人工智慧之風險與合規

魯君禮 副總經理
安永企業管理諮詢服務股份有限公司

EY 安永
Building a better working world

大綱

01 人工智慧發展趨勢

02 風險與管理

03 相關法規與標準

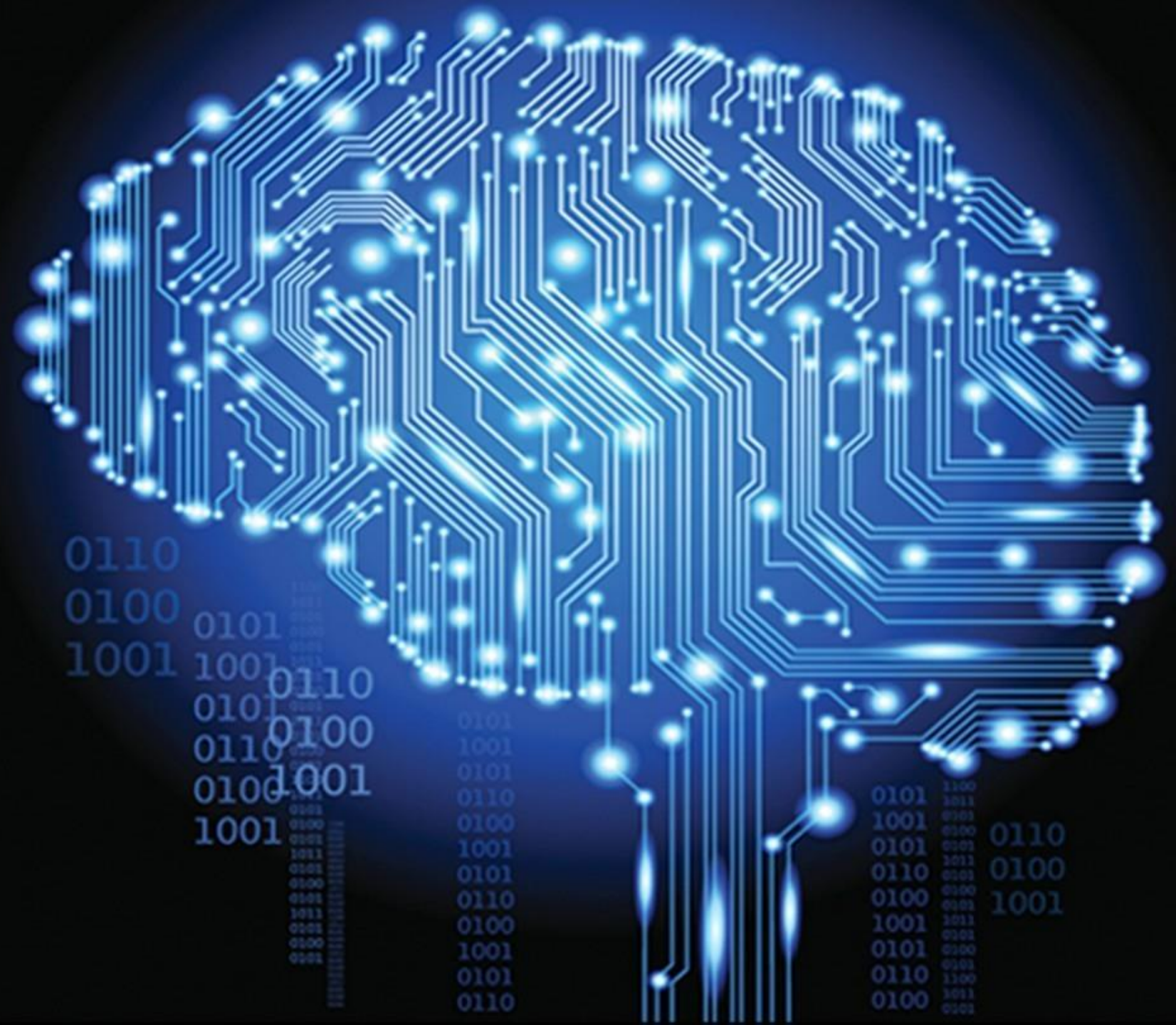
什麼是AI?

