Universitatea POLITEHNICA din București

Facultatea de Automatică și Calculatoare,
Catedra de Calculatoare

REZUMAT

TEZĂ DE DOCTORAT

Explorarea în Invățarea prin Recompensă pentru a rezolva probleme din lumea reală. Explorarea este tot ceea ce avem nevoie.

Conducător Științific:
Prof. Adina Magda Florea

Autor:
Andrei Cristian Nica

București, 2023

Cuprins

Rezumat

Acest document oferă un rezumat cuprinzător al unei teze de doctorat axată pe explorarea în Învățare prin Recompensă (RL, *English:Reinforcement learning*) pentru rezolvarea problemelor din lumea reală. Teza aprofundează aspecte fundamentale ale explorării, care implică căutarea și descoperirea de noi cunoștințe și abilități. Scopul lucrării de cercetare a fost acela de a aborda provocările cu care se confruntă agenții RL în explorare, îmbunătățindu-le în cele din urmă capacitățile de învățare. Pe parcursul tezei, au fost luate în considerare diverse perspective asupra explorării în RL, de la evaluarea progresului de învățare în sarcini cu recompense rare până la valorificarea învățării ierarhice și explorarea mediilor cu agenți multipli. În plus, teza își extinde explorarea în aplicații din lumea reală, cum ar fi navigarea roboților și conducerea autonomă, precum și învățarea automată pentru generarea și descoperirea de noi medicamentelor. Prin examinarea acestor domenii diverse, teza oferă perspective, îmbunătățiri și recomandări practicienilor și cercetătorilor care lucrează în domeniul RL.

Învățare prin Recompensă este o abordare puternică pentru construirea de agenți inteligenți și are potențialul de a fi utilizat într-o mare varietate de aplicații din lumea reală. Cu toate acestea, în ciuda potențialului său și a progreselor recente, există încă multe provocări care trebuie abordate înainte că RL să poată fi adoptat pe scară largă. Explorarea este un aspect fundamental al învățării. Explorarea este actul de căutare și descoperire și este esențială pentru dezvoltarea de noi cunoștințe și abilități. Prin urmare, este o parte integrantă a învățării prin recompensă, deoarece permite agentului să exploreze diferite acțiuni și stări pentru a-și îmbunătăți politica și a se adapta la mediile în schimbare. Scopul acestei teze a fost cercetarea diferitelor metode și strategii menite să abordeze provocările de explorare cu care se confruntă agenții de învățare prin recompensa, îmbunătățindu-le astfel capacitățile de învățare. Activitatea de cercetare include studierea compromisurilor dintre explorare și exploatare, dezvoltarea de noi tehnici de explorare, analiza impactului explorării asupra performanței și învățării, evaluarea și îmbunătățirea eficacității diferitelor metode de explorare în diverse activități și medii. Obiectivul a fost să identifice și să înțeleagă factorii cheie care contribuie la explorarea eficientă în RL și să ofere noi perspective și recomandări pentru practicienii și cercetătorii care lucrează în acest domeniu. Pe parcursul tezei, au fost luate în considerare diverse perspective de explorare și de învățare prin recompensa. Când unui agent i se dă o sarcină cu recompense rare, am căutat modalități de a evalua progresul de învățare al agentului. În timp ce recompensele bazate pe activități nu ar influența schimbările de politică în sarcinile solicitante de explorare, am evaluat recompensele intrinseci pentru îmbunătățirea învățării agenților. Am explorat utilizarea și îmbunătățirea învățării prin consolidare ierarhică pentru sarcini cu orizont lung. Am luat în considerare rezolvarea sarcinilor cu mai mulți agenți de învățare prin întărire și am subliniat dificultățile și soluțiile pentru îmbunătățirea învățării în aceste configurații. De asemenea, am examinat modul în care o perspectivă mulți-agent ar putea spori provocările de explorare cu un singur agent. În cele din urmă, susținem importanța generalizării politicilor în învățarea prin consolidare și relația acesteia cu problema de explorare. Pe lângă cercetarea învățării prin întărire în domenii simulate, motivată de proiectarea atentă a ipotezelor investigate, am luat în considerare și două probleme principale din lumea reală care ridică problema explorării la un nou plan. Pe de o parte, luăm în considerare politicile de învățare pentru navigația roboților și vehiculele autonome. Pe de altă parte, cercetăm învățarea automată pentru generarea și descoperirea de noi medicamente.

1 Introducere

Inteligența artificială (AI, *English: Artificial Intelligence*) a început să prindă contur în anii 1950, când denumirea a devenit populară printre oamenii de știință. În 1956, John McCarthy a inventat termenul de „inteligență artificială" și a stabilit prima conferință AI. Chiar și înainte, ideea de a construi „inteligență" a fost cel puțin o idee contemplata multe decenii, dar de atunci, a devenit într-o oarecare măsură chiar un scop de urmărit științific. O idee promițătoare și intrigantă, că am putea noi înșine să construim mașini inteligente, care ar putea avea potențialul de a rezolva probleme dincolo de ceea ce oamenii ar putea rezolva, atât în scară, cât și în complexitate.

Învățarea automată (ML, EN: Machine Learning) a făcut progrese semnificative în ultimele decenii, iar în ultimii ani, domeniul a avut parte de dezvoltarea multor tehnici și tehnologii noi care au extins foarte mult capacitățile. Câteva exemple ar fi: rețele generative adversare (GAN, EN: Generative adversarial networks) [12], Transformers [54], rețele convoluționale grafice (GCN, EN: Graph convolutional networks)

[23] antrenarea adversariala [13], invățare federată [29], invățare auto-supravegheată [52], căutare automata de arhitecturi neuronală (NAS) [62], automatizarea invățare automate [51] (AutoML), formare bazată pe populație [18], modele de difuzie [45]. Una dintre cele mai importante evoluții în domeniu, care a stat la baza progreselor menționate mai sus, este apariția "Învățării adânci" (Deep Learning) [25]. Aceasta este un tip de învățare automată care utilizează rețele neuronale mari pentru a învăța din date. Acest lucru a condus la crearea multor modele puternice de învățare automată care sunt capabile să atingă performanțe de ultimă generație pentru o gamă largă de sarcini, cum ar fi înțelegerea imaginii, a limbajului și a vorbirii. Folosind acesta abordare putem rezolva probleme pe care nu le-am rezolvat înainte. Învățarea automată a avansat semnificativ în ultimul deceniu datorită unui număr de factori, inclusiv disponibilitatea unor cantități mari de date, dezvoltarea unui hardware de calcul mai puternic și creșterea cercetării interdisciplinare și a colaborării între diferite domenii. Unul dintre factorii cheie care a contribuit la progresul învățării automate în ultimii ani este disponibilitatea unor cantități mari de date. Algoritmii de învățare automată necesită date din care să învețe, iar disponibilitatea unor seturi de date mari și diverse a permis cercetătorilor să antreneze modele mai precise și mai sofisticate. Acest lucru a fost facilitat de adoptarea pe scară largă a tehnologiilor digitale, care a condus la generarea de cantități mari de date într-o gamă largă de domenii, inclusiv inbunatatirea sănătății, finanțe, transporturi și rețelele sociale. Celălalt factor foarte important care a contribuit la progresul învățării automate este dezvoltarea unui hardware de calcul mai puternic. Algoritmii de învățare automată necesită foarte multă putere de calcul, iar dezvoltarea unor procesoare mai rapide și mai eficiente, precum și utilizarea pe scară largă a unităților de procesare grafică (GPU-uri) și a altor hardware-uri specializate, au făcut posibilă instruirea modele mai mari și mai complexe într-un timp mai scurt. În cele din urmă, progresul învățării automate în ultimii ani a fost, de asemenea, facilitat de creșterea cercetării interdisciplinare și a colaborării între diferite domenii. Învățarea automată are aplicații într-o gamă largă de domenii, inclusiv informatică, statistică, biologie, fizică și inginerie, iar creșterea cercetării interdisciplinare a permis cercetătorilor din diferite domenii să împărtășească idei și să colaboreze la probleme comune. Acest lucru a condus la dezvoltarea de noi algoritmi și tehnici care au fost aplicate pentru o gamă largă de probleme din lumea reală.

Învățarea prin Recomensa este un tip de tehnică de învățare automată care permite unui agent să învețe într-un mediu interactiv prin încercări și erori, folosind feedback-ul provenit din mediu (recompensa), acțiunile și experiențele sale proprii. În timp ce învățarea automată este un subdomenie a inteligenței artificiale, care se referă în general la algoritmi care permit mașinilor să învețe și să se adapteze continuu la date "neasistati", învățarea prin recompensa stabilește un paradigmă generală în care mașinile, sau agenții, învață să acționeze într-un mediu în așa fel încât să maximizeze un semnal de recompensă. Originile învățării prin recompensa pot fi atribuite următoarelor două subiecte studiate independent: o zonă ce investiga soluții pentru problemele de control optimal folosind funcții de valoare și programare dinamică, în timp ce a doua zonă majoră poate fi identificată în studiile psihologice ale învățării animalelor, unde au fost propuse modele de învățare prin încercari și erori. O istorie mai detaliată a învățării prin recompensa este descrisă de Sutton și Barto (1998) ([46]).

În această teză ne-am concentrat asupra aspectului explorării în rezolvarea problemelor cu învățarea profundă prin recompensa (DRL, EN: Deep Reinforcement Learning). Învățarea profundă prin recompensa combină tehnicile de învățare prin consolidare și învățarea profundă, folosind rețele neurale că modele computaționale pentru învățarea comportamentului agenților. Explorarea, în contextul învățării prin consolidare, se referă la procesul prin care un agent învață despre mediul său și adună informații care pot fi folosite pentru a lua decizii mai bune. Acest lucru poate implica luarea unor acțiuni care nu conduc imediat la cele mai mari recompense, dar care pot furniza agentului informații valoroase despre mediul său. De exemplu, un agent care învață să navigheze într-un labirint poate lua acțiuni care îl duc pe căi necunoscute, în scopul de a aduna mai multe informații despre configurația labirintului și de a-și îmbunătăți performanța în viitor. Explorarea este un aspect important al învățării prin recompensa, deoarece permite agenților să învețe și să se adapteze la mediul lor într-un mod mai eficient. De asemenea, ajută la abordarea echilibrului dintre exploatare (utilizarea cunoștințelor deja dobândite de către agent pentru a maximiza recompensele) și explorare (adunarea de informații noi pentru a-și îmbunătăți performanța). În aceasta teza: vom analiza unele dintre provocările explorării în învățarea prin consolidare pentru rezolvarea diferitelor probleme, vom cerceta perspective diferite în rezolvarea acestor provocări, vom propune noi modalități de îmbunătățire a învățării și direcții promițătoare de cercetare în viitor.

2 Motivație

Învățarea prin Recompensa a obținut deja rezultate impresionante în multe aplicații, inclusiv jocuri (Go, StarCraft, Dota) [43, 55, 56, 4], robotică (prinderea obiectelor), procesarea limbajului natural (dialog îmbunătățit folosind feedback uman) [34] și chiar în rezolvarea mai multor probleme din lumea reală [9, 61]. Acest succes întărește ideea că învățarea prin consolidare poate fi o metodă puternică și eficientă pentru abordarea unei varietăți de probleme din lumea reală. În timp ce algoritmii de învățare automată au fost aplicați într-o gamă largă de sarcini și industrii, inclusiv finanțe, îngrijirea sănătății, transport și social media, învățarea prin consolidare, pe de altă parte, are potențialul de a fi aplicat în probleme din lumea reală mai, mult mai complexe și dinamice, cum ar fi mașinile autonome și descoperirea de medicamente. Atât învățarea supervizată, cât și învățarea prin consolidare pot fi eficiente în rezolvarea unor sarcini sau probleme specifice, dar nu implică neapărat inteligența în același mod în care o poate avea un om sau un animal. Inteligența este un concept complex și multifacetic, iar nu există un consens în ceea ce privește o definiție precisă. Cu toate acestea, mulți cercetători ar fi de acord că inteligența adevărată implică capacitatea de a gândi și raționaliza, de a învăța din experiență, de a se adapta la situații noi și de a înțelege și folosi limbajul. Învățarea prin recompensa cuprinde multe dintre aceste idei și are potențialul de a permite dezvoltarea de sisteme inteligente care pot învăța și se pot adapta la mediu într-un mod mai asemănător omului. Învățarea prin recompensă este un cadru general pentru învățare, care ar putea avea multe aplicații importante și ar putea fi chiar suficient ([44]) pentru a învăța inteligența. În plus, flexibilitatea algoritmilor de învățare prin consolidare le permite să fie aplicați atât în medii bazate pe simulare, cât și în mediile din lumea reală, ceea ce le mărește și mai mult aplicabilitatea.

