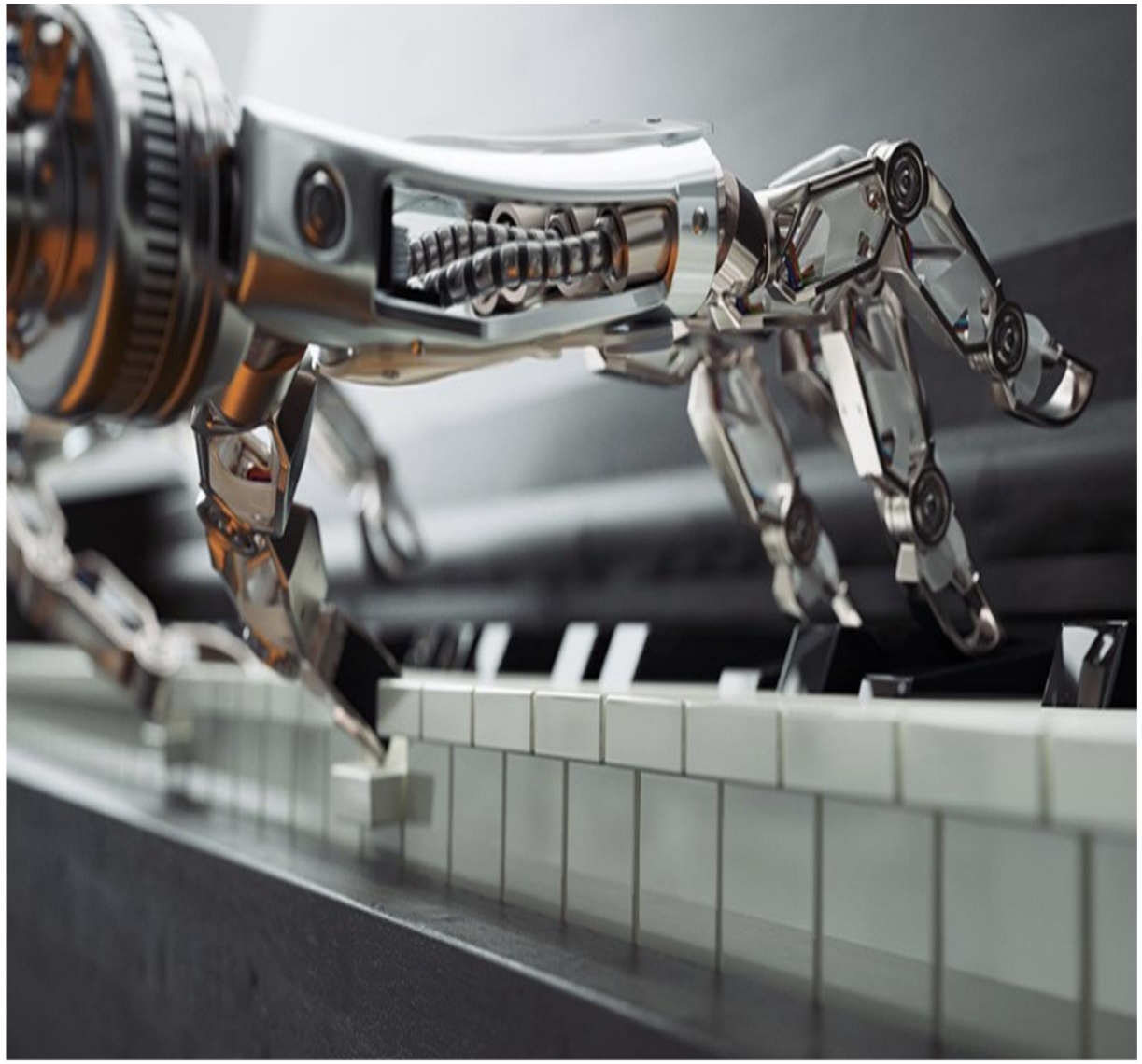
理解機器行為：
我們在人機共生世界中的生存方式

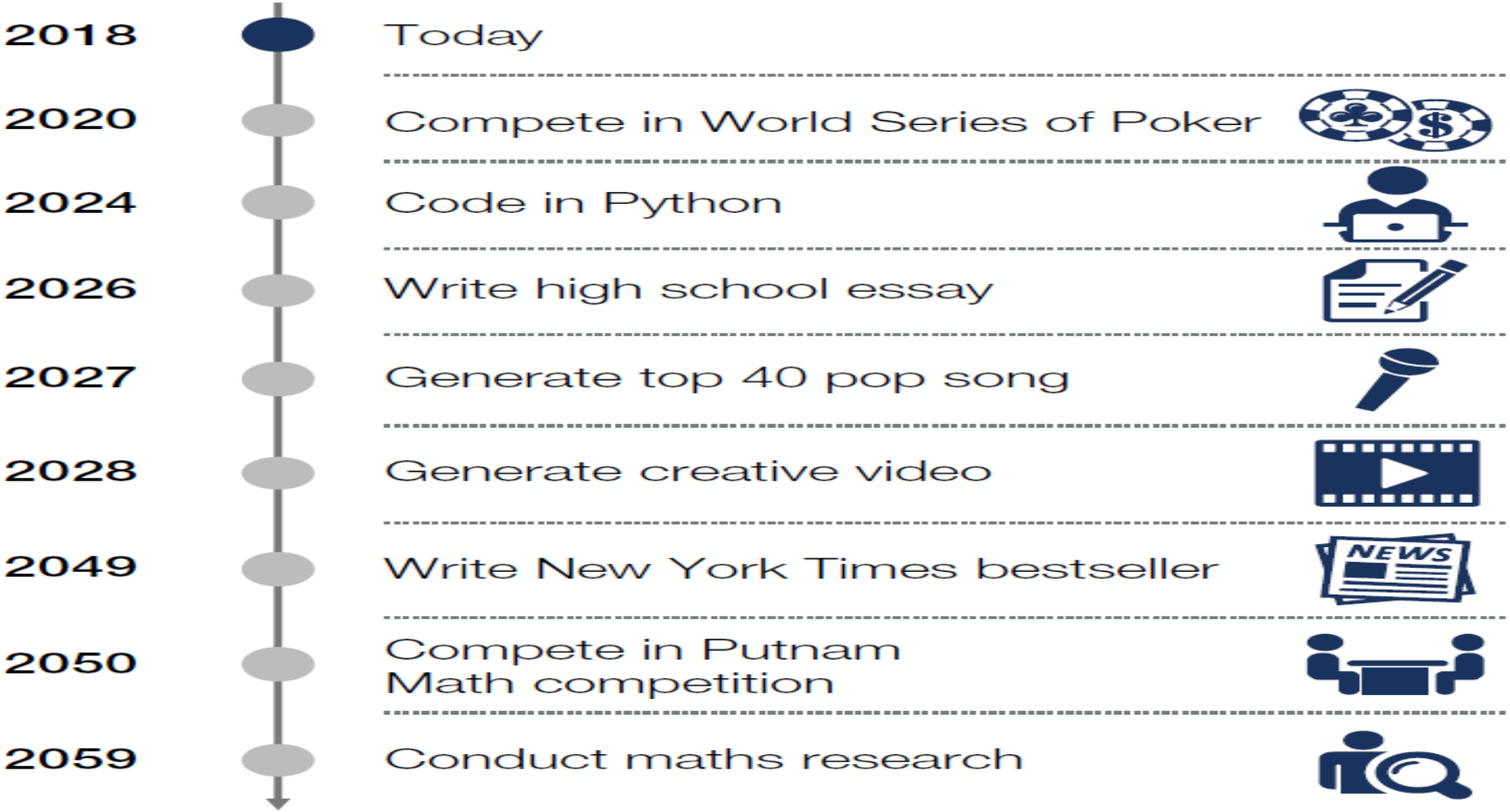
郭毅可 教授　FREng MAE

香港浸會大學副校長
英國皇家工程院院士
歐洲科學院院士

人工智慧的發展：人機共生世界的到來

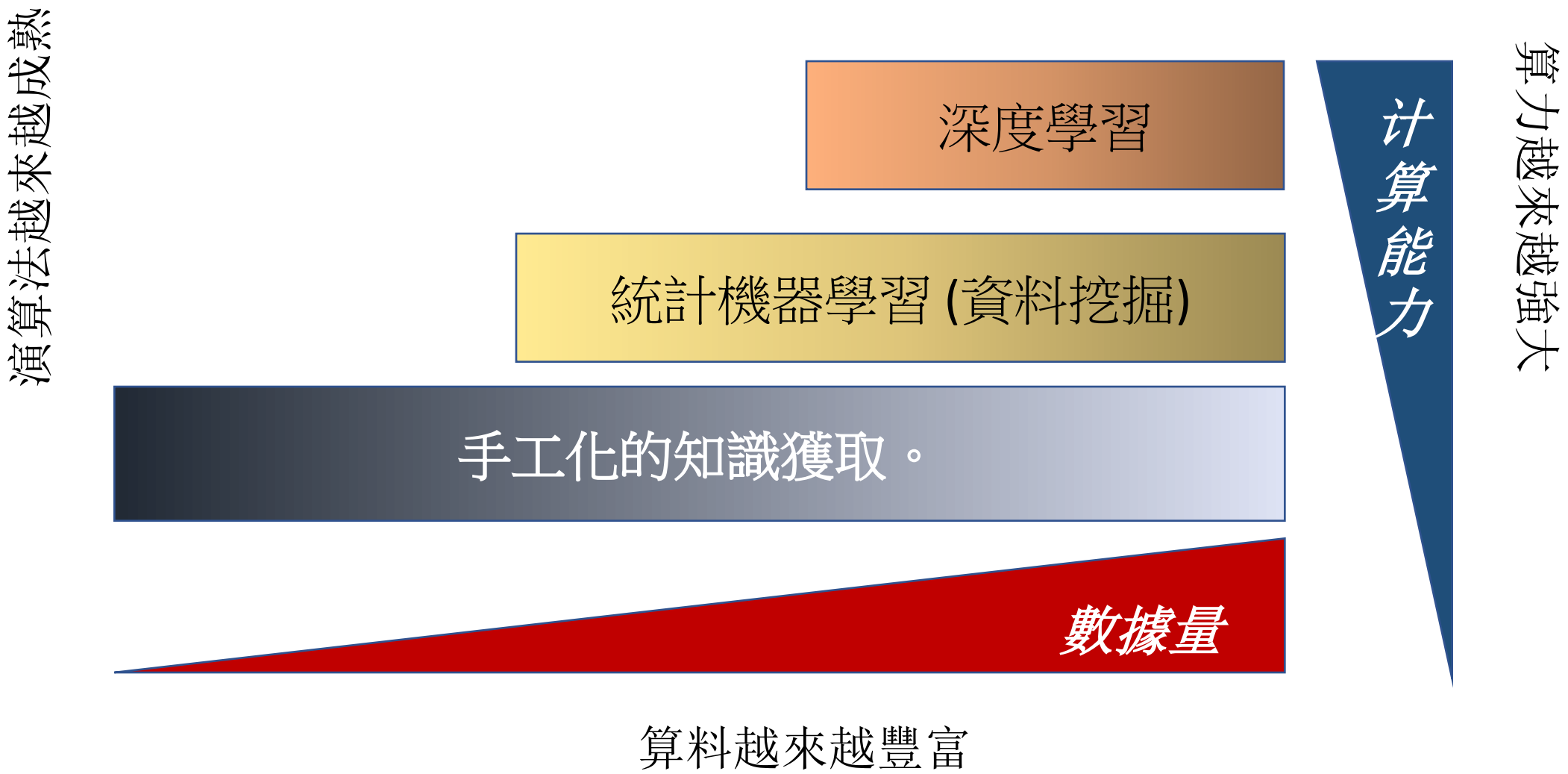


人工智慧发展的時程表

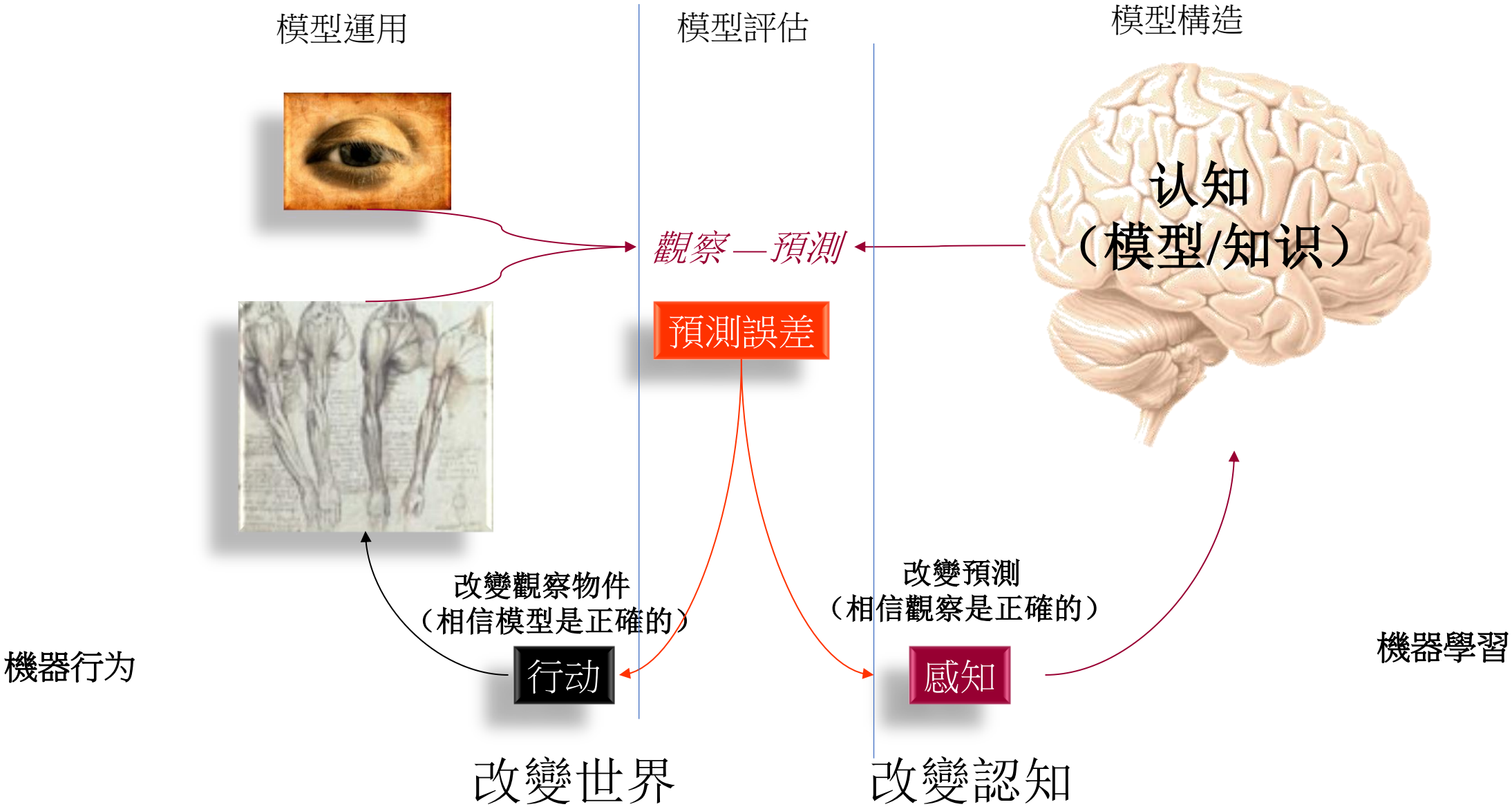


Source: World Economic Forum, Future of Humanity Institute, Oxford University, Department of Political Science, Yale University

機器學習:知識獲取能力的演變



機器學習的認知學原理：觀察和模型



5

什麼是機器行為?


AB139
Vol
RSI
AB1239
TIME VOL PRICE
12:15 10,000 110.00
12:10 20,000 110.00
12:09 8,000 105.00
12:03 5,000 110.00
11:50 500 105.00
11:45 500 110.00
11:30 300 105.00
11:21 11,000 95.00
11:10 8,000 100.00
11:07 7,000 95.00
10:59 5,000 85.00
10:30 5,000 95.00
10:25 4,000 85.00
10:20 900 85.00
10:19 800 95.00
10:15 10,000 100.00
AB1239
+110 (+10%)
Vol/Value(K)
17,2589/153,586
Vol Bid Of