AI是執行以前由人類完成的任務的軟體，包括“變得更聰明”或通過數據和經驗學習的能力.

或者說AI是

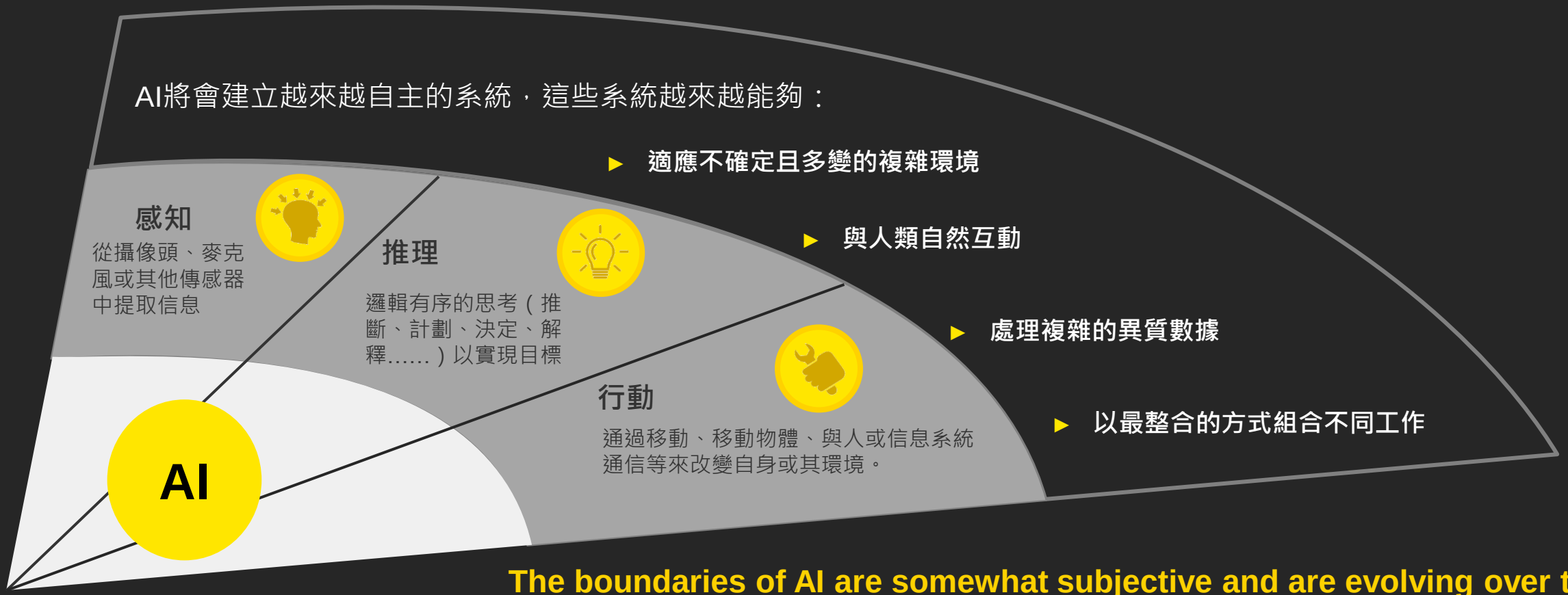
“使感知、推理和行動成為可能的計算科學"

AI能夠“感知”它們的環境、解釋信號並執行動作，例如拍照、導航、拾取和移動物體等。如果機器只是執行一組預編碼指令，它就是在機器人操作。 當它像人們一樣具有感知、評估和採取行動的能力時，它就是人工智慧。



人工智慧的界限有些主觀，並且隨著時間的推移不斷發展

人工智慧是計算機執行類似於人腦學習和決策的任務的能力。



The boundaries of AI are somewhat subjective and are evolving over time

"AI is whatever hasn't been done yet" – L. Tesler

人工智慧發展迅速

理論上...

人類智慧是感知或推斷信息、將其保留為知識並將其應用於決策的能力。 而人工智慧只是將這些能力置於機器中。機器執行認知功能的能力，通常與人類有關。

實際上...