RL este o zonă activă și în plină expansiune a cercetării, cu multe dezvoltări și progrese interesante. Aceasta il face un domeniu palpitant și plin de satisfacții pentru cercetători, nu doar datorită perspectivei sale de impact viitor în știință și societate, ci și datorită oportunităților de învățare pe care le oferă cercetatorilor. Este un domeniu divers al cercetării, capabil să se inspire dintr-o gamă largă de domenii. Algoritmii sunt adesea inspirați de evoluții și perspective dintr-o varietate de domenii, inclusiv psihologie, neuroștiințe, teoria controlului și informatică. De exemplu, conceptul de învățare prin recompensa se bazează puternic pe domeniul psihologiei, unde a fost studiat pentru prima dată în contextul învățării animalelor. În plus, dezvoltarea algoritmilor moderni de învățare prin consolidare a fost puternic influențată de domeniul teoriei controlului, care furnizează un cadru matematic pentru analizarea comportamentului sistemelor care primesc intrări externe. Cercetarea în acest domeniu ne poate ajuta, de asemenea, să înțelegem mai bine modul în care sistemele inteligente învață și iau decizii. Această înțelegere poate oferi indicii valoroase despre natura inteligenței și a oamenilor.

Dilema de explorare versus exploatare este o provocare fundamentală în RL. Acest lucru se datorează faptului că, în multe situații, un agent trebuie să aleagă între explorarea mediului său pentru a aduna noi informații sau exploatarea cunoștințelor sale existente pentru a-și maximiza recompensele. Dilema apare deoarece explorarea poate fi riscantă și poate să nu conducă la recompense imediate, în timp ce exploatarea îi permite agentului să-și folosească cunoștințele existente pentru a-și maximiza recompensele pe termen scurt. Cu toate acestea, dacă agentul își exploatează doar cunoștințele existente, poate pierde oportunitățile de a învăța noi strategii care ar putea duce la recompense și mai mari pe termen lung. Că rezultat, găsirea echilibrului potrivit între explorare și exploatare este o provocare crucială în învățarea prin consolidare. Explorarea este esențială pentru că agenții de învățare prin întărire să poată învăța să se adapteze noi medii. Explorarea poate duce la soluții mai bune la problema în cauză, precum și la un comportament mai robust și mai fiabil. Prin studierea explorării, cercetătorii pot dezvolta strategii și algoritmi care pot ajuta agenții să atingă echilibrul potrivit între aceste două obiective.

Deși cercetarea învățării prin întărire a obținut rezultate incredibile în ultimii ani, credem că încă nu și-a demonstrat potențialul, confruntându-se cu provocări serioase pentru scalarea la probleme mai mari, din lumea reala [10]. Unele dintre provocările cheie includ eficiența eșantionului de experiențe, generalizarea, explorarea în general și în siguranță. Aceste provocări deduc puternic necesitatea cercetării în explorare. Se recunoaște, de asemenea, că multe dintre rezultatele în domeniu au fost atinse până acum datorită amplorii de calcul, a dimensiunii datelor și a utilizării rețelelor neuronale că extractoare de caracteristici. Având în vedere dificultatea de a învăța soluțiile cu RL, se ridică astfel întrebarea dacă doar scalarea acestor algoritmi nu este de ajuns și cel puțîn nu este de dorit, ceea ce motivează noi dezvoltări și schimbări mai fundamentale în domeniu.

Credem că această formă de ghidare a proceselor de luare a deciziilor în timp real, care reprezintă

caracteristică central a Învățării prin Recompensă, vă constitui în viitor fundamentul algoritmilor de rezolvarea a multor probleme din lumea reală. Acești algoritmi vor fi construiți probabil prin combinarea învățării nesupravegheată și supravegheată, iar utilizarea recompenselor pentru ghidarea procesul de învățare va fi o abordare fundamentală și necesară scalabilității pentru atingerea obiectivelor în viitor.

3 Obiective

Obiectivul principal al acestei teze a fost de a investiga problema explorării în algoritmii de învățare prin recompensă și de a explora noi abordări pentru rezolvarea acestei provocări. În teza am avut o abordare holistică pentru analiză acestei întrebări de cercetare. Există mai multe motive pentru care ar fi importantă o astfel de abordare în cercetărea acestei probleme.

În primul rând, explorarea este o problemă complexă și cu mai multe fațete, care implică o serie de factori și considerații diferiți. Examinând problema explorării din mai multe unghiuri și perspective, este posibil să obținem o înțelegere mai profundă și mai cuprinzătoare a provocărilor și oportunităților implicate în această sarcină.

În al doilea rând, o abordare holistică poate ajută la identificarea conexiunilor și relațiilor dintre diferitele aspecte ale problemei și la identificarea de noi abordări și strategii pentru abordarea acesteia. Acest lucru poate fi deosebit de important în situațiile în care abordările tradiționale pot să nu fie eficiente, caz în care sunt necesare soluții noi.

În cele din urmă, o abordare holistică poate ajută la asigurarea faptului că cercetarea privind explorarea în învățarea prin recompensă este relevantă și aplicabilă unei game largi de probleme și aplicații din lumea reală. Luând în considerare numeroasele fațete ale explorării, este posibil să identificăm abordări care sunt generalizabile și care pot fi aplicate într-o varietate de contexte diferite.

În general, o abordare holistică a cercetării explorării în învățarea prin recompensă poate ajută la furnizarea unei înțelegeri mai profunde și mai cuprinzătoare a acestei probleme complexe și la identificarea unor abordări noi și inovatoare pentru rezolvarea acesteia.

Concret, această teză urmărește:

1. Examinarea rolulului explorării în învățarea prin recomensa și provocările în fața necunoscutelor și incertitudinii.
2. Investigația stadiului actual al explorării în domeniul învățării prin recomensa și identificarea domeniilor în care sunt necesare cercetări suplimentare.
3. Examinarea problemei explorării din perspective diferite, inclusiv recompensă intrinsecă, învățare ierarhică, sisteme mulți-agenți și aplicații din lumea reală.
4. Identificarea oportunităților de îmbunătățire a explorării în învățarea prin recompensă și propunerea unor noi abordări pentru atingerea acestui obiectiv.

În general, această teză își propune să contribuie la înțelegerea mai largă a explorării în învățarea prin recompensă și să ofere noi perspective și abordări pentru diminuare acestei provocări importante.

4 Structura tezei

Capitolul **2** începe cu o scurtă introducere pentru Învățarea prin recompensă, prezentând contextul și informațiile de bază necesare pentru a înțelege problema și abordările potențiale pentru îmbunătățirea acestor tipuri de algoritmi de învățare automată. Acest lucru permite cititorului să înțeleagă mai bine problema și rațiunea din spatele cercetării efectuate. Descrierea cunoștințelor și teoriilor fundamentale ne ajută să stabilim expertiză în domeniu, să identificăm lacunele în literatură existentă și să oferim direcție pentru cercetările viitoare.

În continuare, în teză, cercetăm problema explorării în Învățarea prin recompensă din diferite perspective, în diferite configurații și pentru rezolvarea diferitelor probleme din lumea reală. În fiecare capitol vom oferi un rezumat al cunoștințelor prealabile pentru subiectul respectiv, vom analiză posibilele provocări,

vom identifica diferite perspective de abordare a soluțiilor, vom propune noi metode și vom analiză rezultatele experimentale pentru ideile prezentate.

Algoritmii actuali de învățare prin recompensă pot fi cu ușurință ineficienți în medii cu recompensă foarte rară. În astfel de configurații, recompensă mediului este un semnal foarte slab pentru agenții de a învăța comportamente utile pentru rezolvarea sarcinii. Această este una dintre principalele probleme care ridică provocarea explorării. În capitolul **3** vom începe prin a analiză această configurație și una dintre cele mai comune și intuitive abordări pentru îmbogățirea semnalului de învățare, prin utilizarea recompensei intrinseci. Vom investiga cum să evaluăm progresul învățării în explorare, să comparăm diferite metode și să propunem un semnal de învățare bazat pe cunoașterea ordonării stărilor în timp.

În capitolul **4** vom investiga modul în care învățarea abstracțiilor temporale și utilizarea lor în continuare pentru a rezolva sarcinile din aval poate ajută la explorare. Prezentăm avantajele unui mecanism de atenție față de acțiunile abstracte extinse temporal și propunem un algoritm online pentru învățarea opțiunilor și utilizarea lor în continuare pentru a învățarea sub-obiectivelor.

Examinăm în continuare în capitolul **5** modul prin care agenții de învățare prin recompensă pot învăța să rezolve sarcini împreună cu mai mulți agenți. Mai întâi investigăm abordări specifice și îmbunătățiri utile pentru învățarea prin recompensă cu agenți multiplii, prin îmbunătățirea performanței într-un joc care reprezintă dilema vânătorii de cerb (stag hund dilemma). Apoi investigăm și propunem o modalitate diferită de a consideră problemele cu un singur agent că setări cu mai mulți agenți. Acest lucru poate ajută exploatarea avantajele mai multor agenți, care pot colabora și comunica și pentru a îmbunătăți învățarea prin consolidare cu o explorare mai bună.

În capitolele **6** și **7** cercetarea noastră se concentrează pe rezolvarea problemelor din lumea reală și a noilor provocări care vin odată cu explorarea.

În capitolul **6** luăm în considerare problema navigației în realitate atât pentru roboți, cât și pentru vehiculele autonome. Începem prin a analiză provocările politicilor de învățare cap-coadă (end-to-end) pentru navigarea roboților în spații interioare. În continuare, investigăm modul în care putem folosi simulatoare și în care putem transferăm politicile în realitate, arătand cum memoria și informațiile semantice pot ajută la localizarea în spații interioare. Continuăm abordând procesul necesar pentru construirea și testarea unui vehicul autonom în lumea reală care utilizează modele bazate pe învățarea automată pentru a direcționa mașînă. Descriem mai întâi procesul de colectare și procesare a unui set de date local pentru conducere autonomă. Apoi, prezentăm rezultatele acestui set de date pentru învățarea modelelor end-to-end pentru direcție și propunem un proces semi-automat pentru învățarea politicii de conducere numai din înregistrările camerelor monoculare.

În capitolul **7** studiem problema explorării în contextul descoperirii medicamentelor, în special modul în care algoritmii de învățare automată și de învățare prin recompensă pot identifica potențiali compuși noi de medicamente prin căutarea mai eficientă a spațiului de molecule. Cercetăm și propunem îmbunătățiri atât pentru algoritmii RL, cât și pentru GFlowNet pentru sarcina de a genera molecule mici și analizăm importanța generalizării pentru un design molecular mai bun. În cele din urmă, cercetăm importanța evaluării adecvate a generalizării designului moleculelor bazat pe învățarea automată folosind metrici de simulare pentru a o transferă eficient la afinitatea de andocare în realitate.

În urmă cercetărilor noastre privind provocările explorării în RL, atât din perspective diferite, cât și în diferite configurații ale acestei probleme, propunem în Capitolul **8** un set de opinii discutabile despre ceea ce credem că va fi esențial pentru scalarea învățării prin întărire la probleme complexe din lumea reală. În primul rând, discutăm limitările abordărilor tradiționale ale învățării prin recompensă și de ce acestea ar putea să nu fie potrivite pentru rezolvarea problemelor de spațiu infinit de stări, pentru care propunem un cadru mai general pentru RL. În al doilea rând, subliniem legătura puternică dintre generalizare și explorare în RL, care este esențială pentru progresul algorithmic.

Capitolul **9** încheie lucrarea, prezentând principalele contribuții ale tezei și direcțiile viitoare de cercetare.

5 Rezumatul capitolelor

5.2 Baza teoretică

Acest capitol oferă o introducere în Învățarea prin recompensă (RL), subliniind semnificația acesteuia că subiect de cercetare. Prezintă originile RL și discută motivele fundamentale ale importanței sale. De asemenea, capitolul prezintă elementele de bază și cadrele matematice esențiale pentru înțelegerea RL. Dintr-o perspectivă informatică, evidențiază provocările majore întâlnite la antrenarea unui agent într-o configurație RL. Capitolul explorează diverse abordări, algoritmi și criterii de clasificare care sunt cruciale pentru rezolvarea problemelor RL.

Învățarea prin recompensă este punctul central al capitolului 2, care începe cu o examinare a motivației sale. Capitolul analizează conceptul de comportament al agentului bazat pe recompensă și prezintă configurarea problemei și cadrul de învățare. Acesta explică procesele de decizie Markov (MDP, EN: Markov Decision Processes) și oferă fundalul și notația necesare pentru înțelegerea RL. Sunt explorate provocările teoretice din RL, inclusiv catastrofă dimensionalității, dilema de explorare versus exploatare și probleme legate de observabilitatea parțială, stocasticitatea, discreția, atribuirea creditelor, mediile nestaționare și generalizarea. În plus, capitolul acoperă integrarea învățării profunde în RL, discutând rețelele neuronale, învățarea prin recompensă profundă și provocările asociate cercetărilor viitoare. De asemenea capitoluk subliniază și semnificația explorării în RL.

Abordând aceste aspecte fundamentale, capitolul de fundamte pune bazele pentru o înțelegere mai profundă a RL și creează scenă pentru explorarea în capitolele următoare a subiectelor specifice din domeniu.

