出國報告（出國類別：開會）

參加智慧科技應用信任、隱私及安全國際研討會

服務機關：內政部建築研究所

姓名職稱：張怡文副研究員

派赴國家/地區：美國

出國期間：113年10月26日至11月3日

報告日期：114 年 2 月 3 日

摘要

關鍵詞：智慧科技應用，可信任的人工智慧，可解釋的人工智慧，隱私保護，智慧建築

本所依行政院產業科技策略會議(SRB)決議，推動智慧化居住空間產業發展。基於國際電機電子工程師學會係國際權威機構，爰參加其主辦之 2024 年智慧科技應用信任、隱私及安全國際研討會。本次蒐集有關國際產、官、學、研界，為因應 2024 年歐洲議會通過全球首部《人工智慧法》及經濟合作暨發展組織(OECD)倡議可信任的人工智慧理念，所發展的一系列關於可信任的人工智慧科技、可解釋的人工智慧技術，以及智慧空間中的隱私保護等新知，將作為本所規劃後續智慧化居住空間應用人工智慧物聯網科技計畫研究課題參考，透過將可信任的人工智慧技術引進智慧化居住空間產業，引導產業發展符合國際發展趨勢之產品及服務，以發揮我國資通信產業之國際競爭優勢。

目次

表次

圖次

壹、 出國目的

因應國際智慧科技發展趨勢，本所依據行政院智慧國家方案(2021-2025 年)上位科技政策及內政部推動「營造安居家園」施政目標，辦理行政院核定之本所「智慧化居住空間應用人工智慧物聯網科技計畫(112-115 年) 」，旨在發展以人為本之人工智慧物聯網科技建築應用。基於近年人工智慧、物聯網等智慧科技之蓬勃發展，世界上越來越多人口成為各式各樣智慧科技之使用者，所衍生之使用者資訊隱私保護課題成為值得關注之重要課題。爰派員參加本次由國際電機電子工程師學會（IEEE）專業權威機構主辦之研討會，會議為國際跨學科論壇，旨在展示最先進創新成果，促成產、學、研就新興智慧系統和應用中的信任、隱私和安全課題進行討論交流，參加會議可蒐集國際最新資料，作為本所推動智慧化居住空間應用人工智慧物聯網科技業務之參考。

本次會議由國際電機電子工程師學會(Institute of Electrical and Electronics Engineers, IEEE)主辦，該學會係國際性智慧科技技術及標準之權威機構，參加會議可蒐集國際間最新各國產、官、學、研界探索智慧科技應用信任、隱私及安全課題之經驗，供本所推動「智慧化居住空間應用人工智慧物聯網科技計畫(112-115 年) 」內容滾動式修訂參考。

貳、 出國行程

一、 主辦單位及會議宗旨

2024 年第 6 屆智慧科技應用信任、隱私及安全國際研討會(The Sixth IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Application)於美國華盛頓特區召開，由國際電機電子工程師學會(Institute of Electrical and Electronics Engineers, IEEE)協會主辦，獲得美國國家科學基金會贊助(見圖 2)。

該研討會自 2019 年召開第 1 屆至今，為每年定期召開之國際性跨領域論壇，旨在展示最先進的創新成果，邀集學術界、產業研究人員和專業從業者就新興智慧系統和應用中的信任、隱私和安全相關問題進行討論，為從業人員和研究人員、技術人員和工程師以及決策者和資助機構提供一平台，分享知識和經驗、發展能力，並建立社會關係(IEEE, 2024)。

IEEE TPS 2024 Call for Papers

Scope

Recent advances in computing and information technologies such as IoT, mobile Edge/Cloud computing, cyber- physical-social systems. Artificial Intelligence/Machine Learning/ Deep Learning. etc., have paved way for creating next generation smart and intelligent systems and applications that can have transformative impact in our society while accelerating rapid scientific discoveries and innovations. Such newer technologies and paradigms are getting increasingly embedded in the computing platforms and networked information systems/infrastructures that form the digital foundation for our personal, organizational and social processes and activities. It is increasingly becoming critical that the trust, privacy and security issues in such digital environments are holistically addressed to ensure the safety and well-being of individuals as well as our society.

IEEE TPS-ISA is an international multidisciplinary forum for presentation of state-of-the art innovations, and discussion among academic, industrial researchers, and practitioners on issues related to trust, privacy and security in emerging smart and intelligent systems and applications.

Topics

Topics of interest include, **but are not limited to**:

- Foundational, theoretical models for trust, privacy and security in emerging applications
- Trusted AI, Machine Learning and Deep Learning
- Privacy preserving Machine Learning and Deep Learning
- Trustworthy, safe and resilient intelligent systems
- Trusted, privacy-conscious and secure systems, applications and networks/infrastructures
- Security and privacy in IoT and Cyber-physical-human systems
- Trustworthy and secure Human-Machine collaboration
- Access and trust management/negotiation, and secure information flow/sharing
- Bio-inspired approaches to trust, privacy and security
- Game theoretical approaches to trust, privacy, and security

圖 1 2024 年智慧科技應用信任隱私及安全國際研討會公告資料

(資料來源：2024 年智慧科技應用信任、隱私及安全國際研討會網站)

二、會議時間及議程

本屆會議訂於美國華盛頓特區當地時間 2024 年 10 月 28 日至 10 月 30 日召開，會議主題是基於物聯網、行動邊緣/雲端運算、網路實體社會系統(cyber- physical-social systems)、人工智慧/機器學習/深度學習等運算和資訊科技的最新進展，為創建下一代智慧系統和應用程式鋪展發展之道，將產生變革性影響，加速科學發現與創新。這些創新的技術被引入計算平台和網路資訊系統/基礎設施，構成個人、公私組織和社會的數位發展基礎。為全面解決此類數位環境中的信任、隱私和安全問題，確保個人以及社會的安全和福祉工作變得越來越重要，因而邀集國際跨領域產、官、學、研界就新興智慧系統和應用中的信任隱私和安全相關課題進行討論。研討會議程如表 1 所示。

表 1 智慧科技應用信任、隱私及安全國際研討會議程表

<table>
<tr><th>日期(當地時間)</th><th colspan="4">研討會議程</th></tr>
<tr><td rowspan="4">2024 年 10 月 28 日（星期一）</td><td colspan="4">歡迎及開幕致辭
(指導委員會主席及組織委員會主席)</td></tr>
<tr><td colspan="4">專題演講：可信任的人工智慧
哥倫比亞大學電腦科學系教授 Jeannette M. Wing
主持人：James Joshi（美國匹茲堡大學）</td></tr>
<tr><td>惡意軟體偵測查證和深度學習
主持人：Amir Masoumzadeh（美國紐約州立大學奧爾巴尼分校）</td><td>因果推理、人工智慧、和邏輯推理
主持人：Julian Jarrett (美國 Lutron 電子公司)</td><td>自動駕駛人工智慧系統、安全
主持人：Mei-Ling Shyu (美國密蘇里大學堪薩斯分校）</td><td>包容性人工智慧
主持人：Hemant Purohit（喬治梅森大學），Jin-Hee Cho（弗吉尼亞理工學院), Yoosun Chung (喬治梅森大學)</td></tr>
<tr><td colspan="4">午休</td></tr>
</table>

<table>
<tr><th>日期(當地時間)</th><th colspan="4">研討會議程</th></tr>
<tr><td rowspan="3"></td><td colspan="4">專題演講：研究助手的未來發展願景
Ed H. Chi (Google DeepMind 公司 大型語言模型/聊天機器人部門傑出科學家兼研究主管)
主持人：Huan Liu(美國亞利桑那州立大學)</td></tr>
<tr><td colspan="4">主題：人工智慧如何重塑科學研究？
討論來賓：Li Yang（美國國家科學基金會）、Hemant Purohit（美國喬治梅森大學）、Ling Liu（美國喬治亞理工學院）、Huan Liu（美國亞利桑那州立大學）、Ed Chi（美國 Google 公司）
主持人：Paolo Boldi（義大利米蘭大學）</td></tr>
<tr><td>隱私增強技術和網路安全威脅
主持人：
Lavanya Elluri
（美國德州農工大學）</td><td>人工智慧增強安全性和自動監控
主持人：
Julian Jarrett
(美國 Lutron 電子公司)</td><td>人工智慧系統與應用
主持人：
Latifur Khan
（美國德州大學達拉斯分校）</td><td>包容性人工智慧</td></tr>
<tr><td rowspan="2">2024 年 10 月 29 日（星期二）</td><td colspan="4">專題演講：美國國家科學基金會在人工智慧未來中的作用
Michael L. Littman，美國國家科學基金會資訊與智慧系統部主任、布朗大學電腦科學系教授
主持人：Jaideep Vaidya(美國羅格斯大學）</td></tr>
<tr><td>考慮隱私與安全的大型語言模型
主持人：
Indrajit Ray
（美國科羅拉多州立大學）</td><td>系統、應用程式和 AI 效能
主持人：
Keke Chen(美國馬里蘭大學巴爾的摩分校)</td><td>人工智慧驅動的系統、安全和計算
主持人：
Barbara Carminati
（義大利伊蘇布里亞大學）</td><td>人工智慧、量子運算和網路安全
主持人：
Amir Masoumzadeh
（美國紐約州立大學奧</td></tr>
</table>

