2021 級 信工計算機組 學生干皓丞 2101212850 人工智慧作業報告

Email : zxdfgcv@gmail.com
About me : https://kancheng.github.io/

0. 作業說明

此報告為人工智慧課程五篇閱讀報告中的第一篇 agent，全報告包含 agent、search、Markov decision process、Bayesian Network、Reinforcement Learning。第一篇所佔的領域涵蓋 agent 和 Reinforcement Learning。因為考量自身到對該領域知識的掌握程度不足，全報告採心得與翻譯。

GitHub Project : https://github.com/kancheng/kan-readpaper-cv-and-ai-in-2021

1. 原文獻資訊與作者
2. 報告內容心得與講述
3. 原研究文獻

1. 原文獻資訊與作者

A game-theoretic analysis of networked system control for common-pool resource management(公共池資源管理) using multi-agent reinforcement learning
使用多智能體強化學習進行公共池資源管理的網絡系統控制的博弈論分析

Arnu Pretorius; InstaDeep; Cape Town, South Africa
Scott Camerony; University of Oxford; Oxford, UK
Elan van Biljony; Stellenbosch University; Stellenbosch, South Africa
Tom Makkinky; University of Cape Town; Cape Town, South Africa
Shahil Mawjeey; University of Witwatersrand; Johannesburg, South Africa
Jeremy du Plessis; University of Cape Town; Cape Town, South Africa
Jonathan Shock; University of Cape Town; Cape Town, South Africa
Alexandre Laterre; random InstaDeep random; London, UK
Karim Beguir; rando InstaDeep rando; London, UK

https://arxiv.org/abs/2010.07777
Comments: 17 pages, 16 Figures, to appear in Advances of Neural Information Processing Systems (NeurIPS) conference, 2020
Subjects: Machine Learning (cs.LG); Computer Science and Game Theory (cs.GT); Multiagent Systems (cs.MA)

2. 報告內容心得與講述

(1) Motivation 动机

該研究因應近來多智能體強化學習(Multi-agent reinforcement learning)的提出，並且在網路系統的運用中充滿巨大潛力，研究者們想要將此跟所謂的博奕論進行結合，也就是所謂的賽局理論與囚犯困境等議題。

該研究博奕論運用不同工具，來分析在網絡多代理系統(multi-agent systems) 的設計中，若採用不同信息結構而產生的解決方案所產生出來的差異。這些信息結構(information structure)與代理之間共享的信息類型以及採用的通信協議和網絡拓撲(network topology)有關。

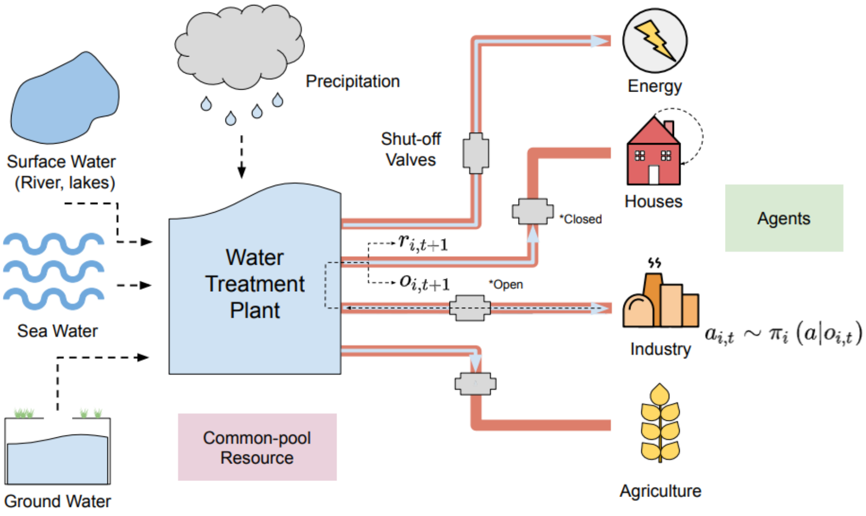
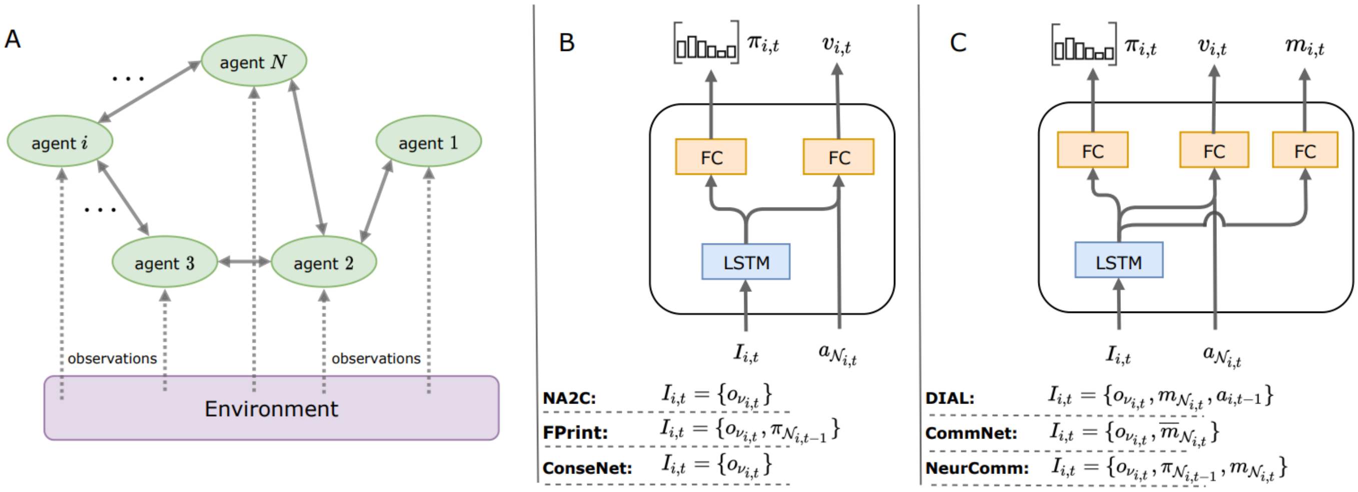


Figure 1: *Water management system as a common-pool resource environment for networked multi-agent reinforcement learning.* Shut-off valve controllers are treated as agents in the system and are responsible for providing water to different sectors of society, such as industry and agriculture.

(2) Intuition 直觉


A
agent N
agent i
agent 1
agent 3
agent 2
...
...
observations
observations
Environment
B
$\pi_{i,t}$
$v_{i,t}$
FC
FC
LSTM
$I_{i,t}$
$a_{\mathcal{N}_{i,t}}$
NA2C: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}\}$
FPrint: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}, \pi_{\mathcal{N}_{i,t-1}}\}$
ConseNet: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}\}$
C
$\pi_{i,t}$
$v_{i,t}$
$m_{i,t}$
FC
FC
FC
LSTM
$I_{i,t}$
$a_{\mathcal{N}_{i,t}}$
DIAL: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}, m_{\mathcal{N}_{i,t}}, a_{i,t-1}\}$
CommNet: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}, \overline{m}_{\mathcal{N}_{i,t}}\}$
NeurComm: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}, \pi_{\mathcal{N}_{i,t-1}}, m_{\mathcal{N}_{i,t}}\}$

在研究在 CPR 管理的背景下對用於網絡系統控制的實用 MARL 系統進行了第一次博弈論分析。CPR 在這裡可以視為牧民的草原，也就是所謂公共財的存在。這篇研究讓我們知道 MARL 系統顯示出不同的平衡剖面(distinct equilibrium profiles)，這取決於它們所採用的信息結構。雖然可微通信協議的系統(Systems with differentiable communication protocols)傾向於改善代理合作，但是，大多數系統即便獲得平衡時仍表現出低效率的狀態。而當該研究者使用鄰域加權獎勵函數時，他們發現新提出的 NeurComm 算法(Chu et al., 2020)能夠達到穩定的平衡，它對於整個系統以及每一個的個人代理、合作。總體上，該研究的發現強調了網絡控制系統設計對於有效 CPR 管理的重要性，且揭示了系統設計和由此產生的博弈論解決方案概念之間的相互依賴關係，這些概念是由於系統內採用的信息結構而產生。

該突圍該系統用水來進行實驗，在每一集開始時(At the beginning of each episode,)，水廠以 w0 = 0.5 的水量開始，並在每個時間步以 wt+1 = wt + c 的恆定速率再生（的實驗中使用不同的 c 值進行實驗） 實驗）。系統中的代理總數為 N = 4，每個代理的觀察空間由一個二維實值向量組成(each agent’s observation space consists of a two-dimensional real-valued vector containing)，其中包含值 wt 以及代理迄今為止輸送到其扇區的水量。如果水耗盡，則情節結束。 每集最多包含 1000 個步驟。(If the water is depleted, the episode ends. Each episode consists of a maximum of 1000 steps.)

(3) Justification 理由

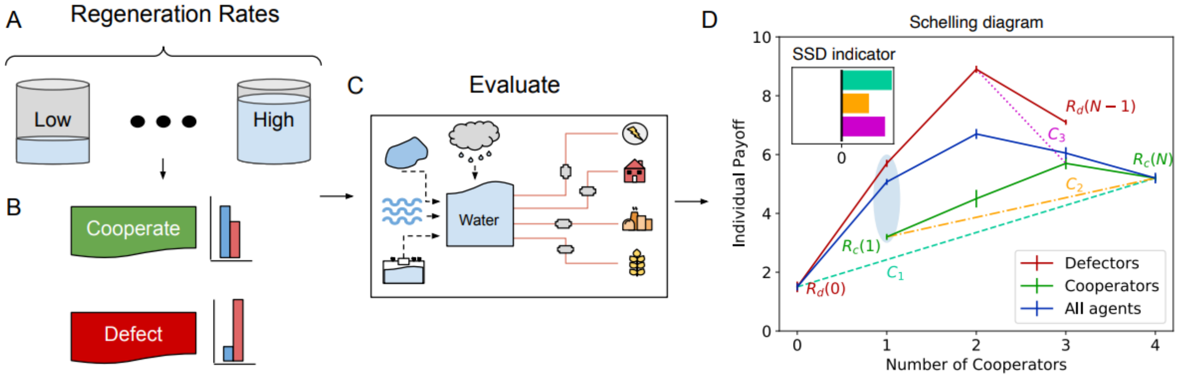


Figure 3: *Empirical game-theoretic analysis pipeline*. **(A)** Agents train on a common resource pool under different regeneration rates, where they may act greedily or show restraint. **(B)** Cooperative policies show restraint, while defecting policies act greedily. **(C)** Evaluate sampled policies on the environment. **(D)** Plot the Schelling diagram, which can be used to visually classify systems.

學生認為過往此問題在微觀經濟學中，使用此方式來廣泛討論公共財問題，比如草原牧民上的草原，因為牧民為了追求各自利益最大化，最後導致草原枯竭，又比如海洋漁業資源，各個漁民追求各自利益最大化捕撈，導致漁業資源消失，甚至河水的上下游廠商與外部性問題也用於此，比如上游鋼鐵廠排放廢水，下游魚群養殖業問題。而在系統設計上也應用此概念，比如線上遊戲系統，玩家之間的對戰，在能力分配上可以將實力相當的玩家，排在同一場對戰上，這樣可以讓玩家能夠獲得更好的遊戲體驗，從而讓該玩家更願意花更多的時間投入在這個遊戲上。在這裡博奕論用來討論在網路系統下的資源控制的問題，比如網路頻寬、計算資源的分配與內存(記憶體)的分配上。

如果可以學生會想要嘗試不同的方法，看看當中如圖二的每個節點中加入判斷。當整體的效率有所改進。

(4) Framework 框架

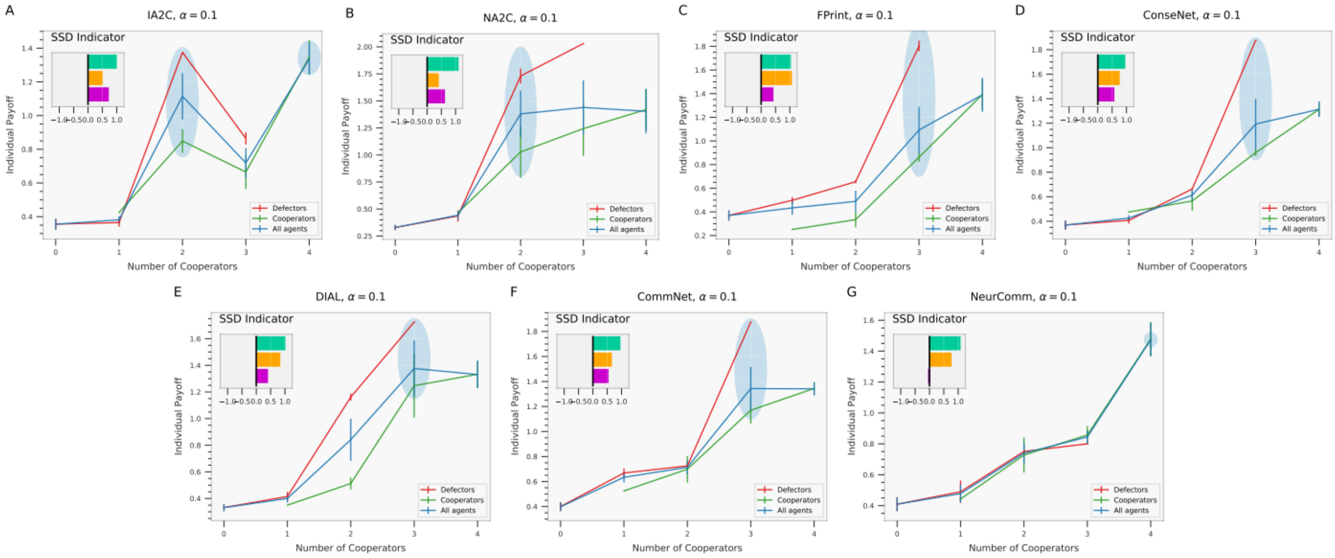


Figure 5: *EGTA for networked system control, with $\alpha = 0.1$.*(**A-G**) Schelling diagrams for each approach with sequential social dilemma (SSD) indicators given as insets. Potential equilibria are shaded in blue.

該研究，最重要的主軸，為下述三項。

(4)-1 多智能體強化學習(Multi-agent reinforcement learning) 與 博奕論

(4)-2 Common-pool resource management (公共池資源管理)，受 MARL 技術發展及其在系統控制中的適用性的推動，我們將 CPR 作為一項特別困難但非常重要的任務(highly important task.)

(4)-3 對序列社會困境 (SSDs; sequential social dilemmas) - Engineered networked systems as sequential social dilemmas(作為連續社會困境的工程網絡系統)，在 SSD 中，合作或背叛的戰略行動不再是原子性的(no longer atomic)，而是與代理通過政策學習跨時提取的戰略相關聯。(每個 agent)

從上圖的序列社會困境 (SSD) 指標作為插圖可以看到主要趨勢多數都是合作，直到所有的 agent 都合作時，主要策略才換成 defect。
A trend easily observed using this parameterisation is that for most algorithms the dominant strategy for a learned agent is to cooperate until all agents are cooperating, where at this point, the dominant strategy switches to defect.