5.3 Explorarea cu recompensă intrinsecă

În capitolul următor vom defini și investiga teoriile pentru îmbunătățirea explorării în algoritmii de învățare prin recompensa folosind recompensă intrinsecă. Vom revizui algoritmii de ultimă generație pentru generarea unei astfel de recompense, vom analiza limitările actuale pentru explorare în învățarea prin intarire pe un set de medii eficiente din punct de vedere computațional, vom analiza opțiunile pentru evaluarea progresului învățării în astfel de configurații și vom investiga soluții de îmbunătățire. În cele din urmă, vom introduce un nou mecanism intrinsec de recompensă care valorifică, auto-supravegheat, cunoașterea stărilor de ordonare bazate pe timp pentru a aborda problemele de explorare în RL. Arătăm cum învățarea ordinii temporale a starilor și utilizarea scorului de „ordonare” a unui set permutat aleator ca recompensă intrinsecă determină agenții să convergă către politici mai puțin haotice.

Din punct de vedere al RL-ului, putem defini explorarea ca fiind politica care ajută agentul să interacționeze cu noi stări sau noi tranziții în mediu. Acest lucru l-ar ajuta pe agent să exploreze spațiul problemei suficient pentru a încerca să rezolve sarcini și să nu rămână "blocat" pe parcurs. Pentru această problemă nu ne vom ocupa de dilema explorării versus exploatarii, dar vom încerca să definim o măsură pentru noutatea experienței agentului și să dezvoltăm un algoritm RL pentru maximizarea acestui lucru. Totuși, acest lucru nu respinge exploatarea acțiunilor temporale pentru a debloca noutatea la o rezoluție temporală mai mare.

Problema pe care ne vom concentra în acest capitol este de îmbunătățire a explorări. Un algoritm clasic de învățare prin întărire va încerca să maximizeze recompensă extrinsecă și astfel nu va avea niciun semnal pentru învățare până când nu primește o recompensă. Există multe medii și situații în care recompensele sunt foarte rare și agentul s-ar confrunta cu o explorare aleatorie în speranța de a ajunge la o recompensă. Este ușor de înțeles de ce ne-am dori că un agent inteligent să poată învăța mult mai multe din aceste experiențe chiar și fără nicio recompensă extrinsecă. Ne-am putea îmbunătății algoritmul, de exemplu, în astfel de situații, prin elaborarea manuală a unei euristicii pentru explorare sau prin proiectarea unor funcții suplimentare de recompensă, cunoscute și sub denumirea de modelare a recompensei, pentru a face o problemă mai ușor de învățat. Această însă nu este o soluție, deoarece cercetăm un algoritm universal de învățare prin recompensă care să poată învăța cu cât mai puține cunoștințe anterioare, iar designul maual al funcției de recompensă sau euristică manuale pot chiar împiedica învățarea celei mai bune soluții [1, 17].

Cu toate acestea, putem căuta să transferăm teorii din teoria informației, biologie și alte domenii și să

definim alte obiective pentru agenții noștri. Am putea măsura progresul învățării [41], putem reducem disonanța cognitivă (incompatibilitatea între două sau mai multe structuri cognitive, experiențe sau comportamente [21], să maximizăm împuternicirea [38] (cantitatea de control pe care un agent îl are asupra mediului său), codarea predictivă [33]. Schmidhuber și colab. [40] descriu, de exemplu, că agenții ar trebui să cuprindă mai multe ingrediente: ar trebui să comprime sau să prezică în mod continuu experiențele, să măsoare progresul și să optimizeze pentru o recompensă mai mare prin îmbunătățirea politicilor de acţionare.

Am argumentat până acum de ce RL este un subiect de cercetare important, cum explorarea este o provocare importantă în acest domeniu și am oferit motive pentru care motivația intrinsecă pare să fie prezentată și la oameni. Având în vedere toate acestea, pare să existe o bază bună pentru cercetarea îmbunătățirilor pentru explorare folosind recompensă intrinsecă (IR). Până acum, există diverse moduri care au fost studiate în literatură de specialitate pentru a echipa agenții cu impulsul de a explora medii simulate [6]. De exemplu, agenții pot manifestă curiozitate în funcție de cât de "plictisitor" este un mediu, evitând stările complet previzibile [35]. Alți agenți pot fi clasificați că fiind în căutare de informații. Există studii care analizează conceptul de flux în care un agent încearcă să mențînă o stare în care învățarea este provocatoare, dar nu copleșitoare [**?**]. Principiul energiei libere afirmă că un agent caută să minimizeze incertitudinea prin actualizarea modelului său intern al mediului și selectând acțiuni de reducere a incertitudinii. Împuternicirea se bazează pe principii teoretice informaționale și cuantifică cât de mult control are un agent asupra mediului său.

În acest capitol, am aprofundat o provocare semnificativă cu care se confruntă algoritmii actuali RL în procesul de învățare, și anume explorarea. Am definit obiectivele metodelor de explorare într-o configurație RL și am examinat modul în care acestea sunt abordate în prezent în cercetare. Pentru a măsura evoluția unor astfel de algoritmi, am selectat o suită de medii de referință (7 medii MiniGrid) cu recompense rare, care pot fi folosite pentru a evalua îmbunătățirile folosind resurse de calcul reduse. Ne-am concentrat pe conceptul de recompensă intrinsecă și am explorat modul în care teoriile din teoria informației, învățarea ierarhică, codificarea predictivă și politicile condiționate de obiective pot fi folosite pentru a genera astfel de semnale.

În a două parte a capitolului, oferim un ghid de bază pentru evaluarea performanței algoritmilor cu IR. De asemenea, propunem soluții pentru evaluarea progresului fără a măsura recompensă extrinsecă. Totodată identificăm provocările cu care se confruntă metodele IR în literatură și identificăm unele dintre defectele pe care le prezintă în mediile selectate. Am examinat două metode preexistente de IR pentru îmbunătățirea explorării în rezultatele întârziate sau necunoscute și am prezentat rezultatele noastre.

În secțiunea finală, propunem un nou semnal auto-supravegheat pentru a învăța politici mai puțîn haotice. Mai întâi demonstrăm modul în care scorul de „ordonabilitate” al experiențelor agenților este corelat cu politicile cu o recompensă mai ridicate și cu o explorarea îmbunătățită. Prezentăm apoi rezultate experimentelor în care acest semnal poate fi utilizat că recompensă intrinsecă pentru îmbunătățirea agenților RL în sarcinile de recompensă rare. Oferim dovezi convingătoare că scorul de ordonabilitate nu este doar o măsură utilă pentru analiză progresului de învățare al agentului, ci și pentru a fi folosit că IR, în special în mediile de explorare provocatoare. Este important de remarcat discuția privind definirea progresului învățării pe care o descriem în Sec. **3.4.2.**. În timp ce recompensă intrinsecă rămâne o problemă deschisă și nerezolvată, există mai multe aspecte ale acestei metode care arată un potențial promițător. Deși nu am furnizat dovezi experimentale definitive ale acestor avantaje, am dori să le subliniem după cum urmează. Credem că această metodă are proprietăți de generalizare puternice în comparație cu predicția stării următoare sau alte metode similare de obținere a informațiilor IR. Susținem că acest lucru se datorează proprietăților de invarianță a procesării datorate arhitecturii și spațiului de stări a predicțiilor [19]. Propunem această ipoteză de generalizare îmbunătățită deoarece prezicerea stărilor următoare este o problemă considerabil mai dificilă decât ordonarea observațiilor de intrare. De exemplu, modelele de predicție pentru starea următoare se pot lupta cu observații sau părți ale observațiilor nevăzute până la inferență, în timp ce această metodă poate încă să facă distincția între stările furnizate în intrare și să le ordoneze corect. De asemenea, credem că această metodă nu va fi afectată de problema zgomotului televizorului, deoarece acele secvențe vor avea un scor de ordonabilitate redus.

5.4 Valorificarea învățării ierarhice

Agenții cu inteligență artificală de luare a deciziilor se confruntă adesea cu două obstacole semnificative: amploarea orizontului de planificare și factorul de ramificare care decurge din numeroase alternative. Tehnicile de învățare prin întărire cu structura ierarhică urmăresc să le abordeze pe cele dintâi prin introducerea de comenzi directe care depășesc mai mulți pași de timp. Pentru a gestiona amploarea, este ideal să limităm concentrarea agentului în fiecare etapă la un număr gestionabil de opțiuni posibile. Teoria accesibilității (Gibson, 1977) propune că numai anumite acțiuni specifice să fie viabile în anumite stări. În acest capitol, introducem un algoritm online, fără model, pentru învățare accesibilității, care poate fi apoi folosit pentru a învăța opțiunile sub-obiectivelor. Simulăm accesibilitatea printr-un mecanism de atenție care restrânge gama de opțiuni extinse temporal disponibile. Analizăm rolul atenției stricte versus relaxată în culegerea datelor de învățare. Analizăm învățarea valorilor abstracte în sarcini cu orizonturi lungi și gestionarea unei game extinse de opțiuni. Discernem și demonstrăm empiric situații în care paradoxul alegerii iese la suprafață - adică atunci când un număr mai mic de opțiuni mai semnificative sporește rată de învățare și eficacitatea unui agent de învățare prin recompensă.

Luarea deciziilor în medii complexe poate fi o provocare, deoarece avem multe opțiuni de luat în considerare la fiecare pas de timp. Învățarea unui mecanism de atenție care limitează aceste alegeri ar putea duce la o performanță mult mai bună. Oamenii au o capacitate remarcabilă de a acorda o atenție selectivă anumitor părți ale input-ului vizual [20, 5], adunând informații relevante și combinând secvențial observațiile lor pentru a construi reprezentări pe diferite scale de timp [14**?**]201, care joacă un rol important în ghidarea percepției și acțiunii ulterioare [32, 2]. În această lucrare, explorăm idei pentru dotarea agenților de învățare prin recompensa (RL) cu acest tip de capabilități.

Un agent RL interacționează cu un mediu printr-o secvență de acțiuni, învățând să-și maximizeze randamentul așteptat pe termen lung [47]. Abstracția temporală, de ex. opțiunile [48] îi permit să ia în considerare deciziile la scări de timp variabil. Opțiunile permit agentului să reducă profunzimea anticipării sale atunci când efectuează planificarea și să propage creditul recompensei pe o perioadă mai lungă de timp. Cu toate acestea, în spațiile mari de acțiune și stări, agentul se confruntă în continuare cu multe opțiuni la fiecare pas de decizie, iar opțiunile pot de fapt agrava această problemă, deoarece alegerile se înmulțesc atunci când se iau în considerare diferite intervale de timp. Prin urmare, abstracția temporală poate duce la un factor de ramificare mai mare.

O soluție binecunoscută este să *selectăm* un număr mic de opțiuni pe care să se concentreze, de exemplu, aplicarea unei acțiuni numai atunci când anumite condiții prealabile sunt îndeplinite în planificarea clasică [11] sau utilizarea seturilor de inițiere pentru opțiuni [48]. Lucrări recente [22] au propus o generalizare a seturilor de inițiere la *funcții de interes* [49, 57], care oferă o modalitate de a învăța opțiunile care se specializează în anumite regiuni ale spațiului de stări. Toate aceste abordări pot fi privite că oferind un mecanism de atenție asupra opțiunilor de acțiune.

În timp ce atenția a fost studiată pe scară largă în domeniul viziunii computerizate [16, 5], rolul diferitelor tipuri de atenție asupra opțiunilor de acțiune nu a fost studiat prea mult, în special pentru acțiunile abstracte temporal. În această lucrare, explorăm atenția asupra opțiunilor de acțiune în agenții RL cu un interes special pentru accesibilitatea asociată cu opțiuni, care pot fi privite că o formă de atenție *strictă*. Pe de altă parte, a avea preferințe relaxate peste toate acțiunile fără a elimina niciuna poate fi privit că o atenție *relaxată*.

În acest capitol, studiem rolul atenției *stricte* versus *relaxată* în luarea deciziilor cu abstractizare temporală. Presupunem că, în anumite setări, restricționarea atenției unui agent la ceea ce este accesibil duce la mai puține, dar utile, alegeri în comparație cu utilizarea atenției relaxate sub formă de funcții de interes. Demonstrăm empiric acest *paradox al alegerilor* mai puține care pot contribui la o învățare mai rapidă, rezultând mai multe recompense. După cum era de așteptat, acest efect este mai pronunțat atunci când agentul se ocupă de alegeri care duc la o utilitate mai bună pe termen lung. Participarea la alegeri utile într-o anumită stare are adesea un efect compus, ceea ce duce la alte alegeri bune în viitorul apropiat.

Contribuții cheie: Măsurăm impactul atenției *stricte* versus *relaxată* asupra opțiunilor opțiunilor temporale: 1) atunci când generăm date de învățare, 2) asupra învățării valorilor abstracte în sarcinile cu orizont lung și 3) la creşterea numărului de opţiuni. Prezentăm un algoritm online, fără model, pentru a învăța opțiunile sub-obiectivelor care țin seama de accesibilitatea opțiunilor, cu sub-obiective date și o sarcină de nivel înalt. Politicile de accesibilitate și opțiunile sunt învățate simultan, informându-se reciproc. Folosirea posibilităților învățate pentru a genera date de antrenament facilitează învățarea

eficientă a eșantionului de experiențe a opțiunilor. Demonstrăm empiric că mai puține opțiuni pot fi mai bune atât pentru viteză de învățare, cât și pentru performanța finală pentru un agent de învățare cu recompensă și structura ierarhică.