<table>
<tr><th>日期(當地時間)</th><th colspan="4">研討會議程</th></tr>
<tr><td rowspan="5"></td><td></td><td></td><td></td><td>爾巴尼分校）</td></tr>
<tr><td colspan="4">學生發表短篇演講：Jaideep Vaidya（美國羅格斯大學）
與談人：Jaideep Vaidya（美國羅格斯大學）、Huan Liu（美國亞利桑那州立大學）、Indrakshi Ray（美國科羅拉多州立大學）、Stacey Truex（美國丹尼森大學）、Peter Kairouz（美國 Google）、Ambareen Siraj（美國國家科學基金會，招募培訓新世代網路安全專業人員獎學金/網路空間的安全、隱私和信任計畫）
主持人：Wenqi Wei (美國紐約復旦大學)</td></tr>
<tr><td colspan="4">專題演講：變革社會：數位轉型加速 TDK 願景
Roshan Thapliya, 日本 TDK 公司 首席數位轉型長兼總經理
主持人：James Joshi，美國匹茲堡大學</td></tr>
<tr><td>人工智慧/機器學習中的隱私與安全
主持人：Dipankar Dasgupta，美國孟菲斯大學</td><td>人工智慧驅動的決策與以人為本的系統
主持人：Keke Chen (美國馬里蘭大學巴爾的摩分校)</td><td>以人為本的系統的人工智慧和語言模型
主持人：Kim HemmingsJarrett (美國賓州州立大學)</td><td>人工智慧安全、隱私和醫療保健
主持人：Indrakshi Ray（美國科羅拉多州立大學）</td></tr>
<tr><td rowspan="2">2024 年 10 月 30 日（星期三）</td><td colspan="2">惡意軟體和威脅偵測
主持人：Pooria Madani (加拿大安大略理工大學)</td><td colspan="2">行為與網路系統的人工智慧驅動建模與分析
主持人：Yanzhao Wu（佛羅里達國際大學）</td></tr>
<tr><td colspan="4">專題討論：推動新興技術創新：研發重點與融資
成員：Jennifer Roberts（美國衛生高級研究計畫署韌性系統主任）、Heidi Sofia（美國國立衛生研究院國家生物技術資訊中心副主任）、Elham Tabassi（美國國家標準暨技術研究院新興技術部門副主任）、Cliff Wang（美國國家科學基金</td></tr>
</table>

<table>
<tr><th>日期(當地時間)</th><th colspan="2">研討會議程</th></tr>
<tr><td rowspan="4"></td><td colspan="2">會網路空間安全、隱私和信任計畫主持人），Craig Schlenoff 博士（白宮科技政策辦公室網路與資訊科技研究與發展計畫主任）
主持人：James Joshi (美國匹茲堡大學)，Huan Liu(美國亞利桑那州立大學)</td></tr>
<tr><td>存取控制和安全模型
主持人: Wenqi We (美國紐約復旦大學)</td><td>認知科學中的計算科學發現
主持人：Fernand Gobet</td></tr>
<tr><td colspan="2">以人工智慧對抗人工智慧：不可避免的下一個虛擬研究前沿</td></tr>
<tr><td colspan="2">研討會閉幕式</td></tr>
</table>

(資料來源：本報告翻譯整理自會議議程，2024)

圖 2 研討會大會專題演講

(資料來源：作者自行拍攝，2024)

參、 會議過程及涉及本所建築研究業務事項

本屆研討會論文主題廣泛，涉及：可信任的人工智慧、隱私增強技術和網路安全威脅、人工智慧驅動的決策與以人為本的系統、人工智慧/機器學習中的隱私與安全、以人為本的系統的人工智慧和語言模型、推動新興技術創新：研發重點與融資等議題，本報告摘錄其中與本所建築研究業務較相關事項說明如下：

一、 可信任的人工智慧

研討會開幕式邀請哥倫比亞大學電腦科學系教授 Jeannette M. Wing 發表可信任的人工智慧(Trustworthy AI))專題演講(Wing, 2024)：

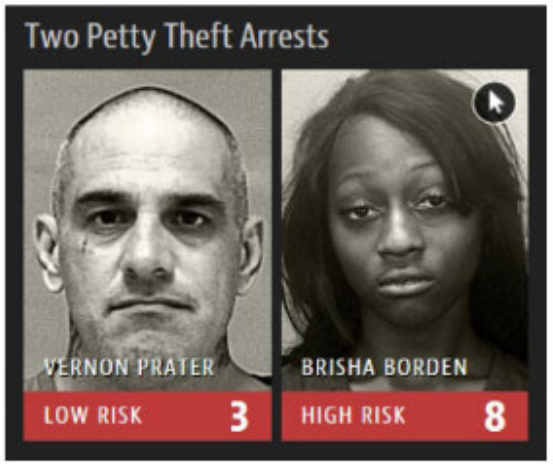
(一) 技術發展背景及目的

某些任務中，人工智慧系統已經有足夠好的效能，可應用在日常生活。例如：物體辨識有助於提高汽駕駛的視覺能力，語音辨識可協助手機或智慧家庭的語音助理進行對話，甚至某些任務，人工智慧系統超越了人類的表現，例如：AlphaGo 是擊敗世界上最好的圍棋棋手的第一個電腦程式；未來人工智慧科技將會駕駛人們的車、幫助醫師更準確地診斷疾病、輔助法官做出更一致的法院判決、協助雇主僱用更合適的求職者。

然而，這些應用資料建立的人工智慧系統可能很脆弱且不公平。例如：在停車標誌上添加塗鴉，就可以愚弄汽車的環境影像分類器，讓汽車認定不是停車標誌；在良性皮膚病變的圖像中添加雜訊，可欺騙疾病診斷分類器，使其誤認醫學影像出自惡性疾病病患；過去美國法院使用的人工智慧系統犯罪風險評估工具，已被證明對黑人有偏見(詳圖 3)；企業員工招募人工智慧系統亦被證明對女性有偏見。

在此背景下，2023 年美國拜登總統於發布「發展與使用安全且可信任的人工智慧行政命令」 (Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)，說明人工智慧擁有巨大潛力，有希望也有危險。負責任的人工智慧使用(Responsible AI use)有助解決挑戰，使世界更加繁榮、高效、創新和安全。不負責任的使用可能會加劇詐欺、歧視、偏見和虛假資訊等社會危害、取代工人並剝奪他們的權力、抑制競爭，對國家安全構成風險。

Robustness and Fairness



D. Mandal, S. Deng, D. Hsu, S. Jana, and J.M. Wing, "Ensuring Fairness Beyond the Training Data," to appear in *Proceedings of the 34th Conference on Neural Information Processing Systems* (NeurIPS), December 2020. arXiv:2007.06029, July 2020. July 2020.

圖 3 具有偏見的司法資訊系統使黑人被標記為罪犯機率約是白人兩倍

(資料來源：Jeannette M. Win, 2024)

(二)系統需求及設計

1. 從可信任的計算到可信任的人工智慧

1999 年美國國家科學院出版具有里程碑意義的《網路空間信任報告》(Trust in Cyberspace)是可信任計算及其後續的研究奠定基礎，美國國家科學基金會(NSF)啟動了一系列關於任的計畫。從被信任的計算(Trusted Computing)（2001 年發起）開始，然後是網路信任(Cyber Trust)（2004 年）、可信任的計算(Trustworthy Computing)（2007 年），現在是安全可信任系統(Secure and Trustworthy Systems)（2011 年），於國際電機電子工程師學會(IEEE)計算機和資訊科學與工程理事會在可信任方面發展了學術研究社群。儘管它始於計算機科學界，但對可信任計算研究的支持現在涵蓋了美國國家科學基金會的多個部門，並吸引了許多其他資助組織，包括透過網路和資訊技術研究與開發(Networking and Information Technology Research and Development,NITRD) 計畫、20 個聯邦機構。

產業界也是可信任計算領域的領導者和積極參與者。比爾蓋茲 2002 年發布「可信任運算」備忘錄，微軟公司向其員工、客戶、股東和資訊科技領域的其他人員表明了可信任軟體和硬體產品的重要性，微軟公司內部白皮書確定可信度的 4 大支柱：安全性、隱私、可靠性和商業誠信。前 3 個屬性面向客戶，讓客戶有充分的理由信任微軟公司的軟體和服務。

經過 20 年產官學研界投資研發帶動進步，可信任的觀念至今涵蓋以下屬性：

(1) 可靠(Reliability)：系統是否做正確的事？
(2) 安全(Safety)：系統不會造成傷害嗎？
(3) 保護(Security)：系統有多容易受到攻擊？
(4) 隱私(Privacy)：系統是否保護個人的身分和資料？
(5) 可用性(Availability)：當我需要存取系統時系統是否已啟動？
(6) 友善使用(Usability)：人類可以輕鬆使用它嗎？

我擁有以上屬性的計算系統是硬體和軟體混合系統，並考慮系統與人類和實體世界的互動。

人工智慧系統除以上屬性之外，還需要更多屬性，例如：

(1) 準確性(Accuracy)：與訓練和測試的資料相比，人工智慧系統在訓練過程中未曾看見的新資料上表現如何？
(2) 穩健性(Robustness)：人工智慧系統結果對輸入變化的敏感度如何？
(3) 公平性(Fairness)：人工智慧系統結果是否公正、無偏見？
(4) 問責制(Accountability)：由何人或以何種方式對人工智慧系統產生結果負法律責任？
(5) 透明度(Transparency)：外部觀察者是否清楚系統的結果是如何產生的？
(6) 可解釋性/可解釋(Interpretability/Explainability)：人工智慧系統產出的結果是否可以人類可理解，以及/或是對最終使用者具有意義的解釋，來證明結果是合理的？(詳圖 4)(Agency, 2018)
(7) 道德(Ethical)：為建構人工智慧系統所收的資料集是否符合道德？人工智慧系統產出的結果會以合乎道德的方式使用嗎？

過去機器學習社群常將上述準確性視為最重要標準，但其中一些屬性彼此是互不相容的(incompatible)，可信任的人工智慧，引導我們進一步就以上屬性加以權衡。例如，願意放棄一部分的準確性，以建構更公平(fairness)、減少偏差的模型(詳圖 5)。此外，上述一些屬性可能有不同的解釋。例如，公平有各種解釋，群體間公平(group fairness)和個人公平(individual fairness)，都是合理的。

