UNIVERSITATEA BABEŞ-BOLYAI
FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ

Abordări ale învățării prin întărire multi-agent în scenarii complexe bazate pe teoria jocurilor

Rezumatul tezei de doctorat

Student-doctorand: Mali Imre Gergely
Coordonator științific: Prof. dr. Czibula Gabriela

2026

Conținut

Lista publicațiilor

Clasamentul publicațiilor a fost realizat conform standardelor CNATDCU (Consiliul Național de Atestare a Titlurilor, Diplomelor si Certificatelor Universitare) aplicabile pentru studenții doctoranzi înscriși dupa 1 octombrie 2018. Toate clasamentele sunt listate conform clasificării jurnalelor[1] și a conferințelor[2] în Informatică.

Publicații în volume ale conferințelor

[Mal25] **Mali Imre Gergely** *A sequence-modeling approach to cooperation in Public Goods Games via Multi-Agent Transformers*. 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025) pp. 3708–3717 (**indexare Web of Science**)

Rank B, 4 puncte.

[Mal24] **Mali Imre Gergely**. *Multi-Agent Deep Reinforcement Learning for Collaborative Task Scheduling*. The 16th International Conference on Agents and Artifcial Intelligence (ICAART 2024), pp. 1076–1083 (**indexare Web of Science**)

Rank B, 4 puncte.

[MC23] **Mali Imre Gergely**, Czibula Gabriela. *Policy-Based Reinforcement Learning in the Generalized Rock-Paper-Scissors Game*. The 31th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2023), pp. 345–350 (**indexare Scopus**)

Rank B, 4 puncte.

[Mal22] **Mali Imre Gergely** *Finding Cooperation in the N-Player Iterated Prisoner's Dilemma with Deep Reinforcement Learning over Dynamic Complex Networks*. 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022) pp. 465–474 (**indexare Web of Science**)

Rank B, 4 puncte.

Publicații în jurnale internaționale

[Mal26] **Mali Imre Gergely** *Self-Play Meta-Reinforcement Learning in Multi-Agent Games*, Acta Universitatis Sapientiae, 18 (5), 2026, pp. 1–17 (**indexare Web of Science**, Q4)

Rank C, 2 puncte.

Scorul publicațiilor: 18 puncte.

[1] https://uefiscdi.ro/premierea-rezultatelor-cercetarii-articole
[2] http://portal.core.edu.au/conf-ranks/

Introducere

Scopul acestei teze este de a investiga problema ingineriei sistemelor multi-agent care dispun de inteligență socială prin utilizarea instrumentelor învățării prin întărire și a teoriei jocurilor.

Împreună cu învățarea supervizată și nesupervizată, paradigma învățării prin întărire (RL, *reinforcement learning*) este una dintre paradigmele de bază ale învățării automate. Dacă învățarea supervizată și cea nesupervizată sunt concepute pentru a descoperi tipare în datele etichetate și neetichetate, date care sunt în general deja gata produse, RL își are baza pe ideea de a învăța din interacțiunea cu un mediu, unde un agent decide cum să acționeze cel mai bine pentru a maximiza un semnal numeric numit recompensă [SB18].

RL este o intersecție a multor domenii științifice. Inspirată de neuroștiință, psihologie și teoria controlului, dezvoltată prin intermediul instrumentelor matematicii și ingineriei, RL a fost larg influențat și de economie și informatică, cu aplicații de succes în lumea reală, de la jocuri, cercetare operațională, managementul resurselor, robotică și multe altele [ADBB17]. Recent, RL a avut parte de o atenție extraordinară și în domeniul Sistemelor Multi-Agent (MAS, *multi-agent systems*), unde mai mulți agenți care interacționează urmăresc simultan obiective individuale sau comune. MAS oferă instrumente puternice pentru a modela situații din lumea reală cu mai multe entități care interacționează, în timp ce instrumentele RL oferă un cadru riguros din punct de vedere matematic pentru învățare. Combinația celor două este denumită Învățare prin Întărire Multi-Agent (MARL, *multi-agent reinforcement learning*), care a devenit recent una dintre cele mai populare domenii de cercetare bazate pe agenți [NX24]. De asemenea, setul de instrumente MARL este potrivit pentru a servi obiectivele acestei teze: proiectarea de agenți social inteligenți care pot urmări obiective comune, ținând cont de obiectivele individuale, dar fără a cădea victime miopiei egoiste de a urmări doar obiective individuale.

În acest scop, domeniul teoriei jocurilor este extrem de util. Dezvoltată inițial în anii 1950, teoria jocurilor oferă modele matematice pentru a studia interacțiunile strategice dintre entități cu abilitatea de a lua decizii raționale [Smi79]. Aceste modele sunt etichetate jocuri, în timp ce decidenții/actorii sunt adesea denumiți jucători. Cu concepte cheie precum strategii, echilibre și optimalitate, teoria jocurilor oferă instrumente matematice riguroase pentru a analiza situațiile sociale de interes.

O problemă centrală a construirii sistemelor cu inteligență socială este problema cooperării și conflictului, care a reprezentat o enigmă de lungă durată în multe domenii ale științei. Biologii investighează prezența rezolvării structurate a conflictelor, a cooperării reciproce și a altruismului în regnul animal prin intermediul instrumentelor teoriei evolutive a jocurilor [Daw16, SP73, Smi79]. Economiștii studiază războaiele prețurilor, proiectarea licitațiilor și strategiile tarifare. Politica este interesată de embargouri, construirea de coaliții, negocierea tratatelor și uneori de războaie propriuzise [AH81]. Teoria jocurilor poate oferi perspective valoroase și poate ajuta la tragerea unor concluzii utile în toate aceste domenii, în timp ce RL oferă o bază teoretică solidă pentru dezvoltarea agenților de învățare.

Domeniile enumerate anterior sunt interesate în principal de interacțiunile, strategiile, intențiile și acțiunile entităților existente, cum ar fi animalele și oamenii (sau grupuri formate de acestea). Cel mai

important, însă, este că studierea interacțiunilor strategice devine din ce în ce mai relevantă în zilele noastre, când epoca agenților autonomi este în ascensiune în diverse domenii, cum ar fi asistenții personali, roboții de fabricație, agenții comerciali și conducerea autonomă [LZG+17]. Pe măsură ce agenții autonomi sunt din ce în ce mai integrați în societate, înțelegerea apariției modelelor de interacțiune dintre agenții de învățare devine din ce în ce mai relevantă. Într-adevăr, proiectarea unor agenți de învățare care pot internaliza modul de urmărire a obiectivelor individuale, menținând în același timp cooperarea la scară mai largă, este una dintre problemele cheie studiate de această teză.

MARL cu medii simulate bazate pe teoria jocurilor este o metodă utilă pentru a studia problemele menționate anterior din câteva motive cheie. Calculul matematic al rezultatelor situațiilor studiate ar fi dificil de rezolvat din cauza problemei dimensionalității *(curse of dimensionality)*. Metodele MARL abordate sunt simulări bazate pe agenți ale adaptării la diferite situații. Această metodă științifică este similară în unele aspecte cheie cu cele două metode matematice principale, inducția și deducția. Metodele bazate pe agenți sunt concepute pe baza unui set de ipoteze care servesc drept principii directoare pentru proiectarea și rularea simulărilor, care apoi produc date ce pot fi analizate inductiv (și, de asemenea, utilizate în cadrul buclelor de adaptare ale agenților în cadrul experimentelor) [BBD18].

Direcții de cercetare

Cele trei direcții principale abordate sunt următoarele:

- o direcție teoretică în care ne propunem să înțelegem și să încurajăm cooperarea în dilemele sociale;
- o direcție orientată spre performanță, integrată în problemele lumii reale și motivațiile inginerești ale sistemelor MARL bazate pe teoria jocurilor;
- o abordare generală atotcuprinzătoare în care meta-învățarea este utilizată pentru a produce agenți robuști și adaptabili nu numai la situații de cooperare, ci și la cele competitive.

O temă recurentă întâlnită în toate direcțiile menționate este problema nestaționarității induse de adversari. Multitudinea agenților care interacționează introduce o nestaționare inerentă [LWT+17, ZYB21] ceea ce duce fie la o varianță mai mare în aproximatorii funcției valorice (ceea ce la rândul său duce la performanțe mai slabe) sau la neconvergență din cauza naturii în continuă schimbare a funcțiilor valorice [FFA+17]. Adresarea acestui tip de nestaționaritate este recurentă în toate lucrările noastre, cu un impact semnificativ asupra apariției cooperării în dilemele sociale, asupra aplicațiilor din lumea reală, și nu în ultimul rând asupra performanței și convergenței agenților de meta-învățare.

Cele trei direcții principale de cercetare sunt detaliate mai jos.

Învățarea cooperării în dileme sociale cu MARL. Ca primă linie de cercetare, am dezvoltat abordări de învățare cu întărire profundă multi-agent pentru *îmbunătățirea algoritmilor existenți bazați pe RL*, cu scopul de a găsi politici cooperative în dilemele sociale.