使用這種參數化很容易觀察到的一個趨勢是，對於大多數算法，學習代理的主導策略是合作，直到所有代理都合作，此時，主導策略切換到 defect。

The only exception is the NeurComm algorithm, which is shown that have cooperation as the dominant strategy for any configuration of the system.
唯一的例外是 NeurComm 算法，該算法表明，對於系統的任何配置，都將合作作為優勢策略。

(5) Result 结果

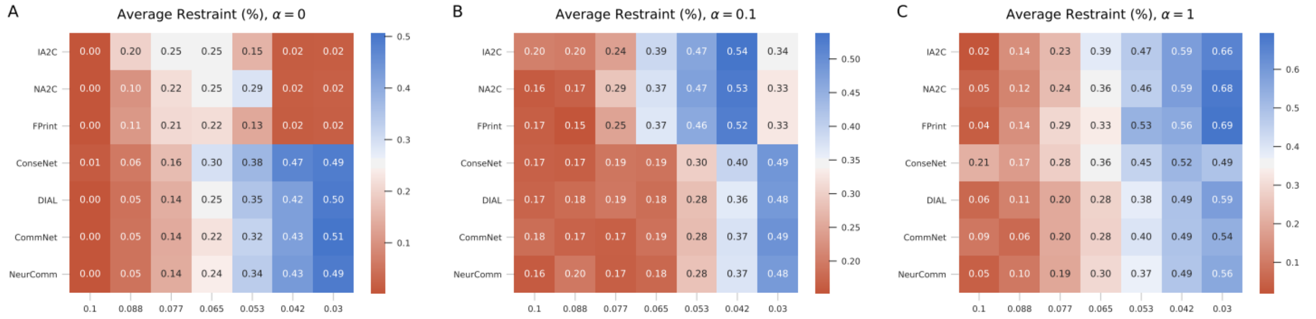


Figure 7: Heatmaps of average restraint percentage as a function of the regeneration rate for different MARL algorithms, from high (0.1) to low (0.03). **(A)** $\alpha = 0$, **(B)** $\alpha = 0.1$, **(C)** $\alpha = 1$.

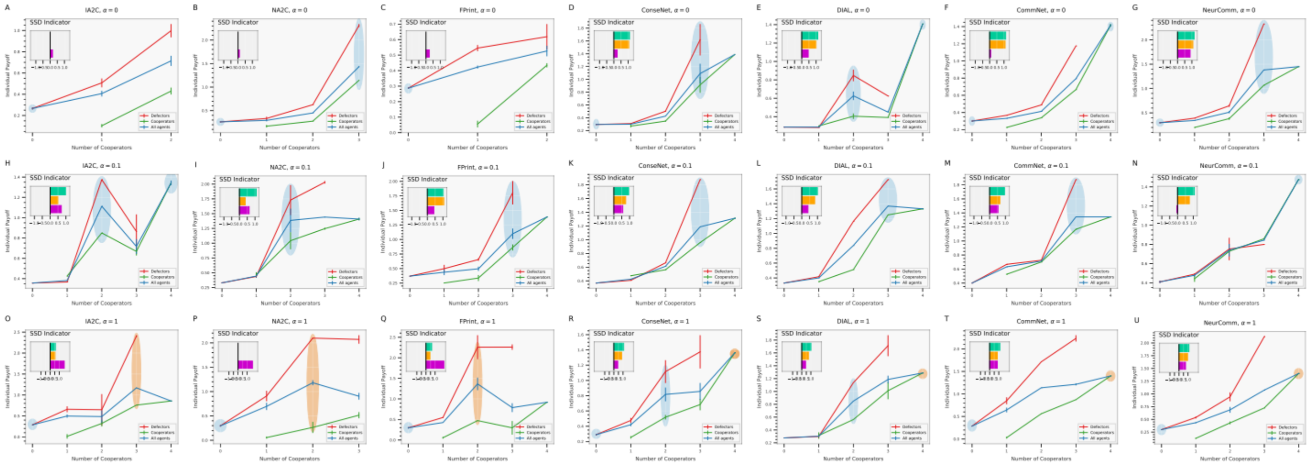


Figure 8: Schelling diagrams for each approach with network system sequential social dilemma (SSD) indicators given as insets. Potential Nash equilibria are shaded in blue. **Top row (A-G)** $\alpha = 0$, **Middle row (H-N)** $\alpha = 0.1$, **Bottom row (O-U)** $\alpha = 1$. Here we include orange shaded regions indicating configurations corresponding to the highest average payoff for all connected agents.

上面可以看到不同方法應用進去後的效果，並且使用熱力圖進行呈現。就如該研究所提到的，我們的 N 人網絡馬爾可夫遊戲中的每個玩家都尋求最大化連接加權收益，其中γ 是選擇的折扣因子，α是相鄰節點的權重 玩家獎勵，在這當中若α = 1，連接的玩家們會尋求最大化共享的全局獎勵，也就是所謂的整體，而如果 α = 0，玩家只最大化他們自己私人的獎勵。

最後，當 0 < α < 1 時，玩家在最大化自己的獎勵的同時，還要考慮他們所連接的鄰域獲得的一定比例的獎勵。

3. 原研究文獻

0. Abstract 摘要

多智能體強化學習(Multi-agent reinforcement learning)最近顯示出作為網絡系統控制方法的巨大希望。

可以說，適用於大規模網絡系統控制的最困難和最重要的任務之一是公共池資源管理。

重要的公共池資源包括耕地、淡水、濕地、野生動物、魚類種群、森林和大氣，其中的適當管理與社會面臨的一些最大挑戰有關，例如糧食安全、不平等和氣候變化。在這裡，我們從最近的一項研究項目中汲取靈感，該項目調查了人類在社會困境（例如眾所周知的公地悲劇）中的博弈論激勵。

然而，與其專注於生物進化的類人代理(human-like agents)，我們更關心的是更好地理解包含通用強化學習代理的工程網絡系統的學習和操作行為，僅受非生物約束，如內存、計算和通信帶寬(communication bandwidth)。

利用經驗博弈論(game-theoretic analysis)分析中的工具，我們分析了由於在網絡多代理系統(multi-agent systems)設計中採用不同信息結構而產生的解決方案概念的差異。這些信息結構(information structure)與代理之間共享的信息類型以及採用的通信協議和網絡拓撲(network topology)有關。

我們的分析為與某些設計選擇相關的後果提供了新的見解，並提供了超越效率、穩健性、可擴展性和平均控制性能的系統之間比較的額外維度。

1. Introduction 前言

基於多智能體強化學習 (MARL; multi-agent reinforcement learning) 的智能控制系統在解決以前手工設計的啟發式系統無法解決的困難任務方面(heuristic based systems)具有巨大潛力。
這部分是由於許多最近的 MARL 創新利用了新穎的信息結構，例如集中式訓練方案來克服非平穩性問題 (Lowe et al., 2017)和學習可擴展和有效合作的網絡通信協議(Foerster et al.,2016; Sukhbaatar et al., 2016; Chu et al., 2020)。

此外，負責社會福祉的關鍵基礎設施已被證明適合多代理系統控制(multi-agent system control)，包括電力、電信和交通系統的管理(Herrera et al., 2020; Haydari and Yilmaz, 2020; Chu et al., 2020)。
傳統的基於規則的系統與基於 MARL 的系統之間的根本區別在於，MARL 系統的操作行為完全依賴於學習到的代理策略(learned agent policies)。

這些策略反過來又是環境中代理交互(agent interaction)過程的產物，取決於系統設計者所使用的特定信息結構。

在網絡化 MARL 系統的上下文中，我們使用類似於 Zhang et al 的術語信息結構(information structure)(2019)。指的是在代理之間共享的信息類型以及採用的通信協議(communication protocol)和網絡拓撲(network topology)。

網絡系統控制的博弈論分析(Game-theoretic analysis of networked system control)
鑑於網絡 MARL 系統由學習的代理組成，這樣一個控制系統的設計者能夠理解控制代理之間交互和整個系統運行的潛在激勵就變得很重要 ，尤其是在安全關鍵控制場景中，例如運輸(transportation)或生命支持資源(life-supporting resource)的管理。

為此，我們轉向博弈論並考慮使用經驗博弈論分析 (EGTA) (Walsh et al., 2002; Wellman, 2006; Tuyls et al., 2018) 作為更好地理解網絡 MARL 系統的可行方法。

具體來說，我們分析了用於公共池資源 (CPR; common-pool resource) 管理的網絡 MARL 系統的學習操作行為 (Gardner et al., 1990)，並調查了以下研究問題 (Q*)：

Q: What are the game-theoretic solution concepts that arise from different information structures within a networked multi-agent reinforcement learning system used for common-pool resource management?*
問：用於公共池資源管理的網絡化多智能體強化學習系統中的不同信息結構所產生的博弈論解決方案概念是什麼？

Common-pool resource management (公共池資源管理)

受 MARL 技術發展及其在系統控制中的適用性的推動，我們將 CPR 作為一項特別困難但非常重要的任務(highly important task.)。

CPR 是難以或不可能被排除在訪問之外的資源，並且其中一個代理提取的資源會減少所有剩餘代理(all remaining agents)可提取的資源，至少在特定時間段內(Ostrom, 1990)。

這些資源通常是可再生的（可再生的），自然和/或人為因素決定了它們的更新速度，但如果以不可持續的速度佔用，仍然會完全耗盡。
CPR 構成了重要的生命支持資源的很大一部分，包括耕地、淡水、森林、大氣和氣候。

在自利代理人的情況下，CPR 面臨一種特殊的社會困境 (Dawes, 1980)，被稱為公地悲劇 (Hardin, 1968; Lloyd, 1833)。

通過遞歸推理，獨立的理性代理達到他們最好的 (Nash equilibrium)策略，即盡可能快地佔用 CPR 到用盡的程度 (since all other agents are very likely to do the same)。

當意識到如果代理人為更可持續地使用資源而合作時，從長遠來看，每個代理人(every agent)(即使從自私的角度(a selfish perspective))都會變得更好，悲劇就完全顯現出來了。

Engineered networked systems as sequential social dilemmas
(作為連續社會困境的工程網絡系統)

Many recent works have focused on cooperation in the context of social dilemmas.
最近的許多工作都集中在社會困境背景下的合作(Kleiman-Weiner et al., 2016; Peysakhovich and Lerer, 2017; Lerer and Peysakhovich, 2017; Peysakhovich and Lerer, 2018a,b; Foerster et al., 2018).

這包括一個成熟的研究計劃，該計劃調查使用 EGTA 在類人(humanlike) MARL 代理中出現合作(Leibo et al., 2017; Perolat et al., 2017; Hughes et al., 2018; Jaques et al., 2018; Köster et al., 2020)。

特別是， Leibo et al. (2017)使用 MARL 和 EGTA 將 2-player 重複矩陣博弈社會困境框架(2-player

repeated matrix game social dilemma framework)擴展到對序列社會困境 (SSDs; sequential social dilemmas) 進行建模。

在 SSD 中，合作或背叛的戰略行動不再是原子性的(no longer atomic)，而是與代理通過政策學習跨時提取的戰略相關聯。

然後，可以根據經驗將涉及兩個代理的建模社會困境情況的戰略收益和相應的均衡估計為一組抽樣政策的預期回報，每個政策代表合作或背叛。

可以使用 Schelling (1973) 的方法對 N-player SSD 進行類似的分析，正如 Perolat et al. (2017) 所做的那樣和休斯等人。 (2018) 研究人類在社會社區中互動時 MARL 代理的動態。

在這裡，我們純粹從網絡系統工程的角度考慮 SSDs。

Our work is related to the study of games on networks (Jackson and Zenou, 2015), but is more focused on learning agents in the context of MARL.
我們的工作與網絡遊戲的研究有關(Jackson and Zenou, 2015)，但更側重於 MARL 背景下的學習代理。
具體來說，我們調查了 SSDs 從用於 CPR 管理的網絡 MARL 控制系統中出現的可能性。

經典博弈論文獻(Classical game theory literature)表明，直接信息共享和某些類型的交流等信息結構可以改變與不同代理策略相關的均衡(Roth and Malouf, 1979; Myerson, 1986; Farrell and Gibbons, 1989; Compte, 1998)。

因此，由於 SSD 是一種環境和策略相關的現象，我們也期望 MARL 系統中用於 CPR 管理的不同信息結構具有不同的博弈論解決方案。

Summary of our findings and contributions - 我們的發現和貢獻總結

To the best of our knowledge, we conduct the first game-theoretic analysis of practical MARL systems for networked system control in the context of CPR management.
據我們所知，我們在 CPR 管理的背景下對用於網絡系統控制的實用 MARL 系統進行了第一次博弈論分析。

具體來說，我們的分析揭示了以下幾點

Answer to Q*:

MARL 系統顯示出不同的平衡剖面(distinct equilibrium profiles)，這取決於它們所採用的信息結構。

具有可微通信協議的系統(Systems with differentiable communication protocols)傾向於改善代理合作，但是，大多數係統配置文件在平衡時仍然表現出低效率。

*有趣的是，當使用鄰域加權獎勵函數時，我們發現新提出的 **NeurComm** 算法(**Chu et al., 2020**)能夠達到穩定的平衡，它對於整個系統以及每個 個人代理，合作。*

*總的來說，我們的發現強調了網絡控制系統設計對於有效 **CPR** 管理的重要性。*

特別是，我們揭示了系統設計和由此產生的博弈論解決方案概念之間的相互依賴關係，這些概念是由於系統內採用的信息結構的特定選擇而產生的。

2 Methodology 方法

我們的調查涉及從系統工程的角度對 MARL 進行 CPR 管理的理解。

儘管受到先前工作的啟發(Perolat et al., 2017; Hughes et al., 2018)，該工作構成了一個更大的博弈論驅動研究計劃的一部分，該研究計劃是關於在類人代理的社會社區中出現合作的。

相反，我們研究了由通用強化學習代理(reinforcement learning agents)組成的工程網絡系統，僅受內存、計算和通信帶寬等約束。

然而，與上述工作類似，當我們試圖回答 answer Q* 時，我們的議程僅在描述科學領域。

換句話說，我們沒有為網絡化 MARL 系統控制提供任何規範性創新，並將其作為基於我們的發現的未來工作的考慮因素。

相反，我們在這項工作中的目標是獲得對學習網絡 MARL 系統的運行條件以及與其設計相關的博弈論後果的新見解和理解。

Networked N-player partially observable Markov games 聯網 N 玩家部分可觀察馬爾可夫博弈

作為建模設備，我們考慮部分可觀察馬爾可夫博弈，其中玩家通過網絡連接。

我們這裡的馬爾科夫博弈構造類似於 Chu et al. (2020)定義的網絡化 MDP。

具體來說，我們考慮由一組節點 (vertices)V 以及一組邊連接組成的圖 G(V, E)
E = {(i, j)|i, j ∈ V, i ≠ j}，其中每個玩家是圖中的一個節點，本地連接到其他玩家節點。

Players i and j from V are connected if the tuple (i, j) is in the set E.
如果元組 (i, j) 在集合 E 中，則來自 V 的玩家 i 和 j 是連接的。
Each player has it’s own local d-dimensional view of the global state S obtained through an observation function
每個玩家都有自己的全局狀態 S 的局部 d 維視圖，通過觀察函數獲得

$$O_i : \mathcal{S} \to \mathcal{O}_i \subseteq \mathbb{R}^d, \text{ where } \mathcal{S} = \prod_{i=1}^{|\mathcal{V}|} \mathcal{O}_i$$

Players take actions from their respective action sets Ai,
玩家從他們各自的行動集 Ai 中採取行動，for i = 1, … |V|.

