信賴AI：將人工智慧融入陸軍專業知識

TRUSTING AI: INTEGRATING ARTIFICIAL INTELLIGENCE INTO THE ARMY'S PROFESSIONAL EXPERT KNOWLEDGE

作者：C. Anthony Pfaff C·安東尼·普法夫, Christopher J. Lowrance 克里斯多福·J·勞倫斯, Bre M. Washburn 布雷·M·沃什伯恩, Brett A. Carey 布雷特A凱里
譯者：蔡婷媜
出版商：USAWC Press 美國陸軍戰爭學院2023年2月8日發行
出處網址：https://press.armywarcollege.edu/monographs/959，因本刊篇幅有限，僅節錄「信賴AI：將人工智慧融入陸軍的專業知識」之「摘要」部分

提要

一、在過往的軍備競賽中，科技專業主要應用於國防部門，但在AI軍備競賽中，科技專業卻主要應用於工業界和學術界。
二、有效利用AI不應倚賴少數專家的努力，幾乎所有官兵都必須具備一定程度的AI和數據素養。
三、影響未來戰備的關鍵因素之一，就是軍隊成員整體的AI素養。

關鍵字：人工智慧（AI）、致命目標處理、人才管理、倫理

Executive Summary 摘要

Introduction 前言

Integrating artificially intelligent technologies for military purposes poses a special challenge. In previous arms races, such as the race to atomic bomb technology during World War II, expertise resided within the Department of Defense. But in the artificial intelligence (AI) arms race, expertise dwells mostly within industry and academia. Effective employment of AI technology cannot be relegated to a few specialists. Not everyone needs to know how to fly a plane to have an effective air force, but nearly all members of the military at every level will have to develop some

level of AI and data literacy if the US military is to realize the full potential of AI technologies. Thus, a critical component of future readiness will be the *AI literacy* of the force.

將AI技術應用於軍事領域，面臨特殊挑戰。在過往的軍備競賽中，例如第二次世界大戰期間各國競相研究原子彈之時，科技專業主要應用於國防部門，但在AI軍備競賽中，科技專業卻主要應用於工業界和學術界。有效利用AI不應倚賴少數專家的努力，在軍中不必所有人都得學會開飛機，才能讓空軍發揮有效戰力，但若要讓美軍充分發揮AI科技潛能，幾乎所有官兵都必須具備一定程度的AI和數據素養。因此，影響未來戰備的關鍵因素之一，就是軍隊成員整體的AI素養

In this context, *AI literacy* means more than simply understanding how to use, design, and engineer AI- and data-enabled systems. Rather, data, algorithms, and the systems they support interact in complex ways that change even familiar processes, such as targeting, into something much more complicated and unfamiliar. Making matters more difficult, from a professional perspective, mastering new technology requires adequately understanding how the technology works and how its application affects organizational, ethical, and political concerns for the military and the US government, its international partners, and American society.

AI素養不僅是理解如何使用、設計和開發AI系統，更要瞭解數據與演算法的複雜互動，這種互動會使得像目標選定這樣的熟悉過程變得更為複雜和陌生。從專業角度來看，掌握新技術不僅要瞭解技術的運作方式，還需瞭解其應用對美軍、政府、國際夥伴和社會在組織、道德及政治方面的影響。

Challenge of Integrating AI and Data Technologies 整合AI和數據科技的挑戰

Often, the problems associated with employing AI, especially in a lethal targeting process, arise from the perceived trade-off between taking advantage of the machine's speed and maintaining meaningful human control. To the extent humans give up control, they give up responsibility. To the extent they give up responsibility, they

undermine accountability, and undermining accountability creates reasons to distrust the machine and the humans who employ it.

使用AI（特別是在致命目標處理的過程中）所面臨的問題，主要來自於如何利用電腦運算速度與保持有效人類控制之間取得平衡。當人類放棄對機器的控制時，也意味著放棄了相應的責任。責任的放棄會削弱問責機制，進而引發對機器及操作人員的不信任

Thus, the central question is: On what basis commanders, staffs, and operators can trust AI technologies and the systems they enable? *Trust*, as used here, entails multiple conditions. First, one expects the system to be effective—that is, able to produce the intended effect at least as well as, if not better, than human-only systems. Moreover, as a report by the UN Institute for Disarmament Research pointed out, AI-enabled systems must be predictable and understandable, where *predictability* entails the system consistently fulfilling its intended purpose and *understandability* entails the machine acting for intelligible reasons. In a professional context, however, professionals trusting the technology is not enough. Clients must further trust professionals to use AI in their interests and in a way that reflects their values and other ethical commitments.