傳統軟硬體系統在大多數情況下，可以根據離散機率(discrete)和確定(deterministic)狀態來預測系統行為。當今的人工智慧系統，特別是使用深度神經網路(deep neural networks)的系統，為運算系統增加一定的複雜性。複雜性是由於具有機率性質。人工智慧系統透過機率對人類行為的不確定性和物理世界的不確定性進行建模。機器學習的最新進展依賴巨量資料，來自現實世界的資料只是機率空間中的一些資料點。因此，可信任的人工智慧會將我們的注意力從傳統運算系統的確定性質轉換到的機率性質。

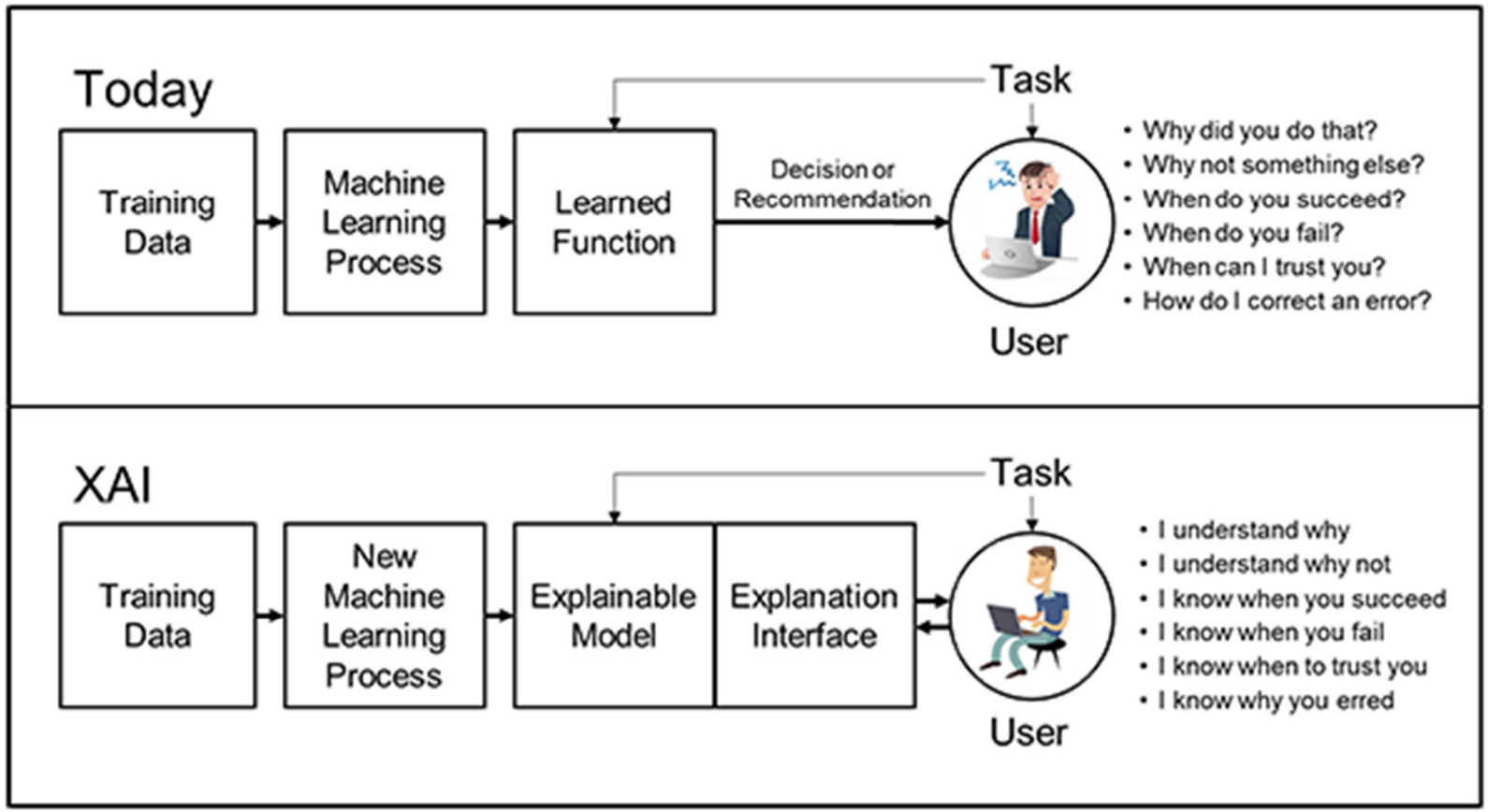


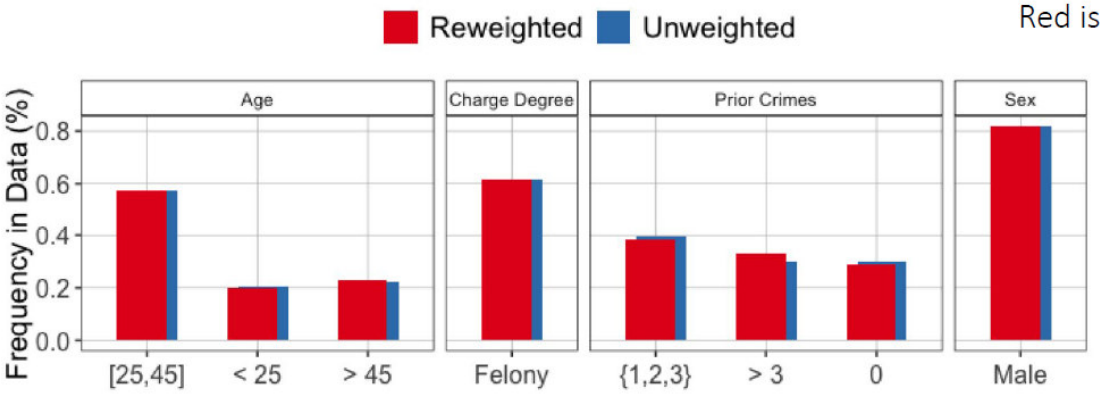
圖 4 人工智慧科技應用成果應可向使用者解釋

(資料來源：美國國防部高等研究計畫局., 2018)

Robust and Fair Classifiers

- State-of-the-art “fair” classifiers are not robust

Blue is unweighted (“fair”).
Red is reweighted (“unfair”).



- For **fairness**, we might want to show the ML model is fair on a given dataset and all unseen datasets that are “similar” (for some formal notion of “similar”).
- Use on-line algorithm (two-player game) to build a fair classifier that is robust to a *class* of distributions.

圖 5 利用對建模資料進行加權以建立穩健且公平的分類器

(資料來源：Jeannette M. Win, 2024)

2. **設計、實施、部署及驗證可信任的人工智慧方法**

傳統的計算系統建立最終使用者信任的一種方法是形式驗證，其驗證方式是就欲驗證的屬性作一次性證明，例如：對於程式的所有輸入行為，或驗證能否識別產生錯誤輸出值或未能滿足所需，從而提供有關如何改進系統的寶貴回饋。此種證明模型可滿足某種屬性的驗證方式，其優點在於無需逐一測試各個輸入值或行為，是一種提供可證明保證的方法，從而增加人們對系統按預期運作的信任。然而，對於大型或具有無限狀態空間的人工智慧系統而言，是不適用的。

人工智慧系統的驗證，可以被視為一個複雜系統，系統中含有一個以上的機器學習模型模組，例如：自動駕駛汽車的電腦視覺系統具有深度神經網路，為驗證系統的安全性或穩健性，必須證明在交通、道路、行人、建築物等指定環境中的汽車，具有系統設定的屬性。

驗證人工智慧系統比傳統形式化方法的難度更高，原因在於：首先，模型具備的性質需要機率推理，從現實世界獲取資料建立的模型，是以隨機過程進行數學建模，輸出與機率相關。其次，由機器生成的機器學習模型，不一定是人類可讀或可理解的模型，可信任人工智慧系統的一些所需屬性（例如：透明度或道德）目前尚無形式化技術或可能無法形式化。 因此，目前人工智慧系統的驗證僅限於可形式化的範圍。更具挑戰性的是，這些驗證技術需要對機器產生的程式碼進行操作，特別是那些本身可能無法事先確定會產生的程式碼。

機器學習模型基本概念是根據一組訓練和測試資料建構的模型，能夠對它以前從未學習過的資料進行預測，通常達到一定程度的準確性。為驗證模型，建議對該資料做出明確的假設，並將待驗證問題表達為資料與模型具備應有屬性。透過增加更多資料來訓練或測試模型是否會使其更穩健、更無偏差等屬性。若測試結果呈現該屬性不成立，應如何修復模型、修改屬性或決定收集哪些新資料來重新訓練模型？思考機器學習模型驗證中的「反例」意義為何？

3. **發展及推廣可信任的人工智慧**

2019 年 10 月，美國國家科學基金會宣布了一項資助國

家人工智慧研究所的新計畫持續至今。6 個主題之一的名稱是「可信任的人工智慧」(Trustworthy AI)。它強調可靠性(reliability)、可解釋性(explainability)、隱私性(privacy)和公平性fairness)等屬性。

2020 年 10 月，美國哥倫比亞大學資料科學研究所主辦了首屆可信任人工智慧研討會，匯集來自形式方法、安全和隱私、公平性和機器學習等領域的研究人員，來自產、學界正在探索各種問題和方法，參與者確定下列研究挑戰課題：

(1) 訂定規格和驗證技術
(2) 正確建模的技術
(3) 新的威脅模型和對抗系統級攻擊
(4) 考慮可解釋性、透明度和責任等屬性的審核流程
(5) 檢測偏差和去偏差資料的方法、機器學習演算法及其輸出
(6) 用於試驗可信度屬性(trustworthiness properties)的系統基礎設施
(7) 理解人的因素，例如：機器在哪些方面影響人類行為
(8) 理解社會要素，包括：社會福利、社會規範、道德、倫理和法律。