Accentul cercetare are obiectivul de a studia accesibilitatea că atenție strică și de a o contrasta cu atenția relaxată pentru acțiunile extinse în timp. Demonstrăm acest lucru prin medii non-tabulare, constând atât din acțiune/spațiu de stare discret cât și continuu în Minigrid [7] și, respectiv, Open-AI Robotics [42]. Alegerea acestor configurații experimentale a fost motivată de o proiectare atentă cu scopul de a investiga și izola rolul accesibilității că atenție strictă, putând varia în același timp complexitatea în ceea ce privește numărul de opțiuni și sarcini. De asemenea, subliniem că abordarea noastră se extinde atât la spații de acțiune discrete, cât și continue și poate fi utilizată în orice domeniu în care sub-obiectivele sunt cunoscute a priori.

5.5 Explorarea mediilor cu mai mulți agenți

În acest capitol examinăm modul în care agenții de învățare prin recompensă pot învăța să rezolve sarcini împreuna cu mai mulți agenți. În primul rând, prezentăm o abordare de învățare prin întărire profundă pentru a învăța un joc de dilemă socială extinsă în timp într-un mediu simulat. Agenții se confruntă cu diferite tipuri de adversari cu diferite niveluri de angajament față de o strategie de colaborare. Metodă noastră se bazează pe progresele recente în formarea cu gradient a politicilor folosind rețele neuronale profunde. Investigăm mai mulți algoritmi de gradient stocastic pentru învățarea cu recompensă, cum ar fi Reinforce sau Actor Critic, cu sarcini auxiliare pentru o convergență mai rapidă. În al doilea rând, propunem o modalitate nouă de a prezența probleme cu un singur agent că configurații multi-agent și prezentăm metode de antrenare a acestora într-o astfel de paradigmă. Motivăm această configurație prin faptul că putem exploata avantajele mai multor agenți, care pot colabora și comunica și pot beneficia de învățare prin recompensă când întâmpină provocări de explorare. Oferim dovezi empirice preliminare pentru a susține această ipoteză.

Cercetarea învățării automate în anii 2016 și 2017 a produs un progres continuu în ceea ce privește algoritmii de învățare profundă prin recompensă pentru agenții plasați în scenarii simulate. O mulțime de algoritmi noi, fără model, au explorat avantajele utilizării rețelelor neuronale profunde pentru prezicerea stării sau a valorilor de acțiune și pentru aproximarea politicilor de control continuu în diferite medii vizuale (de exemplu, Atari [30], Vizdoom [24] , sau Minecraft [50] sau jocuri de societate (Go). Puține studii recente s-au concentrat pe învățarea prin consolidare în configurații cu mai mulți agenți în care mai multe entități de învățare trebuie să coopereze în jocuri competitive sau colaborative. Problema non-staționarității, una dintre provocările de bază în învățarea prin întărire, este agravată de comportamentele în continuă schimbare ale celorlalți actori. Cu toate acestea, rezolvarea problemelor într-o paradigmă de învățare prin recompensă, multiplii agenții ar putea aduce performanțe îmbunătățite, datorită schimbului de informații și coordonării sporite a mai multor actori.

Schemă de returnare a unui scenariu bazat pe recompense mulți-agenți este uneori cel mai bine descrisă din perspectiva teoriei jocurilor. O matrice a profiturilor rezumă câștigurile pentru toți jucătorii în funcție de strategiile alese de ei. Există un corpus mare de cercetări în inteligența artificială cu privire la modul în care agenții își pot maximiza profitul așteptat prin învățarea din interacțiuni repetate. Toate aceste rezultate s-au concentrat pe învățarea în configurații față stare urmând diverse scheme de recompensă (de exemplu, dilema prizonierului). De asemenea, proprietățile teoretice precum echilibrul Nash sau optimitatea Pareto sunt evaluate pentru strategiile învățate. O problemă mai dificilă este identificarea unei astfel de situații într-un scenariu mai complex (de exemplu, în timpul unui joc de șah sau a unei vânători de pradă în colaborare) în care recompensă este o consecință a unei secvențe (posibil mare) de decizii bazate pe observații neprocesate parțiale ale mediului.

În prima parte a acestui capitol prezentăm o abordare de învățare prin consolidare profundă a unui astfel de scenariu, în care agenții sunt situați într-un joc episodic cooperativ bazat pe vânătoarea de cerb. Jucătorii nu au memorie persistentă de la un episod la altul. Unul dintre obiectivele cercetării noastre a fost să vedem dacă agenții înțeleg macroschema subiacentă (care este o instanță a jocului de vânătoare de cerb) prin tehnici de învățare profundă cu recompensă și nu prin învățarea explicită să aleagă între două strategii. Abstracția unei descrieri înalte a interacțiunilor cu mediul ar fi benefică în multe privințe. De exemplu, o performanță bună în acest joc ar putea oferi o experiență anterioară valoroasă înainte de a învăța o a două sarcină similară (învățare prin transfer). A ști cum să faci față unei dileme a încrederii, în general, necesită că un agent să se confrunte cu o situație nouă doar să învețe cum să interpreteze

percepția și cum să afecteze acel mediu specific, reducând povara procesului de învățare. Abordarea luată în muncă noastră se bazează pe metode de gradient de politici on-line, mai precis variații ale algoritmilor Reinforce și Actor-Critic cu sarcini auxiliare. Toți predictorii sunt rețele convoluționale profunde urmate de straturi recurente antrenate folosind actualizări bazate pe gradient per episod.

Scenariul pe care am încercat să-l rezolvăm, numit Malmo Platform PigChase[1] este unul dintre scenariile din Malmo Platform, un mediu de învățare cu recompense construit pornind de la Minecraft. Complexitatea mare a acestui mediu duce la un mediu lent cu resurse mari, care la prima vedere este incompatibil cu antrenarea predictorilor neuronali profundi ce necesită milioane de mostre înainte de a obține rezultate. Din acest motiv am implementat o replică simplificată a configurației originale aproximând dinamică setării originale pe care o vom numi de acum încolo mediul PigChase Replică. Am avut în vedere două obiective pentru acest mediu simplificat: să fim de multe ori mai rapid decât originalul și să lucrăm în modul batch. Apoi am instruit agenți bazați pe gradient de politici în această configurare rapidă doar pentru a-i ajusta la sfârșit pe cea originală. Intuiția noastră a fost că, odată ce agentul înțelege dinamică generală și dilema de încredere subiacentă, mutarea lui într-un mediu similar necesită o mică adaptare.

În RL cu un singur agent, explorarea poate fi o provocare, deoarece agentul trebuie să echilibreze între explorarea acțiunilor noi și exploatarea acțiunilor care au fost deja învățate că a fi satisfăcătoare. Agenții multipli, în învățarea prin întărire mulți-agenți (MARL), pot ajută totuși la îmbunătățirea explorării prin valorificarea diversității de comportament, coordonare, competiție și alinierea intentilor agenților. Având mai mulți agenți să exploreze mediul, există o probabilitate mai mare că unii agenți să exploreze noi stări și acțiuni, crescând explorarea globală a mediului. Agenții își pot coordona eforturile de explorare, cum ar fi împărțirea spațiului de explorare între ei sau împărtășirea informațiilor despre stările pe care le-au vizitat. Prezența mai multor agenți într-un cadru competitiv poate conduce explorarea, deoarece agenții se străduiesc să găsească un avantaj unul față de celălalt. MARL poate fi folosit pentru a alinia recompensele între agenți, încurajându-i să colaboreze și să exploreze împreună, ceea ce duce la o explorare îmbunătățită. Aceste idei motivează muncă de cercetare efectuată în Secțiunea **5.4** unde investigăm modul de ajustare a algoritmilor populari de învățare cu întărire profundă, pentru a pune probleme cu un singur agent că setări cu mai mulți agent. În această secțiune prezentăm rezultate preliminare care arată că metodă noastră este capabilă să exploateze avantajele mai multor agenți, colaborare și comunicare, pentru a reduce provocările de explorare RL.

5.6 Explorarea și navigarea la scară în realitate

În acest capitol, considerăm problema navigației în realitate atât pentru roboți, cât și pentru vehiculele autonome. Începem prin a analiză provocările politicilor de învățare cap-coadă (end-to-end) pentru navigarea roboților în spații interioare. În continuare, investigăm modul în care putem folosi simulatoare și să transferăm politicile în realitate și arătăm cum memoria și informațiile semantice pot ajută la localizarea în interior. Continuăm abordând procesul necesar pentru construirea și testarea unui vehicul autonom în lumea reală care utilizează modele bazate pe învățarea automată pentru direcție. Descriem mai întâi procesul de colectare și procesare a unui set de date local de conducere autonomă. Apoi, prezentăm rezultatele acestui set de date pentru învățarea modelelor cap-coadă pentru direcție și propunem un proces auto-supravegheat pentru învățarea politicii de conducere folosind doar înregistrările camerelor monoculare.

Evoluția rapidă a roboților mobili a propulsat explorarea unor medii vaste și diverse, cu un accent principal pe navigarea autonomă. În aceast capitol, aprofundăm cercentarea în două domenii, cel al navigației roboților de interior și conducerii autonome, explorând complexitățile și provocările pe care le prezintă fiecare domeniu.

Domeniul roboților mobili este în curs de dezvoltare rapidă, cu mașini autonome capabile să îndeplinească o gamă variată de sarcini. Cu toate acestea, navigarea prin aceste medii - în special spațiile interioare și zonele exterioare dens populate, cum ar fi străzi și autostrăzi. Datorită progreselor în algoritmii de învățare automată, utilizarea datelor senzoriale neprocesate, cum ar fi camerele, a devenit fezabilă în roboții mobili, permițând mașinilor să aibă acces la cât mai multe informații pentru a depăși dificultatea scenariilor din lumea reală. Deși sursele de date pot fi utilizate împreună cu alți senzori precum odometria sau măsurătorilor de proximate cu laser, bazându-se pe date senzoriale brute, roboții pot învăța și

[1] https://www.microsoft.com/en-us/research/project/project-malmo/

generaliza dintr-o gamă variată de experiențe, făcându-i mai bine echipați pentru a gestiona situații noi și neprevăzute. Comunitatea științifică este în prezent concentrată pe abordarea diferitelor provocări critice, în special în domeniul navigării sarcinilor, care implică adesea dificultatea explorării. Pentru a depăși această provocare, este imperativ că roboții mobili să aibă capacitatea de a construi modele robuste ale mediului lor, de a rezolva problemele de localizare și de a naviga eficient. Pentru o autonomie îmbunătățită, roboții ar fi echipați cu politici de explorare adecvate, care pot naviga și învăța continuu în contextul lumii reale.

În epoca modernă, progresele tehnologice remodelează în mod constant lumea noastră, scutind oamenii de sarcini laborioase și sporind productivitatea prin automatizare. Această evoluție a înregistrat o creștere exponențială a aplicării roboților mobili și a vehiculelor autonome în diverse medii. Aceste entități autonome se găsesc în gospodării, spații industriale, instituții de învățământ și chiar și pe drumuri, îndeplinind o multitudine de sarcini. Pași semnificativi în ceea ce privește percepția și abilitățile de calcul au contribuit la extinderea domeniului de aplicare a mediilor în care aceste sisteme automate pot funcționa. Navigația, o proprietate crucială atât a roboților mobili, cât și a conducerii autonome, este vitală pentru funcționalitatea acestor sisteme. Inteligența acestor mașini se manifestă în abilitățile lor de navigare, deoarece sunt folosite în numeroase aplicații pentru transport, industrie și misiuni de salvare. Planificarea traseelor, o componentă fundamentală a navigației autonome, a făcut obiectul unor cercetări intense în ultimele două decenii. Numeroși algoritmi au fost dezvoltați și testați pe diverse sisteme robotizate, cum ar fi microvehicule aeriene [59, 39], roboți de mișcare, roboți care urcă pe pereți și roboți subacvatici [60]. În mod similar, vehiculele autonome de pe drumurile noastre se bazează, de asemenea, în mare măsură pe planificarea avansată a traseului pentru a asigura o călătorie sigură și eficientă. Înțelegerea tradițională a navigației în roboții mobili și vehiculele autonome constă în abordarea a trei întrebări de bază (Levitt et. al. [26]): Unde este robotul sau vehiculul, unde sunt alte locuri și lucruri în relație cu acesta și cum poate ajunge robotul în alte locuri pornind de la locația să actuală. Cu toate acestea, în organismele inteligente, aceste reguli nu trebuie să fie separate neapărat linear. În schimb, ele pot fi profund împletite, reflectând realități complexe și dinamice pe care trebuie să le parcurgă aceste sisteme autonome. Această explorare a navigației autonome atât în spațiile interioare, cât și în exterior ne propulsează înțelegerea către soluții mai cuprinzătoare și adaptabile.

