應用強化式學習於兵棋推演之探討

賀增原

網路安全與決策推演研究所

壹、前言

隨著技術的不斷進步，未來的電腦兵棋將不再僅僅依賴傳統模型，而是導入高度智慧化、能夠自我學習和適應的系統，使得軍事訓練和決策輔助過程更加有效率。因此，本研究以電腦兵棋系統「指揮：現代作戰專業版（Command：Modern Operation Professional Edition, CPE, 後文以 CPE 稱之）為平台，設計數個反制無人機的想定場景。並參考海軍研究院[1]與其它關於「層級強化學習」（Hierarchical Reinforcement Learning, HRL）方法的文獻[2]， 希望藉由強化式學習協助處理各種不同戰場複雜狀況，以了解不同戰術運用的差異，並作為後續電腦兵棋導入人工智慧的研究基礎。

貳、研究方法

一、強化學習

層級強化學習（HRL）是基於強化學習所發展出來，強化學習（Reinforcement Learning, RL）為機器學習中的一種方法，其目的在於研究代理者或者智能體（agent）在面對不同環境（environment）狀態（state）中的挑戰，將會採取什麼樣的行動（action）。智能體決定採取的任一種行動，其實就是在選擇一種策略（policy）。強化

[1] Scotty Black and Christian Darken, “Scaling Intelligent Agents in Combat Simulations for Wargaming”, *Interservice/Industry Training, Simulation, and Education Conference*, 2023 Paper No. 23302；Scotty Black and Christian Darken, “Scaling Artificial Intelligent for Digital Wargaming in Support of Decision-Making”, *NATO Science and Technology Organization*, 2023, Page 23-1 to 23-18；George Ellison and Andrew Shepherd, “Might Wargaming be Another Instance Where “Anything You Can Do, AI Can Do Better”?”, *Concept Paper*, 2024, doi: 10.20944/preprints202401. 1311. v1.

[2] 張倩、李天皓、白春光，〈基於多智能體強化學習的分層決策優化方法〉，《電子科技大學學報》，第 24 卷第 6 期， 2022 年 12 月，頁 91-96。

學習特別之處在於評估智能體採取行動後的狀態，藉由量化的大小給予獎勵（reward），期望在眾多策略中尋找一項最佳的策略。以上描述相關架構如圖 4-1。

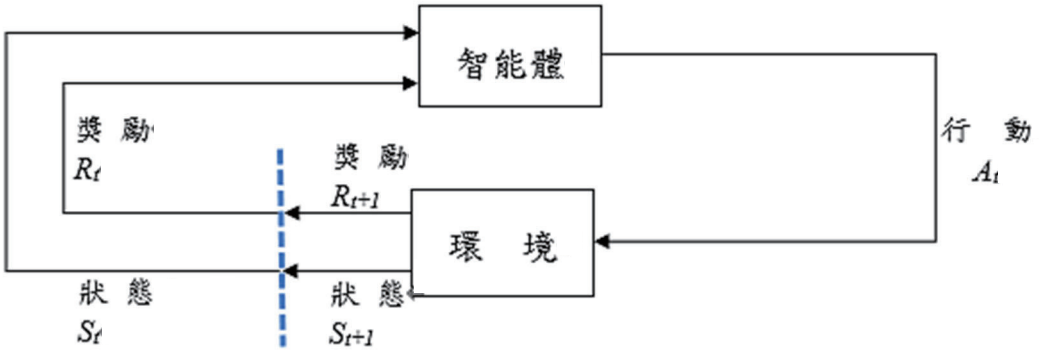

圖 4-1、強化學習的架構

資料來源：Scotty Black and Christian Darken, "Scaling Intelligent Agents in Combat Simulations for Wargaming," *Interservice/Industry Training, Simulation, and Education Conference*, 2023.

基於上述的說明，智能體採取哪種策略(π)，在不同的狀態(S)以獲得的獎勵（R），一般可以採用動作價值函數（action-value function）$Q^{\pi}(s,a)$來表示，[3]如下：

$$Q^{\pi}(s,a) = E[R(n)|s(t) = s, a(t) = a] \quad (1)$$

方程式(1)當中，當狀態(S_t)智能體會依據經驗採取一個行動(A_t)，方程式皆以小寫來代表。因此在時間(t)之後，環境依據行動給予獎勵，所以最初的獎勵方程式$R(t) = r(t+1) + r(t+2) + \cdots + r(T)$，不過智能體所採取的行動，不可能對未來發生的行動都有相同的影響，而是會隨著時間而遞減，因此需要一個介於 0 與 1 的衰減因子(discount factor) λ。所以獲得新的獎勵方程式：

