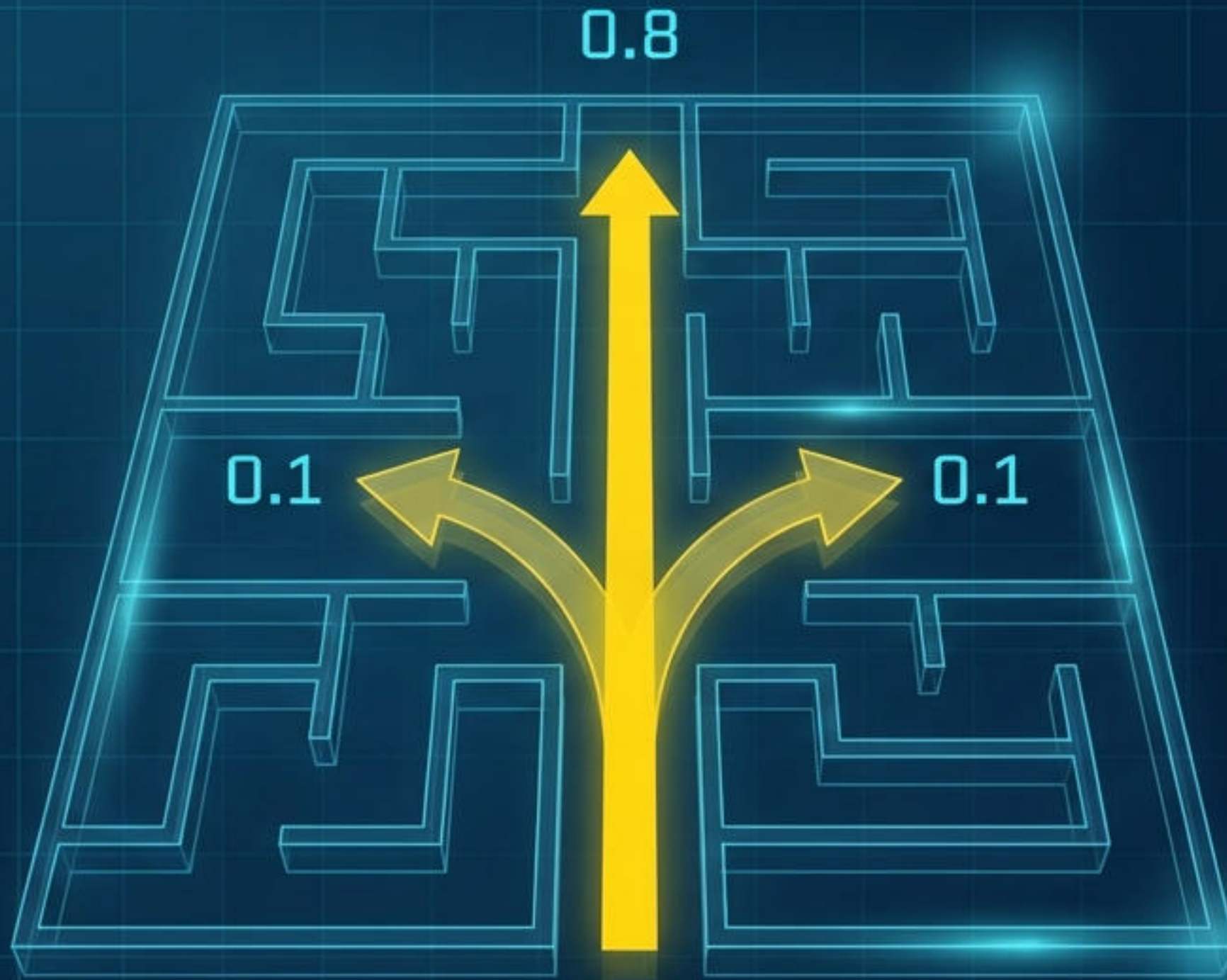


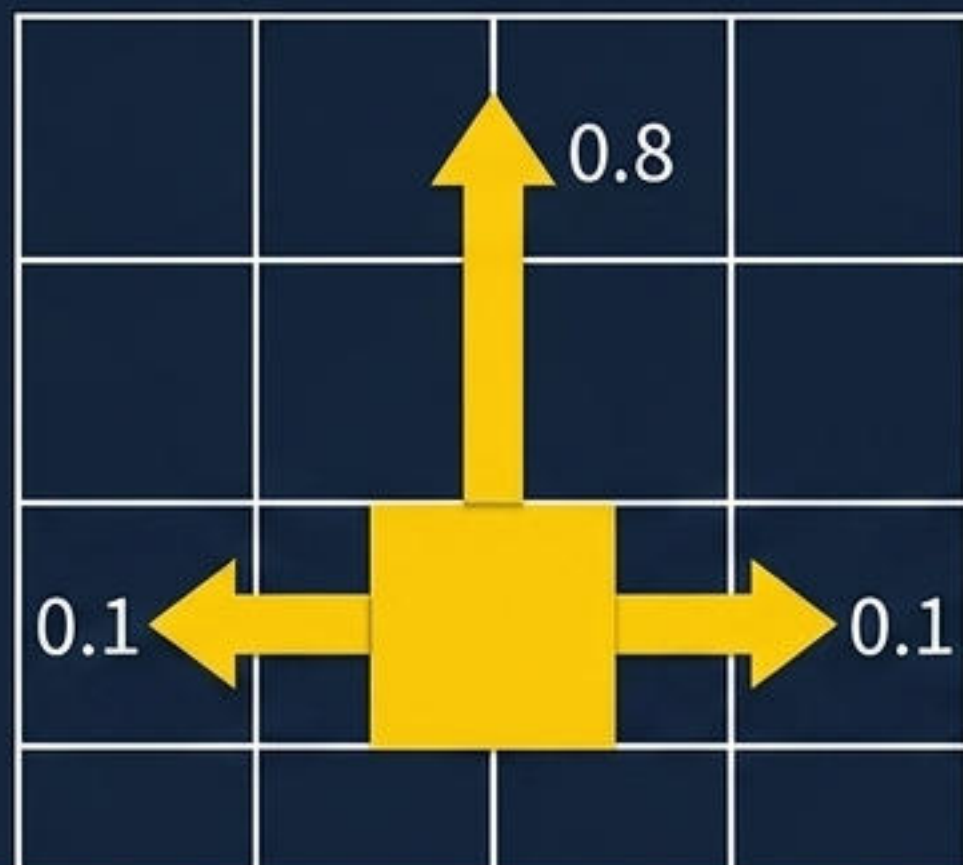


ניווט תחת חוסר ודאות: ארכיטקטורה של החלטות

- מבוא לתהליכי החלטה מרקוביים (MDP).
- המדריך המלא: כיצד סוכנים חכמים מחשבים את המסלול האופטימלי בעולם בלתי צפוי.

הבעיה הבסיסית: העולם אינו דטרמיניסטי

