מוח ונפש במבט קוגניטיבי

סיכום של ניצן ברזילי

תש"פ 2020, מספר קורס 15420, מרצה: פרופ' אורון שגריר

תוכן עניינים

שיעור 1 – מהי קוגניציה?

סוגים של מצבים נפשים

- תחושות גופניות bodily sensations - כאבים, תענוגות גופניים, דגדוגים, חום וקור
- התנסויות חושיות sensory experience – ראיית צבע, שמיעת צליל, טעימת יין.
- רגשות emotions – פחד, כעס, חרדה, זעזוע, אהבה...
- מחשבון, אמנות (belief ולא faith), רצונות, תקוות, הערכות, כוונות, שיפוטים. מה שמאפיין את התכונות האלו היא שניתן להגדיר אותן כאמיתיות או שקריות.
- למידה, קבלת החלטות.
- טיפשות, גאונות, קמצנות, נדיבות, אטימות.
- דמיון מנטאלי, הזיות, חלומות.
- בינה (אינטליגנציה).
- מושגים.

נניח לצורך השיעור שלמרות שכל המצבים האלו שונים מאוד זה מזה אך יש תכונה המאחדת את כולם והופכת את כולם ל"תופעה נפשית", ונרצה לדון במהי התכונה הזו.

נדון במהו הסמן mark של קוגניציה – מודעות? התנהגות כמו למידה או התאמה לסביבה? יכולת לייצג את העולם סביבם? יכולות לוגיות או לשוניות? מבנה המוח שלהם?

דקארט – מהי מחשבה thought? דקארט טען שהמונח "מחשבה" חל על כל המצבים הנפשיים- "מחשבה היא מונח המזוהה לא רק עם הבנה, רצון ודיון, אלא גם עם מודעות תחושתית". הוא טען כי המאפיין המובהק של מחשבה הוא מודעות – "את המונח 'מודעות' אני משיך לכל דבר שקרה לנו שיש לנו תודעה awareness שלו" .

מהי מודעות?

מהי בכלל מודעות? האם יש סוגים שונים של מודעות? האם מודעות היא אכן הסמן המובהק של קוגניציה?

- מתוך ספר פסיכולוגיה במאה ה-19 - "לא ניתן להגדיר מהי מודעות – כל אחד יודע מהי מודעות כי כל אחד הוא מודע".
- מתוך פרויד – "אין מה לדון במהי מודעות כי זה מעבר לכל ספק".
- מתוך ספר פסיכולוגיה מ-1995 – "מודעות היא להחזיק בתפיסות, מחשבות, רגשות ומודעות. המונח בלתי אפשרי להגדרה כיוון שהוא בלתי ניתן לתפיסה מבלי מודעות עצמית. מודעות היא מושג מרתק אך חמקמק, לא ניתן לתאר מהי, ה היא עושה או איך היא התפתחה. שום דבר ראוי לקראיה לא נכתב עליה מעולם".

<u>סוגים של מודעות</u>

- **מודעות פנומנלית**: מאפיינת בעיקר התנסויות ותחושות. לאלו יש פן חוויתי שמייחד אותן – אנחנו חווים את העולם בצורה מסוימת. פילוסופים מדברים על תכונות פנומליות או qualia (ביחיד quale).
- **מודעות רפלקטיבית**: מודעות עצמית (היכולת לעשות רפלקציה על מצבים שקרו לנו), סוג מורכב יותר של מודעות. דקארט דיבר על כך שמודעות היא היחס בין הסובייקט לדברים (מחשבות), הוא השווה בין אנשים (שיש להם מודעות) לחיות (שאין להן מודעות).

- **מודעות גישתית:**
- **מודעות מסדר גבוה:**

האם מודעות היא הכרחית לקיומה של קוגניציה?

- **מצבי זכרון:** ג'ון לוק (מעט אחרי דקארט) – טען שמחשבות, האמנות, כוונות, וזכרונות הם מצבים שלא נלווית אליהם מודעות. אולי היינו מודעים אליהם בעבר, אך כרגע הם בזכרון ואנחנו לא מודעים אליהם.
- **מצבים מודחקים תת-מודעים:** פרויד (תחילת המאה ה-20) – דיבר על מצבים מודחים שהם תת-מודעים. לפי פרויד, יש דחפים או נטיות שאנחנו לא מודעים אליהם, אך יש להם פוטנציאל להיות מודעים (למשל בטיפול פסיכולוגי). ההבדל בין זה לבין המצבים שלוק דיבר עליהם (זכרון), הוא שפרויד מרחיב את המצבים הנפשיים גם למצבים שמעולם לא היינו מודעים אליהם.
- **מצבים לא מודעים:** במדעי הקוגניציה מדברים על מצבים לא מודעים כמו תהליכי ראייה, קשב, קבלת החלטות, הבנת שפה – אלו מצבים שהם לא מודעים, ולא ניתן להפוך אותם למודעים. רוב התהליכים שמדענים קוגניטיביים עוסקים בהם הם מהסוג הזה.

דוגמא – ניסוי המדגים primimg. בניסוי חשפו את הנבדקים ל-2 סוגים של גירויים. בפעם אחת מראים לנבדק מילה (לחם) ואז מילה קרובה סמנטית (חמאה). בפעם אחרת נציג לנבדק מילה (לחם) ואז מילה רחוקה סמנטית (מכונית). לאחר מכן בתוצאות ראו כי הנבדק יזהה את המילה "חמאה" מהר יותר מאת המילה "מכונית".

אינטנציונליות INTENTIONALITY

נקבע ע"י פרנץ ברנטאנו, פסיכולוג אוסטרי, בסוף המאה ה-19 עם צמיחתה של הפסיכולוגיה כמדע.

המונח אינטנציונליות מתייחסת לכל המצבים הנפשיים הקוגניטיביים, ולא מתייחסת רק לכוונה או התכוונות (intention) – כוונה היא סוג אחד של מצב אנטנציונלי.

המונח מתייחס למחשבות עם כיווניות direction, כלומר המחשבות מוכוונות, הן מחשבות "על משהו" – מצבים, תכונות, ודברים שנמצאים בעולם החיצוני או הפנימי. לדוגמא, מחשבה על הטעם של יין שאנחנו שותים או הצבע שלו, או מחשבה על אדם אחר, או האמנה על מצבים והתרחשויות בעולם. המחשבות שלנו יכולות להיות גם מוכוונות פנימה – אנחנו יכולים להרהר באמונות והמחשבות שלנו. ברנטנו טען שיש למחשבות תוכן (אהבה, שנאה, שיפוט, הסכמה וכו') – ושהתוכן והכיווניות האלו הם אלו שמייחדים תופעות נפשיות (לדוגמא – התפרצות הר געש היא חסרת כיווניות / אודותיות, ולכן אינה תופעה נפשית).

דוגמאות למצבים אינטנציונליים:

- האמנה
- תקווה
- מחשבה
- כוונה
- רצון
- השערה

לכל המצבים האלו יש תכנים פרופוזיציונליים – תכנים שאפשר לתאר אותו באמצעות משפט – לדוגמא "דונלד טראמפ הוא נשיא ארצות הברית". זאת לעומת מושגים כמו "כלב" שהם לא משפט בודד אלא אוסף תכונות. כל המצבים הנפשיים ברשימה למעלה ניתנים לתיאור באמצעות משפט – אני מאמינה ש... אני רוצה ל...

הסברים אינטנציונליים – הסברים באמצעותם אנחנו מנסים להסביר מצבים (לדוגמא – למה יוסי ירה בשכנו? ההסבר – כי יוסי רצה שהשכן ימות והאמין שאם ירה בו אז הוא ימות. ההסבר מבוסס על היסק לוגי – ירייה מובילה למוות, ולכן אם רוצים מוות יש לבצע יריה).

לנושאים הבאים יש משמעות אינטנציונלית (בלומר אודותיות ותוכן):

- מושגים (לדוג' כלב)
- רגשות (לדוג' פחד)
- חלומות
- התנסויות חושיות
- מצבים נפשיים תת מודעים
- מצבים נפשיים לא מודעים

ייצוגיות

במחקר הקוגניטיבי המודרני הראו שיש ממש נוירונים במוח שאחראים על ייצוג של מושגים מסוימים. נדון בשאלה האם אינטנציונליות היא הסמן המובהק של תופעות נפשיות, והאם גם לחוויות יש מימד אינטנציונלי. **בעיית האינטנציונליות** – כיצד אינטנציונליות ממומשת במצבים נפשיים? איך יתכן שמצבים נפשיים עוסקים באירועים, אובייקטים ומצבים הנמצאים חיצונית לנו? התשובה היא יצוגיות – למצבים נפשיים יש אלנט ייצוגי – הם בעצמם או שהם מכילים ייצוגים של מידע על העולם. בעית היצוגיות – מה הופך משהו לייצוג מנטלי? מה קובע את התוכן של היצוג?

האם אנטנציונליות הכרחית לקיומן של תופעות נפשיות? 3 גישות שטוענות שלא –

- **גישות דינמיות** – אומרות שאנחנו יכולים לתאר מערכת קוגניטיבית כמו מערכות דינמיות בטבע, באמצעות משוואות עם פרמטרים שמייצגים מצבים נפשיים. כלומר, אין צורך להשתמש בייצוגים.
- **תפיסה ישירה** – טענה שאנחנו תופסים את העולם באופן ישיר ולא מייצגים אותו.
- גישה הטוענת שאנחנו לא מיצגים את העולם אלא חווים אותו, שוהים בו ומהווים שלוחה שלו.

שיעור 2 – דואליזם

מה מקומה של הנפש בעולם פיזיקלי?

מהו עולם פיזיקלי?

- עצמים חומריים (גופיים) תופסים מקום במרחב החלל והזמן
- התכונות של גופים (מסה, גודל, צורה, תאוצה וכו') ניתנים לכימות ותיאור מתמטי.
- חוקי הטבע הם אוניברסליים וחלים על כל התופעות החומריות באותה צורה.

ננסה לענות על השאלה **"מה מקומה של הנפש בעולם פיזיקלי?"**. ראשית נסכים כי בקיימת הרמוניה בזמן בין אירועים נפשיים באירועים פיזיקליים (אני חושבת על להזיז את היד, ואז היד שלי זזה בעולם הפיזיקלי, וזה קרה באותו הזמן). ניתן למקם אירועים נפשיים בזמן, אך קשה למקם אותם בחלל הפיזיקאלי (בין אם במיקום בעולם החיצוני או אפילו במיקום בתוך המוח). כמו כן, לא ברור שמערכת המדידה (משקל, גובה, מהירות וכו') והחוקים הפיזיקאליים חלים על אירועים נפשיים (כמה שוקל כאב? מה הצורה של רצון?).

יש שתי עמדות הנוגעות לשאלה זו:

1. **דואליזם (קרטזיאני):** גוף ונפש הם עצמים נפרדים, אירועים נפשיים ואירועים פיזיקליים הם שונים במהותם.
2. **מטריאליזם (הובס, לה מטרי):** גוף ונפש הם אחד, אירועים נפשיים הם חלק בלתי נפרד מהעולם הפיזיקלי.

דקארט

דקארט מייצג בצורה הכי מובהקת את העמדה הדואליסטית. נולד בצרפת במאה ה-17, כתב ספרים על מתמטיקה ופיזיקה, והיה פילוסוף שנחשב לאבי הפילוסופיה המודרנית. בפילוסופיה יש שני ענפים מרכזיים – מטאפיזיקה / אונתולוגיה: ענף העוסק בשאלות של מה קיים בעולם (האם יש אלוהים? האם יש נפש? וכו'), ואפיסטמולוגיה / תורת ההכרה: ענף העוסק בשאלה "מה אנחנו יודעים" (בניגוד ל"מה קיים").

דקארט חולל את מה שכונה "המהפכה האפיסטמולוגית" ששמה את האפיסטמולוגיה במרכז העיסוק הפילוסופי, ואמר שאין מה לדבר על מה קיים בעולם מבלי לדון בשאלת הידיעה. דקארט השתמש במתודה ספקנית – הוא הטיל ספק בהרבה דברים (העולם החיצוני, אמיתות מסוגים שונים) אך טען שהוא אינו יכול לפקפק בדבר אחד, וזוהי העובדה שיש לו מחשבות, וטבע את המונח המפורסם "אני חושב משמע אני קיים".

מהי מחשבה?

דקארט זיהה את המחשבה עם מודעות, ואמר שהן דברים פנימיים המכילים מושגיות מסוימת המבדילה בני אדם מחיות (לדוגמה היכולת למחשבה רפלקטיבית). מאפייני המחשבה לפי דקארט:

- **ייחוס של מחשבות** – ההבדל בין יחוסי גוף ראשון (אין אנחנו מיחסים מחשבות לעצמנו) ליחוסי גוף שלישי (איך אנחנו מיחסים מחשבות לאחרים). לפי דקארט, אנחנו מיחסים לאחרים מחשבות לפי מה שהם אומרים והאופן בו הם מתנהגים. כשזה נוגע לעצמנו, אנחנו יודעים מה המחשבות שלנו באופן ישיר וללא צורך בעדויות.
- **אוטוריטה של גוף ראשון** – לפי דקארט, ייחוסי גוף ראשון לא יכולים להיות מוטעים. כלומר, אנחנו לא יכולים לטעות לגבי מה אנחנו חושבים על משהו – אם אני חושבת שטראמפ הוא דביל, אני אדע שאני חושבת שטראמפ הוא דביל, ואין סיכוי שאטעה ואחשוב שאני חושבת שטראמפ גאון.
- **אחדות התודעה** – דקארט אמר שהחוויה שלנו היא אחודה ולא ניתנת להפרדה לחלקים (בניגוד לגופים פיזיים).

דואליזם קרטזיאני גורס שהתכונה המהותית של גופים פיזים היא תפיסת מקום פיזי בחלל, לעומת מחשבות שהתכונה המהותית שלהן היא מודעות. דקארט טען שגופים ומחשבות הם ישויות נפרדות – לגופים אין מודעות, ומחשבות אינן תופסות מקום בחלל.

דגשים לגבי דואליזם:

- לפי הדואליזם, **נפש וגוף נמצאים בקשר הדוק**. כלומר הם לא טוענים שאין קשר בין הגוף לנפש. גם דואליסטים וגם מטריאליסטים יניחו שהקשר בין הנפש לעובדות חומריות הוא הכרחי ולא קונטיגנטי (ראו בהמשך).
- **מצבי תודעה נגרמים ע"י מצבים מוחיים**, כלומר הם לא טוענים שהפיזיולוגיה לא משפיעה על המצב הנפשי.
- **ניתן לחקור מודעות באופן מדעי**, אך לא ניתן להסביר מודעות באופן מלא ע"י מדעי המוח.
- **מצבים קוגניטיביים אינם בהכרח מודעים**.
- טיעונים דואליסטיים מבססים (במקרה הטוב) טענה לגבי מערכת המושגים שלנו, ולא לגבי העולם עצמו – זוהי החולשה שלהם.

טיעונים בעד העמדה הדואליסטית

- **"טיעון ההתקבלות על הדעת"** (דקארט, קריפקי, צ'אלמרס): נתמקד בגישה של קריפקי. נניח שמצאנו קורלציה בין קיומו של כאב לבין התופעה הפיזיולוגית של היותם של עצבי ה-C של האדם מגורים – זוהי הנחה מתקבלת על הדעת. לכן אפשרי שחווית הכאב אינה עצבי C מגורים אלא תופעה נפרדת מהם. הכוונה ב"מתקבל על הדעת" היא שלא ניתן להסיק על סמך הידע הקיים שהטענה היא שקרית. הכוונה ב"אפשרי" היא שיתכן מצב (גם אם היפותטי) בו הטענה מתקיימת. נסביר את הקפיצה מ"**אפשרי** שכאב הוא נפרד מגירוי עצבי C" ל"כאב הוא **אכן** נפרד מגירוי עצבי C". נפריד בין טענות הכרחיות (שהן אמיתיות תמיד), לטענות קונטינגנטיות (שלפעמים אמיתיות ולפעמים לא). ההבדל בין טענה הכרחית לטענה קונטינגנטית היא אם התכונה שמדברים עליה היא מהותית, כלומר האם המושא שהיא מתארת יכול להתקיים בלעדיה. אם היא מהותית, הטענה הכרחית, ואם לא – היא קונטינגנטית. קריפקי טען שיש טענות העוסקות בזהות של דברים שהן גם הכרחיות וגם אמפיריות (כלומר ניתנות למדידה). הטענה היא ש"כאב **הוא** עצבי C מגורים" היא טענה הכרחית / מהותית, כלומר – אם המשפט נכון הטענה אמיתית בהכרח, ואם הוא שגוי היא שקרית בהכרח. כלומר – אם אפשרי שכאב הוא נפרד מעצבי C מגורים, אז הטענה אינה נכונה. לכן, כיוון שמדובר בטענה מהותית, ויתכן מצב בו היא שקרית – קיבלנו שכאב הוא נפרד מעצבי C מגורים.
- **"טיעון הידע"** (ג'קסון): בנוגע לניסוי המחשבתי על המדענית מארי – אומר שאם נצא מנקודת הנחה שלמארי יש את כל הידע הפיזיקלי לגבי העובדות שקשורות בראיית צבעים, אך במקביל ישנו ידע לא-פיזיקלי לגבי ראיית צבעים שלמארי אין אותו (איך זה מרגיש לחוות צבע מסוים). מנקודות ההנחה האלו ניתן להסיק שתי מסקנות – המסקנה הראשונה היא שיש ידע לא-פיזיקלי המעורב בראיית צבעים. המסקנה השניה היא שיש עובדות לא פיזיקליות הקשורות לראיית צבעים. זהו טיעון נגדי למטריאליזם, שכן מטריאליזם לא יצליח להסביר את התופעה של תודעה של ראיית צבעים באופן נפרד מהתפקוד הפיזיולוגי ומהידע בנושא.

רנה דקארט – קטעים מתוך הגיונות

דקראט מתאר בספרו כי נסיון חייו גרם לו לפקפק במידע שהוא מקבל מהחושים (לדוגמא בגלל אשליות אופטיות, או עדויות של קטועי גפיים על כאבי פאנטום), בין היתר כי הוא לעולם לא יכול לדעת בוודאות שהוא לא בתוך חלום. עם זאת, הוא לא מפקפק בכך שהעולם והדברים המוחשיים שנמצאים בו הם אמיתיים. הוא אומר שזה שהוא תופס משהו בצורה מסוימת שני דברים כנפרדים זה מזה, אומרת שהם נבראו ע"י אלוהים באופן שמאפשר להבחין ביניהם. הוא טען נחרצות כי נפשו מגדירה את אישיותו, והיא נפרדת מגופו ויכולה להתקיים גם בלעדיו. הוא התמקד ביכולות המחשבתיות של דמיון ותחושה, אשר מהוות חלק מתפיסתו העצמית, אך אינן מתקימות בלעדיו.

דייוויד צ'אלמרס – THE PUZZLE OF CONCIOUS EXPERIENCE

צ'אלמרס טוען שהחוויה של תודעה מתחלקת לשניים – חלק אובייקטיבי (נוירונים פועלים ושולחים אותו חשמליים) וחלק סובייקטיבי (תחושת המודעות, מחשבות, התפיסה המנטלית הישירה של גירויים חיצוניים). יש מדענים החושבים שניתן להסביר את התודעה בשיטות מחקריות סטנדרטיות של פסיכולוגיה ונוירולוגיה, ואחרים שחושבים שלעולם לא נצליח להבין מהי מודעות. צ'אלמרס מאמין שהאמת נמצאת איפה שהוא באמצע בין שתי התפיסות האלו – ואומר שכלים בנוירולוגיה אמנם לא יכולים לתפוס את החוויה המודעת במלואה, אך יש להם הרבה מה להציע למחקר בנושא.

מדענים משתמשים במילה "מודעות" בהרבה דרכים שונות, וצ'אלמרס מבקש להפריד את בעיית המודעות לבעיות נפרדות – **"בעיות קלות" Easy ו"הבעיה הקשה" Hard**. בעיות קלות הן לא טריוויאליות, והן

מאתגרות מאוד את המחקר הפסיכולוגי והביולוגי, אך צ'אלמרס מאמין שהמשך המחקר בתחום הנוירולוגיה והפסיכולוגיה הקוגניטיבית יאפשר מענה עליהן. שאלות שנכללות לדעתו תחת "בעיות קלות" -

- איך אדם מתייחס לגירוי סנסורי ומגיב אליו בהתאם?
- איך המוח חוקר מידע מהרבה מקורו ומשתמש בו כדי לשלוט בהתנהגות?
- איך אנשים יכולים לתאר מילולית את מצבם הפנימי?

הבעיה הקשה היא השאלה "איך תהליכים פיזיולוגיים במוח מעוררים חוויות סובייקטיביות?".

הניסוי המחשבתי של הפילוסוף פראנק ג'קסון – מארי היא נוירולוגית מהמאה ה-23 המובילה בעולם במחקר על התהליכים המוחיים במוח שראיים לראייה בצבע – אבל מארי חייתה כל חייה בחדר שחור-לבן ומעולם לא ראתה צבעים. היא יודעת כל מה שצריך לדעת על התהליכים במוח האחראיים לראיית צבע – אך חסרה לה ההבנה של **איך זה מרגיש** לראות צבע. הניסוי המחשבתי הזה מדגיש את הבעיה הקשה – שכן לא ידעו למה תהליכים פיזיים מלווים מחוויה מודעת – למה כשהמוח מעבד אור באורך גל מסוים, נוצרת לנו חוויה של הצבע הסגול? האם ידיעה שמדובר בצבע הסגול לא היתה מאפשרת לנו לתפקד באותה הצורה מבלי שתתקיים החוויה המודעת?

התאוריה של קריק וקוך Crick & Koch: טוענים שמודעות נוצרת מחזרתיות בקורטקס, שהופכת לסינכרונית כשנוירונים יורים בקצב של 40 פעמים בשניה. קריק וקוך מאמינים שהתופעה הזו עשויה להסביר איך מאפיינים שוים של אובייקט מסוים (צבע וצורה, לדוגמא), שמעובדים באזורים שונים במוח, מאוחדים לתפיסה אחידה וקוהרנטית. לפי התאוריה שלהם, שתי פיסות מידע מתאחדות באופן מדויק כאשר הן מיוצגות ע"י תדרי ירי מסונכרנים של נוירונים. קריק וקוך טענו שהבעיה הקשה לא יכולה להפתר ע"י המדע.

הפער ההסברתי The Explanatory Gap : יש שהציעו שכדי לפתור את הבעיה הקשה יש להשתמש בכלים חדשים כמו מכניקת קוונטים. אך גם אם ניתן יהיה להסביר באמצעותה דברים כמו איך המוח מקבל החלטות, עדיין היא לא תסביר איך התהליכים האלו תורמים לחווית המודעות. גם תאוריות פיזיקליות מספיקות – נראה שמודעות אינה נובעת מחוקים פיזיקליים. פיזיקה יכולה להסביר את התופעות שקורות בקורלציה למודעות (לדוגמא תדר הירי של נוירונים), אך לא את המודעות עצמה. בפיזיקה יש חוקים בסיסיים Fundamental Laws – במקרה של התודעה, החוקים האלו עשויים לקשר בין חווית המודעות לבין התאוריה הפיזיקלית. החוקים האלו יכולים להוות גשר שמגשר על הפער ההסברתי. תאוריה שלמה המסבירה את המודעות צריכה להסביר את ההתנהגות של המערכת הפיזיקלית, ולהסביר איך המערכת הזו מקושרת לחווית המודעות.