Deși algoritmii RL au fost mult timp subiectul analizei academice, domeniul încă prezintă lacune în mai multe aspecte teoretice. Mai precis, este un fapt cunoscut despre algoritmii RL tradiționali utilizați în medii în care agenții se confruntă cu situații dilematice de cooperare și conflict, de interes personal sau beneficiu colectiv că aceștia converg spre puncte optime locale în spațiul politicilor posibile care sunt mioape, egoiste, și pe termen lung suboptimale.

Propunerea acestei direcții este obținerea unui comportament cooperativ comun în dilemele sociale cu MARL. Dilemele sociale sunt situații care creează un conflict între interesele individuale

și interesele colectivului. Aceste situații în general expun structuri de recompensă în care indivizii egoiști sunt recompensați, însă egoismul colectiv duce la rezultate nefavorabile pe termen lung pentru colectiv și, prin urmare, pentru individ. Deși astfel de situații apar în principal în societate, agenții individuali din contexte MARL se confruntă adesea cu situații similare (în special în sisteme MARL descentralizate, unde agenții învață și acționează independent). În cele mai multe cazuri, aceste situații se reduc la dileme de tip cooperare-competiție în sisteme multi-agent, cum ar fi sistemele peer-to-peer, partajarea de fișiere, rutarea pachetelor, rețelele de senzori wireless etc., probleme care sunt adesea destinate a fi abordate cu RL [SRAA+11]. Deși detaliile exacte ale problemelor specifice au numeroase particularități, cerințe și detalii, cum ar fi echipamentul și configurarea, problema simplă a dilemelor poate fi abstractizată și studiată separat. Acest lucru se realizează de obicei prin utilizarea teoriei jocurilor. Cea mai faimoasă dilemă socială modelată de teoria jocurilor este Dilema Prizonierului, însă multe alte modele au fost propuse pentru a surprinde diferite dileme sociale, cum ar fi jocul snowdrift, jocul șoimului și porumbelului, jocul bunurilor publice etc.. O mare parte din aceste jocuri merită investigate în cadrul MARL. Dintre acestea, cercetarea noastră se concentrează pe Dilema Prizonierului și Jocul Bunurilor Publice.

În plus, deși algoritmii RL au atins sau chiar au depășit nivelul de performanță atins de oameni în jocurile cu informații perfecte [SSS+17, SHS+17] este încă dificil să se găsească echilibre Nash aproximative în jocuri cu informație imperfectă și de mare amploare. Metodele care adresează această problemă sunt în general derivate din Neural Fictitious Self-Play [HS16], care totuși adesea suferă de blestemul dimensionalității, când sunt folosite împreună cu metode RL cu funcții de valori [ZWLP19].

Problema devine și mai accentuată în sistemele complet descentralizate, unde agenții nu au nicio modalitate de a partaja informații în timpul antrenamentului, iar singura sursă de semnal este recompensa.

Această subdirecție se materializează în cercetarea noastră privind abordarea problemei MARL descentralizat cu agenți în rețea (o rețea de agenți fiind un graf cu agenți ca și noduri și conexiuni între aceștia fiind muchii). MARL descentralizat utilizează algoritmi RL cu un singur agent separat și independent pentru fiecare agent. O altă provocare din lumea reală inclusă în setările agenților în rețea este comunicarea limitată dintre agenți. [FFA+17]. Deși majoritatea abordărilor se concentrează explicit pe metodele RL utilizate, natura grafurilor de interconectare ale agenților are, de asemenea, o influență semnificativă asupra comportamentului general și, prin urmare, asupra performanței sistemului [BPB11]. Mai mult, modelele de interconectare dinamică (în timp, agenții își schimbă pozițiile relative unul față de celălalt și, prin urmare, față de vecinii cu care comunică) pot avea, de asemenea, o influență semnificativă [PBARA14]. Este deosebit de interesant să analizăm influența interconectării agenților asupra performanței generale și, într-adevăr, studiul nostru [Mal22] al Dilemei Iterate a Prizonierului cu N Jucători (NIPD) studiază exact acest lucru: apariția cooperării în rețele dinamice de agenți RL unde interconexiunile se schimbă stocastic în funcție de ratele de cooperare ale agenților.

Aplicații și benchmark-uri ale MARL-ului aplicat în teoria jocurilor. Deși MARL a avut un succes larg într-o multitudine de domenii, testarea și evaluarea diferiților algoritmi MARL au fost și sunt un punct slab al domeniului pentru mult timp. Efectuarea experimentelor în medii standard este îngreunată de implementările personalizate, care nu numai că reprezintă un efort ingineresc irosit, dar și împiedică comparabilitatea între diferite studii. Pentru a combate acest lucru, au fost propuse mai multe standarde, API-uri, medii și benchmarking-uri, cele mai remarcabile fiind OpenAI Gym [BCP+16], RLLab [DCH+16] și AxelrodPy [pd16] pentru teoria jocurilor. În afară de Proiectul Axelrod, însă, reperele teoriei jocurilor sunt rare, iar cele existente vizează în mare parte dileme

sociale celebre, precum Dilema Prizonierului, iar jocurile cu echilibre cu strategii mixte sunt mai puțin populare cu mai puține abordări existente. Una dintre propunerile acestei teze își propune să elimine exact această lacună: generalizăm celebrul joc cu proprietate ciclică numit Piatră-Hârtie-Foarfecă (Rock-Paper-Scissors) la un număr mai mare de jucători și acțiuni pentru a crea un mediu care este reglabil în complexitate strategică, dar aplicabil unui număr arbitrar de mare de agenți, oferind un punct de referință util pentru agenții MARL care își propun să învețe echilibre strategice mixte [MC23].

O altă sferă de cercetare în cadrul acestei direcții este aplicarea MARL bazat pe teoria jocurilor la probleme din lumea reală, în primul rând pentru a ilustra cum apar în scenarii reale probleme de cooperare și conflict. În al doilea rând subliniem necesitatea dezambiguizării prin modele teoretice ale jocurilor, cum ar fi jocurile Agent-Environment-Cycle. Pentru a realiza acest lucru, aplicăm MARL la problema planificării sarcinilor în cloud computing, unde mai multe mașini consumă dintr-o coadă de joburi partajată.

Meta-învățare aplicată în teoria jocurilor. Deși o parte semnificativă a cercetării se concentrează pe rezolvarea unor medii individuale, rezolvarea unor probleme specifice lumii reale sau concentrarea pe rezolvarea unor dileme sociale singulare, există o nevoie de sisteme MARL care să se poată adapta la situații noi cât mai repede posibil. De exemplu, cu condiția ca circumstanțele să fie adecvate, sportivii își pot aplica cu succes abilitățile pe orice teren de joc, un bucătar poate pregăti mese în orice bucătărie, un muzician poate cânta pe orice scenă, iar un chirurg poate opera în orice sală de operație. În schimb, în RL, mediile de testare și validare sunt aceleași, astfel încât agenții sunt proiectați să se supraadapteze la un anumit mediu. Cu toate acestea, dezvoltarea unor agenți care nu sunt doar capabili să țină cont de nestaționarea indusă de adversari, ci sunt și capabili să facă față schimbărilor ale mediului, este esențială în economie, politică și multe alte domenii supuse unor schimbări constante.

Pentru a aborda această problemă, au fost propuse cadre de meta-învățare, care în domeniul RL schimbă paradigma RL tradițională și introduc antrenamente în bucle interne și externe. Buclele interne sunt etape tradiționale de adaptare RL, în timp ce bucla externă suplimentară eșantionează diferite medii dintr-o anumită distribuție, în timp ce cunoștințele internalizate sunt reținute prin stări ascunse. Rezultatul este reprezentat de agenți de meta-învățare capabili să se adapteze la diferite medii printr-un număr minim de interacțiuni și depășesc limitele RL tradiționale prin evitarea uitării catastrofale.

Astfel, o altă direcție promițătoare pe care o urmărim în această teză este studiul MARL în jocuri multi-agent, unde căutăm algoritmi care produc agenți capabili să se co-adapteze în diferite jocuri, jocuri în care recompensele și peisajele strategice sunt supuse unor schimbări constante.

Contribuții originale

Un obiectiv general al acestei teze este de a avansa către o inteligență adaptivă, cooperativă și conștientă social, care să reușească să navigheze productiv în problemele multi-agent. Pentru a demonstra potențialul diferitelor metode MARL în acest sens, au fost luate în considerare dileme sociale, jocuri ciclice, aplicații din lumea reală și configurații de meta-învățare, care sunt detaliate în Capitolele 2, 3 și 4, și rezumate mai jos:

Cooperare bazată pe MARL în Dileme Sociale Metodele RL tradiționale sunt cunoscute pentru convergența lor către echilibre Nash defectuoase, fără mecanisme suplimentare care să încurajeze

cooperarea. Lucrarea de față explorează mecanismele de cooperare în cadrul Dilemei Prizonierului Iterată cu N Jucători și a Jocului Bunurilor Publice:

- Metoda de tip RL bazată pe funcții de valoare, complet descentralizată și integrată cu rețele complexe dinamice, prezentată în Secțiunea 2.1 pe baza lucrării noastre originale publicate în [Mal22] este aplicată asupra NIPD. Acesta reprezintă primul studiu care investighează această dilemă socială specifică în contextul rețelelor complexe dinamice, unde agenții sunt dispuși inițial în diverse tipuri de grafuri complexe, dar unde muchiile evoluează probabilistic în timp, în funcție de ratele de cooperare. Agenții aleg să coopereze sau să dezerteze în funcție de istoricul interacțiunilor cu vecinii lor din graf. În acest context, demonstrăm în premieră emergența cooperării și oferim o analiză detaliată a amplorii acesteia, raportată la acțiunile trecute și la structurile inițiale de vecinătate. O altă contribuție a acestei lucrări constă în propunerea unei reprezentări flexibile a stării (non-rigid state representation) pentru metodele de tip Deep RL, capabilă să se adapteze natural unui număr variabil de oponenți.