為了便於使用圖中的連接進行通信，我們定義了一個通信空間 C。

具體而言，讓玩家 i 周圍的連通鄰域為

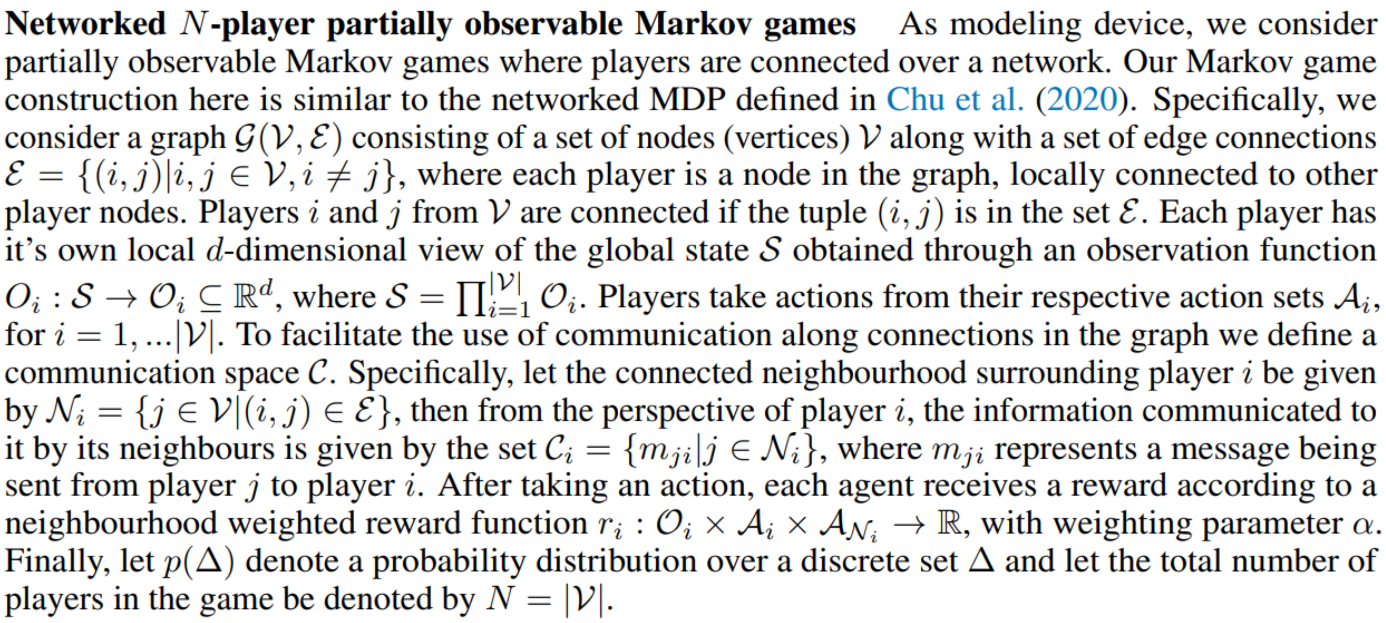
by $\mathcal{N}_i = \{j \in \mathcal{V} | (i, j) \in \mathcal{E}\}$

, then from the perspective of player i, the information communicated to it by its neighbours is given by the set Ci = fmjij, where mji represents a message being sent from player j to player i.

例如，那麼從玩家 i 的角度來看，其鄰居傳達給它的信息由集合 $\mathcal{C}_i = \{m_{ji} | j \in \mathcal{N}_i\}$ 給出，其中 mji 表示從玩家 j 發送給玩家 i 的消息。

採取行動後，每個智能體根據鄰域加權獎勵函數 $r_i : \mathcal{O}_i \times \mathcal{A}_i \times \mathcal{A}_{\mathcal{N}_i} \to \mathbb{R}$ 獲得獎勵，並帶有加權參數。

Finally, let p(∆) denote a probability distribution over a discrete set ∆ and let the total number of players in the game be denoted by N = |V|.
最後，讓 p(∆) 表示離散集合 ∆ 上的概率分佈，並讓遊戲中的玩家總數表示為 N = |V|。



Networked N-player partially observable Markov games As modeling device, we consider partially observable Markov games where players are connected over a network. Our Markov game construction here is similar to the networked MDP defined in Chu et al. (2020). Specifically, we consider a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ consisting of a set of nodes (vertices) $\mathcal{V}$ along with a set of edge connections $\mathcal{E} = \{(i, j) | i, j \in \mathcal{V}, i \neq j\}$, where each player is a node in the graph, locally connected to other player nodes. Players i and j from $\mathcal{V}$ are connected if the tuple (i, j) is in the set $\mathcal{E}$. Each player has it's own local d-dimensional view of the global state $\mathcal{S}$ obtained through an observation function $O_i : \mathcal{S} \to \mathcal{O}_i \subseteq \mathbb{R}^d$, where $\mathcal{S} = \prod_{i=1}^{|\mathcal{V}|} \mathcal{O}_i$. Players take actions from their respective action sets $\mathcal{A}_i$, for $i = 1, \ldots |\mathcal{V}|$. To facilitate the use of communication along connections in the graph we define a communication space $\mathcal{C}$. Specifically, let the connected neighbourhood surrounding player i be given by $\mathcal{N}_i = \{j \in \mathcal{V} | (i, j) \in \mathcal{E}\}$, then from the perspective of player i, the information communicated to it by its neighbours is given by the set $\mathcal{C}_i = \{m_{ji} | j \in \mathcal{N}_i\}$, where m_{ji} represents a message being sent from player j to player i. After taking an action, each agent receives a reward according to a neighbourhood weighted reward function $r_i : \mathcal{O}_i \times \mathcal{A}_i \times \mathcal{A}_{\mathcal{N}_i} \to \mathbb{R}$, with weighting parameter α. Finally, let $p(\Delta)$ denote a probability distribution over a discrete set Δ and let the total number of players in the game be denoted by $N = |\mathcal{V}|$.

Figure 1: Water management system as a common-pool resource environment for networked multiagent reinforcement learning.
圖 1：水管理系統作為網絡多智能體強化學習的公共池資源環境。

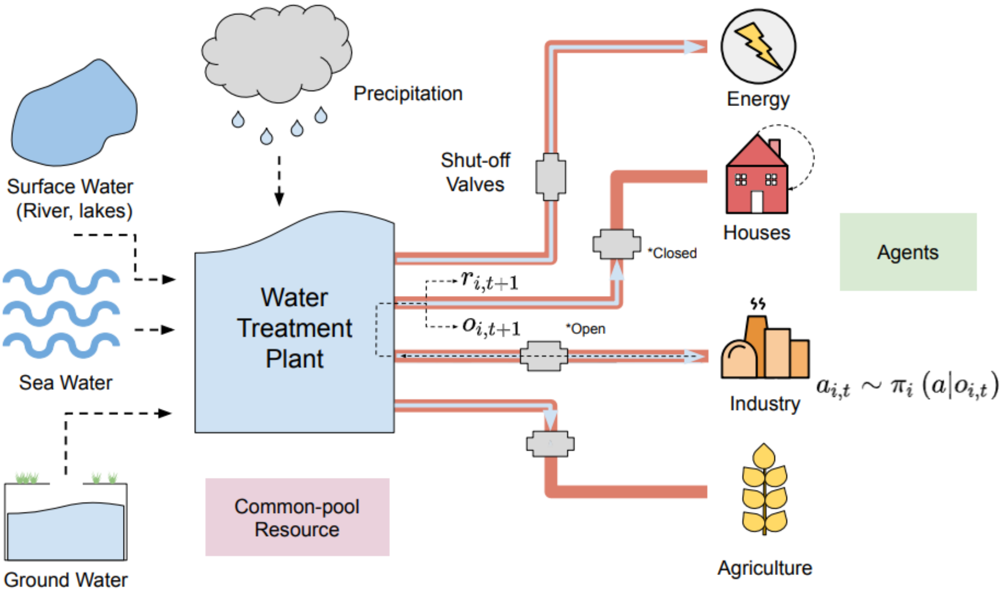
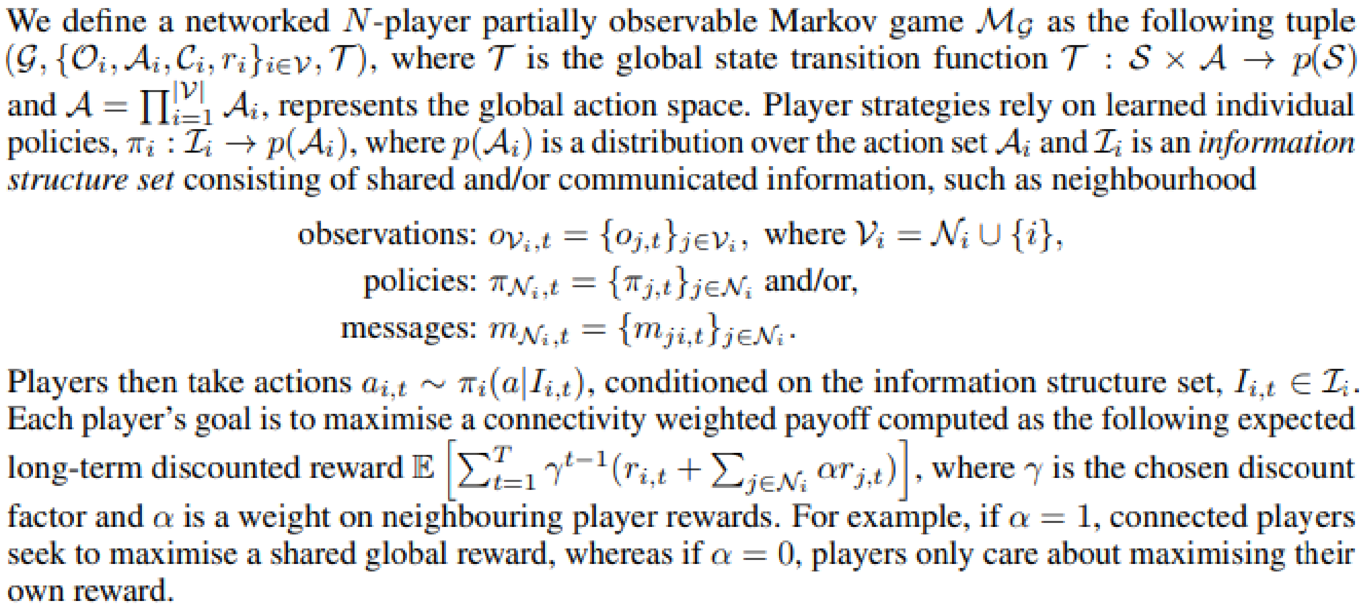


Figure 1: *Water management system as a common-pool resource environment for networked multi-agent reinforcement learning.* Shut-off valve controllers are treated as agents in the system and are responsible for providing water to different sectors of society, such as industry and agriculture.

截止閥控制器被視為系統中的代理，負責向工業和農業等社會不同部門供水。



We define a networked N-player partially observable Markov game $\mathcal{M}_{\mathcal{G}}$ as the following tuple $(\mathcal{G}, \{\mathcal{O}_i, \mathcal{A}_i, \mathcal{C}_i, r_i\}_{i \in \mathcal{V}}, \mathcal{T})$, where $\mathcal{T}$ is the global state transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow p(\mathcal{S})$ and $\mathcal{A} = \prod_{i=1}^{|\mathcal{V}|} \mathcal{A}_i$, represents the global action space. Player strategies rely on learned individual policies, $\pi_i : \mathcal{I}_i \rightarrow p(\mathcal{A}_i)$, where $p(\mathcal{A}_i)$ is a distribution over the action set $\mathcal{A}_i$ and $\mathcal{I}_i$ is an *information structure set* consisting of shared and/or communicated information, such as neighbourhood

$$\begin{aligned} \text{observations: } & o_{\mathcal{V}_i,t} = \{o_{j,t}\}_{j \in \mathcal{V}_i}, \text{ where } \mathcal{V}_i = \mathcal{N}_i \cup \{i\}, \\ \text{policies: } & \pi_{\mathcal{N}_i,t} = \{\pi_{j,t}\}_{j \in \mathcal{N}_i} \text{ and/or,} \\ \text{messages: } & m_{\mathcal{N}_i,t} = \{m_{ji,t}\}_{j \in \mathcal{N}_i}. \end{aligned}$$

Players then take actions $a_{i,t} \sim \pi_i(a|I_{i,t})$, conditioned on the information structure set, $I_{i,t} \in \mathcal{I}_i$. Each player's goal is to maximise a connectivity weighted payoff computed as the following expected long-term discounted reward $\mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1}(r_{i,t} + \sum_{j \in \mathcal{N}_i} \alpha r_{j,t})\right]$, where γ is the chosen discount factor and α is a weight on neighbouring player rewards. For example, if $\alpha = 1$, connected players seek to maximise a shared global reward, whereas if $\alpha = 0$, players only care about maximising their own reward.

We define a networked N-player partially observable Markov game Mg as the following tuple

我們將一個聯網的 N 玩家部分可觀察馬爾可夫博弈 Mg 定義為以下元組

Player strategies rely on learned individual policies,

玩家策略依賴於學習到的個人策略

Players then take actions ai;t   i(ajIi;t), conditioned on the information structure set, Ii;t 2 Ii.

玩家然後採取行動 ai;t i(ajIi;t)，以信息結構集 Ii;t 2 Ii 為條件。

Each player's goal is to maximise a connectivity weighted payoff computed as the following expected long-term discounted reward E

每個玩家的目標是最大化連接加權收益，計算如下預期 .. 長期折扣獎勵 E

, where γ is the chosen discount factor and α is a weight on neighbouring player rewards.

，其中 γ 是選擇的折扣因子，是相鄰玩家獎勵的權重。

For example, if α = 1, connected players seek to maximise a shared global reward, whereas if = 0, players only care about maximising their own reward.

例如，如果 α = 1，連接的玩家尋求最大化共享的全局獎勵，而如果 α = 0，玩家只關心最大化他們自己的獎勵。

Water management system as multi-agent CPR environment - 作為多代理 CPR 環境的水管理系統

一個重要的維持生命的 CPR 的例子是水。

地表水和地下水的製度層面管理已被證明特別困難，同時其重要性也在增加(Baudoin and Arenas, 2020)。

這在很大程度上是因為越來越需要係統更適應來自全球氣候變化等影響的外部壓力(Schlager and Heikkila, 2011)。

然而，有證據表明，至少在機構層面，CPR 管理系統可能會從適當的設計中受益匪淺 (Sarker and Itoh, 2001; Sarker et al., 2009)，基於深思熟慮的設計原則 (Ostrom, 1990; Ostrom et al., 1991; Ostrom, 1993)。

儘管上述這些研究涉及許多不同的利益相關者（遠遠超出單個工程系統），但我們設想精心設計的網絡控制系統將作為更大機構級智能 CPR 管理系統的組成部分發揮越來越重要的作用，而不僅僅是水 ，但適用於任何類型的 CPR。