שיעור 3 – מטריאליזם

העמדה המטריאליסטית אומרת שעובדות נפשיות (קוגניטיביות) אינן אלא עובדות חומריות, למשל פעילות מוחית.

סוגים של מטריאליזם:

- **מטריאליזם מוחי**: "העמדה הטריויאלית ביותר", טוענת שעובדות נפשיות הן עובדות מוחיות. היתה לה אחיזה כבר במאה ה-17 וקיבלה ביסוס בשנות ה50 של המאה ה20.
- **עמדות מטריאליסטיות אחרות**: לפיהן עובדות נפשיות הן עובדות מסוג אחר, למשל נטיות התנהגותיות או עובדות פונקציונליות-חישוביות.

הבסיס של העולם הוא עובדות פיזיקליות (קוואנטים, אלקטרונים, פוטונים) ואלו מתוארים ע"י תחום הפיזיקה (הקוואנטית, החלקיקית וכו'). כל שאר העובדות נקבעות ע"י אותן עובדות פיזיקליות:

- **נסמכות**: עולם אחר שבלתי נבדל משלנו מבחינה פיזיקלית (כל התכונות הפיזיקליות זהות) יהיה זהה גם מכל בחינה אחרת (כולל עובדות נפשיות). דואליזם (בשונה מפיזיקליזם) יתמוך בשלילת הנסמכות – כלומר שיתכן מצב בו הפיזיקה תהיה זהה אך העובדות הנפשיות יהיו שונות.

סוגי פיזיקליזם:

- **פיזיקליזם רדוקטיבי**: כל העובדות (כולל עובדות נפשיות) ניתנות לרדוקציה לעובדות פיזיקליות.
- **פיזיקליזם לא רדוקטיבי**: כל העובדות נקבעות ע"י עובדות פיזיקליות אך חלקן אינן פיזיקליות בעצמן.

ריאליזם

מטריאליסטים לא מכחישים את קיומן של עובדות נפשיות. רוב המטריאליסטים יטענו שעובדות נפשיות הן קיימות – אך יטענו שהן לא יותר מפעילות מוחית. כלומר, רוב המטריאליסטים הם ריאליסטים – טוענים שהעובדות הנפשיות הן עובדות פיזיקליות. "אם אינטנציונליות היא אכן אמיתית, אז היא חייבת להיות משהו ממש אחר (בכוונה לעובדה חומרית כל שהיא)".

יש מטריאליסטים מעטים שמכחישים את קיומן של עובדות נפשיות- הגישה נקראת **מטריאליזם אלימינטיביסטי**. זהו זרם של המטריאליזם. חלקם יטענו שמצבים נפשיים הם מושגים המציינים גורמים מסוימים אך אינם מצביעים על קיומן של עובדות (כמו מושגים שהמדע עסק בהם בעבר אך התגלה בדיעבד שהם שגויים).

סיבתיות הנפשית

גורס שאירועים נפשיים הם **סיבות** של אירועים פיזיקליים, לדוגמא:

- הרצון שלי להרים את היד (סיבה) גם להרמת היד (תוצאה)
- הידיעה שהדרך מסוכנת (סיבה) גרמה לי לבחור בדרך אחרת (תוצאה).

למה סיבתיות נפשית הוא מושג מרכזי בפילוסופיה:

- סיבתיות נפשית מסבירה לפחות חלק מההרמוניה הזמנית בין הנפשי והחומרי (כלומר מסבירה למה הם קורים ביחד).
- היא מציעה אבחנה בין שני סוגים של תנועות גופניות – פעולות (רצוניות, שקדם לה רצון נפשי) ורפלקסים (לא רצוניים).
- עוסקת ברצון חופשי free will – היכולת להחליט ולפעול כרצוננו. לפחות חלק מההנחה שיש לנו רצון חופשי מתבססת על סיבתיות נפשית – היא הכרחית אך לא מספיקה לקיומו.
- אחריות מוסרית – אנחנו מטילים על אנשים אחריות מוסרית, כיוון שאנחנו יוצאים מנקודת הנחה שהפעולות שלו מקורן בפעולות נפשיות (לדוגמא מישהו ירה בשכן שלו כי הוא החליט לירות בו, לא כי היה לו התקף רעידות שגרם לו ללחוץ על ההדק)
- קיום וסיבתיות – כשמדברים על קיומם של אירועים, תכונות ועובדות, אנחנו כמעט תמיד מקשרים אותם לסיבתיות מסוימת – אם הוא לא יכול לגרום לשום דבר, לא בטוח שהוא קיים.

לכן סיבתיות נפשית משחקת תפקיד חשוב בפילוסופיה וגם בחיינו כבני אדם. בד"כ מסתכלים על סיבתיות בתור יחס של שני אירועים – סיבה ותוצאה. לרוב רואים את הסיבה כקודמת לתוצאה מבחינת זמנים. אך ישנם

יחסים אחרים שקודמים זה לזה מבחינת זמנים, אך זה לא אומר שיש סיבתיות בניהם. לכן מלבד לקדמה בזמנים צריך להוסיף פרמטר נוסף על מנת להגדיר סיבתיות – יש הרבה הצעות בנושא. ההצעה של דיויד יום היא "חוקיות", כלומר אם אחרי כל אירוע מסוג א' יופיע אירוע מסוג ב', יש חוקיות ולכן יש סיבתיות. הצעה נוספת של דיויד לואיס היא דומה להצעה הראשונה, ואומרת שהפרמטר המגדיר סיבתיות הוא "אם אירוע א' לא היה מתרחש, גם אירוע ב' לא היה מתרחש".

טיעון הסיבתיות הנפשית

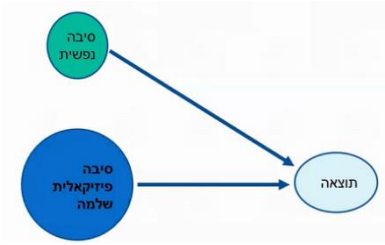
מורכב משלושה חלקים:

- **סיבתיות נפשית**: אירועים נפשיים הם סיבות של אירועים פיזיקליים.
- **הסגור הסיבתי של העולם החומרי**: לכל תוצאה פיזיקלית **ב'** יש סיבה פיזיקלית מספיקה / שלמה (כלומר שלא צריך תנאים נוספים) שהיא הסיבה **א'**.
- **היקבעות יתר**: אם אירוע **א'** הוא סיבה מספיקה / שלמה של אירוע **ב'**, ואם אירוע **ג'** הוא סיבה של אירוע **ב'**, אז אירוע **ג'** זהה או חלק מאירוע **א'**.

באופן ציורי:

<u>שתי ההנחות הראשונות:</u>

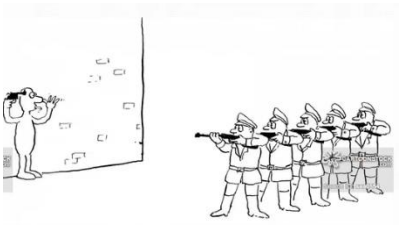
אם הופיעה תוצאה, סימן שיש לה סיבה נפשית שהיא סיבה פיזיקלית שלמה.



<u>ההנחה השלישית:</u>

אם לתוצאה יש סיבה נפשית וסיבה שלמה, אזי הסיבה הנפשית היא זהה או חלק מהסיבה הפיזיקלית.

ערעור על הנחת הקבעות היתר (ההנחה השלישית)



המערערים על טיעון הסיבתיות יוצאים נגד הנחת הקבעות היתר, ואומרים שישנם מקרים שלמרות שיש בהם הקבעות יתר של סיבות, ובכל זאת כל אחת מהסיבות היא נפרדת ועומדת בפני עצמה. דוגמא מפורסמת – כיתת היורים (הדגמה בציור מימין). כל אחת מהיריות של כל אחד מהיורים היא מספיקה על מנת להרוג את האיש, כולם ירו ביחד והאיש מת. כלומר, יש כאן מקרה בו כל אחת מהסיבות למוות (יריה מצד אחד היורים) היא מספיקה, אך אין זה אומר שהן לא נפרדות זו מזו.

<u>תשובת המטריאליסט לערעור על הנחת הקבעות היתר –</u>

מטריאליזם מבחין בין:

- **הקבעות יתר סינגולרית** (הקבעות יתר שמתרחשת באופן לא שיטתי אלא באופן מקרי). לדוגמא בביתת היורים – יש מקרים בהם ירו באיש 5 אנשים ויש מקרים שבהם יש רק יורה אחד, והוא הסיבה למוות של האיש. לכן יריה היא הקבעות יתר סינגולרית.
- **הקבעות יתר שיטתית** (הקבעות יתר שמתרחשת תמיד). לדוגמא, בכל פעם שאני מחליטה להרים את היד, היד שלי באמת תתרומם.

מטריאליסטים יטענו כי אם ישנה הקבעות יתר שיטתית, על הדואליסטים לתת הסבר ללמה הרצון שלי להרים את היד מלווה **תמיד** בסיבה פיזיקלית שלמה אם לא מדובר באותה הסיבה.

ערעור על הנחת הסגור הסיבתי של העולם החומרי (ההנחה השניה)

יש המערערים על ההנחה השניה, ואומרים שהעולם אינו דטרמיניסטי – כלומר ישנו אלמנט של אקראיות בעולם, לכן לא ניתן לקבוע שאירוע א' יגרום להופעתו של אירוע ב' – יתכן שאירוע ב' לא יתרחש, או שיתרחש אירוע אחר ממה שציפינו ועוד.

התשובה המטריאליסטית לטיעון זה היא שאלמנט האקראיות בעולם הוא חלק מהעולם החומרי ונלקח בחשבון בהגדרת הסיבתיות בין האירוע הפיזיקלי והתוצאה.

ערעור על הנחת הסיבתיות הנפשית (הסיבה הראשונה)

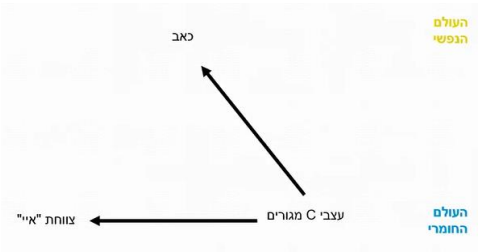
לייבניץ שלל את ההנחה הראשונה, ואמר שאין קשר סיבתי בין אירועים נפשיים ואירועים חומריים. כדי להסביר את ההרמוניה בין קיומן של תופעות נפשיות ופיזיות במקביל, הוא טען שיש הרמוניה קבועה מראש (פרלליזם). לפי פרלליזם, אירועים נפשיים נמצאים בקשר סיבתי רק עם אירועים נפשיים, ואירועים פיזיקליים נמצאים בקשר סיבתי רק עם אירועים פיזיקליים, וההרמוניה בין אירועים פיזיקליים ונפשיים היא קבועה מראש ע"י אלוהים. לדוגמא –

אירועים נפשיים	כאב	←	זכרון
אירועים פיזיקליים	עצבי C מגורים	←	צווחת "איי!!"

כלומר זה שהצווחה "איי" הופיעה לאחר כאב לא אומרת שהיא נבעה ממנו, אלא שהתופעות התרחשו במקביל בגלל שכך אלוהים קבע בבריאת העולם.

אפיפנומנליזם

גישה שעלתה בסוף המאה ה-19 המכחישה את ההנחה הראשונה (שאירועים נפשיים הם סיבות של אירועים פיזיקליים). הגישה האפיפנומנליסטית מתייחסת בעיקר לתופעות נפשיות מודעות (חוויות) וטוענת שיש קשר סיבתי בין תופעות נפשיות לתופעות פיזיקליות, אך הוא חד סטרי – כלומר תופעות פיזיקליות גורמות לתופעות נפשיות אך לא הפוך. כלומר, אירועים נפשיים הם חסרי כוח סיבתי. זוהי לא גישה מטריאליסטית או דואליסטית במהותה.



ניתן לראות בשרטוט כי יש הפרדה בין העולם החומרי לעולם הנפשי, וכי "עצבי C מגורים" (בעולם החומרי) הוא שגרם לשתי תוצאות – אחת בעולם החומרי ואחת בעולם הנפשי, אך הכאב (בעולם הנפשי) לא היה זה שגרם לצווחת "איי" כי הוא חסר כוח סיבתי.

הניסוי של ליבט libet 1985 - נחשב כאישוש לעמדה האפיפנומנליסטית. בניסוי גילו כי יש אות חשמלי ניתן לזיהוי בקורטקס המוטורי במוח, אשר מופיע בערך כחצי שניה (הרבה מאוד זמן) לפני התנועה המוטורית עצמה. ליבט רצה לבדוק מתי מתרחשת החוויה המודעת של קבלת ההחלטה להזיז אצבע – האם היא מתרחשת לפני האות הפיזיולוגי הניתן לזיהוי או אחריו. הוא הושיב נבדקים מול שעון מתקתק, ואמר להם לזכור על מה הצביע השעון כשהם החליטו להזיז את האצבע שלהם. תוצאות הניסוי הראו כי ההחלטה המודעת להזיז את האצבע התרחשה בערך 0.2 שניות לפני הזזת האצבע, כלומר לאחר הסמן הפיזיולוגי.

דיוויד פאפיניו – THE CASE FOR MATERIALISM

טיעון הסיבתיות שמציג Paineau מורכב משלוש הנחות:

1) הנחה ראשונה: מופעים מנטליים מודעים הנם בעלי השפעה פיזית. לדוגמא, תחושה מודעת של צמא יכולה לגרור להשפעה פיזית – הליכה למקרר ופתיחתו לטובת הוצאת מים.

2) הנחה שניה: כל התופעות הפיזיות נגרמות אך ורק ע"י תופעות פיזיות קודמות. לדוגמא, הליכה למקרר נוצרת ע"י כיווץ השרירים ברגליים, שנוצרו בעקבות מסרים חשמליים מהקורטקס המוטורי, שנוצרו בעקבות פעילות בקורטקס הסנסורי וכך הלאה.

3) הנחה שלישית: תופעות פיזיות מודעות אינן תמיד רב-סיבתיות, כלומר הן לא תמיד נוצרות ע"י סיבות ברורות ונפרדות.

טיעון הסיבתיות בא לתמוך בטיעון המטריאליסטי בשאלת המוח-נפש. מסקנת הטיעון היא שישנם מופעים מודעים אשר נוצרים מאחת או יותר משתי הסיבות הבאות: סיבה מודעת (כמו למשל תחושת צמא) המוזכרת בהנחה 1, וסיבה פיזית (כמו למשל ירי של נוירונים במוח) המוזכרת בהנחה 2. הנחה 3 קושרת את ההנחות הראשונות יחדיו, ולפיה המופעים המודעים אינם בהכרח נוצרים ע"י סיבות ברורות ונפרדות. Paineau הסיק מכך שהסיבות המודעות צריכות להיות חופפות, לפחות באופן חלקי, לסיבות הפיזיות. בכך הוא תמך בטיעון המטריאליסטי שטוען כי מצבים מודעים ומצבים פיזיים אחד הם.

עם זאת, בהיעדר ההנחה השלישית (היעדר רב-סיבתיות), המסקנה המתבקשת היא שישנן שתי סיבות נפרדות הגורמות למופעים מודעים – סיבה מודעת, וסיבה פיזית. על שתי סיבות נפרדות אלו להתרחש יחדיו על מנת לגרום להיווצרותו של המופע המודע, שכן כל אחת מהן בפני עצמה אינה מספיקה. מסקנה זו אינה מספיקה כדי לחזק את הגישה המטריאליסטית, שכן היא טוענת שהסיבה המודעת והסיבה הפיזית הן נפרדות – גישה הקרובה יותר לגישה הדואליסטית, לפיה אירועים נפשיים ואירועים פיזיים הם נפרדים במהותם. לכן, רק לאחר הוספת הטענה השלישית המבטלת את המסקנה לפיהן הסיבה המודעת והסיבה הפיזית הן נפרדות, הטיעון אכן מהווה חיזוק לגישה המטריאליסטית.

שיעור 4 – בהוויוריזם – נפש כהתנהגות

בהוויוריזם – גישה **המזהה עובדות נפשיות עם התנהגות** בכלל ונטיות התנהגותיות (דיספוזיציות) בפרט.

למה בהוויוריזם?

- קיים **קשר הדוק בין מצבים נפשיים והתנהגות** - אנחנו מייחסים מצבים נפשיים לאחרים על סמך ההתנהגות שלהם (התנהגות גופנית – שפת גוף, או התנהגות לשונית – מה הם אומרים).

- בהוויוריזם הוא **אלטרנטיבה מטריאליסטית לגישה לפיה מצבים נפשיים הם בעצם פעילות עצבית**.
- **המוטיביציה האפיסטמולוגית** – בעיית ה -other minds (גישה אנטי קרטזיאנית). למוטיבציה הזו יש שורשים בעבודתו של דקארט (מעצם העיסוק במה אנחנו יודעים וההסקה מכך לגבי מהן עובדות נפשיות), אך היא למעשה מתנגדת לדקארט. זאת כיוון שדקארט התמקד בכך שמה שאנחנו יודעים זה קודם כל ה"אני" (החוויה התודעתית), אך הבעיה בגישתו היתה שלא ניתן היה להסיק ממנה שיש עולם, ושיש יצורים קוגניטיביים שאינם "אני" בעולם, ולכן גם אי אפשר לדעת מה אותם יצורים קוגניטיביים אחרים חושבים ומאמינים. לכן הגישה הבהוויריסטית יוצאת מנקודת הנחה שיצורים קוגניטיביים אחרים קיימים ואנחנו יודעים מה הם מאמינים וחושבים (בזכות ההתנהגות שלהם), ורק צריך להבין איך אנחנו יודעים את זה.
- **המוטיבציה הלשונית-סמנטית** – הגדרת המשמעות של מונחים נפשיים. אלו נסיונות מתחילת המאה ה-20 להגדיר משמעות של מונחים, ובכלל כך מונחים נפשיים. כלומר, הכרה בכך שמשמעות נגזרת מהתקשורת בין אנשים וממה שאנחנו מבינים, לכן גם את המשמעות של מונחים נפשיים יש להבין דרך פעולות או אמירות.
- **המוטיבציה הפסיכולוגית** – הבהוויוריזם בפסיכולוגיה צמח במקביל לבהוויוריזם בפילוסופיה, אך בעוד הפילוסופים ניסו "לתת דין וחשבון" למהם מצבים נפשיים, הפסיכולוגים "התעלמו" מקיומם של מצבים נפשיים, וטענו שפסיכולוגים צריכים לתת הסברים על התנהגות ולשם כך אינם צריכים להתייחס למצבים הנפשיים (ובמקום זאת להתייחס רק להתנהגות עצמה).

כיצד הבהוויוריזם מגדיר מצבים נפשיים

נדון לדוגמא במצב הנפשי של כאב. בהוויוריזם עשוי היה להגדיר כאב בתור "זעקת איי" (הביטוי ההתנהגותי של כאב) – כלומר הכאב הוא ממש הביטוי ההתנהגותי עצמו. ההתנגדות הבולטת ביותר לתפיסה זו היא שהיא לא מתקיימת תמיד – לא בכל מקרה של כאב נצעק "איי", למרות שהכאב כן קיים, הביטוי ההתנהגותי לא קיים, לכן לא יכול להיות שהם זהים. לכן, בהוויוריזם לא יזהה את הכאב (או כל מצב נפשי אחר) כזהה להתנהגות, אלא כדיספוזיציה התנהגותית, כלומר הנטיה לבצע את ההתנהגות כשהמצב הנפשי מתקיים, כתלות בגירויים ומשתנים סביבתיים. כלומר, בהוויוריסטית לא תגיד שמצבים נפשיים **הם** ההתנהגות עצמה, אלא **הם הנטיה** להתנהג אופן מסוים (והנטיה הזו לא חייבת להתממש תמיד, כמו שלכוס זכוכית יש נטיה להיות שבירה, אך היא לא תשבר אם לא יתקיימו תנאים מסוימים, ויתכן שלא תשבר אף פעם).

סוגי בהוויוריזם

- **בהוויוריזם פילוסופי (לוגי)** – טוענת שמצבים נפשיים הם לא מצבי מוח (כפי שטוען המטריאליזם) או "רוחות במכונה" (כפי שטוען הדואליזם), אלא דיספוזיציות (נטיות) התנהגותיות תחת תנאים מסוימים. אם ננסה לבצע ניתוח לשוני של מונחים דיספוזיציוניים (כמו "גמיש", "מסיס", "שביר"), לרוב נעשה את הניסוח באמצעות משפטי תנאי ("אם הכוס תיפול על הרצפה"- תנאי, "אז הכוס תשבר" – תגובה). הרישא של המשפט תהיה הגירוי, והסיפא תהיה התגובה ההתנהגותית. הביקורת על הבוהויוריזם הפילוסופי:
  - הבהוויוריזם הפילוסופי מתאים יותר לתיאור של "מצבים אינטנציונליים" כמו אמונות, תקוות, כוונות, אך לא לחוויות תחושתיות (כמו כאב).
  - ניתוח של מצב נפשי חייב לציין את הקשר בינו לבין מצבים נפשיים אחרים, ולכן לא ניתן להפטר בניתוח מהמונחים הפסיכולוגיים. לדוגמא, אם שואלים את יוסי "אתה רוצה לאכול גלידה?" ויוסי עונה שלא, זה לא בהכרח אומר שהוא לא רוצה לאכול גלידה – יכול להיות שהוא מאוד רוצה, אבל – הוא מפחד להשמין / הוא לא רוצה לשבוע לפני ארוחת הערב וכו'. לכן למרות שהתקיימו התנאים שגרמו לו לרצות גלידה, הוא ההתנהגות שלו לא תבטא

את הרצון הזה. הבעיה היא שהתנאים החדשים שהשפיעו על ההתנהגות (האמירה שהוא לא רוצה גלידה) הם תנאים נפשיים (מפחד להשמין, לא רוצה לשבוע), כלומר אנחנו מנתחים מונח מנטלי באמצעות מונח מנטלי אחר, ונוצרת הגדרה מעגלית חסרת משמעות.