- Rezultatele noastre obținute în cadrul Jocului Bunurilor Publice, prezentate în lucrarea [Mal25] și discutate în Secțiunea 2.2, sunt, din câte cunoaștem, primele care aplică cu succes arhitecturi bazate pe modele de tip Transformers, cum ar fi Multi-Agent-Transformers (MAT) [WKL+22], la dileme sociale. Modelele MAT consideră selecția acțiunilor multi-agent ca o problemă de modelare secvențială și utilizează mecanisme de atenție pentru a urmări modul în care alegerile agenților se afectează reciproc. Deoarece MAT folosește teorema descompunerii avantajului pentru a descompune obiectivul comun într-o secvență de obiective individuale fără a pierde din domeniul de aplicare, utilizarea metodei în dilemele sociale este bine motivată. Similar lucrării noastre asupra Dilemei Prizonierului [Mal22], observațiile agenților sunt construite din istoriile adversarilor. În aceste condiții, demonstrăm empiric pe de o parte, că cooperarea poate fi realizată prin mecanisme bazate pe atenție, și pe de altă parte, că introducerea în observații a istoriilor mai lungi ale adversarilor ajută la promovarea cooperării într-un grad considerabil de mare.

Benchmark-uri și aplicații ale MARL bazate pe teoria jocurilor MARL a devenit extrem de popular, deși multidudinea de probleme asupra cărora se aplică algoritmi generici de RL face dificilă compararea fiabilă a rezultatelor. Pentru a adresa această problemă, mai multe benchmarkuri au fost propuse precum SMAC [SRDW+19], OpenSpiel [LLL+19], unele fiind chiar bazate pe teoria jocurilor [LSB+23]. Cu toate acestea, în timp ce reperele și platformele de testare teoretice oferă o structură pentru interacțiunea strategică, utilă pentru demonstrarea conceptului, fiabilitatea și viabilitatea metodelor MARL trebuie verificate și în scenarii mai realiste, care se îndepărtează de ipotezele idealiste pe care teoria jocurilor le presupune adesea. În urmărirea acestor linii de cercetare, rezultatele noastre sunt următoarele:

- Propunem un platformă de testare strategic diversă, versatilă, reglabilă și fundamentată matematic pentru cercetarea MARL în [MC23] și o prezentăm în Secțiunea 3.2. O contribuție crucială a acestei lucrări este generalizarea cu succes a notoriului joc Piatră-Hârtie-Foarfecă la un număr arbitrar de agenți și un număr impar arbitrar de mare de acțiuni și demonstrarea matematică a echilibrului Nash al jocurilor. Aceasta oferă o platformă de testare mai generală, deoarece interacțiunea simultană a mai multor agenți (care este esențială pentru MARL și nu a fost abordată anterior în lucrări similare, cum ar fi [LSB+23]) este încorporabilă în mod natural. O altă contribuție notabilă a acestei lucrări o reprezintă rezultatele noastre empirice, care demonstrează că algoritmul PPO standard [SWD+17] prezintă un comportament analog

ciclului prețurilor economice atunci când este aplicat acestui benchmark independent în fiecare agent.

- În [Mal24] demonstrăm că MARL bazat pe teoria jocurilor poate fi aplicat cu succes în aplicații din lumea reală și detaliem contribuțiile noastre în Secțiunea 3.1. Mai exact, generalizăm simulatorul de planificare a sarcinilor (*task scheduler*) propus în [MAMK16] la mai multe mașini și aplicăm o abordare Agent-Environment-Cycle-games [TBG+21] care este capabilă să gestioneze mai mulți agenți cu mai multe mașini per agent. Rezultatele noastre arată că politicile de planificare comune găsite de agenți sunt la niveluri comparabile cu cele găsite prin euristici, oferind o abordare care poate fi scalată în mod colaborativ la mai mulți agenți simultan.

Abordări Meta-MARL în scenarii bazate pe teoria jocurilor O critică frecventă al RL-ului este aceea că mediile de testare și evaluare sunt aceleași [Wen19] și, prin urmare, antrenamentul RL standard este inerent supraadaptat la un singur mediu. Pentru a combate acest lucru, au fost propuse abordări de meta-învățare, însă această problemă este și mai relevantă în MARL, unde, pe lângă ținta mobilă a mediului, nestaționarea inerentă abordărilor multi-agent face ca învățarea și convergența să fie și mai dificile de obținut. Pentru a adresa aceste probleme, am propus o abordare de meta-învățare după cum urmează:

- În lucrarea noastră originală, [Mal26], propunem o metodă meta-MARL complet descentralizată care prezintă un comportament adaptiv atât în scenarii cooperative, cât și în scenarii adverse și detaliem rezultatele noastre în Capitolul 4. Mai exact, suntem primii care adaptează o metodă Meta-RL standard numită RL^2 la o configurație complet descentralizată de self-play meta-RL și oferim două clase de medii teoretice ale jocurilor. Folosim distribuții de jocuri cu formă normală pentru a crea seturi diverse de situații de cooperare (jocuri de coordonare aleatorii) și conflict (jocuri aleatorii cu sumă zero) similare din punct de vedere strategic, dar concret diferite. Rezultatele noastre demonstrează empiric capacitatea abordării propuse de a produce agenți eficienți din punct de vedere al eșantionării, robuști la schimbări și capabili de generalizare strategică pe diverse seturi de scenarii teoretice ale jocurilor, măsurate prin metrica Nash-regret sau Nikaido-Isoda [NI55].

Structura tezei

Restul acestei teze este organizat în felul următor: Capitolul 1 prezintă fundamentele teoretice necesare, oferind o introducere generală în domeniile învățării prin întărire, sistemelor multi-agent, meta-învățării, teoriei jocurilor și rețelelor complexe.

Apoi, capitolul 2 introduce motivația detaliată pentru studierea cooperării și urmărirea unui comportament cooperativ comun prin intermediul agenților RL în dilemele sociale. Secțiunea 2.1 prezintă prima noastră abordare RL, bazată pe RL cu funcții valoare, într-o formă generalizată și iterată a Dilemei Prizonierului [Mal22]. Secțiunea detaliază experimentele noastre pe mai multe tipuri de rețele, lungimi ale istoricului agenților și comportamentul de recablare al agenților, împreună cu o comparație cu lucrări conexe. În continuare, Secțiunea 2.2 prezintă următoarea noastră lucrare privind căutarea cooperării în dilemele sociale. Această secțiune se bazează pe contribuțiile originale conform lucrării noastre [Mal25], detaliind motivația alegerii modelului, discutând rezultatele obținute și comparându-le cu lucrări conexe.

Capitolul 3 discută propunerea noastră de benchmark și rezultatele obținute în aplicații din lumea reală. Secțiunea 3.1 discută propunerea noastră de planificare colaborativă a sarcinilor cu mai mulți agenți, prezentată în [Mal24]: discutăm motivațiile, prezentăm cadrul nostru general și discutăm

rezultatele noastre în comparație cu euristicile existente de planificare a sarcinilor. Apoi, Secțiunea 3.2 detaliază propunerea noastră de benchmark publicată în [MC23], oferă o demonstrație matematică a echilibrelor sale și prezintă rezultatele evaluării noastre experimentale.

Capitolul 4 prezintă contribuțiile noastre originale în domeniul meta-învățării, integrate în MARL și teoria jocurilor, așa cum sunt prezentate în [Mal26]. Acest capitol discută relevanța meta-învățării, prezintă propunerea noastră de distribuție a mediului bazată pe jocuri în formă normală și evaluează performanța pe baza Nash-regret.

În final, sunt schițate concluziile tezei și posibilele direcții de cercetare viitoare.

Capitolul 1

Fundamente teoretice

Scopul acestui capitol este de a oferi o introducere în cadrul teoretic al subiectelor și problemelor abordate în această teză și de a prezenta stadiul actual al metodelor recente de rezolvare, respectiv al problemelor deschise. Mai exact, domeniile învățării prin întărire, împreună cu sistemele multi-agent, vor fi prezentate ca și cadru teoretic central pentru învățarea prin întărire multi-agent, ale cărui instrumente cuprind întreaga teză. Un alt concept central al prezentei lucrări este teoria jocurilor, a cărei putere descriptivă și intuitivă este utilizată pentru a clarifica diferite aspecte ale acelor interacțiuni multi-agent, a căror înțelegere constituie unele dintre principalele contribuții ale acestei teze. În plus, se va oferi o introducere în domeniul rețelelor complexe și al meta-învățării, ca domenii auxiliare utilizate în unele dintre direcțiile de cercetare abordate, în timp ce o introducere în domeniul planificării sarcinilor este oferită ca context pentru partea aplicativă a lucrării noastre.

