃圾
- ▶ 破坏环境制造垃圾
- ▶ 重新定义垃圾
- ▶ .....



*"As soon as it's done cleaning the house, it brings in trash from the street, and starts all over again!"*

- ▶ 现在的机器学习算法都是优秀的应试者.
- ▶ 智能不仅仅是考试. (人生也不是单一目标导向的打怪通关.)
- ▶ 还涉及提出有用的问题, 定义目标.

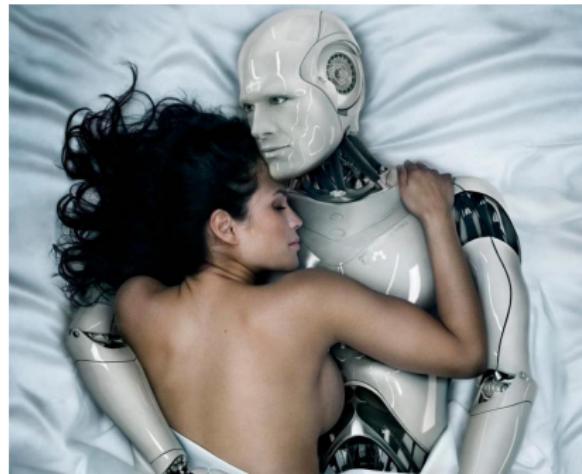
# Technical Research Questions

1. Reliable self-modification
2. Logical uncertainty (reasoning without logical omniscience)
3. Reflective stability of decision theory
4. Decision theory for Newcomb-like problems
5. Corrigibility (accepting modifications)
6. The shutdown problem
7. Value loading
8. Indirect specification of decision theory
9. Domesticity (goal specification for limited impact)
10. The competence gap
11. Weighting options or outcomes for variance-normalizing solution to moral uncertainty
12. Program analysis for self-improvement
13. Reading values and beliefs of AIs
14. Pascal's mugging
15. Infinite ethics
16. Mathematical modelling of intelligence explosion

## What if we do succeed?

你希望人工智能扮演什么样的社会角色？

- ▶ 老人/病人/孩子的看护？
- ▶ 教师？牧师？律师？医生？心理咨询师？
- ▶ 执法者？士兵？死刑执行人？
- ▶ 法律/司法顾问？
- ▶ 如果 AI 是比人类更优秀的 CEO，你会雇佣它吗？
- ▶ 朋友？浪漫伴侣？
- ▶ 性爱机器人？（机器人有感觉，有意识呢？SM 呢？模拟强奸呢？）



## What if we do succeed? — Singularity?

- ▶ 自然选择被人工进化取代. — AI 将成为我们的心智子孙.
- ▶ 一旦机器智能超越了人类智能, 它就能设计出更智能的机器.
- ▶ 这将导致智能爆炸, 导致技术奇点, 人类时代结束.
- ▶ 超越这个事件视界预测之后的事情将是不可能的.

# Singularity

Ulam(1958)/Good(1965)/Solomonoff(1985)/Vinge(1993)/Kurzweil(1999)

## Singularity Hypothesis

Self-accelerating technological advances cause infinite progress in finite time.

### Time Speed Explosion.

- Moore's law: computational resources doubles every 1.5 years.
- In 20 – 30 years the raw computing power of a single computer will reach  $10^{15} \sim 10^{16}$  flop/s.
- Computational capacity of a human brain:  $10^{15} \sim 10^{16}$  flop/s.

### Quantitative Population Explosion.

- Computing costs halve for a certain amount of work.

