VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INTELIGENTNÍCH SYSTÉMŮ
DEPARTMENT OF INTELLIGENT SYSTEMS

HEURISTIKY PRO HRANÍ HRY SCOTLAND YARD

HEURISTICS FOR THE SCOTLAND YARD BOARD GAME

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **MICHAL CEJPEK**
AUTHOR

VEDOUCÍ PRÁCE **doc. Ing. FRATIŠEK ZBOŘIL, Ph.D.**
SUPERVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání bakalářské práce

156221

Ústav: Ústav inteligentních systémů (UITS)
Student: **Cejpek Michal**
Program: Informační technologie
Název: **Heuristiky pro hraní hry Scotland Yard**
Kategorie: Umělá inteligence
Akademický rok: 2023/24

Zadání:

1. Seznamte se s pravidly deskové hry typu "Scotland Yard", kdy pozice jedné z figur bývá protihráčům ukázána jen v některých kolech hry. Také nastudujte výsledky, které pro automatické hraní této hry byly dosaženy.
2. Určete metody, které by měly být důvodně vhodné pro realizaci systému, který bude hrát tuto hru autonomně. Zaměřte vedle klasických metod hraní her i metody pro počítačové učení, jako jsou například metody posilovaného učení a hlubokého učení.
3. Pro jednotlivé role figur ve hře implementujte algoritmy řízení a ověřte jejich schopnost plnit zadané cíle.
4. Vyhodnoťte úspěšnost obou stran hry pro různé míry zapojení metod strojového učení a diskutujte zjištěné výsledky.

Literatura:

- Nijssen, J., A., M., Winands, H., M.:"Monte Carlo Tree Search for the Hide-and-Seek Game Scotland Yard", IEEE Transactions on Computational Intelligence and AI in Games 4(4):282 - 294, 2012
- Norvig, P., Russel, S. :"Artificial Intelligence, A Modern Approach", Prentice Hall, 2020
- Daniel Borák: "Heuristic Evaluation in the Scotland Yard Game", Bakalářská práce, 2021, ČVUT

Při obhajobě semestrální části projektu je požadováno:
První dva body zadání

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Zbořil František, doc. Ing., Ph.D.**
Vedoucí ústavu: Hanáček Petr, doc. Dr. Ing.
Datum zadání: 1.11.2023
Termín pro odevzdání: 9.5.2024
Datum schválení: 6.11.2023

Abstrakt

Tato práce se zabývá možností použití algoritmů hlubokého a posilovaného učení pro řešení problémů s neúplnou informací. Konkrétně je hlavním zkoumaným algoritmem je PPO – Proximal Policy Optimization (optimalizace proximální politiky).

Práce se zabývá jeho teoretickými základy a následně jeho aplikací na hru Scotland Yard. Výsledky jsou porovnány s jinými algoritmy a je provedena analýza výhod a nevýhod zhotovené implementace.

[Výsledky jsou . . .]

Abstract

asd Přeložit asds

Klíčová slova

Sem budou zapsána jednotlivá klíčová slova v českém (slovenském) jazyce, oddělená čárkami.

Keywords

Sem budou zapsána jednotlivá klíčová slova v anglickém jazyce, oddělená čárkami.

Citace

CEJPEK, Michal. *Heuristiky pro hraní hry Scotland Yard*. Brno, 2023. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. Ing. František Zbořil, Ph.D.

Rozšířený abstrakt

Úvod

Hra Scotland Yard

Experimenty

- *Porovnání s náhodnou politikou*
- *Porovnání s jinými algoritmy - implementace DQN - v knihovně rlib*
- *Porovnání s monte carlo algoritmem - zdroj na internetu*

Heuristiky pro hraní hry Scotland Yard

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana X... Další informace mi poskytli... Uvedl jsem všechny literární prameny, publikace a další zdroje, ze kterých jsem čerpal.

.....
Michal Cejpek
20. března 2024

Poděkování

V této sekci je možno uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc (externí zadavatel, konzultant apod.).

Obsah

Seznam obrázků

Kapitola 1

Úvod

Umělá inteligence je obor, který nás postupem let všechny obkopuje čím dál tím více. Dokonce je navždy spjata i s naším českým národem, když Karel Čapek dal zrodu slova robot.

