האוניברסיטה

חוברת קריאה

בינה מלאכותית

איך זה עובד באמת, ועל מה באמת מתווכחים.

20 יחידות · 135 קישורים מאומתים · 50 ספרים

uni.zikeet.co.il

על החוברת

פקולטה בשני חלקים שנכתבה כדי להחזיק כמה שנים, ולכן היא נמנעת משמות מוצרים ומהכרזות של חברות. החלק הראשון מסביר את המנגנון מאפס: מה זה מודל, איך הוא לומד, ולמה ניחוש של מילה אחת בכל פעם מייצר טקסט שנראה כמו מחשבה. החלק השני לוקח עשרה ויכוחים שחוזרים שוב ושוב, ומציג כל אחד מהם בתבנית קבועה — קודם מה שני הצדדים מסכימים עליו, ואחר כך כל צד בשיאו ובמילים שלו.

החוברת הזאת היא אותו חומר שנמצא באתר, ערוך לקריאה רצופה. כל יחידה עומדת בפני עצמה ואפשר לקרוא אותן בכל סדר, אבל הן מסודרות בסדר שיש בו היגיון — מי שקורא מן ההתחלה יראה קו מתפתח ולא רשימה.

בסוף כל יחידה יש סילבוס: מה לקרוא, לצפות ולהאזין כדי להעמיק. כל קישור שם נפתח ונבדק לפני שנכנס, וכתובתו המלאה מודפסת מתחתיו כדי שאפשר יהיה להקליד אותה גם מדף מודפס. היכן שיש ספרים, הם מופיעים בנפרד — שם, מחבר ושנה, בלי קישורי קנייה שמתיישנים.

> היחידות שעוסקות במחלוקת מציגות את שני הצדדים בטיעון החזק ביותר שלהם, ואינן מכריעות ביניהם. זו בחירה מכוונת: המטרה היא שתוכלי להחזיק את הוויכוח בעצמך, לא לקבל ממני מסקנה מוכנה.

תוכן העניינים

איך זה עובד

יסודות

01 מה זה בכלל מודל 5
02 איך מכונה לומדת מטעות 9
03 שכבות וייצוגים 12
04 מאיפה באים הנתונים 15

שפה

05 מילים כמספרים 19
06 למה ניחוש המילה הבאה מייצר טקסט 22
07 קשב והארכיטקטורה שהחליפה את כולן 25

מה יוצא מזה

08 למה זה ממציא 28
09 הטיה 31
10 איך בכלל מודדים 35

הוויכוחים החוזרים

מה זה כן ומה זה לא

11 האם זה מבין 38
12 האם אפשר להגיע מכאן לאינטליגנציה כללית 42
13 האם זה יוצר 45

14 יוצרים נתונים וזכויות 49

15 מה זה עושה לעבודה 53

16 אחריות החלטות ומעקב 57

17 אנרגיה מים ומה זה באמת עולה 62

כמה זה מסוכן

18 כמה זה מסוכן 66

19 פתוח מול סגור 69

ההווה

20 איפה זה עומד — נכון לספטמבר 2026 72

מה זה בכלל מודל

1958 · המכונה הראשונה שאיש לא כתב לה את הכלל

ב־8 ביולי 1958 פרסם הניו יורק טיימס ידיעה על מכשיר של חיל הים האמריקאי תחת הכותרת "מכשיר חדש של חיל הים לומד תוך כדי עשייה". בגוף הידיעה, על סמך דברים שנאמרו במסיבת עיתונאים, נכתב שמדובר ב"עובר של מחשב אלקטרוני שחיל הים מצפה שיוכל ללכת, לדבר, לראות, לכתוב, לשכפל את עצמו ולהיות מודע לקיומו". האיש מאחורי ההצהרה היה פרנק רוזנבלט, פסיכולוג במעבדה האווירונאוטית של קורנל, והדבר שהוא הציג נקרא פרספטרון.

מה שעמד שם באמת היה קופסה. מארק 1 פרספטרון הורכב ונבדק בין יוני לדצמבר 1959 והוצג לציבור ביוני 1960. בקצה אחד שלו מערך של ארבע מאות תאים פוטואלקטריים מסודרים עשרים על עשרים — זו הייתה העין. כל תא היה מחובר ליחידות עיבוד, וחוזק החיבור לא נשמר בזיכרון אלא בווסת חשמלי פיזי, פוטנציומטר, שאפשר לסובב. מנועים חשמליים הם שסובבו אותם. מכונה עם ארבע מאות חיישנים והמון ידיות, ומנועים שמכוונים את הידיות.

וכאן ההבחנה שכל הפקולטה הזאת נשענת עליה. בתוכנה רגילה מישהו יושב וכותב את הכלל: אם הפיקסל הזה כהה ואותו אחד בהיר, זו האות ק. בפרספטרון איש לא כתב את הכלל. מראים למכונה דוגמה, היא עונה, ואם טעתה — מסובבים קצת את הידיות בכיוון שמקטין את הטעות. ואז עוד דוגמה. אחרי חמישים סיבובים כאלה הוא כבר הבחין בין שני סוגי כרטיסים מנוקבים. הכלל לא נכתב. הוא נמצא.

זה בדיוק מה שהמילה "מודל" אומרת כאן, ולא יותר מזה: פונקציה עם מספרים מתכווננים, והמספרים נבחרו כך שהיא תתאים לדוגמאות שראתה. המספרים האלה נקראים משקלים, או פרמטרים, והם הידיות. כל השינוי מאז הוא בכמות: במקום ארבע מאות ידיות ומנועים שמסובבים אותן, מאות מיליארדים של ידיות ותוכנה שמסובבת אותן. הרעיון לא השתנה. התאמת עקום לנקודות — רק במרחב עם מיליארד צירים במקום שניים.

לפרספטרון יש גם הבטחה מתמטית שהוכחה, והיא חשובה כי היא גם המגבלה: אם קיים קו ישר שמפריד בין שתי הקבוצות, ההליך מובטח למצוא קו כזה, במספר סופי של טעויות. אם לא קיים קו כזה — אין הבטחה, ואין פתרון.