因此，在這項工作中，我們只專注於 CPR 管理的網絡控制系統。

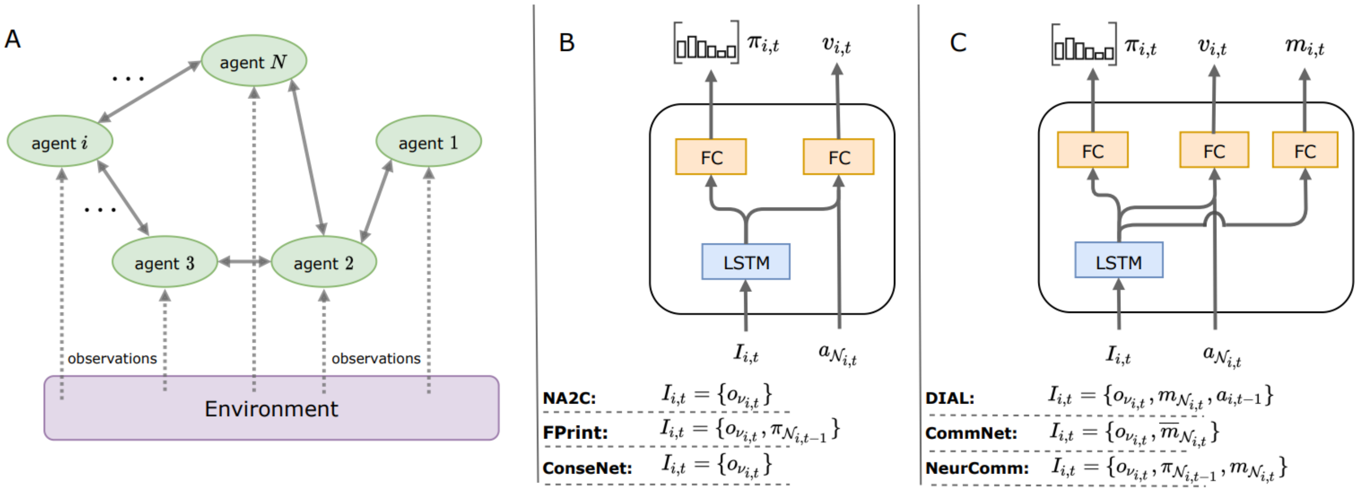
在我們的實驗中，我們設計了一個簡化的水管理系統模型，如圖 1 所示。

在這裡，控制組件被視為系統中的代理，負責向工業和農業等社會不同部門提供水，沿著連接到水處理廠的管道供水。

從每個代理控制器的角度來看，處理廠內的水代表 CPR，即控制器不能被其他控制器排除在外，並且一個控制器將水管道輸送到特定部門會減少其他控制器可用的水。

代理行為包括打開或關閉截止閥，並根據代理能夠向其各自部門供應的水量給予獎勵。

具體來說，在每個時間步長 t，打開閥門允許有 xt = 0.1/N 的水流過管道，而關閉閥門會限制所有流量，即在每個時間步長 xt = 0。


A
agent N
agent i
agent 1
agent 3
agent 2
observations
observations
Environment
B
$\pi_{i,t}$
$v_{i,t}$
FC
FC
LSTM
$I_{i,t}$
$a_{\mathcal{N}_{i,t}}$
NA2C: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}\}$
FPrint: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}, \pi_{\mathcal{N}_{i,t-1}}\}$
ConseNet: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}\}$
C
$\pi_{i,t}$
$v_{i,t}$
$m_{i,t}$
FC
FC
FC
LSTM
$I_{i,t}$
$a_{\mathcal{N}_{i,t}}$
DIAL: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}, m_{\mathcal{N}_{i,t}}, a_{i,t-1}\}$
CommNet: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}, \overline{m}_{\mathcal{N}_{i,t}}\}$
NeurComm: $I_{i,t} = \{o_{\mathcal{V}_{i,t}}, \pi_{\mathcal{N}_{i,t-1}}, m_{\mathcal{N}_{i,t}}\}$

在每一開始時(At the beginning of each episode,)，水廠以 w0 = 0.5 的水量開始，並在每個時間步以 wt+1 = wt + c 的恆定速率再生（我們在我們的實驗中使用不同的 c 值進行實驗） 實驗）。

我們系統中的代理總數為 N = 4，每個代理的觀察空間由一個二維實值向量組成(each agent's observation space consists of a two-dimensional real-valued vector containing)，其中包含值 wt 以及代理迄今為止輸送到其扇區的水量。

If the water is depleted, the episode ends. Each episode consists of a maximum of 1000 steps.
如果水耗盡，則情節結束。 每集最多包含 1000 個步驟。

Figure 2: Networked multi-agent reinforcement learning algorithms.
圖 2：聯網的多智能體強化學習算法。

(A) Decentralised networked multi-agent system.
(A) 去中心化網絡多代理系統。

Agents receive observations from the environment and share information with their neighbours.
代理接收來自環境的觀察結果並與鄰居共享信息。

(B) Agent module for networked non-communicative algorithms (NA2C, FPrint, ConseNet).
(B) 網絡非通信算法的代理模塊（NA2C、FPrint、ConseNet）。

The agent module receives neighbourhood observation, action and/or policy information in order to compute a local policy and state-value estimates.
代理模塊接收鄰域觀察、動作和/或策略信息以計算本地策略和狀態值估計。

(C) Agent module for networked communicative algorithms (DIAL, CommNet, NeurComm).
(C) 網絡通信算法的代理模塊（DIAL、CommNet、NeurComm）。

The agent module has an additional fully-connected (FC) layer for message passing along differentiable

communication channels.
代理模塊有一個額外的全連接 (FC) 層，用於沿著可微分通信通道傳遞消息。

Networked multi-agent reinforcement learning - 網絡化多智能體強化學習

我們網絡控制系統中的代理使用強化學習來學習如何從狀態映射到動作(reinforcement learning to learn how to map from states to actions)。

為了研究在網絡化 MARL 系統中利用不同信息結構的影響，我們考慮了 Chu et al. (2020) 最近工作中的六種最先進的方法。 誰調查了 MARL 在網絡系統控制中的使用，應用於自適應交通信號控制(adaptive traffic signal control)以及協作自適應巡航控制(adaptive cruise control)。

所有方法都利用具有深度神經網絡的優勢 actor-critic (A2C) 進行函數逼近 (Sutton et al., 2000; Mnih et al., 2016)。

第一組三個算法是非通信的，其中代理在執行和訓練期間僅限於觀察本地信息，但不明確通信：NA2C，這是 Lowe et al. (2017)提出的 MADDPG 算法的網絡 A2C 實現。 ; FPrint 具有策略指紋的 NA2C (Foerster et al., 2017)； 和 ConseNet, NA2C 以及 (Zhang et al., 2018) 提出的共識更新。

其餘三個是通信算法，其中代理既觀察本地信息又通過可區分的通信渠道進行明確的通信：DIAL (Foerster et al., 2016); CommNet (Sukhbaatar et al., 2016); and NeurComm (Chu et al., 2020)，其中 NeurComm 特別受到 Gilmer et al. (2017) 等人提出的通信協議的啟發。

最後，我們還考慮使用 A2C、IA2C (#2) 的完全獨立學習，其中所有代理都相互斷開並且沒有信息共享，類似於 IQL (Tan, 1993)。

###
#2
In Chu et al. (2020), NA2C is referred to as IA2C and our IA2C (fully independent/disconnected learning) was not considered in that work.
Chu et al. (2020)，NA2C 被稱為 IA2C，我們的 IA2C（完全獨立/斷開學習）在該工作中沒有考慮。

所有算法都是去中心化的，因為它們都沒有使用中心化的評論家或全球政策網絡。

However, the value estimates Vi,t from the critic network for each algorithm are conditioned on shared neighbourhood actions $a_{\mathcal{N}_i,t} = \{a_{j,t}\}_{j\in\mathcal{N}_i}$.

然而，來自每個算法的評論家網絡的價值估計 i;t 以共享鄰域動作 $a_{\mathcal{N}_i,t} = \{a_{j,t}\}_{j\in\mathcal{N}_i}$ 為條件。

圖 2 (A-C) 中提供了對上述每種算法的系統設計和網絡架構以及代理之間共享的信息類型的概要描

述。

系統設計人員使用 MARL 的一個關鍵動機是讓系統通過強化學習來學習其行為，並且潛在的自我發現解決方案比那些由手工設計的基於規則的系統設計和實現的解決方案效率更高。

此外，除了系統設計和實現方面的考慮，使用 MARL 的另一個論據基於以下假設(hypothesis.)。

假設 Autocurriculum hypothesis (Leibo et al., 2019)：多智能體系統中競爭與合作產生的動態提供了一種自然湧現的課程，其中一項任務的解決方案通常會導致新任務，不斷產生新的挑戰，從而促進創新 系統內。

有證據表明，這種假設對於強化學習代理可能是正確的，例如 Baker et al. (2019)，因此 MARL 可以在網絡系統控制的持續創新中發揮關鍵作用。

Empirical game-theoretic analysis (EGTA) - 經驗博弈論分析 (EGTA)

隨著多智能體系統越來越依賴於學習的複雜行為，它們也將變得更加難以分析。

此外，這些系統可以部署用於安全關鍵操作或管理關鍵的生命支持資源，如我們的水資源管理示例。

這使得理解驅動這些系統中代理行為和交互的潛在機制變得更加重要。

在這裡，我們使用經驗博弈論分析 (EGTA) (Walsh et al., 2002; Wellman, 2006; Tuyls et al., 2018)作為分析聯網 MARL 系統的工具。

博弈論 (von Neumann and Morgenstern, 1944) 關注與不同代理人在以遊戲表示的特定環境中採取的戰略行動相關的回報和激勵的數學研究。

然而，直接應用博弈論中的理論工具來分析複雜的跨期學習的多智能體系統被證明是一項非常具有挑戰性的任務。

在我們這裡考慮的一般和馬爾可夫博弈的背景下，博弈論中的經典分析工具通常更適合分析戰略形式的矩陣博弈，例如囚徒困境，即代理策略由原子級動作組成的博弈 直接標記為合作或背叛的，例如 坦白或保持沉默。(e.g. confess or stay quiet.)

因此，我們在這項工作中考慮的網絡控制系統代表了經典分析的重大挑戰，然而，我們能夠使用更現代的 EGTA 克服這一挑戰。

In EGTA, complex intertemporal strategic play in a multi-agent system becomes condensed into a meta-level game, where the atomic actions of the meta-level game correspond to learned agent policies.
在 EGTA 中，多智能體系統中復雜的跨期戰略遊戲被濃縮為元級遊戲，其中元級遊戲的原子動作對

應於學習的智能體策略。

為了實現這一點，我們遵循 Leibo et al. (2017)，我們調整環境 CPR 的再生率，以誘導與合作或背叛相關的學習行為，然後通過評估這些提取的策略來估計不同代理的平均收益(see steps A-C in Figure 3) .

Specifically, we consider the level of restraint an agent is able to display under different regeneration rates as an indicator of cooperation or defection.
具體來說，我們將代理在不同再生率下能夠顯示的約束水平作為合作或背叛的指標。

We define restraint as the percentage of time spent with a closed shut-off valve.
我們將約束定義為關閉截止閥所花費的時間百分比。

Under high regeneration rates, agents learn policies showing little restraint and continually extract from the resource, without any need to cooperate.
在高再生率下(Under high regeneration rates)，代理學習幾乎沒有約束的策略並不斷從資源中提取，無需任何合作。

We label these learned policies as defecting.
我們將這些學習到的政策標記為叛逃。

In contrast, under low regeneration rates, it is possible for agents to learn policies that display higher levels of restraint by not extracting from the resource at certain times, which requires a larger degree of cooperation so as to not have the resource deplete.
相比之下，在低再生率下，代理有可能通過在某些時候不從資源中提取來學習表現出更高水平約束的策略，這需要更大程度的合作，以免資源枯竭。

We label these learned policies, which display a high degree of restraint, as cooperative. Armed with these labelled policies, we consider a meta-level game, where binary choice analysis becomes possible.
我們將這些表現出高度克制的學習策略標記為合作的。 有了這些標記策略，我們考慮了一個元級別的遊戲，其中二元選擇分析成為可能。

In particular, by having agents with different strategic incentives interact during play, we can make use of a meta-level analysis to find potential equilibria as well as strategic inefficiencies within the system.
特別是，通過讓具有不同戰略激勵的代理在遊戲過程中進行交互，我們可以利用元級分析來發現系統內的潛在均衡和戰略低效率。

Intertemporally learned sequential social dilemmas (SSDs) - 跨期學習的順序社會困境（SSD）

我們的目標是確定可能的

networked system SSDs: operating conditions of a learned networked system under which at least one

inefficient equilibrium exists among all the potential equilibria identified within a system.
網絡系統 SSD：學習網絡系統的運行條件，在該條件下，系統內識別的所有潛在均衡中至少存在一個低效均衡。
That is, there exists at least one instance where all agents could be made better off or the collective total payoff made larger by improved system organisation and cooperation.
也就是說，至少存在一種情況，通過改進系統組織和合作，可以使所有代理人的狀況變得更好，或者集體總收益更大。

Similar to Perolat et al. (2017) and Hughes et al. (2018), we use EGTA combined with the binary choice analysis of Schelling (1973), to distill the complex interaction of a multi-agent system into a meta-level game from which the potential equilibria can easily be obtained.
類似於 Perolat et al. (2017)和 Hughes et al. (2018)，我們使用 EGTA 結合 Schelling (1973) 的二元選擇分析，將多智能體系統的複雜交互提煉為元級博弈(a meta-level game)，從中可以輕鬆獲得潛在的均衡。

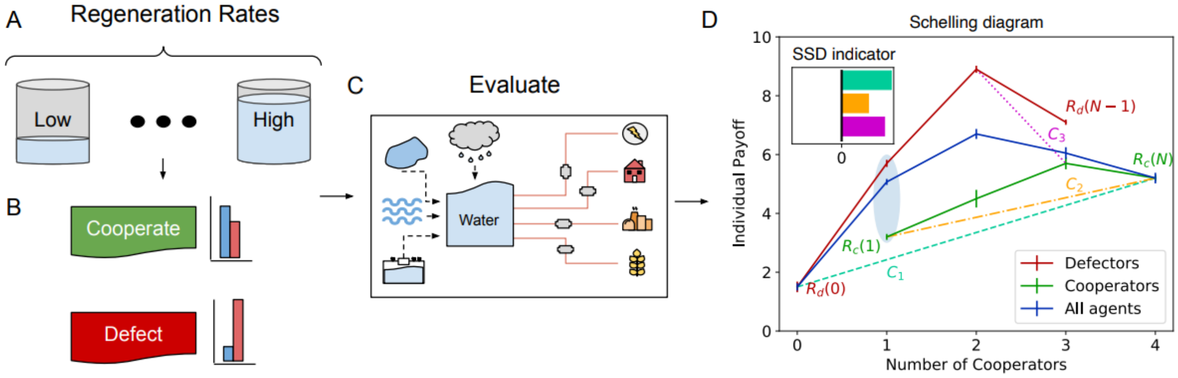


Figure 3: *Empirical game-theoretic analysis pipeline*. **(A)** Agents train on a common resource pool under different regeneration rates, where they may act greedily or show restraint. **(B)** Cooperative policies show restraint, while defecting policies act greedily. **(C)** Evaluate sampled policies on the environment. **(D)** Plot the Schelling diagram, which can be used to visually classify systems.