理解機器行為的重要性：政治

政治：

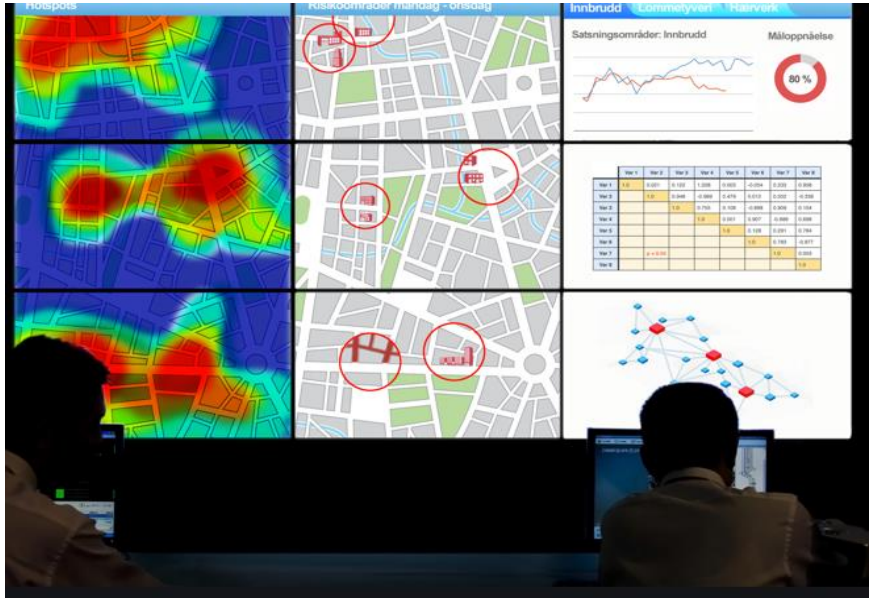
媒體推薦系統:推薦和審查的選擇性

新聞撰寫： 新聞生成算法的可靠性

法律：

機器助審的可能的偏見

警力投入優化系統的自適應性



理解機器行為的重要性：經濟

營銷：

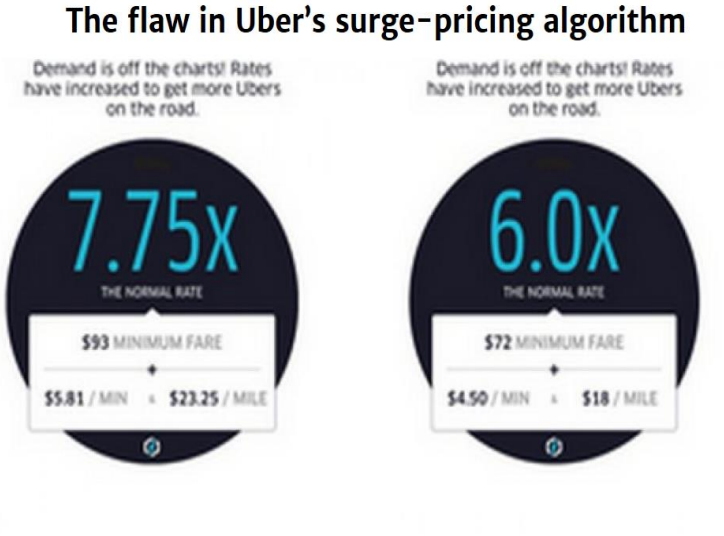
算法交易： 對市場的操縱？

推薦系統：對消費者的損害？

金融：

算法定價：是否會形成惡性競爭或價格操縱？

算法風控：如何避免不合理的歧視？



理解機器行為的重要性：社會

社會：

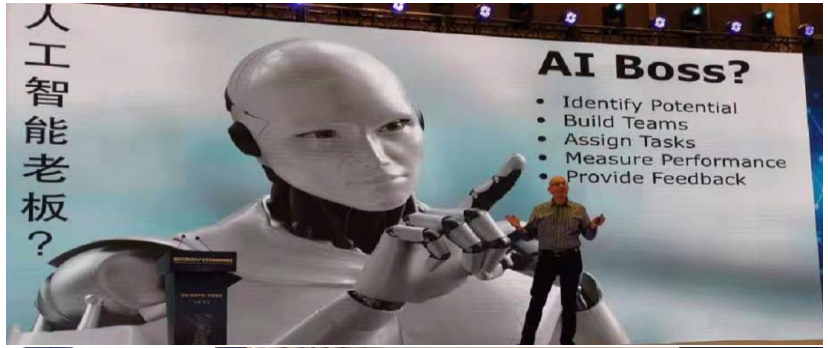
社會信用：是否剝奪他人的權力？

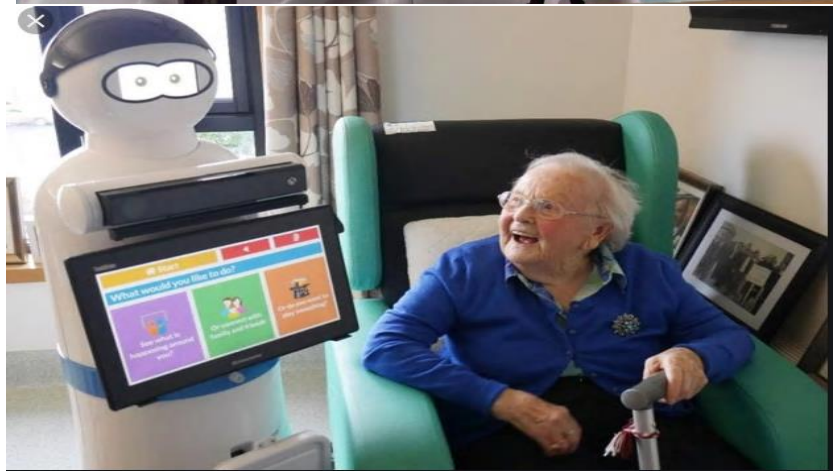
交友推薦：是增加還是減少了社會關係的同質性？

家庭：

聊天機器人：對聊天者的心理影響？

性愛機器人：是否會改變人類的自然感情？





理解機器行為的重要性：生活

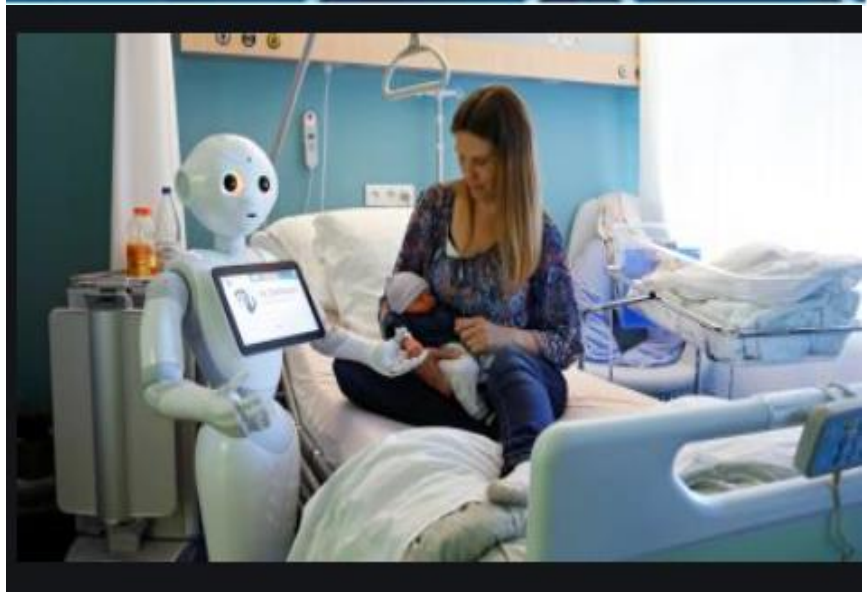
交通：

自動駕駛： 自動駕駛的路權選擇
城市管理：高效交通管理和碳排放量控制的平衡

醫療：

機器診斷：診斷是否可以驗證？

健康監護 ：各種錯誤的代價?