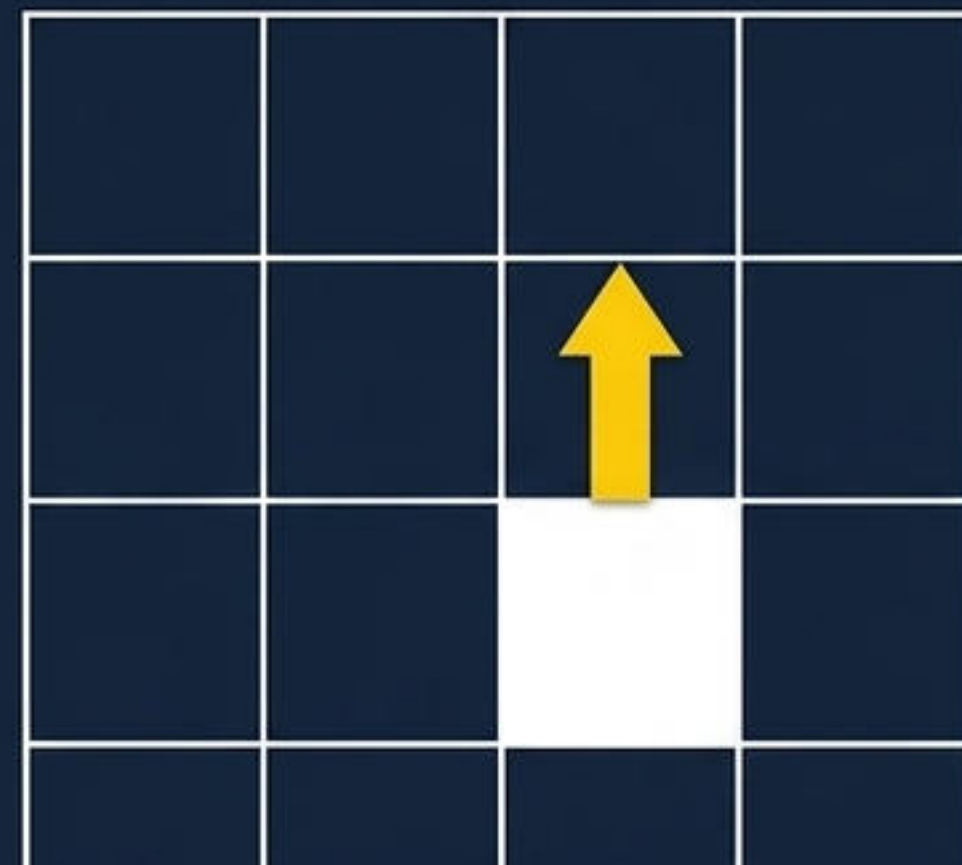
סביבה סטוכסטית



המסקנה: בעולם סטוכסטי, תכנון מסלול בודד יקרוס מיד עם הסטייה הראשונה.

אנו זקוקים לאסטרטגיה מלאה שתאמר לסוכן מה לשות בכל מצב אליו הוא עלול להיקלע.

סביבה דטרמיניסטית



הפעולה מתבצעת בדיוק כפי שתוכננה (100%) ניתן לתכנן מסלול יחיד מראש (Sequence).