為了實現這一點，我們專門使用了 Schelling diagrams (Schelling, 1973)。

謝林圖(A Schelling diagram)顯示了單個代理選擇合作或缺陷的平均收益，作為選擇合作的代理總數的函數。

圖 3 的步驟 4 中顯示了此類圖表的示例，其中 x 軸(x-axis)表示選擇合作的代理總數。

Concretely, if there are N = 4 agents and x = 2 are cooperating, this implies that N - x = 2 of the other agents are defecting.
具體而言，如果有 N = 4 個代理並且 x = 2 個正在合作，這意味著 N⍰x = 2 個其他代理正在叛逃。

此時，圖 3 中的謝林圖(the Schelling diagram)將具有三個 y 值：
the average payoff (in a system with two cooperators and two defectors) for an agent choosing to either (1)

cooperate (green line denoted Rc(x)) or (2) defect (red line denoted Rd(x)) as well as (3) the average (blue line) for all agents (cooperators and defectors).
代理選擇（1）合作（綠線表示為Rc(x)）或（2）缺陷（紅線表示為Rd(x)）以及(3)所有代理（合作者和叛逃者）的平均值（藍線）。

To identify an SSD, we use similar criteria as in (Macy and Flache, 2002; Leibo et al., 2017; Hughes et al., 2018), but collapse the specific conditions referred to as “fear” and “greed” in prior work, into a single criterion.(#3)
為了識別SSD，我們使用與(Macy and Flache, 2002; Leibo et al., 2017; Hughes et al., 2018)類似的標準，但取消了之前稱為“恐懼”(“fear”)和“貪婪”(“greed”)的特定條件 工作，變成一個單一的標準。(#3）

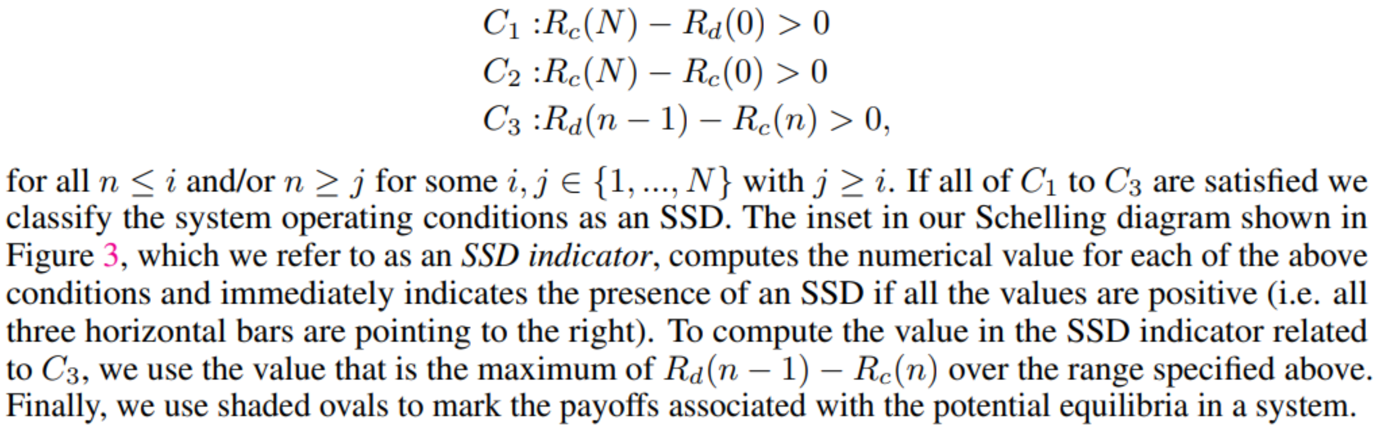
為此，我們利用了以下三個需要滿足的條件。

Condition 1: full system cooperation is preferred to full system defection;
條件1：全系統合作優於全系統叛逃；

Condition 2: full system cooperation is preferred to being exploited by defectors; and
條件2：全系統合作優先於被叛逃者利用；

Condition 3: for a certain range of system configurations there is a stronger reward-driven incentive to defect than to cooperate.
條件3：對於一定範圍的系統配置，有比合作更強的獎勵驅動的背叛動機。

這些條件總結如下：



$$C_1 : R_c(N) - R_d(0) > 0$$
$$C_2 : R_c(N) - R_c(0) > 0$$
$$C_3 : R_d(n-1) - R_c(n) > 0,$$

for all $n \leq i$ and/or $n \geq j$ for some $i, j \in \{1, ..., N\}$ with $j \geq i$. If all of C_1 to C_3 are satisfied we classify the system operating conditions as an SSD. The inset in our Schelling diagram shown in Figure 3, which we refer to as an *SSD indicator*, computes the numerical value for each of the above conditions and immediately indicates the presence of an SSD if all the values are positive (i.e. all three horizontal bars are pointing to the right). To compute the value in the SSD indicator related to C_3, we use the value that is the maximum of $R_d(n-1) - R_c(n)$ over the range specified above. Finally, we use shaded ovals to mark the payoffs associated with the potential equilibria in a system.

for all n <= i and/or n>=j for some i, j ∈ {1, …,N} with j >= i. If all of C1 to C3 are satisfied we classify the system operating conditions as an SSD.
對於所有n <= i和/或n >= j對於某些i, j ∈ {1, …,N} with j >= i。如果所有C1到C3都得到滿足，我們將系統運行條件歸類為SSD。

圖3所示謝林圖中的插圖，我們將其稱為SSD指標，計算上述每個條件的數值，如果所有值均為正（即所有三個水平條），則立即指示存在SSD都指向右邊）。

To compute the value in the SSD indicator related to C3, we use the value that is the maximum of Rd(n - 1) - Rc(n) over the range specified above.
為了計算與 C3 相關的 SSD 指標中的值，我們使用 Rd(n - 1) -Rc(n) 在上述指定範圍內的最大值。

最後，我們使用帶陰影的橢圓來標記與系統中潛在均衡相關的收益。

###
#3
In the context of an engineering system, these two conditions would seem an anthropomorphisation of the incentive to defect.
在工程系統的背景下，這兩個條件似乎是對缺陷動機的擬人化。

3 Results 結果

我們分析的第一階段涉及提取在不同 CPR 再生率下學習的策略，c = { 0.1, 0.088, 0.077, 0.065, 0.053, 0.042, 0.03}，使用我們的水管理環境。

The values in the heatmap shown in Figure 4 are the average percentage restraint displayed by the tested MARL algorithms for different regeneration rates, with a neighbourhood reward weighting set at α = 0.1 (we consider α ={ 0. 1} in the supplementary material).

圖 4 中顯示的熱圖中的值是經過測試的 MARL 算法針對不同再生率顯示的平均百分比約束，鄰域獎勵權重設置為 α = 0.1（我們在補充材料中考慮 α ={ 0. 1}）。

From a system design perspective, a value of α = 0.1 ensures that agents remain focused on providing water to their respective sectors, while at the same time taking into consideration the water needs of their connected neighbourhood.
從系統設計的角度來看，α = 0.1 的值可確保代理繼續專注於為其各自的部門供水，同時考慮其相連社區的用水需求。

As expected, agents show less restraint at higher regeneration rates.
正如預期的那樣，代理在較高的再生率下表現出較少的約束。

However, at lower regeneration rates, differences start to emerge.
然而，在較低的再生率下，差異開始出現。

This can be seen mostly between the algorithms that employ differentiable communication protocols and those that do not.
這主要體現在採用可微通信協議的算法和不採用可微通信協議的算法之間。

Specifically, communicating algorithms seem to show less restraint at lower regeneration rates compared

to other approaches, which could possibly be attributed to more sophisticated cooperation strategies.
具體而言，與其他方法相比，通信算法似乎在較低的再生率下表現出較少的約束，這可能歸因於更複雜的合作策略。

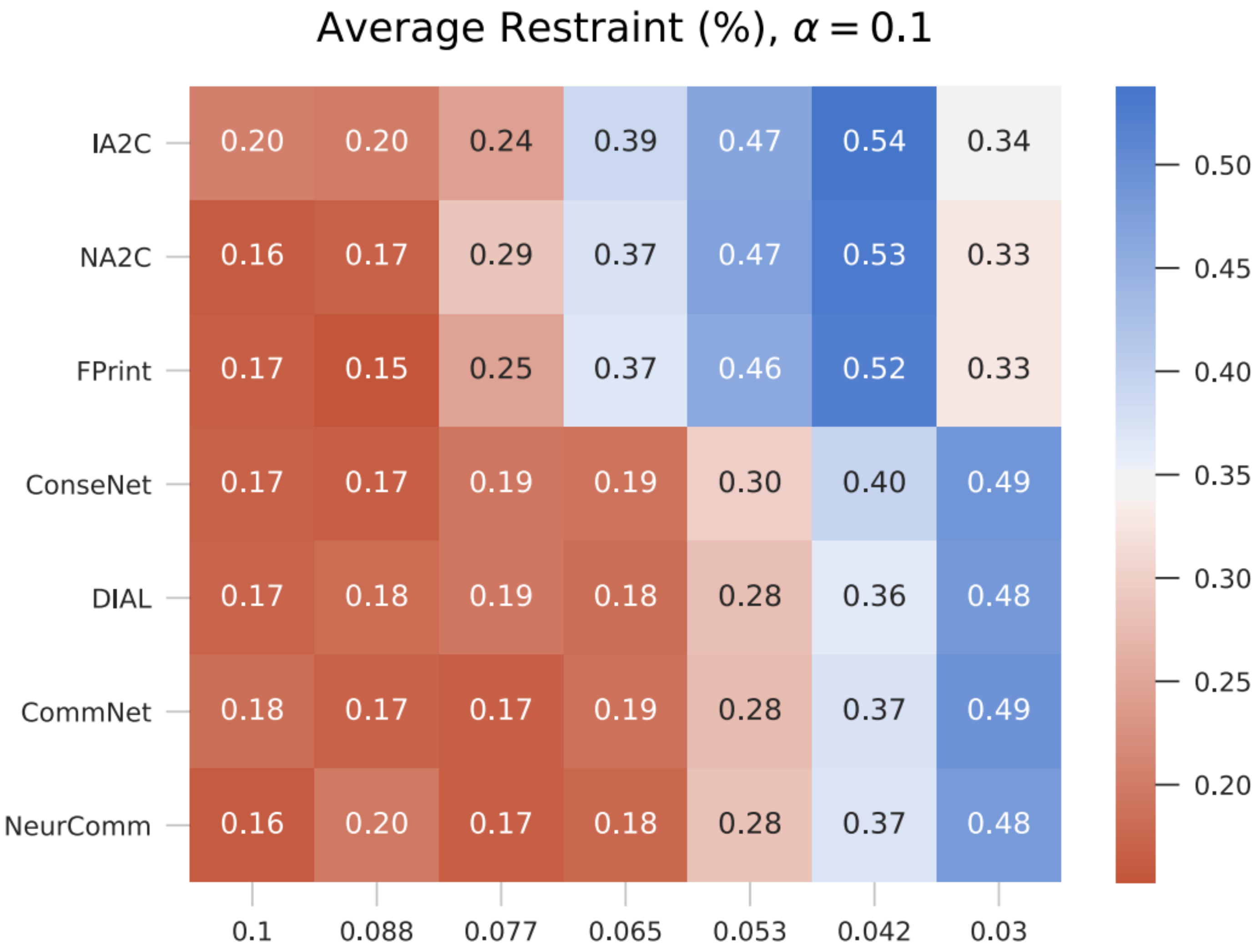
In contrast, for algorithms that do not communicate, but only directly share local information, it might prove more difficult to coordinate agent behaviour in the low resource setting and as a result each agent is forced to show a higher level of restraint to keep the resource from the depleting.
相比之下，對於不進行通信而僅直接共享本地信息的算法，在低資源設置中協調代理行為可能會更加困難，因此每個代理被迫表現出更高水平的約束以保留資源 從枯竭。

That said, we also observe a marked decrease in restraint between the rate = 0.042 (second to last column) and the rate = 0.03 (last column) for three of the non-communicating algorithms (IA2C, NA2C and FPrint).
也就是說，對於三種非通信算法（IA2C、NA2C 和 FPrint），我們還觀察到 rate = 0.042（倒數第二列）和 rate = 0.03（最後一列）之間的約束顯著下降。

This shows that at an extreme level of scarcity, these agents find it difficult to learn restraint, leading to myopic behaviour and the tragedy of the commons to manifest.
這表明，在極度稀缺的情況下，這些智能體(learn restraint)很難學會克制，從而導致近視行為和公地悲劇的表現。

In the second stage of our analysis, we classified policies into two sets: cooperating or defecting, based on their level of restraint.
在我們分析的第二階段，我們根據政策的克製程度將政策分為兩組：合作或背叛。

We then estimated the expected payoffs related to each policy combination in our four-agent water management system, with a constant regeneration rate of c = 0.055 and 100 steps per episode.
然後我們估計了與我們的四代理水管理系統中每個策略組合相關的預期收益(each policy combination in our four-agent water management system)，恆定的再生率 c = 0.055 和每集 100 步。



At this rate of regeneration, we labelled policies as defecting if they had a percentage of restraint below 25% and cooperating if this percentage was instead above 35%.
按照這種再生速度，如果政策的克制百分比低於 25%，我們將其標記為叛逃，如果該百分比高於 35%，則我們將其標記為合作。

The results of our analysis are presented in Figure 5 (A-G) in the form of Schelling diagrams, one for each algorithm, from which the game theoretic solution concepts can be visually obtained.
我們的分析結果以謝林圖的形式呈現在圖 5 (A-G) 中，每種算法一個，從中可以直觀地獲得博弈論解決方案的概念。

我們對這些圖表的參數化是根據代理（合作或背叛）相對於合作者總數（在 x 軸上）的潛在回報（顯示在 y 軸上），如 Perolat et al. (2017) 然而，我們還使用 x 軸上其他合作者的數量提供了替代參數化，如 Hughes et al. (2018)，在補充材料中。

Figure 4: EGTA for networked system control, with α = 0:1.
圖 4：用於網絡系統控制的 EGTA，α = 0:1。

Heatmap of average restraint percentage as a function of the regeneration rate for different MARL

algorithms, from high (0:1) to low (0:03).
作為不同 MARL 算法的再生率函數的平均約束百分比熱圖，從高 (0.1) 到低 (0.03)。

The shaded blue ovals in the Schelling diagrams of Figure 5 mark the various payoffs for cooperating and defecting agents, associated with potential Nash equilibria for each system.
圖 5 謝林圖中的藍色陰影橢圓標誌著合作和背叛代理人的各種回報，與每個系統的潛在納什均衡相關。

For example, in IA2C shown in panel (A), there exist two Nash equilibrium points.
例如，在圖 (A) 所示的 IA2C 中，存在兩個納什均衡點。

These are best response strategies from the perspective of each individual agent, given the strategies of all other agents.
考慮到所有其他代理的策略，從每個代理的角度來看，這些是最佳響應策略。

At the first equilibrium point, the system consists of two cooperating and two defecting agents.
在第一個平衡點，系統由兩個合作代理和兩個背叛代理組成。

In this situation, a cooperating agent will not receive a higher payoff for switching to defect, and neither will a defecting agent by switching to cooperate.
在這種情況下，合作的代理人不會因為轉向背叛而獲得更高的回報，而背叛的代理人也不會因為轉向合作而獲得更高的回報。

However, note that the global expected payoff is not optimal in this case and had all agents instead started out by cooperating (represented by the second equilibrium point with payoffs shown at the top right of the plot), agents would have no incentive to switch and at the same time achieve the global maximum expected payoff for the system as a whole.
但是，請注意，在這種情況下，全局預期收益並不是最優的，而是讓所有代理開始合作（由第二個均衡點表示，收益顯示在右上方） 情節），代理將沒有動力切換，同時實現整個系統的全局最大預期收益。

That said, both these equilibrium points are unstable.
也就是說，這兩個平衡點都是不穩定的。

For instance, if the system was to be perturbed towards a configuration of three cooperators and one defector, it is unclear whether the system would return to the same equilibrium point.
例如，如果系統受到三個合作者和一個叛逃者的配置的干擾，則係統是否會回到相同的平衡點尚不清楚。

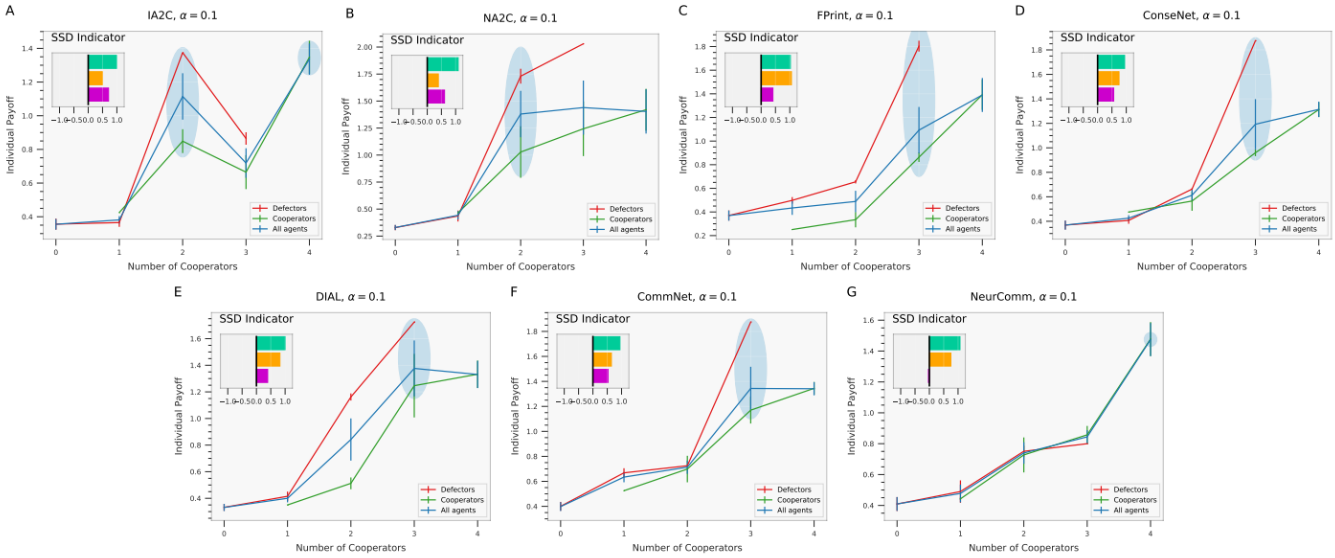


Figure 5: *EGTA for networked system control, with* $\alpha = 0.1$.**(A-G)** Schelling diagrams for each approach with sequential social dilemma (SSD) indicators given as insets. Potential equilibria are shaded in blue.