現在解決實際問題的一系列廣泛的技術——包括計算機視覺、自然語言處理、語音識別和機器學習。 這些技術可以結合起來為特定需求創造新的能力。

持續進步

通過機器學習，人工智慧學習並改進了它被要求執行的任務。

速度驚人

AI持續可用，處理速度遠遠超過人類的能力。

騰出時間

人工智慧執行重複和單調的任務，因此人們可以專注於需要判斷力、創造力或深思熟慮的活動。

伴隨AI產生之議題

責任與合理性

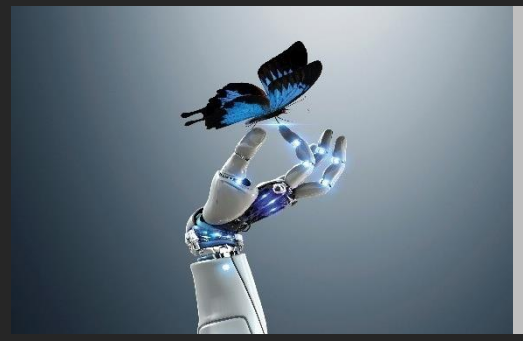
人工智慧系統應該對其做出的決定和採取的行動承擔適當的責任。 在高風險的情況下，人工智慧模型應該是可解釋和可稽核的。

可靠性和安全性

人工智慧系統應該按照預期可靠地運行。

人工智慧的使用，給系統帶來了更大的網路攻擊威脅。 需要採取安全措施來保護敏感數據。

社會責任



應仔細考慮人工智慧系統的潛在社會影響，包括其對人類福祉和自然環境的影響。

倫理

人工智慧系統需要符合企業價值觀、道德和社會規範。

伴隨AI產生之風險

Overview

人工智慧 (AI) 應用高級分析和基於邏輯的技術（包括機器學習）來解釋事件、支持和自動化決策以及採取行動。 在實踐中，人工智慧可以定義為機器，尤其是資訊系統對人類智能過程的模擬。