Inteligența artificială își propune să îmbunătățească sistemele cu inteligență care pot facilita luarea deciziilor și execuția optimă a sarcinilor. Acest concept de bază se dovedește crucial în diverse scenarii, de la navigarea robotului în interior până la conducerea autonomă. Peisajul AI contemporan este dominat de algoritmi de învățare automată care permit mașinilor să învețe în mod independent. O ramură a ML-ului, Învățarea prin Recompensă, conturează un cadru de învățare în care un agent, fie că este un robot de navigație în interior sau un vehicul autonom, efectuează acțiuni într-un mediu și se străduiește să-și maximizeze recompensă pe baza observațiilor acumulate. Această paradigmă de învățare, investigată pe larg în cercetările noastre anterioare privind explorarea în spațiul Procesului de decizie Markov (MDP), poate fi adaptată cu constrângerile necesare pentru a aborda provocările de navigație în ambele domenii. Dintr-o perspectivă RL, explorarea poate fi văzută că o politică care ghidează agentul să interacționeze cu noile stări sau tranziții din mediu. Această explorare permite robotului de interior sau vehiculului autonom să înțeleagă suficient împrejurimile și să își îndeplinească eficient sarcinile fără a rămâne blocat. Astfel, metodele RL nu numai că permit roboților să navigheze în spații interioare complicate, ci și echipează vehiculele autonome cu capacitatea de a traversa rețele rutiere complexe în siguranță și eficient.

Obținerea corectă a software-ului este o provocare în special pentru sistemele autonome, cum ar fi vehiculele autonome care navighează în rețele complexe de trafic sau roboții de interior care se deplasează prin spații dinamice precum aeroporturi, mall-uri și depozite. Aceste zone conțîn adesea căi înguste, obstacole schimbătoare și modele în evoluție care necesită o planificare complicată a rutei. Cheia constă în conceperea unui software care poate gestiona aceste probleme, păstrând în același timp experiența utilizatorului în prim plan. Abordările bazate pe reguli pot deseori să nu aibă scalabilitate în astfel de scenarii imprevizibile. Este esențial că sistemul autonom, fie că este un vehicul sau un robot, să necesite o pregătire minimă pentru operatori, să evite configurarea excesivă a mediului, să învețe rapid din demonstrație și să ofere raportări de productivitate. Variabilele care influențează navigația autonomă nu sunt doar obstacole fizice care ocupă mediul de lucru al sistemului. Terenuri fără caracteristici, momente diferite ale zilei și nenumărați alți factori adaugă complexitate navigației autonome, fie pe șosea, fie în interior. Multe dintre aceste provocări sunt cazuri rare dar care împing limitele strategiilor convenționale, adesea ies la suprafață numai odată ce software-ul este complet dezvoltat și sistemele autonome sunt testate în situații reale. Aceste complexități subliniază necesitatea unei abordări de învățare și a utilizării simulatoarelor din lumea reală. Cu toate acestea, sistemele de navigație autonome cu adevărat funcționale

nu sunt cultivate în întregime într-un laborator; succesul lor depinde de capacitatea sistemului autonom și de software-ul pe care rulează, de a naviga cu abilități în mediile reale. Acest lucru este valabil atât pentru vehiculele autonome care înfruntă incertitudinile drumurilor, cât și pentru roboții de interior însărcinați să navigheze eficient în spațiile comerciale foarte dinamice.

În acest capitol, am investigat provocările și potențialele soluții pentru explorarea pe scară largă folosind învățarea prin recompensa, cu accent atât pe navigarea roboților în interior, cât și pe conducerea autonomă în aer liber. Am început prin a examina procesul de învățare end-to-end a politicilor de navigare a robotului pentru spațiile interioare folosind RL. Aceasta a implicat transferul de cunoștințe din simulator în medii din lumea reală, implementarea atenției semantice și abordarea provocărilor de localizare și planificare pe termen lung. În a doua parte a capitolului, ne-am îndreptat atenția către domeniul mașinilor cu conducere autonomă, aprofundând în complexitatea colectării și procesării unui set de date robust, precum și dezvoltării modelelor de politică de direcție end-to-end pentru vehiculele autonome. Pe măsură ce încheiem capitolul, reflectăm asupra perspectivelor obținute din aceste studii și discutăm căile potențiale pentru activități viitoare atât în navigarea roboților în interior, cât și în conducerea autonomă în aer liber.

În secțiunile **6.2, 6.3, 6.4,** abordăm problema navigării roboților în medii interioare și explorăm potențialul utilizării rețelelor neuronale artificiale ca o abordare complet, de la starea senzorilor la actiuni, bazată pe învățare pentru a depăși limitările metodelor clasice. Analizăm aceste abordări clasice și, pe baza literaturii de specialitate, identificăm domeniile în care învățarea unei politici de navigație ar putea atinge performanțe de ultimă generație. Prezentăm un set de sarcini de navigație care pot fi investigate individual, evidențiind modul în care încorporarea cunoștințelor semantice despre mediu ar putea îmbunătăți abordările geometrice care au neglijat în mod tradițional acest aspect. În timpul acestei analize, provocarea explorării spațiului apare ca o sarcină critică în sine. Restul capitolului analizează această problemă, concentrându-se pe pregătirea agenților RL capabili să navigheze în medii reale. Deoarece antrenamentul exclusiv pe un robot este în prezent imposibil de realizat și ar fi imprudent să nu ținem cont de beneficiile și cunoștințele oferite de simulatoare, ne concentrăm pe trei provocări principale în explorarea robotului de interior: navigarea folosind doar informații RGBD (camera color si adancime) fără localizare precisă; construirea unei reprezentări robuste cu cartografiere implicită și localizare pe distanță lungă prin investigarea mecanismelor de atenție; și explorarea metodelor de îmbunătățire a transferului în medii din lumea reală. În această cercetare am evaluat în mod independent fiecare metodă de augmentare, zgomotul de adâncime îmbunătățind în mod clar scorurile testelor. Cu toate acestea, alte metode demonstrează îmbunătățiri minore și ne propunem să explorăm beneficiile potențiale ale combinării lor. Deși este posibil ca unele metode să nu contribuie în mod semnificativ la generalizarea într-o configurație supravegheată, nu ar trebui să neglijăm impactul potențial al acestora asupra învățării politicilor, așa cum sugerează lucrările anterioare. În plus, intenționăm să investigăm o abordare mai randomizată a creșterii „Noop", ocolind condiționarea bazată pe distanța până la un obstacol. Testarea amănunțită și rafinarea mecanismului de memorie sunt esențiale, deoarece este o abilitate critică pentru un agent de explorare eficient. Metoda recurentă simplă pare să eșueze atunci când se calculează odometria pentru distanțe de cinci pași, indicând importanța acestei componente în arhitectura de învățare. În Secțiunea **6.2** vom introduce problema navigației robotului, ce sarcini pot fi definite în acest cadru și câteva dintre soluțiile actuale. În Secțiunea **6.3** prezentăm resursele necesare pentru reproducerea experimentelor explicate în acest raport. În ceea ce privește Secțiunea **6.4**, ne concentrăm mai mult pe problema explorării, cum să o scalam la orizonturi de timp mai lungi și cum să transferăm mai bine o politică de robot într-un scenariu real.

Până acum, au fost efectuate doar teste preliminare pentru formarea unui agent de explorare în mediu. Pașii viitori includ integrarea sarcinilor auxiliare și a metodelor de augmentare, evaluarea celei mai bune politici pentru un robot real și dezvoltarea unei metrici pentru evaluarea capacității de explorare a politicii în scenarii din lumea reală. Scorurile suprafeței explorate pot varia semnificativ în funcție de nivelul dezordinei din mediu, chiar și cu politici de calitate similară, necesitând dezvoltarea unei metrici de evaluare adecvate.

Analiza noastră continua apoi în domeniul conducerii autonome, unde prezentăm cercetările noastre privind crearea și procesarea seturilor de date pentru conducerea autonomă, explorarea modelelor end-to-end pentru vehicule autonome și investigarea tehnicilor de etichetare auto-supravegheate. Ne propunem să demonstrăm modul în care tehnicile de învățare automată și de învățare prin recompensa pot fi adaptate pentru a aborda aceste probleme complexe și cum aceste soluții pot fi scalate și transferate în scenarii din lumea reală. În Secțiunea **6.5**, vom detalia procesul de colectare și procesare a unui set de date autonom

în campusul UPB. Subliniem importanța calității setului de date, în special în domeniile critice pentru siguranță, cum ar fi conducerea autonomă, și oferim un ghid cuprinzător care acoperă fiecare pas, de la configurarea hardware-ului până la validarea datelor. Considerăm că un proces de colectare a datelor curat și eficient poate îmbunătăți evaluarea comparativă a soluțiilor de conducere autonomă, ținând cont de nuanțele locale de mediu. În secțiunea **6.6**, explorăm cercetarea asupra modelelor end-to-end pentru mașini cu conducere autonomă pe drumurile campusului UPB. Transformând planificarea traiectoriei vehiculului într-un proces care poate fi învățat, investigăm posibilitatea de a integra predicțiile de direcție în cadrele existente de conducere autonomă. Abordarea noastră ia în considerare trăsăturile geografice specifice, folosește tehnici de augumentare a datelor și evaluează calitatea modelului pe baza unui set de date colectat din campusul UPB. Descoperirile noastre sugerează că modelele propuse funcționează bine pentru setul de date adunat. În Secțiunea **6.7**, propunem un proces de instruire pentru etichetarea autosupervizată pentru conducerea autonomă. Această abordare extrage informații valoroase despre direcție, accelerație și frânare numai din datele de provenite din camerele monoculare de pe masina. Conectăm aceste date la un algoritm de învățare prin recompensa capabil să prezică comenzile necesare pentru conducerea autonomă pe baza unei reprezentări abstracte a drumului. Contribuțiile cheie ale acestei secțiuni includ o metodă pentru generarea de adnotări de direcție din date neetichetate și un proces inovator care produce etichete de traseu într-un mod complet nesupravegheat.

Până acum, am prezentat și prototipat un proces pentru învățarea și producerea unghiurilor de virare folosind înregistrări auto, fără etichete ale scenei de conducere. Aceast proces poate distila un comportament valoros de conducere din cantități mari de date ușor de colectat. Propunem posibile idei de îmbunătățire pentru aceast proces, care au fost testate doar într-o măsură limitată. În primul rând, planul de sol al adevărului la sol din lucrarea noastră a fost estimat aproximativ folosind pozițiile camerei, dar ar putea fi generat folosind estimarea adâncimii și estimarea mai avansată a planului de sol pentru a reduce erorile produse de pozițiile înclinate și diferitele camere de înregistrare. Etichetele traseului pot fi, de asemenea, utilizate pentru a determina viteză mașinii, luând în considerare lungimea traseului etichetat în termeni de distanță de la mașînă până la cea mai îndepărtată etichetă în proiecția planului de sol. Pentru a face acest lucru, ar trebui să determinăm cea mai îndepărtată cale „disponibilă" pe care o poate identifica mașina pe baza imaginii curente. Considerăm o potecă disponibilă drept drumul deschis pe care mașina poate circulă fără obstacole. Credem că acest lucru poate fi determinat auto-supravegheat din videoclipuri folosind erori de predicție atât înainte, cât și înapoi în timp. Utilizarea a două rețele neuronale pentru a prezice calea viitoare atât înainte, cât și înapoi prin videoclip ar putea fi folosită pentru a determina obstacolele pentru care mașina le-a evitat sau pentru care a încetinit. Acest lucru poate fi realizat prin determinarea părților din imagine care nu pot fi generate din ambele direcții, ajutându-ne să generăm trasee etichetate doar pentru spațiul carosabil. Predicțiile de traseu pot fi îmbunătățite și evaluate utilizând rețele de segmentare a zonei carosabile. Deși modelul de direcție Ackermann și parametrii camerei ar trebui cunoscuți pentru mașina de testare, calibrarea corectă a acestora rămâne o provocare semnificativă care necesită întreținere constantă. Am propus o metodă prin învățare prin recomepnsa care poate învăța să conducă folosind hărți de segmentare și să se auto-adapteze la diferite modele de direcție ale mașinii sau la diferite poziții sau parametri ai camerei. Acest model poate fi antrenat complet utilizând căi prezise pe datele colectate și spațiul acesta de observație face mult mai simplu augmentare și simulare să acopere distribuția reală a datelor pentru testare.

5.7 Explorarea spațiului chimic pentru generarea de noi molecule

În capitolul următor vom studia problema explorării în contextul descoperirii medicamentelor, în special modul în care algoritmii de învățare automată și de învățare prin recompensă pot identifica potențiali compuși noi de medicamente prin căutarea mai eficientă a spațiului moleculelor dorite. Cercetăm și propunem îmbunătățiri atât pentru algoritmii RL, cât și pentru GFlowNet pentru sarcina de a genera molecule mici și analizăm importanța generalizării pentru o mai bună proiectare moleculară. În cele din urmă, cercetăm importanța evaluării adecvate a generalizării designului moleculelor bazate pe învățarea automată folosind metrici de simulare pentru a o transferă eficient la afinitatea andocare în lumea reală.