Figure 5: EGTA for networked system control, with α = 0.1.(A-G) Schelling diagrams for each approach with sequential social dilemma (SSD) indicators given as insets.

圖 5：用於網絡系統控制的 EGTA，= 0:1。（A-G）每種方法的謝林圖，序列社會困境 (SSD) 指標作為插圖給出。

Potential equilibria are shaded in blue.

電位平衡用藍色陰影表示。

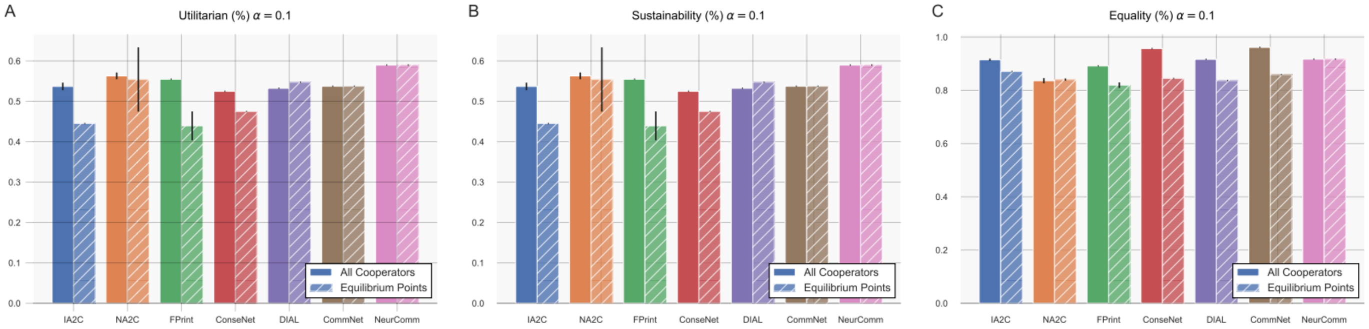


Figure 6: *Social metrics at equilibrium*. **(A)** Utilitarian **(B)** Equality **(C)** Sustainability

Figure 6: Social metrics at equilibrium. (A) Utilitarian (B) Equality (C) Sustainability

圖 6：處於均衡狀態的社會指標。 (A) 功利 (B) 平等 (C) 可持續性

Although Nash equilibria are useful for investigating optimal best response strategies from an individual agent perspective, with α = 0.1, the reward for each agent in our case depends on the rewards of its entire neighbourhood.

儘管納什均衡對於從單個代理的角度研究最佳響應策略很有用，$\alpha = 0.1$，但在我們的案例中，每個代理的獎勵取決於其整個鄰域的獎勵。

Therefore, our interest mostly concerns the expected payoff of the system, represented by the average payoff for all agents (the blue lines in Figure 5).
因此，我們的興趣主要關注系統的預期收益，由所有代理的平均收益表示（圖 5 中的藍線）。

Even though many of the equilibria in Figure 5 are inefficient from this perspective, we find that for communicating algorithms their equilibria do in fact coincide with what is optimal for the system as a whole (see B,E and F).
儘管從這個角度來看，圖 5 中的許多均衡是低效的，但我們發現，對於通信算法，它們的均衡實際上與整個系統的最佳均衡一致（參見 B、E 和 F）。

However, in the case of DIAL (E) and CommNet (F) we still consider these equilibria to be inefficient as indicated by our SSD conditions.
然而，在 DIAL (E) 和 CommNet (F) 的情況下，我們仍然認為這些均衡是低效的，如我們的 SSD 條件所示。

This is because if we make the assumption that all water demanding sectors of society are considered as being equally important, these equilibria still represent an unequal distribution of the resource (the difference between cooperators and defectors).
這是因為如果我們假設社會的所有需水部門都被視為同等重要，那麼這些均衡仍然代表著資源的不平等分配（合作者和叛逃者之間的差異）。

Interestingly, NeurComm (G) is the only non-SSD among all the algorithms tested (C3 < 0, i.e. there exist no reward-driven incentives for agents to defect).
有趣的是，NeurComm (G) 是所有測試算法中唯一的非 SSD（C3 < 0，即不存在獎勵驅動的代理缺陷激勵）。

That is, NeurComm is the only algorithm able to achieve a stable equilibrium point, where what is optimal for each individual agent is also optimal for the entire system.
也就是說，NeurComm 是唯一能夠達到穩定平衡點的算法，其中對每個個體最優的也是對整個系統最優的。

In Figure 6, we compute the metrics from Perolat et al. (2017), namely, the Utilitarian, Equality and Sustainability metric for each algorithm for the two cases where (1) all agents cooperate and (2) agents play their equilibrium strategy.
在圖 6 中，我們計算了 Perolat et al. (2017)，即每個算法的功利、平等和可持續性度量，適用於 (1) 所有代理合作和 (2) 代理髮揮其均衡策略的兩種情況。

When all agents are cooperating, the equality for NeurComm is lower (i.e. the distribution of reward among agents is less uniform) than that of the other communicating algorithms, however, it is still able to achieve

both a higher score for the group (more utilitarian) and have higher levels of sustainability.
當所有代理都合作時，NeurComm 的平等性低於其他通信算法（即代理之間的獎勵分配不那麼均勻），但是，它仍然能夠為群體獲得更高的分數（更實用 ）並具有更高水平的可持續性。

This provides some supporting evidence that NeurComm is perhaps better able to learn how to coordinate effectively.
這提供了一些支持證據，表明 NeurComm 或許能夠更好地學習如何有效地協調。

4 Discussion 討論

網絡化多智能體強化學習已被證明是解決大規模控制問題的可行選擇。

然而，由於學習到的多智能體系統的複雜性，深入了解它們的操作行為可能被證明是困難的。

此外，這些系統可能會越來越多地部署在廣泛的重要環境中，這使得它們的理解變得更加重要。

在這項工作中，我們考慮了公共池資源管理的設置。

此類控制問題與社會面臨的一些最大挑戰有關，包括糧食安全、不平等和氣候變化。

我們對用於公共池資源管理的網絡多代理系統進行了經驗博弈論分析。

該分析強調了由這些系統中使用的不同信息結構引起的解決方案概念的差異。

Specifically, we found that differentiable communication protocols play an important role in driving the system dynamics to equilibrium points associated with optimal payoffs, both for the individual and the system as a whole.
具體來說，我們發現可微通信協議在將系統動態驅動到與最佳收益相關的平衡點方面發揮著重要作用，無論是對於個人還是整個系統。

However, in terms of maintaining an equitable distribution of resources amongst agents, most of these equilibria were still regarded as being inefficient.
然而，就保持代理人之間資源的公平分配而言，這些均衡中的大多數仍然被認為是低效的。

In the end, NeurComm (Chu et al., 2020), a newly proposed networked multi-agent reinforcement learning algorithm, was the only approach able to obtain an optimal equilibrium that was both stable and efficient.
最後，新提出的網絡多智能體強化學習算法 NeurComm (Chu et al., 2020) 是唯一能夠獲得既穩定又高效的最優平衡的方法。

Finally, in addition to highlighting the interaction between system design and behaviour, we consider this work as demonstrating a viable evaluation pipeline for complex multi-agent systems beyond efficiency, robustness, scalability and mean control performance.

最後，除了強調系統設計和行為之間的相互作用外，我們認為這項工作證明了複雜多代理系統的可行評估管道，超越了效率、魯棒性、可擴展性和平均控制性能。

Broader Impact
更廣泛的影響

It could be extremely costly should a critical multi-agent system completely fail.
如果一個關鍵的多代理系統完全失敗，這可能會非常昂貴。

However, in this work, we have shown that even in a very simplified environment, a seemingly working system might convergence towards equilibriums that are still inefficient or unequal.
然而，在這項工作中，我們已經表明，即使在一個非常簡化的環境中，一個看似有效的系統也可能會收斂到仍然低效或不平等的均衡狀態。

這些結果顯然也是不受歡迎的，但它們並沒有為系統故障提供明確的信號。

Given the complexity of MARL system operation, failure modes related to inefficiencies at equilibrium are more subtle, and more difficult to detect.
鑑於 MARL 系統操作的複雜性，與平衡低效相關的故障模式更微妙，更難以檢測。

We consider our work a contribution towards better understanding networked multi-agent reinforcement learning systems for common-pool resource management.
我們認為我們的工作有助於更好地理解用於公共池資源管理的網絡化多智能體強化學習系統。

As mentioned in the main text, we envision these systems becoming more widespread in their use as effective control systems for managing critical life-supporting resources.
正如正文中提到的，我們設想這些系統在用作管理關鍵生命支持資源的有效控制系統方面將變得更加廣泛。

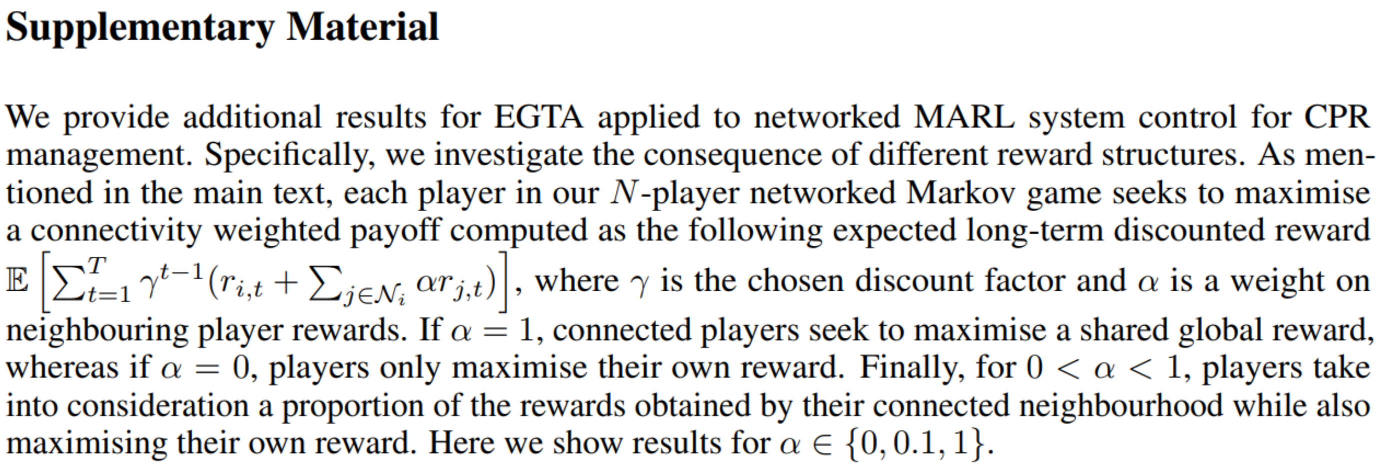
However, a specific challenge facing future systems is to have them be highly adaptable.
然而，未來系統面臨的一個具體挑戰是讓它們具有高度的適應性。

Multi-agent reinforcement learning offers this capability, but it also makes systems far more difficult to analyse and understand.
多智能體強化學習提供了這種能力，但它也使系統更難以分析和理解。

Therefore, even though we consider our environment a considerable simplification over a real institutional level CPR management system, we nevertheless hope the broader impact of our work is to demonstrate EGTA as a viable approach to analysing practical multi-agent reinforcement learning systems.
因此，儘管我們認為我們的環境相對於真正的機構級別的 CPR 管理系統來說是相當簡單的，但我們仍然希望我們工作的更廣泛影響是證明 EGTA 作為分析實際多智能體強化學習系統的可行方法。

Supplementary Material

補充材料



Supplementary Material

We provide additional results for EGTA applied to networked MARL system control for CPR management. Specifically, we investigate the consequence of different reward structures. As mentioned in the main text, each player in our N-player networked Markov game seeks to maximise a connectivity weighted payoff computed as the following expected long-term discounted reward $\mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}(r_{i,t}+\sum_{j\in\mathcal{N}_i}\alpha r_{j,t})\right]$, where γ is the chosen discount factor and α is a weight on neighbouring player rewards. If $\alpha=1$, connected players seek to maximise a shared global reward, whereas if $\alpha=0$, players only maximise their own reward. Finally, for $0<\alpha<1$, players take into consideration a proportion of the rewards obtained by their connected neighbourhood while also maximising their own reward. Here we show results for $\alpha\in\{0,0.1,1\}$.

We provide additional results for EGTA applied to networked MARL system control for CPR management.

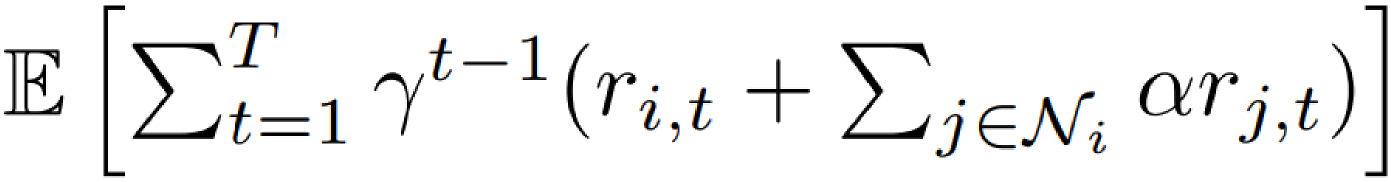
我們為應用於 CPR 管理的網絡 MARL 系統控制的 EGTA 提供了額外的結果。

Specifically, we investigate the consequence of different reward structures.

具體來說，我們研究了不同獎勵結構的後果。

As mentioned in the main text, each player in our N-player networked Markov game seeks to maximise a connectivity weighted payoff computed as the following expected long-term discounted reward

正如正文中提到的，我們的 N 人網絡馬爾可夫遊戲中的每個玩家都尋求最大化連接加權收益，計算為以下預期的長期折扣獎勵 E ...



$$\mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}(r_{i,t}+\sum_{j\in\mathcal{N}_i}\alpha r_{j,t})\right]$$

, where γ is the chosen discount factor and α is a weight on neighbouring player rewards.

, ，其中γ 是選擇的折扣因子，α是相鄰節點的權重 玩家獎勵。

If α = 1, connected players seek to maximise a shared global reward, whereas if α = 0, players only maximise their own reward.

如果α = 1，連接的玩家尋求最大化共享的全局獎勵，而如果 α = 0，玩家只最大化他們自己的獎勵。

Finally, for 0 < α < 1, players take into consideration a proportion of the rewards obtained by their connected neighbourhood while also maximising their own reward.