ב־1969 פרסמו מרווין מינסקי וסימור פפרט את הספר "פרספטרונים", ובו הוכחות. הידועה שבהן: פרספטרון בן שכבה אחת אינו יכול לחשב זוגיות, ובמקרה הפשוט ביותר — את הפעולה שמחזירה "כן" כששני הקלטים שונים זה מזה ו"לא" כשהם זהים. אין קו ישר שמפריד. הוכחה שנייה הראתה שכדי לקבוע אם צורה היא רצופה, הסיבוכיות גדלה עם גודל התמונה. הגרסה הפופולרית אומרת שהספר הוכיח שרשתות עצביות הן מבוי סתום ובכך הרג את התחום. זה לא מה שקרה. שני המחברים ידעו היטב שרשת רב־שכבתית כן פותרת את אותה בעיה בדיוק, והספר עצמו מציין שאפשר לבנות מחשב אוניברסלי מיחידות סף. מינסקי ופפרט טענו מאוחר

יותר שהמחקר דעך בשנות השבעים לא בגלל אלא בגלל קשיים אמיתיים שלא נפתרו. הקושי שלא נפתר אז היה איך מכוונים את הידיות ברשת שיש בה יותר משכבה אחת — וזו בדיוק היחידה הבאה.

רוזנבלט לא הספיק לראות את התשובה. הוא מת ב־11 ביולי 1971, ביום הולדתו הארבעים ושלושה, בתאונת שיט במפרץ צ׳ספיק.

נותרה שאלה אחת שהיא אולי החשובה ביותר בתחום, ואין עליה תשובה טובה: למה זה בכלל עובד על דוגמאות חדשות. מודל שמתאים את עצמו לדוגמאות יכול פשוט לשנן אותן. ב־2016 בדקו את זה ישירות: לקחו רשתות מובילות לסיווג תמונות, ערבבו את כל התוויות באקראי — כלב סומן חתול, מטוס סומן עוגה — והרשתות התאימו את עצמן גם לרעש הזה, בדיוק מושלם. כלומר יכולת השינון היא מוחלטת, והיא לא מסבירה דבר. מה שצריך הסבר הוא למה, כשהתוויות כן אמיתיות, המודל מצליח גם על תמונה שלא ראה. זו שאלה פתוחה.

וכשזה לא מצליח, הכישלון נראה כך. ב־2018 פורסם ב־PLOS Medicine מחקר על מודל לזיהוי דלקת ריאות בצילומי חזה, על 158,323 צילומים משלושה מוסדות רפואיים. מודל שאומן על שניים מהם הגיע בבדיקה פנימית לביצועים גבוהים, ובמוסד השלישי צנח. הסיבה התבררה כשבדקו מה המודל בעצם למד: הוא היה מסוגל לזהות מאיזה בית חולים הגיע הצילום בדיוק של כמעט מאה אחוז, ובתוך בית חולים אחד אפילו מאיזו מחלקה. שכיחות המחלה הייתה שונה בין המוסדות, ולכן ״מאיזה בית חולים זה״ היה ניחוש מצוין — במוסדות האלה. זו לא תקלה. זו המכונה עושה בדיוק את מה שביקשו ממנה, ומגלה שביקשו משהו אחר ממה שהתכוונו.

הערה אחרונה, כי הסיפור הזה מסופר בכל הרצאה שנייה: הסיפור על הרשת שאומנה לזהות טנקים וגילו שהיא מבחינה בין יום שמשי ליום מעונן — הוא אגדה. מי שחיפש את המקור הראשוני לא מצא אותו; הגרסאות סותרות זו את זו בשנה, בזרוע הצבאית ובתעלול שהתגלה. חבל, כי הדוגמאות האמיתיות טובות ממנו.

> למה זה ברשימה
>
> כי כל השאר בפקולטה הזאת נשען על הבחנה אחת — בין תוכנית שמישהו כתב ובין מודל שהותאם לדוגמאות — וההבחנה הזאת בת שבעים שנה.

שנה	1958
חלק	איך זה עובד
שער	יסודות
נושא	התאמה לדוגמאות

להעמיק

הפרספטרון שהקדים את זמנו בשישים שנה

לקריאה

קורנל מספרת את סיפורו של רוזנבלט: ההכרזה של 1958, הכותרת בניו יורק טיימס, והספר שסתם את הגולל. באנגלית.

https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

כבר היינו כאן ב־1958

לקריאה

השוואה בין ההבטחות של אז לאלה של היום, כולל הציטוט המלא על מכונה שתהיה מודעת לקיומה. באנגלית.

https://theconversation.com/weve-been-here-before-ai-promised-humanlike-machines-in-1958-222700

מחשבים החושבים בכוחות עצמם

לקריאה

מאמר עברי בדוידסון שמסביר למידה מונחית והתאמת יתר דרך תחרות ההמלצות של נטפליקס.

https://davidson.org.il/read-experience/scientificamerican/מחשבים-החושבים-בכוחות-עצמם/

הסיפור על הטנקים ולמה הוא לא קרה

לקריאה

חיפוש שיטתי אחרי המקור הראשוני של האגדה הכי מצוטטת בתחום והמסקנה שאין כזה. באנגלית.

https://gwern.net/tank

הניסוי של התוויות האקראיות

מקורות וארכיון

המאמר שהראה שרשתות מתאימות את עצמן בדיוק מושלם גם לתוויות מעורבבות לגמרי. באנגלית.

https://arxiv.org/abs/1611.03530

המודל שלמד לזהות בית חולים

מקורות וארכיון

המחקר על צילומי החזה: הביצועים בתוך המוסד מול מחוצה לו ומה המודל באמת מדד. באנגלית.

https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683

ספרים

ההיסטוריה הקצרה של ה-AI

טובי וולש · 2023

סקירה קצרה של התחום מהפילוסופים היוונים ועד היום. יצא בעברית בהוצאת תכלת.

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

חוקרת ותיקה מסבירה מה המכונות האלה עושות ומה לא ובאיזו נקודה בדיוק ההסברים הפופולריים נשברים. לא אותר תרגום עברי.

The Alignment Problem

2020 · Brian Christian

חלקו הראשון עובר על הפרספטרון ועל התחנות שאחריו בשפה נקייה מאוד. לא אותר תרגום עברי.

איך מכונה לומדת מטעות

1986 · ההליך שהתגלה מחדש ארבע פעמים עד שתפס

היחידה הקודמת נעצרה בקושי: בשכבה אחת קל לדעת איזו ידית לסובב, כי כל ידית מחוברת ישירות לתשובה. ברגע שיש שכבות באמצע, שאיש לא אמר להן מה תפקידן, השאלה נעשית קשה. אם התשובה בסוף יצאה שגויה בשתי נקודות, איזו מבין מיליון הידיות בשכבה השלישית אשמה, ובכמה. בלי תשובה לזה, רשתות עמוקות הן רעיון על הנייר.