Pokrok umělé inteligence je často měřen aplikací v oblasti her. Hry jsou vhodným ukazatelem pokroku v oblasti umělé inteligence, protože mají jasná pravidla, výkon je snadno měřitelný a pokrok dokáže vidět i lajk. Umělá inteligence již dokázala porazit nejlepší hráče v šachu [8], Dota 2 [9] a Go [3].

Hra studovaná v této práci je hra Scotland Yard. Je to hra pro tři až šest hráčů. V této hře obvykle hraje jeden hráč jako Pan X, který se snaží uniknout policistům, ovládanými ostatními hráči. Policisté avšak nevědí, kde na herním poli se Pan X nachází. Musí tedy odhadovat jeho pozici a spolupracovat mezi sebou, aby ho mohli polapit. Pozice Pana X je odhalena pouze v určitých kolech. Hra končí, když je Pan X chycen (vyhrávají policisté), nebo když je dosažen maximální počet kol (vyhrává Pan X). Scotland Yard je ideální hrou pro studování umělé inteligence, protože je hra s neúplnou informací a k vítězství policistů je zapotřebí spolupráce.

Zaměření této práce jsem si vybral jelikož mi vždy byla umělá inteligence blízká a vždy jsem chtěl . Avšak jsem nikdy nenašel odhodlání ponořit se do této oblasti.

Rozhodl jsem se pro bližší zkoumání algoritmů posilovaného učení, konkrétně algoritmu PPO (Proximal Policy Optimization). Tento algoritmus je často používán pro řešení problémů se spojitými veličinamy a ve 3D prostoru. Dle provedených studií je vhodný pro řešení problémů s neúplnou informací [1] a je vhodný pro hry na schování a hledání [2]. Proto je pro mě zajímavý a zkušenost s tímto algoritmém by se dala využít v mém pracovním životě.

Pro zpracování práce byl využity tyto hlavní knihovny:

- *Ray.Rlib* - knihovna s implementací algoritmu PPO
- *PyTorch* - podpůrná knihovna Ray.Rlib
- *TensorFlow* - podpůrná knihovna Ray.Rlib
- *Gym* - knihovna pro vytvoření prostředí pro učení
- *Pygame* - knihovna pro vytváření uživatelského rozhraní

Kapitola 2

Shrnutí dosavadního stavu

- *40-50 % rozsahu práce*
- *Hodně citovat literaturu*
- *Vysvětlit všechno, už ne pro plebíky*
- *Je vhodné na začátku této části uvést, co obsahuje a proč a taky že „není encyklopedickým přehledem“*
- *Asi tak ze 2 kapitol?*
- *Existující řešení (implementace scotlandu, říct že se implementuje pomocí tamtoho algoritmu a proč jsem vzal PPO)*