1.1 Învățare prin întărire

Învățarea cu întărire (RL) este una dintre cele trei paradigme fundamentale ale învățării automate. În timp ce celelalte (învățarea nesupervizată și cea supervizată) iau în considerare date deja colectate și selectate, învățarea prin consolidare obține date și experiență din interacțiunea directă. Învățarea din interacțiune este una dintre ideile fundamentale care cuprinde majoritatea teoriilor despre învățare și inteligență [SB18]. De asemenea, este cea mai simplă și concisă definiție a ceea ce numim *învățare cu întărire*. Din cauza lipsei unor rezultate date fixe, proiectate și etichetate de om, performanța RL nu este, în general, limitată de calitatea datelor experte existente, ci de capacitatea algoritmului RL de a identifica și exploata cu succes modelele care stau la baza mediului, care pot fi utilizate pentru a lua decizii care, în final, maximizează recompensa. Domeniul RL a cunoscut progrese extraordinare și a atras o atenție enormă în ultimii ani. O multitudine de probleme de luare a deciziilor secvențiale din lumea reală au fost rezolvate cu succes prin abordări de învățare cu întărire. Astfel de probleme includ conducerea autonomă, ingineria software, zborul cu drone, robotica, jocuri precum Go, șah și poker etc. [Li19].

În timp ce în alte paradigme de învățare automată conceptele centrale sunt algoritmul de învățare și seturile de date, RL concepe, în general, o viziune asupra lumii formată din agenți și medii. Componentele de învățare ale acestei configurații se află în cadrul agentului, care acționând în interiorul mediului primește observații și recompense, apoi decide asupra următoarei acțiuni, creând astfel o buclă închisă (Figura 1.1).

Obiectivul principal al agentului este de a alege acțiunile într-un mod care maximizează recompensa cumulativă. Din cadrul pe care îl urmează un agent pentru a lua decizii, se pot distinge agenții bazați pe funcții valoare (*eng. value-based agents*), agenții bazați pe politici și agenții actor-critici.

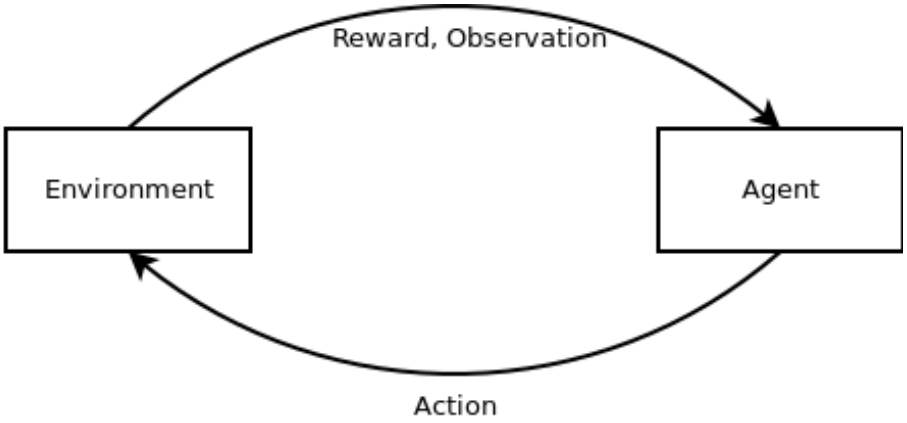


Figure 1.1: Interacțiunea agent-mediu

Politica unui agent bazat pe funcții valoare este implicită, deoarece poate citi direct cea mai bună acțiune din funcția sa valoare. În schimb, un agent bazat pe politici stochează direct o politică și selectează acțiuni fără a stoca vreodată explicit funcția de valoare. Agenții actor-critici combină cele două anterioare (politici și funcții Q/V) și încearcă să obțină beneficiile amândurora.

O altă clasificare cheie a agenților RL se poate face pe baza cunoașterii mediului: agenți fără model și agenți bazați pe model. Un agent bazat pe model are cunoștințe complete despre dinamica și stările mediului, modelul său al mediului este complet și trebuie doar să învețe comportamentul adecvat în acesta, în timp ce agenții fără model au un model incomplet al mediului (sau deloc) și trebuie să aproximeze mediul în timp ce învață să maximizeze recompensa. Majoritatea problemelor din lumea reală abordate de RL utilizează agenți fără model, deoarece majoritatea problemelor de interes sunt imposibil de descris complet și tractabil și/sau prea mari pentru a fi reprezentate explicit.

Tot ce se află în interiorul sistemului, în afara agentului, se spune că face parte din mediu. Există mai multe tipuri de medii, în funcție de care este definită și acuratețea modelului agentului asupra acestuia. Russel și Norvig [RN16] propun următoarele proprietăți ale mediului:

- determinist sau nedeterminist;
- discret sau continuu;
- episodic sau secvențial
- static sau dinamic;
- complet sau parțial observabil;
- de agent unic sau multi-agent.

1.2 Sisteme multi-agent și învățare cu întărire multi-agent

Sistemele multi-agent (MAS) sunt configurații în care mai mulți agenți acționează într-un mediu partajat. Agenții interacționează cu mediul partajat, dar acționează și între ei și urmăresc un anumit obiectiv, care poate fi comun tuturor, unora sau niciunului. În general, sistemele bazate pe agenți și, de asemenea, MAS sunt descrise de agenți autonomi, adică fără intervenție umană directă. Această presupunere se aliniază organic cu natura agenților RL. Astfel, sistemele autonome multi-agent în care agenții se bazează pe algoritmi de RL, vorbim de învățare cu întărire multi-agent. (*multi-agent reinforcement learning, MARL*). De-a lungul timpului, MARL a demonstrat un potențial și o aplicabilitate

imense în numeroase aplicații din diverse domenii, cum ar fi robotica, medicina, agricultura, economia etc. [ZYB21, HY20, HPHKPI20]. Aplicarea MARL la aceste probleme și la altele este justificată de eficacitatea sa în depășirea blestemului dimensionalității, în valorificarea toleranței naturale la erori a MARL și uneori prin respectarea constrângerilor specifice mediului. În ciuda acestor avantaje, MARL nu duce lipsă de propriile provocări, cele mai răspândite dintre acestea fiind nestaționaritatea, scalabilitatea, observabilitatea parțială, atribuirea de credite și coordonarea/competiția.

În primul rând, nestaționaritatea este o problemă centrală care provine direct din multitudinea de agenți. Pentru fiecare agent, mulțimea de colegi este considerată ca parte a mediului. Întrucât acești agenți, la rândul lor, învață simultan cu aceștia, problema învățării în MARL devine o țintă în continuă mișcare, ceea ce face ca garanțiile de convergență să fie considerabil mai dificil de oferit.

În al doilea rând, cu cât un sistem MARL are mai mulți agenți, cu atât spațiile comune de acțiune și observare devin mai mari. Deși un număr rezonabil de agenți ajută la depășirea problemei dimensionalității (*eng. curse of dimensionality*) în unele cazuri, această problemă poate reveni atunci când numărul de agenți crește prea mult și convergența poate fi împiedicată.

În al treilea rând, majoritatea sistemelor MARL iau în considerare agenți care sunt capabili să observe doar o porțiune locală a mediului. Acest lucru face ca aplicarea algoritmilor RL să fie deosebit de dificilă atunci când vine vorba de cooperare și/sau competiție.

În al patrulea rând, atribuirea creditelor este o problemă centrală a MARL. Atunci când recompensele sunt acordate într-un mod comun, devine dificil de decis acțiunile căror agenți duc la recompensa respectivă. Atribuirea necorespunzătoare a creditelor poate face ca unii agenți să învețe, în timp ce alții nu, atribuind în mod eronat recompensa greșelilor sau invers.

În cele din urmă, problema coordonării și a competiției este probabil cea care a devenit cea mai notorie. Această notorietate se datorează în mare parte intuițiilor directe și paralelelor cu situațiile și dilemele sociale. În multe cazuri, agenții trebuie să coopereze pentru a atinge un obiectiv comun. Întrucât toți agenții învață în același timp (cu alte cuvinte, problema nestaționarității), învățarea politicilor comune cooperative devine din ce în ce mai dificilă. În alte situații, obiectivele unor agenți ar putea fi în conflict unele cu altele, ceea ce înseamnă că ajung în situații de competiție între ei: acest lucru duce adesea la cicluri neconvergente sau la agenți supraadaptați la adversari cu performanțe obiective sub așteptări.