הפתרון מתחיל בהגדרה של "כמה טעית" כמספר אחד. לוקחים את מה שהרשת פלטה, מחסרים ממנו את מה שרצינו, מעלים בריבוע כדי שסטייה למעלה ולמטה תיחשב שווה, ומחברים הכול. המספר הזה נקרא פונקציית ההפסד, והוא כל מה שהמכונה יודעת על "טוב" ו"רע". הכול מכאן והלאה הוא ניסיון להקטין אותו.

עכשיו דמיינו נוף הררי שבו הגובה בכל נקודה הוא המספר הזה, והמיקום שלכם נקבע לפי כל הידיות יחד. למידה היא ירידה במדרון. אתם בערפל, רואים רק את השיפוע מתחת לרגליים, ועושים צעד קטן בכיוון היורד. ואז עוד אחד. זה נקרא ירידה בגרדיאנט, ובניגוד לרוב האנלוגיות בתחום היא לא מטפורה אלא תיאור מדויק, פרט לכך שהנוף הוא לא תלת־ממדי אלא בעל מיליארד מימדים.

מה שהיה חסר הוא הדרך לחשב את השיפוע הזה ביעילות עבור כל ידית בבת אחת, ולזה קוראים התפשטות לאחור. את ההליך גילו כמה פעמים בלי שאיש ישים לב. ב־1970 פרסם סֶפּוֹ לינַיינמַא את הצורה הכללית שלו כשיטה לגזירה אוטומטית. ב־1974 כתב פול וֶרבּוֹס בהרווארד דוקטורט בשם "מעבר לרגרסיה: כלים חדשים לניבוי ולניתוח במדעי ההתנהגות" שהכיל את הרעיון, והתקשה לפרסם אותו במשך שנים. ב־1985 תיאר אותו דיוויד פרקר. ורק באוקטובר 1986, כשדייוויד רוּמלהַארט, ג׳פרי הינטון ורונלד ויליאמס פרסמו ארבעה עמודים בכתב העת "ניצ׳ר", זה נתפס.

המאמר ההוא פותח בשני משפטים שמסכמים את כל התחום: "אנו מתארים הליך למידה חדש, התפשטות לאחור, לרשתות של יחידות דמויות נוירון. ההליך מכוונן שוב ושוב את משקלי החיבורים ברשת כדי למזער מדד של ההפרש בין וקטור הפלט בפועל לבין וקטור הפלט הרצוי". וזהו. אין שם כוונה, אין ייצוג של המטרה, אין תכנון. יש מדידה של פער וצעד קטן שמקטין אותו.

המשפט השני שחשוב במאמר הוא מה שקורה לשכבות האמצע: "כתוצאה מכוונון המשקלים, יחידות פנימיות 'חבויות' שאינן חלק מהקלט או מהפלט מגיעות לייצג תכונות חשובות של תחום המשימה". איש לא אמר להן מה לייצג. הן נדחפו לזה מפני שזה מה שהקטין את הטעות. בדוגמה שהמחברים הביאו, רשת שהתבקשה לזהות אם רצף מספרים סימטרי "גילתה פתרון אלגנטי בעזרת שתי יחידות ביניים בלבד". זה הרגע שבו רשת מפסיקה להיות מסווג ומתחילה לבנות לעצמה מושגים — וזו היחידה הבאה.

ההגינות של המאמר ההוא ראויה לציון, כי היא הפוכה לטון של היום. המחברים כתבו בעצמם שני סייגים. האחד: "החיסרון הברור ביותר של הליך הלמידה הוא שמשטח הטעות עשוי להכיל מינימה מקומיות, ולכן אין ערובה שהירידה בגרדיאנט תמצא מינימום גלובלי" — כלומר אפשר להיתקע בגומה ולחשוב שהגעת לעמק. והשני: "הליך הלמידה, בצורתו הנוכחית, אינו מודל סביר ללמידה במוחות". הדימוי של "רשת עצבית" הוא שם היסטורי, לא טענה ביולוגית.

מה שהשתנה מאז הוא גודל, וכדאי לראות אותו בשתי תחנות מתוארכות. ב־2012 הרשת שזכתה בתחרות זיהוי התמונות הגדולה הייתה בעלת שישים מיליון פרמטרים ושמונה שכבות, ואומנה תשעים מחזורים על שני כרטיסים גרפיים ביתיים במשך חמישה עד שישה ימים. ב־2020 פורסם מודל שפה בעל מאה שבעים וחמישה מיליארד פרמטרים ותשעים ושש שכבות, שאומן על שלוש מאות מיליארד יחידות טקסט. אותה פעולה בדיוק, מוכפלת בהרבה.

ודבר אחד שמצאו כשהגדילו, שמסביר למה "גדול יותר" הוא לא התשובה. ב־2022 פורסם מחקר שאימן יותר מארבע מאות מודלים בגדלים שונים, מ־70 מיליון פרמטרים ועד למעלה מ־16 מיליארד, על כמויות טקסט שונות. המסקנה: לאימון יעיל צריך להגדיל את המודל ואת כמות הטקסט באותו יחס — על כל הכפלה של גודל המודל, הכפלה של הנתונים. מודל בן שבעים מיליארד פרמטרים שאומן לפי הכלל הזה עקף מודלים גדולים ממנו פי שניים עד פי שבעה שאומנו בלי לשים לב אליו. כלומר חלק גדול מהעדיפות שיוחסה לגודל נבע פשוט מכך שמודלים קודמים היו רעבים.

> למה זה ברשימה
>
> כי "למידה" כאן היא שם קוד לפעולה אחת פשוטה — למדוד כמה טעית ולזוז קצת בכיוון שמקטין את הטעות — וכל מה שנראה אחר כך כמו מחשבה נבנה מחזרה עליה.