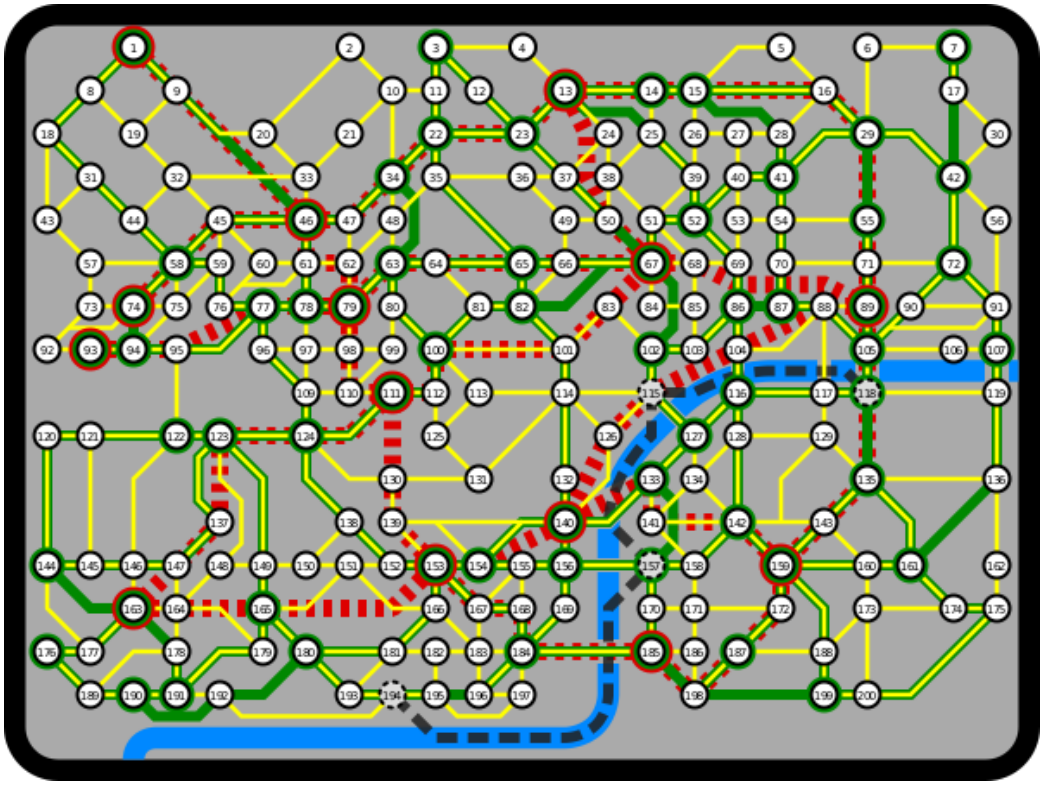
2.1 Desková hra Scotland Yard

Scotland Yard je populární hra pro tři a více hráčů, která kombinuje prvky schovávané a hry na honěnou. Jeden hráč hraje za Pana X, který se snaží uniknout policistům, ovládanými ostatními hráči. Hra končí, když je Pan X chycen (vyhrávají policisté), nebo když je dosažen maximální počet kol (vyhrává Pan X). Originální hra se odehrává v Londýně. Na herní mapě se nachází 200 polí, které jsou vzájemně propojené náhodnými cestami. Každá cesta povoluje určitý způsob pohybu (např. pouze taxíkem, pouze autobusem, atd.). Jednotliví hráči využívají prvky veřejné dopravy k pohybu po herní ploše, kterými jsou:

- *Taxi*
- *Autobus*
- *Metro*
- *Trajekt*

Každému hráči je na začátku hry přidělen pouze určitý počet jízdenek na tyto dopravní prostředky. Pro využití dopravy je potřebná právě tato jízdenka. Pokud ji hráč nemá, nemůže již tento způsob přepravy použít. Hra se dělí na kola, ve kterých se hráči střídají.

Hlavní myšlenkou hry je, že po většinu kol je pozice Pana X je policistům utajena. Odhaluje se jim pouze určená kola. To znamená, že policisté musí odhadovat další kroky Pana X aby ho mohli polapit. Tímto se ze hry Scotland Yard stává hra s neúplnou informací, jelikož policisté nevidí přesnou pozici Pana X. Tento fakt ji činí vhodnou pro studování a rozvíjení oboru umělé inteligence.



Obrázek 2.1: Ukázka herní mapy pro hru Scotland Yard. Zdroj:[6]

2.2 Bayesovské hry: Hry s neúplnou informací

V oblasti umělé inteligence hraje důležitou roli modelování a řešení her. Hry představují abstraktní formalizaci konfliktních interakcí mezi aktéry, tzv. hráči. Klasická teorie her se zaměřuje na hry s úplnou informací, kde mají všechny strany v daném okamžiku přístup ke všem relevantním informacím z herního prostředí. V praxi se však častěji setkáváme se situacemi kde jednotlivým stranám chybí určité informace. Tyto případy lze modelovat pomocí her s neúplnou informací, kde hráči nemají úplné znalosti o prostředí či soupeřích. Neúplnou informaci můžeme sledovat například v:

- *Ekonomii* - kde se jedná o neúplnou informaci o trhu, cenách, situačních výkyvech, atd.
- *Vojenské strategii* - kde se jedná o neúplnou informaci o pozici nepřítele, jeho vybavení, strategii, cíli, atd.
- *Sportovní hry* - kde se jedná o neúplnou informaci o taktice soupeře, jeho schopnostech, atd.

Bayesovská hra s neúplnou informací [7] je definována pětící $(N, A_i, \theta_i, p(\theta_i), u_i)$, kde:

- N je konečná množina hráčů, $N = \{1, 2, \ldots, n\}$.
- A_i je neprázdná množina strategií hráče i.
- θ_i je neprázdná množina typů hráče i.

- $p(\theta_i)$ je apriorní pravděpodobnostní rozdělení typu hráče i na θ_i.
- $u_i : A_1 \times \cdots \times A_n \times \theta_1 \times \cdots \times \theta_n \to \mathbb{R}$ je výplatní funkce hráče i.

Bayesovské hry představují formální rámec pro modelování her s neúplnou informací.