機器行爲

行爲的目的 (What for) : 按照人給出的意圖去改變周圍的環境

行爲的原則 (What is wrong): 不違背人類的倫理和規範

行爲的結果 (What is right): 滿足目的的要求

行爲的根據 (What is the way): 爲什麼會得出這樣的結果

如何向機器表達我们的目的？

效用函数:

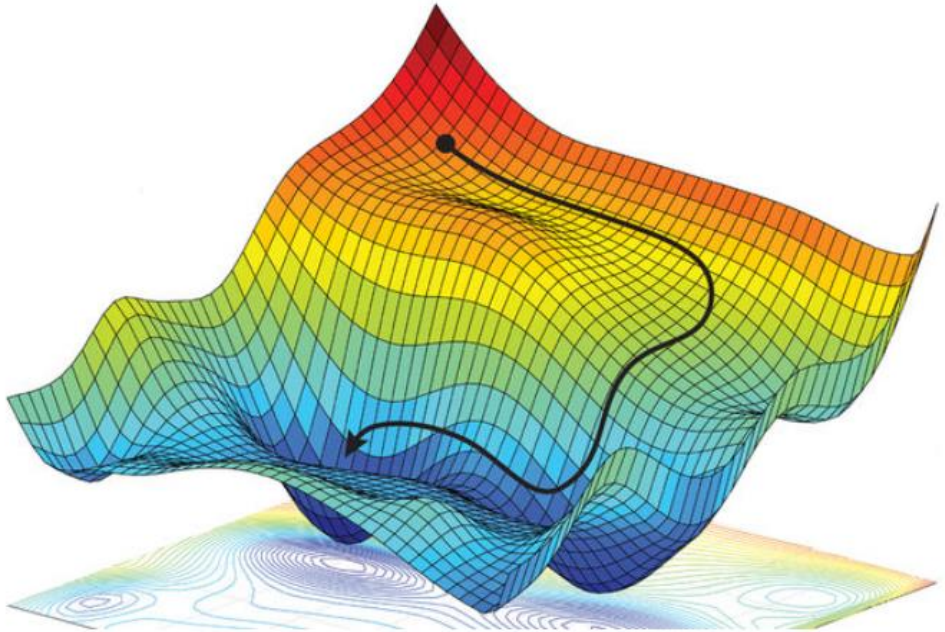
Average under predictions p(x|a)

Cost of 'x' happening

$$\underset{a}{\operatorname{argmax}} \; \underset{p(x|a)}{\mathbb{E}} \left[\ell(x)\right]$$

Possible actions

Possible outcomes 'x'

