ARIEL UNIVERSITY

תזמון ניתוחים לחדרי ניתוח בעזרת למידת חיזוקים

קורס: מבוא ללמידת חיזוקים

מרצה: ד"ר טדי לזבנק

מגישים:

נועה ענקי

רז אלבז

אושר דיגורקר

מבוא ומוטיבציה

- תזמון ניתוחים הוא אתגר תפעולי מורכב.
- נדרש פתרון שמפחית זמני המתנה ודחיית מקרים דחופים.
- RL מאפשר גישה דינמית וגמישה לבעיה.

הגדרת הבעיה והסביבה

State (מצב):

- סטטוס חדרי הניתוח (פנוי/תפוס)
- רשימת הממתינים עם סוג (רגיל/דחוף) וזמן המתנה
- מספר המנותחים שכבר טופלו
- עומס מצטבר/רשימת ממתינים נוכחית

Reward (תגמול):

- תגמול חיובי על שיבוץ מוצלח וחסכון בזמן המתנה
- תגמול שלילי (עונש) על דחיית חולה דחוף או יצירת שעות נוספות
- עידוד ליעילות תפעולית (לדוג' בונוס על ניצול מלא של חדרי ניתוח)

הגדרת הבעיה והסביבה

Action (פעולה):

- בחירת חולה (רגיל או דחוף) לשיבוץ בניתוח הבא
- אפשרות לבחירת סדרי עדיפויות (מי קודם ומתי)

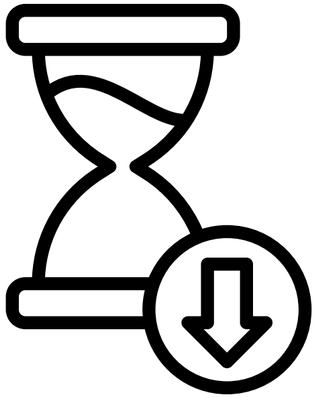
Dynamics (דינמיקה):

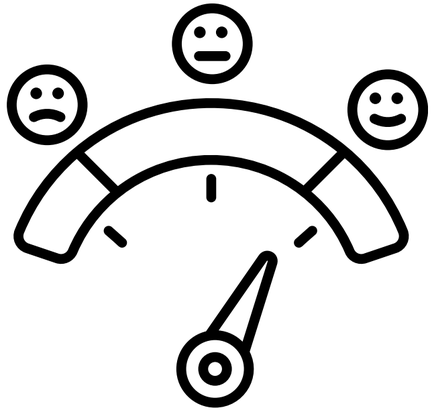
- הגעת חולים חדשים לאורך הזמן (סימולציה)
- אילוצים: מספר חדרי ניתוח, מגבלת זמן יומי, סדרי עדיפות
- תורים שמתארכים יוצרים לחץ ומגדילים עונש

הגדרת הבעיה והסביבה

<u>מטרת הסוכן:</u>

למקסם שביעות רצון, לצמצם זמני המתנה ולמזער דחיות של ניתוחים דחופים





סקירת ספרות

<u>גישות קיימות:</u>

- ניהול תורים לרוב מבוסס כללים ידניים או סטטיים.
- RL מאפשר תעדוף דינמי וחכם.

<u>השראה מהספרות:</u>

Xu ואחרים (2023): יישום RL לניהול תורי ניתוחים ודחיפות רפואית.

<u>הפרויקט שלנו:</u>

השראה עקרונית בלבד – הסביבה והיישום בפועל נבנו מאפס.

פיתוח סביבת הסימולציה

- פיתחנו סביבה ייעודית (env.py) המדמה תהליך תזמון ניתוחים לפי אילוצים אמיתיים.
- המערכת כוללת חדרי ניתוח, תור חולים (רגיל ודחוף), אילוצי זמן, ודינמיקה של כניסת חולים חדשים.
- כל צעד: הסוכן בוחר את החולה הבא לניתוח – בהתאם לחוקים והגבלות הסביבה.