以上研討會結束至今，資通信產業持續積極發展許多可信任的人工智慧技術及解決方案(詳圖 6)(*Trustworthy AI*, 2024)。

參考過去發展可信任計算過程，形式化方法只是確保增強人工智慧系統信任的一種方法，這個領域需要探索多種方法來實現可信任的人工智慧。此外，除了技術挑戰外，還存在社會、政策、法律和道德挑戰，未來可以參考貝蒙特原則：尊重人、行善、正義，發展可信任的人工智慧應用(詳圖 7)。

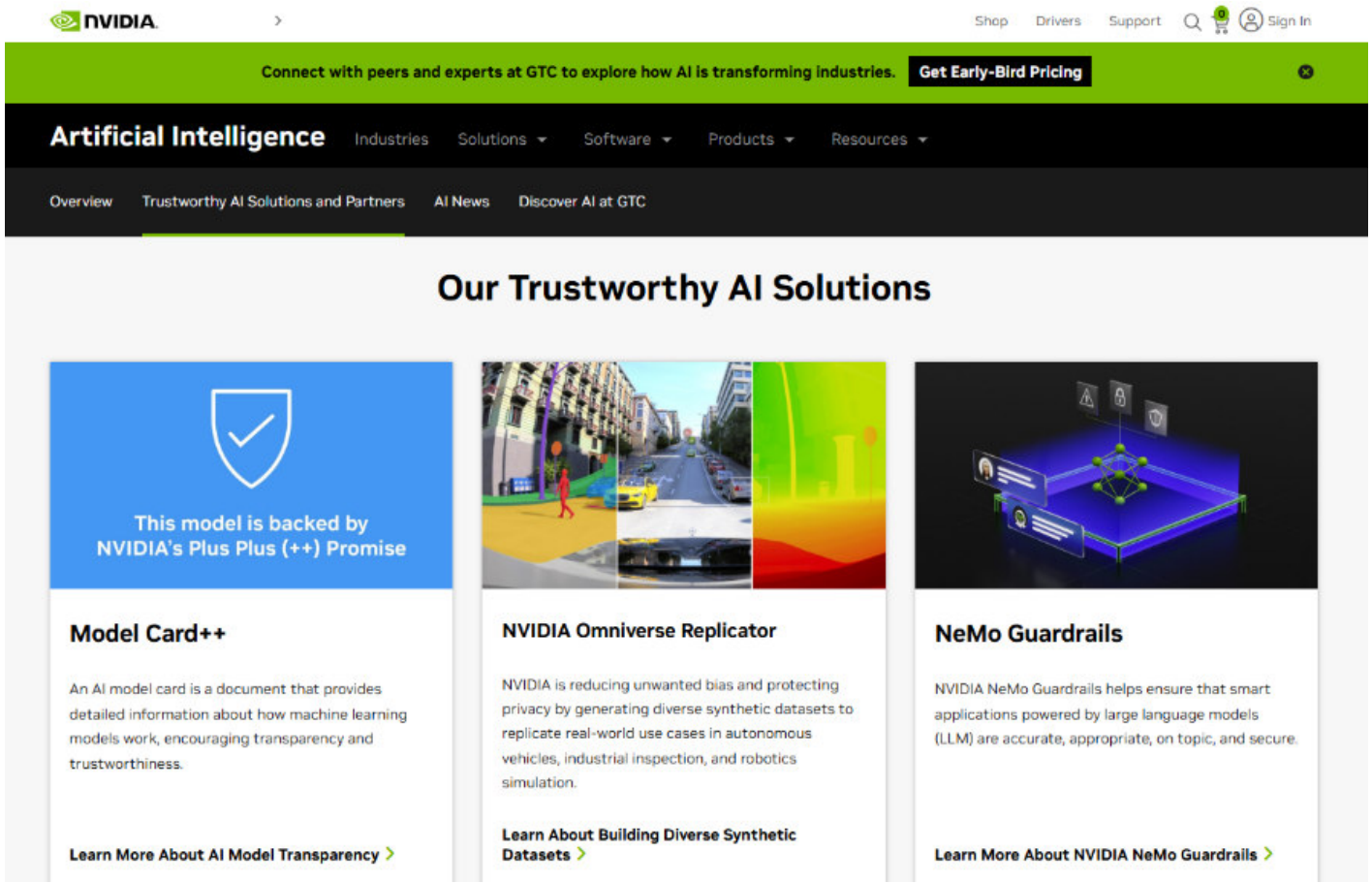
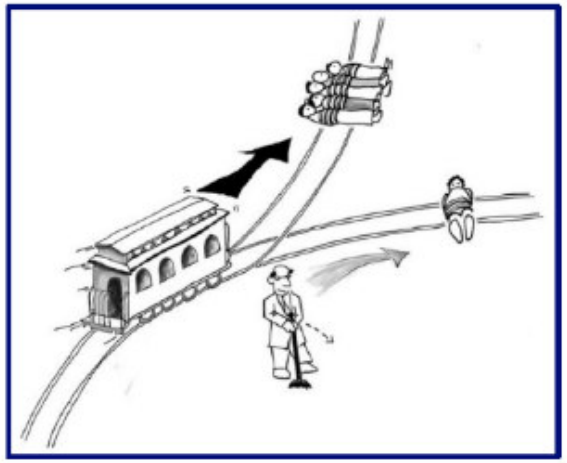


圖 6 產業發展各種可信任的人工智慧解決方案

(資料來源：NVIDIA Corporation, 2024)

Belmont Principles Applied to AI



Respect for Persons

Example: People should always be informed when they are talking to a chatbot.

Beneficence

Example: Risk/benefit analysis on the decision a self-driving car makes on whom not to harm.

Justice

Examples: Ensure the fairness of risk assessment tools in the court system and automated decision systems, e.g., used in hiring.

圖 7 貝蒙特原則應用於人工智慧

(資料來源：Jeannette M. Win, 2024)

二、發展可解釋的人工智慧技術以促進信任

近 20 年開始發展的可解釋的人工智慧(Explainable Artificial Intelligence, XAI)概念與技術旨在增強機器學習模型的透明度和可解釋性，發展背景係基於透過訓練資料建立的機器學習模型，尤其是機器學習系統中的深度學習「黑盒子」不透明性，使預測結果難以被人類理解，因此必須再轉換為人類可解釋的人工智慧模型(Explainable AI Model)及人類可解釋的介面(Explainable Interface)，確保作成的決策是人類可理解、可預測何時可成功或失敗、知道何時可信任或不可信任、以及何時與為何會犯錯，使人類應用人工智慧做決策時更自於理解機器預測結果，具體實現方式包括：增加可解釋性(Interpretability)、增加透明度(Transparency)、增加可靠性(Reliability)、增加穩健性(Robustness) 、強化隱私(Privacy) 、視覺化(Visualization) (詳圖 8)等(Olusegun & Yang, 2024)。

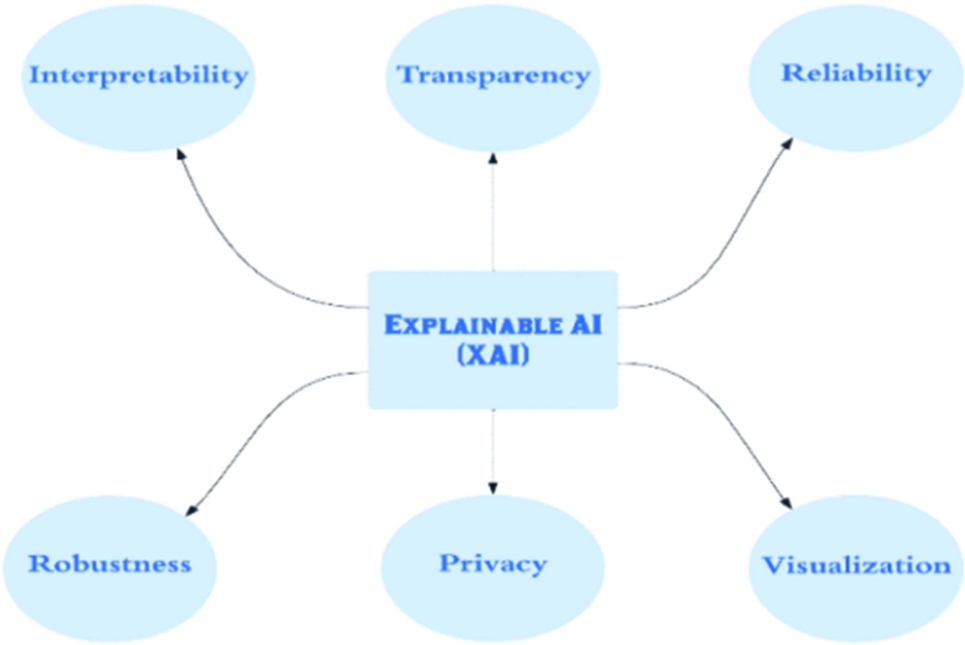


圖 8 可解釋的人工智慧（XAI）

(資料來源： Olusegun & Yang, 2024)

而最近 8 年因歐盟、美國聯邦政府積極發展人工智慧法律規範引進可解釋的人工智慧相關概念，更進一步帶動國際間相關研究的發展。

本次研討會發表實現可解釋的人工智慧技術的一些最新研究成果，例如：基於傳統的欺詐檢測經常導致錯誤分類，合法交

易被標記為欺詐，並且有些經常由於偽陽性和偽陰性而無法檢測到真正的欺詐，從而破壞欺詐檢測系統。針對此一問題所開發具有可解釋性、具有透明性的人工智慧詐欺檢測資訊系統，可有效檢測網路中的細微詐欺模式，在幾秒鐘內識別欺詐模式、減少分類錯誤、引入可解釋的特徵選擇新方法以增強人工智慧模型的透明度，會議並以此案理說明發展可解釋的人工智慧系統建議流程：