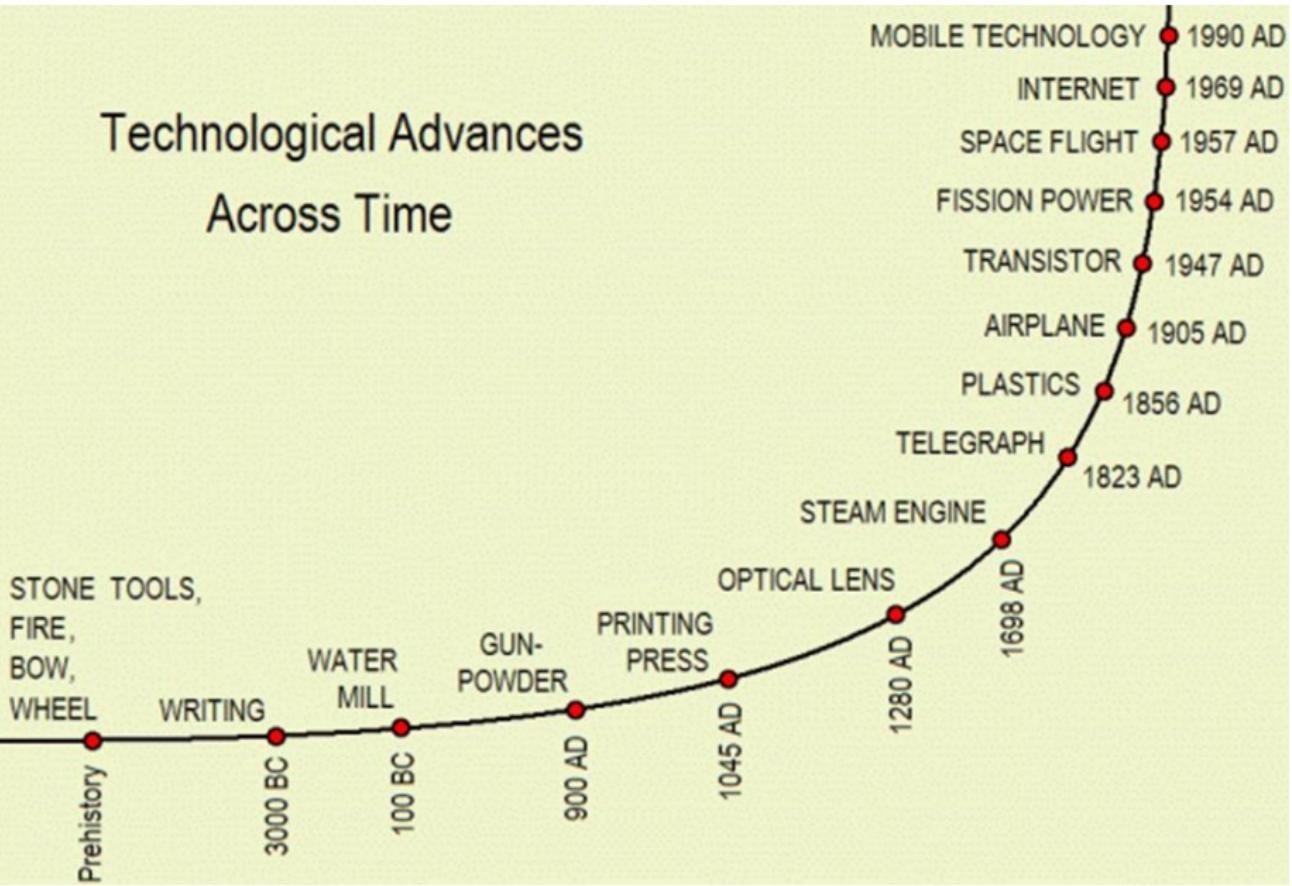
### Qualitative Intelligence Explosion.

- Proportionality Thesis: An increase in intelligence leads to similar increases in the capacity to design intelligent systems.