最後，當 0 < α < 1 時，玩家在最大化自己的獎勵的同時，還要考慮他們所連接的鄰域獲得的一定

比例的獎勵。

Here we show results for $\alpha \in \{0,0.1,1\}$.
這裡我們展示了 α∈{0,0.1,1} 的結果。

Restraint percentages under different regeneration rates
不同再生率下的約束百分比

The heatmaps in Figure 7 (A-C) highlight the differences in restraint percentage for different values of as the regeneration rate is changed from high (0.1) to low (0.03).
圖 7 (A-C) 中的熱圖突出顯示了不同值的約束百分比差異，因為再生率從高 (0.1) 變為低 (0.03)。

In the case where agents are completely self-interested (α = 0) shown in (A), the majority of algorithms without communication display very low levels of restraint for all rates of regeneration.
在 (A) 中顯示的代理完全自利 (α = 0) 的情況下，大多數沒有通信的算法對所有再生率都顯示出非常低的約束水平。

To some degree, this could be seen as a manifestation of the tragedy of the commons where the equilibrium strategy for self-interested agents is to have zero restraint, especially when the regeneration rate is low, i.e. 0.042 or 0.03.
在某種程度上，這可以看作是公地悲劇的體現，其中自利代理的均衡策略是零約束，尤其是當再生率較低時，即 0.042 或 0.03。

In contrast, when connected agents completely share their reward (α = 1), shown in (C), all algorithms display lower levels of restraint when the regeneration rate is high, and higher levels of restraint when the regeneration rate is low.
相比之下，當連接的智能體完全共享他們的獎勵（α = 1）時，如（C）所示，所有算法在再生率高時表現出較低水平的約束，而在再生率低時表現出較高水平的約束。

However, there is still some difference in the level of restraint between agents that communicate and those that do not, with the former showing slightly less restraint for lower levels of regeneration than the latter.
然而，交流的代理和不交流的代理之間的約束水平仍然存在一些差異，前者對較低再生水平的約束略低於後者。

This could possibly be attributed to better coordination as well as cooperation between agents that communicate, making them able to extract more of the resource closer to the limit of its capacity without depleting the resource completely.
這可能歸因於通信代理之間更好的協調和合作，使他們能夠在不完全耗盡資源的情況下提取更多接近其容量極限的資源。

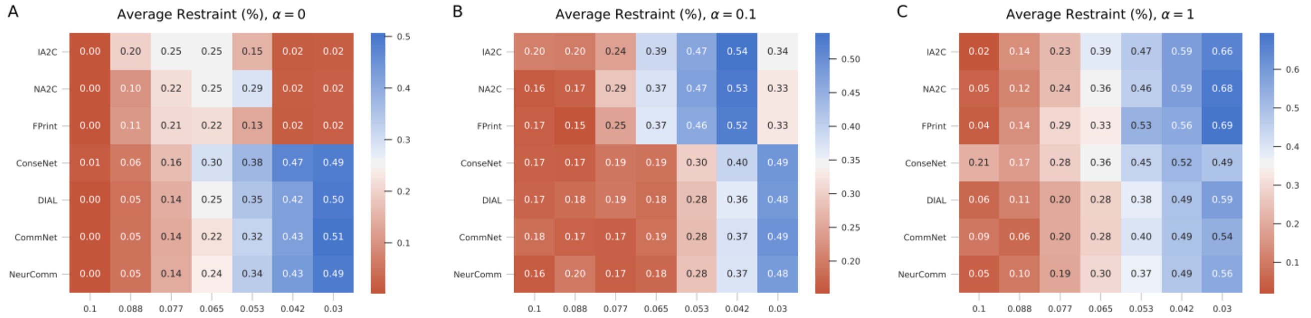


Figure 7: Heatmaps of average restraint percentage as a function of the regeneration rate for different MARL algorithms, from high (0.1) to low (0.03). **(A)** $\alpha = 0$, **(B)** $\alpha = 0.1$, **(C)** $\alpha = 1$.

Figure 7: Heatmaps of average restraint percentage as a function of the regeneration rate for different MARL algorithms, from high (0:1) to low (0:03). (A) α = 0, (B) α = 0.1, (C) α = 1.
圖 7：從高 (0:1) 到低 (0:03) 的不同 MARL 算法的平均約束百分比作為再生率函數的熱圖。（一種）(A) α = 0, (B) α = 0.1, (C) α = 1.

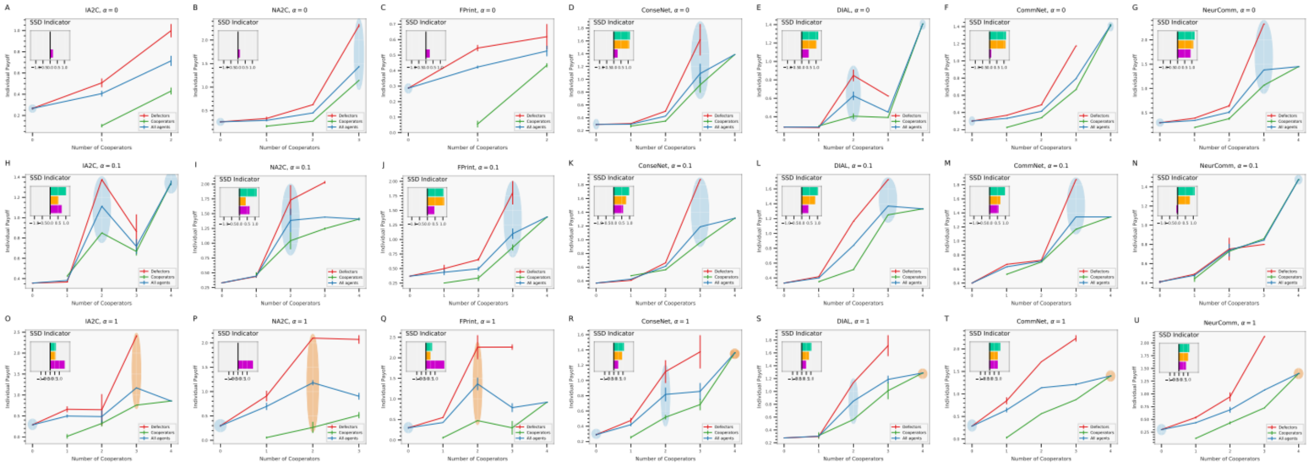


Figure 8: Schelling diagrams for each approach with network system sequential social dilemma (SSD) indicators given as insets. Potential Nash equilibria are shaded in blue. **Top row (A-G)** $\alpha = 0$, **Middle row (H-N)** $\alpha = 0.1$, **Bottom row (O-U)** $\alpha = 1$. Here we include orange shaded regions indicating configurations corresponding to the highest average payoff for all connected agents.

Figure 8: Schelling diagrams for each approach with network system sequential social dilemma (SSD) indicators given as insets.
圖 8：每種方法的謝林圖，其中以插圖形式給出網絡系統順序社會困境 (SSD) 指標。

Potential Nash equilibria are shaded in blue.
潛在的納什均衡以藍色陰影表示。

Top row (A-G) α = 0, Middle row (H-N) α = 0.1, Bottom row (O-U) α = 1.

Here we include orange shaded regions indicating configurations corresponding to the highest average

payoff for all connected agents.
在這裡，我們包括橙色陰影區域，指示對應於所有連接代理的最高平均收益的配置。

Schelling binary choice analysis for different α values
不同值的謝林二元選擇分析

We performed the N-player binary choice analysis of Schelling (1973) for the different reward structures corresponding to $\alpha \in \{0,0.1,1\}$ and plot the Schelling diagrams for each value of α in Figure 8.
我們對與 α∈{0,0.1,1} 對應的不同獎勵結構進行了 Schelling (1973) 的 N 玩家二元選擇分析，並繪製了圖 8 中每個 α 值的 Schelling 圖。

The diagrams in the top row, panels (A-G), are for the different algorithms with α = 0 (self-interested agents), the middle row (H-N) with α = 0.1 (as in the main text) and the bottom row (O-U) with α = 1 (global shared reward for connected agents).
頂行中的圖表，面板 (AG)，針對不同算法，α = 0（自利代理），中間行 (HN)，α = 0.1（如正文）和底行（ OU），α = 1（連接代理的全局共享獎勵）。

For self-interested agents (α = 0) without communication, the insensitivity to the regeneration rate can cause the restraint threshold for classifying agents as cooperative or defective, to never be low enough to obtain all possible configurations of agents.
對於沒有通信的自利代理（ α = 0），對再生率的不敏感會導致將代理分類為合作或缺陷的約束閾值永遠不會低到足以獲得所有可能的代理配置。

This can be seen in the top row, panels (A-C), where for IA2C, NA2C and FPrint there were no instances where all agents could be considered as being cooperative.
這可以在頂行的面板 (A-C) 中看到，其中對於 IA2C、NA2C 和 FPrint，沒有任何實例可以將所有代理視為合作。

However, even for the communicating algorithms where cooperation is seen to emerge more easily, the majority of potential equilibria are still inefficient.
然而，即使對於合作被認為更容易出現的通信算法，大多數潛在的均衡仍然是低效的。

In fact, only DIAL and CommNet have potential equilibria points that correspond to full system cooperation with expected payoffs that are optimal for the individual as well as the group.
事實上，只有 DIAL 和 CommNet 具有潛在的平衡點，對應於全系統合作，預期收益對個人和團體都是最佳的。

Also worth noting is that the equilibrium profile for ConseNet is similar to the communicating algorithms, DIAL, CommNet and NeurComm (across all values of α), which is likely due to its consensus update mechanism.

同樣值得注意的是，ConseNet 的均衡分佈類似於通信算法 DIAL、CommNet 和 NeurComm（跨所有 α 值），這可能是由於其共識更新機制。

The Schelling diagrams for the different algorithms with connected agents sharing a global reward (α = 1) are shown in the bottom row of Figure 8.
圖 8 的底行顯示了具有共享全局獎勵 (α = 1) 的連接代理的不同算法的謝林圖。

The orange ovals in these diagrams indicate which system configurations correspond to the highest expected payoff for all agents.
這些圖中的橙色橢圓表示哪些系統配置對應於所有代理的最高預期收益。

In the case of non-communicative algorithms, IA2C, NA2C and FPrint, agents received on average the highest payoff when the system consisted of a mixture of cooperative, as well as defective, agents.
在非通信算法 IA2C、NA2C 和 FPrint 的情況下，當系統由合作代理和有缺陷代理的混合組成時，代理平均獲得最高的回報。

In contrast, ConseNet and the communicating algorithms, DIAL, CommNet and NeurComm, had their highest payoffs coincide with systems operating at full cooperation.
相比之下，ConseNet 和通信算法 DIAL、CommNet 和 NeurComm 的最高回報與完全合作運行的系統一致。

Schelling diagrams using a different parameterisation
使用不同參數化的謝林圖

An alternative parameterisation for a Schelling diagram is to plot payoffs for a particular agent (cooperating or defecting) with respect to the number of other cooperators on the x-axis, instead of the total number of cooperators.
謝林圖的另一種參數化是繪製特定代理（合作或背叛）相對於 x 軸上其他合作者數量而不是合作者總數的回報。

We find the latter (which we use in the main text) more suitable for highlighting payoffs associated with potential equilibria, but note that the former provides an easier visual interpretation of dominant strategies for any given situation.
我們發現後者（我們在正文中使用）更適合突出與潛在均衡相關的收益，但請注意，前者為任何給定情況下的主導策略提供了更簡單的視覺解釋。

We provide this version of the diagram for each algorithm for the case of α = 0.1 in Figure 9.
對於圖 9 中 α = 0.1 的情況，我們為每種算法提供了此版本的圖表。

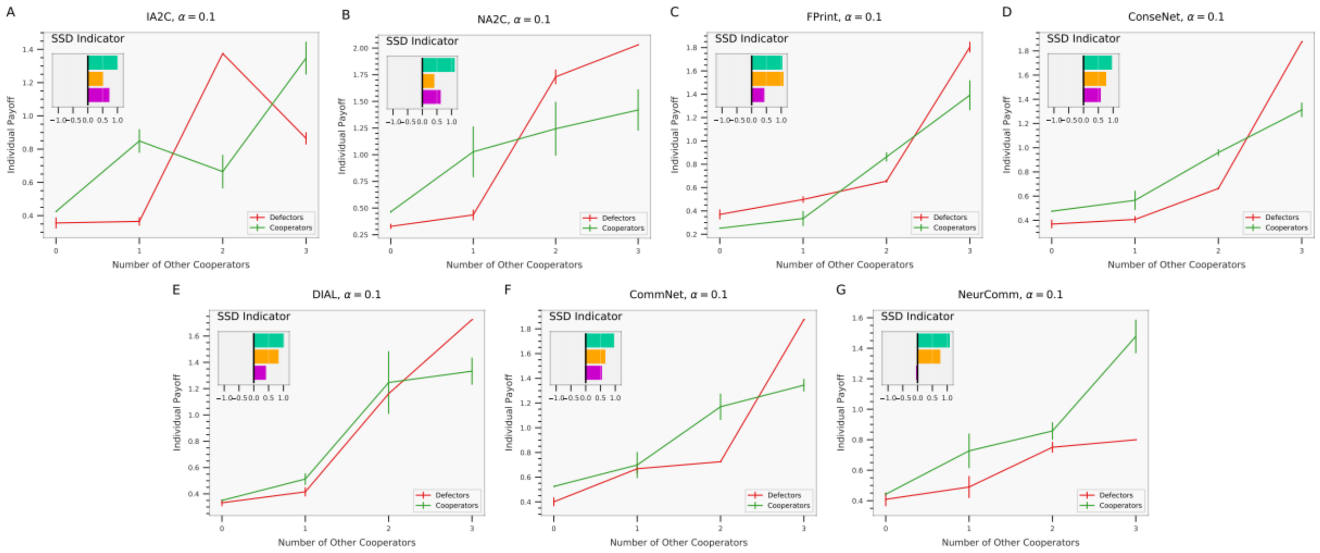


Figure 9: *EGTA for networked system control ($\alpha = 0.1$) with the number of other cooperators shown on the x-axis.***(A-G)** Schelling diagrams for each approach with sequential social dilemma (SSD) indicators given as insets.

Figure 9: EGTA for networked system control (α = 0.1) with the number of other cooperators shown on the x-axis.(A-G) Schelling diagrams for each approach with sequential social dilemma (SSD) indicators given as insets.

圖 9：用於網絡系統控制的 EGTA (α = 0.1)，x 軸上顯示了其他合作者的數量。（A-G）每種方法的謝林圖，序列社會困境 (SSD) 指標作為插圖給出。

A trend easily observed using this parameterisation is that for most algorithms the dominant strategy for a learned agent is to cooperate until all agents are cooperating, where at this point, the dominant strategy switches to defect.

使用這種參數化很容易觀察到的一個趨勢是，對於大多數算法，學習代理的主導策略是合作，直到所有代理都合作，此時，主導策略切換到缺陷。

The only exception is the NeurComm algorithm, which is shown that have cooperation as the dominant strategy for any configuration of the system.

唯一的例外是 NeurComm 算法，該算法表明，對於系統的任何配置，都將合作作為優勢策略。

Finite sample analysis using bootstrap estimation

使用 bootstrap 估計的有限樣本分析

It is possible to connect our analysis to the underlying Markov game.

可以將我們的分析與基礎馬爾可夫博弈聯繫起來。

More specifically, a key result for EGTA is given by the finite sample analysis in Tuyls et al. (2018), which states that given enough samples it is possible to bound the difference between the empirical equilibrium payoff estimates and those obtained in the original game.