האנטומיה של תהליך החלטה מרקובי (MDP)

תהליך מרקובי מוגדר על ידי הרביעייה: $\langle S, A, P, R \rangle$
הנחת מרקוב: העתיד תלוי רק בהווה, ולא בהיסטוריית העבר.



States (S)

אוסף המצבים האפשריים
בהם הסוכן יכול
להימצא בעולם.



Actions (A)

הפעולות הזמינות לסוכן
לבחירה בכל מצב S .

$$P(s'|s, a)$$

Model (P / M)

פונקציית המעברים.
פונקציית המעברים. ההסתברות להגיע
למצב החדש s' אם ביצענו פעולה a
במצב s (ה'פיזיקה' של העולם).



Reward (R)

התגמול $R(s)$ המיידית
שהסוכן מקבל עם
ההגעה למצב.

קו האופק: התמודדות עם רצפים אינסופיים

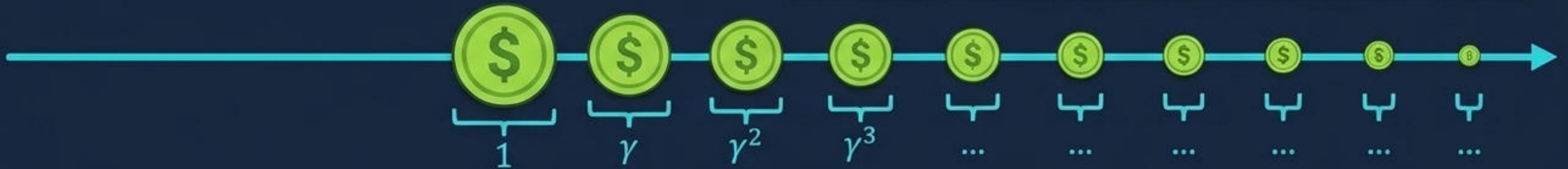
$$R(s_0) + R(s_1) + R(s_2) \dots = \infty$$

סכום רגיל מתבדר לאינסוף כשאין מצב סיום.



$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots$$

שימוש במקדם היוון מבטיח סכום סופי ובר-השוואה.

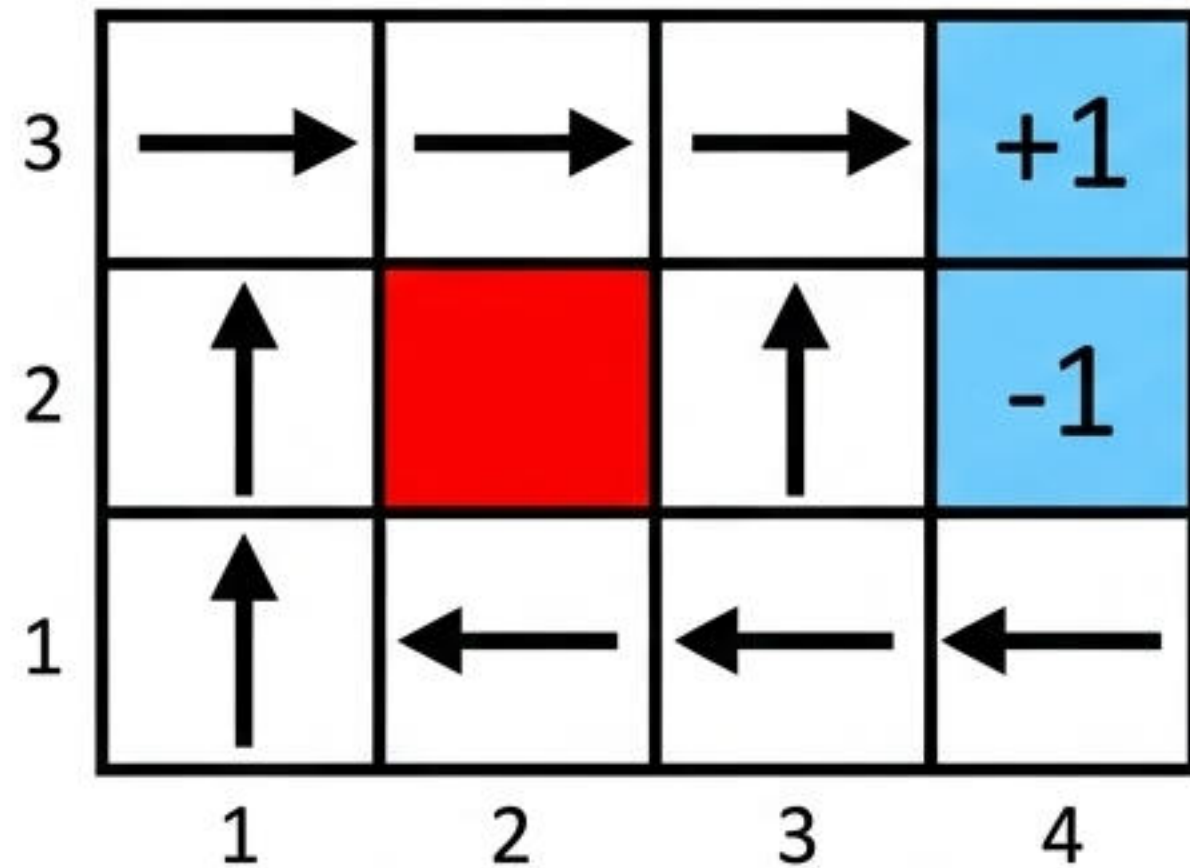


מקדם ההיוון (γ - Discount Factor): ערך בין 0 ל-1 ($0 < \gamma \leq 1$).