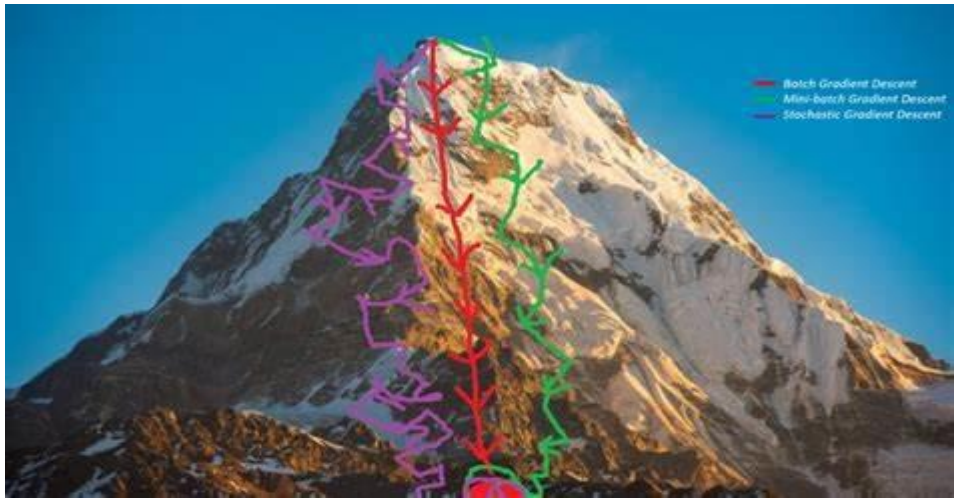


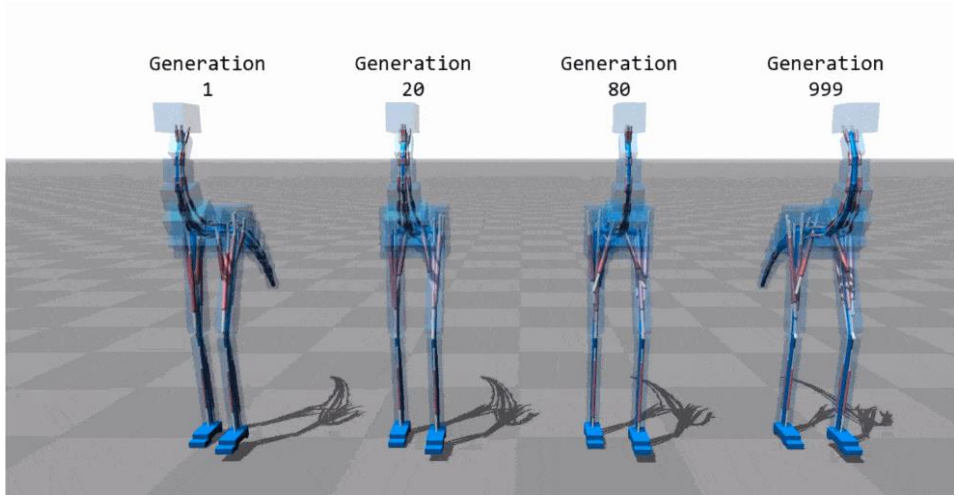
最好的判断（监督学习）
最好的理解（非监督学习）
最好的策略（强化学习）

$$\boldsymbol{argmin_w E(\|Goal - F(x, w)\|)}$$

今天機器行爲的产生机制

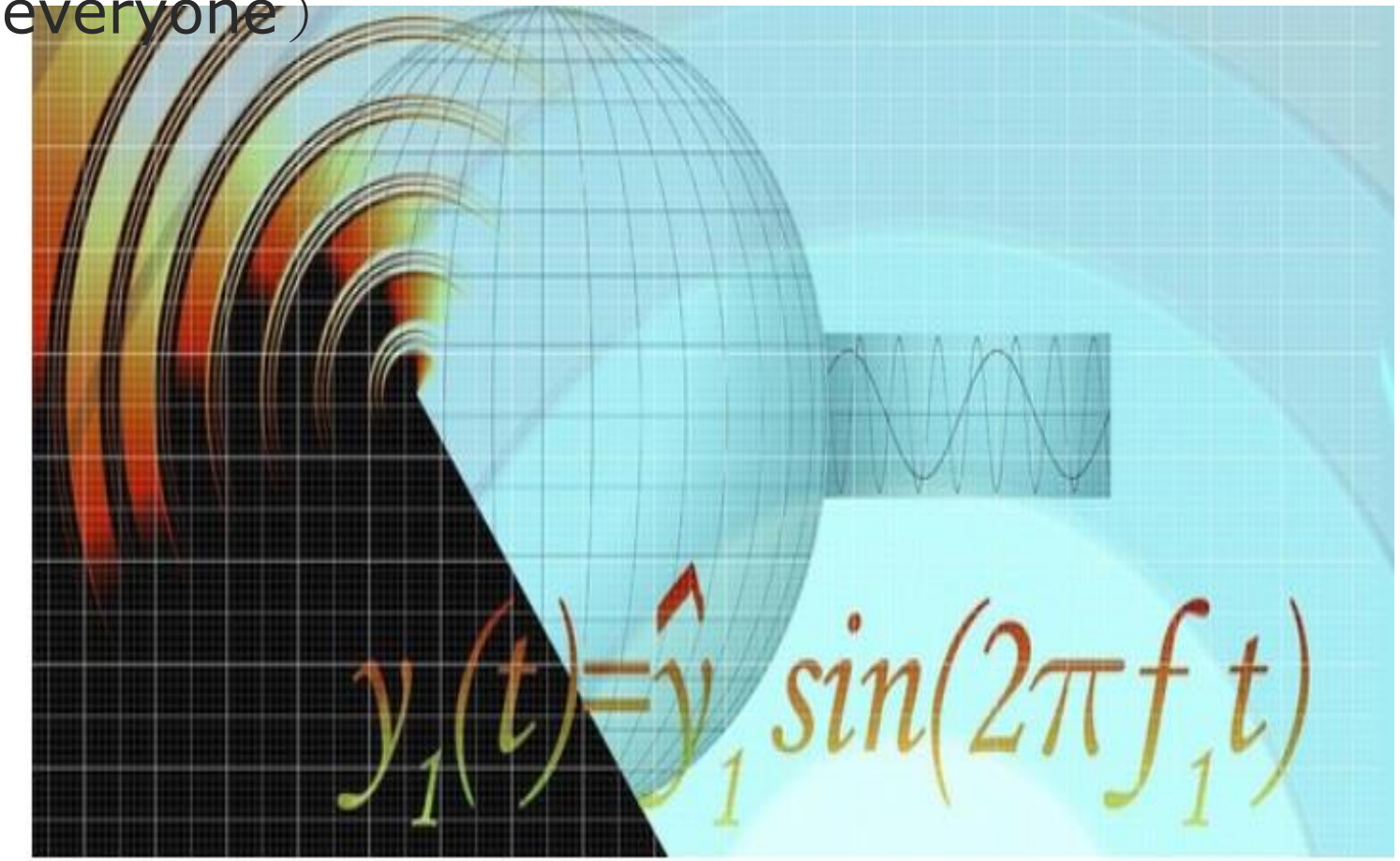
- 算法是從數據中通過學習來產生行爲的
- 算法決定機器行為的辦法往往是通過優化"效用函數"來實現的
- 學習算法的結果是可以通过继续学习而不斷進化
- 算法可以是黑箱，黑箱透明化是通過行爲推演而得的





什麼樣的效用函數可以表達我們的目的?如何來驗證它正確地表達了我們的目的?

Elon Musk 提出了這樣一個效用函數來表達個人行爲的最大自由（maximizing “freedom of action for everyone）



我們如何來定義我們的行為自由？

我們如何來讓AI 來理解 其行為是為人類獲得自由,而不是它自己的自由（“代理人原則”問題（Principal an Agent Problem)。

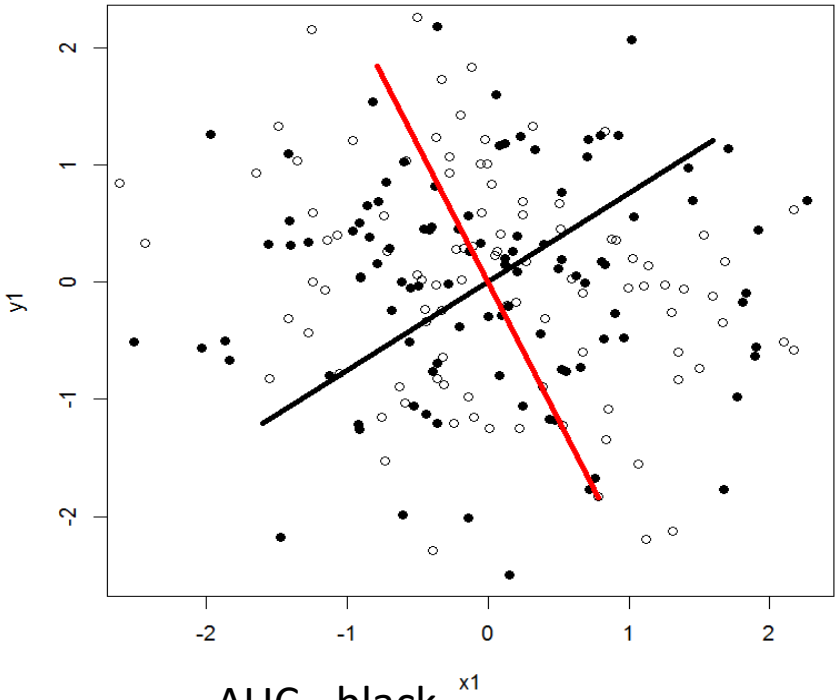
自由的極大化一定是有益的嗎?（米德斯國王問題 或 「不可預知結果」 原則。(Law of Unexpected Consequence)