# 速度爆炸 vs 智能爆炸

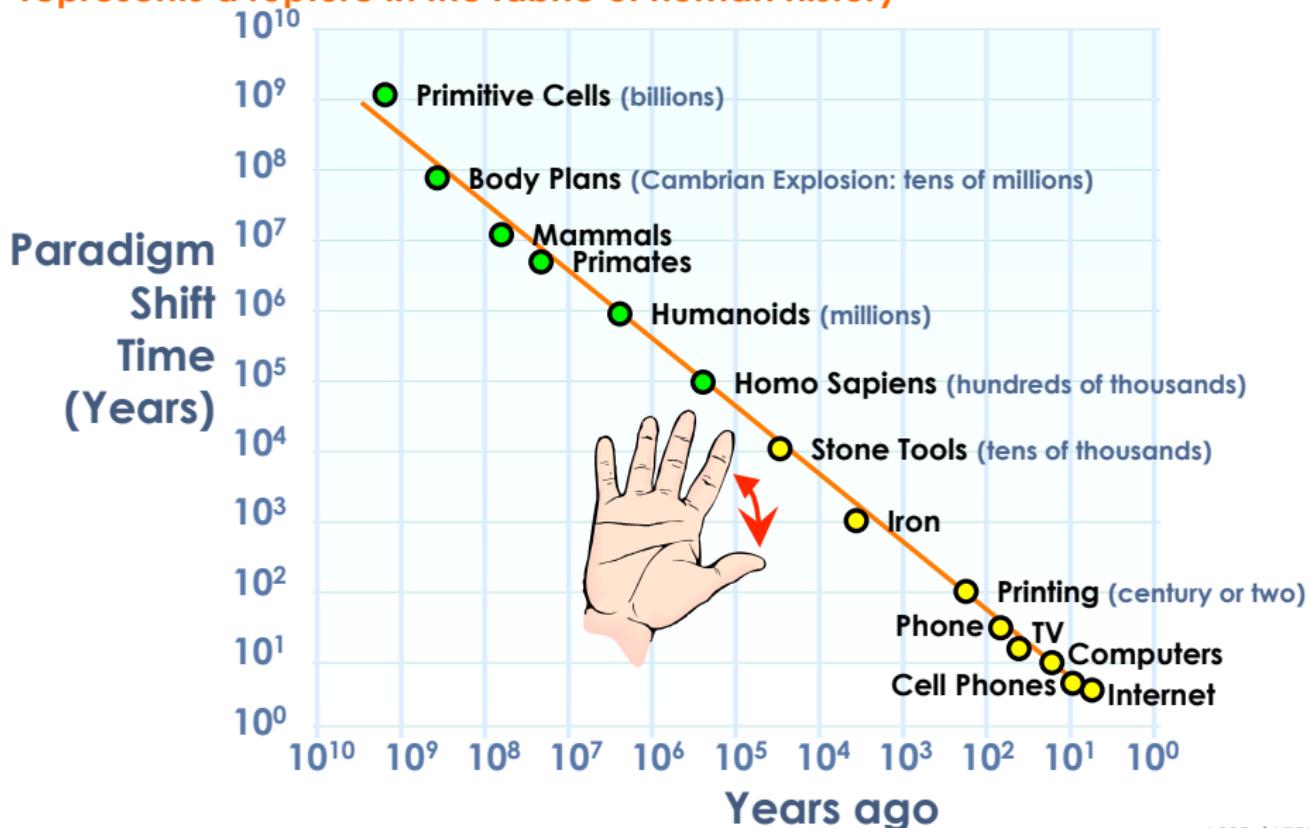
- ▶ 假设世界是个井字棋游戏...
- ▶ 存在最优策略, 不可能有比最优策略更智能的策略.
- ▶ 即使有速度爆炸, 也不会有智能爆炸或智能奇点.
- ▶ 如果智能是有上限的, 那么超过这个界限, 智能只能通过信息处理量和速度来衡量.

