סם ש. רקובר

**כיצד ניתן להבין התנהגות אם לא מבינים את הסברה? נקודת ביקורת מתודולוגית על הפסיכולוגיה הקוגניטיבית**

**How can behavior be understood if its explanation is not comprehended? A methodological critique of cognitive psychology**

 **Abstract**

המאמר הנוכחי דן במודלים של מחשב, הDeep AI models, שבגלל היותם מאוד מתוחכמים ומסובכים קשה מאוד להבין איך הם פועלים, ואם הבנתם לא קיימת הרי שקשה מאוד להבין בעזרתם את התופעות ההתנהגותיות שלמטרת הסברן הם נוצרו. זה לאמתו של דבר המצב הנוכחי של הפסיכולוגיה הקוגניטיבית המושתתת על האנאלוגיה לפעולות המחשב: אם אין מבינים את התוכנות הכי מתוחכמות והכי מצליחות, ה Deep AI models המיועדות להסביר התנהגות, אזי לא ברורה מהי המוטיבציה להמשיך ולבנות מודלים כאלה המבוססים על תהליכים חישוביים שאינם מביאים להבנתה של ההתנהגות? המאמר דן בשאלה זו כשהוא מתאר בקיצור את הנחות הבסיס לפרוצדורה (מודל) של הסבר, את בעיית אי-ההבנה של Deep AI models, את האימפליקציות המתודולוגיות-פילוסופיות הקשורות בבעיה זו, ומסיים בהשלכות של אי הבנה זו על הפסיכולוגיה הקוגניטיבית.

**Key words**: explanation, understanding, explainability of AI models, cognitive psychology.

**הקדמה**

הפסיכולוגיה הקוגניטיבית, שבעשרות השנים האחרונות הינה הגישה השלטת בפסיכולוגיה, מושתתת על האנלוגיה בין תהליכי פעולתו של המחשב לבין תהליכי תחושה, תפיסה, חשיבה וכדומה המתרחשים בראשו של היחיד (ואף בראשם של בעלי חיים). למשל, הזיכרון האנושי נתפס כבנוי על תהליכים מקבילים לתהליכים המתרחשים במחשב: קליטה, קידוד, אחסון, ושליפה של אינפורמציה. אנלוגיה זו מציעה גם דרך לפתרון בעיית הגוף-נפש: כמו התוכנות במחשב שמקבלת בסופו של דבר עיגון פיזיקאלי בזרם החשמלי (יש או אין מתח חשמלי) כך, סבורים מספר חוקרים, ניתן לעגן את התהליכים המנטליים בנוירופיזיולוגיה של המוח (ראו דיון אצל למשלDennett, 1979; Rakover, 2018 ).

הבעיה המתעוררת כאן היא שדווקא מודלים מתקדמים ביותר של אינטליגנציה מלאכותית (AI) להבנת התנהגותו של היחיד, Deep AI models, הינם כה מתוחכמים ומסובכים, שאין מצליחים להבין את דרכי פעולתם והם לאמתו של דבר נחשבים ל Black Boxes, כלומר, מודלים אלה נתפסים כמערכת הכוללת פלט שנכנס לקופסה שחורה (אטומה), עובר איזשהו עיבוד אינפורמציה לא-מובן המסתיים בפלט שקשה מאוד להסבירו. זה מצב משונה ביותר, משום שהוא מעורר מיד את שאלת כותרת המאמר: כיצד ניתן להבין התנהגות אם לא מבינים את הסברה? אם דווקא המודלים המתקדמים והמוצלחים ביותר, מודלים המבוססים על תהליכי מחשוב, אינם מובנים, אזי כיצד תצליח הפסיכולוגיה הקוגניטיבית לממש את יעודה, להסביר התנהגות? כאן עולה מכשול רציני העומד בדרכה של גישה פסיכולוגית זו. (במאמר הנוכחי אני אשתמש ב Deep AI models כשם כללי למספר גדול של סוגי תוכנות מחשב כמו machine learning and neural networks, שכאן לא המקום לעמוד על ההבדלים החשובים ביניהם.)

המאמר הנוכחי בנוי מהקדמה קצרה המסכמת את הנחות הבסיס לפרוצדורה (מודל) של הסבר, תיאור של בעיית אי-ההבנה של Deep AI models, דיוןבאימפליקציות מתודולוגיות-פילוסופיות הקשורות בבעיה זו, ודיון בהשלכות של אי הבנה זו על הפסיכולוגיה הקוגניטיבית.

**הבסיס לפרוצדורה (מודל) של הסבר**

המדע המודרני מתאפיין בניסיון להגשים את המטרה החשובה הבאה: להסביר באופן רציונאלי את התופעות הבלתי מובנות. זו מטרה שהלכה והתבהרה וחשיבותה הלכה וגברה מימי קופרניקוס וגלילאו עד היום. במתן הסבר יש לענות על כמה דרישות מתודולוגיות הרווחות במדעים (כולל גם מדעי החברה). לצרכי המאמר הנוכחי אני אציין את הדרישות הבאות (see Rakover, 2018):

1. *לשם מתן הסבר לתופעה בלתי מובנת יש להשתמש במודל הסבר, בפרוצדורה מיוחדת המתאימה לסוג התופעה הנידונה.*