Stratego

Stratego je desková strategická hra pro dva hráče, která se odehrává na hracím plánu rozděleném do políček a využívá se k ní sada figurek reprezentujících armádu. Vychází z dřívějších her, jako je Šachy a Go, a kombinuje strategické plánování, taktické manévry.

Cílem hry je porazit soupeře nalezením a obsazením jeho vlajky. Hráči to dělají tak, že se navzájem utkávají se svými figurkami na herním plánu.

Každý hráč má 40 figur, rozdělených do 11 hodností (generál, plukovník, skaut, atd.). Figury lze rozeznat jen z jedné strany, proto oponent neví o jakou figuru se jedná. Hráči se střídají v tazích, kdy se snaží najít oponentovu vlajku. Hra začíná tím, že každý hráč rozmístí své figury na herní pole. Hráči se střídají v tazích, kdy se snaží najít oponentovu vlajku. Pokud hráč táhne na pole, kde se nachází oponentova figura, nastává souboj. Souboj spočívá v odkrytí obou figur a vyhrává ta s vyšší hodností. Figura, která vyhrála zůstává, poražená figura je odstraněna z hry. Ve hře Stratego je důležité blafování a odhadování soupeřových tahů.

Obrázek 2.2: Ukázka rozestavěných figur ve hře Stratego. Zdroj:[5]

Z pohledu umělé inteligence je Stratego zapeklitý problém. Nejenže je hra s neúplnou informací a je tedy zapotřebí odhadovat oponentovy tahy a blafovat, ale také je hra s velkým prostorem stavů 10^{535}[10]. Až do roku 2022 se AI nepodařilo porazit expertního hráče v této hře. To se změnilo příchodem *DeepNash*[10], kdy se tato metoda umístila mezi 3 nejlepšími hráči světa.

2.3 Klíčové koncepty posilovaného učení

Posilované učení (Reinforcement Learning, RL) je oblast strojového učení, která se zaměřuje na učení agentů v dynamickém prostředí. Agent se učí strategii chování, která maximalizuje kumulativní odměnu.

2.3.1 Agent

Je komplexní entita, která interaguje s prostředím. Prostředí poskytuje agentovi informace o stavu a agent na základě těchto pozorování vykonává akce. Tyto akce mohou ovlivnit stav prostředí a agent obdrží odměnu na základě odměnové funkce. Agent volí takové akce aby maximalizoval kumulativní odměnu.

2.3.2 Prostředí

Je vše s čím agent interaguje. Prostředí je buď fyzické (entity z reálného světa, ovládání chytré domácnosti, ovládání reaktoru, atd.) nebo virtuální (simulace, například hra). Prostředí reaguje na akce agenta poskytuje mu zpětnou vazbu ve formě odměny či trestu (záporná odměna). Pokud v prostředí existuje více agentů, může mít každý agent jiné pozorování. Diky tomuto můžeme například schovat agentu A určité informace, které agent B vidí.

2.3.3 Model

Je matematická funkce, která popisuje chování prostředí v závislosti na agentových akcích. Model může být známý nebo neznámý, to následně rozděluje metody učení posilovaného na 2 základní kategorie: metody *s modelem* a metody *bez modelu*.

2.3.4 Politika

Pomocí posilovaného učení vzniká takzvaná politika. Politika je matematická funkce, která definuje agentovo chování na základě jeho pozorování (stavu). Snaží se definovat takové chování, které vede k maximální kumulativní odměně. Politika může být deterministická nebo stochastická.

Deterministická politika

Deterministická politika přesně definuje cílový stav přechodu pro každý stav. Agent tedy pro jeden stav vždy volí stejnou akci. Tato politika je vhodná pokud je zapotřebí v každém stavu reagovat konzistentně, bez odchylek. Například, pokud agent ovládá termostat v domě a teplota je pod požadovanou hladinu. Nemůže se stát aby byla šance, že agent zvolí akci, která teplotu ještě sníží. Další výhoda, je že je jednoduchá na interpretaci a implementaci.[4]

Stochastická politika

Zato stochastická politika definuje pro každý stav pravděpodobnostní rozdělení nad množinou akcí. Výsledná akce je tedy náhodná dle rozdělení pravděpodobnosti. Může tedy nastat situace kdy pro jeden stav agent zvolí jinou akci. Tato politika je vhodná pro situace, kdy je potřeba zkoumat různé strategie a kdy agent nemá úplnou informaci o prostředí. Například tam kde by deterministická politika zvolila jasnou akci A, stochastická politika by mohla s

malou pravděpodobností zvolit akci B. Čímž ale může odhalit, že stav B je s ohledem na komulativní odměnu lepší než stav A.[4]

2.3.5 Akce

Akce je přechod z aktuálního stavu, do následujícího stavu z množiny možných stavů. Zjednodušeně, je to rozhodnutí, které agent vykonává v prostředí a toto rozhodnutí ovlivňuje prostředí. Akce zvolena dle politiky a je závislá na pozorování agenta.