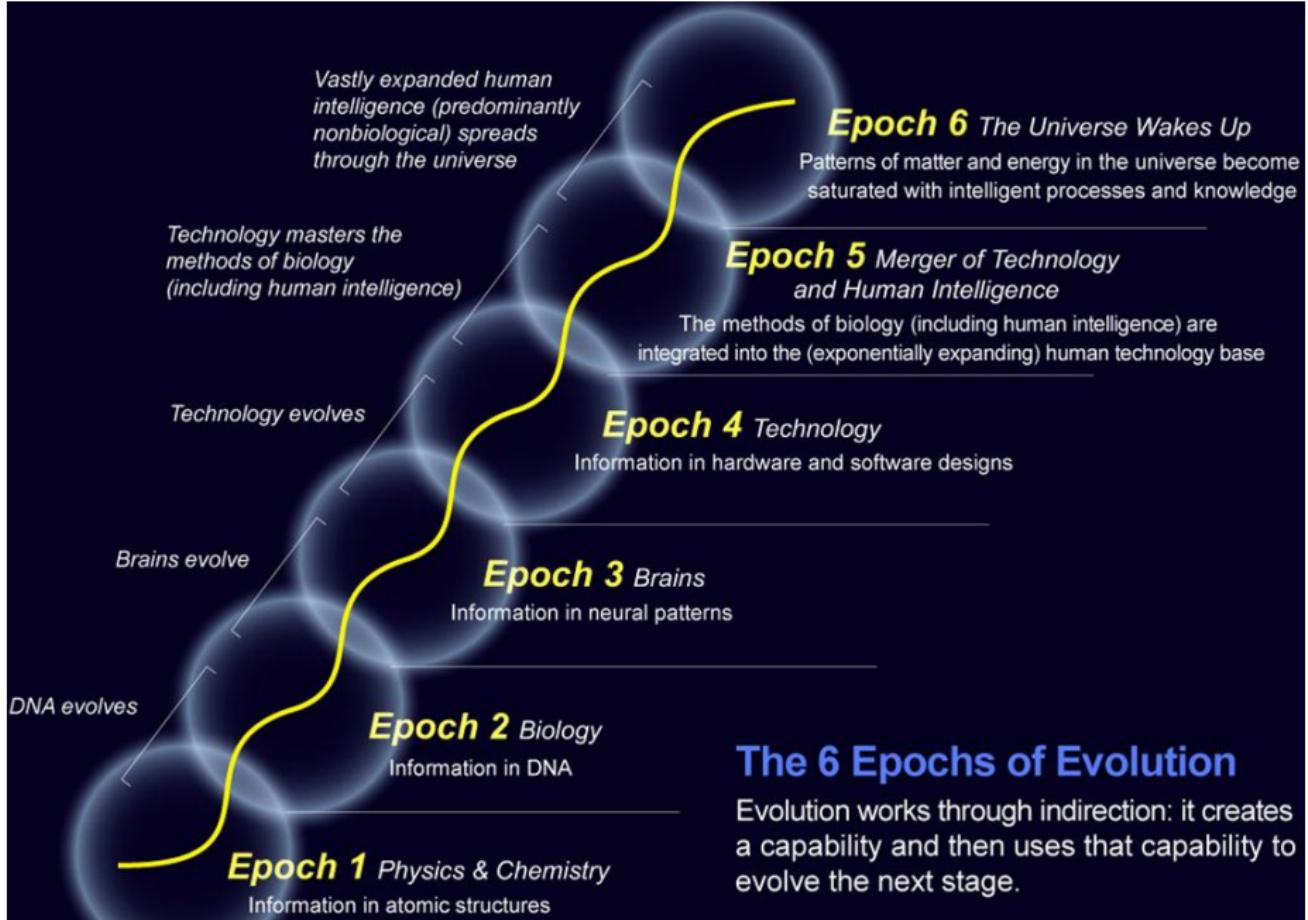
# Technological Advances Across Time



# Countdown to singularity

Singularity is technological change so rapid and so profound that it represents a rupture in the fabric of human history





## The Singularity from the Outside

- ▶ 奇点之外的局外人会看到什么?
- ▶ 奇点会如何影响局外人?
- ▶ 局内人制造出更智能的 AI, 从而以更快的速度生成更智能的 AI'...
- ▶ 局外人只能被动地观察到某种巨大的但无法理解的物质转化.
- ▶ 越来越多的物质被转化为计算机器.
- ▶ 局外人很快将被迫与不断扩张的计算机器进行资源竞争.
- ▶ 扩张速度如此之快, 以至于将接近光速, 逃脱变得完全不可能, 最终将局外的观察者也转化为机器.
- ▶ 最终将没有局外人来观察奇点.
- ▶ 局外人无法经历奇点.
- ▶ 对局外人来说, 奇点类似黑洞.
- ▶ 最大程度的信息压缩与随机噪声没有区别.
- ▶ 一个越来越智能的结构在局外人看来越来越像噪声.
- ▶ 信息量太多会崩塌: 一个包含所有可能书籍的图书馆的信息量为零.