קובע את ההעדפה של הסוכן להווה לעומת העתיד.
ככל ש- γ קרוב ל-0, הסוכן אופורטוניסט (מחפש רווח מיידי). ככל שהוא קרוב ל-1, הסוכן מתכנן לטווח ארוך.

המטרה: להבחין בין תועלת למדיניות

Maximum Expected Utility: עקרון העל של קבלת החלטות.
אנו מחפשים את המדיניות האופטימלית π^* שתבטיח את התועלת הצפויה המקסימלית בכל מצב נתון.



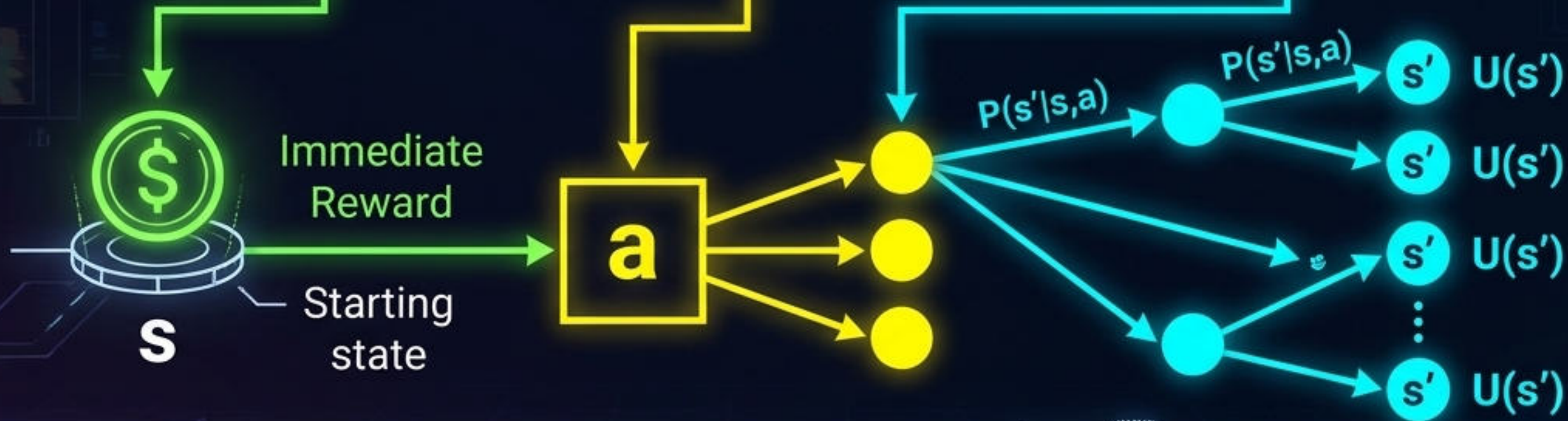
מדיניות (π) : החץ. פונקציה
הממפה כל מצב לפעולה
ספציפית: $\pi(s) = a$

3	0.812	0.762	0.660	+1
2	0.719		0.513	-1
1	0.655	0.556	0.372	0.104
	1	2	3	4

תועלת מצב $(U(s))$: המספר. הערך
(Desirability) של הימצאות במצב מסוים.
מייצג את תוחלת התגמול העתידי.

הלב הפועם: פירוק משוואת בלמן (Bellman Equation)

$$U(s) = R(s) + \gamma \max_a \sum P(s'|s, a) U(s')$$



האינטואיציה: התועלת האמיתית של מצב שווה למה שמקבלים עכשיו, פלוס מה שמצפים לקבל מהצעד הבא והלאה (בהנחה שפועלים באופן אופטימלי).

כיצד פותרים? שתי אסטרטגיות חישוב

שתי השיטות נועדו לפתור את מערכת משוואות בלמן הלא-לינאריות (בגלל פונקציית ה-MAX), אך ניגשות לבעיה מזווית שונה.



אלגוריתם במיקוד: Value Iteration



אלגוריתם במיקוד: Policy Iteration

שלב 1: הערכה (Policy Evaluation)

בהינתן המדיניות הנוכחית π_i ,
מחשבים את תועלות המצבים U .
הסוד המתמטי: ללא מפעיל ה-MAX,
משוואות בלמן הופכות למערכת
משוואות לינאריות פשוטות שניתן
לפתור ישירות.

שלב 2: שיפור (Policy Improvement)

עוברים על כל מצב ובודקים: האם יש
פעולה a שונה מזו שמוגדרת
במדיניות הנוכחית שמניבה תוחלת
גבוהה יותר? ($\arg\max$)
אם כן - מעדכנים את המדיניות
וחוזרים לשלב 1.