2.3.6 Odměna

Odměna je hodnota, kterou agent obdrží od prostředí po vykonání akce. Může být kladná, záporná nebo nulová. Dle této zpětné vazby se agent učí, jak moc byla jeho zvolená akce v daném stavu vhodná.

2.3.7 Hodnotová funkce

Hodnotová funkce vyhodnocuje, jak dobrý je stav tím, že predikuje budoucí odměnu. Čím vzdálenější odměna je, tím více je snížena. Jelikož, čím je odměna vzdálenější tím více je nejistá.

Existují dva typy hodnotových funkcí:

Hodnotová funkce stavu $V(s)$

Hodnotová funkce stavu $V(s)$ vyhodnocuje očekávanou komulativní odměnu, pokud se agent nachází v tomto stavu. Tato funkce je závislá na politice, kterou se agent řídí. Vyhodnocuje tedy jak příznivý je daný stav pro agenta.

Hodnotová funkce akce $Q(s, a)$

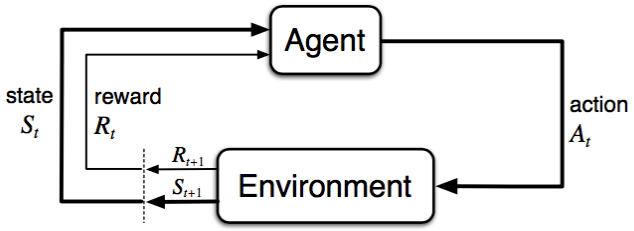
Hodnotová funkce akce $Q(s, a)$ vyhodnocuje očekávanou komulativní odměnu, pokud se agent nachází v tomto stavu a zvolí tuto akci. Tato funkce je opět závislá na politice, kterou se agent řídí. Vyhodnocuije tedy jak příznié je zvolení dané akce v aktuálním stavu.

2.3.8 Markovský rozhodovací proces

Teměř všechny problémy, řešené posilovaným učením, mohou být označeny jako Markovy rozhodovací procesy (Markov Decision Process). Tato abstrakce je základním kamenem pro modelování algoritmů posilovaného učení. Markovský rozhodovací proces značí, že následující stav není závislý na stavech minulých, nýbrž pouze na aktuálním stavu.

Markovský rozhodovací proces je definován pětic (S, A, P, R, γ)[11], kde:

- S je množina stavů.
- A je množina akcí.
- P je pravděpodobnostní přechodová funkce
- R je odměnová funkce
- γ je diskontní faktor pro budoucí odměny



Obrázek 2.3: Interakce mezi prostředím a agentem podle Markova rozhodovacího procesu. Zdroj:[11]

2.3.9 Bellmanova rovnice

Belmanova rovnice se zamuřuje na rozložení hodnotových funkcí na menží snadněji zpracovatelné celky. Dociluje toho tak, že rozděluje hodnotovou funkci na dvě části: *okamžitou odměnu* a postupně snižovanou *budoucí odměnu*.

$$V(s) = E[R_{t+1} + \gamma V(S_{t+1})|S_t = s] \tag{2.1}$$

$$Q(s,a) = E[R_{t+1} + \gamma E_{a\sim\pi} Q(S_{t+1}, a)|S_t = s, A_t = a] \tag{2.2}$$

2.3.10 Rovnováha mezi explorací a exploatací (exploration-exploitation)

Rozdíl mezi nekompletní a neúplnou informací

2.4 Vhodné algoritmy pro řešení her s neúplnou informací

Tato kapitola se zaměřuje na algoritmy, které jsou vhodné pro řešení her s neúplnou informací, s důrazem na metody posilovaného učení a srovnáním s klasickými metodami jako Monte Carlo.

Monte Carlo

Metoda Monte Carlo (MC) je pravděpodobnostní metoda, která využívá simulaci hry anásledné odhady hodnoty stavu.