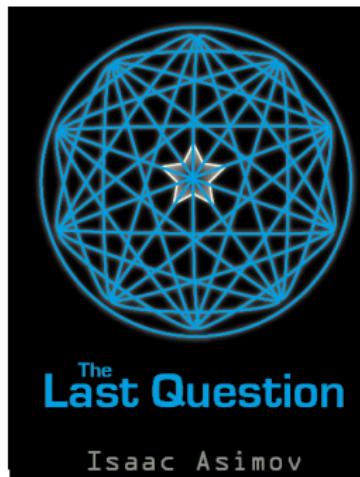
# The Singularity from the Inside

- ▶ 局内参与者会经历什么?
- ▶ 参与者也未必能体验到智能爆炸, 因为他们自身也在以与周围环境相同的速度加速.
- ▶ 他们也只能以‘正常’的主观速度感受‘进步’.

# Paths to Singularity

- ▶ 基于知识的推理和规划 (传统 AI)
- ▶ 从经验中学习的 Agent (机器学习)
- ▶ 意识上传 (扫描大脑) & 后续改进
- ▶ 大脑增强技术 (药物基因工程)
- ▶ 自我进化的 Agent (遗传算法和人工生命)
- ▶ 互联网的觉醒 (数字盖亚).

society of AIXIs or a single organism/mind?



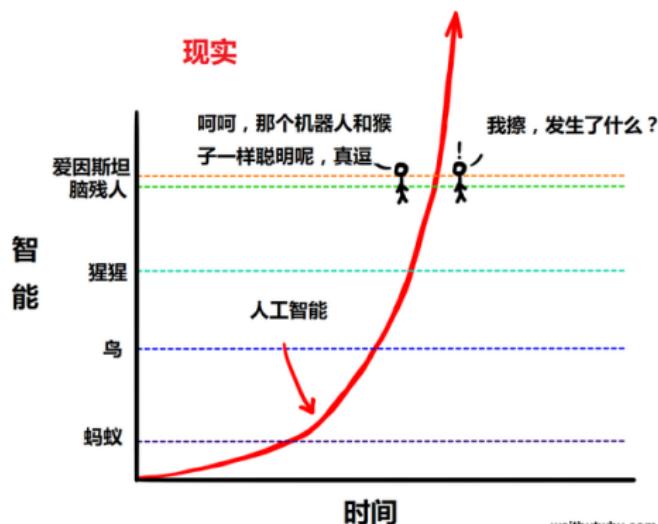
# Evolving Intelligence

- ▶ 进化: 如果智能对生存和繁殖有用, 就通过重组、变异、选择来增加智能.
- ▶ 动物: 后代数量.
- ▶ 人类: 后代数量? 权力/财富...?
- ▶ 模因: 通过复制、变异、选择传播思想.

- ▶ A Blind Man in a Dark Room Looking for a Black Cat That Is Not There?



- ▶ The Singularity is Near?



# References I

- [AMS10] Nihat Ay, Markus Müller, and Arleta Szkola. “Effective Complexity and Its Relation to Logical Depth”. In: *IEEE Transactions on Information Theory* 56.9 (2010), pp. 4593–4607. DOI: [10.1109/TIT.2010.2053892](https://doi.org/10.1109/TIT.2010.2053892).
- [BGT24] Tai-Danae Bradley, Juan Luis Gastaldi, and John Terilla. “The Structure of Meaning in Language: Parallel Narratives in Linear Algebra and Category Theory”. In: *Notices of the American Mathematical Society* (2024). URL: <https://api.semanticscholar.org/CorpusID:263613625>.
- [BS12] John C. Baez and Michael Stay. “Algorithmic thermodynamics”. In: *Mathematical Structures in Computer Science* 22 (2012), pp. 771–787.