我們如何保證机器行为符合倫理原則？

- 行为伦理的研究不是一個技術問題,也不只是一個科學問題,更是一個哲學問題
- 人工智慧的倫理性是外在的, 還是內生的？
- 把倫理性作为对行为的限制有明显的局限性
- 我們是否能把人工智慧的倫理性作為出發點而不是限制項

如何判斷機器做的對不對？

什麼叫做最優：



AUC - black

ER – red

相差76度角

最優未必合理：

模型:有哮喘的肺炎患者死於肺炎的可能性較小

合理嗎?

原因是: 哮喘患者被送到重症監護室,因此傾向於生存

求优并不能揭示隐藏的因果关系

驗證與確認

94% of diagnostic AI studies don't adequately validate results

Anicka Slachta | April 12, 2019 | Research

我們要知道AI系統應該具有那些性質才能贏得信任,系統怎樣才能擁有這樣的性質。

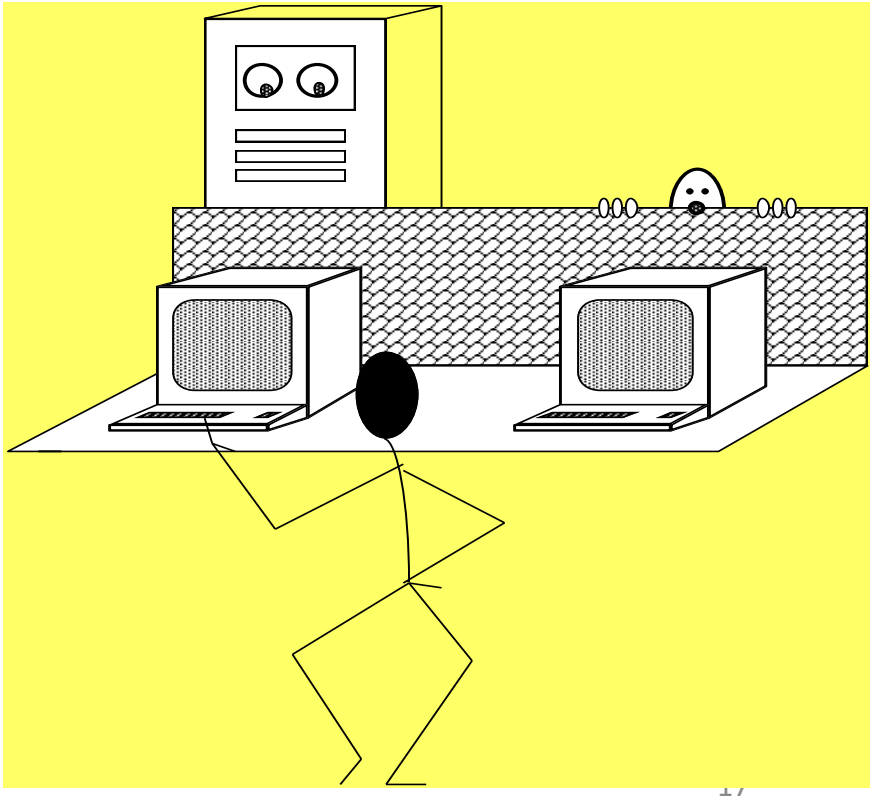
如何理解機器做的對的緣由？

- 現有的深度學習模型不能向使用者解釋其行為,設計者也無法理解其行為過程
- 可解釋性是人類對智慧系統設計的要求,智慧系統必須能夠向用戶解釋其行為。

不可理解

可理解

- 当我们向人和机器共处的黑屋提问，并无法区分得到的回答来自人还是机器，这时候我们可以加上一句："请告诉我你是如何得到这个问题的答案？"，人是能回答这个问题的，而机器则一定很茫然



理解機器的根本困难在于机器不理解人

- 以效用函數向機器表達機器的行為目的:我們無法告訴機器什麼是對的行為
- 機器行為的倫理性應該是機器智慧的內生特性
- 優化效用函數而產生的機器行為不保證和提出的行為目的相吻合
- 優化過程的缺乏無層次性抽象使得人們無法 理解機器行為的緣由
- 理解機器行為的困難在於今天的機器不理解人

人工智慧的重新思考:從機器行為出發

1）行為的目的:有益於人類 (論理原則的內生性—Built-in Ethics)