Aplicarea tehnicilor de învățare automată la descoperirea medicamentelor și proiectarea moleculară deține un potențial semnificativ pentru a revoluționa modul în care identificăm și dezvoltăm noi agenți terapeutici [27]. Metodele tradiționale de descoperire a medicamentelor sunt adesea costisitoare din punct de vedere computațional și forță de muncă. Mulți potențiali compuși nu reușesc să ajungă pe piață din cauza problemelor de eficacitate, siguranță sau farmacocinetică. Timpul mediu de dezvoltare clinică a medicamentelor ajunge la mai mult de nouă ani, iar costul mediu de dezvoltare se apropie

de 1 miliard USD [58]. Descoperirea preclinice de medicamente în stadiu incipient este de obicei un proces iterativ de optimizare care constă în ciclul de generare, testare, învățare. Mai mult, vastitatea și complexitatea spațiului chimic îngreunează abordările convenționale să exploreze și să identifice în mod eficient molecule noi, active biologic. Învățarea automată oferă oportunitatea de a depăși aceste provocări prin valorificarea avantajelor algoritmilor bazați pe date pentru a accelera procesul de descoperire a medicamentelor, optimizând proprietățile moleculare și descoperind noi structuri chimice cu activități biologice dorite.

Procesul de descoperire a medicamentelor [37] este un efort complex și cu mai multe fațete, care constă de obicei din mai multe etape cheie, inclusiv identificarea țintei, validarea țintei, identificarea hit-urilor, optimizarea lead-ului și vaidarea preclinică. În identificarea țintei, cercetătorii se concentrează pe găsirea unei ținte biologice, cum ar fi o proteină sau o genă, care este implicată într-o anumită boală. Validarea țintei implică confirmarea rolului țintei în boală și adecvarea acesteia pentru intervenția terapeutică. Identificarea hit-urilor implică găsirea unui compus cu activitate confirmată împotriva unei ținte biologice (cei care interacționează cu ținta). În această etapă, noi candidați sau molecule de medicamente sunt proiectate sau identificate folosind diferite metode, cum ar fi proiectarea de medicamente asistată de computer, chimia combinatorie sau screening-ul cu randament ridicat al bibliotecilor de compuși existenți. Candidații nou generați sunt apoi testați în vitro (în eprubete sau culturi celulare) sau în vivo (în modele animale) pentru a evalua activitatea lor biologică, potența, selectivitatea și alte proprietăți relevante pentru utilizarea lor terapeutică potențială. Acest pas îi ajută pe cercetători să înțeleagă cât de bine interacționează candidații cu proteinele sau căile lor țintă și efectele lor secundare potențiale. Optimizarea lead-urilor rafinează aceste „loturi" inițiale în compuși „lead" (potențiali) mai puternici și selectivi, cu proprietăți îmbunătățite asemănătoare medicamentelor. În cele din urmă, dezvoltarea preclinică implică testarea riguroasă a compușilor lead în modele celulare și animale pentru a evalua siguranța, eficacitatea și proprietățile farmacocinetice ale acestora înainte de a trece la studiile clinice. Învățarea automată și învățarea prin recompensă pot fi integrate în acești pași pentru a le îmbunătăți eficiența și eficacitatea [8, 53].

Prin automatizarea și îmbunătățirea diferitelor aspecte ale procesului de descoperire a medicamentelor, aceste tehnici de calcul avansate pot ajuta la reducerea timpului și costurilor asociate cu aducerea de noi medicamente pe piață, îmbunătățind, de asemenea, probabilitatea de a descoperi agenți terapeutici eficienți și siguri. În contextul designului molecular de novo, atunci când se proiectează noi molecule cu proprietățile dorite de la zero, învățarea prin recompensa (RL) și alte tehnici bazate pe învățarea automată (ML) au apărut ca abordări promițătoare pentru explorarea și optimizarea eficientă a spațiului chimic. Aceste metode urmăresc să abordeze limitele tehnicilor tradiționale prin automatizarea și accelerarea descoperirii de noi structuri moleculare cu proprietăți dorite. Modelele generative au adus o revoluție în domeniul descoperirii de novo a medicamentelor, introducând abordări inovatoare ale procesului. Tehnici cum ar fi rețelele neuronale recurente cu memorie pe termen scurt (LSTM-RNN), autoencodere variaționale (VAE), rețele generative adversare (GANs), adversarial autoencoders (AAE), algoritmii evoluționari, unitățile recurente cu porti (GRU-RNN) și modelele de difuzie au adus contribuții semnificative la acest progres [28, 31, 15]. De mulți ani, algoritmii de învățare de întărire, cum ar fi metodele Q-learning și gradient de politici, au fost propuși ca o abordare complementară a designului molecular [36]. Tehnicile emergente în domeniu, cum ar fi rețelele neuronale grafice și mecanismele de atenție conduc în mod continuu progresului. Aceste tehnici permit optimizarea iterativă a obiectivelor specifice prin interacțiuni de încercare și eroare în spațiul chimic, oferind o perspectivă valoroasă pentru abordarea provocărilor complexe de proiectare moleculară. Integrarea RL cu modele generative profunde poate produce o explorare mai țintită și mai eficientă, deoarece agentul RL ghidează modelul generativ pentru a produce molecule cu proprietăți optime. În plus, învățarea prin întărire multi-obiectivă poate aborda provocarea de a optimiza mai multe proprietăți fizico-chimice simultan, ceea ce duce la un proces de proiectare mai cuprinzător.

În secțiunea **7.3** am efectuat experimente extinse privind algoritmii de învățare prin recompensa pentru a îmbunătăți generarea de molecule mici, concentrându-ne în mod special pe îmbunătățirea scorurilor de andocare și a diversității. O secțiune esențială a activității noastre a implicat, de asemenea, analiza unui model de rețea neuronală proxy care servește drept mecanism de recompensă, o abordare comună în descoperirea medicamentelor cu ML, datorită proceselor costisitoare de colectare a datelor de antrenament prin experimente de laborator sau simulări costisitoare **7.2**. Ca urmare a cercetării îmbunătățirilor generației de molecule cu GflowNet [**?**] în secțiunea **7.4**, am introdus, de exemplu, o nouă măsurătoare, TopKDiverse, pentru a ne alinia mai bine cu obiectivele noastre principale de a descoperi molecule diverse și cu scor înalt. Luând în considerare provocările unice prezentate de acest domeniu, am propus o

nouă suită de evaluare adaptată descoperirii moleculelor mici, oferind astfel un standard robust pentru cercetările și dezvoltarea viitoare în acest domeniu. În cele din urmă, ne-am angajat într-o analiză profundă a generalizării în cadrul algoritmilor GFlowNet, îmbunătățind și mai mult aplicarea acestora în generarea de molecule mici **7.5**.

Investigația sistematică a designului molecular de novo utilizând metodologii de învățare prin recompensa și algoritmi generativi GFlowNet a dat rezultate substanțiale. După cum sugerează principalele descoperiri, atunci când se confruntă cu o funcție de recompensă nedefinită, algoritmii RL ar putea încă sa maximiza anumite regiuni ale spațiului nostru de stari. Cu toate acestea, calitatea funcției de recompensă proxy a fost direct proporțională cu cantitatea de date de antrenament utilizate.

În contextul antrenamentului RL pentru diversitate, a devenit evident că mai mulți factori erau de importanță critică, inclusiv reducerea recompensei cu explorarea bazată pe numărul de vizitari, determinarea eficientă a orizontului și calibrarea atentă a parametrului limitare și a coeficientului de entropie. De asemenea, am observat că dimensiunea rețelei neuronale ar putea avea un impact semnificativ asupra performanței, având în vedere suficiente date de antrenament.

GFlowNet, pe de altă parte, a prezentat un setul unic de provocări și concluzii. Evaluarea performanței modelului a fost un aspect critic, iar alinierea scopului cu obiectivul adevărat a fost fundamentală. Confruntarea cu recompense multi-obiective a reprezentat o provocare semnificativă, dar folosirea celei mai bune traiectorii descoperite ca reluare a experienței în timpul antrenamentului s-a dovedit a fi benefică.

În ceea ce privește generalizarea, cercetarea noastră a dat rezultate solide pentru evaluarea performanței GFlowNet pe un set de testare. Am găsit o corelație puternică între o măsură specifică și capacitatea de a genera loturi diverse, cu scoruri ridicate, dovedindu-se a fi o zonă promițătoare pentru studii ulterioare. În această lucrare am prezentat o evaluare amănunțită a performanței de generalizare a GFlowNet-urilor pentru proiectarea moleculelor, folosind metricile noastre propuse *GFNEval* și *pHighestKbins*. O limitare cheie a metricilor prezentate în această lucrare este costul de calcul al probabilității exacte de eșantionare a unei molecule sub GFlowNet. Acest lucru poate face dificilă evaluarea cu seturi mari de testare. O direcție de explorat ar fi modalitățile de a aproxima această cantitate mai eficient, păstrând în același timp proprietățile metricilor propuse. De asemenea, demonstrăm puterea discriminatorie a valorilor pe parcursul învățării. Această proprietate poate fi utilă în setările de învățare activă cu un generator GFlowNet [3] pentru oprirea timpurie a antrenării generatorului. Lucrările viitoare ar trebui să se concentreze, de asemenea, pe utilizarea acestor valori pentru analiză ulterioară a dinamicii de învățare GFlowNet, pentru a face recomandări practice pentru formare.

Cu toate acestea, ca în orice efort științific, cercetarea noastră nu a fost lipsită de provocări. Tranziția la scoruri reale și optimizarea rețelelor neuronale pe grafuri au reprezentat obstacole semnificative, iar maximizarea multi-obiectivă s-a dovedit dificil de rezolvat. Pentru viitor, am descoperit căi potențiale pentru cercetări ulterioare. Generarea de proteine, folosind mediul Rosetta, oferă o extindere potențială interesantă a activității noastre. Suita sofisticată Rosetta de algoritmi pentru modelarea computațională și analiza structurilor proteinelor ar putea oferi un teren bogat pentru aplicarea lecțiilor noastre învățate din designul molecular de novo. În plus, transformatoarele de decizie au prezentat o zonă promițătoare pentru construirea de politici pentru explorarea spațiului proteic. Valorificarea capacității lor de a modela traiectorii întregi ar putea permite o explorare mai eficientă și o mai bună aliniere a politicilor RL cu obiectivele noastre.

5.8 Generalizarea și contestarea status quo-ului

În acest capitol, întreprindem o explorare a provocărilor inerente învățării prin recompensă (RL), un domeniu de studiu care a cunoscut progrese semnificative, dar care se confruntă încă cu obstacole considerabile. Această explorare este efectuată dintr-o varietate de perspective și în cadrul diferitelor configurații ale problemelor, oferind o înțelegere largă și nuanțată a problemelor în cauză. Capitolul prezintă o serie de opinii argumentate cu atenție, care considerăm să fie esențiale pentru scalarea învățării prin recompensă pentru a aborda probleme complexe din lumea reală.

Una dintre discuțiile principale din acest capitol se concentrează în jurul limitărilor abordărilor tradiționale ale învățării prin recompensă. Aceste metode, deși sunt fundamentale pentru domeniu, s-au dovedit a fi potențial nepotrivite pentru rezolvarea problemelor care implică spații de stare infinite. Acestea sunt probleme complexe care necesită un nivel de adaptabilitate și scalabilitate pe care abordările tradiționale

RL nu le pot oferi. Mai mult, susțin că abordarea actuală și configurația de evaluare în cercetarea RL ar putea împiedica progresul dezvoltării și urmăririi algoritmilor pentru rezolvarea problemelor complexe din lumea reală. Fără o schimbare în modul în care abordăm și evaluăm RL, riscăm să deviem progresul și să nu reușim să realizăm întregul potențial al RL în abordarea problemelor din lumea reală.

Pentru a aborda acest lucru, propun un cadru mai general pentru RL. Acest cadru este conceput pentru a fi mai flexibil și mai adaptabil decât metodele tradiționale, capabil să învețe din complexitățile și nuanțele problemelor de spațiu infinit de stare și din progresul învățării în sine în timp ce agenții interacționează în acele medii. Dezvoltarea și argumentarea acestui cadru formează o parte semnificativă a capitolului, oferind atât o critică a metodelor existente, cât și o potențială cale de urmat în cercetare pentru viitor.

Un alt punct cheie de discuție în acest capitol este legătura puternică dintre generalizare și explorare în RL. Această este o relație critică căreia, susțin, nu i s-a acordat atenția pe care o merită în domeniu. Generalizarea, capacitatea unui agent RL de a aplică cunoștințele învățate în situații noi și explorarea, procesul de căutare a informațiilor noi, sunt profund relaționate. Înțelegerea și valorificarea acestei conexiuni este, cred, esențială pentru progresul viitor în RL. În acest capitol, ofer, de asemenea, dovezi experimentale care susțin ipoteza că configurarea evaluării poate determina ce fel de avantaje algoritmice maximizăm, care nu sunt aliniate cu rezolvarea problemelor cu spații complexe.