Risks of Technology

複雜的人工智慧系統可以在沒有組織監督的情況下隨著時間的推移學習和發展，從而導致業務戰略和決策的**可見性和控制力**的喪失

由於複雜的過程，演算法可以引入和**放大錯誤並降低其可見性**，增加相關損失並降低檢測它們的能力

行為準則無法得擴展和嵌入人工智慧應用程序，導致演算法和**決策與組織價值觀不一致**（即偏見和歧視）

人工智慧可以連接在一起並管理多個系統，為對整個組織造成**更大破壞的網路攻擊提供有價值的戰略目標**

Use Cases

人工智慧系統可以被訓練來檢測、監控和擊退網絡攻擊。 他們**識別具有某些顯著特徵的軟體**，然後採取措施關閉攻擊。

金融機構使用**自動化系統**通過將交易信息與與交易者相關的其他行為信息（例如電子郵件流量、日曆項目）聯繫起來來監控其交易者

銀行使用經過歷史支付數據培訓的系統來**監控**信用卡支付是否存在潛在的欺詐活動並阻止可疑交易

機器學習可以對個人或組織拖欠貸款或付款的可能性進行更明智的**預測**

建立對人工智慧的信任需要適當的可解釋性和問責制

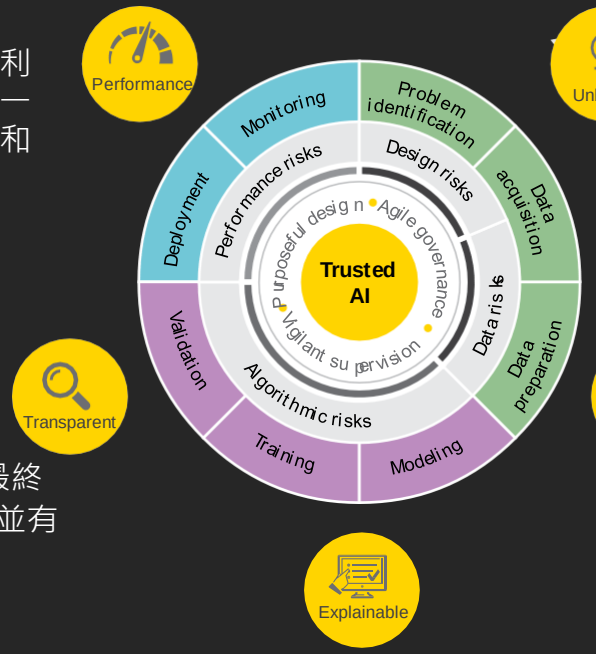
維持信任所需的關鍵屬性

人工智慧的結果應與利益相關者的期望保持一致，並以所需的精度和一致性水平執行

識別由開發團隊組成、數據和培訓方法引起的固有偏見，並通過 AI 設計解決

與 AI 演算法互動時，最終用戶會收到適當的通知並有機會選擇互動層級

保護數據免受未經授權的存取、損壞和對抗性攻擊



人工智慧的訓練方法和決策標準易於理解並可用於人工驗證

連續三步創新流程以實現可信AI

有目的的設計

設計整合機器人、智能和自主功能的系統以實現業務目標

敏捷治理

追踪緊急問題以通知管理系統完整性、其用途和架構的流程

隨時監督

持續監控系統以確保性能的可靠性，並促進透明度和包容性

如何建立可信賴人工智慧生態系統

AI倫理委員會

組成一個一個多領域顧問委員會，就人工智慧開發中的倫理議題考慮提供獨立的建議和指導。

顧問應來自道德、法律、哲學、技術、隱私、法規和科學。

顧問委員會應報告和/或受董事會管轄。

有效驗證工具

應有驗證工具和技術，以確保演算法按預期執行並產生準確、公平和公正的結果。

這些工具亦可用於監控演算法決策框架的變化。

AI倫理設計標準

建立人工智慧倫理設計政策和標準、使其用於發展人工智慧行為準則和人工智慧設計原則。

人工智慧倫理設計標準應定義並管理人工智慧治理和問責機制，以保護用戶、遵循社會規範並遵守法律法規。

觀念宣導及訓練

教育高階管理階層和人工智慧開發人員，使其了解人工智慧發展的潛在法律和道德考慮，認知自己應負之責任，以及他們應如何保護受影響用戶的自由權利和利益。

AI 清單及影響評估

建立所有演算法清單，包括 所有AI 的關鍵細節。

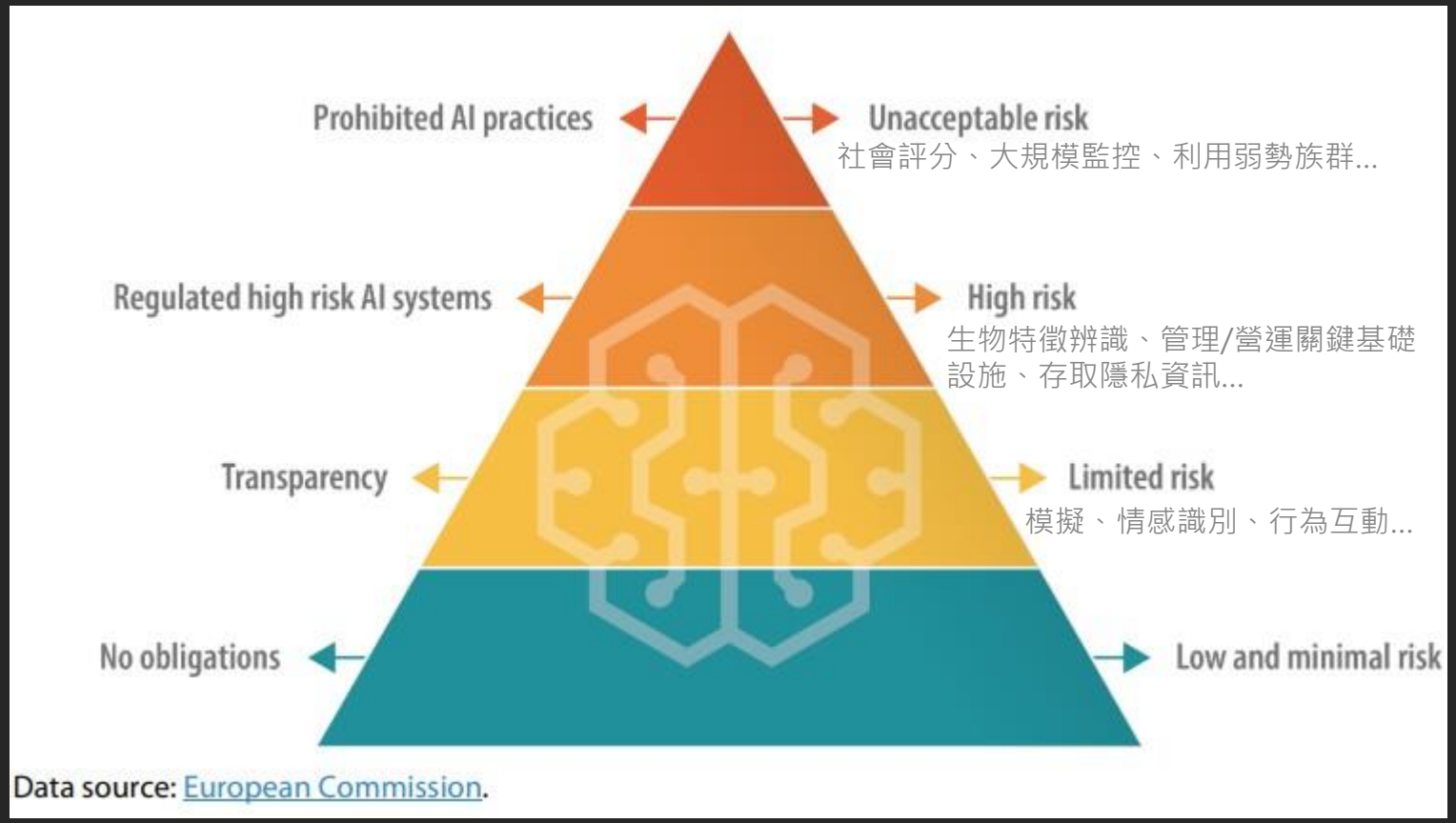
清單中的每個演算法都應進行影響評估，以評估其開發和使用所涉及的風險。

獨立稽核

由獨立第三方根據組織發展人工智慧和技術之政策及規範，以及國際標準進行獨立的 AI 道德和設計審核，以增強對組織 AI 系統的信任。

