王琍瑩律師文章「打破 AI 黑箱，各國政府極力發展的 Trustable AI 是什麼？」刊登於 TechOrange、INSIDE 專欄，針對人工智慧應用領域的全球最新規範提供詳實解讀

這是我家前陣子出現的真實對話，一刀未剪。

我：如果有一輛無人車，出車禍的時候會保護比較多人，另一輛會傷害比較多人，你選那一輛？
孩子：我選保護很多人的。
我：如果為了保護比較多人，有可能害車子裡的你死掉，你選哪一輛？
孩子：還是選保護很多人的。
我：如果車子裡不只有你，還有媽媽呢？
孩子：還是選保護很多人的，媽媽跟我一起去天堂沒關係。
我：如果車子裡只有媽媽呢？
孩子：那我要選保護媽媽。

孩子天真的回答，卻是典型人工智慧應用的 Trolley Problem 倫理兩難，假設回答問題的不是使用者而是車廠，加上商業利益的考量，問題將更為複雜。無論保護多數人的選項多麼道德正確，最終，銷量第一的，恐怕還是能優先

資料來源：https://buzzorange.com/techorange/2019/06/06/trustable-ai/
https://www.inside.com.tw/article/16566-Trustable-AI

保護使用者的無人車吧！無人車或許是極端的例子，但人工智慧演算法不知不覺、方方面面介入我們的日常生活，已是創業者、開發者與使用者必須共同面對的真相。

推動 Trustable AI 為什麼又急又重要？

從 Google 搜尋引擎、Siri 語音助理、Facebook 廣告置入、Netflix 影片推薦……到 UberEats 訂餐外送，每一項服務都因為演算法的優化，為使用者帶來方便，同時為企業增加營收。起初，演算法只是約略猜測我們想買什麼、想看什麼、想吃什麼。隨著數據資料的大量累積與交互運作，演算法開始知道一些「不足為外人道」的秘密，例如，誰最近剛懷孕、誰準備離職跳槽、誰又暗戀著誰。

然後，無論願不願意，演算法可能比我們更瞭解我們，這些關於「自己」的一切，會成為貸款的信用分數、看病的用藥依據、犯罪的呈堂證供。如今，我們也都見證了人工智慧演算法回過頭來操縱個人意志，決定我們該買什麼、該看什麼、該吃什麼，甚至票投給誰。終於，我們變成演算法期待的那個模樣，但究竟哪一個才是「真正的自己」？所謂「更好的自己」誰說了算？而眼前活生生的「自己」，有沒有絲毫抗辯的權利？

事實上，所有機器學習 (Machine Learning) 的模型及成效，都取決於人為的建構，尤其現今主流的深度學習 (Deep Learning)，連開發者都只能判斷運算結果的好壞，卻無從得知 AI 如何作成決策，是名符其實的演算法黑箱。即便暫不考慮 False Positive 與 False Negative 的問題，開發者也很難如同賣菜刀的師傅兩手一攤說：「刀子可以切菜也可以殺人，工具是中立的啊！」

更何況，機器學習的每一筆訓練資料 (Training Data)，都是人類社會既已發生的事實，當歷史經驗存在偏見 (Bias) 未經校正，演算法的產出就必定是偏見的重現。試想，如果國家的警力部署，是依據各地犯罪率高低的歷史資料自動演算的結果，那麼部署嚴密的區域因為破案率提高，在後續犯罪率的統

資料來源：https://buzzorange.com/techorange/2019/06/06/trustable-ai/
https://www.inside.com.tw/article/16566-Trustable-AI

計、警力的部署也將形成循環性的偏差。開發者有意或無意的決定，都可能造成難以預期的結果，也因此，「可信任的人工智慧」(Trustable AI) 成為全球開發者亟欲突破的難題。

Trustable AI 主要的探討，是如何盡可能減少演算法黑箱的節點，提升公平性、當責性與透明性 (Algorithmic Fairness, Accountability and Transparency; FAT)，技術難度或營業秘密再也不能作為演算法偏見或歧視的藉口。

美國國防高等研究計劃署 (DARPA) 自 2017 年推動為期長達五年的「可說明的人工智慧計畫」(Explainable Artificial Intelligence Program; XAI)，就是希望在 Trustable AI 與機器學習的成效二者權衡之間，尋求最佳解。