2）行為目的表達:通過和人交流和學習—非確定性和交互性

3）行為正確性的驗證: 機器行爲和目的的符合可以形式驗證

4）行爲的可解釋性:可以匯出行為的因果關係

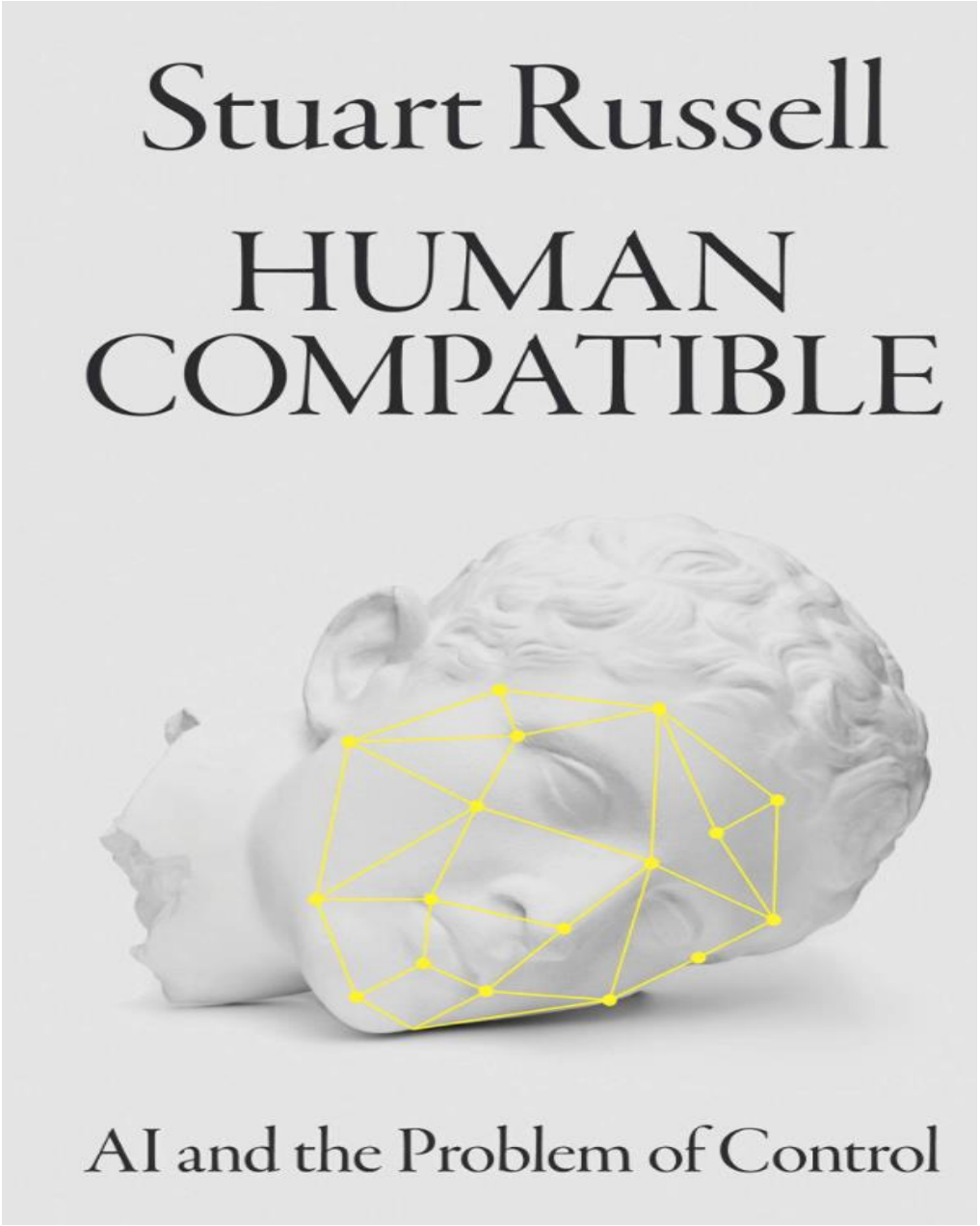
可證明有益人工智慧

- Stuart Russell's 可證明有益人工智慧（provably beneficial AI）三大原則：

1）利他主義（A Law of Altruism）： 機器行為的目標就是用盡全力學習人的需求和喜好，以最大化人类价值

2）謙遜法則：（A law of humility）：機器知道它需要講人類價值最大化,但它不知道這價值究竟是什麼，

3）機器的工作是要瞭解更多人們的需求（A law of learning ）： 它通過對人的行為的觀察和與人的交流來理解人類價值。



人機共生世界

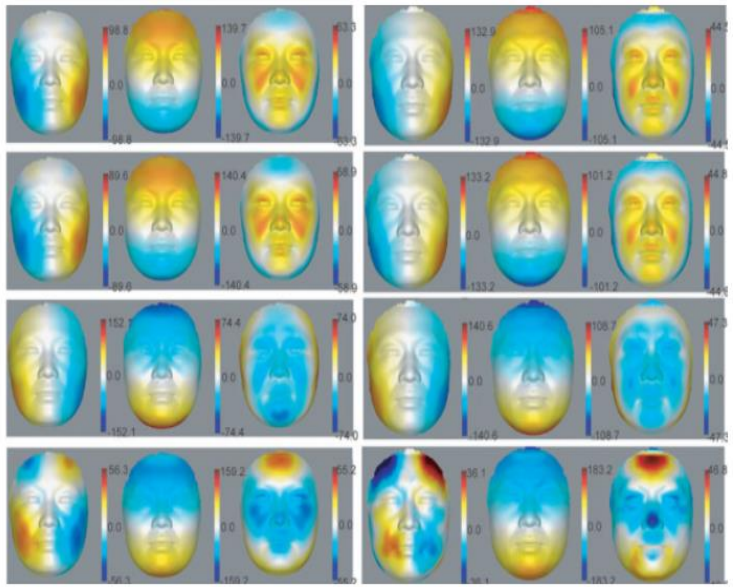
- 在人機共生的社會中，人和機器的互相交流是最重要的
- 人要理解機器的行為,機器要理解人的意圖
- 新一代人工智慧的研究就是以人與機器的有效交流為著眼點的
- 在未來的人工智慧中,對機器行為的解釋,理解和驗證，以及機器行為的倫理性是研究的核心課題

在人机共生世界中機器行為的發展

- 機器行為的發展可以看成是社會對機器行為社會功能的反作用的結果

- 對於人工智慧技術：如何發展新技術，來不斷地在深度和廣度上發展有利行為機制。例：識別人臉這個行爲如何發展--精細地記住每一個人，還是洞察人臉給出的生命信息？

- 對於社會科學：如何提供機制，環境和政策發展有利的機器行為。例：健康數據資源的開發



結語

未來的人工智慧:

- 不求讓機器做的多,但求要機器做的對!
- 我们不是在制造一个仆人，而是一个伙伴
- 我们和机器的互相理解，是人机共生世界的基础