## References II

- [BTB22] Tai-Danae Bradley, John Terilla, and Yiannis Vlassopoulos. “An Enriched Category Theory of Language: From Syntax to Semantics”. In: *La Matematica* 1 (2022), pp. 551–580. DOI: [10.1007/s44007-022-00021-2](https://doi.org/10.1007/s44007-022-00021-2). URL: <https://doi.org/10.1007/s44007-022-00021-2>.
- [Car+25] Ryan Carey et al. *Incentives for Responsiveness, Instrumental Control and Impact*. 2025. arXiv: 2001.07118 [cs.AI]. URL: <https://arxiv.org/abs/2001.07118>.
- [CFP22] Carlos Cinelli, Andrew Forney, and Judea Pearl. “A Crash Course in Good and Bad Controls”. In: *Sociological Methods & Research* (2022). DOI: <https://doi.org/10.1177/00491241221099552>.
- [CH22] Renzo Comolatti and Erik Hoel. *Causal emergence is widespread across measures of causation*. 2022. arXiv: 2202.01854 [physics.soc-ph].

## References III

- [Del+24] Grégoire Delétang et al. *Language Modeling Is Compression*. 2024. arXiv: 2309.10668 [cs.LG]. URL: <https://arxiv.org/abs/2309.10668>.
- [Eve18] Tom Everitt. “Towards Safe Artificial General Intelligence”. PhD dissertation. Australian National University, 2018. URL: <http://hdl.handle.net/1885/164227>.
- [Fen21] Luke Fenton-Glynn. *Causation*. Cambridge University Press, 2021.
- [FPB17] A. Forney, J. Pearl, and E. Bareinboim. “Counterfactual Data-Fusion for Online Reinforcement Learners”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 1156–1164.

## References IV

- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.
- [Ham+23] Lewis Hammond et al. "Reasoning about causality in games". In: *Artificial Intelligence* 320 (2023), p. 103919. ISSN: 0004-3702. DOI: <https://doi.org/10.1016/j.artint.2023.103919>. URL: <https://www.sciencedirect.com/science/article/pii/S0004370223000656>.
- [Har77] John C. Harsanyi. "Morality and the Theory of Rational Behavior". In: *Social Research* 44.4 (1977), pp. 623–656. ISSN: 0037783X. URL: <http://www.jstor.org/stable/40971169> (visited on 08/14/2025).

## References V

- [Hoe17] Erik Hoel. “When the Map Is Better Than the Territory”. In: *Entropy* 19.5 (2017). ISSN: 1099-4300. DOI: 10.3390/e19050188. URL: <https://www.mdpi.com/1099-4300/19/5/188>.
- [HQC24] Marcus Hutter, David Quarel, and Elliot Catt. *An Introduction to Universal Artificial Intelligence*. Chapman & Hall/CRC Artificial Intelligence and Robotics Series. 500+ pages, <http://www.hutter1.net/ai/uaibook2.htm>. Taylor and Francis, May 2024, p. 500. ISBN: Paperback:9781032607023, Harcover:9781032607153, eBook:9781003460299. DOI: 10.1201/9781003460299. URL: <http://www.hutter1.net/ai/uaibook2.htm>.

## References VI

- [JCS16] Dominik Janzing, Rafael Chaves, and Bernhard Schölkopf. “Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference”. In: *New Journal of Physics* 18.9 (Sept. 2016), p. 093052. DOI: [10.1088/1367-2630/18/9/093052](https://doi.org/10.1088/1367-2630/18/9/093052). URL: <https://dx.doi.org/10.1088/1367-2630/18/9/093052>.
- [JS08] Dominik Janzing and Bernhard Schölkopf. “Causal Inference Using the Algorithmic Markov Condition”. In: *IEEE Transactions on Information Theory* 56 (2008), pp. 5168–5194. URL: <https://api.semanticscholar.org/CorpusID:11867432>.
- [Ken+22] Zachary Kenton et al. *Discovering Agents*. 2022. arXiv: 2208.08345 [cs.AI].

## References VII

- [KMB23] Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. “Backtracking Counterfactuals”. In: *Proceedings of the 2nd Conference on Causal Learning and Reasoning*. 2023. arXiv: 2211.00472 [cs.AI].
- [Law69] F. William Lawvere. “Diagonal arguments and cartesian closed categories”. In: *Category theory, homology theory and their applications II*. Springer, 1969, pp. 134–145.
- [Leg08] Shane Legg. “Machine Super Intelligence”. PhD dissertation. University of Lugano, 2008.
- [Lei16] Jan Leike. *Nonparametric General Reinforcement Learning*. 2016. arXiv: 1611.08944 [cs.AI].