因此，核心問題在於指揮官、工作人員和操作人員如何信任AI及其所驅動的系統？在此，信任涉及多個條件。首先，系統必須有效，即能實現預期效果，至少與人工系統相當，甚至更好。此外，正如聯合國裁軍研究所(UN Institute for Disarmament Research)的報告指出，AI驅動的系統必須具備可預測性和可理解性：可預測性意味著系統能穩定達成預期目的，可理解性則代表著系統的行為必須有合理的解釋。然而，在專業背景下，僅僅信任技術是不夠的，客戶還需相信專業人員能夠以符合客戶利益、價值觀和其他倫理標準的方式來使用AI。

Given professionals must ensure these conditions are being met, the question can be reframed as one of professional expertise, which includes educating, training, and certifying the profession's members to use the technology and evolving the profession's institutions to ensure the technology's use is effective and ethical. Knowing how the acquisition of new technologies impacts the profession's

organizational culture and other stakeholders is also critical to meeting the conditions for trust.

考慮到專業人員必須確保這些條件得到滿足，這個挑戰可以被重新定義為專業技能的問題，包括對專業人員進行教育、培訓和認證，確保他們能有效使用技術，並發展相關機構以確保技術的使用既有效又符合倫理。同時，瞭解新技術導入對專業組織文化和其他利益相關者的影響，也是建立信任的關鍵。

To understand how the military can meet these conditions, this project examined Project Ridgway, an effort by the XVIII Airborne Corps to integrate currently available AI, data, and imagery to be *AI-ready*. Project Ridgway is a bottom-up effort wherein the corps engages the private sector directly to take advantage of commercially available data and algorithms to support targeting in the deep fight. This report found trusting an AI-driven system in the professional military context requires: first, understanding the context in which AI is applied; second, understanding what one is trusting AI to do; and, finally, understanding how to interact with the AI-driven system, including how the system receives input and provides output. Meeting these conditions enables one to audit and ensure the authenticity of the data, which is critical for trust.

為了瞭解美軍如何滿足這些條件，本項目研究了第十八空降軍團（XVIII Airborne Corps）推動的雷吉威計畫(Project Ridgway)，該計畫旨在整合現有的AI、數據和影像，以準備好迎接AI技術。該計畫直接直接與民間部門合作，利用商業上可得的數據和演算法來支持縱深作戰中的目標處理。本報告發現，要在專業軍事背景下信任AI驅動的系統，需要做到以下幾點：首先，瞭解AI應用的背景；其次，明白AI能完成的任務；最後，瞭解如何與AI系統互動，包括系統如何接收數據輸入和提供數據輸出。滿足這些條件可以幫助審核並確保數據的真實性，這對建立信任非常關鍵。

Targeting: Why Speed Matters 目標處理：速度的重要性

In this context, targeting is a four-phase process that comprises deciding, detecting, delivering, and assessing. As currently employed in the XVIII Airborne Corps's targeting process, AI primarily applies to the detect phase, wherein sensors provide

input (generally, imagery) to an algorithm, which relies on curated data to predict whether designated objects are present and, if so, their location. In the future, AI may also impact other parts of the cycle, including asset allocation and the assessment of battle damage and effects.

在此過程中，目標處理包括：選擇、偵蒐、打擊和評估四個階段。在第十八空降軍團的目標處理過程中，AI主要應用於偵測階段，此時感測器會提供影像等資料給演算法，演算法依據精確數據來預測目標是否存在以及其所在位置。未來，AI可能還會影響這個過程中的其他部分，包括資源分配和戰損評估。

Targeting is iterative and interactive. The process iterates by learning during each cycle and the cycles within the larger targeting cycle to improve the AI-driven machine's performance. Targeting involves interacting with an adversary engaged in the same cycle. If an adversary is similarly equipped, the one who gets through the cycle faster wins. Since machines are faster than humans, targeting disposes humans to rely on the machine, even if doing so means taking extra risks. Speed matters.

目標處理是一個反覆進行且互動的過程，AI透過每個循環進行學習，從而改進系統性能。這個過程還涉及與敵方互動。如果敵方也擁有類似的裝備，那麼能夠更快速完成整個過程的一方將佔據優勢。由於電腦運算的速度快於人類，人類會依賴機器，儘管這可能帶來額外的風險，卻也突顯了速度的重要性。

Developing Trustworthy AI 建立可信的AI

Given this reliance on machines, one must ask oneself what one is trusting an AI-driven system to do. From a practical and an ethical perspective, lethal targeting requires one to balance the imperatives of defeating an enemy, avoiding noncombatant casualties, and protecting the force. Balancing these imperatives involves answering questions about risk. Put simply, lethal operations expose friendly combatants and noncombatants to risk, avoiding noncombatant casualties exposes friendly combatants or the operation to risk, and protecting the force exposes the operation or noncombatants to risk. Reducing risk to any one imperative thus places risk on the other two. Employing AI can reduce risk to all three. By making fires faster and more

precise, AI makes defeating the enemy more likely while reducing the chance of friendly and collateral harm.