[3] 蘇木春、張孝德，《機器學習-類神經網路、模糊系統以及基因演算法則》（新北市：全華圖書股份有限公司），2004 年，頁 6-12 至頁 6-16。

$$R(t) = r(t+1) + \lambda r(t+2) + \lambda^2 r(t+3) + \cdots = \sum_{k=0}^{\infty} \lambda^k r(t+k+1) \quad (2)$$

此處衰減因子 λ，如果 λ=0.01，隨著次方增加，當它是 2 次方，則第三項獎勵將要乘以 0.0001，就可以忽略；反之 λ=0.9，即使是$\lambda^7 = 0.48$，也仍然將近有一半的影響。由於智能體在不同環境，環境獎勵會隨著採取不同的策略給予考量，因此加入機率使動作價值函數是一個期望值。

本文先以一個例子來說明動作價值函數，[4]在此設計一個迷宮如圖 2。在此迷宮中僅有三種選擇，第一種前進、第二種向右、第三種向左，每一種選擇機率皆為1/3，前進一步獎勵為 6，衰減因子設為 0.5，檢視以下說明如表 4-1，說明各項選擇以及獎勵。從表 4-1 當中可以了解各種不同路徑的獎勵值，當得到獎勵值越大就越可能是迷宮的出口，當然仍然要視實際狀況而定。

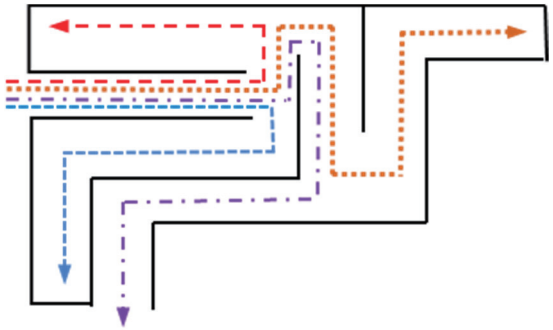

圖 4-2、迷宮

資料來源：作者繪製。

表 4-1、迷宮的路徑（取小數點後兩位）

路　　徑	獎　　勵　　值	期　　望　　值	成 功 與 否
紅線	10.5	3.5	失敗
橘線	11.91	3.97	失敗
藍線	11.25	3.75	失敗
紫線	11.81	3.94	成功

資料來源：作者整理。

[4] 梁瑋倫、林大衛、劉志尉，〈強化學習的簡介及其應用情境與高效訓練法〉，工研院產業學習網，https://college.itri.org.tw/Home/InfoData/f6e19f2d-f81c-421c-bc36-ea6409ba0a5d/e39de825-d056-4100-bbf9-cae46f0fe0aa （檢索日期：2024 年 6 月 20 日）。

二、層級強化學習

然而，更進一步「層級強化學習」的方法，是結合強化學習與階層架構的概念，將複雜的策略與多維度狀態空間劃分至不同階層，[5]各個階層負責不同任務或者行動（此處將任務切割為不同行動組合），藉此共享彼此的訊息。本研究參考 Scotty Black and Christian Darken 的論文，將兵棋推演中不同角色，[6]例如：指揮官、管理者與戰士分別劃分成不同層級，相關架構如圖 4-3。此外，戰場訊息通過不同的「觀察狀態通道」（observation space channel）傳遞給智能體，智能體會根據這些訊息來做出合理的決策。舉例來說，如果我們在訓練一個軍事模擬的層級強化學習模型，觀察狀態通道可能包括：[7]

- 敵方單位的位置；
- 我方單位的位置；
- 資源的狀態(彈藥、糧食、燃料等)；
- 戰場環境變化(氣象、地形、建築物等)；
- 敵我通信狀態(感測資訊、確保指揮與控制)。

每個通道都提供一種特定類型的觀察，智能體需要學會如何整合這些資訊來形成有效的行動策略。層級的概念允許智能體在不同的抽象層次上學習和操作，這可以幫助處理更複雜的任務。依據參考文獻將電腦兵棋推演模擬的狀態轉換成觀察狀態通道，本文依序編排如圖 4-4，真實運用中視實際狀態調整。

---

[5] 孫宇祥、彭益輝、李斌、周佳煒、張鑫磊、周獻中，〈智能博弈綜述：遊戲 AI 對作戰推演的啟示〉，《智能科學與技術學報》，第 4 卷第 2 期，2022 年，頁 157-173。