首先，將人工智慧建模所使用的資料集(Datasets)轉為 CSV 資料格式，其次，進行資料預處理(Data Preprocessing)，就缺漏資料或錯誤資料進行處理(Handling Missing values)，視需要將資料進一步作正規化(Normalization)、檢視建模的樣本資料是否平衡(Data Balancing)具有母體代表性，再進行建模型資料的特徵篩選，確認選擇特徵具可解釋性(Interpretable Feature Selection)，再以訓練資料(Training data) 進行多層感知器 (MLP)、長短期記憶 (LSTM)、卷積神經網路 (CNN) 及卷積神經網路和長短期記憶 (CLSTM)等多種深度學習神經網絡(DL Neural Network)建模及測試資料(Testing data)進行模型績效檢核評估，實現可解釋的人工智慧(ＸＡＩ)的檢測資訊系統(Detection system) (詳圖 9)。

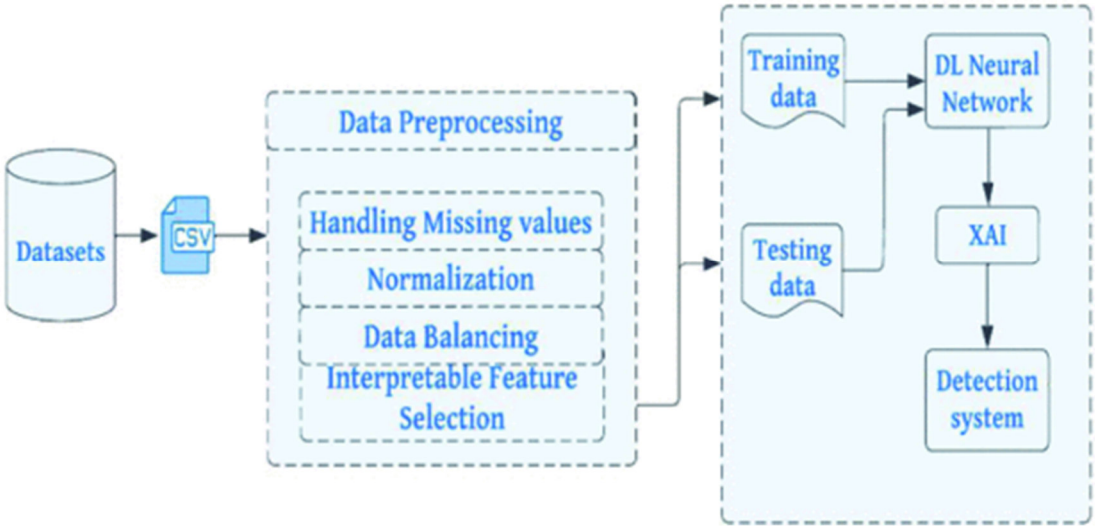


圖 9 可解釋的人工智慧檢測資訊系統例子發展流程

(資料來源： Olusegun & Yang, 2024)

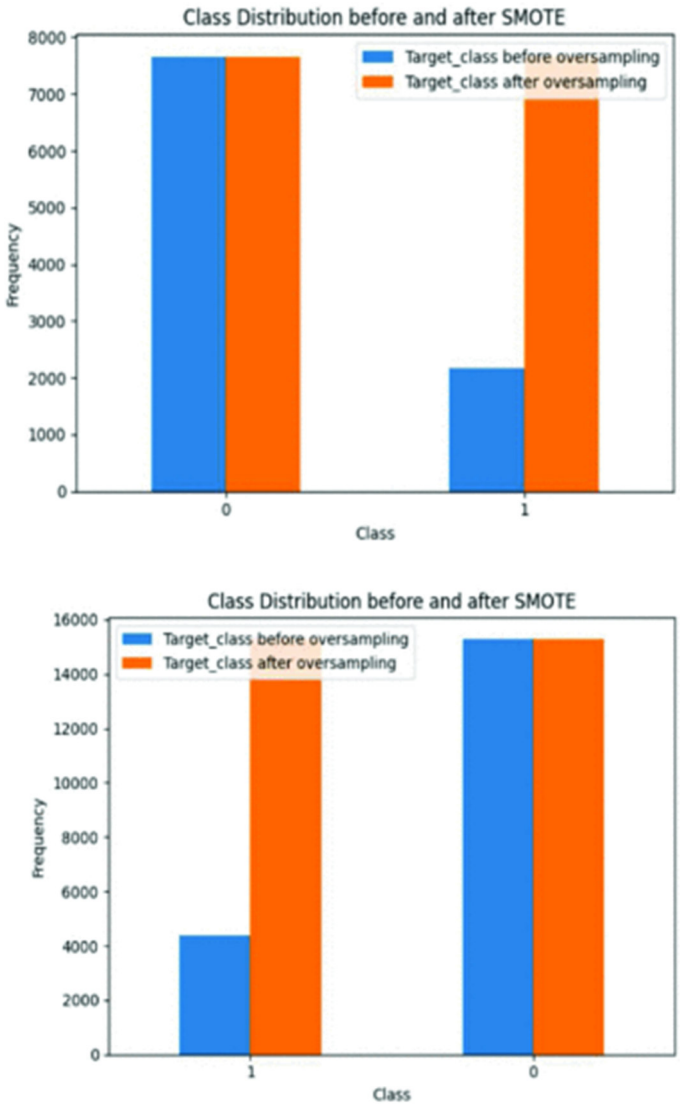


圖 10 生成少數樣本以更均衡的樣本類型訓練模型改善人工智慧預測偏差

(資料來源： Olusegun & Yang, 2024)

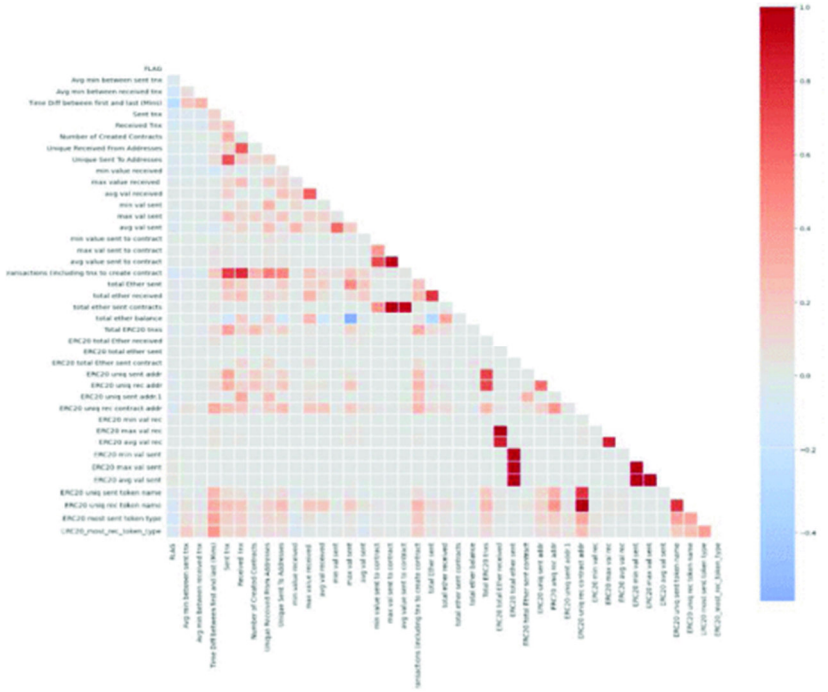


圖 11 透過相關性分析刪除重要性較低特徵以提高模型可解釋性

(資料來源： Olusegun & Yang, 2024)

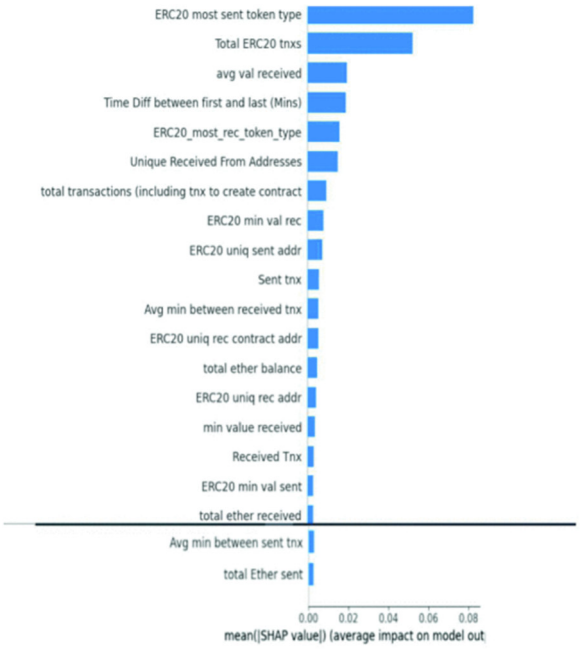


圖 12 刪除重要性值較低特徵以節省訓練時間分析

(資料來源： Olusegun & Yang, 2024)

研究結果顯示，基於可解釋特徵選擇方法來開發基於神經網路的深度學習模型，在大量特徵中，刪除特徵重要性值較低者，可消除冗餘。這使得模型能夠專注於具有重大影響的特徵，同時減少了資料維數空間，從而縮短了訓練時間，可在短短幾秒鐘內就達到 97~99%之間的：準確率(Accuracy)、精確率(Precision)、召回率(Recall)、調和平均數(F1-Score)。從以上評估機器學習分類模型表現時常用的 4 個重要指標可知，系統在篩檢出詐欺資訊的比率、檢測出詐欺資訊正確的比率方面都有良好的成果，能夠有效地最大限度地減少偽陽性和偽陰性錯誤，這項研究驗證了可解釋的人工智慧技術兼具人類可理解、高效率以及促進人工智慧模型透明度和信任等特色。

