אוניברסיטת
אריאל
בשומרון

# קורס מבוא ללמידת חיזוקים

## מבוך פורמולה עם מכשולים

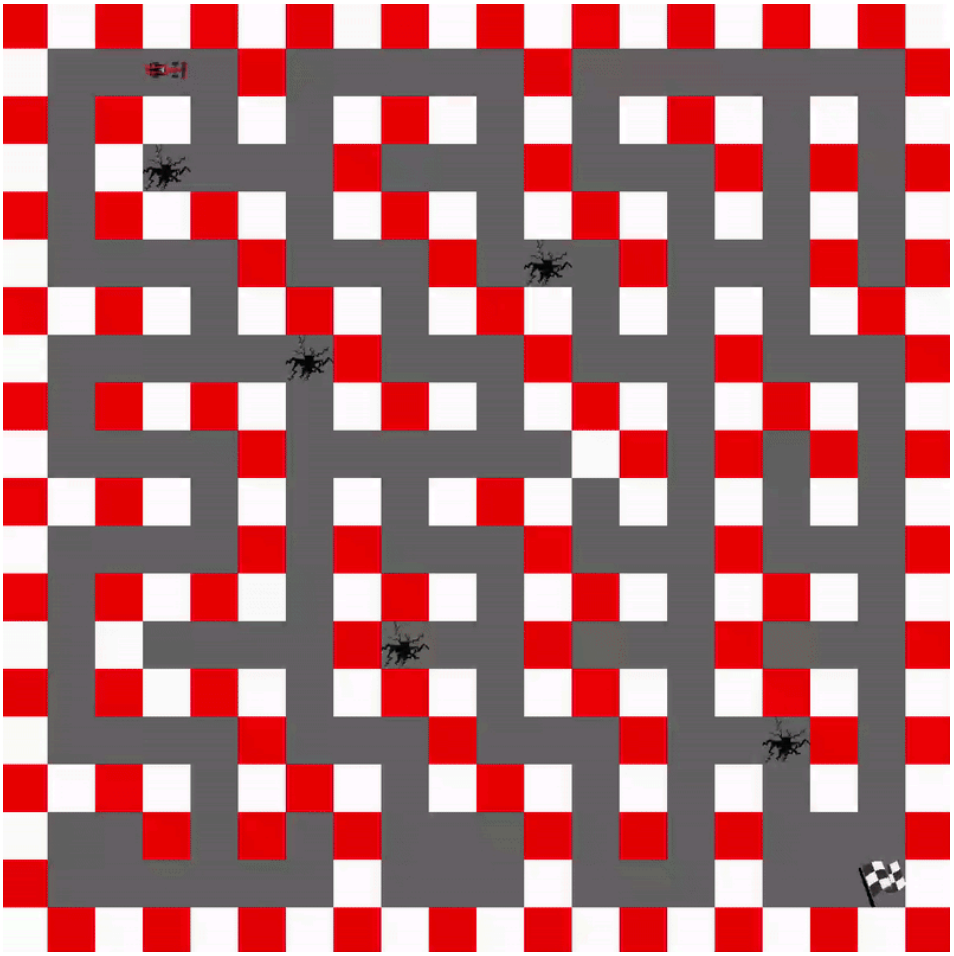

מגישים:

רון שטיינמץ 315578575

שחר תורג'מן 316223307

מנחה:

ד"ר טדי לזבניק

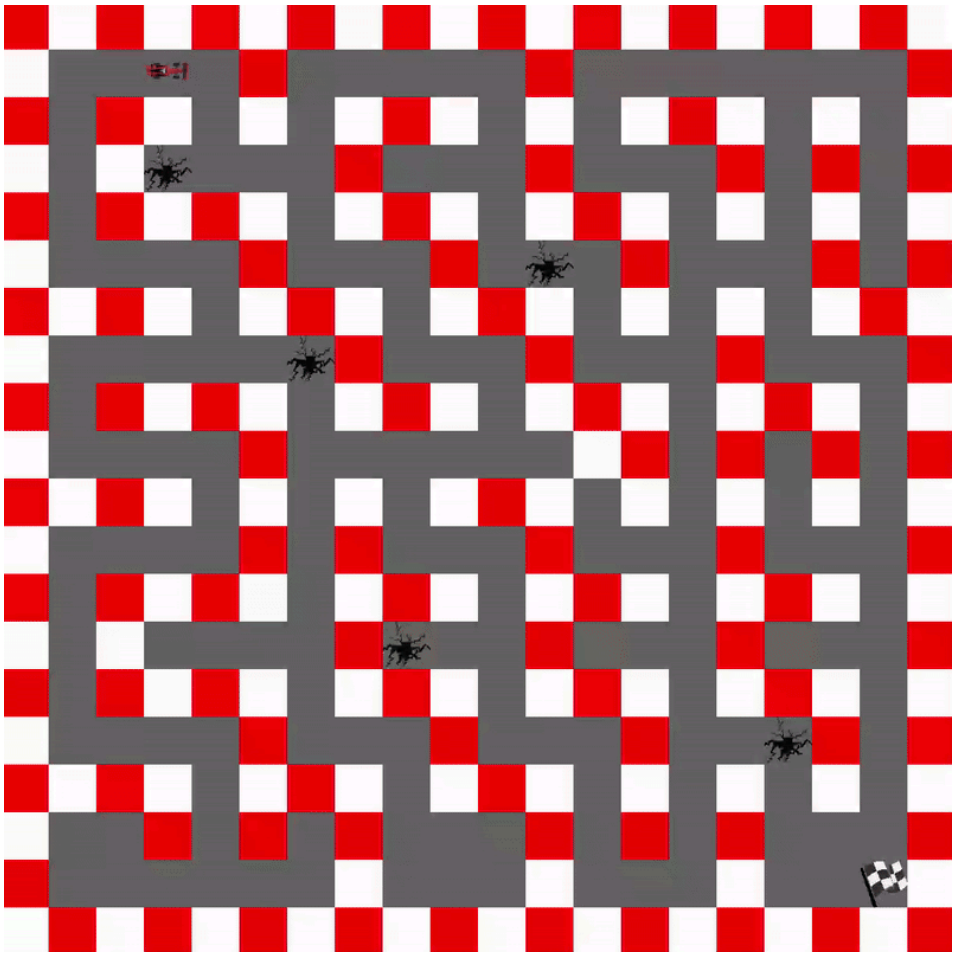



**סקירה כללית של הפרויקט**

במסגרת פרויקט זה, יישמנו אלגוריתם למידת חיזוקים מסוג Q-Learning, כדי ללמד סוכן רובוטי לנווט במבוך דו־ממדי, תוך התמודדות עם אתגרי תנועה, מכשולים, ומטרה קבועה.