דרישה זו מתבססת על כך שהתופעה הנחקרת עברה תצפית העומדת בדרישות המתודולוגיות המתאימות (על התופעה לענות על הדרישה לאובייקטיביות, חזרתיות ופומביות, ראו Rakover, 1990). התצפית עוברת ייצוג בשפה מסוימת (למשל, ייצוג מתמטי) ייצוג שאותו מנסים להסביר על ידי השערה, מודל או תיאוריה רלוונטית. התיאוריה לכשעצמה אינה מסוגלת להסביר את התופעה הנידונה. כדי שההסבר יתאפשר יש ליישם על התיאוריה פרוצדורה של הסבר מתאימה. למשל, כדי להסביר מהו המרחק שיעבור גוף בנפילה חופשית, משתמשים לפי Hempel (1965, 1966) בפרוצדורה הבאה הנקראת Deductive-Nomological (D-N) model: משתי ההקדמות (premises), חוק נפילה חופשית של גופים (חוק גלילאו) וזמן נתון מסוים, מסיקים באופן לוגי (מתמטי) מסקנה מסוימת, כלומר, מנבאים את המרחק שהגוף יעבור בזמן הנתון. אם מתברר באופן אמפירי שאכן הגוף עבר את המרחק המנובא, אזי התופעה מקבלת הסבר בעזרת חוק זה המבוסס על כוח המשיכה של הגרביטציה (ישנם כמובן פרוצדורות, מודלים של הסבר, נוספים שלא אדון בהם כאן. see Rakover, 2018).

 באופן אינטואיטיבי ניתן לעגן את מכלול הפרוצדורות להסבר המופיעות בספרות המקצועית בארבעה סוגים עיקריים של הסבר (ראו דיון אצל Rakover, 2018; Salmon, 1984, 1990):

1. *דוגמה פרטית.* הפרוצדורה להסבר מראה שהתופעה הנידונה הינה מקרה פרטי של חוק או תיאוריה כללית (בין שהתיאוריה הינה דטרמיניסטית או הסתברותית). למשל, נפילה חופשית של 4.9 מטר בשנייה הראשונה של כדור ברזל, היא מקרה פרטי של חוק גלילאו (שניתן להסבירו בעזרת התיאוריה הניוטונית).
2. *סיבתיות.* הפרוצדורה להסבר מראה שניתן להבין את התופעה הנידונה על ידי שיזורה במארג סיבתי. למשל, הופעת מחלת הלוקמיה אצל אדון השימוטו נובעת מקרינה רדיואקטיבית שהוא ספג כתוצאה מהרס התחנה הגרעינית בפוקושימה על ידי גלי צונאמי.
3. *מכניזם.* הפרוצדורה להסבר מראה שניתן להבין את התופעה הנידונה על ידי פירוט המכניזם, התהליך, שמביא בסופו של דבר להופעת התופעה הנידונה. למשל, ניתן להסביר את המכניזם של פנס כיס ליצירת אור על ידי פירוט חלקיו (בטרייה, נורית, וכו') והדרך שבה הם קשורים אחד בשני, כאשר כל חלק וחלק ניתן להסבר על ידי פירוט המכניזם המתפעל אותו. ניתן להסביר זכירה מידית של מספר בן שבעה ספרות על ידי פירוט המכניזם הקוגניטיבי המבוסס על ההבחנה בין short term memory (STM) לבין long term memory (LTM).
4. *אנאלוגיה.* הפרוצדורה להסבר מראה שניתן להבין את התופעה הנידונה על ידי מתיחת קו דמיון בין התופעה הנחקרת לבין תופעה אחרת שהינה מובנת וידועה. למשל, ההשוואה בין פעולת המערכת הקוגניטיבית של האדם לבין פעולתו של המחשב. כנ"ל, הזיכרון של האדם מוסבר באופן כללי על ידי פירוט מנגנוני קליטה, קידוד, אחסון ושליפה, המקבילים למנגנונים אופייניים במחשב.
5. המודלים*, הפרוצדורות להסבר מבוססים על הצדקות הגיוניות ואמפיריות.*

לפי Hempel (1965, 1966) ההצדקות היסודיות למודלים של הסבר שהוא הציע (כדוגמת מודל ה D-N) הינן: רציונאליות ואמפיריות (ראו דיון אצל Rakover, 2018). הדרישה לרציונאליות הינה שהמודל או הפרוצדורה של ההסבר כולל ההסבר עצמו לא יגרמו לסתירות פנימיות, לחוסר קונסיסטנטיות. הדרישה לאמפיריות הינה שמודל, פרוצדורת ההסבר יהיו קשורים באופן איתן לתופעה המוסברת. גם אם ההסבר יהיה מושלם מבחינה לוגית (כלומר, חסר כל סתירות פנימיות) אך מושגיו לא יהיו קשורים לתופעה האמפירית הנחקרת, להסבר לא יהיה כל ערך. מסיבה זו, בין היתר, קשה מאוד להשתמש בתיאוריה הניוטונית כדי להסביר תפיסה פרצופים בבני אדם.

1. *שני תנאים הכרחיים להבנתו של הסבר.*

הבנה של הסבר חייבת למלא שני תנאים הכרחיים (ראו Rakover, 2018, in preparation). לפי התנאי הראשון, הבנה של תוכן מסוים לא תתכן אלא אם כן התוכן הינו מיוצג (או היה מיוצג) בהכרתו של האדם, כלומר, הכרתיות (consciousness) היא תנאי הכרחי להבנה. (הכרתיות אינה תנאי מספיק משום שכדי שהבנה תתרחש צרכים להתממש תנאים רבים אחרים כמו למשל תפקוד תקין של המוח.) לדרישה זו יש חשיבות רבה, משום שהיא מבחינה בין ההסבר עצמו לבין הבנתו של ההסבר. כדי להמחיש זאת נבחן את שתי הדוגמאות הבאות.