סביבת הסימולציה – פירוט ומבנה התגמולים

סביבה: OperatingRoomEnv

- 3 חדרי ניתוח, יום עבודה באורך 480 דקות
- בכל צעד:
- חולים נכנסים למערכת בזמנים שונים (כולל דחופים)
- הסוכן בוחר איזה חולה לשבץ באיזה חדר, או להמתין

מצבים (State):

- מצב כל חדר (פנוי/תפוס, זמן סיום)
- זמן נוכחי ונותר ליום
- רשימת ממתינים: זמן המתנה, דחיפות (1–3)
- מספר ממתינים

פעולות (Action):

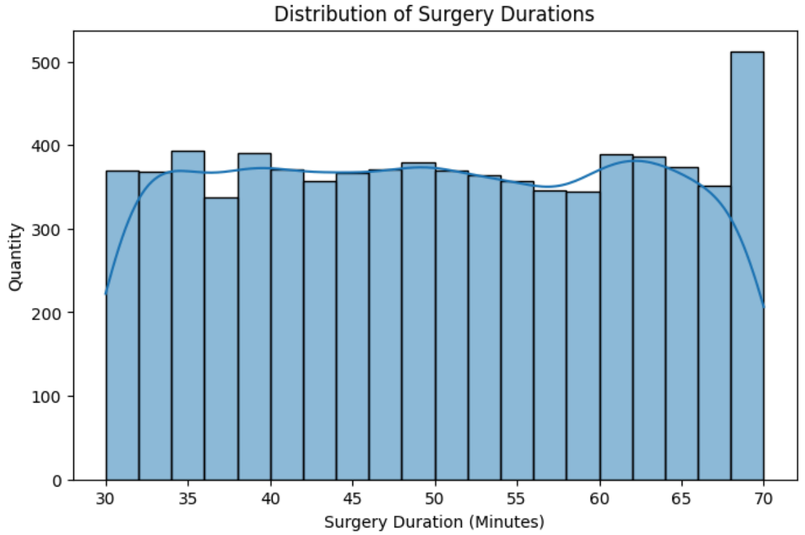
- לשבץ חולה (ספציפי) לחדר מסוים
- או לבחור "המתנה" (לא לשבץ)