מטרת הפרויקט הייתה כפולה: מצד אחד, להבין לעומק את אופן פעולתו של אלגוריתם Q-Learning ואת ההיגיון שמאחוריו ומצד שני, לבחון כיצד ניתן ליישם את האלגוריתם בסביבה גראפית אינטראקטיבית שפיתחנו בעזרת ספריית pygame.

במהלך הפרויקט, בנינו סביבה שבה הסוכן מתחיל כל אפיזודה במקום קבוע (צד שמאל למעלה), כאשר עליו למצוא את הדרך הקצרה והבטוחה ביותר לנקודת הסיום, שהיא יעד קבוע בסביבה(צד ימין למטה). לאורך הזמן, בעזרת למידה חוזרת, אמור הסוכן לשפר את בחירת הפעולות שלו על מנת להגיע ליעד בצורה יעילה יותר. סביבת המבוך כוללת קירות שמפריעים לתנועה החופשית ובורות במסלולים (שמכשילים את הסוכן במידה והגיע לשם), מה שמאלץ את הסוכן ללמוד איך לפעול בצורה חכמה ולא לבחור פעולות בצורה אקראית או פשוטה.



**הבעיה/מטרה**

הסוכן (מכונית המרוץ במקרה שלנו) נמצא בסביבת המבוך שיצרנו בספריית pygame, הסוכן צריך למצוא את הדרך היעילה ביותר מנקודת הפתיחה שלו בצד העליון השמאלי של המבוך ,לצד התחתון ביותר בצד הימני של המבוך (דגל משבצות).
בדרך ישנן אופציות נרחבות לאיפה שהסוכן יכול ללכת ,הקירות (ריבועים בצבע לבן ואדום) מגבילות את הסוכן והוא לא יכול לעבור דרכם, לא כל הדרכים מובילות ליעד ,חלקן אפילו כוללות בורות כדי לאתגר את הסוכן ולכן אם הוא מגיע לבור,הסוכן נפסל ומתחיל מחדש את המשחק.

- **מרחב המצבים**:

כל מצב במערכת מייצג את מיקומו הנוכחי של הסוכן בתוך המבוך, אשר מיוצג כמטריצה דו ממדית של תאים. כל תא במטריצה מהווה מצב (state) ייחודי, והזהות של כל מצב נקבעת על פי מיקומו בקואורדינטות של המפה. לצורך ייעול תהליך הסיווג, קואורדינטות אלו הומרו למספרים סידוריים ייחודיים: התא שבו הסוכן מתחיל מוגדר כ־State מספר 0, ואם הסוכן נוקט בפעולה של צעד אחד ימינה, הוא עובר ל־State מספר 1, וכן הלאה. מבנה זה מאפשר ייצוג חד משמעי של כל מצב באמצעות אינדקס מספרי. בסך הכול, קיימים 189 מצבים שבהם הסוכן עשוי להימצא במהלך תנועתו, כולם תואמים לתאים האפורים במפה, אשר מייצגים אזורים פתוחים ונגישים במבוך.

- **מרחב הפעולות :**

הסוכן פועל במרחב המוגדר כרשת דו ממדית, והוא מסוגל לבחור בין ארבע פעולות תנועה בסיסיות: למעלה, למטה, ימינה ושמאלה. עם זאת, חשוב להדגיש כי לא בכל סטייט קיימת אפשרות לבצע את כל ארבע הפעולות. מגבלות אלו נובעות ממבנה הסביבה. כך למשל, אם הסוכן ממוקם בצמוד לקיר בצדו הימני, האפשרות לנוע ימינה לא תהיה זמינה עבורו באותו מצב. מגבלה זו מדגישה את חשיבות ההתאמה בין הפעולה לבין המצב הנוכחי, ומהווה אתגר נוסף בתהליך הלמידה של הסוכן.