[6] 謝沛學，〈從下棋到作戰：人工智慧在電腦兵棋的運用及其挑戰〉，《戰略與評估》，第 11 卷第 2 期，2021 年，頁 151-178；賀志豪譯，〈中共的人工智慧兵棋推演〉，《國防譯粹》，第 46 卷第 7 期，2019 年，頁 78-81；林傳凱，〈國防安全研究院模式模擬與兵棋推演發展目標〉，《國防情勢特刊-模式模擬專題》，第 25 期，2023 年 3 月 15 日，頁 24-34。

[7] 蘇炯銘、羅俊仁、陳少飛、項風濤，〈海空跨域協同兵棋 AI 架構設計及關鍵技術分析〉，《指揮控制與仿真》，第 46 卷第 2 期，2024 年，頁 35-43；孫怡峰、李智、吳疆、王玉賓，〈作戰方案驅動的可學習兵棋推演智能體研究〉，《系統仿真學報》，第 36 卷第 7 期，2024 年，頁 1525-1535。

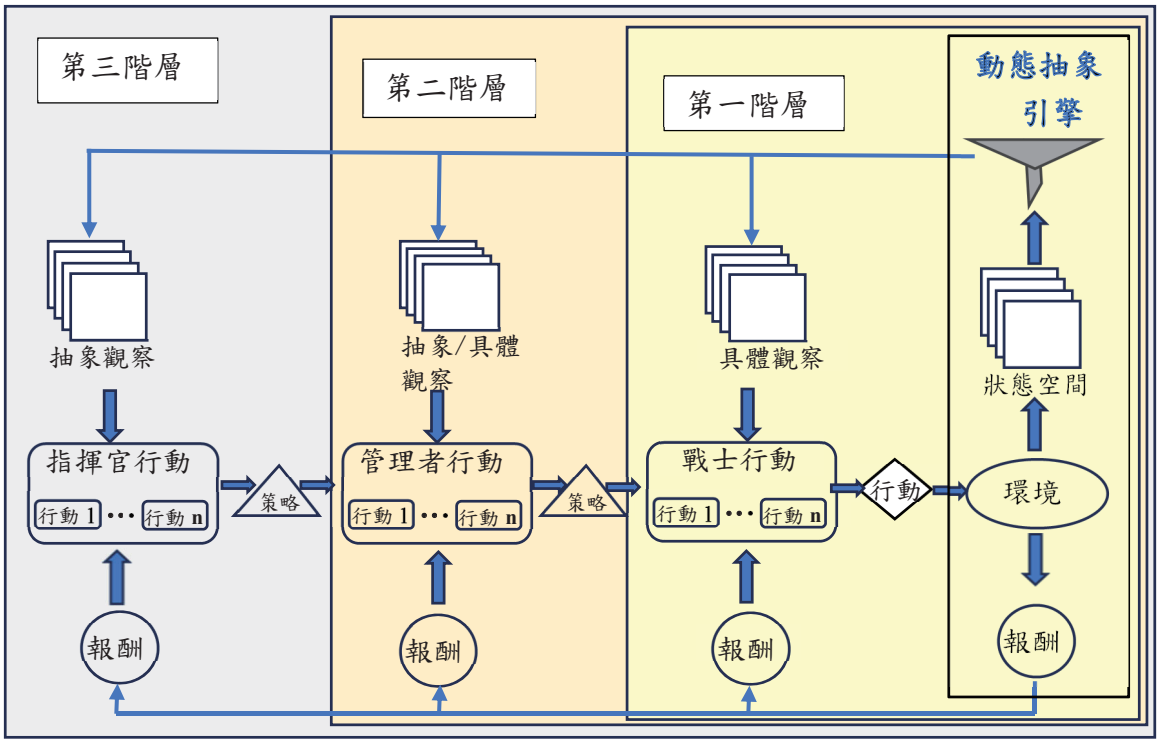

圖 4-3、層級強化學習架構

資料來源：整理 Scotty Black and Christian Darken, “Scaling Intelligent Agents in Combat Simulations for Wargaming,” Interservice/Industry Training, Simulation, and Education Conference, 2023.