שנה	1986
חלק	איך זה עובד
שער	יסודות
נושא	אופטימיזציה

להעמיק

ירידה בגרדיאנט: איך רשת לומדת

לקריאה

השיעור המוכר ביותר בנושא בשפה חזותית: פונקציית ההפסד כנוף ואת הלמידה כירידה במדרון. באנגלית עם איורים ואנימציות.

https://www.3blue1brown.com/lessons/gradient-descent

המכונות הלומדות

לקריאה

יושוע בנג'ו מסביר בעברית איך השכבות והמשקלים עובדים ואיך מתבצע כוונון הדרגתי.

https://davidson.org.il/read-experience/scientificamerican/המכונות-הלומדות/

המאמר שהפך את זה לשימושי

מקורות וארכיון

ארבעת העמודים מ־1986 של רומלהארט הינטון וויליאמס בכתב העת ניצ'ר. באנגלית.

https://www.nature.com/articles/323533a0

אותו מאמר במלואו לקריאה

מקורות וארכיון

כולל הגדרת הטעות משפט היחידות החבויות ושני הסייגים של המחברים על מינימה מקומיות ועל המוח. באנגלית.

https://gwern.net/doc/ai/nn/1986-rumelhart-2.pdf

הכלל שמגדיר מה נחשב אימון יעיל

מקורות וארכיון

המחקר של ארבע מאות המודלים והמסקנה שגודל וכמות נתונים צריכים לגדול יחד. באנגלית.

https://arxiv.org/abs/2203.15556

ספרים

The Alignment Problem

2020 · Brian Christian

מספר את ההיסטוריה של ההתפשטות לאחור כסיפור אנושי כולל עשרות השנים שבהן איש לא רצה לשמוע. לא אותר תרגום עברי.

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

הפרקים על רשתות עצביות מסבירים את המנגנון בלי מתמטיקה ובלי להחמיא לו. לא אותר תרגום עברי.

שכבות וייצוגים

2012 · מה שהרשת בונה לעצמה כשאיש לא אומר לה מה לבנות

בספטמבר 2012 התפרסמו תוצאות תחרות שנתית לזיהוי תמונות. במשך שנתיים היו הציונים הטובים בסביבות עשרים ושמונה אחוזי שגיאה ואז עשרים וחמישה. באותה שנה הוגשה רשת עמוקה, ושיעור השגיאה שלה היה חמישה עשר ושלושה עשיריות האחוז — למעלה מעשר נקודות מתחת למקום השני. בתחום שבו התקדמות שנתית נמדדת בנקודה או שתיים, זה לא היה שיפור אלא הכרזה. התחום כולו עבר לרשתות עמוקות בתוך שנה. לשם קנה מידה: בשנת 2015 ירדה השגיאה ל־3.57 אחוזים, וב־2017 לכ־2.25, בסביבות הרמה של אדם שמשקיע מאמץ מלא באותה משימה.

מה בעצם קרה שם. ברשת עמוקה, כל שכבה מקבלת את מה שהשכבה שלפניה הוציאה ובונה ממנו משהו מורכב יותר. אף אחד לא מגדיר את השלבים. הם נוצרים מפני שהם מקטינים את הטעות. השאלה אם זה באמת מה שמתרחש הייתה במשך שנים אמונה מקצועית, עד שמישהו פתח את הקופסה והסתכל.

ב־2013 פרסמו מתיו זיילר ורוב פרגוס שיטה להציג מה גורם ליחידה מסוימת בעומק הרשת להתעורר. ב־2020 נמשכה העבודה בצורה שיטתית: קבוצת חוקרים לקחה רשת זיהוי תמונות אחת, עברה על 1,056 היחידות הראשונות שלה, וקטלגה אותן אחת אחת. התוצאות מדויקות באופן שקשה להתווכח איתו. בשכבה הראשונה, מתוך שישים וארבע יחידות, ארבעים וארבעה אחוזים הם גלאי קווים בכיוון מסוים וארבעים ושניים אחוזים גלאי ניגוד צבע. זהו. שכבה שלמה שלא עושה שום דבר מלבד לשאול ״יש כאן קו, ובאיזו זווית״. בשכבה השנייה מופיעים גלאי תדר נמוך וצירופי צבע. בשלישית, מתוך מאה תשעים ושתיים יחידות, כבר יש גלאי קו ממש, מרקמים, ועקומות קטנות. ובשכבות עמוקות יותר נמצאו גלאי עקומה בכל זווית, גלאים שמבחינים בין אזור חלק לאזור מחוספס — שימושי מאוד לאיתור גבול של עצם — וגלאי ראש כלב שמגיב לאותו ראש מזוויות שונות.

וזה בדיוק הסדר שהוסבר בלי ראיות במשך עשרים שנה: קווים, אחר כך מרקמים, אחר כך חלקים, אחר כך עצמים. אף אחד לא תכנן אותו. החוקרים כתבו שאפשר ״לקרוא אלגוריתמים בעלי משמעות ישירות מתוך המשקלים״, והם הראו כמה כאלה: איך גלאי עקומה בנוי מגלאי קווים בזוויות סמוכות, ומה מחובר למה.

עד כאן הצד שיודעים. עכשיו הצד שלא, והוא גדול יותר. הבעיה הראשונה נקראת רב־משמעות: יחידה אחת ברשת מגיבה לעיתים קרובות לכמה דברים שאין ביניהם קשר. החוקרים מביאים יחידה שמגיבה לפרצופי חתולים, לחזיתות של מכוניות ולרגלי חתולים גם יחד. ההסבר המקובל היום, שנוסח ב־2022, נקרא סופֶּרפוֹזיציה: לרשת יש הרבה יותר מושגים שכדאי לה להחזיק מאשר יחידות פנויות, ולכן היא דוחסת כמה מהם לאותה יחידה ומסתמכת על כך שנדיר שהם יופיעו יחד. יעיל מאוד. בלתי קריא לחלוטין.

מאז מנסים לחלץ את המושגים הדחוסים בכלים ייעודיים, ויש התקדמות אמיתית. עבודה מ־2024 חילצה ממודל שפה מיליון מושגים נפרדים בניסוי אחד וכשלושים וארבעה מיליון באחר. אבל כדאי לקרוא מה המחברים עצמם כתבו על המגבלה: "גם במפרק בן שלושים וארבעה מיליון אנו רואים עדויות לכך שקבוצת התכונות שחשפנו היא תיאור חלקי של הייצוגים הפנימיים של המודל". באותו ניסוי הגדול, שישים וחמישה אחוזים מהמושגים שחולצו התגלו כמתים, כלומר לא מגיבים כמעט לשום דבר.

ובאותה רוח, המשפט שראוי להיות בכל דיון על "הבנה של רשתות" נכתב בידי מי שעשו את העבודה הטובה ביותר בתחום, בפתח המאמר שלהם: הטענות שלהם "ספקולטיביות במתכוון", ו"האמת היא שרק גירדנו את פני השטח של הבנת מודל ראייה יחיד". זה נכתב ב־2020 על רשת קטנה יחסית, וכל מה שקרה מאז הוא שהמודלים גדלו.