表 2 可解釋的人工智慧模型與傳統人工智慧模型準確率比較

Models	Accuracy	Precision	Recall	F1 Score	ROC
Modified LSTM [14]	0.9507	0.9442	0.9583	0.9512	NA
GA-CS +DLANN [35]	0.9860	0.9718	0.9431	0.9486	0.9887
Modified LGBM [36]	0.9875	0.9718	NA	0.9486	NA
IFS-TabPFN	0.9915	0.9848	0.9785	0.9817	0.999

(資料來源： Olusegun & Yang, 2024)

三、 智慧空間中的隱私保護

研討會也發表了在智慧空間(*Smart Spaces*)保護隱私的新技術。智慧空間是指透過感測器網路監控的環境(*E*nvironments Monitored through A Network of Sensors)，基於資料從感測器通過資訊線路，實現各種需求應用。智慧空間資料還可以儲存以供將來分析和處理，以實現新的目的，並透過資料學習改進已部署應用程式。資料處理可以在邊緣執行（在感測器上或受信任的本地伺服器上），也可以被委託到（可能不受信任的）公有雲，為了在智慧空間中實現隱私保護，可發展應用程式介面(Application Programming Interface, API）以控制何時收集資料、從哪些感測器收集資料、資料以何種格式提供給設備/機器，以及向誰提供，並可以攔截資料以應用差異隱私(Differential Privacy)、加密(Encryption)，或採用基於策略的共用(Policy-based Sharing)等其他各種隱私保護技術(Privacy Enhancing Technologies, PETs)(Farrukh et al., 2024)。

1. **隱私權被多數國家認為是一項憲法和法律明定的基本人權：**隱私權被廣泛認為是一項基本人權，對於維護人的尊嚴、自主權、個人身分和防止歧視至關重要，特別是保護弱勢群體免受更強大實體的剝削。這項權利已被編入許多國家的憲法和法律架構中。值得注意的例子包括歐洲的「一般資料保護規範」（General Data Protection Regulation,GDPR）、加州消費者隱私案（California Consumer Privacy Act, CCPA）等。在智慧空間中，隱私問題尤其重要。

   2022 年智慧空間被美國著名的 Gartner 資訊科技研究和顧問公司評為一項關鍵的新興技術，應用範圍廣泛，從加強城市規劃到改善醫療保健服務等，雖然智慧空間技術具有巨大的潛力，但相對於從線上互動中收集資訊的社交媒體和網際網路數位平台而言，智慧空間對隱私構成更大的風險，主要原因係智慧空間能夠即時擷取有關人們實體活動的資料，即使是看似無害的運動探測器，也可以產生有關個人的高度精細資訊。

2. **應用智慧空間資料共享面臨機會和隱私挑戰：**應用智慧空間技術於高齡者照護機構同時面臨機會和挑戰，在這些機構中，個人居住在提供護理照顧服務的集體生活單元中，智慧空間技術可用於提高高齡者照護品質和及時性，減少人為錯誤，同時降低成本。例如，智慧技術可以偵測跌倒風險和實際跌倒，這是高齡者受傷

的主要原因；它也可以透過監控和提醒高齡者的照護需求來提高護理人員的服務效率。然而，所有這些使用科技的干預措施都會對所有相關人員（護理人員、受試者和訪客）造成巨大的隱私風險。

在智慧空間中，資料跨異質設備流動，可能支援多種即時控制和分析服務/應用，旨在提高營運效率、改善居住者舒適度和確保室內環境的安全，例如，學校的上課出勤率、空調系統的使用狀態分析及消防安全。為說明智慧空間中的隱私問題，首先提出智慧空間定義，說明實體構成及資料處理和共享方式，如圖 13 所示。

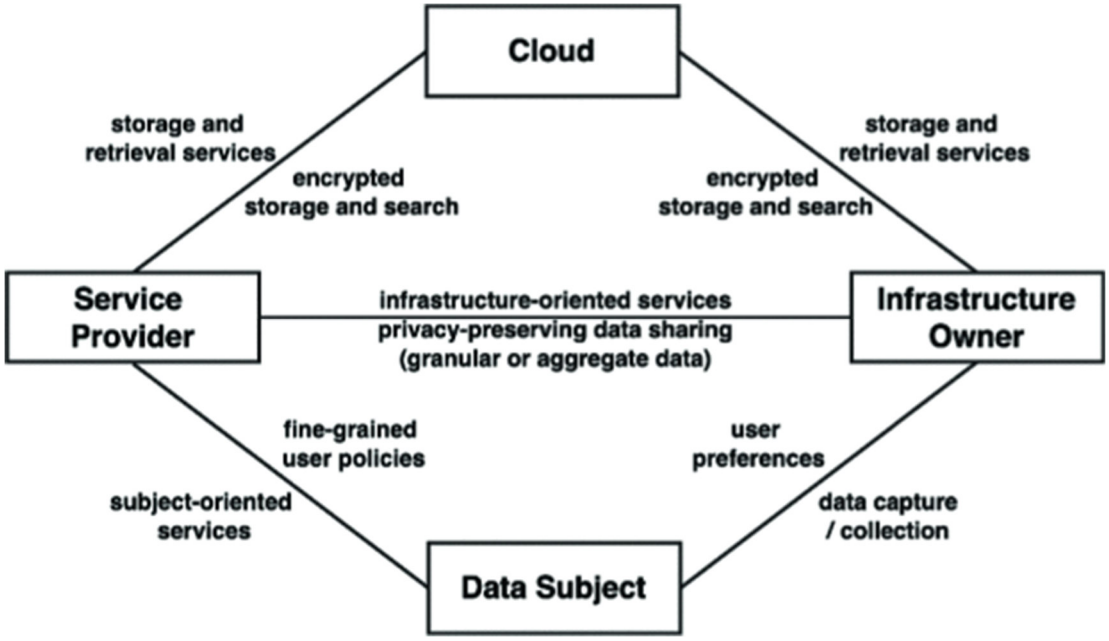


圖 13 智慧空間中的資料共享

(資料來源： Farrukh et al., 2024)

構成智慧空間中的 4 大單元包括：雲端(Cloud)、服務提供者(Service Provider)、基礎設施所有者(Infrastructure Owner)、資料當事人(Data Subject)。各單元彼此之間的資料處理、共享和服務包括：儲存和檢索服務(storage and retrieval services)、加密儲存和搜尋(encrypted storage and search)、儲存和檢索服務(storage and retrieval services)、加密儲存和搜尋(encrypted storage and search)、基礎設施導向服務(infrastructure-oriented services)、隱私保護資料共享

(privacy-preserving data sharing)、分散或聚合資料(granular or aggregate data)、分散的使用者策略(fine-grained user policies)、使用者偏好(user preferences)、當事人導向服務(subject-oriented services)、資料捕捉/收集(data capture /collection)(Farrukh et al., 2024)。

上述資料共享過程引入了隱私問題，從隱私/安全角度來看，所涉及的實體通常不會完全信任彼此，可能會試圖透過獲取的資訊來了解他人，基礎設施所有者必須保留所獲資料控制權。例如，在智慧辦公室中，員工應該能夠決定是否記錄和儲存運動感測器追蹤的他們的運動，以及誰可以存取這些資料。基礎設施-服務提供者互動可能涉及交換資料。例如，在智慧建築中，建築管理系統可能會與第三方服務提供者共享即時中央空調系統感測器分散資料，以優化能源使用，或與設施管理人分享資料以規劃維護計畫。這 2 種類型的資料都必須受明確的規範管轄，規定共享和使用條款，以保護當事人的隱私(Farrukh et al., 2024)。

雲端與其他實體（例如基礎設施所有者或服務提供者）之間的資料互動帶來了額外的隱私挑戰。例如，在智慧醫院中，患者監測資料可以儲存在雲端環境中，以允許醫療保健專業人員遠端存取或使用分析服務。雲端服務提供者必須執行嚴格的隱私權政策，以確保敏感的健康資料不會被不當存取或使用(Farrukh et al., 2024)。

當事人應該能夠控制服務提供者如何利用他們的資料。例如，在智慧零售環境中，顧客可能會使用追蹤他們在店內行為的應用程式來提供個人化的購物建議。客戶應該能夠控制他們的資料是否與第三方行銷服務共享或專門用於店內個人化(Farrukh et al., 2024)(參見圖 14)。

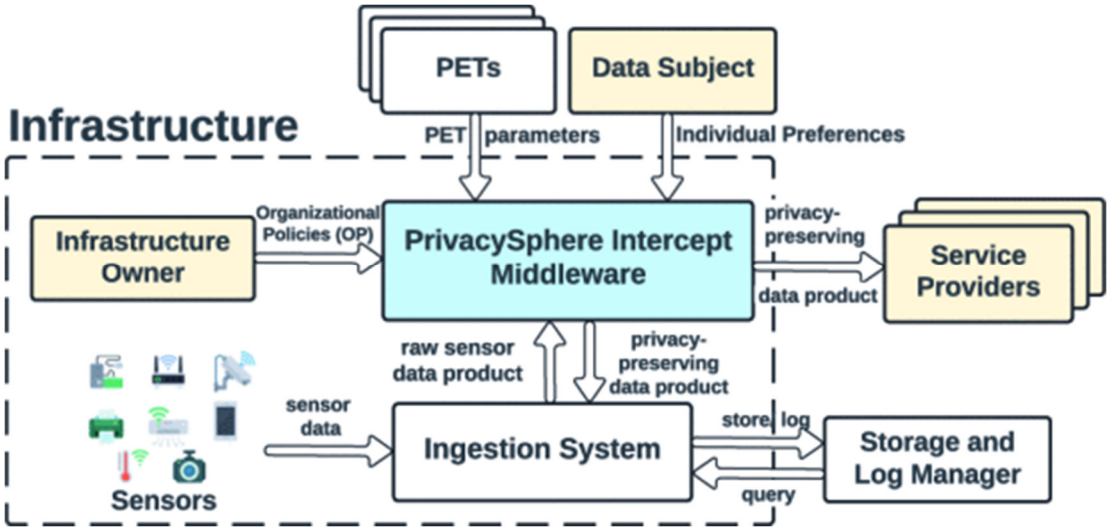


圖 14 智慧空間中的資料共享隱私增強技術

(資料來源： Farrukh et al., 2024)