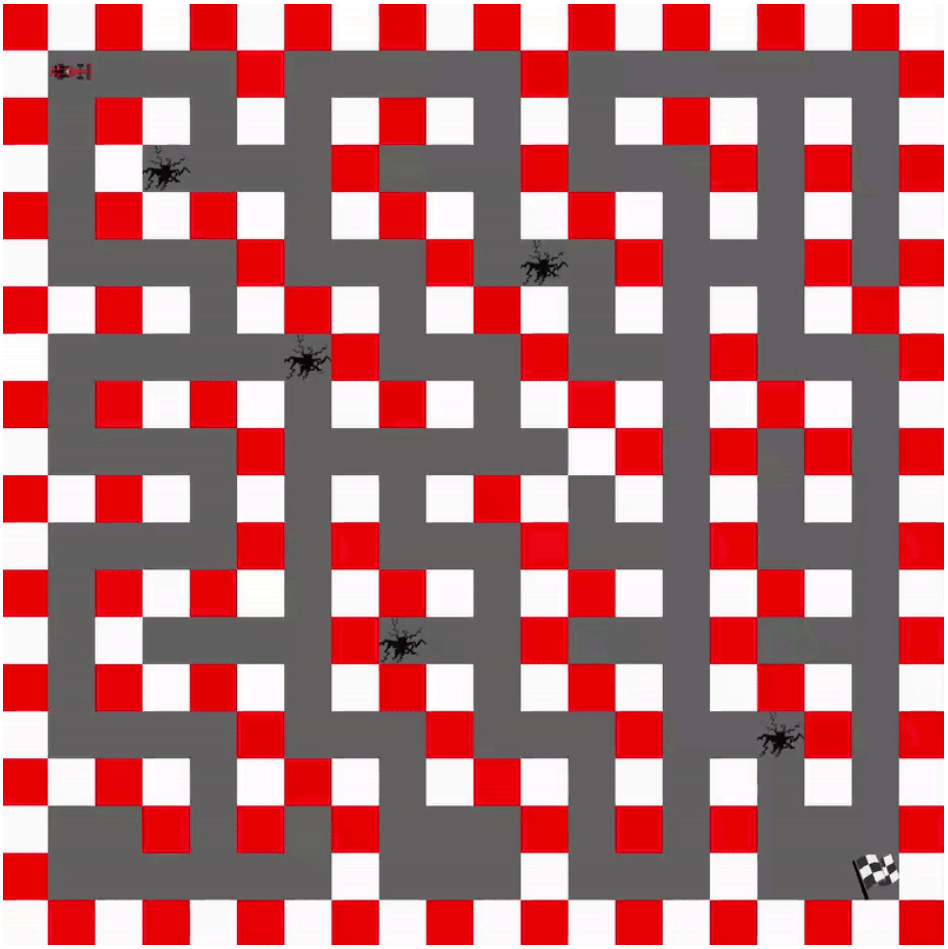
- **פונקציית התגמול**

על כל צעד שהסוכן מבצע במהלך תנועתו במבוך, הוא מקבל תגמול של 0.1-. מטרתו של תגמול שלילי זה היא לעודד את הסוכן לבחור במסלול הקצר והיעיל ביותר האפשרי אל היעד, על מנת לצמצם את מספר הצעדים הננקטים. בנוסף, במרחב המבוך פזורים בורות תאים

המוגדרים כמצבים מסוכנים. כאשר הסוכן מגיע לאחד מהבורות או מגיע למספר המקסימלי לצעדים שהגדרנו לו (350), הפרק מסתיים באופן מיידי והוא מקבל תגמול של 10-. תגמול שלילי גבוה זה נועד להרתיע את הסוכן מלהיכנס לאזורים אלה, וללמד אותו להימנע מהם לאורך הלמידה. מנגד, כאשר הסוכן מצליח להגיע ליעד הסופי, הוא מתוגמל ב +10, תגמול חיובי משמעותי שמטרתו לחזק את הקישור בין פעולות שהובילו להצלחה לבין השגת היעד בפועל. באמצעות מערכת תגמולים זו, הסוכן מפתח מדיניות אופטימלית המעדיפה מסלולים קצרים, בטוחים ומובילים ליעד.

- **דינמיקת הסביבה**

הסביבה שבה פועל הסוכן היא דיסקרטית, כלומר הן מרחב המצבים והן קבוצת הפעולות האפשריות מוגדרים מראש ובעלי גודל סופי. כל מצב מייצג מיקום מסוים במבוך, וכל פעולה מביאה את הסוכן, בתנאים מסוימים, למצב חדש. מעבר בין מצבים מתרחש בצורה דטרמיניסטית. כאשר הסוכן מבצע פעולה חוקית (שאינה חסומה על ידי קיר או בור), הוא עובר למצב החדש בהתאם לאופי הפעולה.



*דוגמה לסוכן שנכשל במשימה ונוסע על בור(נפסל במשחק)

**מתודולוגיה**

האלגוריתם שנבחר לפרויקט הוא Q-Learning, מהסיבה העיקרית שהסביבה שלנו היא דיסקרטית ונתנת לפתירה ע"י האלגוריתם הטבלאי Q-Learning .

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

**תהליך הלמידה ומדיניות הפעולה**

בתחילת האימון, הסוכן מתחיל עם טבלת־Q מאופסת, כלומר כל ערכי ה־Q (עבור כל צירוף של מצב ופעולה) שווים לאפס. במהלך האינטראקציה עם הסביבה, הסוכן מעדכן את ערכי הטבלה בהתאם לתגמולים שהוא מקבל, תוך שימוש בנוסחת עדכון ה־Q.
עם הזמן, ככל שמספר האפיזודות גדל, טבלת ה־Q נעשית מדויקת יותר, ומשקפת בצורה טובה יותר את המדיניות האופטימלית כלומר, את הפעולה הטובה ביותר שיש לבצע בכל מצב נתון.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | -4.377076 | -4.368678 | -4.380066 | 0.618111 |
| 1 | -4.320724 | -4.324702 | -4.357832 | 0.725365 |
| 2 | -4.283813 | -4.256145 | -4.303401 | 0.833702 |
| 3 | -4.210588 | 0.943133 | -4.209938 | -4.215267 |
| 4 | -3.969681 | -3.978949 | -3.953542 | -3.933744 |
| ... | ... | ... | ... | ... |
| 177 | -3.163051 | -3.161088 | -3.159411 | -3.162710 |
| 178 | -3.104758 | -3.106628 | -3.159428 | -3.161881 |
| 179 | -3.161066 | -3.099975 | -3.112055 | -3.104639 |
| 180 | -2.510370 | -2.444788 | -2.451586 | 4.646067 |
| 181 | -2.368997 | -2.365075 | -2.430797 | 4.794007 |