- **בהוויוריזם אפיסטמי** – לפי גישה זו, כל הבדל במצבים נפשיים מתבטא בהבדל התנהגותי או סביבתי. כלומר, הגישה פחות עוסקת בשאלה "מהם מצבים נפשיים", ויותר עוסקת בכך שאנשים אחרים מחזיקים במצבים נפשיים שלכאורה נסתרים מהעין של צופה מהצד. האפיסטמולוגים יטענו כי מצבים נפשיים "חבויים" אינם קיימים, וכי לכל מצב נפשי קיים ביטוי התנהגותי הניתן לצפייה מהצד. הטענה היא לא שניתן לצפות דרך ההתנהגות בכל מצב נפשי או מחשבה, אלא שתחת תנאים מסוימים, צפיה בהתנהגות עשויה להעיד על מצבים נפשיים.
- **בהוויוריזם אלימיניטיביסטי** – סוג של מטריאליזם אלימיניטיביסטי, כלומר כולל הכחשה של עובדות נפשיות. כלומר, עובדות נפשיות אינן קיימות, רק עובדות התנהגותיות קיימות, ולכן פילוסופיה ופסיכולוגיה צריכים להתבסס על מונחים התנהגותיים בלבד (גירויים ותגובות).
- **גישה פרשנית ומערכות אינטנציונליות** – לפי גישה זו, טבעם של מצבים נפשיים קשור באופן הדוק לתהליך פרשנות, כלומר לאופן שבו אנחנו מייחסים מצבים נפשיים לאחרים. התהליך הזה מבוסס על ייחוס מחשבות, האמנות, מטרות, תקוות, רצונות, רגשות וכו' לאחרים, על בסיס האופן שבו הם מדברים ומתנהגים וההקשר הסביבתי שבו הם פועלים. כלומר, המצבים הנפשיים לא **זהים** להתנהגות, אבל אפשר ללמוד על המצבים הנפשיים מתוך הפרשנות של ההתנהגות. הפילוסוף דנט Dennet (1971) טען שהתפקיד החשוב ביותר של מצבים נפשיים הוא תפקיד תיאורטי – הם משמשים בתור הסבר ותחזית להתנהגות. דנט קורא למערכות קוגניטיביות "מערכות אינטנציונליות", שיכולות לחזות או להסביר התנהגות תוך הסתמכות על מערכת של אמונות, פחדים, רצונות וכו'. דנט נותן כדוגמא משחק שחמט למול מחשב מתוחכם – אנחנו עשויים לייחס למחשב רצונות (הוא ירצה להזיז את החייל שלו לשם) ושאיפות (הוא שואף לנצח במשחק), ואז לשער מה הצעד הבא שהמחשב יעשה על סמך התפיסה שלנו של ה"מחשבות" האלו. בדוגמא הזו, אנחנו מתייחסים למחשב בתור מערכת אינטנציונלית – כלומר אין הבדל בין שיוך של אמנות ורציות למכונה מאשר לבני אדם, שניהם הן מערכות אינטנציונליות. כלומר, אנחנו משערים את ההתנהגות העתידית שלהם לפי התפיסה שלנו של המצבים הנפשיים שלהם. **הגישה האינסטרומנטליסטית** אומרת שהעובדה שאנחנו משייכים למחשב שמשחק שחמט ולאנשים רגשות ומחשבות באותה הצורה, והידיעה שלמחשבים שמשחקים שחמט אין באמת רצונות ומחשבות, עשויה להעיד על כך שגם לאנשים אין באמת רצונות ומחשבות. על כך עונה דנט – "להטיל ספק בנודע להאם מחשבים המשחקים שחמט באמת מחזיקים באמונות ורצונות היא לא במקום – שכן ההגדרה שנתתי למערכות אינטנציונליות לא אומרת שבאמת יש לאותן מערכות אמונות ורצונות". כלומר, הוא אומר שלא בהכרח יש לאנשים אחרים רגשות ומחשבות, אבל העובדה שאנחנו **מייחסים** להם את הרצונות והמחשבות האלו מאפשרת לנו לחזות את ההתנהגות שלהם.

הדומה בין אלימיניטיביזם לאינסטרומנטליזם – שתי הגישות לא מחויבות לקיומן של עובדות נפשיות – זהו אלמנט של אנטי-ריאליזם שלא קיים בגישות האחרות שהזכרנו. האלימיניטיביסט יטען שאפשר להתעלם מהעובדות הנפשיות ולא צריך לעשות להן רדוקציה לנטיות התנהגותיות, אלא לעסוק בנטיות ההתנהגותיות עצמן במקום לעסוק בעובדות הנפשיות. האינסטומנטליסט יטען שלמערכות אינטנציונליות לא בהכרח יש רצונות וכוונות, אלא רק שאנשים אחרים משייכים להם כאלו. כלומר – שתי הגישות מסכימות שעובדות נפשיות לא בהכרח קיימות.

השונה בין אלימיניטיביזם לאינסטרומנטליזם – האלימיניטיביסט יתייחס בשלילה לעובדות נפשיות (יגיד שעובדות נפשיות לא עוזרות לנו בשום צורה, גם לא לטובת הסברים מדעיים), בעוד האינסטרומנטליסט מתייחס בחיוב לעובדות נפשיות (ההנחה של קיומן מאפשרת לנו לחזות התנהגות של מערכות מורכבות).

מלבד הגישות הפילוסופיות לבהוויוריזם, יש גם **גישה פסיכולוגית לבהוויוריזם** בה נתמקד יותר מאוחר בקורס.

גילברט רייל – DECARTES' MYTH

פילוסוף מרכזי במאה ה-20 שמייצג זרם שהיה מאוד נפוץ באותן שנים, לפיו בעיות פילוסופיות נוצרות פעמים רבות מעיוות של השימוש היומיומי שלנו במילים (טעויות קטגוריות). כלומר, הטענה לפיה בעיות פילוסופיות שפילוסופים עוסקים בהן אינן "בעיות אמיתיות", אלא תוצרים של בלבול לשוני. לדוגמא, טענות אי-הקיום הבאות לא יכולות להיות אמיתיות – הן או חסרות משמעות או שקריות: "פגסוס אינו קיים", "אין מלך נוכחי לצרפת". הבעיה היא שעל מנת שלפסוק יהיה ערך אמת, השמות "פגסוס" או "המלך הנוכחי של צרפת" צריכים לציין אובייקט כלשהו, שחייב להיות קיים. אם פגסוס קיים, הטענה "פגסוס אינו קיים" שקרית. אם פגסוס לא קיים, הטענה "פגסוס אינו קיים" חסרת משמעות (שכן המושג "פגסוס" אינו מושג אמיתי).

ראסל נותן פתרון לבעיה הזו, בדמותן של טעויות קטגוריות – ומציע שקיום אינו תכונה של אובייקטים, אלא תכונה מבנית של תכונות. טעות קטגורית היא טעות בה אנחנו לוקחים מונח ומשייכים אותו לקטגוריה לא מתאימה. בדוגמא של טענות אי-הקיום, "קיים" הוא לא תכונה של אובייקטים (כמו "אדום" או "חזק"), אלא (אם נחשוב על הטענה הזאת בתחשיב הפרדיקטים) תכונה מסדר שני, שמסומנת באמצעות כמת ולא בתור פרדיקט: ~Ex (כאשר x = פגסוס, E = התכונה "קיים"), אבל מלוגיקה אנחנו יודעים שהמבנה הלוגי העמוק הנכון של הפסוקים הוא - ~∃x(Px) (כש-P = התכונה "להיות פגסוס"). כלומר, "קיום" זה משהו שמתאר **תכונות של אובייקטים** ולא **את האובייקטים עצמם,** לכן זו טעות קטגורית להשמיש אותו על אובייקטים (כמו להגיד "לא קיים פגסוס").

תיאור של אחת הדוגמאות שמספק רייל על מנת להסביר מהי טעות קטגורית: רייל מתאר דוגמה של אזרח זר המבקר באוניברסיטת קיימברידג' בפעם הראשונה. עורכים לאורח סיור ברחבי המתחם של קיימברידג', ומראים לו את ביתות הלימוד, משרדי ההנהלה, המוזיאונים, המעבדות ומגרשי הספורט. לאחר שהסיור מסתיים, שואל האורח "היכן נמצאת האוניברסיטה?". היה צריך להסביר לו כי האוניברסיטה היא לא עוד מוסד או בניין מאותו הסוג כמו ביתות הלימוד והמעבדות שראה, אלא מונח המתייחס למושג מופשט המתייחס למוסד המורכב מכל אותם בניינים, מעבדות, סטודנטים ותוצרי המחקר שראה. הטעות הקטגורית של האורח היתה בהנחה התמימה שלו שזה נכון לדבר על "ספריין בולדין" ו"מוזיאון אשמולן" שנמצאים בשטח האוניברסיטה באותו אופן שבו מדברים על "האוניברסיטה" עצמה, כאילו האוניברסיטה היא אותו סוג של מונח כמו המוזיאון או הספריה. מקור הטעות הקטגורית היתה בשהאורח שייך את האוניברסיטה עצמה לאותה הקטגוריה אליה שייך את שאר המוסדות הנ"ל.

לטענתו של רייל, הטעות הקטגורית שנעשית בנוגע למצבים מנטליים היא שיוכם לאותה הקטגוריה כמו מצבים פיזיולוגים. רייל טוען כי כיוון שמצבים מנטליים אינם יכולים להיות מתוארים באמצעות אותם תיאורים פיזיקליים, כימיים ופיזיולוגיים המשמשים לתיאור מצבים פיזיים בגוף האדם, הרי שלא ניתן לשייך אותם לאותה הקטגוריה כמו מצבים פיזיים. לכן, לטענתו, כיוון שהגוף האנושי הוא מערכת מורכבת ומאורגנת, גם השכל (mind) האנושי צריך להיות מערכת מורכבת ומסודרת אחרת, המורכבת מסוג אחר של מרכיבים ומאורגנת במבנה שונה. כלומר, לא ניתן לשייך מצבים מנטליים לאותה הקטגוריה כמו מצבים גופניים.

כלומר, רייל טוען שדואליזם הוא תוצאה של טעות קטגורית. כלומר, אנחנו מנסים לתאר מצבים נפשיים במונחים של כימיה, פיזיקה וכו', וכיוון שאנחנו לא מצליחים לעשות זאת (נגיד לא מצליחים להגיד שמצבים נפשיים הם עצבי C מגורים), אנחנו מסיקים שהם שייכים לקטגוריה הנפרדת מהקטגוריה אליה שייכים דברים בעולם החומרי. רייל אומר שזו (שמצבים נפשיים שייכים לקטגוריה נפרדת) אינה בהכרח האפשרות היחידה, אלא שאפשר לשקול להתייחס אליהן כחלק מהעולם החומרי, רק צריך להפסיק לנסות לשייך אותן לקטגוריות

כמו "סיבות", "אובייקטים", "מצבים", "תהליכים" וכו'. העמדה הבהוויוריסטית מנסה לענות על השאלה "אז לאיזו קטגוריה בעולם החומרי מצבים נפשיים כן משתייכים".

שיעור 5 – פונקציונליזם וריבוי מימושים

פונקציונליזם היא עמדה שמתכתבת עם המטריאליזם, שמהווה אלטרנטיבה למטריאליזם המוחי והפילוסופי. לפי העמדה הפונקציונליסטית, מה שמשנה לטבען של תכונות קוגניטיביות / מנטליות, הוא המבנה הארגוני / פונקציונלי (מערך הקשרים הסיבתיים) שמקשר בין המצבים המנטליים השונים. יש כמה גישות פונקציונליסטיות, אנחנו נתמקד בפונקציונליזם חישובי.

לפי הפונקציונליזם, סוגים של מצבים נפשיים מוגדרים לפי התפקיד שהם ממלאים בחייו של האדם. תפקיד זה מאופיין ע"י הקשרים הסיבתיים של המצב עם גירויים חושיים, תגובות מוטוריות, ועם מצבים נפשיים אחרים. ניתן לאפיין את מערך הקשרים הפנימיים האלו (כלומר הקשרים בין מצבים נפשיים לבין עצמם) במונחים לא נפשיים (למשל במונחים חישוביים), ומכאן שהאפיון איננו מעגלי. פונקציונליזם "הולך" בד"כ עם ריבוי מימושים, והוא עקבי עם פיזיקליזם לא-רדוקטיבי.

מהו פונקציונליזם

מאפורות המדגימות את הגישה הפונקציונליסטית:

- **מטאפורת "המוח מסיליקון"** – נניח שיכולנו להחליף אחד-אחד את הנוירונים במוח של אדם בנוירונים מסיליקון בעלי תפקוד זהה, והיינו מחליפים את כל הנוירונים שלו בנוירונים מסיליקון, ומקבלים ממש מוח מסיליקון. הטענה הפונקציונליסטית היא שבמקרה כזה תהיה לאדם בדיוק אותה הקוגניציה כמו הקוגניציה המקורית שהיתה לו לפני החלפת הנוירונים, כי מה שמשנה הוא לא הנוירונים עצמם, אלא המבנה – האופן בו הם מקושרים זה לזה (והמבנה לא השתנה כתוצאה מההחלפה).
- **האנלוגיה בין קוגניציה/מוח לתוכנה/חומרה**. ההתיחסות לתוכנה/קוגניציה בתור המבנה הארגוני, שממומש באמצעות החומרה/מוח.

פונקציונליזם לפי אריסטו – אריסטו שם לב שישנן תכונות שמוגדרות ע"י המרכיבים החומריים (אטומיים, מולקולאריים) שלהם – כמו נמר, מים וכו'. אך ישנן תכונות המוגדרות ע"י התפקיד (פונקציה) שהן ממלאות, כמו אמצעים מוניטאריים, לב, קיבה. אנחנו מגדירים את התכונות ע"י התפקיד שהן ממלאות במערכת גדולה ויתר (מערכת כלכלית, גוף האדם) ולא ע"י המבנה החומרי שלהן.

בהמשך (יותר מ-2,000 שנה לאחר אריסטו), פונקציונליזם הוגדר כמייצג את הטענה לפיה מצבים נפשיים מוגדרים ע"י תפקידם במערכת הקוגניטיבית. כלומר, מצבים נפשיים מוגדרים ע"י הקשרים (הסיבתיים) שלהם עם גירויים, עם תגובות, ועם מצבים נפשיים אחרים.

ההבדל למול גישות אחרות

- מטריאליזם מגדיר עובדות נפשיות באמצעות עובדות חומריות (שאינן נפשיות)
- בהוויוריזם מגדיר עובדות נפשיות באמצעות עובדות התנהגותיות (שאינן נפשיות)
- פונקציונליזם מגדיר עובדות נפשיות ע"י עובדות חומריות, עובדות התנהגותיות **וגם עובדות נפשיות**, כלומר מדובר בהגדרה מעגלית.

הטענה הפונקציונליסטית

נתאר את מערך הקשרים הנפשי ע"י הנוסחא הבאה: $F(S_1, \dots, S_{n,} I_1, \dots I_k, O_1, \dots O_l)$ כאשר $S_1, \dots, S_{n,}$ הם מצבים נפשיים, $I_1, \dots I_k$ הם קלטים חושיים, ו-$O_1, \dots O_l$ הם פלטים מוטוריים. מה ש-F עושה הוא לתאר את הקשרים הסיבתיים בין הגורמים השונים המוזנים לה.

נניח לדוגמא כי כאב הוא המצב החמישי S_5, כלומר מוגדר ע"י הקשרים שלו למצבים נפשיים אחרים ("אני שונא שכואב לי"), גירויים ("דקירת מחט") ותגובות (זעקת "איי") . פונקציונליסט יגדיר מערכת קוגניטיבית בתור מערכת עם n מצבים נפשיים, הנמצאים בקשרים עם גירויים, תגובות, ובינם לבין עצמם. כלומר, המצבים הנפשיים מוגדרים כמשתנים - $\exists S_1, \exists S_2, \dots, \exists S_n \mid F(S_1, \dots, S_{n,} I_1, \dots I_k, O_1, \dots O_l) \wedge S_1, \exists S_2, \dots, \exists S_n \in S_5$. כלומר, כל מערכת שיש לה מצבים $S_1, \dots, S_{n,}$ הנמצאים בקשרים עם קלטים מסוימים ופלטים מסוימים, היא מערכת קוגניטיבית. בדוגמא של כאב, נגדיר לדוגמא כי כאב הוא המצב החמישי במערכת המצבים הזו. כל מערכת אחרת, שיש לה n מצבים הנמצאים בקשרים עם גירויים ותגובות, היא מערכת קוגניטיבית, והמצב החמישי בה יהיה כאב. אז זו ההגדרה הפונקציונליסטית לקוגניציה ולעובדות נפשיות.

טיעונים בעד הפונקציונליזם

הטיעון של פטנאם בעד פונקציונליזם – זהו טיעון יחסי – הוא אמר שהעמדה הפונקציונליסטית יותר סבירה ביחס לעמדות האחרות שהוצעו. פטנאם טען כי פונקציונליזם הוא יותר סביר מבהוויוריזם, שכן הוא מתייחס גם לקשרים בין מצבים נפשיים, ולא קושר מצבים נפשיים רק לגירויים ותגובות. הוא טען גם שפונקציונליזם הוא יותר סביר ממטריאליזם מוחי, בעקבות **טענת ריבוי המימושים** – כלומר שסביר שניתן לממש את אותה המערכת הקוגניטיבית ב"מצעים" ביולוגים ופיזיקליים שונים (ואולי אפילו במצעים שאינם ביולוגיים, כמו מכונה). כלומר, טענת ריבוי המימושים טוענת שקיימת אפשרות שאותה תכונה קוגניטיבית תתממש במצבים פיזיקליים שונים. הכאב שלי ממומש בעצבי C מגורים, אך זה לא אומר שזו הדרך היחידה בה כאב יכול להיות ממומש – יתכן לדוגמה כי הכאב של יצורים לא ביולוגיים ממומש באמצעות מצבים לא פיזיולוגים.

פונקציונליזם כעמדה פיזיקליסטית

פונקציונליסטים רואים עצמם כבעלי עמדה מטריאליסטית, ומאמינים שכל העובדות נקבעות ע"י עובדות פיזיקליות. זה עקבי עם פיזיקליזם לא רדוקטיבי – לפיו עובדות נפשיות נקבעות ע"י (אך אינן בעצמן) עובדות פיזיקליות.

נחדד את ההבדל בין פיזיקליזם רדוקטיבי לפיזיקליזם לא רדוקטיבי – **פיזיקליזם רדוקטיבי** גורס שכל תכונה (עובדה) קוגניטיבית M היא תכונה פיזיקלית P, ושקיים "חוק גשר" מהצורה $\forall x(Mx \leftrightarrow Px)$. לעומת זאת, **פיזיקליזם לא רדוקטיבי** מקבל את טענת הנסמכות – שתי ישויות שהן בלתי נבדקות פיזיקלית הן גם בלתי נבדלות קוגניטיבית $\forall x(Px \rightarrow Mx)$, ואת טענת ריבוי המימושים: שתי ישויות שהן בלתי נבדלות קוגניטיבית עשויות להיות נבדלות פיזיקלית $\sim\forall x(Mx \rightarrow Px)$.

ביקורות על פונקציונליזם

המתודה היא להראות שעובדות נפשיות (כמו מודעות או אינטנציונליות) אינן פונקציונליות. כדי להפריך את הטענה הפונקציונליסטית אנחנו צריכים להראות אחד משני דברים:

1) שאותן תכונות נפשיות מתלכדות עם תכונות או מבנים פונקציונליים שונים.
2) שאותו מבנה פונקציונלי מתלכד עם תכונות נפשיות שונות (זו השיטה שנקטו בה רוב הביקורות). כלומר גם אם יצרנו מוח מסיליקון עם מבנה זהה למוח אנושי, עשויות להיות לו תכונות נפשיות שונות. גם בתוך הטיעונים האלו יש שתי גישות:
   a. **הביקורות ממודעות** – אותו מבנה פונקציונלי יכול לממש שתי תכונות פנומנאליות שונות. לדוגמא טיעון החדר הסיני (נגיע אליו בהמשך הקורס).

b. **הביקורות מאינטנציונליות** – אותו מבנה פונקציונלי יכול לממש שתי תכונות אינטנציונליות (ייצוגיות) שונות. תוכן מנטלי תלוי בגורמים סביבתיים. כלומר, הטענה היא שפונקציונליזם ממקם את התכונות הקוגניטיביות שלנו "בתוך הראש" שלנו. **לדוגמא טיעון ה-Twin Earth של פטנאם** – אם אקח את התאום הזהה שלי, ונעביר אותו לסביבה אחרת ("כדוה"א מקביל") בו מים הם חומר עם תכונות שונות מאוד. פטנאם טוען שהמחשבות של התאום שלי שנוגעות למים יהיו מאוד שונות מהמחשבות שלי. הטיעון הזה נוגע להבדל שבין **אקסטרנליזם לבין אינטרנליזם**: אינטרנליזם הן עמדות הממקמות את התכונות הקוגניטיביות בתוך הראש שלנו (הפונקציונליזם, לדוגמא, הוא עמדה אינטרנליסטית). אקסטרנליזם הן עמדות לפיהן לפחות חלק מהמשתנים שקובעים את התכונות הקוגניטיביות שלנו נמצאות בסביבה שלנו (מחוץ לנו). כלומר לפי גישות אלו אם נשנה את הסביבה מבלי לשנות שום דבר פנימי, עדיין נשפיע על תכונות קוגניטיביות. לדוגמה – המטריקס, שבו התכונות הקוגניטיביות (מחשבות וכו') אינן אלא גירויים שהסביבה מכניסה מבחוץ.

תשובת הפונקציונליסט לביקורות – הפונקציונליסט יטען שניתן להגדיר את הקלטים/פלטים בסביבה של האורגניזם, כלומר נגדיר את התכונה ביחס לגירוי ולא באופן בלתי תלוי. טענה נוספת (ומעניינת ונפוצה יותר) היא שעדיין יתכן שחלק מהמצב הנפשי (להיות בעל קוגניציה) נקבע פונקציונלית, ואילו התוכן הספציפי נקבע באופן אחר.

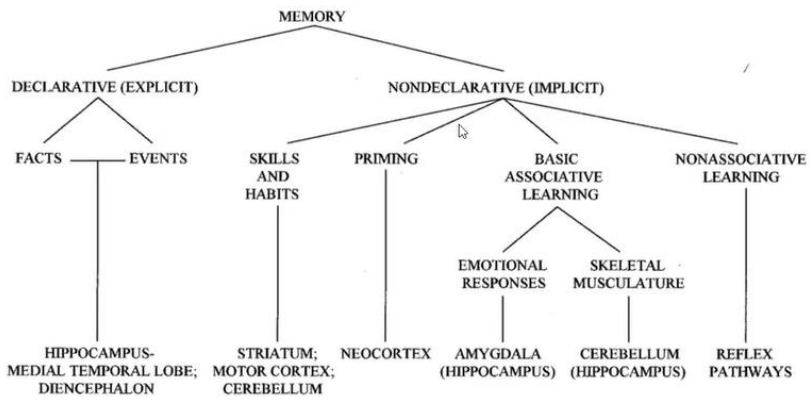
האם יתכן ריבוי מימושים?

פולגר ושאפירו ב-2016 טענו שהטענה של ריבוי מימושים היא מעניינת, אך כמעט ולא קיימת לה תמיכה אמפירית, וכן מבחינה קונספטואלית הרבה יותר קשה לתמוך בטיעון הזה ממה שאנחנו חושבים. כדי להראות שתכונות קוגניטיביות ניתנות לריבוי מימושים אנחנו צריכים להרעות ששתי המערכות זהות מבחינה קוגניטיבית, וששתי המערכות שונות מבחינת הפונקציה הנוירולוגית הרלוונטית (כלומר התכונות הפיזיולוגיות שמממשות את התכונה הקוגניטיבית).