אפשר לתאר רובוט, המורה המושלם רובי, שמסוגל ללמד כל סטודנט בסבלנות אינסופית את הפיזיקה הקלאסית (הניוטונית). והנה מתברר שגם הסטודנט הכי איטי של רובי בסופו של דבר מבין את הפיזיקה הקלאסית ומסוגל כתוצאה מהלימוד אצל מורה מושלם זה לפתור את מרבית התרגילים בציון גבוה מאוד. אולם, האם אפשר להניח שרובי המורה המושלם מבין את הסבריו הוא אף כמו האחרון בתלמידיו? תשובתי הינה שלילית, כי רובוט זה אינו מבין לא את השאלות ולא את התשובות שהוא עצמו ספק לתלמידיו. (ב Rakover, 2018 מוצגת סקירה התומכת בטיעון שעדין לא הצליחו להראות שרובוט מתוחכם ומסובך פיתח הכרתיות בדומה לבני אנוש.)

והינה דוגמה נוספת המעידה על כך שברובוט מסובך ומתוחכם ככל שיהיה לא התפתחה הכרתיות. נניח שרובוט-אם מחקה באופן מושלם את ההתנהגות של אם-אנושית. מיד אחרי הלידה רובוט-אם מתנהגת בדיוק כמו אם-אנושית כשמביאים לה את הרך הנולד. יתר על כן, רובוט-אם תדחה בשעט נפש רובוט-יָלוּד בדיוק כמו אם-אנושית. וכאן עולה הנקודה החשובה: אם רובוט-אם הייתה מפתחת הכרתיות בדומה לבת אנוש, סביר להניח שהייתה דוחה אל הילוד האנושי ומחבקת מאמצת אל ליבה את הרובוט-ילוד שהינו בשר מבשרה. אולם אין היא מתנהגת כך, משום שהיא אינה אלא מכונה שלמדה להגיב בדפוס תגובות מסוים כאשר מציגים לה דפוס גירויים מסוים.

 לפי התנאי השני, הבנה של תוכן מתרחשת כאשר עולות שאלות על אודות התוכן הנידון וניתנות תשובות בהתאם לידע ולפרוצדורות למתן הסברים. למשל, התשובה לשאלה הנ"ל על אודות נפילה חופשית של גופים ניתנה בהתאם לידע הרווח באותה תקופה ובהתאם לפרוצדורה המתאימה, במקרה הנוכחי, מודל ה D-N נוסח Hempel (1965, 1966).

1. *לא ניתן להבין את התופעה הנחקרת אם לא מבינים את הסברה.*

זו הינה הנחה המתקבלת מאליה על ידי השכל הישר: בלי להבין את ההסבר לא נוכל להבין את התופעה הנחקרת. יתר על כן, חוסר הבנת ההסבר עלול להביא אותנו לרגרסיה אינסופית. למשל, נתונה לנו תופעה בלתי מובנת P. מציעים ל P הסבר על ידי E1, אולם מתברר שאין מבינים הסבר זה, לכן אנו זקוקים להסבר נוסף E2 שיסביר את ההסבר, את E1. אולם איננו מבינים גם את ההסבר של ההסבר ולכן זקוקים להסבר נוסף E3 ... וכך עד אינסוף (כאן אני מדלג על האפשרות שלפיה מציעים ל P מספר הסברים שונים ולא מובנים). כלומר, כל עוד שאין אנו מבינים את ההסבר ל P לא נוכל להבין את P, וכמובן שללא הבנת ההסבר לא נוכל גם לשפוט אם ההסבר מממש את מטרתו באופן שלם או חלקי.

רעיון זה, שהוא גם מרכז הדיון של המאמר הנוכחי, יומחש על ידי דיון בבעיות הקשורות באי הבנת תוכנות מחשב מסובכות ועמוקות של אינטליגנציה מלאכותית ( Deep AI models ).

 **אי הבנה של Deep AI models**

הדוגמאות שאני אתייחס אליהם בדיון הנוכחי קשורות לDeep AI models , והן כוללות סוגים שונים של תוכנות programs, softwares)), מתוחכמות ומסובכות שמטרתן, בין היתר, הסבר זיכרון אנושי, זיהוי והיכר של פרצופים, עשיית החלטות, קטגוריזציה, כולל גם מתן דיאגנוזה רפואית. (ראו למשל, Elmahmudi & Ugail, 2019; Kumar, A., 2021; Samek, Montavon, et al, 2019; Samek, & Muller, 2019; Samek, Wiegand, et al, 2017; Zhou, Bau, et al, 2019 ). סוג זה של תוכנות,Deep AI models , מבוסס על רשתות מאוד מסובכות, שמכילות מספר עצום של מרכיבים: input layer, hidden layers (which includes a huge amount of nodes) and output layer . אולם על אף הצלחתן הגדולה בניבוי התצפיות, מתברר שעולה בעיה גדולה בהבנת דרכי פעולתן. וכך כותבים :Samek, Wiegand, et al (2017)

“However, although these models reach impressive prediction accuracies, their nested non-linear structure makes them highly non-transparent, i.e., it is not clear what information in the input data makes them actually arrive at their decisions. Therefore, these models are typically regarded as *black boxes.*”(p. 1)

לתופעה זו כמובן השלכות מרחיקות לכת, כגון, יצירת אי אמון בתקפות הפלטים (החלטות, תגובות וכו) של התוכנות הנידונות. וכך כותבים Samek, & Muller (2019):

“Despite the revolutionary charter of this technology, challenges still exist … lack of transparency and explainability, which reduces the trust in and the verifiability of the decisions made by an AI system.” (p. 6).