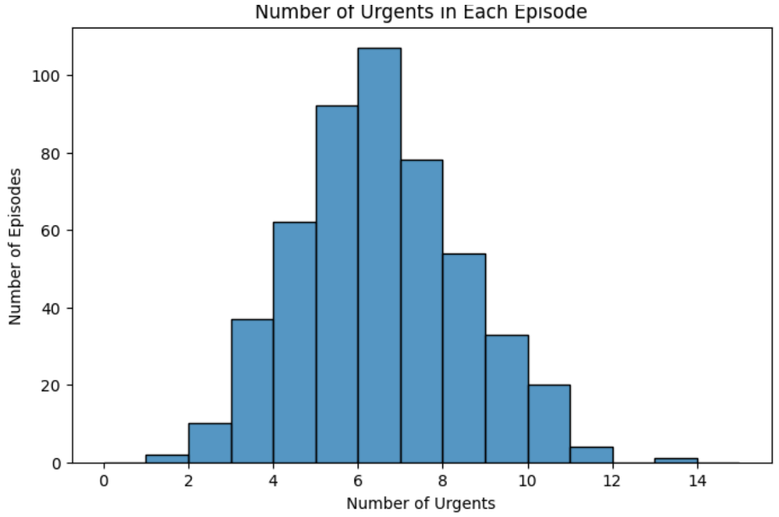
סביבת הסימולציה – פירוט ומבנה התגמולים

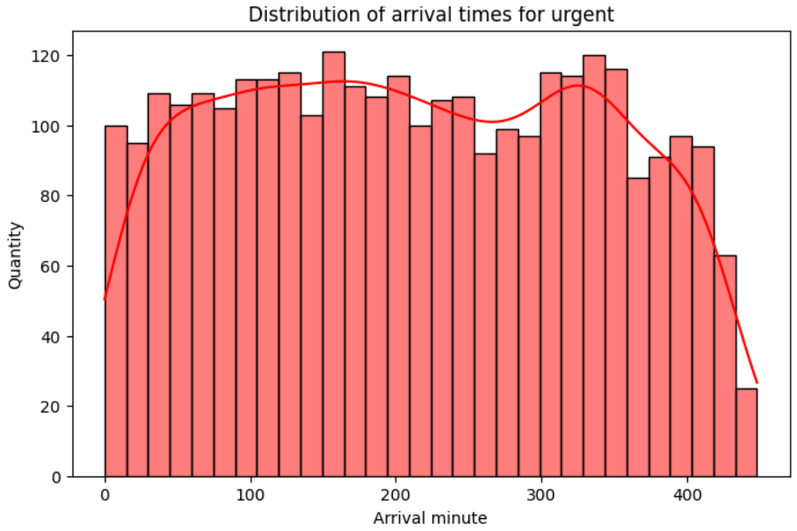
תגמולים ועונשים מרכזיים:

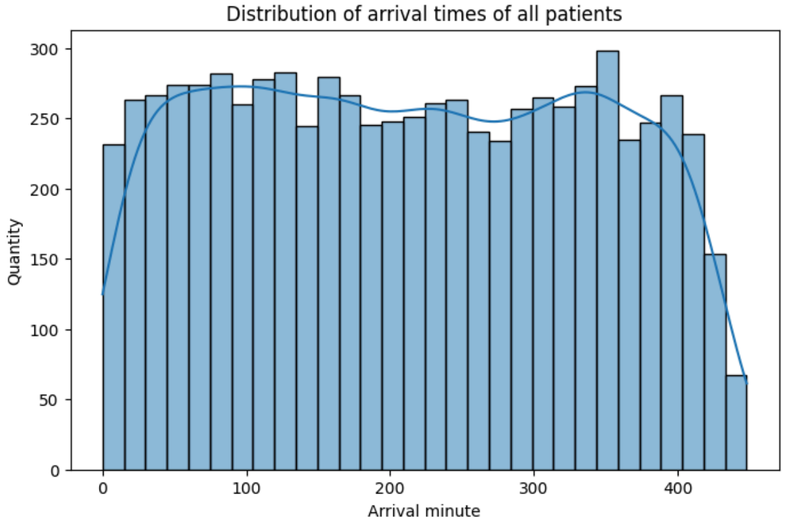
- +60 נק' על שיבוץ מוצלח של חולה
- +40 בונוס לשיבוץ דחוף (דחיפות 3)
- עונש מתמשך על המתנה לחולה (0.1–0.3 נק' לדקה, תלוי דחיפות)
- עונש חמור על דחיית ניתוח מעבר ליום (20–30 נק' לפי דחיפות)
- עונש על שעות נוספות (5 נק' לכל דקה מעבר)
- עונש על פעולה לא חוקית (2–20 נק', גודל קנס משתנה עם התקדמות הסוכן)
- בונוס חד-פעמי ליעילות גבוהה בסוף יום