Subliniez acest punct pe parcursul capitolului, argumentând că o abordare mai concentrată pe înțelegerea relației dintre generalizare și explorare ar putea debloca progrese semnificative în RL. Acest lucru ar putea duce la agenți RL mai eficienți, capabili să învețe și să se adapteze în moduri în care am putea să ne apropie de rezolvarea problemelor complexe din lumea reală.

Pe scurt, al optulea capitol al tezei mele oferă o explorare detaliată și critică a provocărilor și potențialelor soluții în RL. Oferă o serie de opinii și propuneri bine argumentate care cred că vor contribui semnificativ la dezvoltarea continuă a domeniului. Printr-o critică a metodelor tradiționale și propunerea unui cadru nou, mai general, precum și un accent pe legătura dintre generalizare și explorare, acest capitol reprezintă o contribuție semnificativă la discursul despre RL.

6 Concluzii

În această teză, am explorat rolul explorării în învățarea prin recompensă dintr-o varietate de perspective și în diferite probleme din lumea reală. Procesul de explorare este integral și extrem de implicat în întregul proces de învățare prin recompensă al agenților care ating politici optime pentru atingerea obiectivelor. Prin această cercetarea, am dobândit o înțelegere mai profundă a provocărilor și oportunităților asociate cu explorarea în RL și am identificat și dezvoltat o serie de abordări noi pentru reducerea acestor provocări.

În general, muncă noastră a contribuit la domeniul RL prin oferirea de noi perspective asupra mecanismelor de explorare și prin demonstrarea importanței acestui mecanism într-o serie de aplicații din lumea reală. În special, constatările noastre au evidențiat importanța încorporării strategiilor de explorare care pot fi adaptate la caracteristicile specifice ale problemei în cauză, pentru a obține performanțe optime.

În timp ce contribuțiile prezentate în această teză ne avansează înțelegerea învățării prin recompensă, în special în domeniul explorării, este important să recunoaștem că numeroase întrebări și provocări persistă în domeniu. Sunt necesare cercetări suplimentare pentru a elucida cuprinzător rolul explorării în învățarea prin recompensă. În paragrafele următoare, vom rezumă principalele noastre constatări și contribuții (Secțiunea **9.1**) și vom discuta limitele și direcțiile viitoare ale cercetării noastre (Secțiunea **9.2**).

6.1 Contribuții

Contribuțiile cheie ale cercetării prezentate în această teză, inclusiv orice noi perspective sau înțelegeri care au fost obținute că urmare a studiului, sunt înșiruite în continuare:

- O analiză și o evaluare aprofundată a mai multor mecanisme de recompensă intrinsecă pentru îmbunătățirea RL într-o suită de medii cu recompensă rară, care a condus la definirea unor metrici pentru evaluarea progresului explorării, precum și la propunerea unei metode de recompensă intrinsecă nouă.
- Un mecanism original de recompensă intrinsecă care valorifică, auto-supravegheat, înțelegerea ordonarii temporale a stărilor pentru a reduce problemele de explorare în RL. Arătăm cum învățarea

ordinii timpului din stări și utilizarea scorului de „ordonare” al observațiilor permutate aleator că recompensă intrinsecă determină agenții să conveargă către politici mai puțîn haotice.

- Un algoritm online, fără model, pentru a învăța opțiunile sub-obiective care țîn seama de accesibilitatea opțiunilor. Investigăm rolul atenției stricte versus cea relaxată în colectarea datelor de învățare, învățarea valorilor abstracte în sarcinile dificile de explorare și gestionarea unui număr tot mai mare de opțiuni. Identificăm și ilustrăm empiric cazurile în care apare paradoxul alegerii, adică atunci când mai puține opțiuni, dar mai semnificative, îmbunătățesc viteză de învățare și performanța unui agent de învățare prin recompensă.
- Demonstrăm avantajul utilizării Învățării profunde prin recompensă pentru a învăța tabula rasă, o politică robustă într-un mini-joc colaborativ construit pe baza unei extensii a jocului teoretic „vânătoarea de cerb”. Lucrarea noastră de cercetare, care identifică și integrează mai multe îmbunătățiri pentru un algoritm clasic de învățare profundă cu recompensă, a reușit să învețe într-o configurație cu mai mulți agenți o politică care a depășit într-o competiție publică online alte politici, chiar și cele care au folosit cunoștințele de specialitate.
- Propunem o modalitate nouă de a poziționa probleme cu un singur agent drept configurații multi-agent și prezentăm metode de antrenare a acestor cu RL într-o astfel de paradigmă. Motivăm această configurație prin faptul că putem exploata avantajele mai multor agenți, care pot colabora și comunica și pot beneficia de învățare prin recompensă pentru a reduce provocări de explorare. Oferim dovezi empirice preliminare pentru a susține această ipoteză.
- Studiem transfer de cunostinde din simulator în lumea reală pentru RL atât în navigarea interioară a roboților, cât și în controlul vehiculelor autonome. Am colectat mai întâi un set mare de date atât simulate, cât și din lumea reală, pe care le-am folosit pentru a antrena algoritmi RL pentru ambele scenarii. Apoi am îmbunătățit performanța acestor algoritmi printr-o combinație de optimizare a hiperparametrilor, strategii de explorare și i-am adaptat la lumea reală prin ajustarea politicilor pentru a ține cont de diferențele dintre mediile simulate și cele reale. Pentru navigarea robotului în spații interioare propunem o metodă de îmbunătățire a localizării bazată pe transferul de cunoștințe semantice. Pentru politică vehiculelor autonome propunem o metodă autosupravegheată de predicție a traseului de învățare din care extragem unghiurile de virare. În cele din urmă, am evaluat performanța algoritmilor RL în sisteme cu buclă închisă printr-o serie de experimente, inclusiv colectarea de date suplimentare din lumea reală și compararea performanței algoritmilor cu o metodă de bază și cu alte strategii de control. În general, rezultatele noastre au demonstrat eficiența utilizării simulării pentru a antrena algoritmi RL atât pentru navigarea roboților în spații interioare, cât și pentru controlul vehiculelor autonome și au arătat că este posibil să se transfere în mod eficient acești algoritmi în lumea reală printr-o reglare și adaptare atentă.
- Prin cercetarea designului molecular de novo folosind algoritmi de învățare automată, am putut să prezentăm noi perspective pentru generarea de molecule mici și să propunem îmbunătățiri pentru creșterea scorurilor de andocare și a diversității compusilor generați. Am dezvoltat tehnici pentru evaluarea performanței GFlowNet pe un set de testare și am identificat cea mai promițătoare metrică pentru prezicerea generalizării pentru generarea de molecule mici, care este un aspect crucial de luat în considerare pentru eficiența datelor și transferul aplicabilității în lumea reală. Folosind învățarea prin recompensă cu design specifice pentru aceste tipuri de probleme, suntem capabili să obținem atât diversitatea dorită de molecule mici generate, cât și energii de andocare extrem de mari ale compușilor descoperiți (folosind AutoDock Vină).
- Propunem un nou cadru de învățare prin recompensă care poate capta teoretic, cu ușurință, mai multe informații necesare pentru a învăța comportamente mai complexe ale agenților în diferite medii. Această configurație pentru RL ar împuternici fluxul de informații necesar pentru a meta-învăța comportamente de ordin superior printr-un meta-agent care ar învăța să antreneze alți agenți.
- Propunem regândirea generalizării în Învățarea prin recompensă prin evaluare și prejudecăți inductive. Subliniem importanța evaluării generalizării pentru scalarea învățării prin întărire la probleme complexe din lumea reală și prezentăm experimente care motivează acest lucru.

6.2 Explorări viitoare în RL

Această teză a explorat mai multe domenii cheie în învățarea prin recompensă, de la recompense intrinseci, învățare ierarhică, medii mulți-agenți, robotică, până la descoperirea medicamentelor. De-a lungul acestei experiențe, am identificat atât punctele forțe, cât și limitările în starea actuală a RL, care împreună pot pregăti scenă pentru lucrările viitoare în acest domeniu interesant.

Prima noastră zonă de atenție a fost explorarea cu recompense intrinseci. Deși am reușit să evaluăm diferite recompense intrinseci și să îmbunătățim explorarea cu cunoștințe din modele bazate pe informații temporale, rezultate prezentate nu au acoperit o gamă suficient de diversă a mediilor. De asemenea, recompensă intrinsecă temporală nu a ignorat aspectele aflate în afară controlului agenților, deși ar putea fi abordate prin luarea în considerare a unui model de dinamică inversă. Lucrările viitoare ar trebui să urmărească introducerea unor medii mai diverse și mai complexe pentru a evalua mai bine și a spori capacitatea agenților RL de a gestiona diverse situații. În plus, rafinarea mecanismului de recompensă intrinsecă temporală pentru a distinge mai bine și a răspunde la diferite modele complexe este un domeniu promițător pentru investigații ulterioare. Recompensele intrinseci joacă un rol esențial în învățarea prin recompensă, prezentând oportunități nu doar pentru îmbunătățirea RL, ci și pentru o mai bună înțelegere a motivației intrinseci în viitorii agenți din lumea reală sau a sistemelor biologice. Cercetările viitoare ar trebui să urmărească descifrarea tiparelor de motivație în entitățile biologice, ceea ce ar putea oferi perspective în proiectarea mecanismelor artificiale de IR. Cele mai multe metode de recompensă intrinsecă funcționează mai mult pentru dependențe pe termen scurt, dar nu pentru sarcini care necesită gândire strategică pe termen lung. Dezvoltarea modalităților de a gestiona dependențele pe termen lung și chiar combinarea diferitelor metode IR ar putea stimula comportamente de învățare îmbunătățite.

În studierea învățării ierarhice prin recompensă, scalabilitatea politicii de accesibilatate a opțiunilor nu a fost pe deplin studiată. Acest lucru indică o direcție crucială pentru lucrările viitoare: dezvoltarea mecanismelor care să permită scalarea eficientă a învățării acesibilitatii în medii din ce în ce mai complexe, îmbunătățind astfel înțelegerea ierarhică și capacitățile de luare a deciziilor ale agenților RL. Descoperirile noastre din această activitate de cercetare subliniază rolul substanțial al concentrării asupra experiențelor recompensatoare în îmbunătățirea RL-ului. Privind în perspectivă, credem că o direcție promițătoare de cercetare constă în investigarea mecanismelor care sporesc atenția agentului RL față de experiențele cu semnal ridicat, spre deosebire de zgomot. Intenția ar fi de a concepe strategii mai sofisticate pentru a face distincția între informațiile relevante și irelevante din experiențe, optimizând astfel procesul de învățare. Această linie de anchetă ar putea aduce progrese semnificative, permițând sistemelor să învețe mai eficient din mediul lor.

În explorarea mediilor mulți-agenți, am reușit să rezolvăm sarcini individuale distribuind învățarea către mai mulți agenți care comunică. Cu toate acestea, scalarea numărului de agenți și încorporarea unor mecanisme precum distilarea și reinitializarea ponderilor agenților ar putea îmbunătăți eficiența și adaptabilitatea sistemelor mulți-agenți. Domeniile cheie de interes ar putea include tehnici de explorare adaptativă și învățare din experiențe eterogene. Această implică elaborarea de strategii care să permită agenților să-și modifice în mod dinamic explorarea pe baza feedback-ului colegilor, potențial îmbunătățind procesul decizional și învățarea colectivă. În același timp, valorificarea experiențelor diverse ale agenților multipli oferă o oportunitate de a oferi sistemelor RL o înțelegere mai cuprinzătoare a mediului lor. Explorând în continuare modul de asimilare eficientă a acestor experiențe variate, putem îmbunătăți eficacitatea generală și adaptabilitatea sistemelor RL mulți-agent. Putem îmbogăți și mai mult această idee luând în considerare sisteme mulți-agenți, în care mai mulți agenți RL interacționează cu mediile complexe. Fiecare agent ar putea lucra la sarcini individuale, contribuind la un obiectiv colectiv mai mare. Prin comunicare, acești agenți și-ar putea împărtăși ipotezele, constatările și comportamentele învățate, construind în mod eficient o „cunoaștere culturală" comună. Acest concept ar putea revoluționa modul în care înțelegem sistemele mulți-agenți, promovând rezolvarea mai eficientă a problemelor și construirea cunoștințelor printr-o formă de „inteligență colectivă".

Am investigat mecanisme de memorie, învățarea prin imitație și învățarea prin transfer atât în contextul roboticii în spații interioare, cât și al vehiculelor autonome, că parte a investigației noastre privind navigația robotică în lumea reală. În ciuda acestor evoluții, este încă dificil producem agenți care reușesc să se autoadapteze, astfel încât aceștia să se poată integra cu ușurință cu datele din lumea reală și la medii noi după învățarea în simulator. Cercetările viitoare ar trebui să se concentreze pe crearea de modele care să permită agenților RL să se recalibreze în mod independent în timp ce fac inferențe că răspuns la condițiile de mediu. O astfel de avansare le-ar îmbunătăți considerabil capacitatea de a naviga,

le-ar consolida capacitatea de a învăța din experiențele practice și ar face tranziția de la mediile simulate la mediile din lumea reală mai simplă.