Pentru a aborda principalele complexități ale gestionării impedimentelor sistemice în MARL, procesul de învățare al fiecărui agent trebuie să țină cont de comportamentul celorlalți, ceea ce, în termeni de instruire și execuție, duce la mai multe arhitecturi de soluții [HM23]:

- **Instruire centralizată, execuție centralizată**: Această paradigmă reduce practic contextul multi-agent la unul cu un singur agent, unde multiplicitatea agenților, acțiunilor și observațiilor este ascunsă sub un singur agent al cărui comportament este căutat prin procesul de instruire în spațiul politicilor comune.

- **Instruire centralizată, execuție descentralizată**: Agenții învață folosind un creier centralizat, dar fiecare acționează independent, în funcție de observațiile individuale. Această paradigmă are avantajul de a partaja global experiența în timpul antrenamentului, dar beneficiază totodată de separarea agenților prin acțiuni descentralizate. Această paradigmă este predominantă în modelele cooperative, cum ar fi MAPPO. [YVV+22]

- **Instruire descentralizată, execuție descentralizată**: Cunoscută și sub denumirea de complet descentralizată, această paradigmă consideră fiecare agent cu propria unitate de învățare și execuție, separată și independentă. În cadrul acestei paradigme, agenții nu pot partaja direct experiența, iar din perspectiva fiecărui agent, ceilalți sunt părți cu adevărat independente

ale mediului. Această separare vine adesea cu dezavantajul instabilității și divergenței, însă independența agenților face ca acest model să fie mai scalabil și mai realist decât celelalte.

Pentru a surprinde formal interconectarea unui context în care acționează mai mulți agenți RL, se poate utiliza o generalizare MDP, numită Jocuri Markov (uneori denumite jocuri stocastice) [Sha53]. Jocurile Markov au fost introduse în [Lit94] ca un cadru general pentru MARL.

Formal, un joc Markov [Lit94] este definit ca un tuplu $< N, S\{A^i\}_{i\in\mathbb{N}}, P, \{R^i\}_{i\in\mathbb{N}}, \gamma >$. Aici N denotă mulțimea agenților (≥ 2), S denotă mulțimea stărilor observate de toți agenții. A^i și R^i denotă funcțiile de acțiune și recompensă pentru agentul i, în timp ce P și γ au semnificațiile lor obișnuite din Procesele Markov: funcția de probabilitate de tranziție și, respectiv, factorul de discount [ZYB19].

Obiectivul principal al fiecărui agent în parte este în continuare optimizarea recompensei cumulative individuale. La fiecare unitate de timp (pas) indexat cu t, din starea s_t fiecare agent i alege și execută acțiunea a_t^i, apoi tranziționează la starea s_{t+1} și primește recompensă după funcțiile respective $R^i(s_t, a_t, s_{t+1})$. Funcția valoare $v^i : S \to \mathbb{R}$ depinde de politica comună definită ca $v : S \to \Delta(A)$, și scrisă ca $\pi(a, s) := \Pi_{i\in\mathbb{N}}\pi^i(a^i|s)$

Similar cu contextul cu un singur agent, dar ținând cont de multitudinea de actori, se poate defini funcția valoare pentru o politica comună π și starea s după cum urmează:

$$V^i_{\pi^i,\pi^{-i}}(s) := \mathbb{E}\left[\sum_{t\geq 0}\gamma^t R^i(s_t, a_t, s_{t+1}) \middle| a_t^i \sim \pi^i(\cdot|s_t), s_0 = s\right] \tag{1.1}$$

Aici -*i* marchează indecșii tuturor agenților cu excepția agentului i. Faptul care face ca Jocurile Markov să fie exponențial mai complexe decât MDPS simplu este că, în cazul comportamentului optim, acesta depinde de mai mult decât comportamentul agentului curent, iar comportamentul tuturor celorlalți agenți joacă, de asemenea, un rol crucial. Datorită acestui fapt, conceptul de soluție este, de asemenea, diferit de soluțiile MDP convenționale. Cea mai comună soluție este strâns legată de conceptul de echilibru Nash [BO98]. Echilibrul Nash al unui joc Markov este definit ca o politică comună $\pi^* = (\pi^{1,*}, \pi^{2,*}, ..., \pi^{n,*})$ care pentru orice stare $s \in S$ și $i \in N$ respectă condiția următoare:

$$V^i_{\pi^{i,*},\pi^{-i,*}} \geq V^i_{\pi^i,\pi^{-i,*}}, \forall\pi^i \tag{1.2}$$

Acest echilibru este un punct în spațiul politicilor comune, cu proprietatea că niciun agent nu are stimulente pentru abaterea de la acesta. Abstractizarea sarcinilor MARL bazată pe Jocul Markov este suficient de generală pentru a surprinde mai multe scenarii diferite, cum ar fi situații cooperative, competitive sau mixte, după cum urmează [ZYB19]:

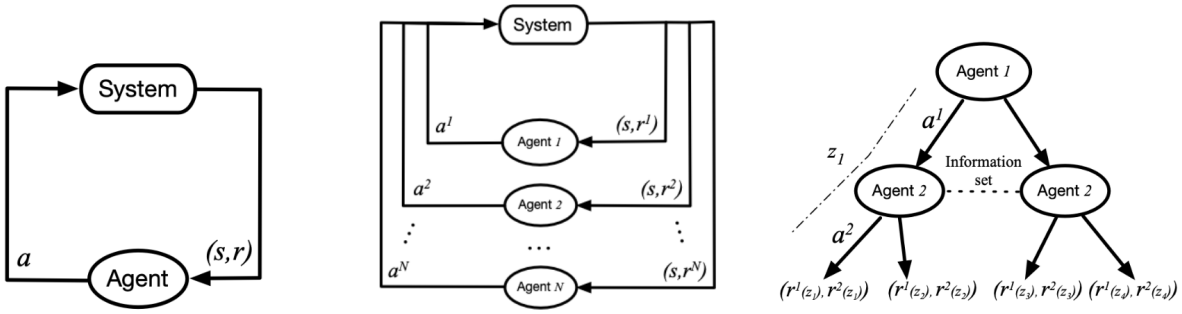
- **Cadru coperativ**: De obicei, toți agenții au aceeași funcție de recompensă, această configurație fiind cunoscută și sub numele de *MDP multi-agent* [LR00]. În astfel de configurații, se pot utiliza algoritmi RL cu un singur agent, deoarece valoarea funcției Q este comună tuturor agenților, însă acest lucru impune și restricția unui anumit nivel de omogenitate asupra agenților. Cu toate acestea, jocurile Markov cooperative pot fi generalizate în continuare la modele care iau în considerare în schimb *recompensa medie a echipei*, ceea ce permite, de asemenea, ca diferiți agenți să aibă funcții de recompensă private și distincte. Această generalizare, spre deosebire de cea anterioară, permite o mai mare eterogenitate între agenți și facilitează dezvoltarea învățării descentralizate cu întărire multi-agent, un obiectiv foarte ambițios urmărit în literatura de specialitate [KMP13, WYWH18].

- **Cadru competitiv**: de obicei modelate ca jocuri Markov cu sumă nulă (*eng. zero-sum games*), cel mai frecvent cu doi jucători [SSS+17, Lit94] unde câștigurile și pierderile celor doi jucători

sunt simetrice. Jocurile cu sumă nulă sunt însă adesea folosite și pentru învățarea robustă, deoarece incertitudinea mediilor care încetinește învățarea sau care poate fi văzută ca un impediment pentru agent poate fi atribuită unui adversar imaginar într-un joc care se desfășoară împotriva agentului.

- **Cadru mixt**: este o generalizare a situațiilor cooperative și competitive, modelate ca jocuri *general-sum* în care fiecare agent acționează într-o manieră egoistă, iar obiectivele diferiților agenți pot intra în conflict sau se pot suprapune. Existența unui set format din agenți complet cooperanți și complet competitivi, de exemplu echipe de agenți cooperanți care concurează unii împotriva altora [BKM+19, BBC+19] poate fi considerată o abordare cu motivație mixtă.

O altă abstracție utilă pentru sarcinile MARL este cea a jocurilor cu formă extensivă (*eng. ex*). Aceste jocuri sunt folosite pentru a modela jocuri fără informație perfectă și sunt definite ca un tuplu $< N \cup \{c\}, H, Z, A, \{R^i\}_{i \in N}, \tau, \pi^c, S >$. Aici N marchează mulțimea agenților (c fiind un agent de natură șansieră introdus special cu o politică stocastică fixă [ZYB19]). A este mulțimea acțiunilor posibile, în timp ce H este mulțimea istoricelor, un istoric denotând o serie de acțiuni întreprinse pe parcursul unui episod. Z este un subset al H și denotă mulțimea istoricelor terminalelor, conform cărora R^i atribuie utilități. τ este funcția de identificare care specifică ce agent a întreprins ce acțiune. Elementele lui S se numesc stări informaționale.