דוגמא – חולץ הפקקים וריבוי מימושים: התכונה "חולץ פקקים" היא תכונה שאנחנו יכולים להגיד שהיא מרובת מימושים (ראו בתמונה מימושים שונים לתכונת "להיות חולץ פקקים" עם ביטויים פיזיקליים שונים). פולגר ושפירו אומרים שהשאלה היא איך מגדירים מהם מימושים שונים – אם היו לנו שני חולצי פקקים מאותו הדגם רק בצבעים שונים, לא היינו אומרים שאלו מימושים שונים, אלא שמדובר באותו המימוש (כי האופן שבו הם חולצים פקקים הוא זהה, והצבע לא משפיע על הפעולה הזו).

עדויות אמפיריות בעד ריבוי מימושים:

- פלסטיות המוח – מימוש של תכונה קוגניטיבית אחת באמצעות מבנים מוחיים שונים. המקרה הקלאסי הוא פגועי מוח שמצליחים לשחזר יכולות קוגניטיביות תוך שימוש באזורים חדשים שלא שימשו לתכונה הזו קודם.
  - התשובה של פולגאר ושאפירו לעדות זו – לא ברור שהפעילות העצבית שונה. מיקום שונה ברחבי המוח אינו מראה שהפעילות העצבין שונה (כי המיקום אינו פונקציה נוירולוגית רלוונטית – כמו שהצבע לא רלוונטי לחליצת פקקים). כמו כן, לא ברור שהפונקציה הקוגניטיבית הממומשת היא זהה.
- Kind-splitting – מקרים בהם אותו תוכן קוגניטיבי ממומש באזורים שונים במוח. לדוגמא, הנה דיאגרמה של סוגים שונים של זכרון ואיפה הם ממומשים במוח:



אפשר לטעון כי זוהי דוגמה לריבוי מימושים – יש לנו סוג קוגניטיבי (זכרון) והוא ממומש באופנים שונים.

- התשובה של פולגאר ושאפירו לעדות זו – זה נכון שזכרון ממומש באזורים שונים במוח, אך זה לא אומר שבאזורים השונים אין תכונה נוירולוגית רלוונטית שהיא משותפת לכל האזורים השונים וזהה בכל המימושים (למרות שהמיקום במוח שונה).

כלומר, ניתן לראות שתמיד יתקיים ויכוח לגבי האם ניתן להראות אמפירית שמתקיים ריבוי מימושים, כי תמיד ניתן יהיה לטעון כי לא מדובר במימושים שונים של אותה פונקציה קוגניטיבית אלא מימושים של פונקציות קוגניטיביות שונות (לדוג' זכרון אקספליציטי לעומת אימפליציטי). כמו כן, העובדה שתכונות שונות ממומשות באזורים שונים במוח היא לא מידע מספיק כשלעצמו על מנת לקבוע שהמימוש שלהם שונה.

מקרים מורכבים – פולגאר ושאפירו מצביעים על מקרים מורכבים, בהם יש שוני מסוים בפונקציה הנוירולוגית הרלוונטית הממומשת, וגם יש זהות מסוימת בפונקציה הקוגניטיבית הממומשת. במקרים כאלו, עדיין צריך לבדוק שהשוני בפונקציה הנוירולוגית אינו מתבטא בשוני (מסוים) בפונקציה הקוגניטיבית. כלומר, צריך לראות אם יש קורלציה בין השוני הקוגניטיבי לשוני הנוירולוגי.

ריבוי מימושים וקוגניציה כחישוב – יש פילוסופים שטוענים שאם קוגניציה מוגדרת ע"י המימוש החישובי שלה, ואנחנו יודעים שאנחנו יכולים לממש את אותה התוכנה בחומרות השונות אחת מהשניה – זו עדות טובה לקיום של ריבוי מימושים. המסקנה של פטנאם היתה ש"אנחנו יכולים להיות עשויים מגבינה שוויצרית וזה לא ישנה".

- ביקורות על הטיעון החישובי – צריך לשאול את עצמנו "האם ניתן לממש את מערכת הקשרים הסיבתיים של המערכת הקוגניטיבית שלנו בגבינה שוויצרית באמת?" – כלומר, נראה שפטנאם מניח מושג חלש מידי של מימוש. עם זאת, זה לא סותר את העובדה שיתכן שאפשר לממש מצבים נפשיים בכל מיני מצבים ביולוגיים (ואולי אף בחומרה מלאכותית). גם אם אפשר ליצור מוח מסיליקון שיהיו לו אותן תכונות קוגניטיביות, זה עדיין לא אומר שאין תכונה פיזיקלית רלוונטית המשותפת לכל המערכות הממומשות (המוח הביולוגי למול המוח מסיליקון), שהיא זו שקובעת את התכונה הקוגניטיבית.

שאלת ריבוי המימושים נשארת פתוחה – העדויות האמפיריות אינן מכריעות לכאן או לכאן, והנימוקים הפילוסופיים אינם חד משמעיים.

שיעור 6 – מבחן טיורינג

טיורינג מתאר את "משחק החיקוי" באופן הבא: במשחק משתתפים שלושה אנשים: איש (המכונה A), אישה (המכונה B) וחוקר/ת (המכונים C). האיש והאישה נמצאים בחדר נפרד מהחוקר/ת, והתקשורת בינם לבין החוקר/ת מתבצעת בכתב בלבד. מטרת החוקר/ת היא לזהות מי מבין שני המשתתפים האחרים הוא האיש, ומי היא האישה. החוקר/ת מעביר/ה למשתתפים האחרים שאלות, והם עונים עליהם כרצונם. מטרתו של האיש (A) היא להטעות את החוקר/ת, כלומר לגרום להם לבחור בו בתור האישה. מטרתה של האישה (B) היא לעזור לחוקר/ת, כלומר לגרום להם לבחור בה בתור האישה.

טיורינג מעלה את השאלה – מה יקרה אם נחליף את שחקן A במכונה? האם החוקר/ת יטעו בזיהוי באותה מידה בה היו טועים לו היו משתתפים במשחק איש ואישה? טיורנג מציע כי השאלות האלו עשויות להחליף את השאלה "האם מכונות חושבות" בשיח על אינטליגנציה מלאכותית.

הטיעון של טיורינג במאמר:

- **הנחה ראשונה:** למחשב דיגיטלי (אוניברסלי) יש את היכולת לעבור את "משחק החיקוי" (מבחן טיורינג). טיורינג אמר "אני מאמין שבעוד בערך 50 שנה (נכתב לפני כ-70 שנה מהיום) זה יהיה אפשרי ליצור מחשבים..... שיוכלו לשחק את משחק החיקוי כה טוב, כך שלחוקר/ת הממוצע/ת לא יהיו יותר מ-70% הצלחה בזיהוי כי נכון של המכונה אחרי 5 דקות של תשאול".
- **הנחה שניה:** הצלחה במבחן טיורינג מעידה על חשיבה (אינטליגנציה, מנטאליות).

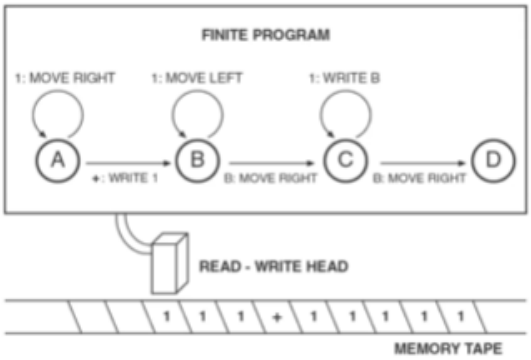
<u>מסקנה:</u> למחשב יש את היכולת לחשוב (בינה מלאכותית).

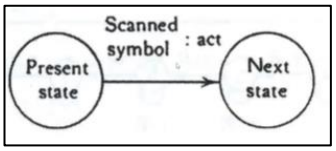
מבחן טיורינג הוא מבחן **התנהגותי**, שמטרתו לחקות התנהגות אינטליגנטית (ובאופן ספציפי – שיחה). נאמר שמכונה עברה את מבחן טיורינג אם החוקר/ת לא הצליח/ה להבחין בין המכונה לבין האדם.

מכונת טיורינג

מכונת טיורינג היא בעלת שלושה חלקים:

1) **תוכנית סופית** – מספר פקודות.
2) **סרט זכרון** – שם מאוחסן המידע, הוא סופי אך ניתן להרחבה כרצוננו.
3) **ראש קריאה-כתיבה** – קורא ורושם סימנים על סרט הזכרון. בכל שלב ראש הקריאה-כתיבה נמצא מעל תא מסוים, קורא את התא, ומבצע פעולה לפי התוכנית.





פעולה של מכונת טיורינג: בכל פעולה מתחלקת ל-4: קיומו של מצב נוכחי, סריקת הסמל על סרט הזכרון, על בסיס התו שנקרא והמצב הנוכחי מתבצעת פעולה, ומתבצע מעבר לקיומו של המצב הבא.

הפעולות האפשריות שמכונת טיורינג יכולה לבצע:

- למחקו תו מהתא הנסרק בסרט הזכרון
- לכתוב תו לתא הנסרק בסרט הזכרון
- להזיז את ראש הקריאה-כתיבה תא אחד ימינה
- להזיז את ראש הקריאה-כתיבה תא אחד שמאלה
- לעצור את החישוב

טיורינג טען שלכל פעולת חישוב אפשרית, ניתן ליצור מכונת טיורינג (כפי שהוגדרה מעלה) שתוכל לבצע את החישוב הזה (מה שנקרא בימינו התחום של אלגוריתמיקה). הוא העלה את הרעיון של **מכונה אוניברסלית**, וכתב "היכולת הייחודית של מחשבים דיגיטליים, היא שהם מחקים כל פעולת חישוב העובדת לפי מצבים בדידים, מתוארת ע"י כינוי שלהם בשם "מכונות אוניברסליות". לקיום של מכונות עם היכולת הזו יש השלכה חשובה – נניח את מהירות החישוב בצד, ואז נוכל להגיד שזה מיותר לתכנן מכונות שונות שכל אחת מהן תעשה פעולת חישוב אחרת, כי כולן יכולות להיות מבוצעות ע"י אותו מחשב דיגיטלי שניתן לתכנת אותו להתמודד עם כל מקרה.".

האם כל פונקציה ניתנת לחישוב?

לפי טיורינג, התשובה היא לא. מספר הפונקציות מ-$\mathbb{N}$ ל-$\mathbb{N}$ הוא C(בלומר נמצא בסדר גודל של המספרים הממשיים), בעוד שמספר הפונקציות הניתנות לחישוב ע"י מכונת טיורינג הוא $\aleph_0$ (שקטן מ-C משמעותית). טיורינג הזכיר במאמר שלו דוגמא לבעיה שאינה ניתנת לחישוב במכונת טיורינג – **בעיית העצירה**. בעיה זו היא: בהנתן מכונת טירוינג א' וקלט כלשהו, האם ישנה מכונת טיורינג ב' שיכולה לקבוע אם מכונת טיורינג א' הפועלת על הקלט הנתון תעצור או לא? טיורינג הראה שהתשובה לשאלה היא לא, לכן מדובר בדוגמה לבעיה שלא ניתנת לחישוב במכונת טיורינג. כמו כן, טיורינג טען שלא ניתן ליצור מכונת טיורינג אחת שתדע לקבוע עבור כל פסוק בתחשיב היחסים, האם הפסוק הוא בעל ערך אמת.

חשוב לציין שלא כל מכונת טיורינג עוצרת בשלב מסוים. לדוגמא, מכונה שיש לה מצב אחד, וכל פעולה מחזירה אותה לאותו המצב, תמשיך לפעול ללא עצירה באופן מעגלי.

נימוקים בעד ונגד הטענה שמחשב יכול לעבור את מבחן טיורינג

נימוקים בעד הטנה:

- ההתנהגות (הלשונית) שלנו נשלטת ע"י חוקיות סופית כלשהי.
- המוח הוא מערכת פיזיקאלית, ולכן מציית לחוקי הטבע, והם ניתנים לחישוב (בקירוב) ע"י מכונת טיורינג.

נימוקים נגד הטענה:

- עד היום לא מצאנו מכונת חישוב שעוברת (אפילו בקירוב) את מבחן טיורינג.
- החשיבה המתמטית שלנו חורגת מזו של מכונת טיורינג - אנו, בניגוד למחשב, יכולים לקבוע האם מכונה עוצרת או לא.

פירוש ההנחה השניה (שהצלחה במבחן טיורינג מעידה על חשיבה)

- **מבחן טיורינג ובהוויוריזם פילוסופי**: נד בלוק כתב כי "בהוויוריסטים מגדירים מצבים מנטליים לא במונחים של התנהגות, אלא באמצעות נטיות התנהגותיות, שהם הנטייה להתנהג באופן מסוים בהנתן גירוי כלשהו. תכלית שלישית, שהכי מתקרבת לכוונה של טיורינג, היא התכלית של "הבהרה קונספטואלית". אין ספק שטיורינג קיווה שהתפיסה של אינטליגנציה באמצעות מבחן טיורינג תפיק את כל מה שהיית יכול לרצות מתוך ההגדרה של אינטליגנציה, וזאת מבלי להיות מעורפלת כמו ההגדרות שלה שאנחנו מכירים".
- **מבחן טיורינג כמבחן אידוקטיבי**: מור כתב כי "אם מכונה עברה את מבחן טירוינג, אז יהיה לנו בסיס אינדוקטיבי כדי לשייך לה אינטליגנציה או מחשבה. לא תהיה לנו ודאות בשיפוט הזה, ואנחנו עשויים לחזור בנו ממנו בהנתן ראיות חדשות, אך יהיו לנו ראיות מספקות כדי להסיק שהמכונה אכן אינטליגנטית". כלומר, התנהגות נותנת לנו מידע שמאפשר לנו לשפוט (גם אם לא באופן ודאי) האם ישות היא אינטליגנטית.
- **בהוויוריזם אלימיניטיביסטי**: טיורינג בעצמו כתב: "במקום להתיימר לתת הגדרה לאינטליגנציה, אחליף את השאלה "האם מכונות חושבות" בשאלה אחרת – שהיא יחסית קשורה אליה, והיא מובעת במילים יחסית חד משמעיות". זה מתקשר לגישה האלימיניטיביסטית כיוון שהטענה אומרת למעשה שהמונחים בהם הגדרנו אינטליגנציה עד כה היו עמומים מידי ולא מדויקים מספיק, וצריך להחליף אותן בהגדרות המבוססות על התנהגות.
- **גישה רגשית (תלוית תגובה)** – קובעת שקיימת הגדרה סביבתית. כלומר, המבחן בודק מתי החוקר/ת ייחסו לישות כלשהי אינטליגנציה. כלומר, אינטליגנציה מוגדרת ע"י האופן בו אחרים מייחסים את המושג לישות, ומבחן טיורינג קובע את הרף לייחוס שכזה. כלומר, אינטליגנציה לא מוגדרת על ידי תכונה שקיימת אצל הישות עצמה, אלא ע"י תכונה שלכאורה שייכת לישות כפי שהיא נתפסת ע"י גורם חיצוני. זוהי גישה הקרובה לבהוויוריזם הפילוסופי – כלומר, אינטליגנציה היא הנטיה לעבור את מבחן טיורינג.

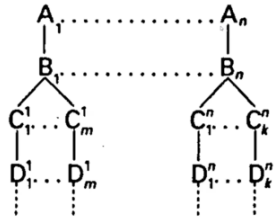
נימוקים נגד ההגדרה של חשיבה כמעבר מבחן טיורינג

- ההגדרה אינה הכרחית – כלומר הרף גבוה מידי, יתכנו ישויות אינטליגנטיות שלא יעברו את המבחן (לדוגמא חיות). כמו כן, המבחן למעשה בוחן יכולת חיקוי (של בני אדם), ועולה התהיה האם לכל יצור אינטליגנטי חייבת להיות יכולת חיקוי .
- ההגדרה אינה מספיקה – כלומר הרף נמוך מידי, כלומר יתכנו ישויות לא אינטליגנטיות שכן יעברו את המבחן. עולה השאלה – האם שיחה מעידה על חשיבה? המבחן למעשה לא בודק סוגי אינטליגנציה אחרים כמו תנועה מוטורית, התמצאות במרחב, יצירתיות אומנותית ועוד. כמו כן, קיימת גם "בעיית החוקר/ת" – הבעיה היא שאנחנו מגדירים את האינטליגנציה של המכונה באמצעות האינטליגנציה של החוקר/ת, כי הם אלו שקובעים אם המכונה אינטליגנטית או לא.

BLOCK'S HEAD

נד בלוק טען כי יש פגם עקרוני בהנחה השניה, כי חשיבה היא יותר מיכולת התנהגותית. ניתן להראות זאת ע"י הצגת מכונה שמצליחה לעבור את מבחן טיורינג למרות שאינה אינטליגנטית. בלוק הציג את המכונה Aunt Bubble (שכונתה בהמשך Block's Head). הנחות היסוד של בלוק בבניית המכונה הזו היו:

- מבחן טיורינג הוא מבחן התנהגותי שאורך זמן סופי נתון (לדוגמא שעה)
- מספר השיחות האפשריות שניתן לקיים באותו זמן סופי נתון, גם הוא סופי (כי יש גבול למספר ההברות שאפשר לבטא בשעה, לכן מספר השיחות האפשריות הוא מספר ההברות השונות שאפשר לבטא במשך השיחה x משך השיחה).
- נסדר את כל השיחות האפשריות בתצורת עץ:



- גם מספר השאלות שניתן לשאול באותו פרק זמן הוא סופי.

לכן, בלוק הציע פשוט להגדיר תשובה מוכנה לכל שאלה אפשרית. בהנתן שאלה של החוקר/ת, המכונה מזהה את השאלה בשורה הראשונה A ומיד מגיבה עם התשובה המתאימה B. בהנתן שאלה נוספת של החוקר/ת, המכונה תזהה אותה בשורה השלישית C, ומיד מגיבה עם התשובה המתאימה D, וכך הלאה. בלוק טוען שהמכונה הזו תעבור את מבחן טיורינג, אך ניכר שהיא אינה חושבת (היא רק שולפת תשובות מוכנות מתוך מאגר).

התנגדויות לבלוק:

- האם Block's Head אכן עובר את מבחן טיורינג?
  - האם היא אכן מכסה את כל השיחות האפשריות? (תשובת בלוק - כן, אם המבחן מוגבל בזמן)
  - האם אפשרי פיזיקלית לכסות את כל השיחות האפשריות? (תשובת בלוק – גם אם זה לא באמת אפשרי פיזיקלי, זה מראה שמה שאנחנו בעצם דורשים זה מאינטליגנציה יותר מאשר שאנחנו דורשים מהתנהגות. כלומר, ההתנהגות לא מספיקה לנו להגדיר אינטליגנציה, ומעניין אותנו עוד דברים מעבר אליה כדי לקבוע אם ישות היא אינטליגנטית).
  - האם מבחן טיורינג צריך להיות מוגבל בזמן?
- למה Block's Head אינה אינטליגנטית?
  - למה טבלה מוכנה אינה אינטליגנציה?
  - אולי גם אנחנו בעצם שולפים תשובות מוכנות מתוך מאגר זכרון?
  - Block's Head מכילה בעצם את האינטליגנציה של מי שתכנת אותה.

מה חסר ל-Block's Head? חסרה לה **גנרטיביות / פרודקטיביות ("יצירתיות")**: היכולת לטפל (לקלוט ולייצר) מספר אינסופי של משפטים תוך שימוש באמצעים סופיים. כלומר, היכולת לטפל בקלטים שהמכונה לא ראתה מעולם.

כשאנחנו חושבים על הקשר בין חשיבה וחישוב, אנחנו יכולים לחשוב על **שני סוגים של שקילות חישובית:**

1) **שקילות התנהגותית (בהוויוריזם)** – שני אובייקטים מחשבים את אותן פונקציות (אך יתכן שעושים זאת בדרכים שונות)

2) **שקילות אלגוריתמית (קוגניטיביזם)** – שני האובייקטים מחשבים את אותן הפונקציות באותם האמצעים (תוכניות, אלגוריתמים). כלומר השקילות היא לא רק ברמת ההתנהגות אלא ברמת האלגוריתם החישובי.

בלוק טען שמה שמשנה הוא לא רק ההתנהגות אלא גם איך ההתנהגות מושגת (התהליך הפנימי). הוא טען ששקילות התנהגותית אינה תנאי מספיק לחשיבה, כי ישנם אובייקטים ששקולים לנו התנהגותית אך אינם חושבים. הוא טען שיתכן ששקילות אלגוריתמית היא כן תנאי מספיק לחשיבה. זה יותר קרוב לעמדה הקוגניטיביסטית, שמזהה אינטליגנציה לא רק עם ההתנהגות אלא גם עם תהליכי עיבוד המידע הפנימיים שעומדים מאחוריה.

שיעור 7 – בינה מלאכותית ומתנגדיה

היפותזת המערכת הסימבולית-פיזיקלית: "המערכת הסימבולית-פיזיקלית היא מאפיין הכרחי ומספיק לקיום פעולה אינטליגנטית" (ניואל וסיימון, 1975).

מהי בינה מלאכותית? יש שתי גישות לבינה מלאכותית: הגישה הראשונה היא להתייחס לבינה מלאכותית בתוך משהו ששייך למערכות שאינן אינטליגנטיות כמונו, אך מציגות משהו שדומה לאינטליגנציה שלנו (כלומר לא מדובר בבינה אמיתית). גישה שניה (זהו מובן חזק יותר לבינה מלאכותית), היא שבינה מלאכותית היא מערכת אינטליגנטית לכל דבר, וההבדל היחיד הוא שהיא מבוססת על "חומרה" מלאכותית ולא ביולוגית. הפילוסוף סרל Searle מתאר – "לפי בינה מלאכותית חזקה, הסימולציה הנכונה היא למעשה המיינד. לפי בינה מלאכותית חלשה, הסימולציה הנכונה היא למעשה מודל של המיינד". כלומר, הוא מפריד בין סימולציה של התנהגות, לעומת סימולציה של המוח/הקוגניציה עצמם.

יש הבדל משמעותי בין העמדה של טיורינג, למול העמדה של ניואל וסיימון. טיורינג קישר בינה מלאכותית עם התנהגות מסוימת (מעבר של מבחן טיורינג), בעוד ניואל וסיימון אמנם מדברים על התנהגות במובן מסוים ("פעולה אינטליגנטית") אך קובעים תנאי נוסף לקיום אינטליגנציה (מערכת לעיבוד סימנים).

מערכות פיזיקליות לעיבוד סימבולים

מהי מערכת פיזיקלית? מערכת העשויה מחומרים פיזיקליים (ביולוגיים או שלא), שהפעולות שלה עקביות עם חוקי הטבע. ההגדרה הזו לא קשורה בהכרח לבני אדם, יתכנו מערכות אינטליגנטיות שאינן אנושיות.