אז מה התשובה הנכונה לשאלה "האם מבינים מה קורה בפנים". לא "כן" ולא "זו קופסה שחורה". התשובה היא שיש מדע צעיר עם שיטות שעובדות, עם ממצאים חוזרים שנמדדו, ועם הודאה מפורשת של העוסקים בו שהם מתארים שבריר. ומי שמצטט את המשפט "אפילו המתכנתים לא יודעים מה קורה שם" כדי לסיים דיון — מצטט נכון ומסיק לא נכון.

> למה זה ברשימה
>
> כי זו הנקודה שבה מכונה מפסיקה להתאים עקום ומתחילה לייצר מושגים משלה — ומפני שאנחנו יודעים לראות חלק קטן מהם ולא את הרוב.

שנה	2012
חלק	איך זה עובד
שער	יסודות
נושא	ייצוגים פנימיים

להעמיק

איך הפסקנו להבין מה המחשב חושב

לקריאה

סקירה עברית בדוידסון על בעיית הקופסה השחורה ועל הקיצורים המוזרים שרשתות מוצאות.

https://davidson.org.il/read-experience/firefly/איך-הפסקנו-להבין-מה-המחשב-חושב/

להתקרב: מבוא למעגלים

לקריאה

המאמר שהציג את שלוש הטענות על תכונות מעגלים ואוניברסליות וגם את ההסתייגות שהן ספקולטיביות במתכוון. באנגלית עם המחשות.

https://distill.pub/2020/circuits/zoom-in/

הקטלוג של 1,056 היחידות הראשונות

מקורות וארכיון

הספירה בפועל: כמה גלאי קו כמה גלאי ניגוד צבע וכמה עקומות בכל שכבה. באנגלית.

https://distill.pub/2020/circuits/early-vision/

השיטה שאפשרה להסתכל פנימה

מקורות וארכיון

המאמר של זיילר ופרגוס שהציג איך מציגים מה מעורר יחידה בעומק הרשת. באנגלית.

https://arxiv.org/abs/1311.2901

מודלים צעצוע של סופרפוזיציה

מקורות וארכיון

ההסבר לכך שיחידה אחת מחזיקה כמה מושגים דחוסים ולמה זה מסבך כל ניסיון לקרוא רשת. באנגלית.

https://arxiv.org/abs/2209.10652

חילוץ מושגים ממודל שפה

מקורות וארכיון

העבודה שחילצה מיליוני מושגים ופירטה במפורש מה עדיין חסר. באנגלית.

https://transformer-circuits.pub/2024/scaling-monosemanticity/

ספרים

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

הפרק על ראייה ממוחשבת מסביר מה רשת באמת מזהה ובאיזה מובן זה שונה מראייה. לא אותר תרגום עברי.

Atlas of AI

2021 · Kate Crawford

לא על הפנים של הרשת אלא על מה שמסביבה אבל הפרק על מאגרי התמונות הכרחי כדי להבין מה היא ראתה. לא אותר תרגום עברי.

מאיפה באים הנתונים

2009 · לא ידע אנושי אלא ערימה מסוימת של דברים עם היסטוריה

ב־2009 הוצג בכנס לראייה ממוחשבת מאגר תמונות בשם אימג'נט. בשלב ההוא היו בו כ־3.2 מיליון תמונות בחמשת אלפים ומאתיים קטגוריות; היום הוא מונה 14,197,122 תמונות ב־21,841 קטגוריות. התמונות לא צולמו לצורך העניין. הן נאספו בשאילתות למנועי חיפוש תמונות, כשכל מילה מורחבת בשמות נרדפים לה ומתורגמת לסינית לספרדית להולנדית ולאיטלקית כדי לגרוף עוד. ואז היה צריך מישהו שיאמת שבתמונה שנמצאה תחת המילה "נמר" יש באמת נמר. המישהו הזה היה 49 אלף עובדים מ־167 מדינות, שעברו על יותר ממאה ושישים מיליון תמונות מועמדות בין יולי 2008 לאפריל 2010, כל תמונה בשלושה אישורים בלתי תלויים.

כדאי לעצור על המשפט הזה, כי הוא חוזר בכל שלב מאז. מאגר אימון אינו ידע. הוא תוצר של החלטות: מה חיפשנו, איפה, באיזו שפה, ומי אישר. בהמשך גם התברר מה זה אומר. בבדיקה שנערכה בין 2018 ל־2020 עברו על החלק במאגר שמתאר בני אדם: מתוך 2,832 קטגוריות אנושיות, 1,593 סומנו כפוגעניות ו־1,081 כקטגוריות שאינן באמת חזותיות — כלומר תכונות שאי אפשר לראות בתמונה. נותרו מאה חמישים ושמונה, ורק מאה שלושים ותשע מהן עם יותר ממאה תמונות.

במעבר לטקסט קנה המידה משתנה אבל ההיגיון זהה. מאגר הטקסט שנבנה מסריקת האינטרנט וניקויה, ופורסם עם ניתוח מפורט ב־2021, מכיל כ־156 מיליארד יחידות טקסט אחרי סינון — מתוך 1.4 טריליון לפני. ומה יש בו בפועל. המקור הגדול ביותר הוא מאגר פטנטים של גוגל. אחריו ויקיפדיה האנגלית, ואחר כך עיתונות: הניו יורק טיימס, הלוס אנג'לס טיימס, הגרדיאן. יש בו גם כמעט שלושים וארבעה מיליון יחידות טקסט מאתרי צבא ארצות הברית. ולמעלה מעשירית מהפטנטים הגיעו ממשרדי פטנטים שאינם דוברי אנגלית, כלומר עברו תרגום מכונה — הטקסט שעליו לומדים אנגלית אינו כולו אנגלית שאדם כתב.

הניקוי עצמו מסתבר כהחלטה טעונה. בין המסננים היה מסנן שמסיר מסמכים שמכילים מילים מרשימת מילים גסות. הבדיקה מצאה מה הוא עשה בפועל: טקסט באנגלית אפרו־אמריקאית הוסר בשיעור של ארבעים ושניים אחוזים, טקסט באנגלית של דוברי ספרדית בשלושים ושניים אחוזים, וטקסט באנגלית "לבנה" ב־6.2 אחוזים בלבד. אף אחד לא התכוון לזה. המסנן פשוט עשה את מה שנאמר לו, ומה שנאמר לו לא היה מה שהתכוונו.