鑑於對機器的依賴，人類需要考慮對AI驅動系統的信任程度。從實際操作和倫理角度來看，目標處理需要在擊敗敵人、避免平民傷亡和保護部隊之間取得平衡，而這種平衡涉及風險的考量。簡單來說，攻擊行動會使友軍和非戰鬥人員暴露於風險之中，避免平民傷亡則可能使友軍或作戰面臨風險，而保護部隊則可能使作戰或平民受到威脅。因此，降低某一方面的風險，會將風險轉移到其他兩方面。運用AI可以在三方面的風險中取得平衡，因為AI能使火力更快發揚且更精準，提高擊敗敵人的可能性，同時減少對友軍和附帶損害的風險。

In a human-only process, trust depends on understanding the capabilities of one's soldiers and the weapons they carry, ensuring they understand and will comply with the law of armed conflict, and being able to hold them accountable when they do not comply. In an AI-driven process, trust depends on knowing how to curate and monitor data, assess and optimize algorithm performance, and secure the system from external manipulation. Artificial intelligence is a process of algorithms operating on data in a specific context. Trusting this process depends, at least in part, on trusting the components. To ensure trust, the data must be auditable and the algorithm adequately understood in its operational context.

在傳統由人類操作的過程中，信任來自於對士兵及其武器性能的瞭解，確保他們理解並遵守武裝衝突法，並在違反時可追究責任，而在由AI驅動的過程中，信任則取決於如何整理和監控數據、評估和優化演算法性能，並防止系統受到外部操控。AI本質上是針對特定情境運行數據的算法過程，信任這個過程，部分來自於對其組成要素的信任。為了建立信任，必須對數據稽核，並且充分瞭解演算法在實際操作中的運作方式。

Barriers to Trusting AI 信任AI面臨的挑戰

Barriers to trusting AI include uncertainty about how to warrant confidence one has curated data correctly, trained and retrained the data and algorithms to be accurate and precise, and protected the system against spoofing or other unwanted manipulation.

信任AI面臨的挑戰包括如何確保正確蒐集資料、妥善訓練和重新調整演算法，以保證其準確性與精確性，以及如何防止系統免受到欺騙或其他不良操作。

Data Challenges AI數據的挑戰

In the context of AI, multiple other factors that are functions of the data, the algorithm, and external interference impact trust. Algorithms are often only as good as the data on which they are trained. Through training, the machine learns to differentiate items of interest from everything else. Collecting accurate, complete, consistent, and timely data sets for the system to train on is extremely difficult and sensitive to the environment in which the targeting will occur. Keeping data sets updated is critical work that must be ongoing. The challenge is that it is extremely difficult to know when one has collected all the necessary data to optimize the system's performance. As a result, the system will make mistakes when the inputs do not closely resemble the data on which the system was trained.

在AI的背景下，影響信任的因素包括數據、演算法和外部干擾。演算法的效果往往取決於訓練所使用的資料，透過訓練，機器學會區分所需的項目和其他內容。蒐集準確、完整、一致且即時的資料集供系統訓練非常困難，且對目標選定的環境高度敏感，保持資料集的持續更新是關鍵工作，必須不斷進行。挑戰在於，難以確定是否已蒐集到所有必要的數據以優化系統性能，因此，當輸入資料與系統訓練所使用的資料差異較大時，系統容易出錯。

Performance Issues 系統性能問題

Performance issues usually come in the form of misclassifications, false positives, and false negatives. For example, when the inputs to AI classifiers do not resemble the training data, prediction mistakes are more likely. Prediction mistakes can occur when a classifier is trained using only images of targets taken during the summer months and then presented images of partially concealed targets taken during the winter. If a classifier that was trained using only images of tanks operating in the desert is asked to classify an image of the tank partially covered in snow, then the classifier will likely

make a mistake. To counter such mistakes, continuously searching for and collecting new, informative data examples as they become available and using them to retrain and update the classifier as needed— especially relative to the environment in which one is operating—is important. Often, retraining and updating the classifier means collecting new data while the system is operating and then identifying which samples can help to improve the AI model's performance.