חלק נכבד מהמאמרים הקשורים בנושא של explainability AI models מציעים תוכנות-הסבר המיועדות להציע הבנה ל Deep AI models . סוגים אלה של תוכנות-הסבר מציעים, בין היתר, תשובה לשאלה מהי התרומה של קבוצות צְמָתִים (nodes) מסוימות ביצירת הפלטים של התוכנות. למשל, זיהוי של ספל קפה או תרנגול מתבסס על איתור קבוצות nodes שמתארת את הצורה העגלגלה של פי הספל או הכרבולת האדומה של התרנגול. במקרה זה ניתן לומר שההסבר מסתמך על מציאת סיבה חשובה לפלט (ראו Samek, Wiegand, et al, 2017); ניסיון לזהות פנים בעזרת אינפורמציה חלקית של הפרצוף כאשר מאמנים את התוכנה על ידי אינפורמציה חלקית של פרצופים (בהשוואה לאי אימון מסוג זה). גם במקרה זה ניתן להציע שנמצאה סיבה חשובה, חלק בפנים, לזיהוי פרצופים (ראו Elmahmudi & Ugail, 2019); (כאן מן הראוי לציין שטבעו של אוסף הנתונים, שבעזרתו מאמנים את הרשת הלומדת, עלול לקבוע בתוכנה הטיות מסוימות, כמו למשל, במקרה שנתוני האימון מבוססים על תגובות של גברים, עלולה הרשת הלומדת להעדיף גבר על פני אישה בתחרות על משרה פנויה.) שימוש ב meta-explanations המבוסס על סיכום של מספר הסברים אינדיבידואליים היוצרים דפוס הסברי, כלומר, ההסבר מסתמך על סכמה או הכללה שבעזרתה מבינים את הפלט (ראו Samek, & Muller, 2019). למרות שתוכנות-ההסבר עוזרות להבנת ה Deep AI models עולה גם לגביהן השאלה המטרידה: האם תוכנות-הסבר אלו עצמן מובנות? שאלה המעלה את האפשרות לרגרסיה אינסופית של הבנת ההסבר – נקודה שהעליתי לעיל. וכך כותבים Samek, Montavon, et al (2019) בקשר לעניין זה במבוא לספרם:

“More recent deep learning based neural networks provide far superior predictive power, but at the price of behaving as a ‘black-box’ where the underlying reasoning is much more difficult to extract.” (p. v).

והם ממשיכים ומעלים פקפוקים בקשר לתוכנות-ההסבר המיועדות לספק הבנה ל Deep AI models:

“However, many questions remain on whether these explanations are robust, reliable and sufficiently comprehensive to fully assess the quality of the AI system.” (p. v).

במאמר הנוכחי אתייחס ל Deep AI models ולתוכנות-ההסבר שאינן מובנות, כלומר, אתייחס לכל אותן התוכנות המתוחכמות והמסובכות שנחשבות ל black-boxes.

**אימפליקציות מתודולוגיות-פילוסופיות של אי הבנת Deep AI models והסבריהן**

במאמר זה איני מעוניין להתעכב על כיוון מחקרי חשוב זה שתואר בקיצור נמרץ לעיל (בניית תוכנות לביאור ולפירוש ה Deep AI models), אלא אעמוד על האימפליקציות המתודולוגיות-פילוסופיות הקשורות בבעיית אי ההבנה של ההסבר הקשור ב Deep AI models ובתוכנות המתיימרות לפרשן. תחילה אדון בדוגמה של הבנה מלאה ולאחר מכן אדון בדרך אפשרית להתייחסות לתוכנות לא מובנות.

***הבנה מלאה של ההסבר***

נניח שחוקר התנהגות אנושית החי בתקופת ניוטון ואשר התרשם עמוקות מגישתו המכניסטית לפתרון בעיות פיזיקאליות, הציע גישה תיאורטית הקובעת שניתן להסביר באופן מכניסטי מושלם גם את התנהגות האדם. כדי לתמוך בתיאוריה המכניסטית-התנהגותית שלו ולהמחישה בנה אותו חוקר רובוט המכונה רוברט והמסוגל לחקות באופן מושלם שתי התנהגויות פשוטות יחסית: למזוג ספל תה ולחתום את שמו על פיסת נייר. המכניזם שביצע שתי התנהגויות אלו היה בנוי ממערכת מכאנית הכוללת קפיצים, מוטות ברזל, חוטי ברזל, גלגלי שיניים, משקלות וכו. ההסבר של מנגנון זה מתברר שהינו פשוט למדי. תחילה יש למתוח היטב את הקפיץ הנמצא בגבו של הרובוט. לאחר מכן יש לסגור את הידית להפעלת החתימה (או למזיגת ספל הקפה). כתוצאה מכך מופעל המכניזם המתאים שניתן להסבירו בעזרת תרשים המתאר במדויק את התנועות של כל אחד מהמרכיבים ברובוט שגורמים בסופו של דבר לחתימה של שמו, רוברט. התיאור המפורט והמדויק של מכניזם החתימה הוא ההסבר השלם של התנהגותו של הרובוט, הסבר שכל אחד מסוגל להבינו. אולם, האם רוברט הרובוט מסוגל להבין את מעשיו הוא? ברור שאינו מבין דבר, אבל כל בן אנוש אכן מבין.