שתי בעיות נוספות שהתגלו באותו גוף מחקרים ראויות לשם. הראשונה, שכפול: במאגרים האלה יש טקסט שחוזר על עצמו פעמים רבות — במקרה אחד משפט בן שישים ואחת מילים שהופיע למעלה משישים אלף פעם. ניקוי הכפילויות הוריד פי עשרה את התדירות שבה מודל פולט טקסט משונן מילה במילה, וגם קיצר את האימון. השנייה, זיהום מבחנים: קטעים שאמורים לשמש למבחן נמצאים בתוך חומר האימון. בבדיקה של מאגרי מבחנים מקובלים נמצאו שיעורי זיהום שנעים בין כמעט אפס ובין קרוב לרבע מהמבחן. כשמודל "מצליח

במבחן", זו השאלה הראשונה שצריך לשאול, וגם המאמר שהציג את אחד המודלים הגדולים של 2020 הודה בעצמו שבאג בסינון גרם לו לפספס חלק מהחפיפות.

יש גם עניין של בעלות. מאגר טקסט פתוח שפורסם בסוף 2020 בגודל 825 ג"יגה־בייט הורכב מעשרים ושניים חלקים, ואחד מהם, כשתים־עשרה אחוז ממנו, היה אוסף ספרים שמקורו באתר פיראטי. ביולי 2023 הוסר החלק הזה בדרישת ארגון זכויות דני. והחלק החשוב: מודלים שכבר אומנו עליו לא שכחו אותו. אותו עניין בדיוק בתמונות — מאגר של כמעט שישה מיליארד זוגות של תמונה וכיתוב, שנאספו מהרשת, הוא הבסיס לחלק גדול ממחוללי התמונות, והוא גם מה שאמנים מצאו את עבודתם בתוכו. זו היחידה על זכויות יוצרים בחלק השני של הפקולטה.

ומי עושה את העבודה. בינואר 2023 פורסם תחקיר על צוות בקניה שהועסק כדי לסמן טקסטים קשים — תיאורים של התעללות, אלימות ופגיעה עצמית — כדי שאפשר יהיה ללמד מודל לזהות ולסנן אותם. השכר נטו נע בין 1.32 לשני דולרים לשעה. החברה שהזמינה את העבודה שילמה לקבלן 12.50 דולר לשעה. המספרים על העומס שנויים במחלוקת בין הצדדים: העובדים דיברו על מאה חמישים עד מאתיים וחמישים קטעים במשמרת של תשע שעות, הקבלן טען שבעים. אחד המסמנים תיאר זאת כך: "זה היה עינוי. אתה קורא שורה של אמירות כאלה כל השבוע, ועד יום שישי אתה מעורער מעצם החשיבה על התמונה הזאת". הרעיון שמערכת כזאת "למדה לבד" הוא אחד הדברים הרחוקים ביותר מהאמת.

ולבסוף, בעיה חדשה יחסית ומעניינת. ככל שיותר טקסט ברשת נוצר במכונה, יותר ממנו נכנס למאגרי האימון של הדור הבא. מחקר שפורסם ב־2024 הראה במודלים שונים מה קורה כשמאמנים דור על פלט של קודמו שוב ושוב: קודם נעלמים המקרים הנדירים בקצוות ההתפלגות, ואחר כך התפוקה מתכווצת לטווח צר. הכינוי שניתן לזה הוא קריסת מודל. כמה זה רלוונטי בפועל שנוי במחלוקת, כי בעולם האמיתי לא זורקים את הנתונים הישנים — אבל הכיוון נמדד ותועד, ולמאמר יצא תיקון מחברים ב־2025.

> למה זה ברשימה
>
> כי כל תכונה של מודל — מה הוא יודע מה הוא מדלג עליו ובאיזה עברית הוא כותב — נקבעת בערימה הזאת ולא בארכיטקטורה.

שנה	2009
חלק	איך זה עובד
שער	יסודות
נושא	מאגרי אימון

להעמיק

מי סימן את הטקסטים הקשים

לקריאה

התחקיר על הצוות בקניה כולל השכר לשעה מול מה ששולם לקבלן והמחלוקת על העומס. באנגלית.

https://time.com/6247678/openai-chatgpt-kenya-workers/

נקמת האמנים

לקריאה

סקירה עברית על איסוף התמונות מהרשת ועל ניסיונות של אמנים להרעיל את חומר האימון.

https://davidson.org.il/read-experience/firefly/נקמת-האמנים-נגד-בינה-מלאכותית/

המאמר שהציג את אימג'נט

מקורות וארכיון

איך נאספו התמונות מהם מנועי החיפוש ואיך נבנה האימות בידי עובדים חיצוניים. באנגלית.

https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf

מה באמת יש בתוך מאגר טקסט ענק

מקורות וארכיון

הבדיקה שמצאה את הפטנטים את אתרי הצבא את התרגום המכונה ואת מה שמסנן המילים הגסות הוריד. באנגלית.

https://arxiv.org/abs/2104.08758

הכפילויות ומה שהן עושות

מקורות וארכיון

הממצא שניקוי כפילויות מוריד פי עשרה פליטה של טקסט משונן וגם מקצר את האימון. באנגלית.

https://arxiv.org/abs/2107.06499

מה קורה כשמאמנים על פלט של מכונה

מקורות וארכיון

המאמר על קריסת מודל: קודם נעלמים הקצוות ואחר כך מצטמצם הטווח. באנגלית.

https://www.nature.com/articles/s41586-024-07566-y

ספרים

Atlas of AI

2021 · Kate Crawford

הפרקים על בניית מאגרים ועל העבודה שמאחוריהם הם הטיפול המקיף ביותר בנושא הזה. לא אותר תרגום עברי.

Ghost Work

2019 · Mary L. Gray and Siddharth Suri

על כוח העבודה הבלתי נראה שמסמן מסנן ומתקן כדי שהמערכת תיראה אוטומטית. לא אותר תרגום עברי.

Artificial Unintelligence

2018 · Meredith Broussard

על ההנחה שמחשב פותר בעיה חברתית ועל מה שקורה כשהנתונים נושאים את הבעיה בתוכם. לא אותר תרגום עברי.