系統性能問題通常包含分類錯誤、誤判和漏判。例如，當AI分類系統的輸入與訓練數據不匹配時，錯誤的可能性會增加。如果分類系統僅用夏季拍攝的目標影像進行訓練，然後再接收到冬季拍攝的部分隱藏影像，就可能會出錯。同樣地，如果分類系統只用沙漠中行駛的戰車影像進行訓練，當遇到部分被雪覆蓋的戰車影像時，也可能會出錯。為了應對這些問題，特別是在當前操作環境下，持續收集新數據樣本並用它們來重新訓練和更新分類系統是非常重要的。通常，重新訓練和更新分類系統意味著在系統運行時蒐集新數據，並確定哪些樣本能幫助改善AI模型的性能。

In short, classifiers can make mistakes given the state of the art and the difficulty of collecting comprehensive data sets. Artificial intelligence can be a "black box" because how it arrives at an output is not always discernible to humans, either due to the complexity of the algorithm or because the AI's output depends on the strength of the connections in the network. Nevertheless, commanders and operators should understand the limitations of AI and observe AI-enabled systems' performance in similar conditions, thereby enabling the commanders and operators to decide, based on risk calculations, how much control to provide to the AI in targeting operations.

簡而言之，由於蒐集最新且完整資料集的困難，分類系統可能會出錯。AI有時被比喻為「黑盒子」，這可能是由於演算法的複雜性，或因為AI的分析結果依賴於網路強度，導致其分析過程不透明，儘管如此，指揮官和操作人員應該瞭解AI的限制，並觀察AI系統在類似條件下的表現，從而根據風險評估決定在目標處理中給予AI多少控制權。

Other issues include the enemy actively attempting to thwart AI by poisoning data sets or changing the enemy's asset signatures. For instance, a poisoning attack can

undermine a machine learning model during the training phase by altering the model's training data. Adversarial poisoning attacks could train a target identification model to ignore one class of object entirely, enabling a high-value target to hide in plain sight. To conduct an input attack, an adversary injects noise into a model's input to produce an incorrect output.

其他問題包括敵方積極嘗試破壞資料集或改變軍事裝備的特徵來干擾AI的應用。例如，敵方可在訓練階段篡改模型的數據，從而影響機器學習模型的準確性。對抗性攻擊可能會使目標識別模型完全忽略某一類物體，讓高價值目標在公開環境中隱形。進行輸入攻擊時，敵方會在模型的數據中加入干擾訊號，導致模型產生錯誤的分析結果。

In one example, a small piece of tape placed on a stop sign caused self-driving cars to misidentify the sign as a 60-mile-per-hour speed marker. Similarly, an adversary could visually modify a tank so a machine-learning model assesses the tank as a truck. Moreover, doing so would not be difficult. Small pixel changes, invisible to the human eye, have caused classification algorithms to misidentify images of pandas as monkeys. Both types of attacks, input and poisoning, can undermine the perceived effectiveness of fielded models and degrade trust. More to the point, one should expect AI-driven systems to be under constant attack, requiring users to find ways to detect the effects of these types of attacks.

舉例來說，將一小段膠帶貼在停止標誌上，會使自駕車將其識別為每小時60英里的速限標誌。同樣，敵人也可以通過視覺修改，使戰車被機器學習模型誤認為卡車，而這樣的修改並不困難。即使是對人眼不可察覺的微小像素變化，也可能使分類演算法將熊貓的圖像辨識為猴子。這兩種攻擊方式，包括輸入攻擊和數據攻擊，都可能削弱實際應用模型的效能，並損害使用者對系統的信任。更具體而言，應預期AI驅動的系統將不斷受到攻擊，因此使用者需要找到方法來檢測這些攻擊的影響。

Taken together, the sensitivity of the data sets, the complexity of the algorithms, and the potential for undetected sabotage give rise to an accountability gap. Accountability depends on intent and action. But harm, including violations of the law

of armed conflict, may occur, despite commanders, staffs, and operators involved in an AI-driven system acting with good intentions and despite the system, with the exception of spoofing, working according to specification. Commanders and staffs may understand the system well but suffer from automation bias, especially with systems that are normally reliable, thus increasing the probability of unaccountable harm.

綜合來看，資料集的敏感性、演算法的複雜性，以及潛在未被察覺的破壞行為，會導致責任漏洞。責任依賴於指揮官、幕僚和操作人員的意圖和行動，即使他們在操作AI系統時有良好意圖，且系統在未遭受欺騙的情況下正常運行，仍可能會發生包括違反武裝衝突法在內的問題。即使指揮官和幕僚熟悉系統，仍可能受到系統自動化的影響，特別是在系統表現可靠時，這會增加無法究責的風險。

Importantly, AI performance is not all about speed. In fact, the machine provides better output when humans interact with it, even during operations. So, the idea that developing and employing AI involves a trade-off between speed and meaningful human control is a false dilemma. The question, then, is how do humans know when and where to interact with a system and provide control while optimizing the system's performance?