מהי מערכת לעיבוד סימבולים? דוגמה למערכת לעיבוד סימבולים היא מכונת טיורינג שדיברנו עליה בהרצאה הקודמת (זו אולי לא מערכת פיזיקלית אך אפשר כנראה לממש אותה גם באופן פיזיקלי). מערכת לעיבוד סימבולים היא מערכת המקבלת כקלט סימבולים (סימנים) מתוך רשימה דיסקרטית (כלומר ניתן להבחין באופן חד משמעי בין סימן אחד לאחר), עושה פעולות על הסימנים האלו (החלפה, הזזה וכו') לפי תוכנית מוגדרת. יש שלוש תכונות לסימבולים:

- <u>חיים פיזיקאליים:</u> הסימבול ממומש במערכת פיזיקאלית.
- <u>חיים פורמליים / סינטקטיים:</u> הסימבול הוא חלק ממערכת פורמלית. מערכת החוקים המופעלת על הסימנים פועלת על ההיבט הסינטקטי שלה (ולא על ההיבט הסמנטי שלה).
- <u>חיים סמנטיים / יצוגיים:</u> לסימבול יכולה להיות משמעות או הוראה (reference).

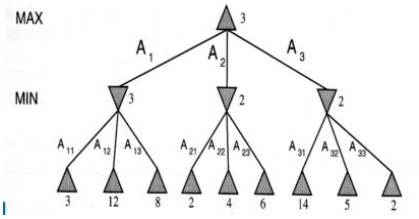
עדויות לטובת היפותזת עיבוד הסימבולים: ההיפותזה טוענת למחשבים דיגיטליים יש את האמצעים ההכרחיים והמספיקים להתנהגות אינטליגנטית. לפי טענה זו, לא רק שמחשבים דיגיטליים יכולים להציג התנהגות אינטליגנטית, **רק** מערכות סימבוליות לעיבוד סימנים יכולות להציג התנהגות אינטליגנטית. אלו העדויות לטובתה:

- עדויות מהעבודה עם **בינה מלאכותית**. נתמקד בתחום של מכונות שח.
- עדויות מתחום **מדעי הקוגניציה** (שמראות שהקוגניציה האנושית / קוגניציה של בע"ח היא למעשה מערכת לעיבוד סימבולים) – נתמקד בזה בהרצאה הבאה.

מכונות שח

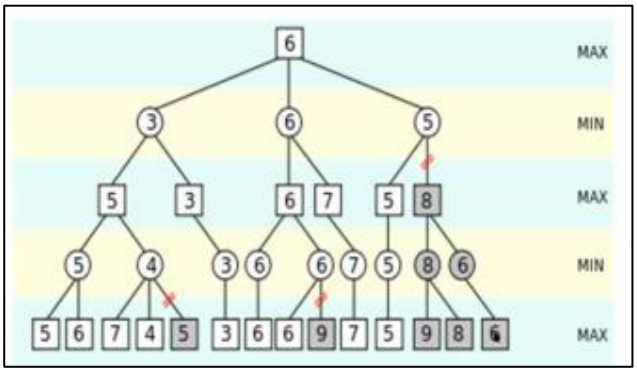
בשנים הראשונות לקיומן של מערכות אינטליגנטיות מלאכותיות, שחמט היה נראה בתור מבחן טוב לבחינת אינטליגנציה. היה שיפור מאוד משמעותי משנות ה-60 ועד שנות ה-90 באיכותן של מכונות שח, כשב-1997 מערכת "כחול עמוק" ניצחה את אלוף העולם האנושי בשחמט. אך לאחר שעברו את המשוכה הזו, עלו קולות שטענו ששחמט אינו מבחן אינטליגנציה מספק, כמו חומסקי שאמר "תוכנת מחשב שמנצחת את אלוף העולם בשחמט היא מעניין כמו בולדוזר שמנצח באולימפיאדה במקצה הרמת משקולות". כלומר, אין למכונה אינטליגניציה גבוהה בהכרח, אלא בעיקר יכולת חישוב.

כיצד פועלות מכונות שח – הן יוצרות עץ חיפוש שכולל את המהלכים העתידיים האפשריים (בראש העץ נמצא מצב הלוח הנוכחי, כל ענף מייצג מסלול משחק אפשרי). הן מעריכות את כל הלוחות בתחתית העץ (העלים) – לוח טוב יקבל ציון גבוה. הן קובעות את המהלך הבא כפונקציה כלשהי של ציוני הלוחות האפשריים מתחתית העץ (לדוגמה באמצעות min-max).



- איך עובד עקרון ה-min-max: יוצרים עץ חיפוש ונותנים ציונים לעלים. נבין שהיריב יעדיף לבחור בלוח עם הציון המינימלי (ציון גרוע עבורינו = ציון טוב עבורו). אז "נוריש" לשכבה שמעל העלים את הציונים הנמוכים ביותר, ונתן למכונה לבחור את הציון הגבוה ביותר מביניהם (בתמונה- A1).

הבעיה – גידול מעריכי. ברוב המשחקים המעניינים (כמו שחמט) מספר הלוחות האפשריים הוא עצום, וגדל בכל רמה באופן מעריכי. אם את כל הלוחות ברמה השניה אפשר להעריך בפחות משניה, הרי שבשביל להעריך את כל הלוחות ברמה העשירית נדרשות כמה שנים. השיפור שנעשה לאורך השנים במכונות השח הוא ביכולת חישוב שלהן (כלומר בקצב החישוב) שמאפשר להן לרדת "יותר עמוק" בעץ האפשרויות. מכונות שח טובות אף פוסלות ענפים חסרי סיכוי כבר בהתחלה ע"י גיזום (pruning), ומאחסנות אסטרטגיות מוכנות למצבים שכיחים.



מאפיינים בולטים של השחקן האנושי: רוב השחקנים האלופים אינם סורקים יותר מהלכים בעץ משחקני השח הבינוניים. ההבדל הוא ששחקנים אלופים מזהים יותר מהר את המהלכים האפשריים המוצלחים יותר (יש להם "אינטואיציה לשח") מבלי להצטרך להתעמק בעץ האפשרויות. האלופים תופסים את הלוח ב-Chunks ומאחסנים בזכרון כמויות גדולות של צ'אנקים כאלה. עדויות לכך:

- הם יכולים לשחזר באופן מושלם לוחות משחק לאחר חשיפה של שניות ספורות, בעוד שחקנים בינוניים לא.
- הם מאחסנים בזכרון כ-10k עד 100k צ'אנקים.

מכונות השח הן דוגמא למערכות שעוברות את "מבחן טיורינג לשח" – הן אולי אינטליגנטיות ביחס לשח, אך החשיבה שלהם שונה מזו של השחקנים האנושיים. כלומר, סוג החשיבה אינו מוגדר רק ע"י התנהגות, אלא גם ע"י התהליך הפנימי. עולה השאלה – האם מכונות שח יכולות לחקות את התהליך הפנימי? היום כבר קיימות מערכות שח לומדות המבוססות על למידת מכונה ורשתות נוירונים.

הביקורת של דרייפוס: "האם מכונות שמשתמשות ב-brute force הן באמת כאלה חכמות? 25 שנים של מחקר אינטליגנציה מלאכותית הצליח לקיים מעט מאוד מההבטחות שלו, ונכשל בנסיון להציג ראיה כלשהי לכך שהוא אי פעם עתיד להצליח לקיים אותן". אחת הטענות של דרייפוס הוא טיעון הידע – בני אדם משתמשים בידע ובהנחות קודמות בשיחות ובפתרון בעיות, ומחשבים מאוד מתקשים עם הקונספט הזה. כמו לדוגמא – אדם ידע להגיד שאם ביל קלינטון נמצא בוושינגטון, אז גם הרגל השמאלית שלו נמצאת בוושינגטון, אבל מחשב לא ידע להסיק זאת ישירות מהידיעה שקלינטון נמצא בוושינגטון.

הביקורת של סרל: סרל מגדיר "אינטליגנציה מלאכותית חזקה" בתור הטענה שמבנה חישובי מכונן ומספיק לקיומם של תכנים מנטאליים. הוא יוצא באופן ישיר נגד הטענה של ניואל וסיימון לפיה עיבוד סימבולים בשלעצמו הוא מספיק לקיומם של תכנים מנטליים. הוא לא יוצא נגד טענת ההכרחיות – כלומר הוא מסכים שיתכן שעיבוד סימבולים עשוי להיות תנאי הכרחי לאינטליגנציה, אך מתעקש שהוא אינו תנאי מספיק – כלומר יש מרכיב נוסף שצריך להתקיים על מנת שתהיה אינטליגנציה. סרל אינו תוקף:

- את השימוש במודלים חישוביים במחקר קוגניטיבי
- את הטענה שאנחנו יכולים ליצור מכונות חושבות שאינן ביולוגיות
- את הטענה שתהליכים קוגניטיביים הם (גם) תהליכים חישוביים
- את הטענה שמבנה חישובי מסוים הכרחי לתכנים מנטאליים

סרל טוען שמבנה חישובי אינו מכונן ומספיק, ואומר שכדי שתתקיים אינטליגנציה צריכה להתקיים גם מודעות. הוא משתמש ב"**טיעון החדר הסיני**" הטוען את הדברים הבאים:

- תוכניות מחשב הן פורמליות (סינטקטיות)
- לנפש אנושי יש תכנים מנטאליים (סמנטיקה)
- סינטקס כשלעצמו אינו מספיק ואינו מכונן סמנטיקה.

המסקנה העיקרית – תכניות מחשב אינן מספיקות ואינן מכוננות נפש אנושית. לפי סרל, תכנים מנטליים הם "מחשבות, תפיסות, הבנות וכו' העוסקים בחפצים, מצבים ועניינים בעולם".

החלק שאינו ברור מאליו מבין שלוש ההנחות הנ"ל הוא ההנחה השלישית, וסרל מדגים אותו באמצעות הסיפור על החדר הסיני – נניח (בשלילה) שתכניות מחשב כן מכוננות תכנים מנטליים. אם כך, קיימת תכנית מחשב המכוננת הבנה של השפה הסינית. נרשום את התוכנית הזו בספר כלשהו. כמו לכל תכנית, יש לפקודות שלה את הצורה הבאה: אם סדרת הסימנים היא ☺□■, שמור את ■ בזכרון, והחלף בין ☺ ו-□. כעת סרל (בתור המעבד) יושב בחדר עם הספר (המייצג את התוכנה), וסלי ניירות המשמשים בתור זכרון. בהנתן קלט בסינית, סרל פותח את הספר, ממלא את פקודות התוכנית, ומחזיר פלט של סימנים בסינית. כלומר, סרל מבצע בדיוק את תכנית המחשב להבנת סינית. אבל, סרל אינו מבין סינית. לכן, ההנחה שיש תכנית מחשב שמכוננת תכנית הקשורים בהבנת סינית היא שקרית. לכן, באופן רחב יותר, סרל טוען כי סינטקס כשלעצמו אינו מספיק ואינו מכונן סמנטיקה.

מסקנות נוספות מטיעון החדר הסיני:

- "אינטליגניציה מלאכותית חזקה" היא שקרית – ישנם שני יצורים (דובר סינית ו"מכונת החדר הסיני") שמריצים אותן תוכניות חישוב שפה, אך יש להן תכונות נפשיות שונות מאוד (האחד מבין סינית והשני לא).
- הפרכה של ההגדרה הבהוויוריסטית דרך מבחן טיורינג – ישנם שני יצורים (דובר סינית והחדר הסיני) שמתנהגים אותו דבר (עוברים את מבחן טיורינג בסינית), אך יש להם תכונות נפשיות שונות.

מה חסר לחדר הסיני?

- לפי סרל (גישה אינטרנליסטית)- כוחות סיבתיים המצויים במוח.
- לפי דרטסקה (גישה אקסטרנליסטית) – קשרים סיבתיים עם העולם החיצוניץ מה שמכונן תוכן של ייצוגים הוא קורלציה סיבתית מסוג מסויים עם העולם החיצוני.

שיעור 8 – הגישה הקלאסית

נדון היום בעדויות מתחום מדעי הקוגניציה שתומכות בהיפותזת עיבוד הסימבולים – כלומר עדויות שטוענות שהמערכת הקוגניטיבית שלנו היא למעשה מערכת לעיבוד סימבולים.

רציונליזם לעומת אמפריציזם

- רציונליסטים (דקארט, לייבניץ, שפינוזה) ראו את המערכת הקוגניטיבית האנושית כמערכת לעיבוד סימבולים – כלומר מבנה לוגי מורכב, בעוד אמפיריציסטים (לוק, ברקלי, יום) חשבו שהיא פחות מורכבת, כלומר היא מבנה פשוט / אסוציאטיבי.
- רציונליסטים לרוב חושבים שחלק משמעותי מהמערכת הקוגניטיבית היא מולדת, בעוד אמפריציסטים טוענים שרוב היכולות הקוגניטיביות שלנו אינן מולדות אלא נרכשות / נלמדות.
- רציונליסטים טוענים שהמערכת הקוגניטיבית היא מודולרית (מערכת נפרדת לעיבוד שפה, זיהוי פנים וכו') בעוד אמפרציסטים טוענים שזוהי מערכת אחודה ולא מודולרית (מערכת קוגניטיבית אחת האחראית על כל התפקודים הקוגניטיביים).

הגישה הקלאסית – פודור ופילישין

פודור (פילוסוף) ופילישין (קוגניטיביסט חישובי) הציגו בשנות ה-70 את המאמר שלהם "השפה של היפותזת המחשבה", בו הם טענו שהמערכת הקוגניטיבית היא "שפת מחשבה":

- היא מערכת ייצוגים
- הייצוגים הם ביטויים לשוניים / סימבוליים (למשל סדרות של אפסים ואחדים)
- הסינטקס והסמנטיקה של הביטויים הלשוניים מוגדרים באופן רקורסיבי
- הפעולות החישוביות מוגדרות על הסינטרס (ולא על הסמנטיקה)
- המערכת ממומשת בפעילות העצבית במוח

מערכות ייצוגים קלאסיות

מערכת ייצוגים קלאסית היא מערכת בה יש הפרדה בין ביטויים סימבוליים פשוטים למורכבים. הסינטקס והסמנטיקה שלה מוגדרים באופן רקורסיבי. חוקי הפעולה מוגדרים על (בלומר רגישים ל) מבנה תחבירי של מערכת הייצוגים. התחביר הוא זה שמניע את הסמנטיקה (המשמעות) והם מכונים "מנועים סמנטיים". מכונת טיורינג היא דוגמא מובהקת למערכת ייצוגים קלאסית.

דוגמא – מערכת בינארית לייצוג מספרים:

- Rules of syntax:
  - '0' and '1' are (simple) expressions.
  - If T is a symbolic structure, then T0 and T1 are also symbolic (complex) expressions.

- Rules of semantics (Interpretation):
  - I('0') = 0.
  - I('1') = 1.
  - I('T1') = 2 * I('T') + I('1').
  - I('T0') = 2 * I('T') + I('0').

חוקי **התחביר** של המערכת הזו עושים אבחנה בין ביטויים פשוטים (אפס / אחד) ומורכבים (סדרות של אפסים ואחדות כאשר יש הגדרה למה מוגדר בסדרה חוקית). ההגדרה הזו נקראת רקורסיבית כיוון שההגדרה של סדרה תקנית מתבססת על ההגדרה של סדרה תקנית.

דוגמא ל**פירוש חוקי הסמנטיקה** של הדוגמה הנ"ל (מדגיש את הרקורסיביות) – מימין.

I('110') = 2 * I('11') + I('0') =
2 * I('11') + 0 =
2 * I('11') =
2 *(2 * I('1') + I('1') =
2(2 * 1 + 1) =
2 * 3 = 6

מערכות ייצוגים לא קלאסיות

מערכות כמו מד מהירות ברכב הן מערכות ייצוגים לא קלאסיות – אין להן את המאפיינים המגדירים מערכת קלאסית (זו לא מערכת של ביטויים לשוניים, אין בה הגדרות רקורסיביות וכו').

נעם חומסקי – על שפה טבעית

לפי חומסקי, בשפה יש מספר אינסופי של פסוקים דקדוקיים (נכונים תחבירית), לדוגמא:

- יהיה טוב.
- יוסי חושב שיהיה טוב.
- דינה חושבת שיוסי חושב שיהיה טוב
- דני חושב שדינה חושבת שיוסי חושב שיהיה טוב.
- מיכל חושבת שדני חושב שדינה חושבת שיוסי חושב שיהיה טוב.

כל אחד מהמשפטים הוא סופי, אך עקרונית אפשר ליצור מספר אינסופי של משפטים, וזו נקודת המוצא של חומסקי. יש כמה גישות לאיך לקבוע האם משפט הוא תחבירי:

- בלשנות קלאסית – הסבר של השפה דרך ההקשר ההיסטורי, התרבותי, הטקסטואלי, והשוואה לשפות אחרות
- בהוויוריזם פסיכולוגי – מחקר של שפה בתור אוסף קשרים בין גירויים ותגובות.
- בלשנות חומסקיאנית – השפה היא יכולת קוגניטיבית הנמצאת במוחנו (היא איבר ביולוגי).

חומסקי מאפיין את היכולת הלשונית הנמצאת במוחנו בשתי תכונות – כשירות competence (היכולת להבין ולייצר מספר אינסופי של משפטים תחביריים), לצד מגבלות הקיימות על הביצוע performance כמו זכרון, זמן וכו'. הכשירות מורכבת ממספר סופי של כללים שניתן לכנותם "כללי דקדוק". כללי הדקדוק חלים על כל השפות הטבעיות, והם כללים אוניברסליים. היכולת הלשונית היא יצירתית / גנרטיבית, כלומר למרות אופיים הסופי של כללי הדקדוק, ניתן ליצור באמצעותם מספר אינסופי של משפטים תחביריים חדשים.

חומסקי והגישה הקלאסית – לפי חומסקי, יש לחוקי הדקדוק מבנה רקורסיבי שחלים על כל משפט בשפה:

A. Basic rules:

- $S \rightarrow$ (that) NP_2 VP_2
- $X_2 \rightarrow$ (det.) X_1
- $X_1 \rightarrow X_0$ (suffixes)

B. Movement rules: (e.g., you cannot move words from NP-structures).

[S = sentence; NP = Noun Phrase; VP = Verb Phrase].

ניתן לחשוב על היפותזות שיצביעו על קיומם של חוקי תחביר והעובדה שאנשים מייצגים אותם באופן רקורסיבי (כמו שעושים בקורס "מבוא לחקר השפה" עם עצים תחביריים).

הגישה הרציונליסטית למול האימפריציסטית

לפי הגישה הרציונליסטית (וגם לפי חומסקי):

- היכולת הלשונית- קוגניטיבית היא **מולדת**. היא טבועה במבנה הגנטי שלנו, ומועברת מהורים לצאצאיהם. זה מציג עמדה **נייטיביסטית**, לפיה כל המושגים (ידע) שלנו מולדים (לדוגמא דקארט טען שלא קיים רעיון שאינו מולד innate למיינד או ליכולת החשיבה).
- היכולת הלשונית-קוגניטיבית היא **מודולרית**. חוקי השפה הם יחודיים לכושר קוגניטיבי מסוים (שפה).

לעומתם, אימפרציסטים טוענים:

- אנחנו באים לעולם בתור לוח חלק (טאבולה ראסה). לדוגמא, לוק טען שעלינו להניח שהמיינד הוא דף חלק, נקי מכל אפיון, וריק מכל רעיון.
- כל הידע שלנו נרכש באמצעות החושים וההתנסויות שלנו לאורך החיים.

עמדות שונות על מולדות של שפה

עולה השאלה – למה הכוונה ב"מולד"? מה מולד (ידע, מושגים, אקסיומות, חוקי מחשבה)? האם מולד הוא ההיפך מנרכש? נבחן עמדות שונות בנושא -

- **דקארט – מולדות כדיספוזיציה**: "הרעיונות האלו טבועים בנו, הם תמיד קיימים בפוטנציאל שלנו, מעולם לא כתבתי או אפילו חשבתי שהרעיונות האלו באמת אמיתיים". כלומר, אנשים נולדים עם הפוטנציאל לדעת דברים, לא עם הידע עצמו, וכאשר אנחנו מגיעים לפוטנציאל מסוים הידע מתממש.
- **לוק – ביקורת על דקארט**: לוק ביקר את דקארט – אם ידע הוא "טבוע", לא ברור באיזה מובן זה ידע או מושג, כיוון שמצב נפשי מזוהה רק עם דברים מודעים (או דברים שהיו מודעים), לכן כיצד משהו שמעולם לא היינו מודעים אליו יכול להיות מצב נפשי (ידע)? "רעיון שמעולם לא נתפס ע"י המיינד מעולם לא התקיים במיינד". הוא גם טוען שההגדרה של דקארט למולדות אינה מעניינת, שכן אם מגדירים מודעות לפי פוטנציאל, אז לפי ההגיון הזה הכל מולד.
- **העמדה האימפריציסטית**: <u>כל</u> מה שבקוגניציה הוא תוצאה של רשמים של החושים שלנו, שעובדו במנגנונים פשוטים (שמאפשרים הכללות של רשמי החושים). מנגנוני ההכללה הם מנגנוני למידה המבוססים על דוגמאות אסוציאטיביות (דמיון וסמיכות בזמנים בין הדוגמאות של רשמי החושים). המנגנונים עצמם הם מולדים. המערכת הקוגניטיבית היא לא מודולרית – החוקיות הבסיסית שמקיימת את כל התכונות הקוגניטיביות מבוססת על אותם מנגנוני למידה מדוגמאות, אין "מודולים"

שונים במערכת הקוגניטיבית, כלומר לדוגמא גם שפה וגם זיהוי פרצופים יבוצעו באמצעות למידה מדוגמאות חושיות.

- **העמדה של חומסקי**: חוקי הדקדוק האוניברסליים הם מולדים (ממש מולדים, כלומר לא "קיימים בפוטנציאל" כמו שדקארט מציע). החוקים האלה הם מבנים קוגניטיביים שדוברי השפה אינם מודעים להפעלתם. היכולת הלשונית היא מודולרית – חוקי הדקדוק יחודיים ליכולת השפתית ולא מאפיינים כשרים קוגניטיביים אחרים.

נימוקים <u>בעד</u> טענת המולדות של שפה:

- האוניברסליות של שפות – חוקי הדקדוק משותפים לכל השפות המדוברות, ואף לשפות סימנים שונות.
- ילדים לומדים שפה בקצב מהיר באופן קיצוני.
- טיעון דלות הגירוי – היכולת הלשונית שלנו (שהיא אינסופית) אינה יכולה להיות מוסברת ע"י גירויים סביבתיים (שהם מוגבלים וסופיים), זהו טיעון נגדי לעמדה האימפריציסטית (שטוענת שהלמידה של השפה מתבצעת באמצעות גירויים בלבד).

יש הרבה דיונים על הטענות האלו, ואין הסכמה לגבי הרעיון שהן מעידות על מולדות.