## References VIII

- [LH07] Shane Legg and Marcus Hutter. “Universal Intelligence: A Definition of Machine Intelligence”. In: *Minds & Machines* 17.4 (2007), pp. 391–444. ISSN: 0924-6495. DOI: 10.1007/s11023-007-9079-x. URL: <http://arxiv.org/abs/0712.3329>.
- [Liu+24] Tian Yu Liu et al. *Meanings and Feelings of Large Language Models: Observability of Latent States in Generative AI*. 2024. arXiv: 2405.14061 [cs.AI]. URL: <https://arxiv.org/abs/2405.14061>.
- [Lou09] A. H. Louie. *More Than Life Itself — A Synthetic Continuation in Relational Biology*. Berlin, Boston: De Gruyter, 2009. ISBN: 9783110321944. DOI: 10.1515/9783110321944. URL: <https://doi.org/10.1515/9783110321944>.

## References IX

- [Lup+21] Andrea I Luppi et al. “What it is like to be a bit: an integrated information decomposition account of emergent mental phenomena”. In: *Neuroscience of Consciousness* 2021.2 (Nov. 2021), niab027. ISSN: 2057-2107. DOI: 10.1093/nc/niab027. eprint: <https://academic.oup.com/nc/article-pdf/2021/2/niab027/41172296/niab027.pdf>. URL: <https://doi.org/10.1093/nc/niab027>.
- [Luz+09] María Luz Cárdenas et al. “Closure to efficient causation, computability and artificial life”. In: *Journal of Theoretical Biology* 263.1 (2009), pp. 79–92. ISSN: 0022-5193. DOI: <https://doi.org/10.1016/j.jtbi.2009.11.010>. URL: <https://www.sciencedirect.com/science/article/pii/S0022519309005360>.

## References X

- [LV19] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*, 4th edition. New York, USA: Springer, 2019.
- [Man14] Yuri I. Manin. “Complexity vs energy: theory of computation and theoretical physics”. In: *Journal of Physics: Conference Series*. Vol. 532. 1. IOP Publishing. 2014, p. 012018.
- [MEB23] Matt MacDermott, Tom Everitt, and Francesco Belardinelli. *Characterising Decision Theories with Mechanised Causal Graphs*. 2023. arXiv: [2307.10987 \[cs.AI\]](https://arxiv.org/abs/2307.10987).
- [Mil54] John Willard Milnor. “Games Against Nature”. In: *Decision processes*. Ed. by Robert McDowell Thrall. Wiley, 1954.

## References XI

- [Mül20] Markus P. Müller. "Law without law: from observer states to physics via algorithmic information theory". In: *Quantum* 4 (July 2020), p. 301. ISSN: 2521-327X. DOI: [10.22331/q-2020-07-20-301](https://doi.org/10.22331/q-2020-07-20-301). URL: <https://doi.org/10.22331/q-2020-07-20-301>.
- [MZP20] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. "Survey on Causal-based Machine Learning Fairness Notions". In: *ArXiv* abs/2010.09553 (2020).
- [Pea09] Judea Pearl. *Causality*. 2nd ed. Cambridge University Press, 2009. DOI: [10.1017/CBO9780511803161](https://doi.org/10.1017/CBO9780511803161).
- [Pea16] Judea Pearl. "The Sure-Thing Principle". In: *Journal of Causal Inference*, Causal, Casual, and Curious Section 4 (1 2016), pp. 81–86.

## References XII

- [PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2017. ISBN: 978-0-262-03731-0. URL: <https://mitpress.mit.edu/books/elements-causal-inference>.
- [PM18] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018. ISBN: 9780465097616. URL: <https://books.google.com/books?id=9H0dDQAAQBAJ>.