היתרון: לרוב מתכנס במספר קטן בהרבה של איטרציות לעומת Value Iteration,
כיוון שהמדיניות (בחירת פעולה מתוך סט סופי) מתייצבת הרבה לפני שהערכים
המספריים הרציפים מתכנסים לחלוטין.

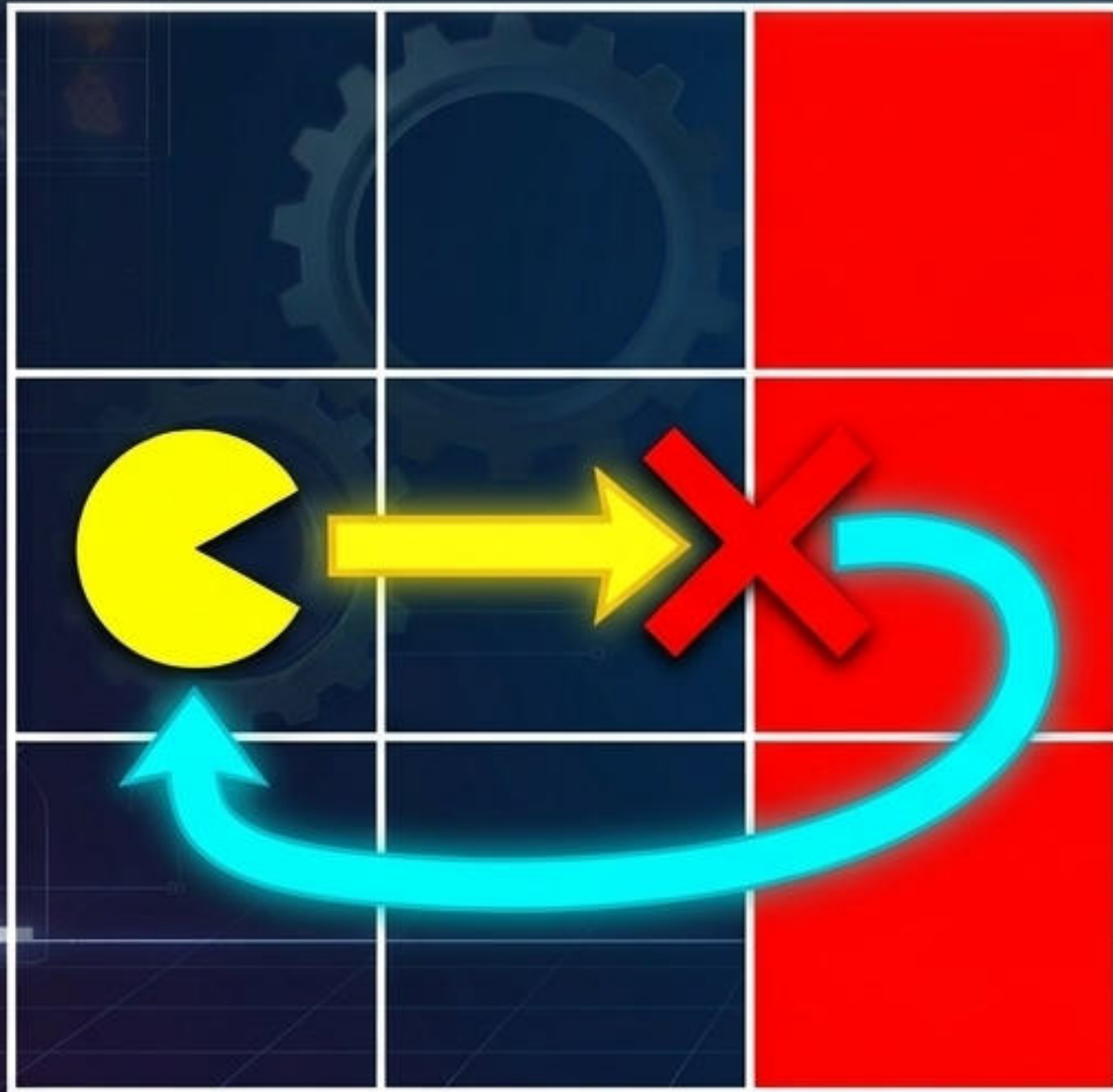
השוואה דיאגנוסטית: מה לבחור **מתי?**

	Value Iteration	Policy Iteration
מה מעדכנים בכל איטרציה?	ערכים מספריים (Utilities)	אסטרטגיה של בחירת פעולות (Actions)
ליבת החישוב (Math)	חילוץ Max מתוך הסתברויות (חישוב לא לינארי)	פתרון מערכת משוואות לינאריות (קבועה מראש)
קריטריון עצירה	כשהערכים המספריים מפסיקים להשתנות משמעותית	כשהמדיניות עצמה מפסיקה להשתנות כלל (שינוי בינארי)
מהירות התכנסות אופיינית	איטי יותר (הרבה איטרציות קצרות)	מהיר יותר (מעט איטרציות עמוקות)

במבחן: כאשר נתונה מדיניות ראשונית כלשהי, זו לרוב רמז ברור
להשתמש ב-**Policy Iteration** כדי לפתור מערכת משוואות.