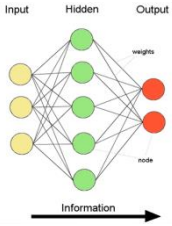
שיעור 9 – חישוביות עצבית

היסטוריה של תחום החישוביות העצבית

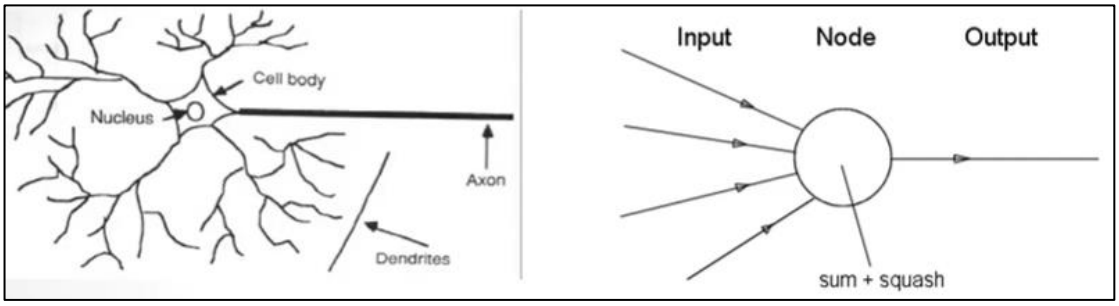
- מקקלוך (נוירוביולוג) ופיטס (תאורטיקן) 1943 – כתבו מאמר על חישוב לוגי של הרעיונות הבסיסיים של פעילות עצבית.
- הב 1949 – הוסיף את אלמנט הלמידה לרשתות נוירונים. הרעיון שלו היה שצריך לחזק את הקשרים בין נוירונים שיורים ביחד.
- וון ניומן 1958 – הבחין בין המחשב הדיגיטלי לאופן החישובי שבו המוח פועל, באופן שמזכיר את האבחנה בין הגישה הקלאסית והגישה האימפריציסטית.
- רוזנבלט 1958 – דיבר על מערכת פשוטה שפועלת לפי חוק למידה, אך למערכת שלו היו מגבלות שהוא לא הצליח להתגבר עליהן וגרמו לנטישה של התחום עד שנות ה-70.
- רומלהרט ומקללנד – הראו שבעזרת חוק למידה מאוד פשוט אפשר ללמד רשת נוירונים לבצע כל פונקציה שניתנת לחישוב.
- הופפילד 1982 – קבוצה של פיזיקאים שמה לב שיש דמיון בין האופן בו אפשר לתאר רשתות נוירונים לא לומדות לאופן שבו פועלות מערכות פיזיקליות מסוימות, ולכך שיש דמיון בין האופן שבו פועל תא עצב בודד לאופן שבו פועל חלקיק. הם הצליחו לתאר באמצעות רשתות נוירונים תופעות קוגניטיביות מורכבות.

הויכוח הקלאסי-עצבי

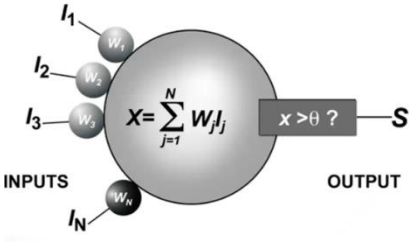
שתי הגישות רואות את המוח כמחשב – חוקים הפועלים על ייצוגים. הויכוח הוא על המבנה (הארכיטקטורה) של המערכת הקוגניטיבית, ועל מידת המולדות של המערכת.



גישת החישוביות העצבית (קונקשניזם, PDP, רשתות נוירונים) – לפי גישה זו, מבנה המערכת הקוניטיבית דומה למבנה המערכת העצבית. מערכת הייצוגים היא אסוציאטיבית וגמישה. אין הפרדה פונקציונלית בין חוקים לבין זכרון. רוב הכשרים הקוגניטיביים מושגים ע"י חוק כללי של למידה אסוציאטיבית-אינדוקטיבית (כלומר הכללה מדוגמאות).



הגישה מקבלת השראה ממבנה הנוירון, וכך יצרו רשתות נוירונים מלאכותיות הבנויות באופן דומה לאיך שפועלים נוירונים במוח. כלומר, התא מקבל קלט מהרבה תאים אחרים, מבצע לו סכימה מסובכת, ואז מוציא פלט יחיד (שיכול להגיד ליותר מתא אחד). כלומר, תא העצב פועל באופן הבא:



התא מקבל n קלטים (I_i עם $1 \le i \le n$ - באיור n=3). כל קלט עובד עיבוד בסינפסה i_N (שיכול להיות מעורר או מעכב), ולאחר מכן בתוך התא מתרחשת סכימה של כל הקלטים (לאחר המודולציה שעברו בסינפסה). הפלט נקבע לפי תוצאת הסכימה הנ"ל ולפי תנאי מסוים (לדוגמא – התנאי יכול להיות "התא ירה אם הסכום גדול מאפס").

רשתות לומדות

נתמקד ברשתות לומדות. ברשתות מסוג זה מתבצעת למידה באמצעות שינוי של הקשרים הסינפטיים (אם הקשרים הסינפטיים משתנים, אופן הפעולה של הרשת כולה משתנה, כלומר התרחש שינוי של ההתנהגות שמבטא למידה).

יש כמה דרכים לשנות קשרים סינפטיים, הדרך הפשוטה והנפוצה ביותר הוא חוק "מזעור השגיאה":

$$\Delta W_i = \eta * (D-Y)\, I_i$$

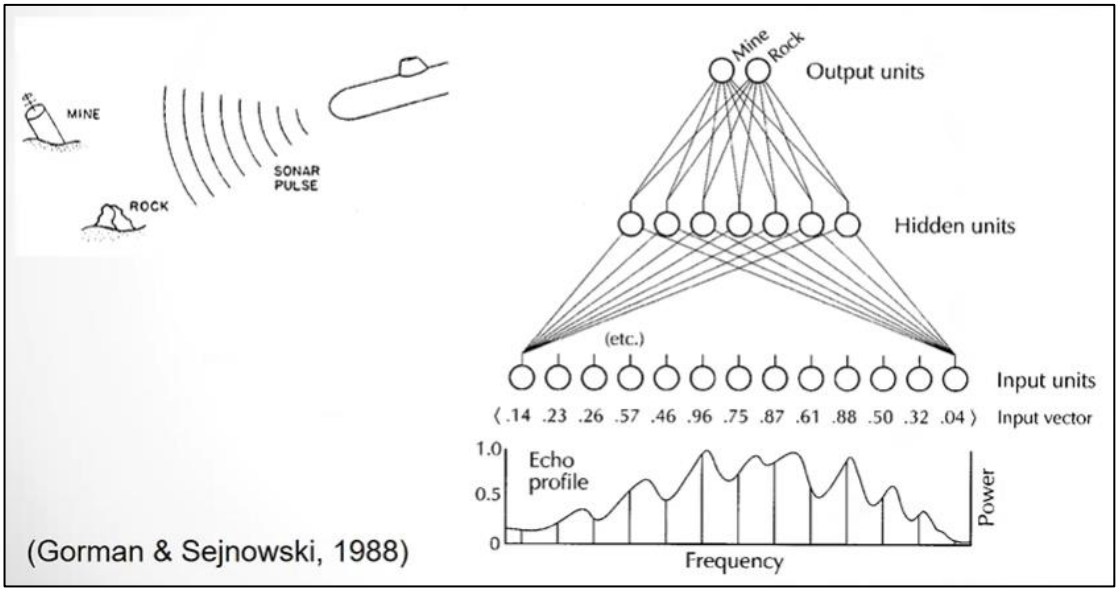
כלומר, אנחנו משנים את הקשרים הסינפטיים לפי הקשר בין ההתנהגות הרצויה של הרשת להתנהגות המצויה של הרשת. מודדים את ההפרש בין הרצוי למצוי, מכפילים אותו בקבוע מסוים, ועל סמך תוצאת החישוב משנים את הקשר הסינפטי. החוק הזה הוא הבסיס ללמידה של רוב הרשתות הלומדות.

רשתות שכבתיות FEED FORWARD

רשתות שכבתיות הן רשתות המורכבות משכבות, כאשר בכל שכבה ישנם תאי עצב, וכל תא מקרין על כל תאי העצב שבשכבה שאחריו (ורק עליהם). בציור, השכבה הצהובה היא השכבה הראשונה (קלט), הירוקה היא השניה (שכבת ביניים חבויה) והאדומה היא האחרונה (פלט).

דוגמא ראשונה (לא קוגניטיבית) – ניסו לאמן רשת לזהות האם תדר ההחזר של סונר של צוללת מעיד על החזר מחפץ מתכתי (מוקש) או מסלע. בשלב הראשון, ביצעו השמה מקרית של ערכים לקשרים – כלומר בנו

את הרשת עם "חוזקי קשרים" אקראיים בין התאים השונים ברשת. בשלב השני, אימנו את הרשת על קצת יותר מ-200 תדרים מתויגים למוקש/סלע. עבור כל דוגמא חישבו את ההפרש בין הפלט הרצוי לפלט שהרשת נתנה בפועל, ועדכנו את עוצמת הקשרים ברשת בהתאם. בשלב השלישי, ביצעו בדיקה של ביצועי הרשת על קלטים חדשים (לא מרשימת הדוגמאות המקוריות). הרשת נראית ככה:



התוצאות – הרשת למדה להבחין בין מוקשים וסלעים. תיקוני הקשרים נעשו בצורה כזו שמערכת הקשרים הצליחה להתכנס לנקודה התואמת את כל התיקונים הקודמים (אם כי יתכן ונדרש מספר רב של הרצות כדי להגיע למצב הזה). החוק המבחין בין מוקשים וסלעים אינו ידוע למתכנני הרשת, ואינו נמצא באופן מפורש ברשת בנפרד מהזכרון. הרשת רכשה את החוק מהדוגמאות בלבד.

הרשת הבחינה בין מוקשים וסלעים שלא נתקלה בהם קודם לכן – ההנחה היא שיש חוקיות כלשהי המבחינה בין גלי מוקדים וגלי סלעים, והמערכת רכשה את החוק הכללי הזה בעזרת לימוד ממספר דוגמאות בלבד.

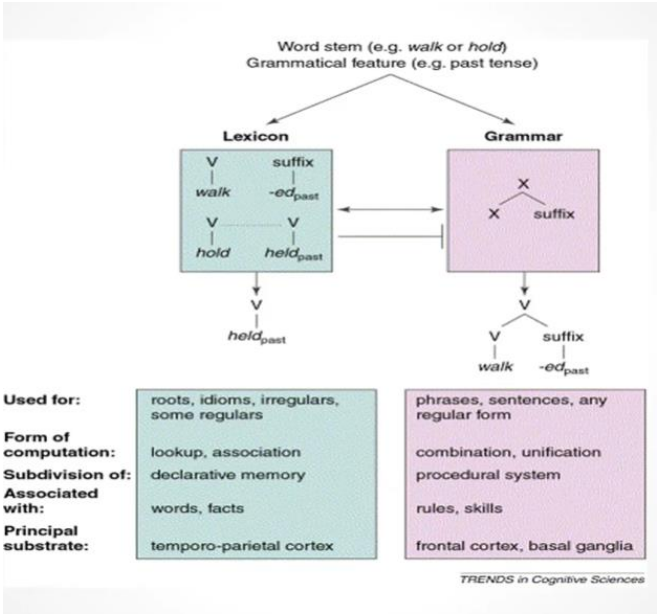
דוגמה שניה (קוגניטיבית) – הניסוי של רומלהרט ומקללנד (מפורט בהמשך).

המודל של פינקר

פינקר ביקר את המאמר של רומהרט ומקללנד:

- גנרליזציה של מילים לא שגרתיות – הרשת של רומלהרט ומקללנד לא תדע איך להתייחס למילה שלא דומה לשום דבר שלא ראתה בעבר, ותחזיר בתשובה הטיה שלה שהיא nonsense.
- הכללות סיסטמטיות – בני אדם יודעים לזהות מקרים שבהם פעלים לא רגולריים כן אמורים להיות מוטים בצורה רגולרית, והרשת לא יודעת לעשות זאת.
- הבדלים ביולוגיים – יש עדויות לכך שבמוח האנושי יש הבדל ביולוגי בעיבוד של פעלים רגולריים ופעלים לא רגולריים (מתרחשים באזורים שונים במוח).

פינקר הציע בשנות ה-2000 מודל שמתכתב עם הגישה האימפריציסטית וגם עם הגישה הקלאסית, ועוסק בהטיית פעלים לזמן עבר באנגלית:



לפי המודל של פינקר, המערכת להטית פעלים בזמן עבר מתחלקת לשניים:

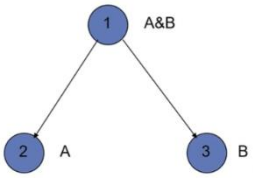
- **לקסיקון** – זכרון של הטיית פעלים לא רגולריים (שעשוי להיות לדוגמא רשת עצבית מהסוג שהציגו רומלהרט ומקללנד).
- **דקדוק** – יחידה נפרדת, שמבצעת עיבוד של הפועל ומפרידה בין צורת הבסיס של הפועל לבין המוספיות המייצגות את ההטיה שלו.

כאשר המערכת רוצה לעבד מילה (כלומר להטות אותה לזמן עבר), היא קודם כל בודקת בלקסיקון אם המילה קיימת בזכרון. אם היא קיימת בזכרון, תוחזר המילה שקיימת בזכרון. אם היא לא קיימת בזכרון, היחידה הדקדוקית מפעילה עליה את החוק הדקדוקי ומוסיפה לה את המוספית הנדרשת, ומה שיוחזר זה התוצאה של הפעולה הזו. זה מתאים לכך שבמוח העיבוד של פעלים רגולריים ואי רגולריים מתרחש במקומות שונים.

ביקורת על המודל של פינקר:

ישנן תופעות קוגניטיביות שאינן יכולות להיות מוסברות ע"י ארכיטקטורה עצבית – כמו פרודוקטיביות (גנרטיביות), קומפוזיציונליות, סיסטמטיות. למה הכוונה בסיסטמטיות – היכולת השפתית היא סיסטמטית במובן שזוהי יכולת להפיק / להבין משפטים (ומחשבות) שמקושרת ליכולת להפיק / להבין דברים אחרים. לדוגמא – אם למישהו יש את היכולת לחשוב שג'ון אוהב את מארי, קשה לדמיין מצב שבו אין לו את היכולת לחשוב שמארי אוהבת את ג'ון. ההסבר הקלאסי הוא שהמחשבות בנויות מאותם חלקים – ג'ון, מארי, והיחס "אוהב/ת את". מכאן את החוקים התחביריים והסמנטיים שמאפשרים מחשבה אחת מבין השתיים, מאפשרים גם את קיומה של האחרת. ניתן ליישם סיסטמטיות גם במערכות עצביות. יתכן שהמקור של סיסטמטיות אינו בארכיטקטורה הקוגניטיבית אלא באינטראקציה עם קלטים שיש להם סיסטמטיות (למשל תחביר של שפה טבעית).

נביט ביכולת לייצג משפט שמורכב משני משפטים המחוברים ביניהם ב"ו". תומכי **הגישה הקלאסית** יטענו שמדובר בשני יצוגים פשוטים יותר המוחברים ביניהם בחיבור הלוגי של הפונקציה "ו", כלומר למי שיש יצוג של המשפט המורכב בהכרח יש ייצוג של שני המשפטים הפשוטים יותר, ואת היכולת לייצג את החיבור "ו".



לעומת זאת, ברשת עצבית (בציור), יכול להיות מצב שבו כל אחד מהמרכיבים (המשפט המורכב עצמו, ושני המשפטים הפשוטים שמרכיבים אותו) הוא תא עצב בפני עצמו, אבל הקישור ביניהם (נגיד שאם קורה A ו B אז קורה A ו/או אז קורה B) הוא לא משהו הכרחי לתפקוד הרשת, ויצירת הקישורים האלו היא לא "טבעית" אלא נוצרת (או לא נוצרת) כתלות בקלט שאנחנו נותנים לרשת הנוירונים.

למידה עמוקה DEEP LEARNING

בשנים האחרונות (החל מבערך 2012) התפרסמו מספר מאמרים שתיארו תכונות חדשות שניתן להוסיף לרשתות נוירונים, שהופכות אותן להרבה יותר חזקות ומאפשרות להן לבצע דברים שלא יכלו לבצע קודם לכן. התכונות העיקריות שהוזכרו הן:

- אילוצים מובנים על הקשרים
- תאי עצב עם פעולות שונות (לדוגמא תאים שמבצעים max pooling שדומה לעקרון הmin-max שראינו בהרצאה על בינה מלאכותית)
- הרבה רמות חבויות (זוהי תכונה שמאוד משפרת את יכולת הקטגוריזציה וכוח החישוב של הרשתות)

זה אפשר להשתמש ברשתות נוירונים לטובת עיבוד תמונות, זיהוי דיבור, תרגום, משחקים ועוד. הרשתות האלו הפכו להיות מאוד מרכזיות בתעשיות שעוסקות בבינה מלאכותית. אך עולה השאלה האם באמת מדובר בבינה אמיתית, או שאולי מדובר ב-brute force מסוג חדש. עולה גם התהייה עד כמה הבינה המלאכותית המשתמשת בלמידה עמוקה אכן משקפת את הקוגניציה האנושית.

דייוויד רומלהרט וג'יימס מקללנד – ON LEARNING THE PAST TENSES OF ENGLISH VERBES

רומלהרט ומקללנד מתבססים על תאוריית שלושת השלבים לרכישת זמן עבר שהציע ברקו במאמרו מ-1958. שלושת השלבים הם:

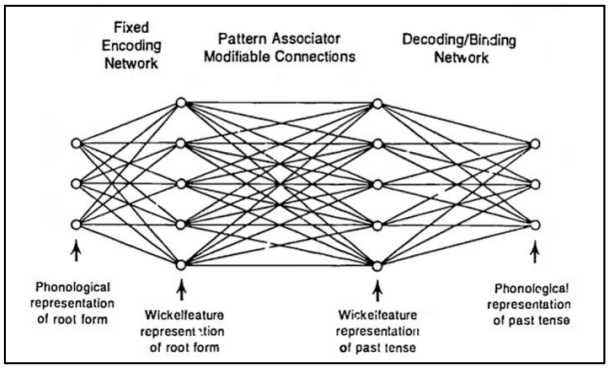
1) שלב ראשון – הלקסיקום של הילדים כולל מספר קטן של פעלים מתוכם רק שניים הם רגולריים.
2) שלב שני – הלקסיקום מתרחב לפעלים נוספים ורבים, שרובם הגדול הוא רגולרי. בשלב זה ילדים יטו מילים מומצאות "לפי החוק", למשל rick ל-ricked. בשלב זה לעתים יהיו לילדים טעויות של הכללת יתר – הם יחילו את החוק גם על פעלים לא רגולריים, כולל כאלו מהקבוצה שהכירו בשלב 1 (כלומר פעלים לא רגולריים שביטאו בעבר באופן תקין).
3) שלב שלישי – הלקסיקון מתייצב על השימוש הנכון, וכלל הטיה נכונה של פעלים חדשים (בין אם הם רגולריים או לא רגולריים).

רומלהרט ומקללנד בנו מודל מלאכותי המדמה את שלושת השלבים האלו. המודל שהם מציעים מכיל שני חלקים עיקריים:

1) מכשיר ליצירת אסוציאציות בין דפוסים, שאחראי לזיהוי הקשר בין צורת ההווה לצורת העבר. בחלק זה מתבצעת כל הלמידה. חלק זה מורכב משני מאגרים של "יחידות" - מאגר קלט (שמכיל את צורת הבסיס של

הפועל) ומאגר פלט (שמכיל את צורת העבר של הפועל), ומאגר של קישורים (במידות חוזק שונות) המחברים בין צורות של פעלים ממאגר הקלט לצורתם התואמת במאגר הפלט.

2) רשת "פענוח" שממירה את הצורה הייצוגית של המילה לצורתה הפונולוגית.



בניסוי שביצעו, הם לימדו רשת נוירונים (מחשב) בעלת מכשיר ליצירת אסוציאציות בין דפוסים כמה מאות של פעלים, תוך שסיפקו לה קלט נכון בלבד, אך היא ביצעה הסקות גם מהידע הפנימי שלה.

הניסוי שביצעו כלל שלושה שלבים:

1) שלב הלימוד – ביצעו 10 הרצאות על קבוצה של 10 פלים שכיחים (מתוכם 8 לא רגולריים) – מקביל לשלב הראשון הקוגניטיבי)
2) שלב הלימוד בהמשך – הוספת 410 פעלים חדשים עם שכיחות ממוצעת (334 רגולריים ו-76 לא רגולריים) למשך 190 הרצות (מקביל לשלבים השני והשלישי הקוגניטיביים) .
3) שלב שלישי – בדיקת המערכת ללא לימוד על 86 פעלים חדשים בעלי שכיחות נמוכה (72 רגולריים ו-12 לא רגולריים).

תוצאות הניסוי הראו כי רשת הנוירונים אכן למדה פעלי עבר בצורה דומה לצורה בה למד ילד אנושי בן שלוש, והיא אכן הדגימה מעבר בין שלושת השלבים המתוארים מעלה, תוך מעבר הדרגתי בין שלב 2 לשלב 3. כל זאת מבלי שיסופקו לה חוקים תחביריים מפורשים. הניסוי שלהם הראה שרשת הנוירונים למדה בהצלחה 460 פעלים, והצליחה לבצע גנרליזציה בעצמה, כלומר להצליח להתמודד עם הטית פעלים לא מוכרים בהתבססות על קישורים קיימים שיצרה קודם לכן.

העמדה בנוגע לרכישת זמן עבר אצל דוברי אנגלית ילדיים אותה מנסים רומלהרט ומקללנד להפריך היא שלמידת פעלי עבר מתבצעת בצורה מפורשת (אקספליציטית), כלומר באמצעות למידה של חוקי ההטיה של פעלים. המודל שלהם מדגים זאת, כיוון שהם טוענים כי רשת הנוירונים שיצרו לא יצרה לעצמה חוקים ופעלה לפיהם, אלא פשוט הסתמכה על חוזק של קישורים דומים שנוצרו באופן הסתברותי. כלומר, המודל לא קיבל ולו חוק תחבירי אחד מקדים (כלומר הקשרים בו נוצרו באופן אקראי), ועדיין הצליח ללמוד בהצלחה את כל ההטיות הנכונות של הפעלים שהוזנו לו, והצליח ללמוד להטות פעלים חדשים בעצמו. לכן, הם מסיקים, כי חוקים מולדים אינם הדרך בה נלמדת הטית פעלים בזמן עבר.

הם מסיקים מכך שילדים לא יוצרים לעצמם חוקים מפורשים שהם מודעים אליהם (וטוענים כי יתכן כי ילדים אפילו אינם מודעים לקיומם של חוקים תחביריים שכאלה), ואומרים שהיכולת שלהם להטות פעלים שלא נתקלו בהם מעולם מבוססת על כך שנתקלו במספיק דוגמאות דומות שביססו דפוס ברור מספיק כדי שיהיו מסוגלים להשתמש בו בעצמם מבלי להיות מודעים לכך שהם עושים זאת.