重要的是，AI的性能不僅僅取決於速度，事實上，在系統分析過程中，當人類與其互動時，機器能提供更好的分析結果。因此，認為開發和使用AI必須在速度與人為控制之間做出取捨的觀點是錯誤的。關鍵問題在於，人類應該判斷何時以及在哪裡與系統互動，並適時控制，以優化系統性能。

Developing Reliable and Capable Systems 建立可靠且功能強大的系統

Trust and risk are central concerns in developing reliable, capable systems. Commanders need a reliable way to know when AI can be trusted and when to execute some stages of the targeting process with less supervision for the benefit of speed but at the cost of more risk. The systems studied here rely on neural networks that provide a measure of probabilistic confidence in each target classification. Commanders can exploit these neural networks during targeting to make informed decisions about the level of human supervision required, especially when the probabilistic confidence is

combined with other information, such as the commander's risk tolerance in the context of the mission.

在開發可靠且功能強大的系統時，信任與風險是主要考量因素。指揮官需要可靠的方法來判斷何時可以信任AI，何時可以在某些目標處理階段中減少監督以提高效率，但這樣做也會增加風險。本文研究的系統依賴神經網路，這些網路能夠評估每個目標分類的可信度。當這些評估結果與其他資訊（如指揮官的任務風險容忍度）結合時，指揮官可以根據這些評估結果來決定對AI系統的監督程度

The commander's risk tolerance can aid in the process of deciding how to handle targets that have been classified by AI. Determining the acceptable level of risk for the operation of the AI is the commander's decision. Therefore, the commander should be given the flexibility and option to assume more risk at times if, based on his or her best judgment, the conditions merit the risk.

指揮官的風險承受度有助於決定如何處理由AI分類的目標，確定可接受的風險等級是指揮官的責任，因此，應賦予指揮官彈性和選擇權，使其能在條件允許時，根據自身判斷承擔更高的風險。

For instance, a commander may be risk averse when providing fire support in a counterinsurgency mission or in a dense urban environment with many civilians nearby. But a commander may be more risk tolerant if facing a high-intensity battle in mostly open terrain or performing final protective fires when friendly forces may be overrun by the enemy. To capture risk tolerance, commanders could be given a rheostat-like device that they can tune and use to convey their risk tolerance directly to the system. One can also run more than one AI model at a time; this approach, which is commonly referred to as an *ensemble*, can be used to increase confidence that inferences drawn are true or to detect errors.

指揮官在執行鎮壓叛亂任務或在人口密集的城市環境中提供火力支援時，可能會傾向於降低風險，但在面對高強度戰鬥的開闊地形，或友軍可能遭到敵軍壓制時，指揮官可能會更願意承擔較高的風險。為了評估風險承受度，指揮官可以使用類似可變電阻的裝置，直接向系統傳達風險容忍度。此外，同時運作多個

AI模型，採用集成學習的方法，有助於提高推斷的準確性，並增強檢測錯誤的能力。

Decision-making Logic within the Control System 控制系統中的決策邏輯

The rheostat would interact with the system through a fuzzy-logic controller that would account for commander risk tolerance and machine certainty to determine the optimal setting for human control. Fuzzy logic can help balance machine confidence and a commander's risk tolerance. Fuzzy logic's purpose is to avoid hard coding single-value thresholds, which specify where certain values belong to certain categories (for example, 34 is moderate, and 32 is low). Rather, the idea is to program transitions between the input classes of low, moderate, and high.

指揮官能透過模糊邏輯控制器、可變電阻與系統互動，設定風險容忍度和機器的可性度，決定人為控制的最佳方案。模糊邏輯有助於平衡機器的可信度與指揮官的風險容忍度，其目的是避免使用固定數值的分類方法（例如，將34歸為中等，32歸為低等）。相反地，它通過設置低、中、高等類別之間的過渡區間，使得處理不確定性時更加靈活。

Programming transitions between input classes makes fuzzy logic more tolerant of uncertainty when measuring and quantifying the inputs into linguistic sets. The regions where the moderate set overlaps with either the low or high set are the ranges where the input would be classified as belonging to multiple sets with partial membership in each, such as 80 percent high and 20 percent moderate. For instance, one could program a freezer thermostat's controller to alert one to intervene to lower the temperature.