(a) MDP (b) Markov game (c) Extensive-form game

Figure 1.2: Markov process generalizations for reinforcement learning [ZYB19]

Scopul într-un asemenea joc se poate formaliza ca un ϵ-*Nash equilibrium*, o aproximare a adevăratului echilibru Nash care este atins când $\epsilon = 0$. Acest ϵ-*Nash-equilibrium* este definit astfel:

$$R^i(\pi^{i,*}, \pi^{-i,*}) \geq R^i(\pi^i, \pi^{-i,*}), \forall \pi^i \; of \; i \tag{1.3}$$

Figura 1.2 rezumă tipurile de abstracțiuni ale proceselor Markov detaliate mai sus, utilizate pentru învățarea prin consolidare, surprinzând de la cele mai simple setări cu un singur agent până la setări mai complexe cu mai mulți agenți.

1.3 Teoria jocurilor

Studiul situațiilor care implică interacțiuni strategice între entități raționale decizionale [Mye91] prin intermediul modelelor matematice se numește teoria jocurilor. Este studiul formal al conflictului și/sau cooperării dintre mai mulți jucători (oameni, grupuri, animale, companii, oameni, computere etc.). Dintr-o perspectivă teoretică, jocurile pot fi considerate modele de situații interactive între

jucători raționali. Modelarea jocurilor oferă instrumentele și cadrul matematic adecvat pentru a înțelege comportamentul jucătorilor atunci când urmăresc un anumit obiectiv în joc. Acest domeniu are aplicații într-o multitudine [Byu14, Pic19, FS03, JSB02, Cow12, Har09] și ajută la formalizarea multor relații comportamentale, este de interes particular și în domeniul inteligenței artificiale.

Jocurile pot fi clasificate în mai multe moduri, printre care cele mai importante sunt:

- cooperative sau necooperative;
- sumă zero vs sumă generală;
- simultane sau secvențiale;
- simetrice vs asimetrice;
- informație perfectă sau imperfectă.

În 1950, John Nash a demonstrat [N+50] că fiecare joc finit administrează cel puțin un punct de echilibru, unde fiecare jucător (cunoscând alegerile adversarilor) face cea mai bună acțiune pentru sine și nu are niciun stimulent să se abată de la strategia sa. Descoperirea lui Nash a făcut din teoria jocurilor un domeniu de cercetare foarte activ pentru o lungă perioadă de timp.

Cel mai bun răspuns (*eng. best response*) este conceptul care denotă strategia unui anumit jucător, presupunând strategia adversarului/adversarilor, care face ca rezultatul cel mai favorabil pentru jucător să fie cel mai probabil. O strategie se numește dominantă dacă depășește întotdeauna celelalte în ceea ce privește utilitatea. Conceptul best response a fost deja introdus în secțiunile anterioare în termeni de RL, aici le vom revedea dintr-o perspectivă pur teoretică a jocurilor.

Un set de strategii între jucători este considerat Pareto-optimal sau Pareto-eficient dacă o utilitate/un câștig mai bun pentru orice jucător nu poate fi obținut prin obținerea unei utilități mai slabe pentru un alt jucător [BBD18]. Aceasta înseamnă că, dacă jucătorii raționali se află într-o situație non-Pareto-optimală, au stimulente raționale pentru a trece la una mai bună, deoarece acest lucru se poate face fără a afecta câștigul niciunui alt jucător. Eficiența Pareto, însă, este o noțiune slabă de optimalitate. Conceptul de eficiență este, în schimb, o noțiune mult mai puternică și denotă setul de strategii care maximizează utilitatea totală a rezultatelor pentru toți jucătorii.

Un **echilibru Nash** se numește a fi setul de strategii de la care nu ar merita să se abată niciun agent. Deoarece agenții sunt considerați egoiști, este foarte important de reținut că un **echilibru Nash** ar putea să nu fie eficient, iar un set eficient de strategii ar putea să nu fie un echilibru Nash.

Deși jocurile pot fi reprezentate formal în mai multe moduri (formă normală, formă extensivă etc.), cea mai simplă formalizare și cea mai frecvent utilizată în domeniul nostru de activitate este Forma Normală.

Un joc în formă normală este un tuplu $< N, A, u >$. Aici $N = \{1, 2, ...n\}$ este mulțimea finită a jucătorilor, $A = A_1 \times A_2 \times ... \times A_n$ este spațiul de acțiuni agregate al tuturor agenților, având fiecare A_i ca mulțime de acțiuni posibile pentru jucătorul i. Atunci, u este mulțimea funcțiilor de utilitate $(u_1, u_2, ..., u_n)$. Fiecare u_i este definit ca $u_i : A \mapsto \mathbb{R}$, mapând un profil de acțiune comună la o anumită valoare numerică a recompensei.

Un profil de strategie (*(eng. strategy profile)*) $\sigma_i \in \Delta(A_i)$ definește modul în care un jucător alege acțiunile, unde $\Delta(A_i)$ este mulțimea fiecărei distribuții posibile de probabilități pe A_i. Dacă o astfel de distribuție conține un singur element cu probabilitatea 1, se numește strategie pură, în timp ce toate celelalte strategii se numesc strategii mixte. Un profil de acțiune a este un set de acțiuni comune $(a_1, a_2, ..., a_n)$ pentru toți jucătorii. Frecvent, notația (a_i, a_{-i}) sau (σ_i, σ_{-i}) este utilizată pentru a desemna acțiunile/strategiile alese de i în contrast cu acțiunile/strategiile alese de toți ceilalți jucători, astfel încât $a_{-i} = (a_1, ...a_{i-1}, a_{i+1}, ...a_n)$ și $\sigma_{-i} = (\sigma_1, ...\sigma_{i-1}, \sigma_{i+1}, ...\sigma_n)$.

Având aceste definiții, un profil de strategie σ_i^* este un cel mai bun răspuns la σ_{-i}, dacă:

$$u_i(\sigma_i^*, \sigma_{-i}) \geq u_i(\sigma_i, \sigma_{-i}), \forall \sigma_i \in \Delta(A_i) \tag{1.4}$$

Dacă formula 1.4 este adevărată pentru toți jucătorii, strategia σ este un echilibru Nash.

Teoria jocurilor a avut un rol esențial în descoperirea unor perspective valoroase asupra diverselor situații sociale, care pot fi modelate riguros prin intermediul setului său de instrumente. Cele mai cunoscute dintre acestea sunt numite dileme sociale, în care binele colectiv și interesul individual al jucătorilor se află într-un conflict constant. Aceste situații prezintă un interes deosebit și în MARL.

În contextul învățării, conceptele de echilibru servesc atât ca repere, cât și ca principii directoare pentru evaluarea și interpretarea comportamentului agenților. Deși jocurile pot fi reprezentate în mai multe moduri (cum ar fi forma extensivă, forma grafică sau formulări bazate pe stări), ne concentrăm pe jocurile de formă normală, în care fiecare jucător selectează o singură acțiune per interacțiune dintr-un set fix de acțiuni reprezentat de așa-numitele matrici de plată. Această reprezentare nu este doar compactă și expresivă pentru jocurile cu mutări simultane, ci se pretează bine și la parametrizare și eșantionare eficientă, care sunt esențiale pentru experimentarea la scară largă, aplicațiile MARL și meta-învățare.

Este de o importanță capitală să se distingă tipurile de dileme sociale luate în considerare, iar acest lucru se realizează cel mai clar printr-o comparație între dilemele sociale care au primit cea mai mare atenție academică, permițând o înțelegere mai profundă a jocurilor în cauză. În timp ce numeroase modele teoretice ale jocurilor își propun să surprindă dilema binelui individual versus cel colectiv, fiecare model o surprinde în mod distinct. Pentru a clarifica aceste nuanțe, vom analiza pe scurt structurile de stimulare din cele mai faimoase modele, cum ar fi Dilema Prizonierului, Jocul Bunurilor Publice (PGG), Stag Hunt și jocul Chicken [FJW24].

Poate cea mai cunoscută, dilema prizonierului, prezintă o situație în care fiecare jucător poate alege între non-cooperare și cooperare. Aici, cooperanții sunt întotdeauna într-o situație strict mai precară decât transfugii. În schimb, jocul Stag Hunt prezintă o situație în care cooperarea este benefică, dar mai riscantă decât non-cooperarea și, deși cooperarea este, de asemenea, un echilibru Nash, este considerabil mai dificil de realizat. Jocul Chicken contrastează alegerea cooperării cu prețul unui câștig mai moderat și escaladarea unui conflict cu prețul unui risc mai semnificativ. Jocul Bunurilor Publice modelează provocarea furnizării colective de resurse din contribuții individuale, unde jucătorii decid dacă să coopereze pentru un bine comun care să beneficieze toți. Spre deosebire de Dilema prizonierului, unde non-cooperanții exploatează direct cooperanții, în PGG, fiecare jucător primește același beneficiu, indiferent de contribuția sa. Totuși, similar jocului Stag Hunt, cooperarea este suboptimală la nivel individual și, deși nu există o confruntare directă precum în jocul Chicken, se creează totuși un stimulent pentru a profita de contribuțiile altora, subliniind un aspect diferit al interesului personal împotriva binelui comun.