更具體地說，EGTA 的一個關鍵結果是由 Tuyls et al. (2018) 的有限樣本分析給出的。其中指出，給定足夠的樣本，可以限制經驗均衡收益估計與原始博弈中獲得的差異之間的差異。

We used this result combined with bootstrap resampling to more tightly bound our estimation difference and used these improved estimates in our presented results.
我們將此結果與自舉重採樣結合使用，以更緊密地限制我們的估計差異，並在我們呈現的結果中使用這些改進的估計。

Figures 10 to 16 show histograms of the bootstrap procedure with mean payoff estimates for the different algorithms for the case of $\alpha = 0.1$.
圖 10 到 16 顯示了在 $\alpha = 0.1$ 的情況下不同算法的平均收益估計的引導程序的直方圖。

We note that not all estimates could be improved.
我們注意到並非所有的估計都可以改進。

In a few cases the original samples we obtained displayed very low variance, sometimes even with an effective sample size of only one and zero variance.
在少數情況下，我們獲得的原始樣本顯示出非常低的方差，有時甚至有效樣本大小僅為 1 且方差為零。

Therefore, in these low sample diversity cases, our estimates are less reliable as an improvement on the original payoff estimate.
因此，在這些低樣本多樣性的情況下，我們的估計作為對原始收益估計的改進不太可靠。

For example, when the number of other cooperators is zero in IA2C (Figure 10), the resampling distribution over payoffs when defecting approaches a normal distribution, whereas for cooperation, it is a constant with zero variance.
例如，當 IA2C 中其他合作者的數量為零時（圖 10），背叛時收益的重採樣分佈接近正態分佈，而對於合作，它是一個零方差的常數。

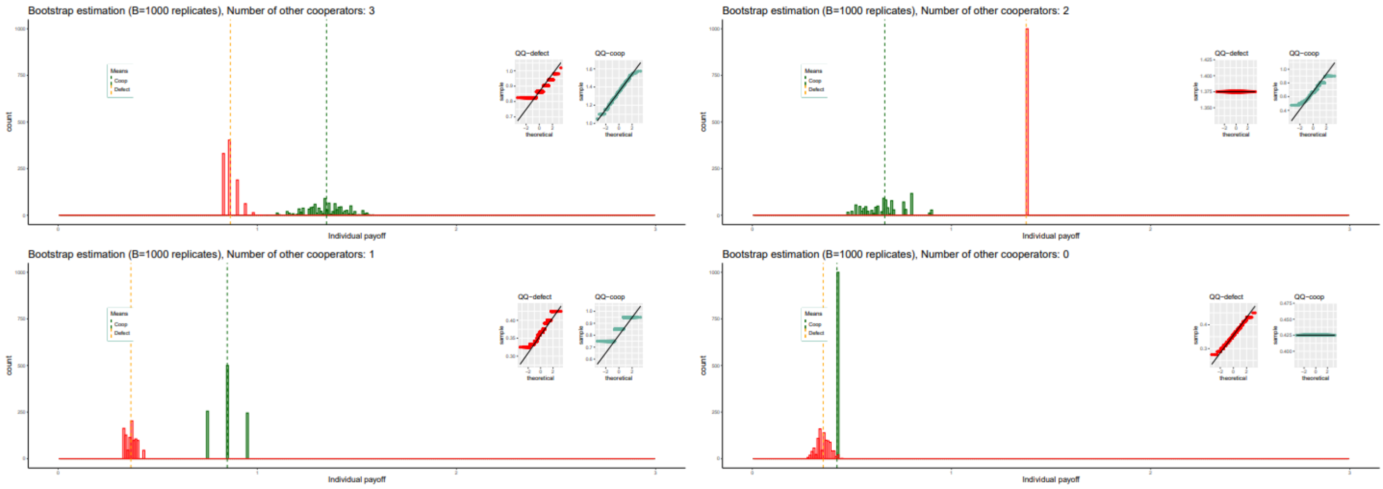


Figure 10: **IA2C**

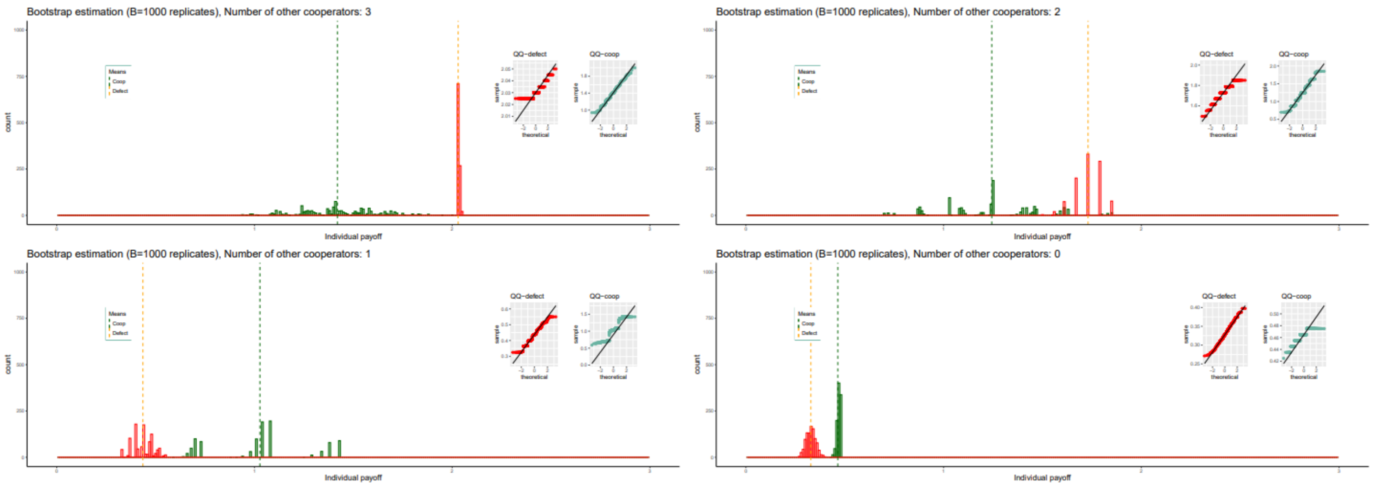


Figure 11: **NA2C**

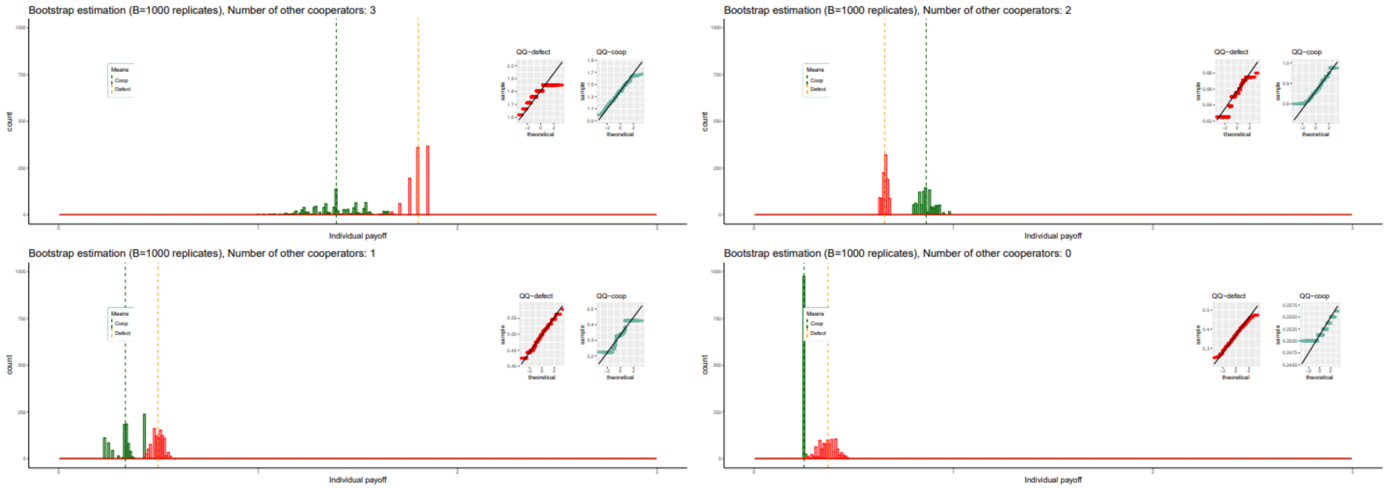


Figure 12: **FPrint**

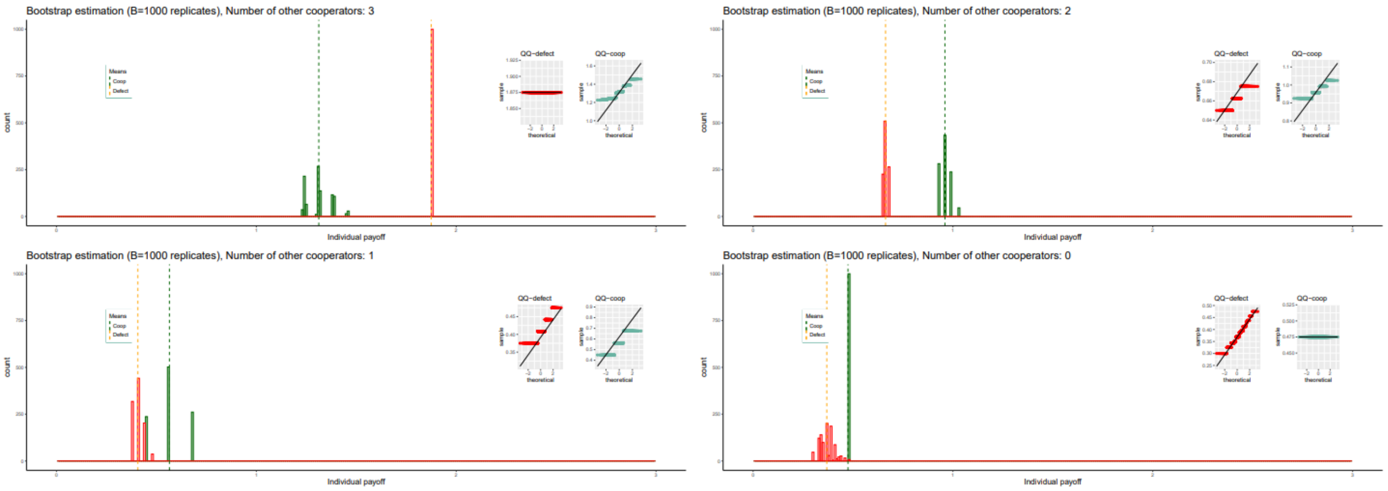


Figure 13: **ConseNet**

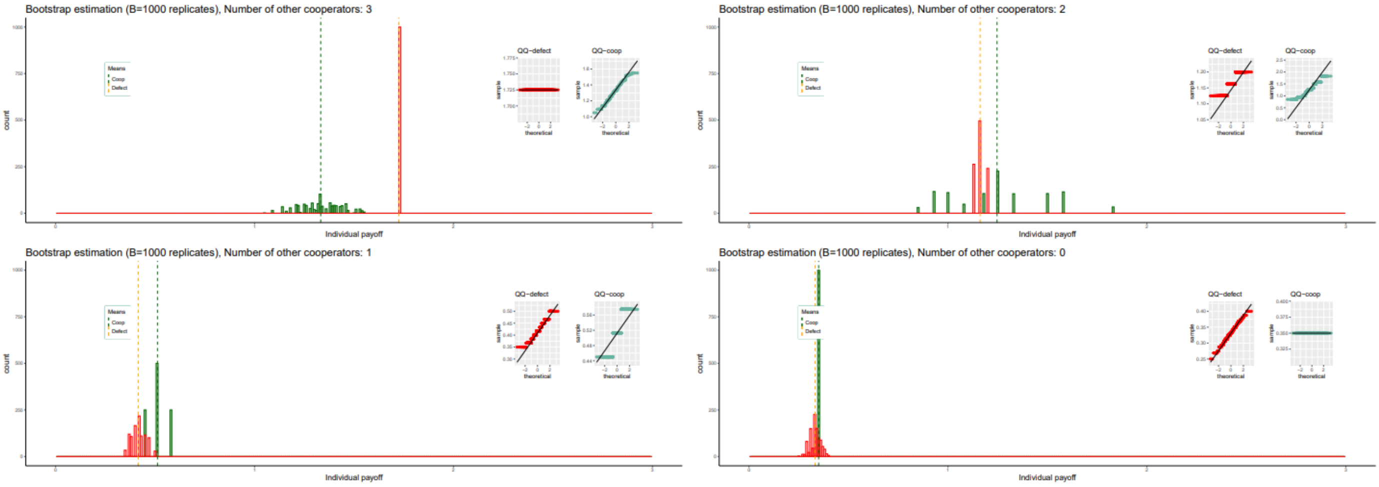


Figure 14: **DIAL**

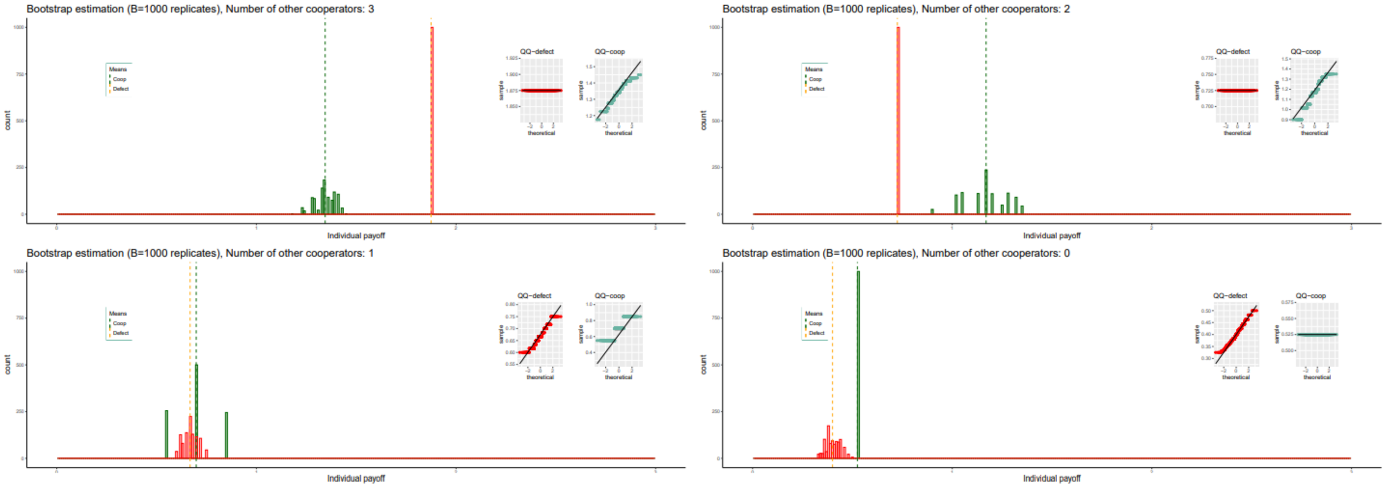


Figure 15: **CommNet**

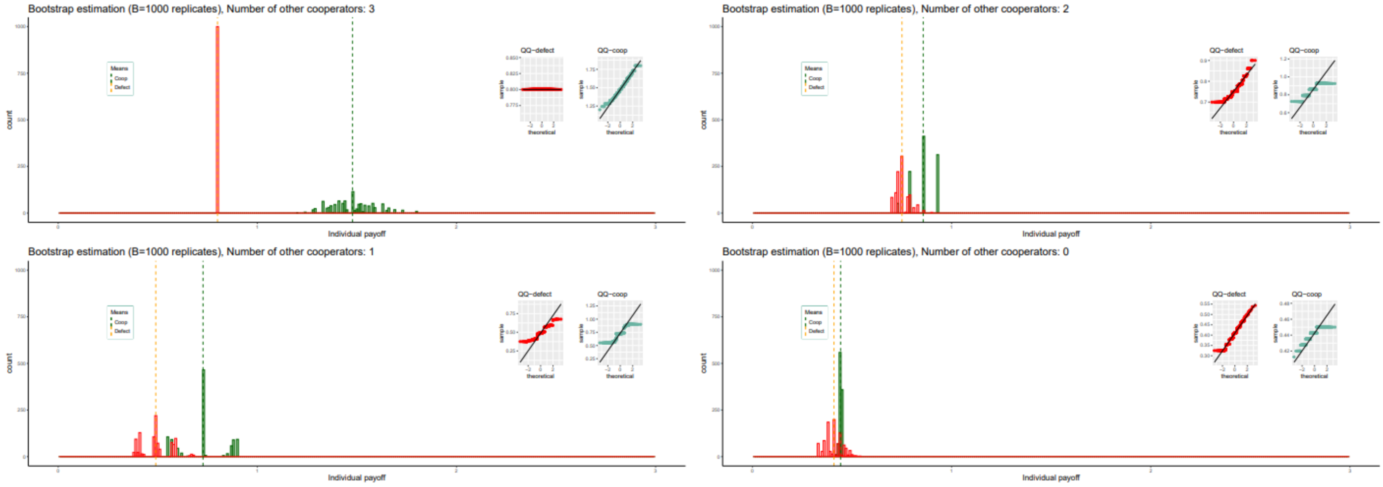


Figure 16: **NeurComm**