 לדוגמה זו שתי אימפליקציות חשובות. ראשית, לא רק שה Deep AI models אינן מסוגלות להבין את שהן מבצעות, אלא גם תוכנות-ההסבר (explanatory-programs) שלהן אינן מבינות את שהן מסבירות (כלומר, מבצעות) כי כולן חסרות הכרתיות בדיוק כמו רוברט הרובוט. שנית, בעוד שאנשים מסוגלים להבין כיצד פועל רוברט, בני אדם כולל המתכנתים עצמם אינם מסוגלים להבין את הפעולות המאוד מסובכות של ה Deep AI models וגם את תוכנות-ההסבר שלהן. וכאן עולה השאלה הבאה: כיצד זה ייתכן? הרי בסופו של דבר המתכנתים הם אלו שכתבו את התוכנות הללו ולכן הם חייבים להבין את מה שהם יצרו. אם כך, כיצד זה יתכן? התשובה שלי היא כפולה. ראשית, אי ההבנה נובעת, כאמור, מהסיבוכיות העצומה של תוכנות אלו, ה Deep AI models. שנית, ניתן לראות בתוכנות אלו מסגרות רחבות שבתחומן מתחוללים התרחשויות והתפתחויות הדורשות הסבר שהוא מעבר לכללים שבעזרתם נוצרו ה Deep AI models. אני אדגים רעיון זה בעזרת משחק השחמט (או כל משחק אחר).

 כמעט כל אחד יודע מהם כללי משחק השחמט וכמעט כל אחד שיחק בימי חייו במשחק יפהפה זה. אולם, למרות שכללי משחק זה הם אלה המבחינים בין שחמט לבין משחקי לוח אחרים (כגון, דמקה, שש-בש) לא ניתן להסביר בעזרת כללי משחק זה בלבד את העובדה שבובי פישר היה מגדולי השחקנים בשחמט. כדי להבין את העובדה שפישר היה שחקן מהדרגה הראשונה אנחנו צרכים לקחת בחשבון מספר גורמים חשובים שאינם מתייחסים ישירות לכללי המשחק, כגון, השליטה המושלמת שלו באסטרטגיה ובטקטיקה של המשחק (פתיחות וסיומים), תפיסה מידית של מצב המשחק ותכנון מהלכי שח למרחוק, עצבי פלדה, והיכר מצוין של דרך המשחק של היריב. ובאופן דומה אפשר לומר, שתִכְנוּת ה Deep AI model אינו אלא בבחינת קביעת כללי משחק שבמסגרתם התוכנה הממשית לומדת לשחק, לפעול, כלומר, להשיג מטרות מסוימות, כמו, זיהוי פרצופים, עשיית החלטות, וקטגוריזציה של פריטים. במילים אחרות, ניתן להציע, שסדרת המשוואות שהמתכנתים משתמשים בהן כדי ליצור Deep AI models אינם אלא בחזקת כללים הקובעים מסגרת שבה תתפתח תוכנה נתונה לרמת סיבוכיות כה גדולה שקשה מאוד יהיה להבינה. העובדה שאין תשובה ברורה ומלאה לשאלה כיצד בדיוק לומדת התוכנה וכיצד היא מתפתחת, מעידה שאכן ה Deep AI model אינו אלא בחזקת black-box, ושלכן זקוקים לתוכנות-הסבר שיפרשו מודלים סתומים אלה.

***אפשרות להבנת תוכנה לא-מובנת: ה Deep AI model כתופעה חדשה ודרגות שונות של הבנה****.*

עצם העובדה ש Deep AI model אינה מובנת ושלכן היא נזקקת לתוכנת-הסבר כולל עריכת ניסויים שונים כדי לפענח את טיבעה (ראו למשל, Elmahmudi & Ugail, 2019; Samek, Montavon, et al, 2019) מעלה את האפשרות שבמסגרת כללי היצירה של תוכנות מסובכות אלו נוצרו תופעות חדשות הדורשות הסבר, כלומר, התוכנות עצמן הפכו לתופעות מעניינות התובעות הבנה. בהינתן עובדה זו של אי-הבנת ה Deep AI model כולל תוכנות-ההסבר שלהן, עולה השאלה הבאה: כיצד יש להתייחס לכך שתוכנות אלו אינן מובנות והן בחזקת תופעות חדשות הדורשות בעצמן הסבר?

אפשר לתפוס את ה Deep AI models ואת התוכנות-מסבירות אותם כמספקים הסברים חלקיים, לא מושלמים, כלומר, הסברים המספקים דרגות שונות ולא גבוהות של הבנה (לתופעה הנחקרת או לתוכנה הלא מובנת). למשל, Samek, Wiegand, et al (2017) משווים שני סוגים של תוכנות-מסבירות: sensitivity analysis (SA) and layer-wise relevance propagation (LRP) , ומוצאים שההסברים המסופקים על ידי LRP טובים יותר מאלו הניתנים בעזרת SA. מכאן אפשר להבין שגם התוכנות-המסבירות אינן מספקות הסברים מושלמים ל Deep AI models ומבחינה זו אין הן שונות מהסברים אחרים הניתנים במדע ושמאופייניים כחלקיים בלבד. ישנם כמה גורמים האחראיים לכך, וכאן אעמוד על שני הגורמים הבאים המסבירים מדוע תיאוריה מדעית מספקת הסברים לא מושלמים:

1. *אישוש והפרכה.* כל תיאוריה אמפירית הינה פרוביזורית והיא נחשבת לתיאוריה מאוששת (confirmed) עד לאותו הרגע שבו היא מופרכת (Popper, 1972; Rakover, 2018). למשל, התיאוריה הניוטונית נחשבה לנכונה במשך זמן ארוך מאוד, עד לאותו הרגע שבו התברר שכמה הנחות בסיס שלה אינן תופסות כאשר מהירות תנועת הגוף מתקרבת למהירות האור.
2. *מסגרת ההתייחסות של ההסבר*. כל תיאוריה אמפירית נתחמת באופן גלוי או סמוי בתחום אמפירי ותיאורטי מסוים. בתחום זה ניתן ליישם את התיאוריה הנידונה על כל התופעות הנכללות בתחום המדובר. למשל, התיאוריה הניוטונית חלה על כל הגופים הנעים במהירויות ארציות (שאינן קרובות לאור). ועוד דוגמה: האם הרובוט רוברט, שתואר לעיל, עשוי להוות הסבר טוב להתנהגותו של האדם? אינטואיטיבית התשובה היא שלילית וכל אחד יכול לחשוב על מספר הבדלים קריטיים הקשורים בתחום הפעילות של האדם לבין תחום הפעילות של הרובוט: רוברט הרובוט פועל בעזרת קפיצים מוטות וגלגלי שיניים ואילו האדם פועל בעזרת מערכת עצמות, שרירים ועצבים; האדם, בניגוד לרוברט, מתאים את פעולתו לסביבה כתוצאה מהיותו ניחן בהכרתיות. ודוגמה אחרונה: Rakover & Cahlon (1989, 2001) פתחו מודל מתמטי לזיהוי פרצופים, “the Catch Model”, שהנחותיו תחמו מסגרת תיאורטית-אמפירית שבה נסגרה התנהגותו של היחיד (העד), מודל שבעזרתה ניסו החוקרים לשחזר את פרצוף-המטרה מזיכרונו של העד. אחת הבעיות החמורות שניצבה בפני המודל הנוכחי, נבעה מכך שמסגרת זו הוציאה מכלל חשבון את העובדה שחשיפה נוספת של היחיד לפרצופים מעבר לתפיסת פרצוף-המטרה (החשיפה לפרצופים נוספים היא חלק מהשיטה של המודל לשחזור פרצוף-המטרה) יצרה הפרעת זיכרון הפוגעת מאוד ביעילותו של המודל הנידון. במילים אחרות, אם אכן זכרונו הפרצופי של האדם לא היה מושפע לרעה מאינפורמציה פרצופית נוספת, סביר להניח שה Catch Model היה מצליח טוב יותר לשחזר את פרצוף-המטרה מזיכרונו של היחיד.

בהנחה הסבירה שכל תיאוריה אמפירית מספקת דרגת הבנה מסוימת של התופעה הנחקרת, אפשר להציע את הרעיון הבא. מצד אחד, ניתן לתפוס את ה Deep AI models כמסובכים כל כך עד שהם נתפסים כלא מובנים, black-boxes. אולם, מצד אחר, מספקים מודלים אלה רמה מסוימת (מוגבלת) של הבנה העולה מעצם הדרך שבה הם תוכנתו. למשל, ניתן להבין בצורה כללית כיצד מפתחים ומאמנים neural networks מסוימים בעזרת אלגוריתם מיוחד הנקרא Backpropagation: בקיצור נמרץ, אלגוריתם זה משתמש בטעות של התוכנה (הפער בין הערך של הפלט לבין הערך ההתנהגותי המצוי) ומשנה את המשקלות (עוצמת הקשרים בין היחידות (ה nodes) השונות המרכיבות את המודל) כך שפער זה ילך ויצטמצם ועוצמת ניבוי ההתנהגות הנחקרת על ידי הרשת תלך ותגבר.

**דיון: האנאלוגיה של הפסיכולוגיה הקוגניטיבית**

אנאלוגיות משמשות כלי חשוב בהסבר התנהגות. נבחן את המקרה הסכמתי הבא המאפיין את הפסיכולוגיה הקוגניטיבית:

אם נתפוס את התנהגות האדם בצורתה הכללית באופן הבא:

 Response (Y) = f{[Unknown Mechanism], Stimulus (X)}

 ואם נמצא מכניזם מסוים, כמו למשל, המחשב או רוברט הרובוט המתנהגים בצורה הבאה:

 Response (Y\*) = f{[Known Mechanism], Stimulus (X\*)}

כאשר, Response (Y) דומה מאוד ל Response (Y\*) (למשל, מזיגת כוס תה), וכאשר Stimulus (X) דומה מאוד ל Stimulus (X\*) (המצב של מזיגת כוס תה),

אזי אנחנו נוטים להציע את המסקנה הבאה:

גם ה [Unknown Mechanism] באדם מתאים, דומה ל [Known Mechanism] ברובוט.

 על אנאלוגיה זו יש להעיר שתי הערות:

הערה ראשונה: העובדה שבין שני דברים, הבנויים ממספר רב של מרכיבים, נמצאו קווי דמיון חשובים במספר מרכיבים מסוים אינו מבטיח שגם ביתר המרכיבים יימצאו קווי דמיון חשובים. במקרה של המחשב, כאמור, קיימים מספר קווי דמיון גדולים (פונקציונאליים) בין התנהגות המחשב לאדם: הקלט והמצב הגירוי, הפלט והתגובה, וגם בין מספר תת-מערכות מסוימות במחשב לבין השערות על תת-מערכות מקבילות הפועלות במוח האדם. למרות קווי דמיון אלה, קל להצביע על הבדלים עצומים בין תפקוד המחשב לבין תפקודו הקוגניטיבי של האדם, למשל, במספר רב של תחומים עולה עצמת החישוב של המחשב אלפי מונים על זה של האדם; טרם נמצא מחשב שנוצרה בו הכרתיות כבאדם.