歐、美、亞三地最新的 Trustable AI 規範在說些什麼？

演算法對人類社會的影響既深且廣，開發者探討 Trustable AI 議題，不能缺少人文與社會科學的研究。所幸，隨著 Trustable AI 技術領域日臻成熟的同時，相關法律論述也在今年上半年逐漸成形。一切的開始，或許要回溯到「家喻戶曉」的歐盟 GDPR (General Data Protection Regulation) 個資保護規範，說到 GDPR，業界對它的認識不外乎「史上最嚴個資法」。

事實上，除了個資的蒐集、處理與利用等隱私保護機制外，GDPR 也相當強調 Trustable AI 在法規上的具體實踐。例如，當「自動化個案決策」(Automated Individual Decision-Making) 完全經由演算法自動產出，且該決策對用戶產生法律效果或類似重大影響時，應賦予用戶請求解釋、表達意見、拒絕適用，或以「工人智慧」人為介入判斷的權利。

此外，GDPR 為了避免人工智慧應用，對於個人的基本人權與自由產生重大危害，針對「特種個資」(Special Categories of Personal Data) 保護更為嚴格，包括種族、政治立場、宗教信仰、工會會籍、基因、生物特徵、健康狀

資料來源：https://buzzorange.com/techorange/2019/06/06/trustable-ai/
https://www.inside.com.tw/article/16566-Trustable-AI

況、性生活或性取向等，只有在符合特殊要件的前提下，才允許採用特種個資做為演算法的輸入資料。

GDPR 影響所及，技術領域長期關注的 Trustable AI，終於在法律層面開啟對話。近期最受全球矚目的三項法案包括：

歐盟：「可信任的人工智慧倫理規範」七大關鍵要素

歐盟執委會 (European Commission) 去年 6 月甫成立的人工智慧高級專家小組(High-Level Expert Group on AI; AI HLEG)，在同年 12 月整理出一份規範草案，廣泛徵詢各界意見之後，於今年 4 月正式頒布「可信任的人工智慧倫理規範」(Ethics Guidelines for Trustworthy AI)，強調七大關鍵要素：人類自主性與監督 (Human Agency and Oversight)、技術穩健性與安全性 (Technical Robustness and Safety)、隱私與數據監理 (Privacy and Data Governance)、透明性 (Transparency)、多元性、非歧視性與公平性 (Diversity, Non-Discrimination and Fairness)、社會與環境福祉 (Societal and Environmental Well-Being)、當責性與咎責機制 (Accountability)。

這份倫理規範不具強制性，各會員國也不急著增修既有法規，使得以上的內容看起來像高大上的精神喊話。但是，執委會將參考現行實務與試辦成果，持續演譯規範內容，預計 2020 年初再度釋出新版。

因此，我們建議 AI 開發者觀察的重點，在於透過目前文件預見政策與法規的走向，包括：將來不同產業細部化的分類規範、自律組織或第三方認證機構的形成等等。尤其，這份倫理規範將 Trustworthy / Trustable AI 區分為合規的 (Lawful)、倫理的 (Ethical) 和穩健的 (Robust) 三種面向，僅僅合規是不夠的，開發者也要努力爭取另外兩個面向的加分題。

此外，草案用語「Trustworthy AI made in Europe」已經更正為「Trustworthy AI for Europe」，明確表示無論 AI 系統的「產地」是不是在

資料來源：https://buzzorange.com/techorange/2019/06/06/trustable-ai/
https://www.inside.com.tw/article/16566-Trustable-AI

歐盟境內，只要「使用」行為在歐盟境內，這份倫理規範都有所適用。事實上，執委會也一再強調，最終目標期待這份倫理規範成為真正接地氣、能實作的「全球共識」。

美國：國會草擬「演算法問責法案」，強化使用者權利

今年上半年，美國國會針對數位隱私、人臉辨識、基因資訊等領域，提出多項法案，都值得開發者關注。其中針對 Trustable AI 議題，美國民主黨 Ron Wyden 等國會議員共同草擬了「演算法問責法案」(Algorithmic Accountability Act of 2019)，特別限制大型企業 (年收入超過 5,000 萬美元，或用戶數、裝置數超過 100 萬件) 的人工智慧應用。