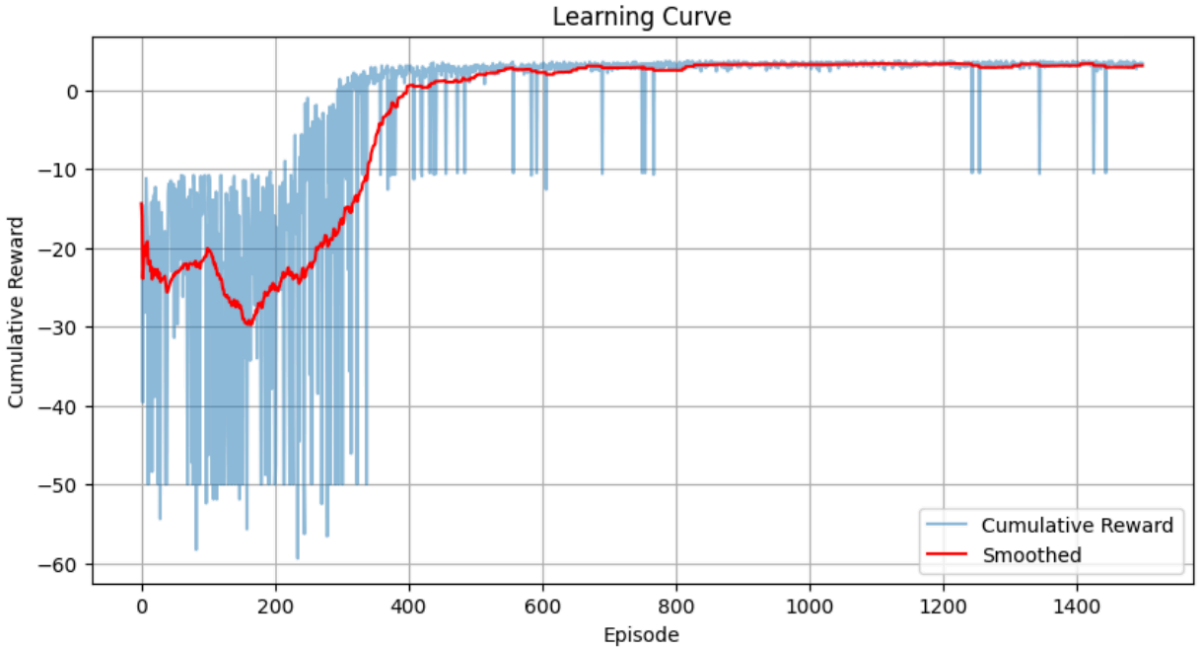
*טבלת ה-Q האופטימלית

כדי לאפשר למידה אפקטיבית, נעשה שימוש במדיניות חמדנית עם הסתברות אפסילון לחקירה. בתחילת הלמידה, ערך האפסילון נקבע ל־1, כך שהסוכן בוחר פעולות באקראי (exploration), דבר המאפשר לו לחקור את הסביבה ולגלות את התגמולים האפשריים. בהמשך, ערך האפסילון יורד בהדרגה לפי נוסחת דעיכה קבועה, עד שהוא מתייצב על ערך סופי של 0.05. בשלב זה, הסוכן ממעט לחקור ומרבה לנצל (exploitation) את הידע שצבר כלומר, לבחור את הפעולה עם ערך ה־Q הגבוה ביותר בכל מצב.

לאורך כל תהליך הלמידה, טבלת ה־Q מתעדכנת באופן רציף. ככל שהסוכן ממשיך לפעול בסביבה, הטבלה מתכנסת לערכים יציבים, והמדיניות שנגזרת ממנה מובילה להשגת תגמולים מצטברים מרביים דבר המעיד על כך שהסוכן פועל בצורה אופטימלית בסביבה הנתונה.

**היפר-פרמטרים ההתחלתיים שנבחרו לניסוי היו:**

- **מספר אפיזודות:** 1500
- **מספר צעדים מקסימלי:** 500
- **קצב למידה :** 0.7
- **גאמא :** 0.95
- **אפסילון התחלתי:** 1.0
- **אפסילון מינימלי:** 0.05
- **קצב דעיכה של אפסילון** 0.0005