יישום במעבדה 1: הפיזיקה של עולם הגריד (קירות)

תנועה למעלה: 0.8, שמאלה: 0.1, ימינה: 0.1



המכשלה הקלאסית (Edge Case):
מה קורה כשהסוכן נתקל בקיר?

לפי חוקי המקור, התנועה מבוטלת והסוכן נשאר במקומו.

משמעות מתמטית בנוסחה:

יש לחבר את ההסתברות של תנועת ההתנגשות לתוך ההסתברות של 'להישאר במקום' באותו מצב.

(לדוגמה: אם תנועה למעלה עם סבירות 0.8 פוגעת בקיר, נוסיף 0.8 למקדם של $U_{current}$).

יישום במעבדה 2: הווה מול עתיד (אפקט ה- γ)

+50	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Start														
-50	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1

בעיית הזיהום: מתי כדאי לסבול עכשיו בשביל להרוויח אחר כך?

$$50\gamma - \frac{\gamma^2(1-\gamma^{100})}{1-\gamma}$$

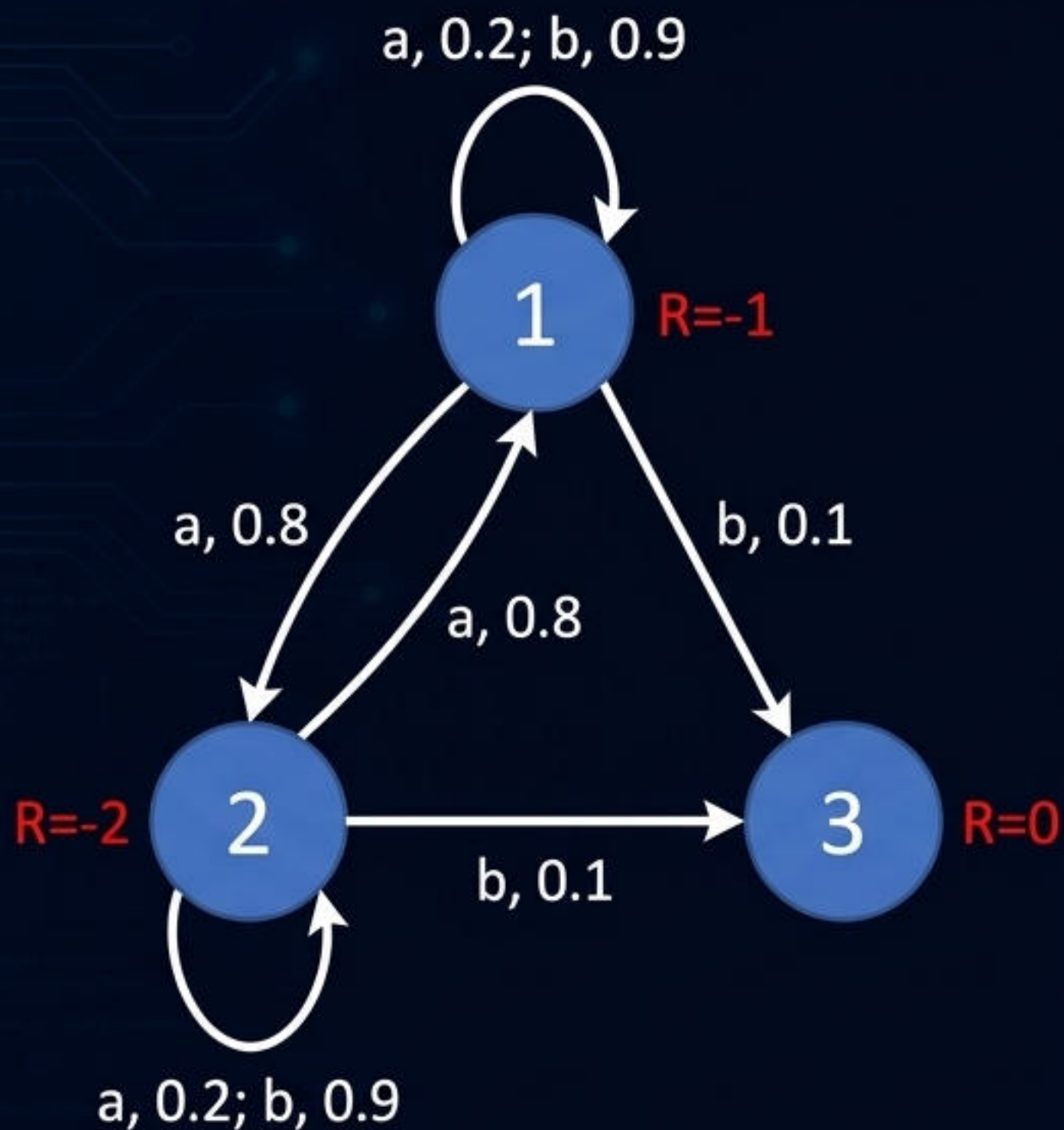
פעולת Up (אקזיט מהיר):
(רווח מיידי גבוה + עונש קטן לנצח)

$$-50\gamma + \frac{\gamma^2(1-\gamma^{100})}{1-\gamma}$$

פעולת Down (השקעה):
(עונש מיידי כבד + רווח קטן לנצח)

על ידי השוואת שני הביטויים, מוצאים את נקודת האיזון: $\gamma = 0.9844$. אם $\gamma > 0.9844$, לסוכן אכפת מספיק מהעתיד כדי לספוג את המינוס המיידי המיידית וללכת למטה. אם γ נמוך מזה – הוא יעדיף את ה-50 המידיים וינוע למעלה.