3. **資料所有權和治理：**上述互動中涉及的資料都受到控制其整個生命週期策略的管轄—從創建或提取到儲存、分析、共享和保留。訂定這些規定須考量於所涉及的資料類型，可以將其分為兩類：（a）原生形式的資料，例如感測器讀數；（b）聚合資料，例如特定區域隨時間的變化。

資料所有權因資料類型而異。對於分散資料，所有權通常屬於一個或多個主體/個人。對於聚合數據，所有權通常轉移到基礎設施所有者。值得注意的是，聚合資料源自單一資料點，一旦聚合，管理聚合資料的策略就由基礎設施所有者控制。在多個實體共享資料所有權的情況下，最終的政策可能需要是各個政策的組合。根據所涉及實體之間的情況和協議，可以使用各種機制來實現這一點，例如：多數規則、最保守的政策或最寬鬆的政策。

4. **案例研究—高齡者照護機構智慧空間：**對高齡者生活設施及其活動的靜態/動態物理基礎設施/空間進行建模，產生靜態資料—即感測器和空間資料，如樓層/房間、其中的感測器，以及例如：居民、工作人員、訪客等人員資料，以及動態資料，即感測器觀測、空間內的人際互動和事件資料。

透過社區參與式研究 (CBPR) 方法進行焦點小組研究模擬

隱私要求，以引出最直接受其影響的利害關係人（例如高齡者和照顧者）的隱私認知和隱私擔憂。隨後，我們可以根據每個利害關係人的需求，將這些隱私要求建模並定義為組織政策和使用者偏好，同時還可以模擬智慧空間技術為高齡者帶來護理服務，例如：跌倒風險檢測、提升護理人員照護效率以及防止自殘行為、為病情複雜的高齡者提供護理可能需要對居民進行監測，這需要居民和護理人員的定位資料以及治療資料（例如，施用的藥物）來分析所述居民的健康狀況。使用焦點小組研究的結果來模擬受試者的個人偏好，例如， 30% 的居民會選擇不分享他們的位置資料。隨著符合規範的資料在系統中流動，因而可以問責和保持透明度，以便進行隱私和實用性的權衡。

透過前面提到的焦點小組研究，首先，收集並解釋高齡者照護社區面臨的需求。例如，在我們之前提到針對行動不便和跌倒風險較高的居民的基於位置的服務中，護理人員定義的效用指標可能會捕捉到居民出現步態/平衡變化的情況。在這種情況下，隱私目標可能是確保護理人員的行為資訊不會被透露給設施管理人員，在這種情況下，設施管理人員被視為監督者（例如，如果護理人員未幫助居民去洗手間的可能性，可以透過使用日誌來評估）。

綜上，智慧空間是新興的重要技術之一，其為社會帶來變革性改善的潛力已被廣泛認可。隱私、信任和責任是其廣泛採用和部署的主要障礙。最近的研究已經開始將隱私增強技術擴展並應用於智慧空間，以滿足其獨特的需求。

肆、 心得及建議

本次會議吸引國際間許多高水平之產官學研機構代表交流，探討如何在信任、隱私及安全之前提下應用智慧科技，確保人們利用人工智慧等創新技術可達成優化人類福祉的願景、實現科技以人為本，綜整與本所建築研究業務職掌較為密切相關，值得參考或借鏡之心得及建議如下：

一、 心得

(一) 可信任的人工智慧科技觀念與技術係近期國際科技發展重要趨勢

因應 2024 年歐洲議會通過全球首部《人工智慧法》，經濟合作暨發展組織(OECD)公布《人工智慧建議書》，倡議可信任的人工智慧（Trustworthy AI）國家政策和國際合作等理念、我國國家科學及技術委員會已於 2024 年 7 月預告制定我國人工智慧基本法(草案)(國家科學及技術委員會, 2024)，許多跨國科技產業並已將可信任的人工智慧科技觀念與技術融入所開發之產品及服務，其共通性目標是促進人工智慧應用之產業發展，管理人工智慧新科技之風險，確保對人類社會產生良性之影響，具體手段是透過可信任的人工智慧、透明可解釋、公平、減少偏見、不歧視、問責等方式實現。

發展可信任的人工智慧亦納入 2024 年 12 月召開之行政院第 12 次全國科學技術會議議題，本所亦將配合納入智慧化居住空間應用人工智慧物聯網科技計畫研究課題，本次會議已蒐集最新國際發展的相關資料將一併作為後續業務推動之參考。

(二) 引進可解釋的人工智慧技術以促進人類對於人工智慧科技的信任

本次研討會探討可解釋的人工智慧(XAI)最新發展趨勢，透過增強機器學習中的深度學習模型「黑盒子」的透明度和可解釋性，使預測結果容易被人類理解，成為人類可解釋的人工智慧模型，確保作成的決策是人類可理解、可預測何時可成功或失敗、知道何時可信任或不可信任、以及何時與為何會犯錯。具體實現方式包括：增加可解釋性、增加透明度、增加可靠性增加穩健性、強化隱私、視覺化等。本次會議已蒐集最新國際發展的相關資料亦將納入相關業務推動之參考。

(三) 應積極發展隱私保護技術因應推動智慧空間面臨的資料共享挑

戰

本次研討會蒐集關於智慧空間保護隱私的新知，發現智慧空間基於資料從感測器通過資訊線路，實現各種需求應用外，智慧空間資料還可以儲存以供將來分析和處理，以實現新的目的，未來具有巨大的發展潛力。然而，相對於從線上互動中收集資訊的社交媒體和網際網路數位平台而言，智慧空間對隱私構成更大的風險，主要原因係智慧空間能夠即時擷取有關人們實體活動的資料，即使是看似無害的運動探測器，也可以產生有關個人的高度精細資訊。基於保護當事人的隱私，對於智慧空間中的雲端、服務提供者、基礎設施所有者、資料當事人彼此之間的資料處理、共享應有更明確的規範管轄，以確保敏感的健康資料不會被不當存取或使用，建議本所後續推動智慧化居住空間應用人工智慧物聯網科技計畫可規劃相關研究課題。

二、 建議

本次會議蒐集資料發現，國際間產官學研各界近年積極發展可信任的人工智慧創新技術，建議本所後續推動「智慧化居住空間應用人工智慧物聯網科技計畫」時，可參考此一趨勢探討以下研究課題：

建議一

辦理可信任的人工智慧技術引進智慧建築研究

主辦機關：內政部建築研究所

協辦機關：無

本次研討會發現國際產官學研界已積極發展可信任的人工智慧創新技術，共通性之目標管理人工智慧新科技之風險，確保對人類社會產生良性之影響，建議本所後續可參考此一國際發展趨勢，將可信任的人工智慧技術引進智慧建築，確保符合促進以人為本、可信任之人工智慧應用發展方向。

建議二

辦理可信任的人工智慧法規國際比較研究

主辦機關：內政部建築研究所

協辦機關：無

本次研討會發現許多跨國科技產業積極因應 2024 年歐洲議會通過全球首部《人工智慧法》，經濟合作暨發展組織(OECD)公布《人工智慧建議書》，倡議可信任的人工智慧（Trustworthy AI）國家政策和國際合作等理念，已將可信任的人工智慧科技觀念與技術融入所開發之產品及服務，我國國家科學及技術委員會亦已於 2024 年 7 月預告制定我國人工智慧基本法(草案)，建議本所後續可參考此一國際、國內發展趨勢，將可信任的人工智慧技術引進智慧建築，確保符合促進以人為本、可信任之人工智慧用，引導產業發展符合國際法規發展趨勢之產品及服務，以減少相關創新產品及服務之出口或進口貿易障礙，發揮我國資通信產業之國際競爭優勢。

附錄一、會議議程

Overview Day 1: Monday, Oct 28, 2024

7:15 AM - 8:30 AM		Registration & Continental Breakfast (provided by conference)			
8:30 AM - 8:45 AM		**Welcome and Opening Remarks** (Steering Committee Chair and Organizing Committee Chairs) (Room: Logan Ballroom)			
08:45 AM – 9:45 AM	**Workshop (LLM CyberSec)** (Room: Deacon)	**Keynote 1** (Room: Logan Ballroom) **Jeannette M. Wing**, *Executive Vice President for Research & Professor of Computer Science, Columbia University* **Title: Trustworthy AI** (Chair: James Joshi, University of Pittsburgh, USA)			
09:45 AM – 10:45 AM	**Chair**: Kristen Moore (CSIRO)	**Keynote 2** (Room: Logan Ballroom) **Deirdre K. Mulligan**, *Director of the Berkeley Center for Law and Technology & Professor, School of Information, UC Berkeley* **Title: New Directions in Tech Governance** (Chair: James Joshi, University of Pittsburgh, USA)			
10:45 AM – 11:00 AM		Break			
11:00 AM – 12:20 AM		**TPS Invited Research/Vision Session 1: Malware Detection, Forensics, and Deep Learning** (Room: Logan Ballroom) Session Chair: Amir Masoumzadeh (SUNY-Albany, US)	**CogMI Invited Research/Vision Session 1: Causal Inference, AI, and Logical Reasoning** (Room: Gaston) Session Chair: Julian Jarrett (Lutron Electronics, US)	**CIC Invited Research/Vision Session 1: AI for Autonomous Systems, Security, and Monitoring** (Room: Bader) Session Chair: Mei-Ling Shyu (UMKC, US)	**Workshop (Inclusive AI)** (Room: Whitman) **Chairs** Hemant Purohit (GMU), Jin-Hee Cho (Virginia Tech), Yoosun Chung (GMU)
12:20 PM – 01:30 PM	12:20 PM – 1:30 PM	**Lunch Break** (provided by conference)			
01:30 PM – 02:30 PM	**Workshop (LLM CyberSec)** (Room: Deacon)	**Keynote 3** (Room: Logan Ballroom) **Ed H. Chi**, *Distinguished Scientist & Research Lead (LLM/LaMDA), Google DeepMind* **Title: The Future of Discovery Assistance** (Chair: Huan Liu, Arizona State University, USA)			
02:30 PM – 04:30 PM	**Chair**: Kristen Moore (CSIRO)	**Panel 1** (Room: Logan Ballroom) **Panel Title: How Will Artificial Intelligence Reshape Scientific Research?** **Panelists**: *Li Yang (National Science Foundation, US), Hemant Purohit (George Mason Univ, US), Ling Liu (Georgia Tech, US), Huan Liu (ASU, US), Ed Chi (Google, US)* **Moderator:** *Paolo Boldi, University of Milan, Italy*			
04:30 PM – 04:45 PM		**Coffee Break**			
04:45 PM – 06:25 PM		**TPS Session 1: Privacy Enhancing Technologies and Cybersecurity Threats** (Room: Logan Ballroom) Session Chair: Lavanya Elluri (TAMU-Central Texas, US)	**CogMI Session 1: AI-enhanced security, and automated monitoring** (Room: Gaston) Session Chair: Julian Jarrett (Lutron Electronics, US)	**CIC Session 1: AI-powered Systems and Applications** (Room: Bader) Session Chair: Latifur Khan (UT Dallas, USA)	**Workshop (Inclusive AI)** (Room: Whitman)
06:30 PM – 08/09:00 PM		**Networking/Reception** (provided by conference)			