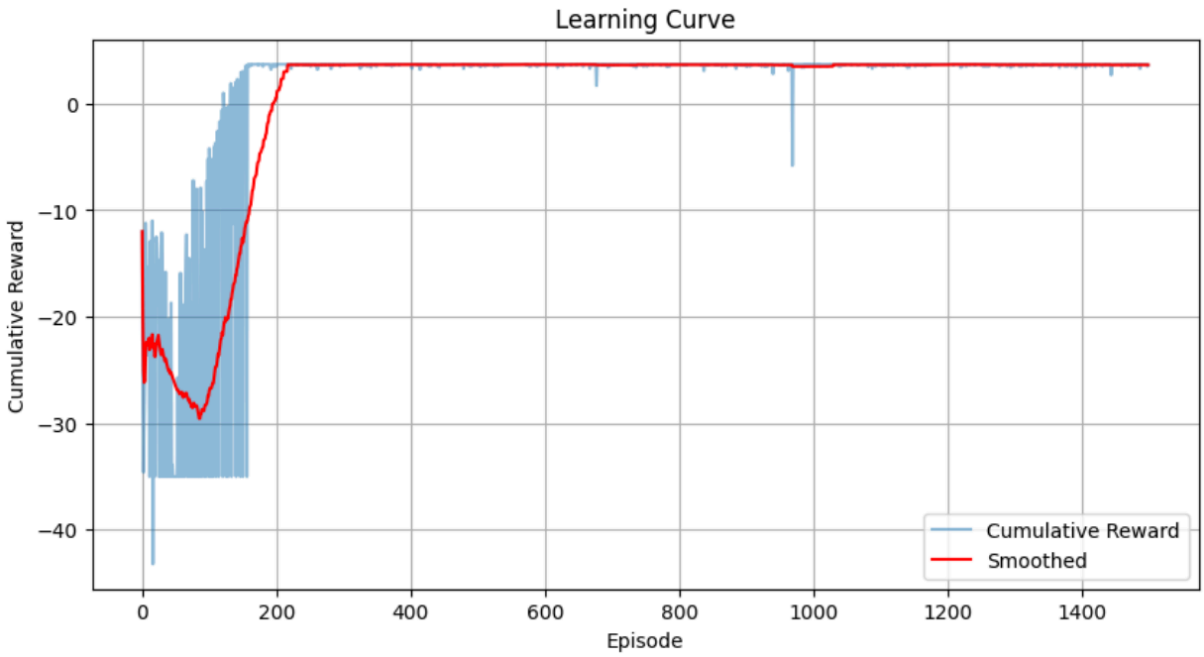

בתחילת שלב האימון, הרצנו את הסוכן לאורך 200 אפיזודות, אך כפי שניתן לראות בתרשים מספר 2, בשלב זה הוא עדיין לא הצליח להתייצב מבחינת סך התגמולים שהשיג בכל אפיזודה. רק לאחר כ־400-500 אפיזודות החלה להופיע מגמת שיפור יציבה, אשר העידה על למידה אפקטיבית של הסוכן. קצב הלמידה שנבחר היה מאוזן, לא מהיר מדי ולא איטי מדי ולכן לא נדרש לבצע בו שינויים מהותיים. עם זאת, כחלק מתהליך החקירה, ניסינו לבחון גם שילובים נוספים של היפר פרמטרים, במטרה לוודא את אופטימליות הביצועים. בנוסף, הוספנו מגבלה של מספר צעדים מקסימלי לכל אפיזודה, כדי למנוע מהסוכן להיתקע בלולאה אינסופית. בתחילה נקבעה מגבלה של 150 צעדים, אך נמצא כי היא מקשה על הסוכן להשלים את המשימה. לכן, הגדלנו את המגבלה ל־500 צעדים, מה שהוביל לשיפור בולט ביכולת שלו ללמוד ולהתמודד עם הסביבה בצורה יעילה יותר.

בחנו היפר פרמטרים נוספים כדי לראות אם אנחנו יכולים לגרום לסוכן ללמוד בצורה יותר טובה ויעילה ולהתייצב הרבה יותר מהר.

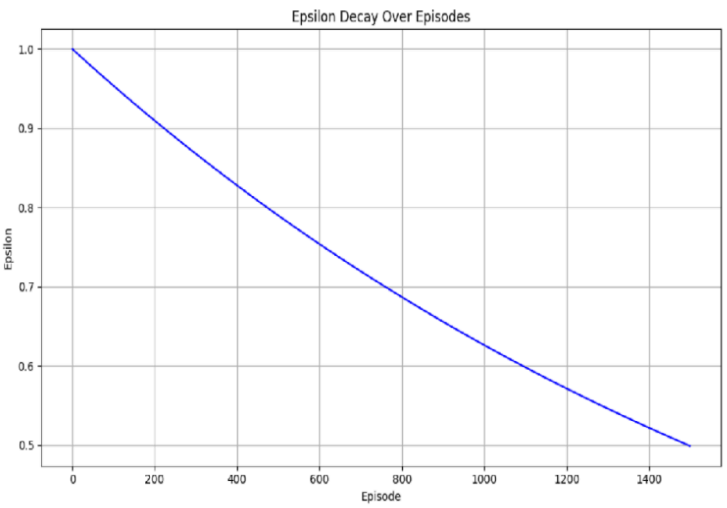
**הפעולות הראשונות שביצענו זה להגדיר:**

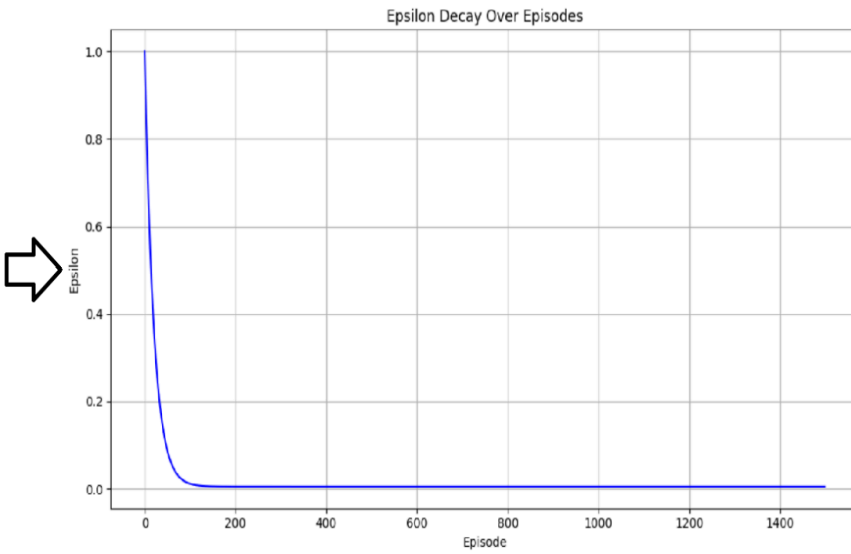
- **מספר אפיזודות:** 1500
- **מספר צעדים מקסימלי:** 350
- **קצב למידה :** 0.9
- **גאמא :** 0.95
- **אפסילון התחלתי:** 1.0
- **אפסילון מינימלי:** 0.005
- **קצב דעיכה של אפסילון** 0.05

**חשוב לציין שאת האפיזודות בסוף הניסוי וטעייה שלנו הורדנו לכמות שמתייצבת , הגדרנו 1500 בעיקר בשביל ויזואליזציה .**




כפי שניתן לראות, ביצועי הסוכן השתפרו משמעותית בהשוואה להגדרות ההתחלתיות של ההיפר פרמטרים. כבר באזור האפיזודות 180–200 הסוכן מצליח להתייצב ולהשיג תוצאות טובות באופן עקבי. השיפור המרכזי נבע מהשינויים שביצענו בפרמטרים: העלאת קצב הלמידה, הפחתת ערך האפסילון המינימלי לרמה נמוכה יותר, והגברת קצב הדעיכה של האפסילון. שילוב זה אפשר לסוכן לחקור את הסביבה בשלבים הראשונים, אך בהמשך לעבור בצורה מדורגת למצב של ניצול הידע שנצבר, ובכך לשפר את היעילות ואת הביצועים שלו.

