歐盟議會發布《可信賴人工智慧倫理準則》

2019年4月9日，歐盟議會發布《可信賴人工智慧倫理準則》（Ethics Guidelines for Trustworthy AI）。此次內容大致延續歐盟人工智慧高階專家小組（High-level Expert Group on Artificial Intelligence）於2018年12月18日發布的《可信賴人工智慧倫理準則草案》（Draft Ethics Guidelines for Trustworthy Artificial Intelligence）之內容，要求人工智慧須遵守行善（do good）、不作惡（do no harm）、保護人類（preserve human Agency）、公平（be fair）與公開透明（operate transparency）等倫理原則；並在4月9日發布的正式內容中更加具體描述可信賴的人工智慧的具體要件，共計七面向概述如下：

1. 人類自主性和監控(Human agency and oversight)：AI係為強化人類能力而存在，使人類使用者能夠做出更明智的決策並培養自身的基礎能力。同時，AI應有相關監控機制以確保AI系統不會侵害人類自主性或是引發其他負面效果。本準則建議，監控機制應可透過人機混合（一種整合人工智慧與人類協作的系統，例如human-in-the-loop, human-on-the-loop, and human-in-command）的操作方法來實現。
2. 技術穩健性和安全性(Technical Robustness and safety)：為防止損害擴張與確保損害最小化，AI系統除需具備準確性、可靠性和可重複性等技術特質，同時也需在出現問題前訂定完善的備援計劃。
3. 隱私和資料治理(Privacy and data governance)：除了確保充分尊重隱私和資料保護之外，還必須確保適當的資料治理機制，同時考慮到資料的品質和完整性，並確保合法近用資料為可行。
4. 透明度(Transparency)：資料、系統和AI的商業模型應該是透明的。可追溯性機制（Traceability mechanisms）有助於實現這一目標。此外，應以利害關係人能夠理解的方式解釋AI系統的邏輯及運作模式。人類參與者和使用者需要意識到他們正在與AI系統進行互動，並且必須了解AI系統的功能和限制。
5. 保持多樣性、不歧視和公平(Diversity, non-discrimination and fairness)：AI不公平的偏見可能會加劇對弱勢群體的偏見和歧視，導致邊緣化現象更為嚴重。為避免此種情況，AI系統應該設計為所有人皆可以近用，達成使用者多樣性的目標。
6. 社會和環境福祉(Societal and environmental well-being)：AI應該使包含我們的後代在內的所有人類受益。因此AI必須兼顧永續發展、環境友善，並能提供正向的社會影響。
7. 問責制(Accountability)：應建立機制以妥當處理AI所導致的結果的責任歸屬，演算法的可審計性（Auditability）為關鍵。此外，應確保補救措施為無障礙設計。

相關附件

Ethics Guidelines for Trustworthy AI [cfm?doc_id=58477]

蔡宜臻
法律研究員 編譯整理

上稿時間： 2019年06月

資料來源：

1. EUROPEAN COMMISSION, Ethics Guidelines for Trustworthy AI, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477 (last visited May 1, 2019)

文章標籤

機器人 人工智慧 巨量資料

推薦文章