במקרה של פעולת רוברט הרובוט והאדם הדמיון בין פעולת החתימה ומזיגת כוס התה אינה קובעת בהכרח שהמכניזם האחראי לפעולותיו של הרובוט הוא המכניזם האחראי לפעולותיו של האדם. במקרה זה ברור מאוד שהמכניזמים שונים לחלוטין. מבחינה לוגית הסיבה שאנאלוגיה אינה מבטיחה בהכרח הסבר יעיל ונכון, נובעת מכך שלכל קבוצה של נתונים ניתן להתאים באופן עקרוני אינסוף של פונקציות (כלומר, תיאוריות) שונות. במקרה הנוכחי לנתונים הבאים: מצב חתימה, ומצב מזיגת כוס תה; תגובת חתימה, ותגובת מזיגת תה, הותאמו שני מכניזמים שונים, אחד מכאני לחלוטין והשני פיזיולוגי, קוגניטיבי ומנטאלי.

הערה שנייה, האנאלוגיה מפתה במיוחד כאשר התנהגות לא מובנת A מוּשוות לפעילות-התנהגותית מובנת B. במקרה זה, אנחנו נוטים ליישם את ההסבר המובן של B על ההתנהגות הלא-מובנת של A. אולם, במקרה שפעילות B אינה מובנת, השימוש בפעילות B כהסבר אנאלוגי ל A הופך להיות בעייתי. וזה לאמתו של דבר המצב הנוכחי של הפסיכולוגיה הקוגניטיבית: אם אין מבינים את התוכנות הכי מתוחכמות ומצליחות, ה Deep AI models, אזי לא ברורה מהי המוטיבציה להמשיך ולבנות מודלים להסבר התנהגות המבוססים בסופו של דבר על תהליכים חישוביים כמוסים וסתומים? כדי לענות על שאלה זו, אנתח עתה מספר יתרונות וחסרונות של האנאלוגיה עליה מבוססת הפסיכולוגיה הקוגניטיבית.

*יתרונות:*

*א) כ*פי שציינתי לעיל, מספר חוקרים סבורים שכמו שניתן להעמיד (reduce) את כל תוכנות המחשב על סדרות של אפס ואחד (אין-יש מתח חשמלי) כך ניתן להעמיד את כל התהליכים הקוגניטיביים והמנטליים המתרחשים בראשו של האדם על הנוירופיזיולוגיה של המוח. במילים אחרות, האנאלוגיה למחשב אפשרה פיתוח של השערה מעניינת ומפתה שלפיה בסופו של דבר הפתרון לבעיית הגוף/נפש בוא יבוא, כלומר, ניתן יהיה לפתח תיאוריה המסבירה כיצד התהליך המנטאלי-הכרתי יועמד על תהליך נוירופיזיולוגי במוח (וראו דיון ביקורתי על כך אצל Rakover, 2018).

ב) אנאלוגיה זו אינה מהווה תיאוריה כללית המיועדת לטפל בסוגים רבים של התנהגויות בדומה לתיאוריות הכוללות של ניוטון, איינשטיין ותורת הקוונטים. האנאלוגיה מהווה מסגרת כללית (בדומה לכללי משחק השחמט) שעל בסיסה ניתן לבנות השערות, תיאוריות, מודלים, שונים המטפלים במגוון תופעות התנהגותיות בשטחי התנהגויות רבים ושונים. כלומר, האנאלוגיה מהווה קרקע פורייה לצמיחה של מודלים ספציפיים ולגילוי של תופעות התנהגותיות רבות ומעניינות.

ג) גם אם תופסים את המוח כמעין מחשב-על טבעי (natural super-computer) וגם אם אין מבינים את דרך פעולתם של מרבית ה Deep AI models, המיועדות להסביר את התנהגות האדם, עדיין תוכנות מחשב אלו מאפשרות רמה מסוימת (נמוכה) של הבנה שעליה דנתי לעיל.

*חסרונות:*

א) לאי ההבנה של תוכנות ה Deep AI models יש חסרון מתודולוגי גדול הנוסף לכך שבמצב הנוכחי תוכנות אלו, המייצגות רמה גבוהה של הפסיכולוגיה הקוגניטיבית, אינן מסוגלות לספק רמה סבירה של הבנת ההתנהגות. מתברר שעיקר הצטיינותם, כוחם, של תוכנות אלו הוא בניבוי המוצלח שלהם: הפלטים שלהן מתיישבים היטב עם התצפיות האמפיריות. אולם דווקא הצטיינות זו מעלה בעיה קשה: מופיע כאן פער בין הניבוי לבין ההסבר. חשיבות הדגש המתודולוגי עוברת מההסבר אל הניבוי (כלומר, הדיוק בחיזוי התוצאות, התצפיות). בצורתו הקיצונית, מעבר זה של הדגש מההסבר, מההבנה, אל הניבוי אומר שנקבל ונשתמש בכל תיאוריה כל עוד היא מצליחה לנבא היטב את התצפיות. גישה קיצונית זו עלולה להוביל לירידה דרמטית באיכות המחקר המדעי, כי ללא הבנה תיאורטית לא נוכל לבנות מבחנים אמפיריים לבדיקתן של המודלים והתיאוריות. אם ניתן להתאים לכל קבוצה של נתונים אינסוף של פונקציות שינבאו היטב נתונים אלה, ואם לא קיימת הבנה של ההסבר התיאורטי, נתקשה מאוד להבחין בין כל התיאוריות השונות שחוזות היטב את אותן התצפיות. לאמתו של דבר, תיאוריות מנבאות-היטב (well predicting theories - ?) אלו יהפכו לאד הוק, משום שלא ניתן לבחון אותן. טיעון זה נשען על התפיסה של השכל הישר האומרת שמהסבר נכון עולה ניבוי מוצלח (כלומר, שבתנאים הרלוונטיים ההסבר התיאורטי הנכון יציע ניבוי מוצלח שהתרחש או יתרחש במציאות). אולם ניבוי נכון אינו אלא תנאי הכרחי להסבר נכון. לא יתכן שמהסבר נכון יתקבל ניבוי לא נכון, אך ניתן לקבל ניבוי מוצלח גם מתיאוריה לא נכונה.