מסקנות מהניסוי (לפי רומלהרט ומקללנד עצמם)

- ניתן ללמוד חלק משמעותי מהפעלים בזמן עבר מבלי שיהיה ללומד שום סוג של ידע מקדים על חוקים תחביריים (בשונה לדוגמא מהגישה החומסקיאנית)
- אין "בעיה אינדוקטיבית" – הבלשנות החומסקיאנית טוענת שבהרבבה מקרים יש שתי דרכים שונות לנסח משפט שרק אחת מהן תקנית (אך שתיהן תואמות להרבה דוגמאות אליהן נחשפנו בעבר), ושלא ניתן לבחור בין השתיים בצורה נכונה מבלי שיהיה ידע / חוק מקדים (מולד) שיסייע בכך. רומלהרט ומקללנד הראו שהדבר לא נכון – שכן הרשת הצליחה לבחור באפשרות הנכונה מבלי שהיה לה שום ידע מקדים.
- חייב להתקיים דפוס כלשהו (דמיון בין הדוגמאות) שהרשת תוכל ללמוד ממנו על מנת שהרשת תצליח לבצע גנרליזציה (הכללה) בהצלחה.
- רומלהרט ומקללנד טענו שסביר שעקרונות דומים יוכלו להיות מושמשים גם על מובנים נוספים של רכישת שפה.

שיעור 10 – פסיכולוגיה כמדע

מהו מדע, ופסיכולוגיה כמדע

מטרות המדע הן:

- גילוי של קורלציות, אפקטים, חוקים, הכללות, קשרים סיבתיים
- תחזית של תופעות עתידיות
- בקרה ושליטה על הטבע
- מידול וסימולציה של תופעות
- הסבר של תופעות, קורלציות, יכולות

כשאנחנו רוצים להסביר תופעות קונגיטיביות, אנחנו משתמשים הרבה ב**הסברים אינטנציונלים**: למה יוסי [ירה בשכנה] – כי יוסי רצה ש[שכנו ימות]. ההסבר מבוסס על היסק לוגי: ההבנה שאם תתרחש יריה בשכן, אז יתרחש מות השכן. הבעייתיות בגישה הזו:

- היא יוצאת מנקודת הנחה שכל המעשים של בני אדם הם רציונליים, כשהרבה מחקרים הראו שיש מקרים רבים בהם אנשים פועלים באופן אי-רציונלי (בעקבות הטיות בשיפוט).
- אפשר לדון גם בהאם הקוגניציה מבוססת על "שפת מחשבה", כלומר האם לרצונות והאמנות שלנו יש ייצוב פרופוזיציונלי (פסוקי, לוגי). הגישה הקלאסית אומרת שהארכיטקטורה הקוגניטיבית שלנו מבוססת ברובה על שפת מחשבה, בעוד גישות אחרות (כמו חישוביות עצבית) אומרת שהיא ברובה **אינה** מבוססת על שפת מחשבה.
- אפשר לדון בכמה הסברים אינטנציונליים הם "מדעיים" – ראינו ביקורת של סקינר על הלגיטימיות של הסברים אינטנציונליים כהסברים מדעיים. סקינר טען כי הרבה פעמים ניתן להתאים את האמנות והרציות להתנהגות מבלי שיש לכך בסיס, או מסיקים שמצבים נפשיים מסוימים קיימים מבלי שיש לכך עדות (לדוגמא – להסיק שמישהו מתנהג באלימות כי הוא הלום קרב – הסקנו שההתנהגות שלו מונעת ממצב נפשי, מבלי שיש לנו גישה לפנימיות שלו, כלומר אנחנו לא יודעים אם המצב הנפשי אכן מתקיים ובכל זאת מסיקים שהוא קיים). לעומת סקינר, פילוסופים אחרים (כמו דנט) העידו שהסברים אינטנציונלים נותנים תחזיות טובות ומאפיינים אותנו כבני אדם, אך אינם מתיימרים להיות הסברים מדעיים.

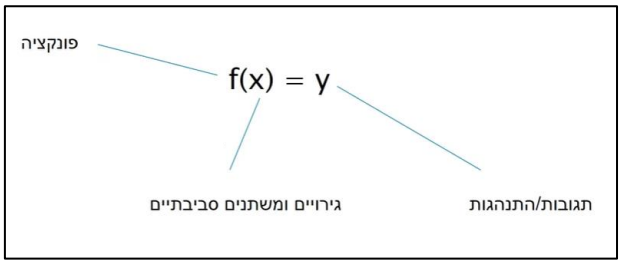
במאה ה-17 הפילוסופיה הופרדה ממדעי הטבע, ובמאה ה-19 הפסיכולוגיה הופרדה מהפילוסופיה. בתחילה, הפסיכולוגיה נתפסה כמדע העוסק במחשבות ורגשות והקשר שלהן להתנהגות, אך לא עוסקת בטבען של

מחשבות, מה קובע את התוכן של מחשבות וכו' – שהן שאלות פילוסופיות. המתודות הראשונות ששימשו למחקר פסיכולוגי –

- המתודה האינטרוספקטיבית – שלטה בסוף המאה ה-19 ותחילת המאה ה-20. במתודה זו חוקרים תיחקרו את המצבים הקוגניטיביים של עצמם, וניסו לאפיין אותם. זו היתה שיטה לא מדויקת שנזנחה מהר.
- המתודה הניסויית – משתמשת בדאטא בקנה מידה גדול תוך שימוש בכלים סטטיסטיים, במטרה לתקן שגיאות והטיות. זוהי המתודה העיקרית שמושמשת עד היום.

בהוויוריזם פסיכולוגי

כבר בתחילת המאה ה-20 התחילה העלייה של הבהוויוריזם הפסיכולוגי. בגישה זו, מושא המחקר של הפסיכולוגיה הוא התנהגות, ומטרת המחקר בו היא בקרה וחיזוי של התנהגות. המחקר מבוצע באמצעות אנליזה פונקציונלית – זיהוי המשתנים שהתנהגות היא פונקציה שלהם (מודגם בתרשים מטה). הנחת הבסיס היא שהמשתנים האלו הם בדיוק התנאים הסביבתיים וההיסטוריים של הנבדק, והם ניתנים לתצפית ומתוארים בשפה פיזיקלית.



הבהוויוריזם הפסיכולוגי משתייך לשתי גישות:

- גישה אמפיריציסטית – החוקים ("פונקציות") מבוססים על הכללות אינדוקטיביות ביחס ליחסי הדמיון בין הדוגמאות ("גירויים").
- גישה אנטי-קוגניטיבית: אין צורך בגילוי התהליכים והמצבים הפנימיים לטובת ביצוע האנליזה הפונקציונלית. זה הופך את הגישה לגישה אמפיריציסטית קיצונית.

נבדיל את הבהוויוריזם הפסיכולוגי מהבהווירזם הפילוסופי – בעוד הפילוסופים ניסו לתת אפיון למצבים נפשיים באמצעות שיח על נטיות / דיספוזיציות התנהגותיות, הפסיכולוגים התעלמו מקיומם של מצבים נפשיים, והתמקדו בהתנהגות בלבד.

סקינר, הוגה הבהווירזם הפסיכולוגי באמצע המאה ה-20. הטענה המרכזית שלו היא שמצבים נפשיים אינם נחוצים לאנליזה הפונקציונלית של ההתנהגות. כדי לתמוך בטענה, הוא נתן את הטיעון הפילוסופי של דילמת התאורטיקן (יוסבר בהמשך), ואת הטיעון האמפירי לפיו ניתן לחזות ולשלוט בהתנהגות במסגרת הבהוויריסטית בלבד.

המסגרת הבהוויוריסטית האמפירית

המסגרת הבהוויוריסטית האמפירית – זוהי מסגרת של מערכות ניסויים, לרוב באמצעות בקרה על בעלי חיים בקופסאות (מבוכים / קופסאות סקינר), ומשפיעים על ההתנהגות שלהם באמצעות חיזוקים ועונשים. המסגרת התאורטית היא חיזוי התנהגות באמצעות שלושה סוגים של קורלציות סטטיסטיות בין גירויים לתגובות פיזיקליות – כאשר תגובות נתפסות כהתנהגות הקשורה סטטיסטית לגירויים שקדמו לה. אלו שלושת הקורלציות בתאוריה:

- **חוק א' – רפלקסים**: קשר לא מותנה בין גירוי לתגובה, זהו קשר קבוע ולא משתנה.

uS (food) → uR(salivation).

uS = unconditional stimulus
uR = unconditional response

- **חוק ב' – התניה קלאסית (פבלובית)**: תהליך בו נוצר קשר בין גירוי לתגובה, ראשית באמצעות הצמדה של גירוי בלתי מותנה לגירוי מותנה, ובהמשך נוצרת תגובה פבלובית שגורמת לתגובה המותנית להופיע גם עבור הגירוי הבלתי מותנה.

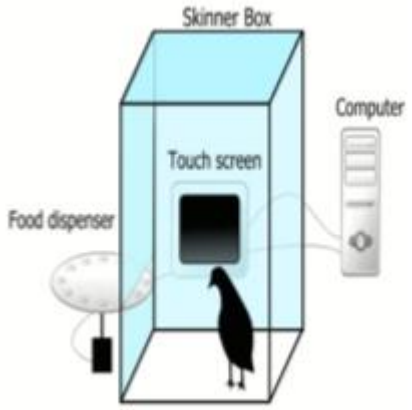
- uS (food) → uR (salivation).
- uS (food) + cS (bell) → uR (salivation).
- cS (bell) → uR (salivation).

cS = conditional stimulus.

- **חוק ג' – התניה אופרנטית**: תהליך בו נוצר קשר בין גירוי לתגובה, שנוצר ע"י מתן חיזוקים ו/או עונשים לאחר הופעת התנהגות רצויה / בלתי רצויה. בדוגמא למטה – נותנים ליונה חיזוק חיובי (אוכל) כשהיא מבצעת פעולה רצויה (ניקור למראה נקודה אדומה). התניה היא שינוי בתדירות של תגובה מסוימת ("אופרנטית" כיוון שההתנהגות פועלת על הסביבה כדי להשיג תוצאות). בחוק ההתניה האופרנטית, אם התכיפות של R (הפעולה הרצויה) עולה, מדובר בחיזוק חיובי, אם היא יורדת, מדובר בחיזוק שלילי. התנהגות היונה מוסברת בכך שהחיזוק החיובי גרם לכך שהיא למדה מהנסיון בעבר (התנהגות מסוג דומה הובילה בעבר לקבלת חיזוק חיובי), והובילה להגברת ההתנהגות.

- R (pecking on the red dot) → rS (food)
- Frequency(R) changes.

rS = reinforcing stimuli



סקינר טען שהתניה אופרנטית ביחד עם התניה קלאסית מסבירות את מכלול ההתנהגויות הנרכשות. התניה קלאסית מסבירה כיצד מקשרים בין גירויים חדשים לתגובות קיימות, והתניה אופרנטית מסבירה כיצד יוצרים תגובות חדשות ("יצירתיות") ו/או משנים את התכיפות של תגובות קיימות. **עקרון הברירה הטבעית –** המבנה של הסבר באמצעות התניה אופרנטית דומה למבנה של הסברים אבולוציוניים, באמצעות עקרון הברירה הטבעית. לדוגמא – למה עכביש טווה קורים? כיוון שלאחר התנהגות דומה בעבר (ע"י האבות הקדמונים שלו) הופיעו חיזוקים חיוביים (הולדת יותר צאצאים), וזה מה שהוביל לחיזוק ההתנהגות.

עולה השאלה – כיצד ניתן לתת הסבר להתנהגות שהתוצאות שלה נמצאות בעתיד? כמו לדוגמא הרתחת מים בדי לשתות קפה, או שחיה כדי להכנס לכושר. יש כמה סוגי הסברים אפשריים:

- הסברים תכליתיים – הסיבה נמצאת בעתיד. הסברים כאלו היו פופולריים בעולם העתיק, וכיום לא משתמשים בהם (כיום לא מסבירים תופעות באמצעות סיבות שמופיעות לאחר קיום התופעה ולא לפניה).
- הסברים אינטנציונליים – הסיבות הן מצבים נפשיים (אינטנציונליים) הקודמים לפעולות.

דילמת התאורטיקן

מדגימה את זה שאפשר למצוא קשר בין משתנים סביבתיים, מצב פנימי,והתנהגות.

- E: משתנים סביבתיים (מיידיים והיסטוריים): תוספת מלח למזון, חימום מוגזם של החדר, מניעת מים לאורך זמן.
- M: מצב פנימי. מצב תיאורטי-היסקי: אינו ניתן לתצפית ציבורית (מצב נפשי, מצב מוחי).
- R:התנהגות: שתייה.

$E \rightarrow M \rightarrow R$

- מונח תצפיתי: 'הרמת היד', 'ליקוי חמה': מונחים המורים על אירועים או ישויות שניתנים לתצפית "ישירה" (בעזרת החושים + מכשור עזר) וציבורית (כל אדם יכול לצפות בהם).
- מונחים תיאורטיים: 'אלקטרון', 'קוואזר', 'המפץ הגדול', 'כאב הראש שלי', 'פעילות עצבית' (?).

לפי סקינר, הבעיה עם מונחים תיאורטיים – הם לא ניתנים לתצפית, אלא אנחנו מסיקים אותם מהתנהגות (אירועים תצפיתיים אחרים). אין לנו יכולת ישירה לצפיה ב"צמא" לדוגמא.

- למה R (יוסי שתה מים)? כי M (יוסי היה צמא, ויוסי מאמין ששתייה תרווה את צימאונו...)
- אבל: M הוא אירוע תיאורטי: אנו מסיקים אותו מתוך אירועים תצפיתיים: חלקם קודמים ל – M (אי-שתייה לאורך זמן, אכילת מזון מלוח) וחלקם עוקבים ל – M (שתיית מים).

בדילמת התאורטיקן יש שתי אפשרויות (שתיהן לא טובות, ולכן זו דילמה):

1) אפשרות ראשונה - ההסקה שלנו היתה שגויה. הסקנו על הצמא של יוסי מהתנאים הסביבתיים שהוא שהה בהם (אכילת אוכל מלוח) או מההתנהגות שלו (שתיית מים), אך למרות שהסקנו שקיים צמא, לא בהכרח קיים צמא – כלומר לא היה קשר סיבתי בין האירועים.

M אינו קשור חוקית (סיבתית) ל- E
או ל- R אז $\sim(E \rightarrow M \rightarrow R)$

2) אפשרות שניה – ההסקה שלנו היתה נכונה, אך המצב הפנימי לא היה נחוץ על מנת לבצע את ההסקה הזו.

אם $E \rightarrow M \rightarrow R$, אז גם מתקיים $E \rightarrow R$,
ומכאן שלא צריך את M.

מה שנובע מהדילמה זה שאם עלינו תמיד להסתכל מעבר לקישור השני על מנת לבצע חיזוי ובקרה, המסקנה מכך היא שעלינו להמנע מניתוח שבוחן את הקישור השלישי בתור פונקציה של הראשון. מכאן נובע שאין צורך במונחים תיאורטים באנליזה הפונקציונלית – או ש-M פיקטיבי, או שהוא מיותר. ההתנגדות לקיומם של מצבים פנימיים היא לא בגלל שהם לא קיימים, אלא בגלל שהם לא רלוונטיים לאנליזה הפונקציונלית.

הביקורת של חומסקי על סקינר

הביקורת של חומסקי – גם אם סקינר צודק בטיעוניו, הרי שבפועל, כשמדובר בהתנהגות מורכבת, לא ניתן למצוא את המשתנים הסביבתיים שההתנהגות היא פונקציה שלהם. "מחקר זהיר של הספר הזה.. חושף..

שהמסקנות שהושגו במעבדות של התאורטיקנים, הן אמנם מקוריות, אך יכולות להיות מושמות על התנהגות מורכבת רק בצורה הכי שטחית שאפשר". חומסקי דיבר על מושג הגירוי, ואמר: "אם נסתכל על כסא אדום ונגיד 'אדום', התגובה היא תחת השליטה של גירוי ה'אדימות'. אם אנחנו אומרים 'כסא', זה תחת השליטה של גירוי ה'כיסאיות" chariness, וכך לכל סוג של תגובה. המילה "גירוי" איבדה את כל האובייקטיביות בשימוש שלה. גירוי הוא לא עוד חלק מהעולם הפיזיקלי החיצוני, הוא מונע מתוך האורגניזם. אנחנו מזהים גירוי כשאנחנו שומעים את התגובה אליו".

ביקורת נוספת היא ביקורת אנטי אמפיריציסטית, שטוענת שלא ניתן לאפיין את ההתנהגות הלשונית של בני אדם באמצעות המונחים הבהוויוריסטים ובעזרת חוקי למידה מתוך דוגמאות.

ביקורת שלישית היא שבכל שההתנהגות מורכבת יותר, אנחנו נוטים לחפש הסבר לאותה התנהגות – כאשר פעמים רבות ההסבר הוא תיאור המנגנון שעומד בבסיס הקורלציה בין הגירוי לתגובה (והוא אינו בהכרח ניתן לתצפית).

שיעור 11 – מהו הסבר קוגניטיבי

הסבר מדעי –

- מהו המוסבר?
- מהם המסבירים?
- מהו הסבר? מה מקשר בין המסבירים למוסבר?

מודלים / גישות להסבר מדעי

- המודל הדדוקטיבי-נומולוגי
- אנליזה פונקציונלית
- הגישה הסיבתית-מכניסטית

המודל הדדוקטיבי-נומולוגי

בנוגע לשאלה "מהו הסבר מדעי טוב", גישה זו טוענת שהסבר הוא טיעון. המסקנה של הטיעון היא המוסבר, וההנחות הן המסבירים. הטיעון צריך להיות תקף ונאות, ואחת ההנחות חייבת להיות חוק טבע. להלן סכמה של הסבר בגישה הדדוקטיבית-נומולוגית:

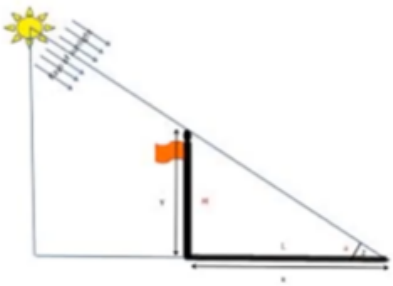
- $E_1, \ldots, E_n$ תנאי התחלה
- $L_1, \ldots, L_k$ חוקים

$O_1, \ldots, O_m$ תופעות נצפות

תנאי ההתחלה והחוקים יסבירו את התופעות הנצפות אם אפשר להסיק מהם את התופעות הנצפות באופן לוגי. לדוגמא, למה אורך הצל הוא 10 מטר?

- גובה המוט הוא 10 מטרים
- השמש היא בזוית של 45 מעלות מעל האופק
- חוקי אופטיקה: אור נע בקווים ישר

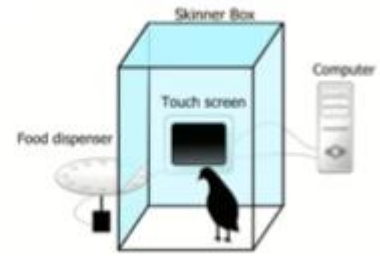
- אורך הצל הוא 10 מטר



דוגמא להסבר דדוקטבי-נומולוגי לתופעה התנהגותית:

- נתונים לגבי תכיפות הניקור של עיגול אדום וריבוע לבן בזמן נתון.
- "חוק האפקט": תכיפות של ניקור עיגול אדום עולה בהינתן גירוי מחזק.

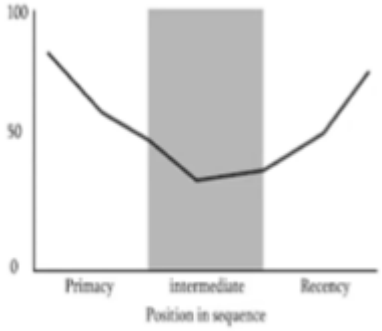
- תכיפות חדשה של ניקור עיגול אדום.



דוגמא להסבר דדוקטיבי-נומולוגי לתופעה קוגניטיבית:

- יוסי קיבל סדרה של 30 מילים והתבקש לאחר 5 דקות לשלוף אותן מהזכרון.
- Recency effect: בני אדם זוכרים יותר מהמילים האחרונות ברשימה.

יוסי שלף יותר מילים שהופיעו בתחתית הרשימה.



בעיות של המודל הדדוקטיבי-נומולוגי:

- ישנם תיאורים שמקיימים את סכמת D-N אך אינם הסברים.
- ישנם תיאורים שאינם מקיימים את סכמת D-N אך הם כן הסברים: לדוגמא, הסברים בפסיכולוגיה (ומדעים אחרים) לא פונים בד"כ לחוקים ותנאי התחלה, אלא מסבירים חוקים ("אפקטים")?.
- **הסברים הם א-סימטריים** – למרות שבלוגיקה אם א' גורר את ב', אז לא-ב' גורר את לא-א'. אבל במציאות זה לא עובד ככה. אם נחזור לדוגמא עם הצל, בכיוון הראשון ראינו שאורך המוט וחוקי האופטיקה מסבירים למה אורך הצל הוא 10 מטר. אבל בכיוון ההפוך, אמנם אפשר לחשב באמצעות

אורך הצל וחוקי האופטיקה את אורך המוט, אבל הם לא מסבירים אותו. כלומר, הם לא מסבירים למה אורך המוט הוא דווקא 10 מטר (התשובה לכך היא שמישהו החליט לשים שם מוט באורך 10 מטר. כלומר הסיבה היא מי ששם אותו שם, לא הצל) – כלומר, הטיעון אינו סימטרי (לא עובד בכיוון ההפוך). זאת למרות שזה הסבר שמבוסס על טיעון תקף, ואחת ההנחות היא חוק טבע. כלומר, ההסבר מקיים את התנאים של הסבר נומולוגי, אך עדיין אין ספק שלא מדובר בהסבר טוב (כלומר אין ספק שהסיבה לכך שאורך המוט הוא 10 מטר היא לא בגלל שאורך הצל הוא 10 מטר).

- **אפקטים** – זוהי ביקורת של קאמינס, שטען שאפקטים אינם חוקים (במובן החמור של הפיזיקה), אלא הכללות סטטיסטיות שיש להן יוצאים מן הכלל. "חוקי אפקט" (כמו התיאור של האפקט של הגברת תופעת הניקור אצל היונה בניסוי הבהוויוריסטי) אינם הסברים (אם כבר, הם בעצם דורשים הסבר).

בגלל הבעיות האלו, המודל הדדוקטיבי נומולגי אינו מספיק בפני עצמו בשביל להגדיר מהו הסבר. כלומר צריך להוסיף עליו עוד תנאים על מנת שיתקיים הסבר מדעי.