יישום במעבדה 3: הרכבת משוואות (Policy Iteration)



בניית משוואות הערכה (Value Determination):

נתון: מדיניות התחלתית כופה פעולה b בכל המצבים.

$$u_1 = -1 + (0.1 \cdot u_3) + (0.9 \cdot u_1)$$

$$u_2 = -2 + (0.1 \cdot u_3) + (0.9 \cdot u_2)$$

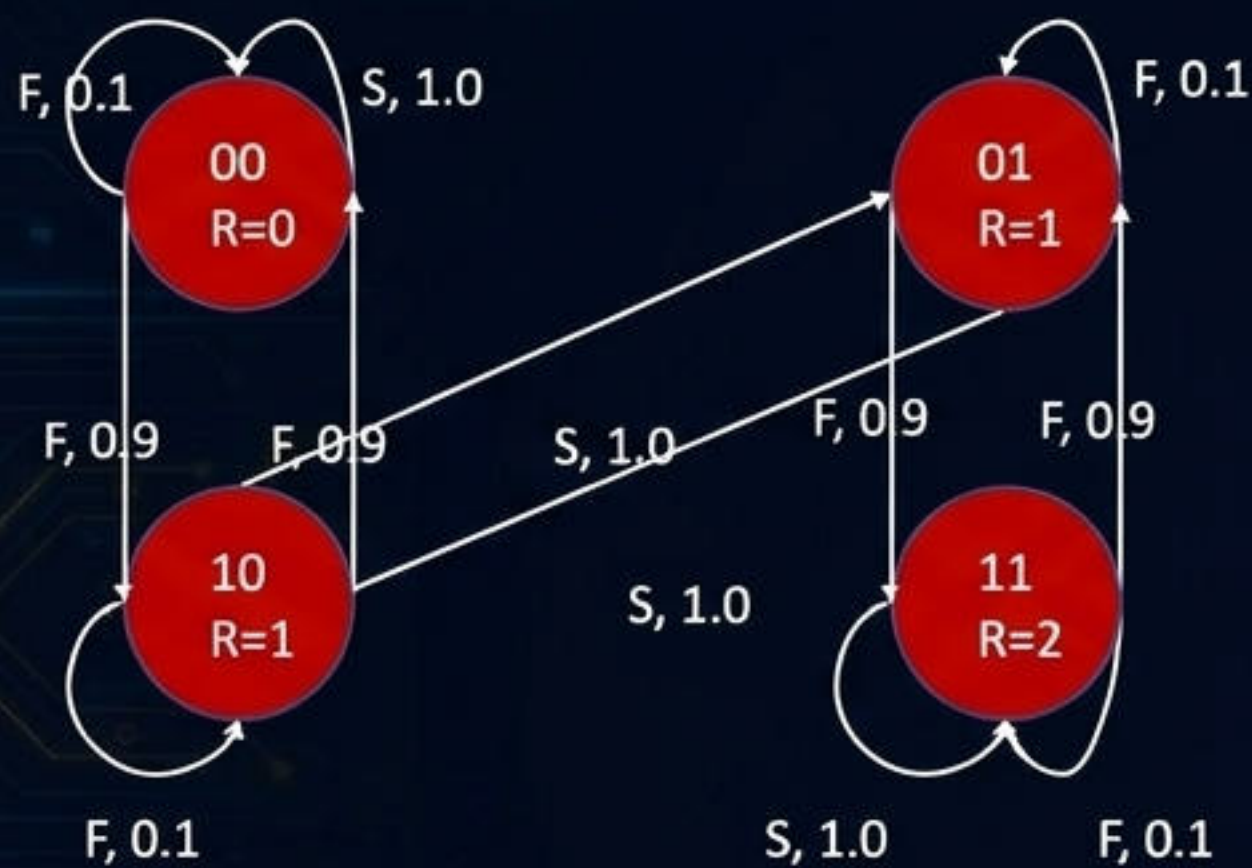
$$u_3 = 0 \text{ (Terminal)}$$

העיקרון:

ברגע שהמדיניות קבועה (בחרו עבורנו את b), נעלים ה-MAX ממשוואת בלמן.

מחלצים מתוך הגרף מערכת פשוטה של משוואות בשלושה נעלמים, פותרים לינארית, ולאחר מכן בודקים בעזרת MAX האם פעולה a הייתה מניבה תוחלת גבוהה יותר לשיפור המדיניות.