法案認為大型企業的產品或服務傾向於依賴 (而非消除) 演算法偏見，因此必須藉由外部拘束力加以導正，包括授權聯邦貿易委員會 (Federal Trade Commission; FTC) 訂定規範，要求高風險自動化決策系統 (High-Risk Automated Decision System) 進行影響評估 (Impact Assessment)，強化使用者的權利，而非相反地予以邊緣化。

這份法案雖然針對大型企業、並以 FTC 為執法單位，新創團隊的開發者仍然可以經由法案所欲處理的具體情境，來檢視並優化自己的開發項目。舉例來說，機器學習領域普遍應用的「剖析建檔」(Profiling) 與「推薦系統」(Recommendation)，可能由於演算法偏見，導致特定弱勢族群即便使用相同平台服務，卻無法接觸某類住宅建案、某些工作機會或某種貸款方案。

事實上，Facebook 在今年 3 月已經與各該領域主管機關協商和解，針對住房 (Housing)、僱傭 (Employment) 和信用貸款 (Credit Offers) 這些敏感項目，正式宣佈新的廣告投放措施，禁止使用年齡、性別或郵遞區號來鎖定投放對象，並對整體分眾功能 (Targeting Categories) 加以限縮，甚至在房屋租售的項目，將一勞永逸地建構新的搜尋工具，讓每一位使用者都能搜尋到美國境內每一則廣告。

亞洲：新加坡居於領先地位

新加坡對於 Trustable AI 的規範進程，在亞洲各國之中儼然居於領頭羊地位。先是去年 11 月由金融管理局 (Monetary Authority of Singapore) 針對金融產業提出「人工智慧與資料分析行為準則」(FEAT Principles)，隨後今年 1 月再由個人資料保護委員會 (Personal Data Protection Commission) 頒布「人工智慧監管模式框架」(A Proposed Model Artificial Intelligence Governance Framework)，為非特定產業的一般性人工智慧應用提供指導方針。無獨有偶地，這份監管框架與前面提到的歐盟倫理規範一樣，強調自己是動態文件 (Living Document)，不具強制性，並將隨著技術進程與各界反饋持續更新。

而相較於高大上的歐盟倫理規範，新加坡的監管框架已經針對內部監理的架構與措施、人工智慧決策的風險管理、營運管理、與使用者關係維護等四個面向，提供開發者具體可資遵循的步驟，甚至引用 UCARE.AI 與新加坡 Parkway Pantai 醫療集團有關醫療費用的預測分析 (Predictive Analysis) 作為實作案例。

這個個案的挑戰性，在於醫療資料的蒐集、處理、利用本來就受到相關法規的嚴格監管，而個案涉及類神經網絡的深度學習演算法黑箱，也正是 Trustable AI 的透明性要求最常卡關的痛點。

依據監管框架的介紹，個案在 Trustable AI 的公平性、當責性與透明性各個環節，都已規劃對應流程，一方面藉由演算法自動化決策，來降低人為主觀評價欠缺一致性的問題 (最常見就是受到病患收入水平或保單承保範圍的影響)，另一方面也確保健全的人為反饋迴路 (Human Feedback Loop) 能夠介入運作，甚至在演算法內建人為撤銷機制 (Manual Override Protocol) ，以便必要時能夠安全地終止演算法的運作。

結語

資料來源：https://buzzorange.com/techorange/2019/06/06/trustable-ai/
https://www.inside.com.tw/article/16566-Trustable-AI

從 2016 年 Microsoft Tay 失控之前的樂觀，到今年初 Open AI 對釋出開源碼的猶豫，人類漸漸學會對機器感到敬畏，也終於明白所謂「工具中立性」其實意味著，它的亦正亦邪完全掌握在人類手裡。眼前這些 Trustable AI 最新規範趨勢，不約而同展現這樣反覆辯證的歷程，提醒我們不要過於相信人類，也不要過於相信機器。

換言之，Trustable AI 所有規範都指向同一個方向——人類應當有權力 (也有責任) 決定什麼情況要信任「人工智慧」、什麼情況要依賴「工人智慧」。要達到這個目標，首先必須認清人類和機器各自的強項與侷限，但是這一題太難，因而需要在制度面加以規劃，包括事前的管理、事中的檢驗與事後的救濟，缺一不可，供各位創業者、開發者參考。

資料來源：https://buzzorange.com/techorange/2019/06/06/trustable-ai/
https://www.inside.com.tw/article/16566-Trustable-AI