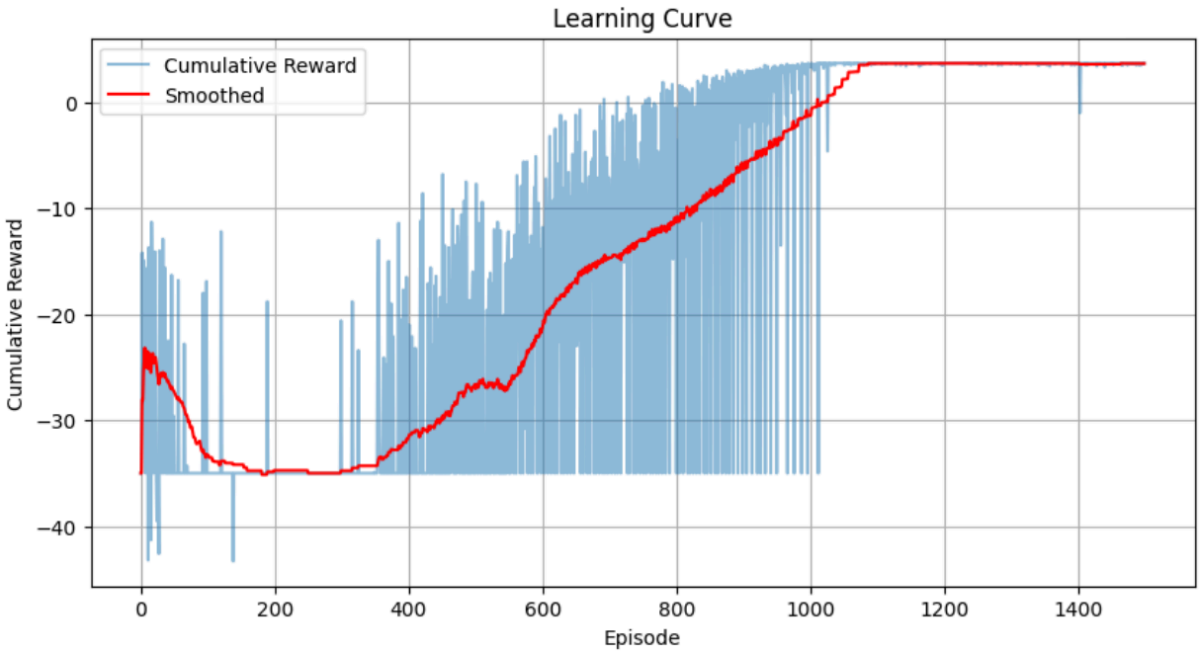


**רצינו לראות גם מה קורה בקצב למידה איטי ולראות איך זה משפיע על הסוכן**

**הפעולות הראשונות שביצענו זה להגדיר:**

- **מספר אפיזודות:** 1500
- **מספר צעדים מקסימלי:** 350
- **קצב למידה :** 0.1
- **גאמא :** 0.95
- **אפסילון התחלתי:** 1.0
- **אפסילון מינימלי:** 0.005
- **קצב דעיכה של אפסילון** 0.05

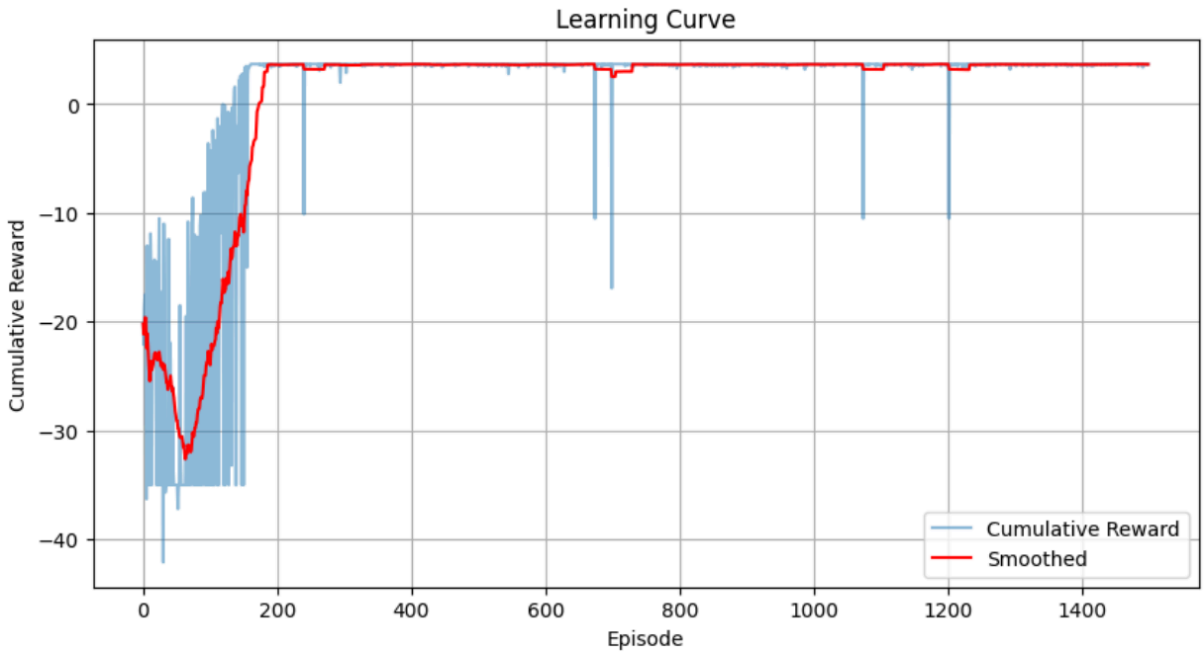



אז ניתן לראות שבאופן הגיוני, הסוכן מתייצב לאחר כ1000 אפיזודות מאחר והקצב למידה שלו מאוד איטי כפי שנקבע ולכן נשאר על קצב למידה של 0.9 .