將輸入分類的過程程式化，使模糊邏輯在處理不確定性時更加靈活。當「中等」類別與「低等」或「高等」類別重疊時，輸入可能同時屬於多個類別，每個類別都有一定的屬性程度，例如，某個數據輸入可能有80%屬於「高等」，20%屬於「中等」。這一原理可以應用於冷凍櫃的溫度控制器，當溫度接近設定範圍的邊界時，控制器會根據輸入數據的屬性程度發出警示，提醒使用者進行調整。

Given two variables (risk and certainty) and three settings (low, medium, and high), a rule base of nine recommended settings for human oversight would logically exist. The rule base would be programmed into the controller's memory using a series of if/then statements and obey the following logic: "If AI's Classification Confidence is low and Commander's Risk Tolerance is low, then human involvement is maximum. If AI's Classification Confidence is high and Commander's Risk Tolerance is high, then human involvement is minimum." Assuming two inputs with three categories each (low, moderate, and high), the complete set of nine rules can be derived by the two-dimensional rule base, expressed by machine-generated probabilities and the commander's risk tolerance.

假設有兩個變數（風險和確定性），以及三個設定（低、中、高），則會有九條建議的設定規則。這些規則將以if/then的形式存於控制器中，並按照以下邏輯執行：「如果AI的分類信心和指揮官的風險容忍度都低，則需要最大程度的人類介入；如果AI的分類信心和指揮官的風險容忍度都高，則只需最少的人類介入。」這九條規則是根據機器的信心度和指揮官的風險容忍度推導出來的。

Risk Profiles and Adaptive Teaming Based on Fuzzy-logic Controllers
利用模糊邏輯控制器進行風險管理和團隊協作

What does this rule base mean in practice? The controller's decision for maximum involvement implies a human-driven targeting process in which humans lead each step. Using a human-driven targeting process does not preclude AI from assisting in these steps. In other words, AI can augment any step, but a human must explicitly verify the output before the target proceeds. On the opposite extreme, minimum involvement translates into AI automating all steps, except for the final validation and authorization process, wherein a leader in the fires cell would review the targeting information and recommendations before giving the order to proceed with a fire mission. The moderate oversight process flow is more nuanced and similar to the minimum oversight process flow, except the classification confidence of the AI algorithm and the risk assessment from the integration stage must meet stringent thresholds. If a threshold is not met in either case, then a human must inspect the output generated by the AI algorithm.

這些邏輯規則在實務應用中的意義如下：

●人為最大參與：當邏輯控制器參數將人為控制設定為最高，意味著整個目標處理過程由人為主導。雖然 AI 可以協助各個步驟，但每一步的結果仍需由人類確認後才能進行下一步。

●人為最小參與：當邏輯控制器參數將人為控制設定為最低，AI 將自動完成所有步驟，僅在最後進行確認和授權。在此階段，火協小組指揮官會審查目標資料和建議，並在確認無誤後下達射擊命令。

●人為中等參與：中等參與的過程介於最大參與和最小參與之間。AI 的分類信心和風險評估必須達到設定的標準。如果未達到這些標準，則需要人員檢查 AI 生成的結果。

Human Development人才培育

The technical component shows soldiers will have to develop varying degrees of AI and data literacy. For this to happen, the US Army must identify what this literacy entails and how to certify it. Although identifying the varying degrees of AI and data literacy falls under the technical component, determining how to recruit, certify, and manage knowledgeable personnel will become a critical professional task. To remedy the lack of personnel with AI and data-science education and skills, the Army has implemented plans to educate selected personnel at the leader, analyst and engineer, and technician levels. Although necessary, these plans may not be adequate to provide the range of skilled personnel required to proliferate capabilities at the corps level Army-wide, especially in the short term.

技術層面顯示，士兵需具備不同程度的AI和數據素養。美陸軍必須明確界定這些素養的具體內容及認證方式，雖然辨識不同程度的AI和數據素養屬於技術範疇，但招募、認證和管理具備相關知識的人員也非常重要。為了填補AI與數據科學領域的人才缺口，陸軍已經制定了針對領導者、分析師、工程師和技術人員的教育計畫，然而，短期內這些計畫可能無法滿足整體軍隊，尤其是軍團層級的需求。

Part of the reason the Army's existing plans may be inadequate is the Army needs soldiers with the right data and AI skills and leaders who know how to employ data

and AI skills effectively. Thus, the Army should also consider integrating AI and data literacy into commissioning and other entry-level education and training.