יישום במעבדה 4: עולם של ביטית (Value Iteration)



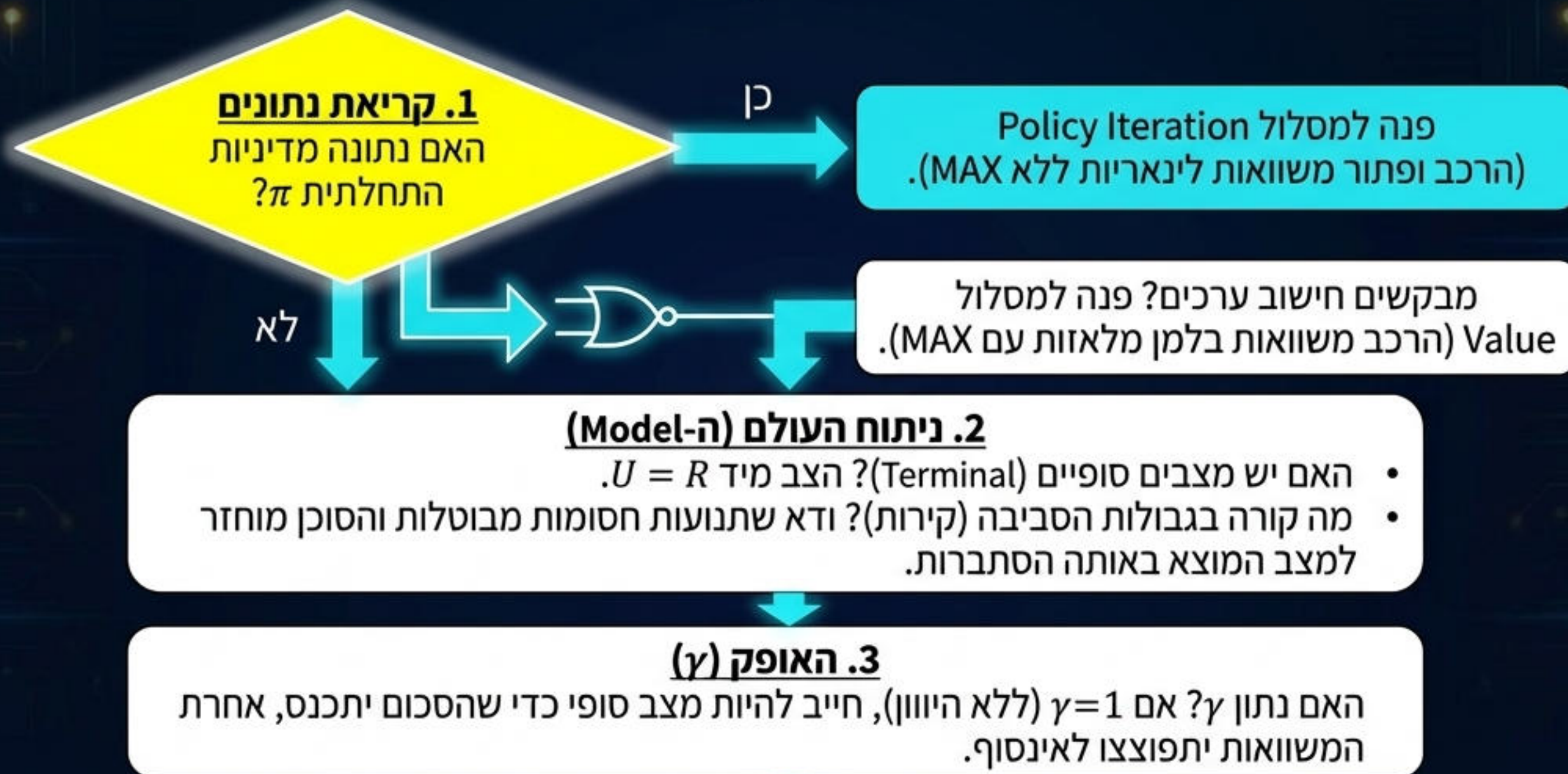
$$U(00) = 0.8 * \text{MAX} (0.9*U(10) + 0.1*U(00) , U(00))$$

לא כל בעיית MDP היא מיקום במבור! מצבים יכולים להיות ערכי ביטים. העיקרון זהה: מזהים עבור כל פעולה (F, S) לאן היא מובילה, מרכיבים תוחלת, ועוטפים בבחירת המקסימום המוכפלת בפקטור ה- γ .

	00	01	10	11
U_0	0	0	0	0
U_1	0	1	1	2
U_2	0.72	2.52	1.8	3.6

בביצוע מעקב איטרציות ידני, מציבים ערכי איטרציה i בצד ימין של הנוסחה כדי לקבל את ערך $i+1$ בצד שמאל.

תרשים זרימה למבחן: פיצוח בעיות MDP



תהליכי החלטה מרקוביים מאפשרים להפוך אי-ודאות למערכת משוואות פתירה.
שלטו בארכיטקטורה, ותשלטו בעולם הסטוכסטי.