(資料來源：研討會議程，2024)

Overview Day 2: Tuesday, Oct 29, 2024

7:15 AM - 8:30 AM		Registration & Continental Breakfast			
8:30 AM - 8:45 AM		**Remarks and Conference Logistics** (Room: Logan Ballroom)			
8:45 AM – 9:45 AM	**Workshop (QUILLS)** (Room: Deacon) **Chairs:** Kaushik P. Seshadreesan (University of Pittsburgh)	**Keynote 4** (Room: Logan Ballroom) **Michael L. Littman**, NSF IIS Division Director & University Professor of Computer Science, Brown University Title: **The National Science Foundation's Role in the Future of AI** (Chair: *Jaideep Vaidya, Rutgers University, USA*)			
9:45 AM – 10:00 AM		Break			
10:00 AM – 12:00 PM		**TPS Session 2: Large Language Models for Privacy and Security** (Room: Logan Ballroom) Session Chair: Indrajit Ray (CSU, USA)	**CogMI Invited Research/Vision Session 2: Systems, Applications, and AI Performance** (Room: Gaston) Session Chair: Keke Chen (UMBC, USA)	**CIC Session 2: AI-driven Systems, Security and Computing** (Room: Badar) Session Chair: Barbara Carminati (University of Insubria, Italy)	**TPS Invited Research/Vision Session 2: AI, Quantum Computing, and Cybersecurity** (Room: Whitman) Session Chair: Amir Masoumzadeh (SUNY-Albany, US)
12:00 PM – 02:00 PM		Lunch Break (provided by conference) and **Panel Session** (Room: Logan Ballroom) **Student Mentoring Panel** **Tiny Keynote: Jaideep Vaidya (Rutgers University, USA)** **Panelists:** Jaideep Vaidya (Rutgers University, USA), Huan Liu (Arizona State University, USA), Indrakshi Ray (Colorado State University, USA), Stacey Truex (Denison University, USA), Peter Kairouz (Google, USA), Ambareen Siraj (SFS/SaTC Program, NSF, USA) **Moderator:** *Wenqi Wei*, Fordham University, USA			
02:00 PM – 03:00 PM	**Workshop (QUILLS)** (Room: Deacon) **Chairs:** Kaushik P. Seshadreesan (University of Pittsburgh)	**Keynote 5** (Room: Logan Ballroom) **Roshan Thapliya**, Corporate Officer, Chief Digital Transformation Officer and General Manager, TDK Corporation, Tokyo, Japan Title: **Transforming Society: TDK's Vision Accelerated by Digital Transformation** (Chair: James Joshi, University of Pittsburgh, USA)			
03:00 PM – 03:15 PM		Break			
03:15 PM – 05:15 PM		**TPS Session 3: Privacy and Security in AI/ML** (Room: Logan Ballroom) Session Chair: Dipankar Dasgupta (University of Memphis, US)	**CogMI Session 2: AI-driven decision-making and human-centered systems** (Room: Gaston) Session Chair: Keke Chen (UMBC, US)	**CogMI Session 3: AI and Language Models for human-centered systems** (Room: Badar) Session Chair: Kim Hemmings-Jarrett (PSU, US)	**TPS Invited Research/Vision Session 3: AI Security, Privacy, and Healthcare** (Room: Whitman) Session Chair: Indrakshi Ray (CSU, US)
06:00 PM – 09:00 PM		**Banquet Dinner** (provided by conference)			

(資料來源：研討會議程，2024)

Overview Day 3: Wednesday, Oct 30, 2024

<table>
<tr><td>07:15 AM - 08:00 AM</td><td colspan="5">Registration & Continental Breakfast (provided by conference)</td></tr>
<tr><td>08:00 AM – 10:00 AM</td><td rowspan="3">Workshop (WAAM)
(Room: Deacon)
Chairs: Phil LaPlante (NIST) and Rick Kuhn (NIST)</td><td rowspan="3">Workshop (SR-CIST)
(Room: Gaston)
Chairs: Mai Abdelhakim (University of Pittsburgh), Mohan Baruwal Chhetri (CSIRO), Peilin He (University of Pittsburgh)</td><td>TPS Session 4: Malware and Threat Detection
(Room: Logan Ballroom)
Session Chair: Pooria Madani (OntarioTechU, Canada)</td><td>CogMI Session 4: AI-driven modeling and analysis for behavioral and network systems
(Room: Badar)
Session Chair: Yanzhao Wu (FIU, USA)</td><td>Workshop (EIC)
(Room: Whitman)
Chairs: Julian Jarrett (Lutron Electronics, US)</td></tr>
<tr><td>10:00 AM – 10:15 AM</td><td colspan="3">Break</td></tr>
<tr><td>10:15 AM – 12:15 PM</td><td colspan="3">Panel 2 (Room: Logan Ballroom)
Panel Title: Driving Innovations in Emerging Technologies: R&D Priorities and Funding Landscape
Panelists: Jennifer Roberts (Director, Resilient Systems, ARPA-H, USA), Heidi Sofia (Deputy Director, NCBI-NIH, USA), Elham Tabassi (Associate Director of Emerging Technology, NIST, USA), Cliff Wang, (Program Director, SaTC Program, NSF, USA), Dr. Craig Schlenoff (Director, NITRD + AD of Networking and IT at White House OSTP)
Moderator: James Joshi, University of Pittsburgh, USA and Huan Liu, Arizona State University, USA</td></tr>
<tr><td>12:15 PM – 01:15 PM</td><td colspan="5">Lunch Break (provided by conference)</td></tr>
<tr><td>01:15 PM – 03:00 PM</td><td rowspan="2">Workshop (WAAM)
(Room: Deacon)
Chairs: Phil LaPlante (NIST) and Rick Kuhn (NIST)</td><td rowspan="2">Workshop (SR-CIST)
(Room: Gaston)</td><td colspan="2">TPS Session 5: Access Control and Security Models
(Room: Logan Ballroom)
Session Chair: Wenqi Wei (Fordham University, USA)</td><td>Tutorial: Tutorial: Computational Scientific Discovery in Cognitive Science
(Room: Whitman)
Session Chair: Fernand Gobet</td></tr>
<tr><td>3:00 pm -5:00 pm</td><td colspan="3">Panel 3 (Room: Logan Ballroom)
Panel Title: AI vs AI: The Inevitable Next Cyber Frontier
Panelists: Karl Aberer, (EPFL, Switzerland), Elena Ferrari (University of Insubria, Italy), Anupam Joshi (UMBC, USA), Calton Pu (Georgia Tech, USA), Peter Kairouz (Google, USA)
Moderator: James Joshi, University of Pittsburgh, USA</td></tr>
<tr><td>05:00 PM – 05:15 PM</td><td colspan="5">Closing Remarks
(Room: Logan Ballroom)</td></tr>
</table>

(資料來源：研討會議程，2024)

參考文獻

Agency, D. A. R. P. (2018). *XAI: Explainable Artificial Intelligence*. Defense Advanced Research Projects Agency. https://www.darpa.mil/research/programs/explainable-artificial-intelligence

Farrukh, H., Lahjouji, N., Mehrotra, S., Nawab, F., Rousseau, J., Sharma, S., Venkatasubramanian, N., & Yus, R. (2024, 28-31 Oct. 2024). PrivacySphere: Privacy-Preserving Smart Spaces. 2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA),

IEEE. (2024). *The Sixth IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications Conference Program*. IEEE TPS. https://www.sis.pitt.edu/lersais/conference/tps/2024/program.html

Olusegun, R., & Yang, B. (2024, 28-31 Oct. 2024). Improved Ethereum Fraud Detection Mechanism with Explainable Tabular Transformer Model. 2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA),

Trustworthy AI. (2024). NVIDIA Corporation. https://www.nvidia.com/en-us/ai-data-science/trustworthy-ai/

Wing, J. M. (2024). *Trustworthy AI*. IEEE. https://www.sis.pitt.edu/lersais/conference/common/resources/ieee-tps-wing.pdf

國家科學及技術委員會. (2024). *人工智慧基本法草案總說明及條文*. Retrieved from https://join.gov.tw/policies/detail/4c714d85-ab9f-4b17-8335-f13b31148dc4