Capitolul 2

Abordări bazate pe învățarea prin întărire pentru cooperare în dilemele sociale

Cooperarea ca fenomen social, în regnul animal și în societățile umane, este adesea privită ca un puzzle evolutiv. De ce ajung indivizii să coopereze, adesea sacrificându-și propriul bine, când a profita de semenii lor pare adesea mult mai rațional? De ce cooperarea eșuează în alte cazuri în mod spectaculos, prăbușindu-se în indivizi egoiști și fără minte care distrug binele comun? Care sunt perspectivele cooperării într-o societate din ce în ce mai mult integrată în sisteme cu agenți artificiali?

Căutarea înțelegerii cooperării și, eventual, a încurajării acesteia în anumite situații este continuă de zeci de ani. Întrebarea este presantă din mai multe direcții: înțelegerea cooperării are un impact direct asupra sistemelor politice și economice, afectând stabilitatea, relațiile internaționale, războaiele comerciale și gestionarea bunurilor publice epuizabile. În al doilea rând, din punctul de vedere al biologiei și psihologiei, este important să înțelegem cum cooperează ființele non-raționale precum insectele și animalele, dar și motivațiile și condițiile în care oamenii dezvoltă un comportament cooperativ. Cea mai relevantă pentru teza noastră este însă perspectiva inteligenței artificiale: pe măsură ce era automatizării înflorește, sunt implementați din ce în ce mai mulți agenți autonomi, care trebuie integrați cu succes într-un sistem în care agenții trebuie să colaboreze între ei și cu alți oameni.

Obiectivul principal al acestui capitol este de a prezenta două rezultate în direcția găsirii unui comportament cooperativ comun în sistemele multi-agent bazate pe agenți RL. De la munca de pionierat a lui Axelrod [Axe81] în Dilema Prizonierului, domeniul de cercetare al simulărilor bazate pe agenți a devenit din ce în ce mai popular. Adăugarea algoritmilor de învățare la această tendință a adus multe rezultate interesante. Majoritatea acestor lucrări explorează modul în care unii agenți de învățare pot coopera sub regulile anumitor dileme sociale, cum ar fi Dilema Prizonierului, Jocul Bunurilor Publice, Jocul SnowDrift sau Jocul Stag Hunt. Cu toate acestea, majoritatea abordărilor bazate pe agenți care învață nu reușesc să convergă către echilibre cooperative. În acest scop, multe mecanisme suplimentare au fost adăugate la configurațiile acestor sisteme, cum ar fi alterarea regulilor jocului, modificarea agenților înșiși sau structurii de recompensă, cu scopul de a încuraja apariția și susținerea cooperării.

Prima noastră lucrare, bazată pe abordarea noastră originală din [Mal22] asupra jocului Dilema Prizonierului, este o astfel de abordare: introducem o abordare bazată pe agenți în rețea, în care interconexiunile agenților sunt modificate astfel încât să încurajeze cooperarea. A doua noastră abordare, aplicată Jocului Bunurilor Publice, bazată pe lucrarea noastră originală din [Mal25], nu utilizează însă mecanisme suplimentare și totuși atinge rate considerabile de cooperare.

Capitolul 3

Benchmark-uri propuse și aplicații

Domeniul MARL a fost subiectul unei dezvoltări rapide în ultimii ani. În acest domeniu s-au dezvoltat perspective semnificative nu numai din aplicațiile la probleme concrete din lumea reală, ci și din domenii de referință (*benchmarks*). Au fost dezvoltate diferite sisteme MARL promițătoare și de succes în domeniile rețelelor celulare, vederii computerizate, conducerii autonome, controlului traficului, managementului energiei, cloud computing și multe altele. În plus, benchmark-urile propuse, cum ar fi mediile Gym [BCP+16] și jocuri precum Atari, șah, Go, poker, au fost extrem de utile în evaluarea performanței MARL, menținând în același timp experimentele comparabile și reproductibile. Asemenea benchmark-uri oferă setări care sunt extrem de controlabile, ușor de utilizat și care reduc costurile inginerești.

Așadar, intenția acestui capitol este de a prezenta două rezultate ale lucrării noastre din direcția de cercetare a benchmark-urilor și aplicațiilor.

În abordarea noastră originală, [Mal24], prezentăm o aplicație directă a MARL bazat pe teoria jocurilor în domeniul planificării sarcinilor. Aceste rezultate demonstrează că modelele teoretice ale jocurilor depășesc simpla analiză teoretică, iar sistemele MARL bine elaborate, inspirate de teoria jocurilor, care prioritizează în mod explicit aspecte precum cooperarea prin utilizarea Jocurilor de tip Agent-Environment Cycle, pot avea un impact semnificativ asupra performanței problemelor din lumea reală, cum ar fi domeniul planificării sarcinilor în sistemele cloud.

În al doilea rând, detaliem propunerea noastră de benchmark pentru algoritmii MARL, bazată pe abordarea noastră originală introdusă în [MC23]. Această propunere se inspiră din jocul clasic Piatră-Hârtie-Foarfecă și oferă un mediu flexibil și controlabil pentru testarea agenților în contexte în care trebuie învățate strategii mixte. În plus, trebuie remarcate două contribuții originale cheie ale acestei lucrări: pe de o parte, generalizarea reușită și solidă a jocului la un număr arbitrar de agenți și acțiuni, menținând în același timp structura motivațională a jocului și, în al doilea rând, demonstrația matematică a existenței și identificării echilibrelor Nash ale jocului Piatră-Hârtie-Foarfecă generalizat.

Capitolul 4

Învățare prin meta-întărire în jocuri multi-agent

Interacțiunile în sistemele multi-agent sunt adesea structurate prin intermediul instrumentelor teoriei jocurilor; cu toate acestea, în scenariile din lumea reală, structura și parametrii jocului subiacent cu care se confruntă agenții sunt frecvent necunoscuți sau nestaționari. Aceasta reprezintă o provocare critică: agenții trebuie să deducă rapid natura mediului lor și să își adapteze strategiile în consecință, chiar și în prezența mai multor alți agenți.

Meta-învățarea prin întărire (meta-RL) a demonstrat capacitatea de a facilita adaptarea rapidă în sarcini precum jocuri multi-agent, procese decizionale Markov și navigare vizuală. În acest capitol, prezentăm rezultatele lucrării noastre originale [Mal26], în care extindem aplicarea meta-RL la jocuri cu mai mulți agenți. Prin antrenarea agenților meta RL bazate pe auto-interacțiune (*self-play meta-reinforcement learning*) pe diverse clase de jocuri cu formă normală, parametrizate de matricile lor de recompense și eșantionate dintr-o distribuție, dezvoltăm algoritmi care nu sunt doar eficienți din punct de vedere al eșantionării (*sample-efficient)* și robuști la schimbări, ci și capabili de generalizare strategică în structuri distincte ale teoriei jocurilor. Deși rămâne limitată la o demonstrație teoretică a conceptului, abordarea noastră reduce decalajul dintre modelarea clasică a teoriei jocurilor și tehnicile moderne de meta-învățare, cu implicații promițătoare pentru comportamentul adaptiv în medii dinamice cu mai mulți agenți.

Concluzii

Învățarea cu întărire este un domeniu de cercetare omniprezent, care a avut o influență semnificativă asupra multor alte domenii. Întrucât învățarea prin întărire (RL) nu este constrânsă de cunoștințele domeniului, ceea ce marchează de obicei un platou limită pentru tehnicile de învățare supravegheată, se spune că RL are un potențial foarte ridicat pentru construirea de sisteme care depășesc performanța umană, în special în domeniile în care se dorește învățarea și comportamentul similar cu cele umane, dar în care intervenția umană manuală era anterior imposibilă.

Cu toate acestea, domeniul încă mai are numeroase probleme deschise și încă așteaptă soluții pentru diverse probleme de interes real. În mod similar, pentru multe probleme care utilizează deja tehnici de RL există încă o multitudine de arii în care se pot face îmbunătățiri sau unde sunt necesare investigații suplimentare, fără a mai menționa faptul că chiar și teoria RL în sine prezintă multe probleme deschise, ale căror soluții finale ar putea avea un impact semnificativ asupra evoluției domeniului.

Mai exact, aplicarea învățării cu întărire în modelele teoretice ale jocurilor în sisteme multi-agent are implicații de anvergură în problemele lumii reale, care se extind mult dincolo de cercetarea teoretică. Adoptarea tot mai mare a agenților cu inteligență artificială face ca importanța modelelor bazate pe teoria jocurilor să fie din ce în ce mai mare. Aplicabilitatea și potențialul RL sunt substanțiale în sfera digitală, fie că este vorba de optimizarea recomandărilor sau de gestionarea dinamică a resurselor. Capacitatea metodelor RL de a naviga și optimiza în medii complexe și în schimbare, cum ar fi logistica și eficiența lanțului de aprovizionare, le face un instrument atractiv de utilizat. Domeniul medical și cel financiar sunt, de asemenea, puternic influențate: optimizarea planurilor de tratament și echilibrarea nevoilor pacienților, gestionarea automată a bugetelor și proiectarea planurilor de investiții care se adaptează cu succes la condițiile de piață în schimbare au, de asemenea, un impact ridicat. Planurile educaționale adaptate RL, cercetarea bazată pe RL privind probleme de tip protein folding și creșterea sistemelor de conducere autonomă dovedesc disponibilitatea tot mai mare a societății de a delega din ce în ce mai mult control inteligenței artificiale, o mare parte din aceasta fiind determinată de RL.