מילים כמספרים

2013 · מה קורה כשמחליפים משמעות בשכנות

ב־1957 כתב הבלשן הבריטי ג׳ון רופרט פירת׳ משפט שהפך למשפט המכונן של התחום: ״תכיר מילה לפי החברה שבה היא מסתובבת״. הוא לא היה הראשון. פירת׳ עצמו מפנה לוויטגנשטיין, שטען שמשמעותה של מילה נמצאת בשימוש בה; ביוונית יש שורה דומה אצל אוריפידס, ובלטינית פתגם משפטי עתיק. אבל הניסוח של פירת׳ הוא שהפך להוראת הפעלה.

וההוראה היא זו: אל תנסה להגדיר מה מילה אומרת. תסתכל במקום זה על כל המילים שמופיעות לידה, בטקסט רב ככל האפשר. ואז תן לכל מילה רשימה של מספרים — כמה מאות מספרים — כך שמילים שמופיעות בחברה דומה יקבלו רשימות דומות. הרשימה הזאת נקראת וקטור, וממנה והלאה אפשר לחשב. ״קרוב״ ו״רחוק״ הופכים למרחק שאפשר למדוד. זה כל התרגום של שפה למתמטיקה, והוא היה מספיק.

התוצאה שהפכה את הרעיון למפורסם פורסמה ב־2013. מתברר שהיחסים בין המילים נשמרים לא רק כקרבה אלא ככיוון: המרחק והכיוון מ״גבר״ ל״מלך״ דומים למרחק והכיוון מ״אישה״ ל״מלכה״. ולכן אפשר לחשב. לוקחים את הווקטור של ״מלך״, מחסרים ״גבר״, מוסיפים ״אישה״, ומחפשים את המילה הקרובה ביותר לתוצאה. יוצא ״מלכה״. אותו טריק עובד גם על דקדוק: מ״הולך״ ל״הלך״ הוא אותו כיוון כמו מ״רץ״ ל״רץ בעבר״. במבחן שבנו המחברים היו שמונת אלפים שאלות כאלה.

וכאן נכנס תיקון שראוי להכיר, כי הוא משנה את מה שמותר להסיק. בכל המימושים הסטנדרטיים של הטריק הזה, שלוש מילות הקלט מוצאות מרשימת התשובות האפשריות. כלומר כשמחשבים ״מלך פחות גבר ועוד אישה״, המערכת אינה רשאית להחזיר ״מלך״, ״גבר״ או ״אישה״. הכלל נכנס משום שהמבחן המקורי בנוי מארבע מילים שונות. אבל מסתבר שהוא לא טכני. במחקר מ־2019 הסירו את המגבלה ובדקו: הדיוק צנח משבעים וארבעה אחוזים לעשרים ואחד. במילים אחרות, ביותר משני שלישים מהמקרים התשובה ״הנכונה״ מתקבלת רק מפני שאסור לתת את התשובה שהחישוב בעצם מצביע עליה, והיא לרוב אחת ממילות הקלט. שם המאמר הוא סיכום העניין: ״גבר הוא לרופא כמו שאישה היא לרופא״. מבקרים נוספים הראו שהמבחן רגיש מאוד לגחמות של מילים בודדות, ושחיפוש מתוחכם יותר מוצא מידע שהשיטה הפשוטה מפספסת. אז הטריק אמיתי והוא הרבה פחות נקי ממה שמראים.

מה שכן עמיד לגמרי הוא מה שהווקטורים חשפו על הטקסט שממנו נלמדו. ב־2016 פורסם מאמר בשם ״גבר הוא למתכנת מחשבים כמו שאישה היא לעקרת בית?״, שהראה שוקטורים שאומנו על ארכיון חדשות מכילים אסוציאציות מגדריות במידה מטרידה. ב־2017 פורסם ב״סיינס״ מחקר שעשה את זה כמו שצריך: לקחו מבחן פסיכולוגי מוכר שמודד אסוציאציות סמויות אצל בני אדם, והריצו את אותו מבחן בדיוק על וקטורי מילים שאומנו על סריקת רשת בת 840 מיליארד יחידות טקסט. כל ההטיות שנמדדו אצל אנשים חזרו, בגדלים דומים

ואף גדולים יותר: פרחים מול חרקים, כלי נגינה מול כלי נשק, שמות פרטיים "לבנים" מול "שחורים", ושמות גברים מול נשים ביחס לקריירה ומשפחה.

ובתוך אותו מחקר נמצא הממצא החזק ביותר, והוא לא על הטיה אלא על מראה. החוקרים בדקו עד כמה החוזק של הקשר המגדרי של כל שם מקצוע בווקטורים מנבא את אחוז הנשים בפועל באותו מקצוע לפי נתוני הלשכה לסטטיסטיקה של העבודה בארצות הברית. המתאם היה 0.90. בדיקה מקבילה על שמות פרטיים מול מפקד האוכלוסין נתנה 0.84. כלומר הווקטורים לא "המציאו" סטריאוטיפ — הם מדדו במדויק את חלוקת העבודה בפועל בחברה שכתבה את הטקסט. וזה בדיוק מה שהופך את השאלה לקשה: מה בדיוק מבקשים מהמכונה, לתאר את העולם או לתאר עולם אחר.

הניסיון לתקן נתקל בקושי משלו. ב־2016 הוצעו שיטות "ניקוי הטיה" שמזהות כיוון מגדרי במרחב ומנטרלות אותו, והמדדים אכן ירדו. ב־2019 הראו חוקרות שהמדדים ירדו אבל המידע נשאר: אפשר עדיין לשחזר את החלוקה המגדרית מתוך המרחקים בין המילים ה"מנוקות". הניסוח שלהן: "האפקט בפועל הוא בעיקר הסתרת ההטיה, לא הסרתה".

למה זה ברשימה

כי ההחלטה להגדיר מילה לפי החברה שבה היא מופיעה היא שאיפשרה למכונה לעבוד עם שפה — והיא גם הסיבה שהמכונה יורשת את העולם שממנו נלקח הטקסט.