התפלגות פרמטרים בסביבה

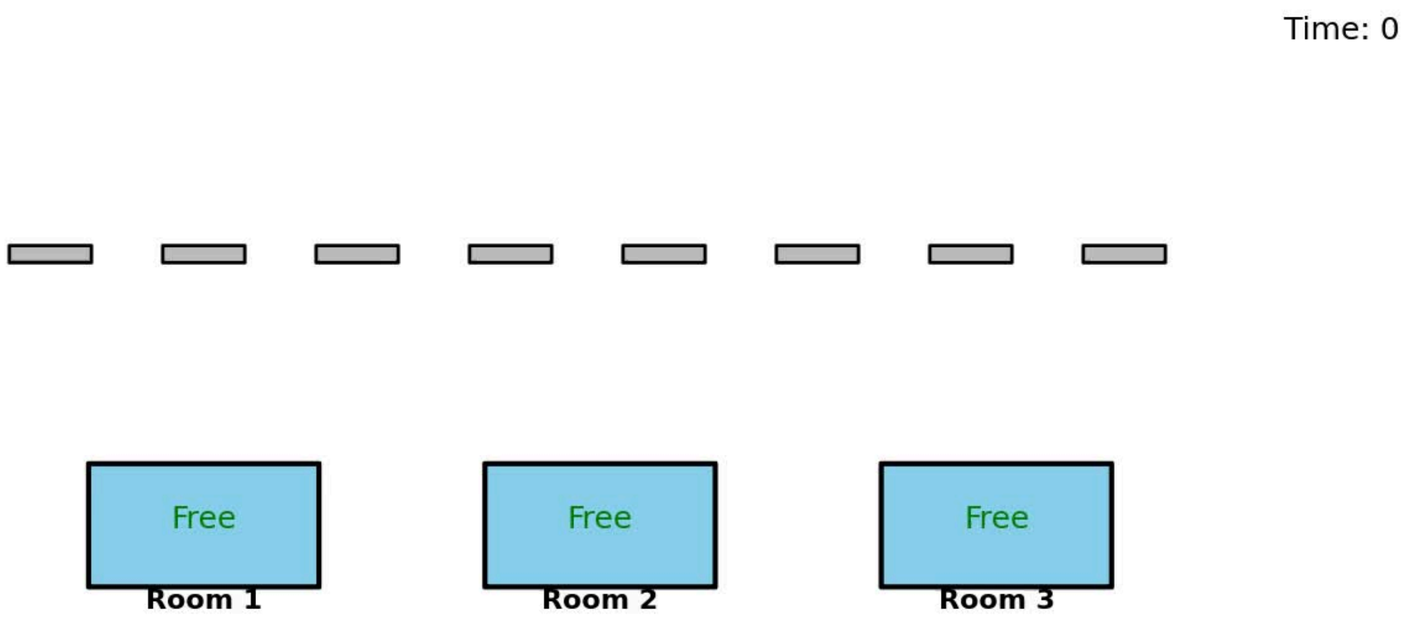






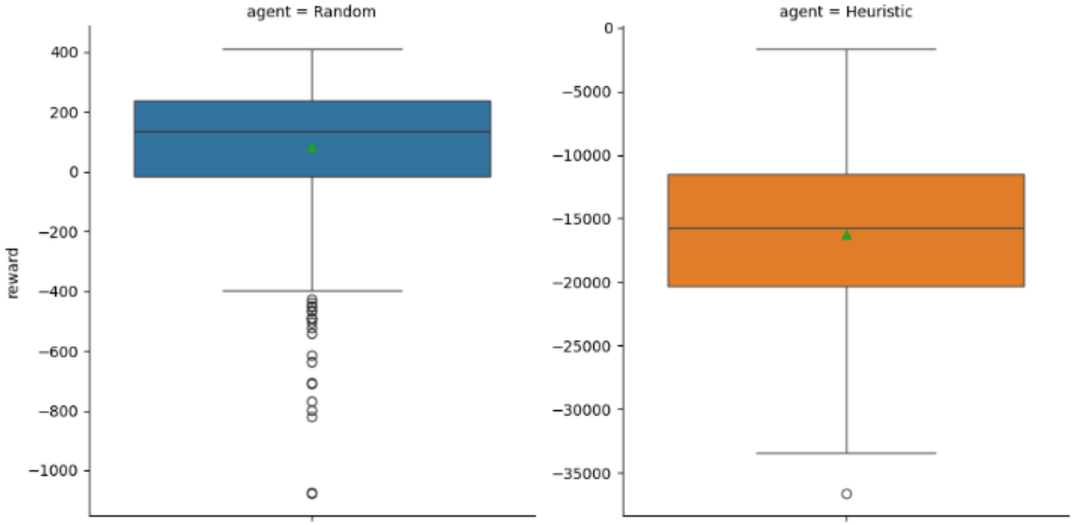


פיתוח סביבת הסימולציה



מימוש הסוכן, אלגוריתמים ו־Baseline

- פיתחנו סוכנים מבוססי RL: DQN, PPO, A2C.
- לצורך השוואה, נבחנו גם שני Baseline:
- סוכן אקראי (Random Agent)
- סוכן היוריסטי חמדני (Heuristic Agent)



מימוש הסוכן, אלגוריתמים ו־Baseline

מודל	עיקרון פעולה	למה בחרנו בו?	יתרון עיקרי	התנהגות/תוצאה בניסוי
Random Agent	בוחר פעולה באקראי	קו בסיס להשוואה	פשטות, מינימום למידה	ביצועים גרועים; הרבה דחיות, ניצול נמוך
Heuristic Agent	כלל פשוט: מי שהגיע נכנס	להשוואה עם שיטות אנושיות	פשטות, יעילות בסיסית	ביצועים בינוניים; מתמודד רק עם מקרים פשוטים
DQN	לומד ערך Q לכל מצב-פעולה בעזרת רשת נוירונים	Benchmark קלאסי ברוב סביבות RL	מהיר, פשוט, אפקטיבי	שיפור ניכר במדדים תפעוליים, למידה מהירה
PPO	עדכון מדיניות ישיר (Policy Gradient) באופן זהיר ומבוקר	יציבות ואפקטיביות בסביבות מורכבות	עמידות לשינויים, יציבות	ביצועים טובים במיוחד בתורים משתנים
A2C	שילוב של Actor (מדיניות) ו-Critic (ערכי מצב), עם עדכון ע"פ יתרון	קונברגנציה מהירה, למידה יעילה	יעילות בשילוב ערכים ומדיניות	למידה מהירה, הצלחה בניהול trade-offs