ב) בהתוויית קווי האנאלוגיה בין המחשב לבין פעילותו ותפקודו של המוח, הולך ומתברר שבעוד שעולם המִחְשוּב תחום באופן מכניסטי (כלומר, אין בו שום מקום להשפעות של תהליכים מנטאליים כמו רצון ואמונה), עולמו של האדם (וגם של בעלי החיים) כולל גם את חווית ההכרתיות (consciousness), כמו רצון ואמונה הרוויים בחלק נכבד מההתנהגות. אולם עד כה טרם הוצעה תיאוריה שמסבירה כיצד ההכרתיות מעוגנת בנוירופיזיולוגיה של המוח ולכן קשה גם לראות כיצד ה Deep AI models, המושתתים על העולם המכניסטי של המיחשוב, עשויים להציע הסבר שלם להתנהגותו ההכרתית של בן אנוש (ראו סקירה בנידון אצל Rakover, 2018).

ג) במעבר האנאלוגי ממושגי המחשב המוגדרים היטב באופן מתמטי אל תחום הפסיכולוגיה הקוגניטיבית, נפרצו מושגי היסוד כמו 'אינפורמציה' ו'עיבוד אינפורמציה'. בעוד שמושגים אלה מוגדרים במדויק במדעי המחשב, תחומיהם נשברו בפסיכולוגיה, למשל, מושג ה'אינפורמציה' חל כמעט על כל דבר, כגון על הברות, מילים, משפטים, קטעי פרוזה או שירה, דימויים ויזואליים וכו (ראו דיון אצל Palmer & Kimchi, 1986; Rakover, 2018).

בהתחשב ביתרונות וחסרונות אלה, חובה עלינו לשאול כיצד יש להתייחס לפסיכולוגיה הקוגניטיבית? התשובה מתבססת על שתי האפשרויות הבאות:

1. אם מתרשמים שהיתרונות עולים על החסרונות, ניתן להמשיך לפעול במסגרת של הפסיכולוגיה הקוגניטיבית, כי ה Deep AI models מספקים רמה מסוימת של הסבר (רמה קלה של הסבר), ובמקרים מסוימים גם תוכנות-ההסבר עשויות לעזור בהבנת מודלים מסובכים אלה.
2. אם מתרשמים שהחסרונות עולים על היתרונות, עולה הצורך להחליף את האנאלוגיה של המחשב, אך בד בבד עולה השאלה הבאה: איזו גישה עשויה להחליף את הפסיכולוגיה הקוגניטיבית?

ומשום שאין בידי תשובה לשאלה קשה אחרונה זו, אשאיר אותה פתוחה ואסיים כאן את המאמר.

 **References**

1. Dennett, Daniel C. (1979). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: The MIT Press.
2. Rakover, Sam S. (2018). *How to Explain Behavior: A Critical Review and New Approach.* Lanham: Lexington Books.
3. Rakover, Sam S. (1990). *Metapsychology: Missing Links in Behavior, Mind and Science*. New York: Paragon/Solomon.
4. Hempel, Carl. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press.
5. Hempel, Carl. G. (1966). *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice Hall.
6. Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*.

Princeton: Princeton University Press.

1. Salmon, W. C. (1990). *Four decades of scientific explanation*. Minneapolis:

University of Minnesota Press.

1. Rakover (in preparation)
2. Popper, Karl R. (1972). *Objective Knowledge: An Evolutionary Approach.* New York: Oxford University Press.
3. Rakover, S. S. & Cahlon, B. (1989). To catch a thief with a recognition

model: The model and some empirical results. *Cognitive Psychology*, 21, 423-468.

1. Rakover, S. S. & Cahlon, B. (2001). *Face recognition: Cognitive and computational processes*. Amsterdam/Philadelphia: John Benjamins.
2. Palmer, Stephan E. and Ruth Kimchi. 1986. “The Information Processing Approach to Cognition.” In *Approaches to Cognition: Contrasts and Controversies*, edited by Terry J. Knapp and Lynn C. Robertson, 3777. New Jersey: LEA.
3. Elmahmudi, A. & Ugail, H. (2019). Deep face recognition using imperfect facial data. *Future Generation Computer System*, 99, 231-225.
4. Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & review*, 28, 40-80.
5. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Muller. K-R (Eds.) (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning.* Springer Nature Switzerland AG.
6. Samek, W. & Muller. K-R. (2019). Toward explainable artificial intelligence. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Muller. K-R (Eds.) *Explainable AI: Interpreting, explaining and visualizing deep learning.* Springer Nature Switzerland AG.
7. Samek, W., Wiegand, T. & Muller. K-R. (2017). Explainable artificial understanding, visualizing and interpreting learning model. *arXiv: 1708.08296v1.*
8. Zhou, B., Bau. D., Oliva, A. & Torralba, A. (2019). Comparing the interpretability of deep networks via network dissection. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Muller. K-R (Eds.) *Explainable AI: Interpreting, explaining and visualizing deep learning.* Springer Nature Switzerland AG.