כעת נבחן ערכים שונים של גאמא :

- **מספר אפיזודות:** 1500
- **מספר צעדים מקסימלי:** 350

- **קצב למידה :** 0.9
- **גאמא :** 0.99
- **אפסילון התחלתי:** 1.0
- **אפסילון מינימלי:** 0.005
- **קצב דעיכה של אפסילון** 0.05




הגדרת ערך גאמא ל-0.99 מבטאת העדפה חזקה לתגמולים עתידיים על פני תגמולים מיידיים. גישה זו מעודדת את הסוכן לפעול מתוך ראייה ארוכת טווח, ולא לבחור רק בצעדים שמניבים תגמול מיידי. כפי שנראה בתוצאות, ביצועי הסוכן תחת פרמטר זה היו טובים במיוחד, ולכן בחרנו לשמר את ההגדרה הזו. השילוב בין גאמא גבוהה לשאר ההיפר פרמטרים שהותאמו, סייע לנו לפתור את המשימה בצורה אופטימלית הן מבחינת הביצועים בפועל והן מבחינת יציבות הלמידה.

**לכן פרמטרים אופטימליים סופיים :**

- **מספר אפיזודות:** 200
- **מספר צעדים מקסימלי:** 350
- **קצב למידה :** 0.9
- **גאמא :** 0.99
- **אפסילון התחלתי:** 1.0
- **אפסילון מינימלי:** 0.005
- **קצב דעיכה של אפסילון** 0.05

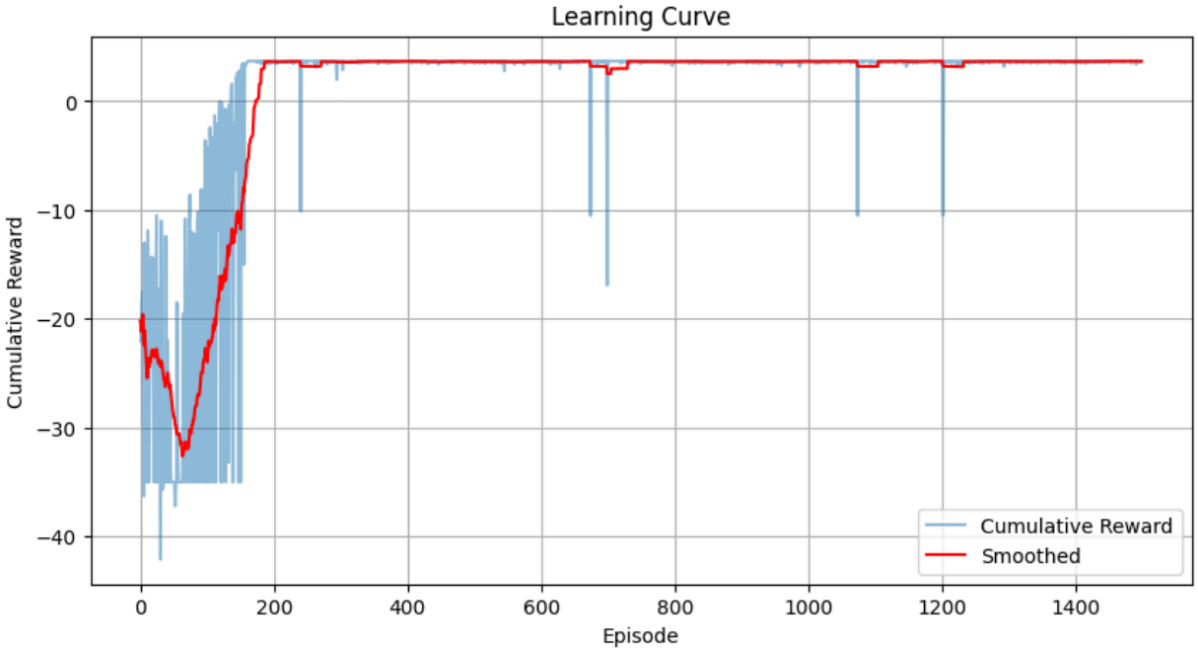
**תוצאות**

לאחר אימון של 200 אפיזודות, ניתן היה לראות שיפור משמעותי בביצועי הסוכן. בתחילת הדרך, כאשר טבלת ה־Q עדיין הייתה ריקה או כמעט ריקה, הסוכן ביצע פעולות אקראיות, נתקל בקירות, חזר על אותם מסלולים ואף לעיתים לא הצליח להגיע ליעד. ככל שהתהליך התקדם, הביצועים הלכו והשתפרו הסוכן התחיל לפתח מדיניות פעולה שמביאה אותו ליעד במספר צעדים נמוך יותר, תוך הימנעות ממכשולים ידועים.

**תצפיות כמותיות:**

אחת הדרכים למדוד את התקדמות הסוכן היא מעקב אחר התגמול המצטבר בכל אפיזודה. בתחילה, התגמולים היו שליליים מאוד לעיתים אף מתחת ל30-, מה שמעיד על פעולות לא יעילות, התקלות בבורות ואי הגעה ליעד. בהמשך האימון, התגמולים השתפרו באופן ניכר, והחלו להופיע אפיזודות עם תגמול חיובי סימן לכך שהסוכן הגיע ליעד ואף עשה זאת בדרך יעילה .

ניתן לראות בתרשים שבהתחלה התגמולים שליליים מאוד ואז לאט לאט הם עולים ומתייצבים מה שמעיד על למידה של הסוכנים למספרים צעדים אופטימלי ומספר צעדים יעילים.