Cercetarea noastră a făcut progrese importante în îmbunătățirea diversității pentru descoperirea moleculară de novo și în facilitarea generalizării într-o varietate de compuși propuși în domeniul descoperirii de medicamente. Cu toate acestea, includerea incertitudinii în procesul de formare a primit puțină atenție. Astfel, studiile suplimentare ar trebui să se concentreze pe încorporarea incertitudinii în modelele RL (învățare prin recompensă) pentru descoperirea medicamentelor. Această dezvoltare ar putea îmbunătăți semnificativ robustețea și generalizarea compușilor propuși. De asemenea, este important să se analizeze modul în care compromisul explorare-exploatare în cadrul RL poate fi ghidat de incertitudinea învățată a unui model, accelerând potențial procesul de descoperire. Metodele cu ansamblu de rețele sau abordărilor bayesiene pot oferi, de asemenea, o modalitate de a captura și de a utiliza eficient incertitudinea în această situație.

Pe măsură ce continuăm să avansăm în domeniul învățării prin recompensă, o zonă critică de explorare viitoare se află dincolo de domenii specifice de cercetare și se extinde în direcții mai largi, generale. Dezvoltarea agenților RL care pot interacționa cu lumi deschise și obiective auto-stabilite promite un mod mai dinamic de explorare și învățare. Această noțiune provoacă preceptele tradiționale RL, propunând o schimbare în care agenții își determină spațiul de stare, mai degrabă decât să fie limitați de medii predefinite. Această tranziție dă putere agenților să-și modeleze lumea, creând un proces de învățare mai adaptabil și mai receptiv. În același timp, există potențialul de a ne regândi dependența de cadrul tradițional al procesului de decizie Markov. Prin adoptarea unui flux continuu de interacțiuni stări-spațiu, am putea îmbunătăți capacitățile agenților noștri de a gestiona situații noi, împingând astfel granițele generalizării și explorării. Această abordare le-ar permite agenților noștri să se ocupe de vastele complexități ale unei lumi deschise, în continuă schimbare, în care s-ar putea să nu întâlnească niciodată aceeași stare de două ori. O altă cale promițătoare pentru cercetări viitoare ar putea fi aplicarea meta-învățării în politică de explorare în sine. Această abordare implică pregătirea agenților RL pentru a învăța cum să exploreze, în loc să genereze motivația intrinsecă care duce la o mai bună explorare. Scopul ar fi de a dota agenții cu capacitatea de a determina în mod adaptativ cele mai eficiente strategii de explorare pe baza experiențelor, mediului și sarcinilor lor. Învățând să genereze o politică de explorare, agenții RL ar putea naviga mai bine în medii sau scenarii nefamiliare, ceea ce duce la o învățare mai eficientă. Cu toate acestea, este important să recunoaștem că nu este neapărat de dorit doar creșterea complexității în toate cercetările RL. Creșterea complexității și capacităților pot introduce redundanță și ineficiență. Ar trebui să se pună accent pe crearea unor sisteme mai eficiente și inteligente, care să valorifice ceea ce au învățat pentru a-și îmbunătăți performanța. Accentul ar trebui să fie pe „a învăța mai bine”, nu doar pe „a învăța mai mult”. Cu această abordare echilibrată, putem reuși un obiectiv aplicat al RL: crearea de sisteme inteligente care să poată naviga și să se adapteze la mediul lor în același mod în care o fac oamenii.

În concluzie, chiar dacă cercetarea RL a avansat semnificativ, arii vaste de cercetare pentru îmbunătățirea acestor algoritmi sunt încă de explorat. Cu toate acestea, recunoscând limitările noastre și bazându-ne pe cunoștințele dobândite, putem continuă să depășim granițele RL, apropiindu-ne tot mai mult de obiectivul nostru de a crea sisteme de învățare inteligente, adaptabile și eficiente.

Bibliografie

[1] Dario Amodei and Jack Clark. Faulty Reward Functions in the Wild. *https://blog.openai.com/faulty-reward-functions/*, 2016.

[2] Ryan Paul Badman, Thomas Trenholm Hills, and Rei Akaishi. Multiscale Computation and Dynamic Attention in Biological and Artificial Intelligence. *Brain Sciences*, 10(6):396, 2020.

[3] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *arXiv preprint arXiv:2106.04399*, 6 2021. URL http://arxiv.org/abs/2106.04399.

[4] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with Large Scale Deep Reinforcement Learning. 2021. URL https://www.facebook.com/OGDota2/.

[5] Ali Borji, D N Sihite, and L Itti. Salient Object Detection: A Benchmark. Computer Vision—ECCV 2012: the 12th European Conference on Computer Vision; 2012 Oct 7-13; Florence, Italy, 2012.

[6] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

[7] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic Gridworld Environment for OpenAI Gym. \url{https://github.com/maximecb/gym-minigrid}, 2018.

[8] Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. Machine Learning in Drug Discovery: A Review. *Artificial Intelligence Review*, 55(3): 1947–1999, 3 2022. ISSN 0269-2821. doi: 10.1007/s10462-021-10058-4. URL https://link.springer.com/10.1007/s10462-021-10058-4.

[9] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature 2022 602:7897*, 602(7897):414–419, 2 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04301-9. URL https://www.nature.com/articles/s41586-021-04301-9https://www.nature.com/articles/s41586-021-04301-9%E2%80%A6.

[10] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of Real-World Reinforcement Learning. *arXiv preprint arXiv:1904.12901*, 2019.

[11] Richard E Fikes, Peter E Hart, and Nils J Nilsson. Learning and executing generalized robot plans. *Artificial Intelligence*, 3:251–288, 1 1972.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. 6 2014. URL http://arxiv.org/abs/1406.2661.

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. 12 2014. URL http://arxiv.org/abs/1412.6572.

[14] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9 (4):188–194, 2005.

[15] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. 3 2022. URL http://arxiv.org/abs/2203.17003.

[16] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[17] Alex Irpan. Deep Reinforcement Learning Doesn't Work Yet. \url{https://www.alexirpan.com/2018/02/14/rl-hard.html}, 2018.

[18] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population Based Training of Neural Networks. 11 2017. URL http://arxiv.org/abs/1711.09846.

[19] Andrew Jaegle, Vahid Mehrpour, and Nicole Rust. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *Current Opinion in Neurobiology*, 58:167–174, 10 2019. ISSN 09594388. doi: 10.1016/j.conb.2019.08.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0959438819300054.

[20] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[21] Jerome Kagan. Motives and development. *Journal of personality and social psychology*, 22(1):51, 1972.

[22] Khimya Khetarpal, Martin Klissarov, Maxime Chevalier-Boisvert, Pierre-Luc Bacon, and Doina Precup. Options of Interest: Temporal Abstraction with Interest Functions. *AAAI*, 2020.

[23] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. 9 2016. URL http://arxiv.org/abs/1609.02907.

[24] Guillaume Lample and Devendra Singh Chaplot. Playing FPS Games with Deep Reinforcement Learning. In *AAAI*, pages 2140–2146, 2017.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015. ISSN 0028-0836. doi: 10.1038/nature14539. URL http://www.nature.com/articles/nature14539.

[26] Tod S Levitt. Qualitative navigation for mobile robots. *Int. J. Artificial Intelligence*, 44:305–360, 1990.

[27] Kit-Kay Mak and Mallikarjuna Rao Pichika. Artificial intelligence in drug development: present status and future prospects. *Drug discovery today*, 24(3):773–780, 3 2019. ISSN 1878-5832. doi: 10.1016/j.drudis.2018.11.014. URL http://www.ncbi.nlm.nih.gov/pubmed/30472429.

[28] Dominic D. Martinelli. Generative machine learning for de novo drug discovery: A systematic review. *Computers in Biology and Medicine*, 145:105403, 6 2022. ISSN 00104825. doi: 10.1016/j.compbiomed.2022.105403. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482522001950.

[29] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data, 4 2017. ISSN 2640-3498. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

[30] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, and others. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[31] Varnavas D. Mouchlis, Antreas Afantitis, Angela Serra, Michele Fratello, Anastasios G. Papadiamantis, Vassilis Aidinis, Iseult Lynch, Dario Greco, and Georgia Melagraki. Advances in De Novo Drug Design: From Conventional to Machine Learning Methods. *International Journal of Molecular Sciences*, 22(4):1676, 2 2021. ISSN 1422-0067. doi: 10.3390/ijms22041676. URL https://www.mdpi.com/1422-0067/22/4/1676.

[32] Anna C Nobre and Mark G Stokes. Premembering experience: a hierarchy of time-scales for proactive attention. *Neuron*, 104(1):132–146, 2019.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[34] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. 3 2022. URL http://arxiv.org/abs/2203.02155.

[35] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

[36] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep Reinforcement Learning for De-Novo Drug Design. *Science Advances*, 4(7), 11 2017. doi: 10.1126/sciadv.aap7885. URL http://arxiv.org/abs/1711.10907http://dx.doi.org/10.1126/sciadv.aap7885.

[37] V Srinivasa Rao and K Srinivas. Modern drug discovery process: An in silico approach. *Journal of Bioinformatics and Sequence Analysis*, 2(5):89–94, 2011. ISSN 2141-2464. URL http://www.academicjournals.org/JBSA.

[38] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment and state-dependent noise-an intrinsic motivation for avoiding unpredictable agents. In *Artificial Life Conference Proceedings 13*, pages 118–125. MIT Press, 2013.

[39] F Schler. 3d path planning for autonomous aerial vehicles in constrained spaces [Ph. D. thesis]. *Department of Electronic Systems, Faculty of Engineering and Science, Aalborg University, Aalborg, Denmark*, 2012.

[40] Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Workshop on Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76. Springer, 2008.

[41] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

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

[43] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 10 2017. ISSN 0028-0836. doi: 10.1038/nature24270. URL http://www.nature.com/articles/nature24270.

[44] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. doi: 10.1016/j.artint.2021.103535. URL www.elsevier.com/locate/artint.

[45] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. 3 2015. URL http://arxiv.org/abs/1503.03585.

[46] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

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

[48] Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[49] Richard S Sutton, Ashique Rupam Mahmood, and Martha White. An Emphatic Approach to the Problem of Off-policy Temporal-Difference Learning. *Journal of Machine Learning Research*, 17: 73:1–73:29, 2016.

[50] Chen Tessler, Shahar Givony, Tom Zahavy, Daniel J Mankowitz, and Shie Mannor. A Deep Hierarchical Approach to Lifelong Learning in Minecraft. In *AAAI*, pages 1553–1561, 2017.

[51] Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-WEKA. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, volume Part F128815, pages 847–855, New York, NY, USA, 8 2013. ACM. ISBN

9781450321747. doi: 10.1145/2487575.2487629. URL https://dl.acm.org/doi/10.1145/2487575.2487629.

[52] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42(2): 245–284, 2 2015. ISSN 0219-1377. doi: 10.1007/s10115-013-0706-y. URL http://link.springer.com/10.1007/s10115-013-0706-y.

[53] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. Applications of machine learning in drug discovery and development. *Nature reviews. Drug discovery*, 18(6): 463–477, 6 2019. ISSN 1474-1784. doi: 10.1038/s41573-019-0024-5. URL http://www.ncbi.nlm.nih.gov/pubmed/30976107http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6552674.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. 6 2017. URL http://arxiv.org/abs/1706.03762.

[55] O Vinyals, I Babuschkin, J Chung, M Mathieu, M Jaderberg, W Czarnecki, A Dudzik, A Huang, P Georgiev, R Powell, and others. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II, 2019.

[56] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, and others. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[57] Martha White. Unifying Task Specification in Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning, {ICML} 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3742–3750, 2017.

[58] Olivier J Wouters, Martin McKee, and Jeroen Luyten. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844–853, 2020.

[59] Fei Yan, Yi-Sha Liu, and Ji-Zhong Xiao. Path planning in complex 3D environments using a probabilistic roadmap method. *International Journal of Automation and computing*, 10(6):525–533, 2013.

[60] Namik Kemal Yilmaz, Constantinos Evangelinos, Pierre F J Lermusiaux, and Nicholas M Patrikalakis. Path planning of autonomous underwater vehicles for adaptive sampling using mixed integer linear programming. *IEEE Journal of Oceanic Engineering*, 33(4):522–537, 2008.

[61] Xianyuan Zhan, Haoran Xu, Yue Zhang, Xiangyu Zhu, Honglei Yin, and Yu Zheng. DeepThermal: Combustion Optimization for Thermal Power Generating Units Using Offline Reinforcement Learning. 2 2021. URL http://arxiv.org/abs/2102.11492.

[62] Barret Zoph and Quoc V. Le. Neural Architecture Search with Reinforcement Learning. 11 2016. URL http://arxiv.org/abs/1611.01578.