陸軍現有的計畫可能不足，原因之一是需要具備適當數據分析和AI技能的士兵，以及能夠有效運用這些技能的領導者。因此，陸軍應考慮將AI和數據素養納入任官及其他入門教育和訓練中。

Further complicating matters, the Army's ability to manage personnel who are skilled in science, technology, engineering, and mathematics in general, much less those with AI and data-related skills, is limited. Indeed, without a more efficient management system, optimizing the assignment of personnel trained by the Army's new educational programs may not be possible, especially at the operational level. Effectively assigning newly trained personnel is critical to taking advantage of new, often commercially available technologies so the Army remains agile relative to its adversaries. Optimizing the Army's talent management will require the service to revise how it identifies educational requirements, aligns talent with operational needs, and tracks talent so personnel are available where they are most needed.

陸軍在管理具備科學、技術、工程和數學技能的人員等方面面臨挑戰，尤其是那些具備AI和數據技能的人。若缺乏有效的管理系統，尤其是在作戰層面，將難以將新教育計畫所培訓的人員做出最佳分配。為了充分利用新興技術並保持靈活性，陸軍必須有效地分配新訓練的人員。因此，陸軍需要重新檢討如何確定教育需求、將人才與作戰需求匹配，以及追蹤人才分佈，確保人員在最需要的地方得到妥善部署。

This study recommends the Army create new skill identifiers to improve the tracking of AI- and data-related expertise, consider establishing a technology corps that would be managed much like the logistics corps to provide expert knowledge where and when it is needed most, and code certain positions for more than one skill to increase assignment flexibility.

本研究建議陸軍採取以下措施以提升對AI和數據相關專業知識的管理：首先，設立新的專長代碼，以更有效地追蹤這些技能；其次，考慮成立一個類似後

勤部隊的技術單位，以利在關鍵時刻提供專業支援；最後，將某些職位設標示為具備多種技能，以提升人員分配的靈活性。

Ethical 倫理規範

From an ethical perspective, targeting requires preventing, or at least mitigating, potential harm to noncombatants as well as friendly forces. Given the potential for friendly and noncombatant casualties, especially in large-scale combat operations, professionals will have to ensure application of the technology represents acceptable risk to protected persons, infrastructure, and other material assets. Artificial intelligence also raises questions of accountability. Having machines play a larger role in decision making may result in bad outcomes, even if both the humans and the machines perform their duties correctly. Understanding how to deal with such outcomes will be critical to applying AI. Here, we might measure success in terms of whether AI-enabled outcomes represent less harm than human-only processes. To meet the requirements for ethical targeting, commanders must ensure staffs and operators are capable of curating and training data and that they do so at appropriate intervals to ensure the system performs as well as a human-only system. Staffs and operators must also develop familiarity with systems to the point that the staffs and operators can explain outcomes intelligibly. Introducing an interface, like the fuzzy-logic controller discussed above, would facilitate meeting the requirements for ethical targeting and allow commanders to take greater advantage of machine speeds without losing the kind of control that might give rise to ethical failure. The interface addresses accountability by making commanders accountable for the accuracy of their risk assessments and ensuring data is properly curated for the context in which commanders employ the algorithm. The interface also addresses automation bias because it provides humans a way to know when the machine itself is, in a sense, uncertain about its output. Whether these measures are good enough depends on how well the system balances the ethical imperatives discussed earlier in comparison to a human-only process. Balancing imperatives is ultimately the responsibility of the humans involved in the targeting process.

從倫理角度來看，目標選定需要防止或至少減輕對非戰鬥人員和友軍的潛在危害。特別是在大規模作戰中，友軍和非戰鬥人員可能會面臨傷亡，因此，專家必須確保科技應用對受保護者、基礎設施和其他資產的風險是可接受的。

AI也引發了責任追究的問題。即使人類和機器都正確執行其職責，機器在決策中的重要角色仍可能導致不良結果。因此，瞭解如何處理這些結果對於AI的應用至關重要。我們可以通過比較AI技術與僅由人類進行決策的結果，來衡量AI是否減少了損害。

為符合倫理要求，指揮官必須確保其幕僚和操作人員能夠選擇和訓練數據，並定期進行此項作業，以確保系統的表現與人類操作系統相當。幕僚和操作人員還需熟悉系統，能夠清楚解釋結果。引入前述的模糊邏輯控制器等介面，將有助於滿足倫理要求，同時允許指揮官充分利用電腦的運算速度，保持必要的控制權，從而降低倫理失誤的風險。

該介面解決了責任問題，使指揮官對風險評估的準確性負責，並確保數據在使用演算法時得到妥善管理，它還解決了系統自動化的問題，讓人類能夠察覺機器對結果的不確定性。這些措施是否足夠，取決於系統是否能在倫理要素與純人為操作的表現中取得平衡。最終，平衡這些要素的責任在於參與目標選定的人員

One can further improve the system's ability to avoid collateral harms— and thus perform ethically—by training data to identify legitimate targets and illegitimate targets (such as hospitals and schools). For example, if the machine could produce a result such as "80 percent tank; 10 percent school bus," the machine could alert commanders and staff that even though the target probability was within the commanders' risk tolerance, they may have additional reasons for scrutiny. Building data sets that can account for legitimate and illegitimate targets may be beyond the resources available in any given system. In these cases, commanders should account for the likelihood of illegitimate targets in their risk assessments.