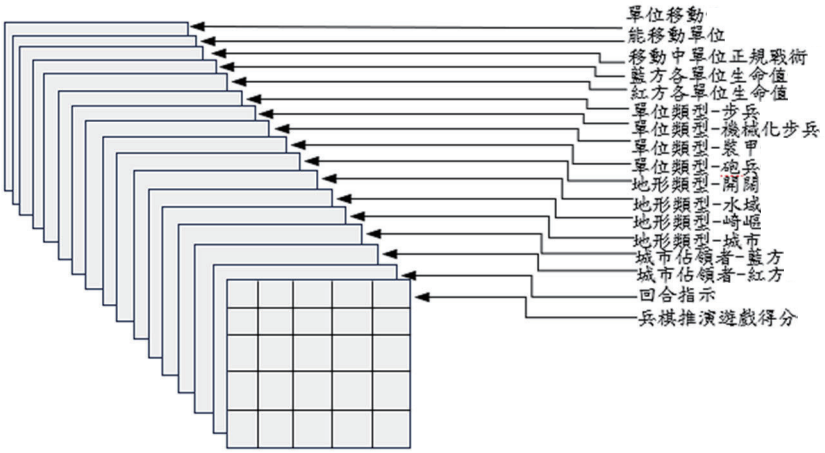

圖 4-4、觀察狀態通道

資料來源：整理 Scotty Black and Christian Darken, “Scaling Intelligent Agents in Combat Simulations for Wargaming,” Interservice/Industry Training, Simulation, and Education Conference, 2023.

參、兵棋推演個案介紹

運用動作價值函數以及層級強化學習，個案探討先從反制紅方「偵打一體」無人機開始，假定紅方運用翼龍-II 無人機 10 架準備偵查藍方關鍵基礎設施以及每架攜帶 2 枚藍箭飛彈進行攻擊；藍方派遣 3 輛車載式高能雷射武器系統如圖 4-5，各項單元的關聯性如圖 4-6，以及 6 輛復仇者防空飛彈系統進行防禦。

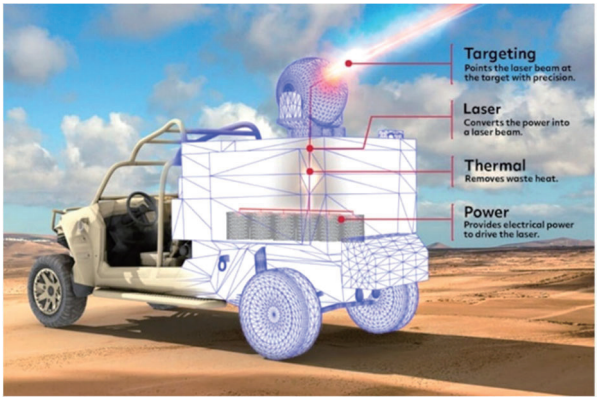

圖 4-5、車載式高能雷射武器系統

資料來源：https://www.rtx.com/raytheon/news/2021/06/22/how-laser-defeats-hostile-drones。

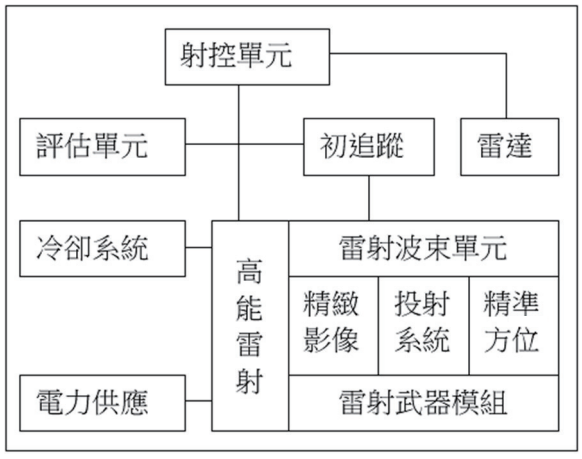

圖 4-6、雷射武器各單元相關性

資料來源：賀增原，〈模擬艦載雷射武器與防空飛彈任務效益之分析〉，第 19 屆軍事作業研究與模式模擬論壇，國立陽明交通大學，中華民國 112 年 9 月 26 日。[8]

[8] 賀增原，〈模擬艦載雷射武器與防空飛彈任務效益之分析〉，第 19 屆軍事作業研究與模式模擬論壇，國立陽明交通大學，2023 年 9 月 26 日，頁 327-335。

說明：雷射武器的射控系統首先經由雷達偵測遠距離的威脅，接著評估單元藉由初追蹤，光電目獲系統利用寬視角放大倍率小，窄視角放大倍率高來辨識目標，並且評估威脅；接著電力供應高能雷射，同時高能雷射會產生大量高溫需要冷卻系統來調節溫度，避免影響光學影像，同時導入波束單元，進行追蹤與攻擊。[9]