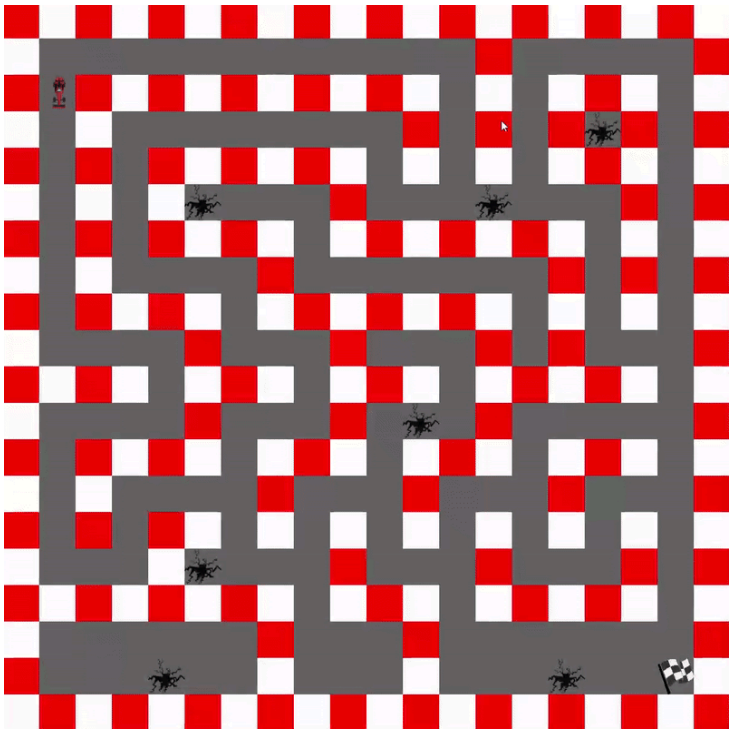


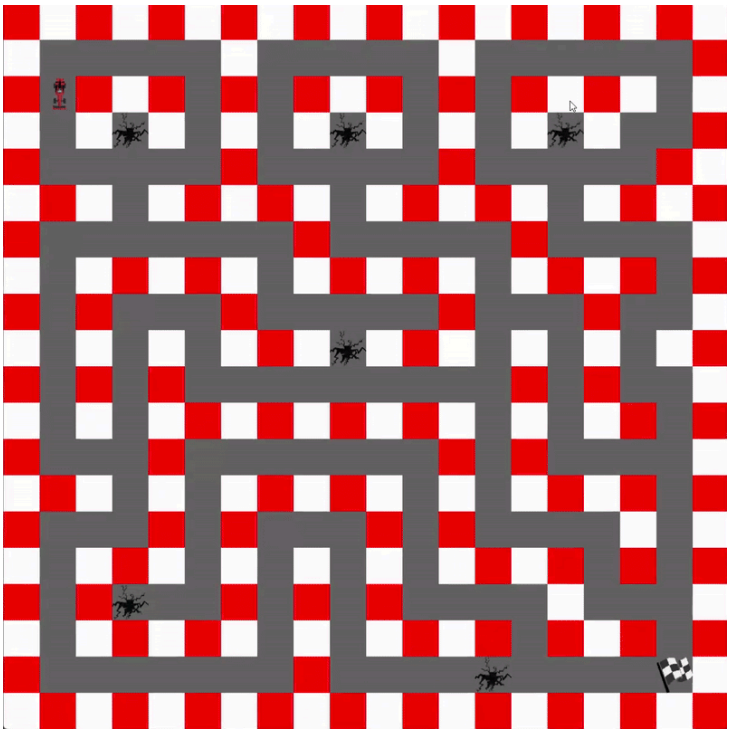
**תצפיות איכותיות**: במהלך הניסויים שערכנו, ניתן היה להבחין בבירור כי לאחר מספר אפיזודות למידה, הסוכן הצליח לפתור את המבוך בצורה עקבית ויעילה. התנהגותו הפכה ממקרית ומבולבלת בתחילת הדרך, לאסטרטגית וממוקדת עם התקדמות האימון. הסוכן פיתח מדיניות פעולה המעדיפה מסלולים קצרים ככל האפשר, תוך הימנעות מהגעה לבורות והגעה מהירה אל היעד. התנהגות זו מעידה על כך שהסוכן הצליח להפנים את חוקי הסביבה ואת מבנה התגמולים, ולפעול בהתאם לאסטרטגיה שממקסמת את התגמול המצטבר. ניתן לומר כי הלמידה אכן הייתה אפקטיבית, והביאה להיווצרות של דפוס פעולה רציונלי ויעיל.

**דיון**

התוצאות שהתקבלו מצביעות על כך שהסוכן מתייצב לאחר כ-200 אפיזודות, ומצליח לפתור את המבוך באופן יעיל ומהיר. ניתן לראות שהתגמול המצטבר מתייצב סביב 4–5 נקודות, דבר שמעיד על הצלחה עקבית בביצוע המשימה. במהלך תהליך האימון זיהינו כי ההגדרות ההתחלתיות של ההיפר-פרמטרים לא היו אופטימליות, שכן הסוכן דרש מספר רב של אפיזודות כדי להגיע ליציבות. בעקבות כך, אימצנו גישה של ניסוי וטעייה ובחנו שילובים שונים של פרמטרים במטרה לשפר את ביצועי הסוכן.

המעבר לפרמטרים מותאמים היטב הביא לשיפור ניכר: הסוכן למד במהירות גבוהה יותר, גילה מדיניות פעולה יעילה יותר, והציג רמת ביצועים גבוהה ועקבית. לאחר תהליך הלמידה, הסוכן לא רק פותר את המבוך בהצלחה בכל פעם שהוא מופעל, אלא אף מסוגל להתמודד עם מבוכים חדשים, תוך שימוש בידע הכללי שצבר במהלך האימון.





**מסקנות**

הפרויקט איפשר לנו להעמיק את ההבנה של עקרונות למידת החיזוקים, הן מההיבט התיאורטי והן מההיבט המעשי. חווינו באופן ישיר כיצד בחירה לא מדויקת של היפר פרמטרים עלולה לפגוע בביצועי המודל, מה שהדגיש את החשיבות הרבה של תהליך ניסוי וטעייה ככלי אופטימיזציה.
הבחירה באלגוריתם Q-Learning התבררה כמתאימה במיוחד לסביבה הדיסקרטית שיצרנו, ואפשרה לנו לעקוב אחר תהליך שיפור מדיניות הפעולה של הסוכן לאורך הזמן. למדנו כיצד הסוכן לומד להימנע מטעויות חוזרות ולגבש אסטרטגיה יעילה דרך חיזוקים בלבד.
ההישג המרכזי של הפרויקט היה היכולת להכשיר סוכן שיכול לפעול באופן מושכל, ללא הוראות מפורשות אלא באמצעות תגמולים בלבד. בנוסף, הבנו את חשיבות האיזון בין חקירה לניצול, ואת התרומה של מבנה הסביבה ליכולת הלמידה והכללה של הסוכן.