Q-learning

Deep Q-learning

Metody gradientu politiky

Věta o gradientu politiky

2.5 Proximální optimalizace politiky

2.6 Algoritmus PPO

Kapitola 3

Zhodnocení současného stavu a plán práce (návrh)

- *Kritické zhodnocení dosavadního stavu*
- *Návrh, co by bylo vhodné vyřešit na základě znalostí dosavadního stavu*
- *Co jste konkrétně udělal s teorií popsanou výše*
- *Volba OS, jazyk, knihovny*
- *Detailní rozbor zadání práce, detailní specifikace a formulace cíle a jeho částí*
- *Popis použití řešení, situace/problémy, které projekt řeší*
- *Postup práce/kroky vedoucí k cíli, rozdělení celku na podčásti*
- *Návrh celého řešení i jeho částí, s odkazy na teoretickou část*

Zkoumaná modifikovaná verze hry Scotland Yard

Tato práce využívá modifikovanou verzi hry Scotland Yard, ve které se hráči pohybují po mřížkové herní ploše ve tvaru čtverce. Na mřížce se nachází [15x15] polí. Hračí se po těchto polích pohybují ortogonálně i diagonálně, vždy o maximálně 1 pole. Hráč se může rozhodnot nezměnit pozici a zůstat na svém aktuálním poli. K pohybu nejsou potřebné žádné jízdenky. Toto zjednodušení herní plochy nijak nemění základní podstatu hry, zachovává neurčitost, ale značně zjednodušuje implementaci. Hra začíná tím, že se vyberou náhodné možné pozice Pana X a policistů. Z těchto možných pozíc se následně náhodná pozice přidělí jednotlivým hráčům.

Implementace umělé inteligence do hry Scotland Yard pomocí algoritmu PPO

Vizuální stránka

Prostředí

Učení

//zmínit, že ray rlib neumí action mapping, takže nejde zamezit zvolení jistých akcí. Proto jsou možné i nevalidní akce. Ale jsou velmi penalizovány.

Kapitola 4

Experimenty

Kapitola 5

Závěr

Kapitola 6

Přílohy

Literatura

[1] Baes, J. *Application of Reinforcement Learning Algorithms to the Card Game Manille*. 2022. Diplomová práce. UGent. Faculteit Ingenieurswetenschappen en Architectuur.

[2] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G. et al. *Emergent Tool Use From Multi-Agent Autocurricula*. 2020.

[3] Borowiec, S. *AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol* [online]. 2019 [cit. 2024-03-18]. Dostupné z: https://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victory-over-grandmaster-lee-sedol.

[4] Carr, T. *Policies in Reinforcement Learning* [online]. March 2024 [cit. 2024-03-20]. Dostupné z: https://www.baeldung.com/cs/rl-deterministic-vs-stochastic-policies.

[5] Guinness, H. *Here's how a new AI mastered the tricky game of Stratego* [online]. 2022 [cit. 2024-03-20]. Dostupné z: https://www.popsci.com/technology/ai-stratego/.

[6] Mliu92. *Scotland Yard schematic* [online]. 2023 [cit. 2024-03-20]. Dostupné z: https://commons.wikimedia.org/wiki/File:Scotland_Yard_schematic.svg.

[7] Narahari, Y. *Game Theory* [online]. 2012 [cit. 2024-03-19]. Dostupné z: https://gtl.csa.iisc.ac.in/gametheory/ln/web-ncp13-bayesian.pdf.

[8] Newborn, M. *Kasparov versus deep blue: computer chess comes of age*. 1. vyd. Springer-VerlagBerlin, Heidelberg, 1996. ISBN 978-0-387-94820-1.

[9] OpenAI. *OpenAI Five defeats Dota 2 world champions* [online]. 2019 [cit. 2024-03-18]. Dostupné z: https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer).

[10] Perolat, J., De Vylder, B., Hennes, D., Tarassov, E., Strub, F. et al. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*. American Association for the Advancement of Science (AAAS). prosinec 2022, sv. 378, č. 6623, s. 990–996. DOI: 10.1126/science.add4679. ISSN 1095-9203. Dostupné z: http://dx.doi.org/10.1126/science.add4679.

[11] Weng, L. *A (Long) Peek into Reinforcement Learning* [online]. February 2018 [cit. 2024-03-20]. Dostupné z: https://lilianweng.github.io/posts/2018-02-19-rl-overview/#key-concepts.