此個案作者運用電腦模擬系統（CPE）進行不同武器交戰，整個過程會執行蒙地卡羅 30 次，會統計不同狀況的機率，CPE 在輸出的檔案內可獲得不同交戰的執行細節，系統可以清楚交代日期與時間（時、分、秒）、執行行動位置（經緯度）、如何接戰情況、以及各單位災損狀況等資訊。例如執行結果災損可以分成四種：擊中（hit）、擊毀（kill）、未命中（miss）以及摧毀（destroyed）等。

進一步說明：本文採用高能雷射抵禦無人機攻擊，由於高能雷射會隨著功率密度形成光束在目標燒熔光斑，因此無法一次照射即擊毀無人機，必須擊中多次或者是持續擊中才能摧毀。另外還會發生無人機脫逃現象，與持續飛行甚至發生油料不足墜地才摧毀的情況。因此，參照之前迷宮的例子，統計蒙地卡羅執行的各項機率，並且將造成來襲威脅三種損失的強弱，分別設計不同的獎勵，其中「未命中」為“0”、「擊中」為“3”、「擊毀」為“6”，「衰退率」可以設定為任何值。並且將輸出的數據放入層級強化學習的動作價值函數進一步比較不同武器的組合，哪一種所能得到的報酬最大，以分析不同戰術的運用。

[9] K. Ludewigt, Th. Riesbeck, A. Graf and M. Jung, “50kW Laser Weapon Demonstrator of Rheinmetall Waffe Munition,” *Proceedings of SPIE - The International Society for Optical Engineering 8898*, October 15, 2013, https://www.researchgate.net/publication/260668879_50_kW_Laser_Weapon_Demonstrator_of_Rheinmetall_Waffe_Munition.

肆、執行成果

一、執行情況

此個案當中，統計紅軍翼龍-II 型共 10 架，遭到藍軍高能雷射（60k W Solid State Laser）3 台的防禦，由於電腦兵棋系統執行的過程，是每隔 30 秒派遣一架無人機，同時順序是從編號 10 至 1，所以編號第 10 號無人機會先接戰。統計 30 次蒙地卡羅，整理各架損害的情況以及藍軍高能雷射車輛損壞情況如圖 4-7，圖形先以前三次的例子來說明，隨後會統計紅軍各架次在蒙地卡羅 30 次中各種損壞的情況如表 4-2。