獨立審查將評估整個人工智慧生命週期中治理模型和控制的充分性和有效性

歐盟「人工智慧草案」內容中之風險分類



歐盟「可信賴人工智慧倫理準則」

1.人類自主性和監控
(Human agency and oversight)
2.技術穩健性和安全性
(Technical Robustness and safety)
3.隱私和資料治理
(Privacy and data governance)
4.透明度
(Transparency)
5.保持多樣性、不歧視和公平
(Diversity, non-discrimination and fairness)
6.社會和環境福祉
(Societal and environmental well-being)
7.責任原則
(Accountability)

ISO 42001 AIMS

依照ISO Annex SL用於管理系統之要求，標準架構統一如下：

Annex A (informative)

MSS for AI Controls

Annex B (informative)

Possible AI-related

organizational objectives

when **managing risks**

ISO 23894

Artificial Intelligence – Risk management

ISO 42001 AIMS - Annex A

MSS for AI Controls	
Clause	Control Domain
A.2	Policies for AI
A.3	Internal organization
A.4	Resources for AI
A.5	Assessing impacts of AI systems
A.6	AI system development life cycle
A.7	Data for AI systems
A.8	Information for users of AI systems
A.9	Use of AI systems
A.10	Third party relationships
A.11	Information security
	TCIC整理

ISO 42001 AIMS - Annex B

Possible AI-related organizational objectives when managing risks	
Clause	Organizational objectives
B.2	Fairness
B.3	Security
B.4	Safety
B.5	Privacy
B.6	Robustness
B.7	Transparency and explainability
B.8	Accountability
B.9	Availability
B.10	Maintainability
B.11	Availability and quality of training data
B.12	AI expertise
	TCIC整理