אנליזה פונקציונלית

קאמינס על מדעי הקוגניציה – המטרה שלהם היא להסביר אפקטים, סדירויות, דיספוזיציות ובעיקר יכולות. האמצעי לכך הוא הסברים קוגניטיביים, שהם בד"כ דקומפוזיציונליים (אנליזה פונקציונלית). לא מדובר באותה אנליזה פונקציונלית שדיברנו עליה בהרצאה הקודמת בהקשר של בהוויוריזם פזיכולוגי. האנליזה הפונקציונלית שקאמינס מדבר עליה היא לקחת חוק מסוים, ולעשות אנליזה למרכיבים שלו על מנת להגיע להסבר. כלומר, הסברים דקומפוזיציונליים מסבירים את היכולת של מערכת, באמצעות היכולת של יכולות בסיסיות יותר של המערכת (תת-מערכות) והקשרים הסיבתיים ביניהם. במילותיו של קאמינס – "אנליזה פונקציונלית מורכבת מאנליזה של דיספוזיציות (נטיות) בתור מספר של דיספוזיציות אחרות שהן בעייתיות פחות".

דוגמאות להסברים דקומפוזיציונליים – אפיית עוגה (ניתן להסביר עוגה באמצעות פיצול התהליך לתתי-תהליכים – חימום התנור, ערבוב החומרים וכו'), או שימוש בתרשימי זרימה שמפרקים תהליך ראשי לתתי-תהליכים.

תנאים על אנליזה פונקציונלית –

- היכולות של המערכת ותת-המערכות המרכיבות שונות במידה מספקת זו מזו.
- יכולות תת המערכות "פשוטות" יותר מיכולות המערכת הראשית.
- המערכת "מסובכת מספיק".

בחלק מהמקרים, חלק מהתת-יכולות הן יכולות של תתי-מערכות, ובחלק מהמקרים אלו יכולות של המערכת הראשית עצמה. קאמינס טען שההסברים האלו מתאימים לקוגניציה, פסיכולוגיה, ביולוגיה ועוד – ולא התחייב לכך שהסברים קומפוזיציונליים יתאימו לכל סוג של מדע.

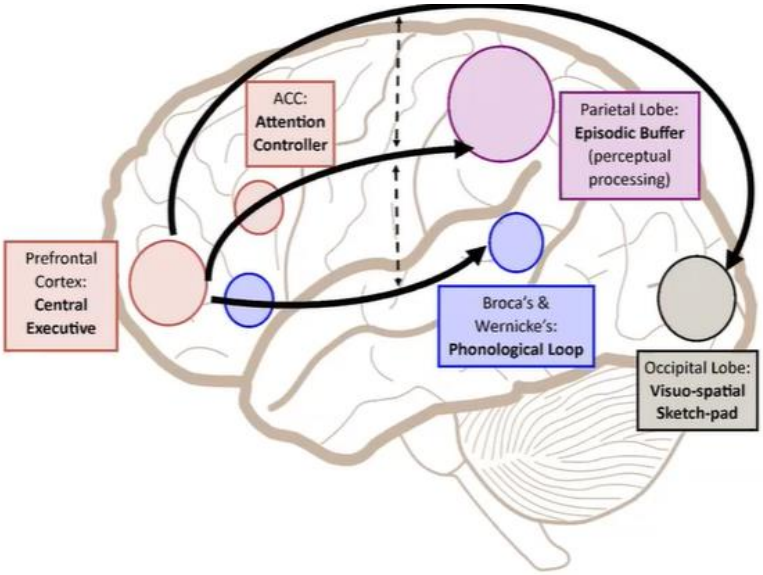
שאלות לגבי ההצעה של קאמינס (אנליזה פונקציונלית) –

- האם תפקיד מדעי הקוגניציה הוא הסברים?
- כיצד מתקשרים הסברים קוגניטיביים להסברים נוירולוגיים? (בזה נעסוק בהרצאה הבאה)
- האם הסברים קוגניטיביים הם אנליזות פונקציונליות? יש הסבורים שלא:
  - **התנגדות ראשונה:** הסברים קוגניטיביים הם הסברים **לא** דקומפוזיציונליים (חלק מרשתות, גישה דינאמית):

- ישנה ביקורת שאומרת שהסברים במדעי הקוגניציה הם לעתים רבות יותר מורכבים מאנליזות פונקציונליות. לדוגמא, רשתות נוירונים הן מורכבות מכדי להיות מורכבת מאנליזות פונקציונליות – עצם כך שאנחנו מצביעים על תתי יחידות עם יכולות משל עצמן, ומעידים על הקשרים ביניהן, זה לא מספיק כדי לתת הסבר מלא לאופן הפעולה של הרשת. חסרים הסברים "גלובאליים" שמסבירים למה ואיך הרשת מבצעת את הפעולה שהיא מבצעת בצורה נכונה.
- התשובה של התמוכים באנליזה פונקציונלית יגידו שזה שיש מידע נוסף מעבר למידע שקולטת האנליזה הפונקציונלית לא סותר את הלגיטימיות של האנליזה הפונקציונלית עצמה. תשובת המתנגדים לכך היא שהמידע החסר הזה הוא קריטי לטובת מתן הסבר מלא לתופעה.

- **התנגדות שניה:** הסברים קונגיטיביים הם מכניסטיים. גישה זו מקבלת את הרעיון שאנחנו עושים דקומפוזיציה לתופעה, אך סבורים שזה לא מספיק, ואנליזות פונקציונליות אינה הסבר מדעי מלא.

- המידע החסר הוא תכונות מבניות של המערכת (איפה הן קיימות, ביצד הן ממומשות – לדוגמא מיקום היכולת הקוגניטיבית במוח), ובלי המידע הזה לא קיים הסבר מדעי מלא לתופעה. התוספת הזו לאנליזה הפונקציונלית נקראת הסברים מכניסטיים.
- אנליזה מכניסטית של מערכת קוגניטיבית נראית דומה מאות לאנליזה פונקציונלית – אנחנו מתכלים על תתי מערכות ועל היכולות שלהן והקשרים ביניהן, רק בשונה מאנליזה פונקציונלית, זה נעשה תוך התחשבות במיקום בו היכולת ממומשת (במוח).
- תומכים של **הגישה המכניסטית** יטענו שהרכיב של התכונות המבניות / מימושיות הוא קריטי למתן הסבר מלא, בעוד מתנגדים של הגישה יטענו כי אמנם הרכיב הזה חשוב לאישוש של הסבר קיים, אך זהו לא חלק הכרחי בהסבר עצמו.



לסיכום, הסבר קוגניטיבי הוא ככה"נ לא הסבר דדוקטיבי-נומולוגי (הביקורת של קאמינס קטלה אותו בצורה די טובה). יתכן כי מדובר באנליזה פונקציונלית (שנראית די מתאימה לתיאור תופעות קוגניטיביות), אך רבים יטענו שהגישה הזו אינה מספיקה (מכניסטים, לדוגמא, יטענו שיש להוסיף אלמנט מימושי להסבר). יתכן כי מדובר בהסברים מסוגים אחרים – כמו הסברים דינאמיים או סביבתיים.

שיעור 12 – רמות הסבר במדעי המוח והקוגניציה

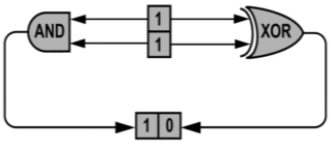
יש שני סוגי היררכיות של רמות (levels):

1. **רמות תיאור / ניתוח**: רמות המתארות את אותן ישויות, רק מהיבטים שונים או תכונות שונות.
2. **רמות אירגון**: רמות המתארות ישויות שונות, כאשר הישויות ברמה הנמוכה הם חלקים של ישויות ברמה גבוהה יותר.

רמות תיאור / ניתוח

בד"כ כשמדברים על רמות תיאור בתחום של מדעי המוח או הקוגניציה, אצל רוב הכותבים בנושא נמצא חלוקה לשלוש רמות. זוהי החלוקה של פילישין, ניואל ודנט (שנות ה-80):

- **רמה סמנטית** (או רמת הידע\ רמה אינטנציונליות): מסבירה מה אנשים (או ישויות קוגניטיביות אחרות כמו בינה מלאכותית) עושים. זוהי הרמה הגבוהה ביותר.
- **רמה סימבולית** (או רמה תחבירית / אלגוריתמית / עיצובית / פונקציונלית): קובעת איך התוכן הסמנטי מקודד ע"י ביטויים סימבוליים. היא גם מפרטת את המבנה של הביטויים האלו, ואת הביטויים הרגולריים שעליהם מתבצעת המניפולציה.
- **רמה פיזיקלית** (או רמה ביולוגית / מימושית): מפרטת איך הביטויים הסימבוליים ממומשים במערכת פיזיקלית (למשל המוח). זוהי הרמה הבסיסית ביותר.



דוגמא – Ned Block's Addition Machine – מכונה שעושה חיבור של שתי ספרות. היא מקבלת שני ארגומנטים בקלט (בציור – 1 ו-1), הם עוברים מצד אחד דרך שער XOR (Exclusive Or, כמו הכמת הלוגי "או" רק שזה "או חזק", כלומר היא תחזיר 1 רק אם רק אחד מהארגומנטים הוא אמיתי, ו-0 אם שניהם אמיתיים או שניהם שקריים), ומצד שני דרך שער And (שמחזיר 1 רק אם שניהם אמיתיים, ו-0 בכל מקרה אחר).

אפשר להציג את המבנה הזה בשלוש רמות שונות:

Physical			Syntactic			Semantic (*plus*)		
Input 1	Input 2	Output	Input 1	Input 2	Output	Input 1	Input 2	Output
5-10V	5-10V	(5-10V, 0-5V)	*1*	*1*	*10*	1	1	2
5-10V	0-5V	(0-5V, 5-10V)	*1*	*0*	*01*	1	0	1
0-5V	5-10V	(0-5V, 5-10V)	*0*	*1*	*01*	0	1	1
0-5V	0-5V	(0-5V, 0-5V)	*0*	*0*	*00*	0	0	0

ברמה הפיזיקלית – נניח שאנחנו מממשים את המערכת הזו במערכת חשמלית, אז המימוש הפיזיקלי יכול להיות לפי המתח של הקלט הראשון והמתח של הקלט השני.

ברמה הסימבולית (סינטקטית) – מרכיבה ספרות לפי הקלט הפיזיקלי. בדוגמה פה, נותנים 1 לקלט גבוה מ-5 וולט ו-0 לקלט נמוך מ-5 וולט.

ברמה הסמנטית – הפירוש שאנחנו נותנים למה שהרמה הסינטקטית מוציאה. בדוגמה פה, ברמה זו מבוצעת פעולת החיבור עצמה ע"ב הקלט מהרמה הסינטקטית. כלומר הפלט של רמה זו יהיה 0, 1 או 2.

כלומר, קיבלנו דרך לתאר את אותה המערכת משלוש "נקודות מבט" שונות. המחשבה היא שאנחנו עושים את אותו הדבר בקוגניציה.

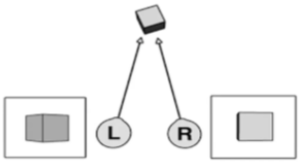
דיויד מאר – חי בין 1945-1980, היה פרופסור לפסיכולוגיה ב-MIT. פרסם תאוריה מאוד מקיפה על תהליך הראייה, מפגיעת הפוטונים ברשתית ועד יצירת ה-mental image המורכב במוח. הוא נתן את תהליך הראיה בתור דוגמה לחלוקת תהליך קוגניטיבי לרמות. מאר הציע חלוקה ל-3 רמות:

1) **הרמה החישובית** – הרמה המופשטת ביותר. זוהי הרמה של מה המערכת עושה ולמה. היכולת הקוגניטיבית נתונה ע"י פונקצית התאמה, והרמה מסבירה את הקשר בין הפונקציה המחושבת לבין משימת עיבוד המידע של המערכת. הקשר הוא יחס בין המערכת (לדוג' מערכת הראייה) לבין הקלט שהיא מקבלת (לדוג' שדה הראיה בעולם החיצוני).
2) **הרמה האלגוריתמית** – מאפיינת את המערכת של הייצוגים שבשימוש, ואת האלגוריתם שממיר קלט לפלט.
3) **הרמה המימושית** – מפרטת איך הייצוג והאלגוריתם ממומשים פיזית (במוח).

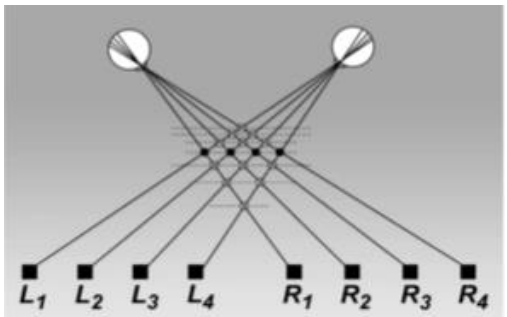
מה ששונה בתאוריה של מאר לעומת החלוקה ה"סטנדרטית" לרמות היא שהוא קובע את הרמה החישובית בתור הרמה הראשונה, בתור האחרים מבליעים אותה בתוך הרמה השניה או מתעלמים ממנה. מה שייחודי ברמה החישובית של מר לעומת החלוקה ה"סטנדרטיים", היא שהוא התעקש על האלמנט החישובי עצמו, ש"הוזנח" לטענתו בחלוקות אחרות שהוצעו ע"י חוקרים אחרים. מאר התמקד בפונקציה שהמערכת מבצעת, שהיא אלמנט חישובי. מה שעניין את מאר הוא לא רק הפונקציה המתמטית של מה המערכת עושה, אלא גם הסיבה שבגללה היא עושה את זה דווקא באופן הזה (**מה** וגם **למה**).

דוגמא ליישום התאוריה של מאר – בעיית ההתאמה

בעיית ההתאמה היא ההתאמה בין אלמנטים מהתמונה השמאלית והתמונה הימנית. כלומר הבעיה היא שהאובייקטים שאנחנו רואים נמצאים במיקום שונה מעט ביחס לעין השמאלית לעומת ביחס לעין הימנית:



זה קורה כי היכולת שלנו לראות עומק מושפעת מאוד מהפער שבין תמונות הרשתית. הפער גדול יותר כאשר אנחנו מתמקדים באובייקט שקרוב אלינו, וקטן יותר כאשר האובייקט רחוק. לפי מאר, הפער הזה הוא זה שמאפשר לנו לבנות ראיית עומק (ראייה תלת מימדית של העולם). מסתבר שבעית ההתאמה היא ממש מורכבת, כי לפעמים יש דו משמעות או רב משמעות בעיבוד האובייקטים:



בדוגמא הזו, יש הרבה אפשרויות לחיבור בין כל אחת מהנקודות השחורות שנתפסות בעין הימנית, לעומת הנקודות השחורות שנתפסות בעין השמאלית. העיגולים בהצלבה בין הקווים מייצגים את כל האפשרויות השונות להתאמה (16 אפשרויות שונות להתאמה). הקווים האופקיים מייצגים את הפער.

כיצד מושגת ההתאמה ברמה החישובית?

לפי מאר, הפונקציה הממוחשבת (כלומר בינה מלאכותית שמנסה לחקות ראיה אנושית) צריכה לקיים שני אילוצים:

1) <u>יחידות –</u> נקודה שחורה מתמונה אחת יכולה להתאים לכל היותר נקודה אחת מהתמונה השניה.
2) <u>רציפות –</u> הפער משתנה ברציפות לאורך התמונה.

בתמונה למעלה, אם נסתכל על כל האפשרויות השונות להתאמה, רק האפשרויות שצבועות בשחור עונות על שני האילוצים האלו, וזוהי באמת גם ההתאמה הנכונה.

בנוגע לשאלת ה"למה" – מאר כתב כי "השאלה האמיתית שצריך לשאול היא למה שמשהו כזה יעבוד. זה קורה מהסיבה הפשוטה שאם אנחנו מסתכלים רק על זוג תמונות, אין שום סיבה למה L1 לא יתאים ל-R3 וכו'...". כלומר, אנחנו יודעים שמערכת הראייה מבצעת את ההתאמה בהתאם לעולם הפיזיקלי, אך מאר סבר כי תפקידו של המדען הקוגניטיבי היא גם לשאול מדוע היא עושה זאת באופן בו היא עושה זאת (כלומר למה נבחרה התאמה אחת ולא התאמות אחרות). מאר אומר שכדי לענות על השאלה הזו, אנחנו צריכים להבין את השדה הפיזיקלי שסביבנו עובד. הוא אומר שהסביבה הפיזיקלית מכילה בתוכה את אותם אילוצים (יחידות ורציפות).

מאר אמר שהנסיון לבודד אילוצים שהם גם חזקים מספיק כדי לגרום לתהליך להיות מוגדר ואמיתי בעולם, הוא משימה מרכזית והכרחית של המדע – כלומר להבין מהם האילוצים הפיזיקליים שגורמים למערכת הראיה לעבוד כפי שהיא עובדת, לדוגמא. הוא אמר שבמהלך יצירתה של תאוריה חישובית, חלק משמעותי הוא הגילוי של הנחות אימפליציטיות שמושמשות ע"י מערכת הראייה - ההנחות שנוגעות לסביבה החיצונית, שהמערכת משתמשת בהן כשהיא מבצעת את החישובים שלה. מאר הציג אילוצים פיזיים שגורמים להיווצרותם של האילוצים של מערכת הראיה (מתוך הנחה שמערכת הראיה התפתחה אבולוציונית כדי להתאים לאילוצים הפיזיקליים האלו):

- <u>יחידות</u> נוצרת ע"י עקביות במיקום בחלל, כלומר "נקודה נתונה בחלל הפיזי היא בעלת מיקום ייחודי בכל זמן נתון".
- <u>רציפות</u> נוצרת ע"י הפער בין הראיה בעין השמאלית לעין הימנית – "הפער משתנה באופן חלק בכל מקום" ומושפעת מהאחידות של החומר – "חומר הוא אחיד, הוא מופרד לאובייקטים נפרדים, והשטח של כל אחד מהאובייקטים הוא יחסית חלק ביחס למרחק שלהם מהצופה".

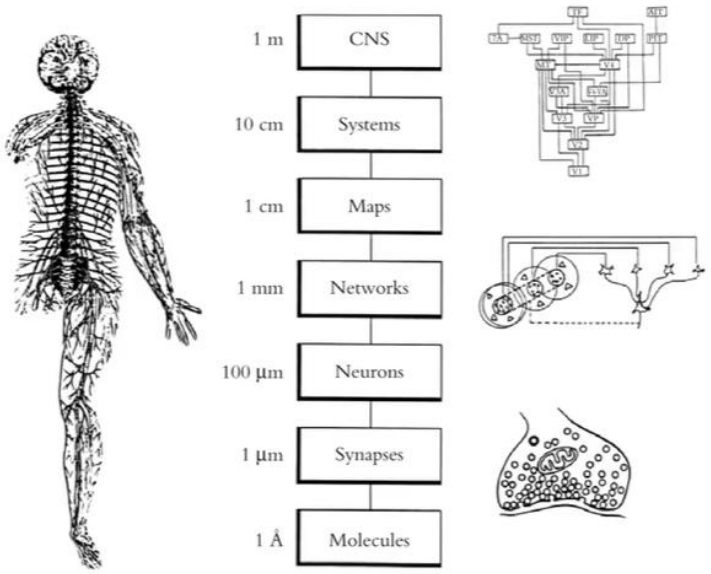
לכן, הדגש של מאר היה שאם אנחנו רוצים ללמוד על מערכת קוגניטיבית (לדוגמא, ראיה) כדאי לנו להתחיל מהסביבה החיצונית, שכן היא יוצרת אילוצים ברורים שהמערכת הקוגניטיבית פועלת (ומחשבת) לפיהם.

כיצד מושגת ההתאמה ברמת האלגוריתם?

מאר ותלמידו פוגיו הציעו אלגוריתם מתאים של רשת נוירונים, בו כל נוירון מייצג אפשרות התאמה כלשהי. האלגוריתם יזהה אם יש התאמה ברורה ויפסול אפשרויות אחרות לפי אילוצי היחידות והרציפות. הרשת ממשיכה לפעול עד שהיא מגיעה לפתרון אופטימלי בהתאם לאילוצים.

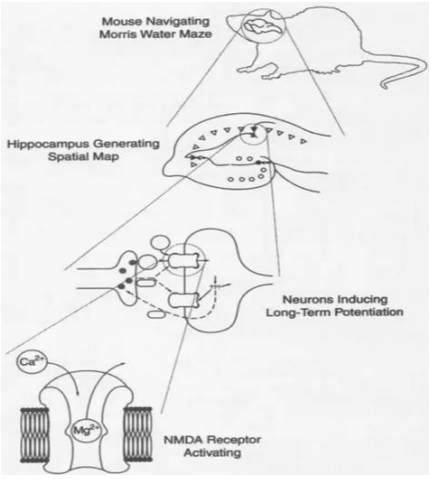
איפה ממומשת ההתאמה? ידוע היכן ההתאמה ממומשת במוח (אזורים מסוימים בקורטקס הראייה).

להלן דוגמא שמציגה ארגון של המוח לרמות:



הרעיון הוא שאפשר לחשוב על רמה מסוימת בתור כזו שמורכבת מהרמות שנמצאות מתחתיה. במקרה כאן (של המוח), ככל שהישות יותר נמוכה בהיררכיה, היא גם יותר קטנה פיזית.

היררכיה פונקציונלית – כפי שדיברנו בהרצאה הקודמת על אנליזה פונקציונלית, אפשר להסתכל על תתי יחידות ואת הקשרים ביניהם, ובאמצעות זאת להסביר את הפעולה של היחידה הראשית.



היררכיה מכניסטית – כפי שדיברנו בהרצאה הקודמת על מכניזם, שהיא אותו דבר כמו ההיררכיה הפונקציונלית, רק שהיא כוללת גם את המימוש הפיזי / מבני של כל אחד מהמרכיבים של המערכת. דוגמא להסבר מכניסטי – הסבר של קארל קרייבר מ-2007 על יכולת הניווט של עכבר במבוך.

ההסבר הוא באמצעות תאי עצב שנמצאים בהיפוקמפוס (תאי מיקום, שכל אחד מהם מייצג מיקום מסוים במפה של העכבר). אפשר לרדת להיררכיה "קטנה" יותר ולשאול איך פועל כל תא, ואז איך פועלת הסינפסה של התא, וכך הלאה עד הרמה המולקולרית. זהו הסבר מכניסטי כיוון שממש מדובר בהיכן ממוקמת כל היררכיה באופן פיזי במוח.

בשתי הגישות (פונקציונלית ומכניסטית), כל שלב בהיררכיה הוא למעשה רמת הסבר. ניתן לדון בכל רמת הסבר בנפרד (לדוגמא לדון בביצד פועלים תאי מיקום), ואפשר לדבר על הפעילות של המערכת כולה באמצעות דיון בפעולה של תתי המערכות שלה והקשרים ביניהן.

מה לגבי הרמות החישובית והמימושית? המרכיבים החישוביים לא נראים כמורכבים מהחלקים המימושיים, כלומר הרמה המימושית היא לא "מתחת" לרמה החישובית. כמו כן, המרכיבים החישוביים לא נראים כמו סקיצה של מכניזם (שחסרים בה המרכיבים המימושיים). השאלה של "מה הקשר בין הרמה החישובית והמימושית" נשארת פתוחה – הן לא יכולות להיות ממומשות באותה היררכיה, ומצד שני הן לא באמת היררכיות נפרדות.