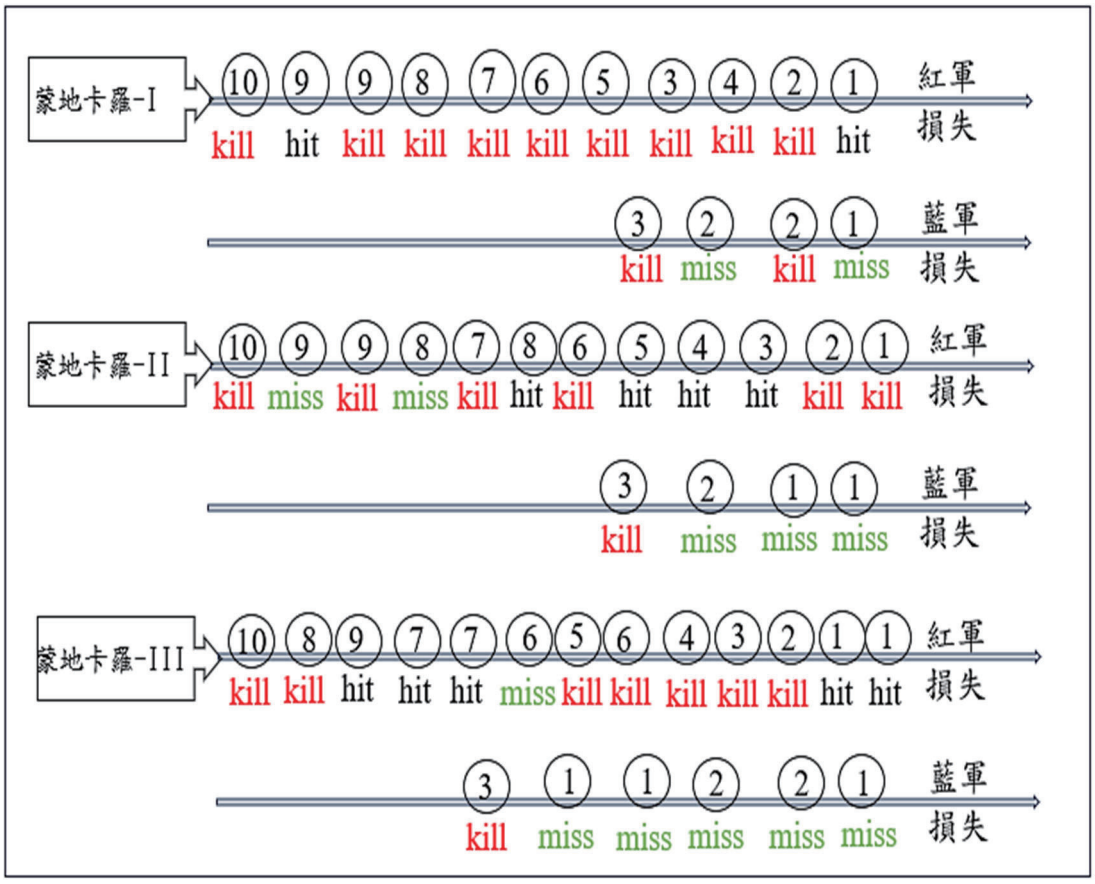

圖 4-7、紅藍軍損害情況

資料來源：作者整理。

在 CPE 執行的時間軸當中，第一次蒙地卡羅，可以發覺編號 10 的無人機首先被擊毀，接下來編號 9 的無人機被擊中，再隔 9 秒，

編號 9 的無人機就被擊毀，另外編號 1 的無人機被擊中，最後在失去引擎的功用，發生墜毀。

表 4-2、各架無人機損害統計

損害 / 無人機	miss	hit	kill	Sum
No. 1	7	18	18	43
No. 2	4	14	18	36
No. 3	5	9	24	38
No. 4	3	10	22	35
No. 5	2	21	14	37
No. 6	6	9	24	39
No. 7	3	14	21	38
No. 8	3	10	23	36
No. 9	3	16	21	40
No. 10	4	17	16	37

資料來源：作者整理。

二、分析方法

個案採用機率為分析方法，每台無人機 miss, hit, kill 的機率表示式：

未命中（miss）的機率

$$P_{miss,i} = \frac{miss\_count_{i,j}}{total\_shoot_{i,j}} \tag{3}$$

擊中（hit）的機率

$$P_{hit,i} = \frac{hit\_count_{i,j}}{total\_shoot_{i,j}} \tag{4}$$

擊毀(kill)的機率

$$P_{miss,i} = \frac{kill\_count_{i,j}}{total\_shoot_{i,j}} \tag{5}$$

方程式（3）至（5）中，i 代表無人機的架數，j 代表第幾回執行，分母則表示武器對某一個特定無人機射擊的枚數。

結合各台無人機不同狀況機率（miss, hit, kill）加上考慮衰減因子 λ 以及獎勵值，可以獲得新的動作價值函數

$$Q_i^{\pi}(s,a) = \sum_{j=1}^{N} \lambda^t \left(P_{miss,i,j} \cdot Rewards_{miss,i,j} + P_{hit,i,j} \cdot Rewards_{hit,i,j} + P_{kill,i,j} \cdot Rewards_{kill,i,j} \right) \quad (6)$$

接下來，運用 CPE 執行結果的數值，代入以上的方程式，來檢視強化式學習的不同功效。

三、執行結果

利用表格 4-2 的數據代入方程式（3）至（5）當中，可以獲得各架無人機的不同狀況機率。接者利用三種不同衰減因子，λ=0.3, 0.5, 0.7，利用圖七的順序，依序代入方程式（6）中，比較其期望值如表 4-3。從表 4-3 可以看出，無論任何的衰減因子，第一回合期望值優於第二回合，即使第二回合次數比第一回合多一次，但是由於未命中的次數達到兩次，造成整體的期望值降低；另外第三回合有兩種情況：第一種衰減因子= 0.3 由於第三回合前兩次攔截，均是造成擊毀，其次方的增加，造成後續的影響比較小，所以第三回合大於第一回合；第二種衰減因子=0.7，其次方的增加，造成後續的影響比較大，所以第三回合期望值比第一回合來得低分，藉由此表可以簡單說明不同衰減因子對應的期望值的大小，也因此在整個複雜的兵棋推演中，我們所期望是整體戰場的防禦能力，因此選擇衰減因子=0.7 較合適。

表 4-3、不同衰減因子對應的期望值(取小數點後兩位)

蒙地卡羅	衰減因子=0.3	衰減因子=0.5	衰減因子=0.7
第一回合	3.38	4.87	8.66
第二回合	2.91	3.65	5.86
第三回合	3.89	5.11	7.56