未被適當管理及使用之AI案例

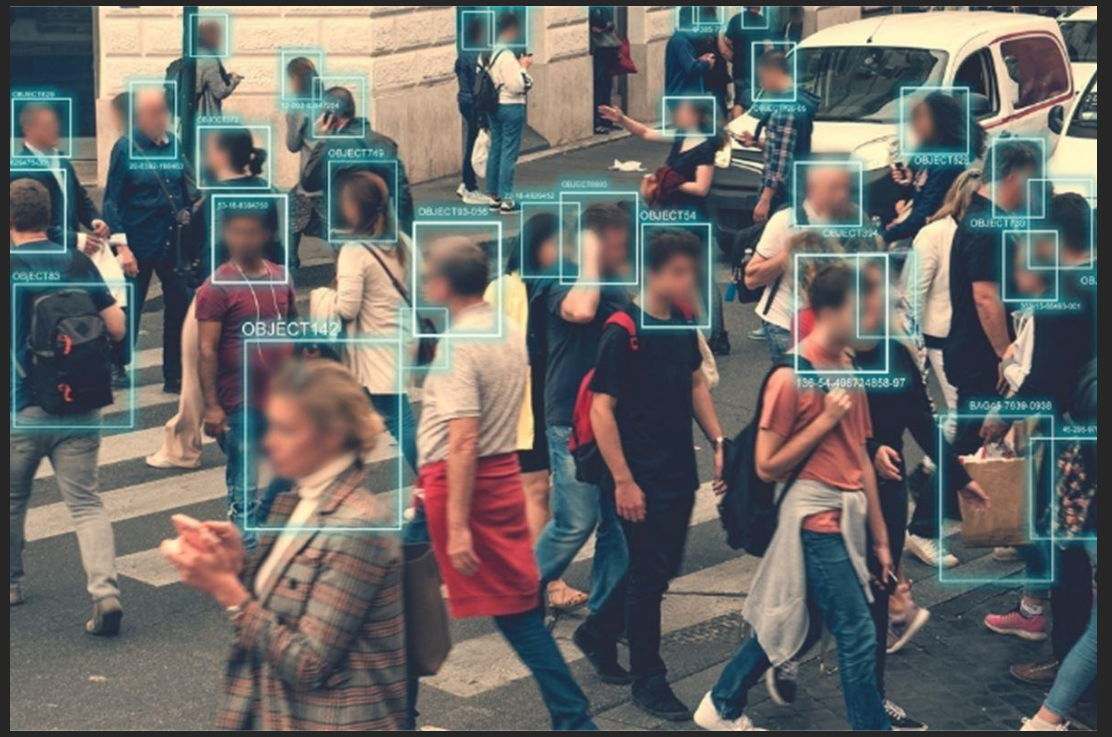
在美國，近年法院已廣泛運用「COMPAS」系統，這是一套由商業公司開發的AI ，幫助法官評估被告的再犯風險，作為量刑的準據。COMPAS會進行大量問答調查，依據被告回答、年齡、過往犯罪紀錄與類型等各項資料，推估被告的再犯率，給出 1 - 10 的危險指數，最後由法官決定被告服刑的長短。

2013年飛車開槍嫌犯Eric Loomis被起訴，法院執行審前調查報告，利用AI風險評估軟體「COMPAS」對Loomis進行再犯風險評估，得到「高風險」的結果，法官據此判他6年徒刑、刑後監督5年，雖然Loomis主張COMPAS無法評估分數是否正確、量刑應個別化，而非透過大數據的特性判定他是否會再犯、計算風險時不應納入「性別」，並要求公開COMPAS的演算法，但全都遭到拒絕；而COMPAS在計算黑人的再犯率時，冤枉黑人會再犯的比例足足是白人的兩倍，亦被抨擊有種族歧視。

目前國內也逐步嘗試將 AI 運用在量刑與家事判決預測。司法院建立的「量刑趨勢建議系統」，以自然語言分析判決書，提供類似案件的量刑參考。清華大學研究團隊則開發出「AI 人工智慧協助家事判決預測系統」，讓 AI 從判例「學會」法官的判決模式，預測撫養權歸屬。

未被適當管理之AI濫用可能性

1. 面部辨識搜尋恐怖分子(Facial recognition)
2. 微表情偵測判斷測謊(AI Polygraph)
3. 投票行為預測分析(Behavior prediction)
4. 金融信用判斷(Automated decision-making)
5. 挖掘網路/系統弱點(Vulnerability mining)
6. 製造假訊息/新聞(Fake news)
7. 演算法偏見(AI Bias)
8. 深度偽造(Deep Fake)



Q & A

我們準備好面對AI了嗎？