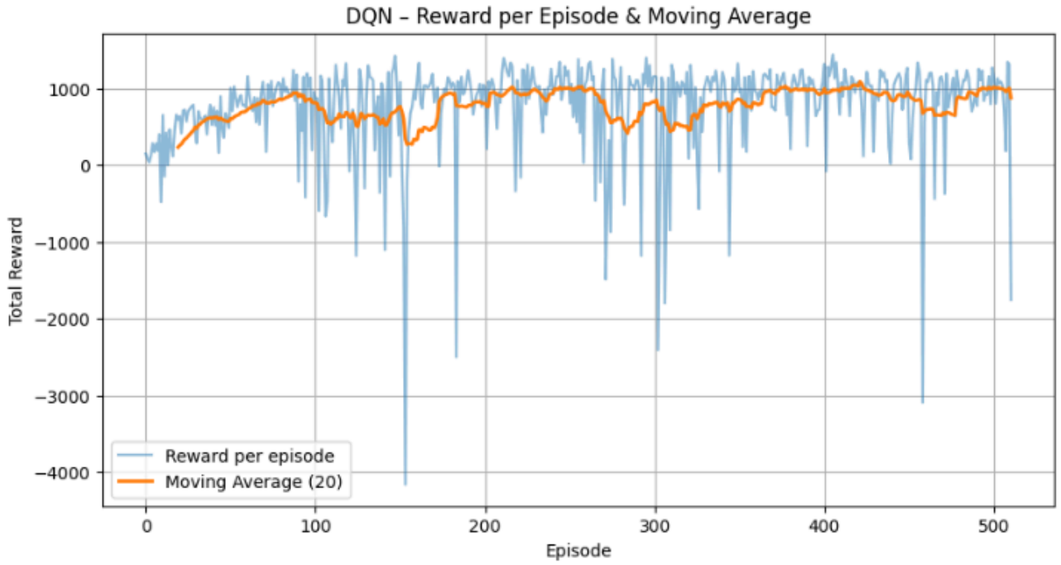
Grid Search

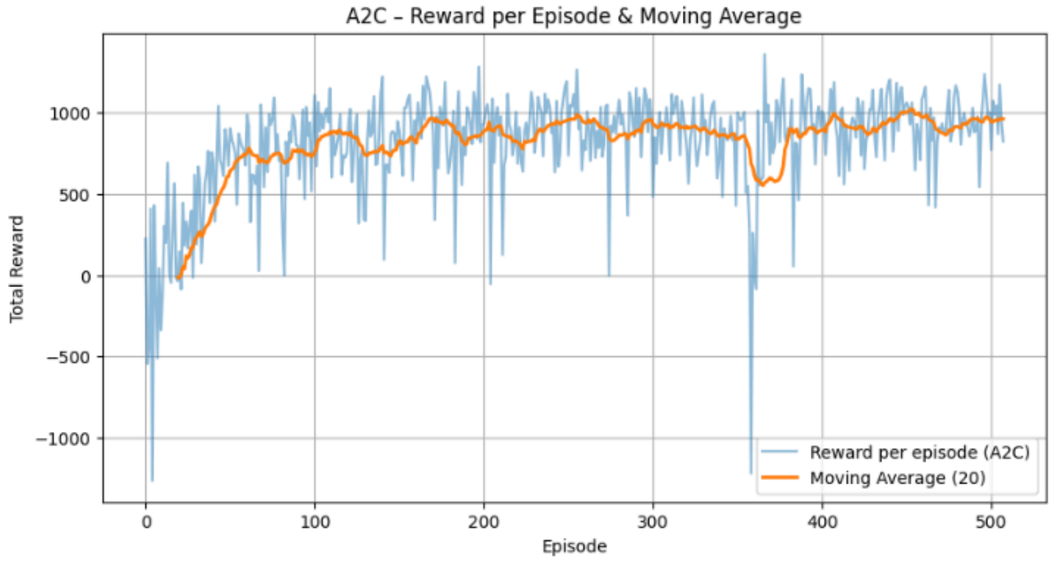
- לכל אלגוריתם (DQN, PPO, A2C) בוצע Grid Search על היפר־פרמטרים עיקריים.
- בדקנו מאות שילובים של: learning rate, batch size, מבנה רשת, exploration rate ועוד.
- כל מודל נבחן לפי ביצועים במדדים עיקריים (Reward, זמן המתנה, יציבות).