資料來源：作者整理。

四、進階分析

由於本文介紹的是層級強化學習，因此除了藍軍高能雷射的防禦，另外部署復仇者防空飛彈系統如圖 4-8 以及圖 4-9。[10]

圖 4-8、復仇者防空飛彈系統

資料來源：參考網頁 https://www.military.com/equipment/avenger-weapon-system。

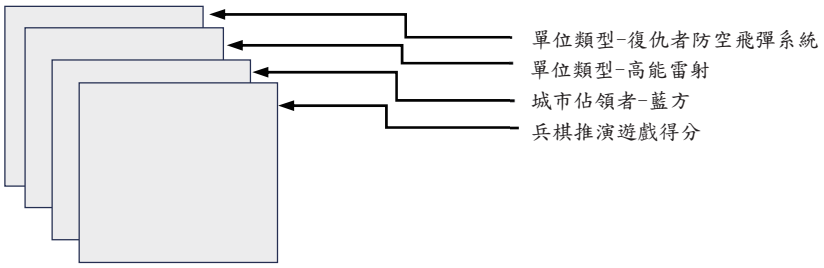

圖 4-9、層級強化學習的架構

資料來源：作者整理。

[10] 楊培毅，〈復仇者飛彈系統發展歷史與未來性能提升之研究〉，《砲兵季刊》，2017 年 6 月，頁 30-47。

整理前三次蒙地卡羅執行的狀況，此部分結合高能雷射與刺針飛彈，檢視各架損害的情況如圖 4-10。由於翼龍-II 型在 CPE 中飛行高度與攻擊高度為 4,572 公尺至 7,620 公尺，復仇者防空飛彈系統防禦高度為 60.9 公尺至 4,876.8 公尺，因此在防禦高度明顯不足的條件下，即便高能雷射未完成對無人機的擊毀，但是刺針飛彈仍是未命中，例如無人機編號 10。