可以透過訓練數據來辨別合法與非法目標（如醫院和學校），提升系統避免傷及無辜的能力，從而更符合倫理標準。例如，若電腦運算能產生「80%是戰車，10%是校車」的結果，即使目標識別的準確度在指揮官的風險容忍範圍內，系統仍能提醒指揮官及其幕僚進行更詳細的審查。建立能區分合法與非法目標的資料集，可能超出系統的資源限制，在這種情況下，指揮官應在風險評估中考慮非法目標的可能性。

Political 政治與文化

Political-cultural knowledge requires knowing how the use of an emerging technology will affect public expectations about the use of force, how these expectations affect society's perception of military service, and how other Department of Defense efforts to employ the emerging technology affect one's own efforts. Moreover, political-cultural knowledge requires senior military leaders to understand how shifts in public expectations will affect civil-military relations and military culture because public expectations will affect who joins the military and how they serve.

政治與文化知識要求瞭解新興技術的應用如何影響民眾對武力使用的期望，以及這些期望如何改變社會對軍事行動的看法。同時，也需掌握其他國防部門在運用新興技術時的策略，並評估其對自身工作的影響。此外，軍中高階領導幹部還需理解民眾期望的轉變將如何影響軍民關係與軍隊文化，因為這將影響民眾是否願意從軍及其在軍中的表現。

To the extent that using technology reduces risks to soldiers and noncombatants, doing so reduces the political risks associated with using force. Thus, senior military leaders will need to manage senior civilian leaders' expectations to ensure using technology does not risk escalation into a wider conflict. In addition, senior military leaders will need to manage public expectations about collateral harms to ensure the public's support. Perhaps most importantly, senior military leaders will need to manage expectations about the effectiveness of the technology so civilian leaders do not rely too much on technology and the public does not become frustrated by a lack of results. The public is not likely to trust a military that cannot deliver results and that imposes risks on soldiers and noncombatants alike.

當科技應用降低士兵和非戰鬥人員的風險時，也減少了使用武力帶來的政治風險。因此，高階軍事領導幹部必須設法做到上級領導人的期望，確保技術的使用不會引發更大規模的衝突。他們還需應對民眾對附帶損害的擔憂，以確保民眾支持。更重要的是，他們需要管理人們對科技效益的期望，避免領導人過度依賴科技，並防止民眾因成效不彰而失望。畢竟，民眾難以信任一支無法達成目標且讓士兵和非戰鬥人員陷於危險的軍隊。

Conclusion 結論

Developing and employing new military technologies is a part of being a military professional. Indeed, military history is a story of technological innovation and soldiers learning how to operate new systems. Many aspects of integrating AI are not new. Artificially intelligent technologies' capability to improve a wide range of military weapons, systems, and applications differentiates this type of technology from others. As this technology expands in application, war will be as much about managing data as it is about managing violence. Thus, commanders of the near future will need to understand how AI-enabled systems will interact with the commanders' judgments about risk to friendly forces and noncombatants. Commanders will also need to know how to ensure staffs and operators can curate and train data effectively. Finally, commanders and staffs will gain experience interacting with the private sector, which will increasingly be relied upon for AI and data technology and aspects of its operation.

發展和應用新軍事科技是軍事專業的重要一環。事實上，軍隊歷史就是技術創新和士兵學習新系統操作的歷程，許多 AI 整合的概念並不新穎。AI 技術提升了各種軍事武器、系統和應用的效能，使其與其他技術不同。隨著這項技術的廣泛應用，未來的戰爭將不再僅是暴力的展現，同時也會高度重視資訊的運用。因此，未來的軍事領導幹部必須瞭解 AI 技術如何影響他們對友軍和非戰鬥人員風險的判斷，他們還需確保幕僚和操作人員能有效管理和訓練數據，並與民間企業密切合作，因為未來將越來越依賴民間企業提供的 AI 和數據技術。

譯者簡介

蔡婷嫕少校，國防大學管理學院資管系99年班、通資電正規班110年班、墨爾本大學國際關係碩士；曾任排長、副連長、資訊官、教官，現任陸軍通信電子資訊訓練中心網路作戰組教官。