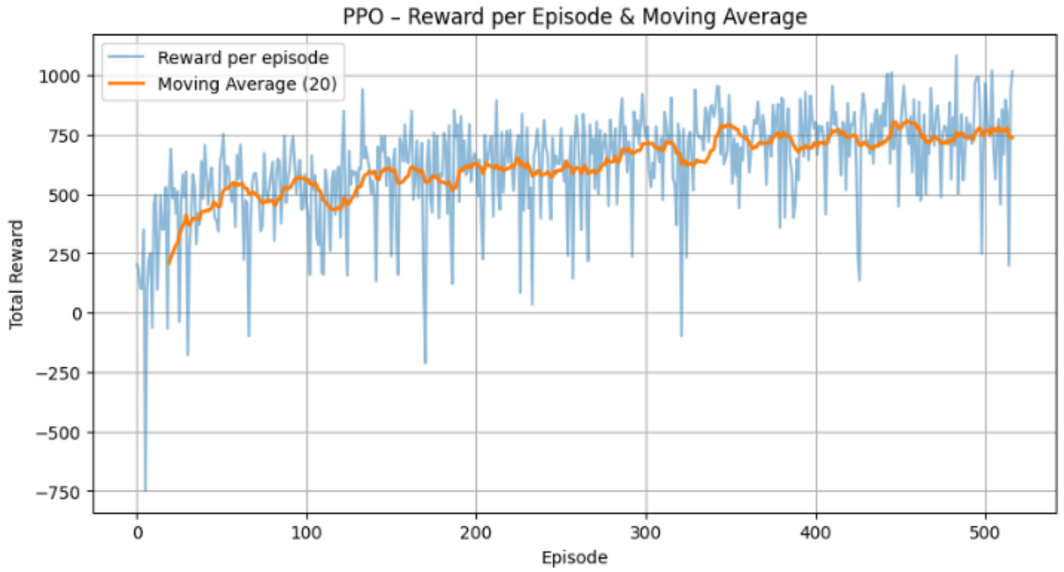
תכנון הניסויים ומדדי הערכה

- לכל מודל (לאחר Grid Search) בוצעו מאות הרצות סימולציה במגוון תרחישים.
- בוצעה השוואה בין DQN, PPO, A2C ו־Baseline (אקראי, היוריסטי).
- מדדי הערכה מרכזיים:
- Reward ממוצע וסטיית תקן (אימון/הערכה)
- זמן המתנה ממוצע
- אחוז מקרים דחופים שבוצעו בזמן
- ניצול חדרי ניתוח
- שיעור אפיזודות עם חריגה מהיום/שעות נוספות