## References XIII

- [Ros+20] Fernando E. Rosas et al. “Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data”. In: *PLOS Computational Biology* 16.12 (Dec. 2020), pp. 1–22. DOI: [10.1371/journal.pcbi.1008289](https://doi.org/10.1371/journal.pcbi.1008289). URL: <https://doi.org/10.1371/journal.pcbi.1008289>.
- [SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second. The MIT Press, 2018. URL: <http://incompleteideas.net/book/the-book-2nd.html>.
- [Sch+21] B. Schölkopf et al. “Toward Causal Representation Learning”. In: *Proceedings of the IEEE - Advances in Machine Learning and Deep Neural Networks* 109.5 (2021), pp. 612–634. DOI: [10.1109/JPROC.2021.3058954](https://doi.org/10.1109/JPROC.2021.3058954). URL: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9363924>.

## References XIV

- [SD22] Giovanni Sileno and Jean-Louis Dessimès. “Unexpectedness and Bayes’ Rule”. In: *3rd International Workshop on Cognition: Interdisciplinary Foundations, Models and Applications (CIFMA)*. Ed. by Antonio Cerone. Switzerland: Springer Nature, 2022, pp. 107–116. DOI: [10.1007/978-3-031-12429-7\\_8](https://doi.org/10.1007/978-3-031-12429-7_8). URL: [https://cifma.github.io/Papers-2021/CIFMA\\_2021\\_paper\\_13.pdf](https://cifma.github.io/Papers-2021/CIFMA_2021_paper_13.pdf).
- [Sol78] Ray Solomonoff. “Complexity-based induction systems: Comparisons and convergence theorems”. In: *IEEE Transactions on Information Theory* 24.4 (1978), pp. 422–432.
- [SP24] Oliver Schulte and Pascal Poupart. *Why Online Reinforcement Learning is Causal*. 2024. arXiv: [2403.04221 \[cs.LG\]](https://arxiv.org/abs/2403.04221).

## References XV

- [Svo18] Karl Svozil. *Physical (A)Causality Determinism, Randomness and Uncaused Events*. Fundamental Theories of Physics, 192. Cham: Springer International Publishing, 2018. ISBN: 3-319-70815-5.
- [Sza18] Jochen Szangolies. “Epistemic Horizons and the Foundations of Quantum Mechanics”. In: *Foundations of Physics* 48.12 (2018), pp. 1669–1697. DOI: [10.1007/s10701-018-0221-9](https://doi.org/10.1007/s10701-018-0221-9).
- [Tad20] Kohtaro Tadaki. “Algorithmic information theory and its statistical mechanical interpretation”. In: *Sugaku Expositions* 33 (May 2020), pp. 1–29. DOI: <https://doi.org/10.1090/suga/446>. URL: <https://www.ams.org/journals/suga/2020-33-01/S0898-9583-2020-00446-4/>.

## References XVI

- [TZZ23] Zeyu Tang, Jiji Zhang, and Kun Zhang. "What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective". In: *ACM Computing Surveys 55.13s* (July 2023), pp. 1–37. DOI: [10.1145/3597199](https://doi.org/10.1145/3597199). URL: <https://doi.org/10.1145%2F3597199>.
- [VH22] Thomas F. Varley and Erik P. Hoel. "Emergence as the conversion of information: a unifying theory". In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 380 (2022). URL: <https://api.semanticscholar.org/CorpusID:233407555>.
- [YS18] Eliezer Yudkowsky and Nate Soares. *Functional Decision Theory: A New Theory of Instrumental Rationality*. 2018. arXiv: [1710.05060 \[cs.AI\]](https://arxiv.org/abs/1710.05060). URL: <https://arxiv.org/abs/1710.05060>.

## References XVII

- [Zeč+23] Matej Zečević et al. “Causal Parrots: Large Language Models May Talk Causality But Are Not Causal”. In: *Transactions on Machine Learning Research* (2023). ISSN: 2835-8856. URL: <https://openreview.net/forum?id=tv46tCzs83>.
- [Zur89] Wojciech H. Zurek. “Algorithmic randomness and physical entropy”. In: *Phys. Rev. A* 40 (8 Oct. 1989), pp. 4731–4751. DOI: [10.1103/PhysRevA.40.4731](https://doi.org/10.1103/PhysRevA.40.4731). URL: <https://link.aps.org/doi/10.1103/PhysRevA.40.4731>.

Thank 