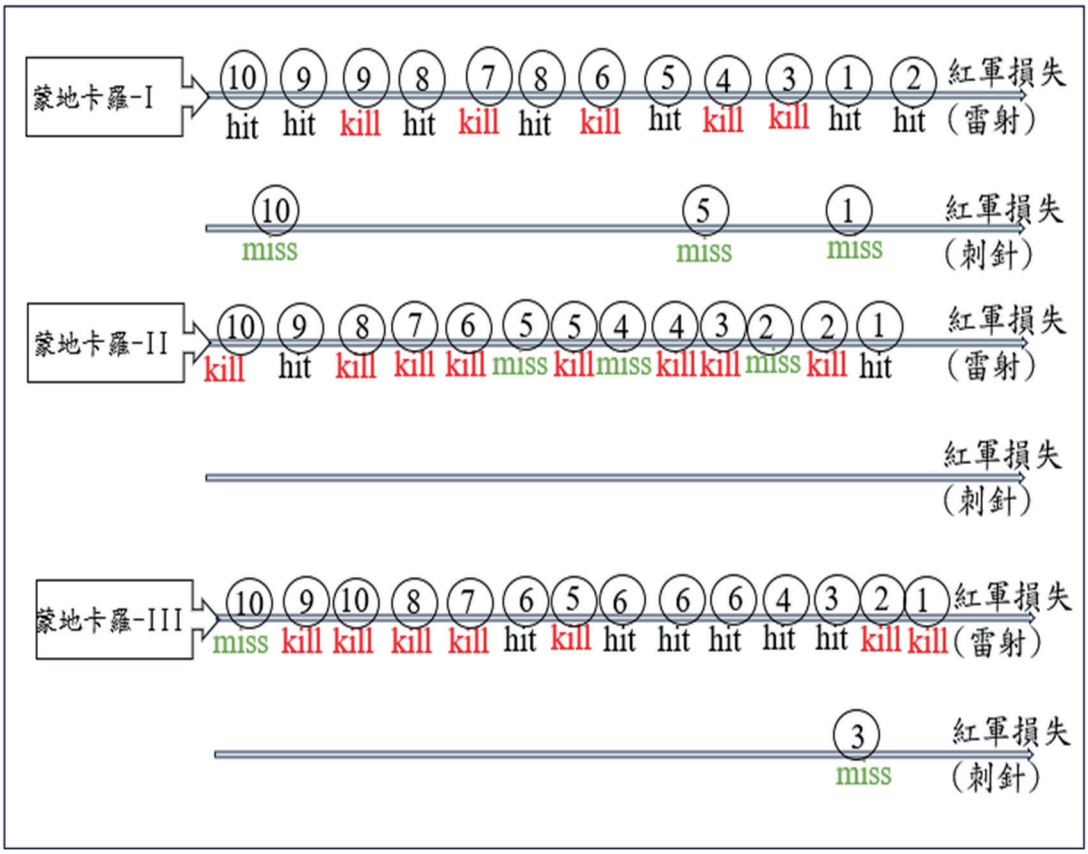

圖 4-10、結合高能雷射與刺針飛彈防禦情況

資料來源：作者整理。

由圖 4-10 可以大致看出伴隨著復仇者防空飛彈系統的加入其成效反而不如單獨高能雷射攔截效率好，除了第二回合之外，將圖 4-10 蒙地卡羅三個回合與圖七沒有部署刺針飛彈以衰減因子=0.7 做比較如表 4-4，此結果與參考文獻相近可以參考表 4-5。[11]

[11] 賀增原、林傳凱、謝沛學，〈運用 CMO 模擬雷射反制無人機威脅之研究〉，《陸軍後勤季刊》，第 4 期，2022 年，頁 62-73。

表 4-4、衰減因子=0.7 刺針飛彈部署前後期望值的差異

衰減因子=0.7	部署前(僅有高能雷射)	部署後(加上刺針飛彈)
蒙地卡羅第一回	8.66	5.99
蒙地卡羅第二回	5.86	8.05
蒙地卡羅第三回	7.56	6.20

資料來源：作者整理 。

表 4-5、紅藍軍毀損情況

行動方案	翼龍-II 毀損 (初始 10 架)	復仇者防空飛彈系統毀損 (初始 6 輛)	雷射系統毀損 (初始 3 輛)
行動方案_1	298/300		27/90
行動方案_2	297/300	20/180	0/90

資料來源：作者整理。

伍、結論

作者以「層級強化學習」的方法，改進過去採用其它機器學習所產生的問題，並結合 CPE 電腦兵棋所模擬的數據，探討不同反制無人機方案。因此。本文盡可能用方程式，配合圖表的說明，以利讀者的閱讀，從分析的結果也可以看到本文所採用的方法與 CPE 所統計的資料沒有相背離，並且藉由不同衰減因子對應的期望值，雖然蒙地卡羅是隨機執行，不過藉著不同大小的衰減因子，將會很明顯檢視出不同回合蒙地卡羅的期望值，未來將朝向使用工具或者撰寫程式代入到整體架構，如此將可以進一步檢視在不同的戰術條件下，何種戰術的效益最好，可供決策者參考。

本文作者賀增原為國防大學中正理工學院國防科學研究所博士，現為財團法人國防安全研究院網路安全與決策推演研究所研究員，研究領域：決策分析、國防武器獲得管理、系統工程。

Discussion on Applying Reinforcement Learning in Wargaming

Tzeng-Yuan Heh

Division of Cyber Security and Decision-Making Simulation

Abstract

As battlefield environments rapidly evolve, the question arises: Can traditional war games effectively address the complexities of today's warfare and assist in personnel training and decision support? With the rapid advancement of artificial intelligence (AI), various AI methods have already been applied to wargaming. This paper explores the intersection of traditional war games and AI, examining their potential synergies and challenges.

This paper introduces hierarchical reinforcement learning (HRL) methods in a systematic manner. Beginning with an overview of reinforcement learning, the discussion proceeds to illustrate various path choices and their corresponding rewards using a maze example. Subsequently, leveraging existing literature, an HRL framework and observation space channels are constructed to enhance the interpretation of complex battlefield environments.

In the final stage of the wargaming case study, different discount factors in each phased and incremental exploration were examined. Initially, results obtained from computer software were utilized to statistically assess the probabilities of various drone damage scenarios. Employing Monte Carlo simulations across different rounds, the situations in which each drone encountered attacks were systematically organized. Subsequently, expected values with varying discount factors were

computed to validate the accuracy of the data. A more advanced analysis involved hierarchical reinforcement learning, wherein interactions between different anti-aircraft missile systems were examined. Graphs and tables derived from this analysis were used to evaluate weapon effectiveness.

This paper aims to enhance readability by incorporating equations and explanatory charts. The analysis results demonstrate alignment between the methods employed in this study and the statistical data generated by computer software. Future study will explore the integration of tools or custom programming within the overall framework. This approach will allow for a more detailed examination of tactical conditions, enabling the identification of optimal strategies for decision-makers.

Keywords: Reinforcement Learning, Wargaming, War Mode