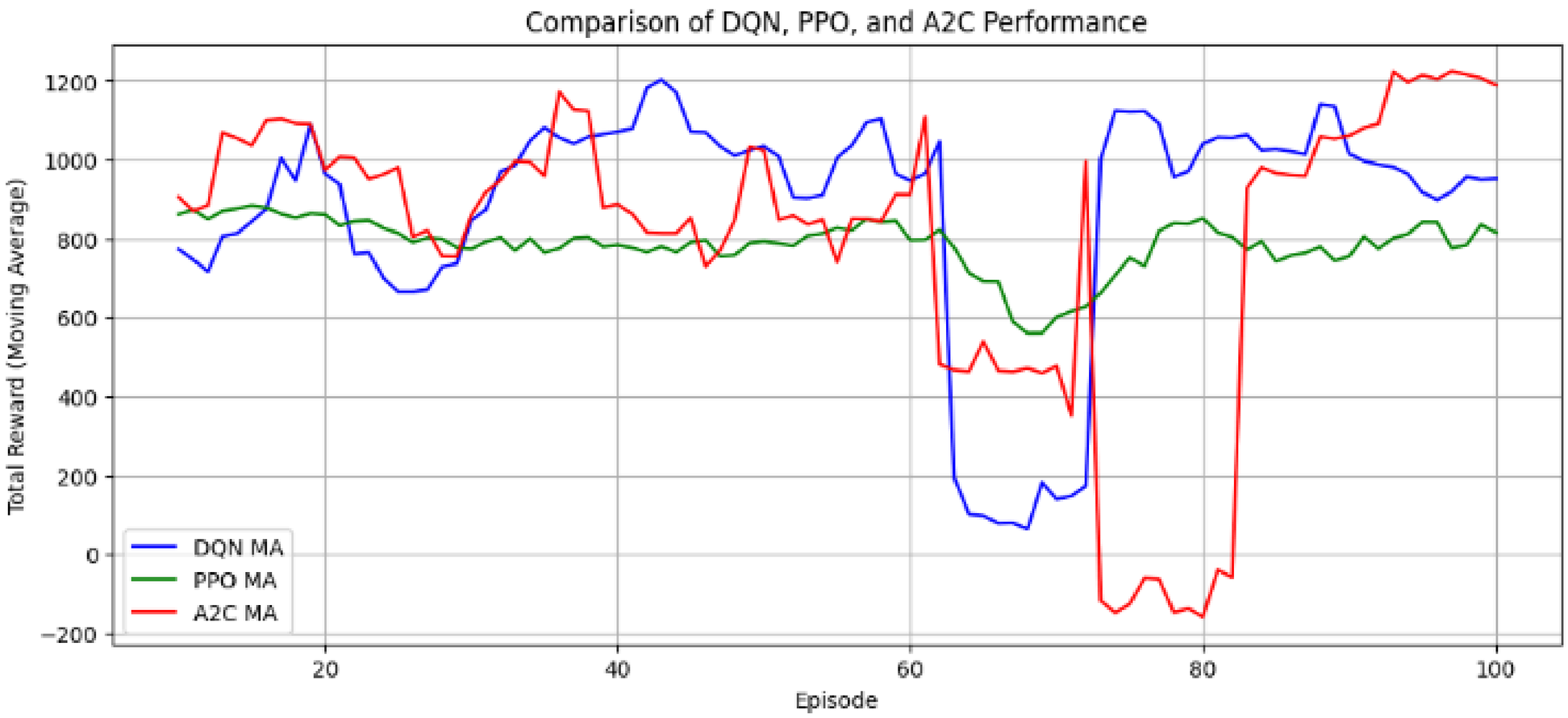
גרפים


DQN – Reward per Episode & Moving Average
Total Reward
Episode
Reward per episode
Moving Average (20)


A2C – Reward per Episode & Moving Average
Total Reward
Episode
Reward per episode (A2C)
Moving Average (20)


PPO – Reward per Episode & Moving Average
Total Reward
Episode
Reward per episode
Moving Average (20)

גרפים


Comparison of DQN, PPO, and A2C Performance
Total Reward (Moving Average)
1200
1000
800
600
400
200
0
−200
20
40
60
80
100
Episode
DQN MA
PPO MA
A2C MA

תוצאות עיקריות – השוואה בין המודלים

- PPO הראה יציבות גבוהה (סטיית תקן נמוכה, ללא "נפילות" קיצון), ממוצע תגמול טוב.
- DQN הגיע לשיאים גבוהים אך סבל מחוסר יציבות וקריסות באפיזודות מסוימות.
- A2C ממוצע דומה ל־DQN אך עם שונות קיצונית (Min Reward נמוך מאוד).
- Baseline (אקראי/היוריסטי): ביצועים נמוכים בהרבה בכל המדדים.

תוצאות עיקריות – השוואה בין המודלים

Urgent Served	% Overtime	Avg Wait	Min / Max	Std Reward	Avg Reward (Eval)	מודל
6.9	39%	12.4	-7229/2915	922	879	DQN
7.4	58%	15.2	-48/1033	185	790	PPO
7.2	52%	13.3	-10221/2668	1359	814	A2C

סיכום והמשך חקירה

סיכום: למידת חיזוקים משפרת משמעותית את ניהול התורים ותזמון הניתוחים.
זוהו אפיזודות "נפילה" חריגות – יש לחקור ולהבין את המקור.

הצעות לשיפור וחקירה:

- ניתוח מפורט של אפיזודות קיצון (תגמול שלילי במיוחד)
- בדיקה האם הנפילות קשורות לסביבה רנדומלית או למדיניות של הסוכן
- בחינת קשר בין פרמטרי הסביבה (עומס, דחיפות, מספר חדרים) לתוצאות קשות
- חקירת מצבים בהם הסוכן "נמנע" באופן גורף מפעולה מסוימת
- השוואה בין Grid Search שונה – האם יש אזורים "מסוכנים" במרחב הפרמטרים?
- בדיקה האם Reward shaping משפיע על התפלגות ה"נפילות"
- בדיקת רגישות של הסוכן לאפיזודות "קשות" (edge cases)

ביבליוגרפיה

u, H., Fang, Y., Chou, C.-A., Fard, N., & Luo, L. (2023). A reinforcement learning-based optimal control approach for managing an elective surgery backlog after pandemic disruption. Health Care Management Science, 26, 430–446