Totuși, se poate observa tendința că multe dintre domeniile menționate anterior încep să utilizeze posibil mai mulți agenți decizionali, ale căror coordonare, cooperare și/sau concurență reprezintă o problemă care crește odată cu adoptarea. În acest context, problemele care apar din cauza interacțiunii mai multor agenți, cum ar fi nestaționaritatea sistemelor multi-agent, devin cruciale de gestionat.

De asemenea, trebuie obținute informații despre motivatoa intrinseci ale acestor agenți, iar motivațiile agenților trebuie înțelese pentru a avea un control granular asupra lor. În mod similar, este o nevoie de a înțelege transparența acestor sisteme pentru a se asigura că funcționarea lor respectă reglementări de siguranță și sunt aliniate cu scopuri umane. Modelele teoretice ale jocurilor ajută enorm la abordarea și investigarea acestor probleme, descriind scenarii care apar în situații reale, fără a fi necesară prezentarea tuturor detaliilor problemelor specifice, precum cele enumerate anterior. Probleme precum transparența, adaptabilitatea, cooperarea, coordonarea, conflictul și motivația sunt preocupări transversale în toate sistemele multi-agent, care pot fi surprinse cu succes de modelele bazate pe

teoria jocurilor.

În această teza ne-am propus să prezentăm contribuțiile noastre originale în utilizarea încătării cu întărire cu accent specific pe deep RL în modelele bazate pe teoria jocurilor, respectiv modelele noastre proprii. Aceste rezultate promițătoare constituie doar o fracțiune dintr-un spectru mult mai larg, deoarece considerăm că întregul potențial al DRL în modelele teoretice ale jocurilor nu a fost încă pe deplin valorificat.

Ne-am concentrat nu doar asupra modului în care aceste modele pot fi exploatate pentru a valorifica procesele decizionale în scenarii cu mai mulți agenți la nivel abstract și generic, ci am arătat și cum încorporarea lor în aplicații din lumea reală poate duce la mai multă claritate și îmbunătățiri ale performanței, cum ar fi programarea sarcinilor în cloud și, în final, am prezentat potențialul modelelor de meta-învățare în aceste contexte.

Varietatea largă de probleme în care MARL este aplicabil este un argument autosuficient pentru potențialul său în construirea unor sisteme de agenți cu adevărat inteligente din punct de vedere social, în timp ce rezultatele obținute, prezentate aici, demonstrează semnificația sa incontestabilă.

Bibliografie

[ADBB17] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[AH81] Robert Axelrod and William D Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

[Axe81] Robert Axelrod. The emergence of cooperation among egoists. *The American Political Science Review*, pages 306–318, 1981.

[BBC+19] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[BBD18] Juan C Burguillo, Burguillo, and Ditzinger. *Self-organizing coalitions for managing complexity*. Springer, 2018.

[BCP+16] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[BKM+19] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*, 2019.

[BO98] Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

[BPB11] Ana LC Bazzan, Ana Peleteiro, and Juan C Burguillo. Learning to cooperate in the Iterated Prisoner's Dilemma by means of social attachments. *Journal of the Brazilian computer society*, 17(3):163–174, 2011.

[Byu14] Chong Hyun Christie Byun. The Prisoner's Dilemma and Economics 101: Do Active Learning Exercises Correlate with Student Performance? *Journal of the Scholarship of Teaching and Learning*, 14(5):79–91, 2014.

[Cow12] CC Cowden. Game theory, evolutionary stable strategies and the evolution of biological interactions. *Nature Education Knowledge*, 3(6), 2012.

[Daw16] Richard Dawkins. *The extended selfish gene*. Oxford University Press, 2016.

[DCH+16] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pages 1329–1338. PMLR, 2016.

[FFA+17] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. corr abs/1705.08926 (2017). *arXiv preprint arXiv:1705.08926*, 2017.

[FJW24] Shaheen Fatima, Nicholas R Jennings, and Michael Wooldridge. Learning to resolve social dilemmas: a survey. *Journal of Artificial Intelligence Research*, 79:895–969, 2024.

[FS03] Tobias Frick and Stefan Schuster. An example of the prisoner's dilemma in biochemistry. *Naturwissenschaften*, 90(7):327–331, 2003.

[Har09] Garrett Hardin. The tragedy of the commons. *Journal of Natural Resources Policy Research*, 1(3):243–253, 2009.

[HM23] Dom Huh and Prasant Mohapatra. Multi-agent reinforcement learning: A comprehensive survey. *arXiv preprint arXiv:2312.10256*, 2023.

[HPHKPI20] Manuel Herrera, Marco Pérez-Hernández, Ajith Kumar Parlikad, and Joaquín Izquierdo. Multi-agent systems and complex networks: Review and applications in systems engineering. *Processes*, 8(3):312, 2020.

[HS16] Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

[HY20] Ammar Haydari and Yasin Yılmaz. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):11–32, 2020.

[JSB02] Dominic DP Johnson, Pavel Stopka, and Josh Bell. Individual variation evades the Prisoner's Dilemma. *BMC Evolutionary Biology*, 2(1):1–8, 2002.

[KMP13] Soummya Kar, José MF Moura, and H Vincent Poor. QD-Learning: A Collaborative Distributed Strategy for Multi-Agent Reinforcement Learning Through Consensus. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.

[Li19] Yuxi Li. Reinforcement learning applications. *arXiv preprint arXiv:1908.06973*, 2019.

[Lit94] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[LLL+19] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

[LR00] Martin Lauer and Martin Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer, 2000.

[LSB+23] Marc Lanctot, John Schultz, Neil Burch, Max Olan Smith, Daniel Hennes, Thomas Anthony, and Julien Perolat. Population-based Evaluation in Repeated RPS as a Benchmark for Multiagent Reinforcement Learning. *arXiv preprint arXiv:2303.03196*, 2023.

[LWT+17] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[LZG+17] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[Mal22] Imre Gergely Mali. Finding cooperation in the n-player iterated prisoner's dilemma with deep reinforcement learning over dynamic complex networks. In *26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES) - Procedia Computer Science*, volume 207, pages 465–474, 2022.

[Mal24] Imre-Gergely Mali. Multi-Agent Deep Reinforcement Learning for Collaborative Task Scheduling. In *Proceedings of the 16th International Conference on Agents and Artifcial Intelligence (ICAART 2024)*, pages 1076–1083, 2024.

[Mal25] Imre Gergely Mali. A sequence-modeling approach to cooperation in public goods games via multi-agent transformers. *Procedia Computer Science*, 270:3708–3717, 2025.

[Mal26] Imre Gergely Mali. Self-play meta-reinforcement learning in multi-agent games. *Acta Universitatis Sapientiae, Informatica*, 18(1), feb 2026.

[MAMK16] Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM workshop on hot topics in networks*, pages 50–56, 2016.

[MC23] Imre-Gergely Mali and Gabriela Czibula. Policy-Based Reinforcement Learning in the Generalized Rock-Paper-Scissors Game. In *Proceedings of the 31th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2023)*, pages 345–350, 2023.

[Mye91] R Myerson. Game Theory: Analysis of Conflict Harvard Univ. *Press, Cambridge*, 1991.

[N+50] John F Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[NI55] Hukukane Nikaidô and Kazuo Isoda. Note on non-cooperative convex games. 1955.

[NX24] Zepeng Ning and Lihua Xie. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence*, 3(2):73–91, 2024.

[PBARA14] Ana Peleteiro, Juan C Burguillo, Josep Ll Arcos, and Juan A Rodriguez-Aguilar. Fostering cooperation through dynamic coalition formation and partner switching. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 9(1):1–31, 2014.

[pd16] The Axelrod project developers. Axelrod: v4.10.0, April 2016.

[Pic19] Elvis Picardo. The Prisoner's Dilemma in business and the economy. *Retrieved May*, 25:2019, 2019.

[RN16] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Sha53] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[SHS+17] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[Smi79] John Maynard Smith. Game theory and the evolution of behaviour. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):475–488, 1979.

[SP73] J Maynard Smith and George R Price. The logic of animal conflict. *Nature*, 246(5427):15–18, 1973.

[SRAA+11] Norman Salazar, Juan A Rodriguez-Aguilar, Josep Ll Arcos, Ana Peleteiro, and Juan C Burguillo-Rial. Emerging cooperation on complex networks. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 669–676, 2011.

[SRDW+19] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

[SSS+17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[SWD+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[TBG+21] J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

[Wen19] Lilian Weng. Meta reinforcement learning. *lilianweng.github.io*, 2019.

[WKL+22] Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems*, 35:16509–16521, 2022.

[WYWH18] Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. *arXiv preprint arXiv:1806.00877*, 2018.

[YVV+22] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

[ZWLP19] Li Zhang, Wei Wang, Shijian Li, and Gang Pan. Monte carlo neural fictitious self-play: approach to approximate nash equilibrium of imperfect-information games. *arXiv preprint arXiv:1903.09569*, 2019.

[ZYB19] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

[ZYB21] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.