שנה	2013
חלק	איך זה עובד
שער	שפה
נושא	ייצוג מילים

להעמיק

מבוא לייצוג מילים בעברית

לקריאה

הסבר עברי של הרעיון ההתפלגותי ושל חשבון הווקטורים כולל הדוגמה של מלך ומלכה. מתחיל מושגי וממשיך לקוד.

https://reshetech.co.il/machine-learning-tutorials/gensim-and-word2vec

גבר הוא לרופא כמו שאישה היא לרופא

לקריאה

הבדיקה שהראתה שהדיוק צונח משבעים וארבעה לעשרים ואחד אחוזים ברגע שמסירים את המגבלה על התשובות. באנגלית.

https://arxiv.org/abs/1905.09866

המאמר שהציג את חשבון הווקטורים

מקורות וארכיון

כולל את הנוסחה המדויקת ואת מבנה מבחן שמונת אלפים השאלות. באנגלית.

https://aclanthology.org/N13-1090.pdf

הטיות אנושיות בתוך וקטורי מילים

מקורות וארכיון

המחקר מ"סיינס" עם גודלי האפקט של כל מבחן ועם המתאם של 0.90 מול נתוני התעסוקה בפועל. באנגלית.

https://faculty.washington.edu/aylin/papers/caliskan_Science_open_access.pdf

גבר הוא למתכנת כמו שאישה היא לעקרת בית

מקורות וארכיון

המאמר שהציג את הבעיה ואת שיטות הניקוי הראשונות. באנגלית.

https://arxiv.org/abs/1607.06520

למה ניקוי ההטיה בעיקר מסתיר אותה

מקורות וארכיון

הבדיקה שהראתה שהמידע המגדרי נשאר שחזיר מתוך הווקטורים גם אחרי הניקוי. באנגלית.

https://arxiv.org/abs/1903.03862

ספרים

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

הפרק על שפה מסביר מה ייצוג התפלגותי לוכד ומה הוא בהכרח מפספס. לא אותר תרגום עברי.

Atlas of AI

2021 · Kate Crawford

על השאלה מה אוסף טקסט בכלל מייצג ומי החליט מה נכנס אליו. לא אותר תרגום עברי.

למה ניחוש המילה הבאה מייצר טקסט

1951 · מטרת אימון אחת ופשוטה ומה היא מכריחה מערכת לבנות

ב־1951 פרסם קלוד שאנון ניסוי שראוי להיות ההתחלה של הסיפור הזה. הוא רצה למדוד כמה מידע באמת יש באות אחת של אנגלית כתובה. אם כל האותיות היו מופיעות בהסתברות שווה, אות אחת הייתה שווה כ־4.7 סיביות. כשלוקחים בחשבון את שכיחות האותיות בפועל, היא יורדת ל־4.14. כשלוקחים בחשבון גם איזו אות באה אחרי איזו, 3.56. כשמסתכלים על מילים שלמות, 2.62.

ואז הוא עשה משהו יפה. הוא לקח ספר — ביוגרפיה של תומס ג׳פרסון — בחר ממנו מאה קטעים אקראיים באורך חמש־עשרה אותיות כל אחד, והושיב אדם לנחש אות אחר אות מה בא הלאה, עם אפשרות לנסות שוב עד שיקלע. מספר הניחושים הוא המדידה. התוצאה, כשנותנים לאדם מאה אותיות של הקשר: בין 0.6 לבין 1.3 סיביות לאות. כלומר יותר משלושה רבעים ממה שכתוב באנגלית הוא עודף — הוא נגזר ממה שכבר נכתב. וזו הנקודה: מערכת שיודעת לנחש היטב את מה שבא הלאה אינה מבצעת תרגיל נחמד. היא מחזיקה, במובן מדויק, מודל של השפה.

זו כל מטרת האימון של מודל שפה. לא לענות נכון, לא להיות מועיל, לא לומר אמת. לקבל רצף ולנחש מה בא אחריו, ולהיענש לפי כמה שפספס. חוזרים על זה על כמות טקסט אדירה, והדבר שנבנה תוך כדי הוא מה שדרוש כדי לנחש טוב.

וכאן מתפצלים הדרכים. הטענה הראשונה אומרת: מה שדרוש כדי לנחש טוב הוא ידיעת צורה בלבד. מאמר משפיע מ־2020 מנסח את זה חד: ״מערכת שאומנה על צורה בלבד אין לה מראש שום דרך ללמוד משמעות״. המחברים מביאים ניסוי מחשבתי על תמנון שמצותת לכבל תקשורת בין שני אנשים, לומד להמשיך את דפוסי האותיות בצורה משכנעת לחלוטין, ואז נשאל מה לעשות כשמתקיפה אותו דוב — ואין לו שום דרך לדעת, כי מעולם לא ראה דוב. ב־2021 ניסחה אותה קבוצה את הביטוי שתפס: מודל שפה הוא ״מערכת לתפירה מקרית של רצפי צורות לשוניות שראתה בחומר האימון שלה, לפי מידע הסתברותי על איך הן מצטרפות, בלי שום התייחסות למשמעות — תוכי סטוכסטי״. בקו דומה נטען ב־2022 שכדאי פשוט להימנע ממילים כמו ״יודע״, ״מאמין״ ו״חושב״ כשמתארים מנבא אסימונים.

והטענה הנגדית אומרת: ניבוי טוב מכריח לבנות מודל של מה שמייצר את הטקסט. יש לה ראיה יפה במיוחד, שפורסמה ב־2022. חוקרים אימנו מודל מאותו סוג בדיוק על דבר אחד: רצפי מהלכים חוקיים במשחק הלוח אותלו. בלי חוקים, בלי לוח, בלי לספר לו שקיים משחק. ואז בדקו מה יש בפנים, ומצאו ייצוג פנימי של מצב הלוח. מעבר לכך, כששינו את הייצוג הזה בכוח — הזיזו ״אבן״ בתוך הרשת — הפלט השתנה בהתאם. כלומר זה לא מתאם, זה רכיב שהמערכת משתמשת בו. ב־2023 הראה חוקר אחר שהייצוג אפילו פשוט משנראה: הוא ליניארי, ומקודד לא ״במשבצת הזאת יש אבן שחורה״ אלא ״במשבצת הזאת יש אבן שלי״ — וזה הגיוני, כי

אותה רשת משחקת את שני הצדדים. בכיוון דומה נמצאו במודלי שפה ייצוגים ליניאריים של מיקום גיאוגרפי ושל זמן היסטורי, ואפילו יחידות בודדות שמקודדות קואורדינטה.

אז מי צודק. שימו לב שהשאלה החלוקה כאן אינה עובדתית אלא מושגית: אף אחד לא חולק על כך שקיים ייצוג פנימי של לוח אותלו. המחלוקת היא אם לקרוא למשהו כזה "מודל של העולם" ואם משמעות דורשת מגע עם דבר מחוץ לטקסט. יש גם עמדה שלישית, שגורסת שמשמעות נקבעת ביחסים הפנימיים בין המצבים של מערכת ולא בהצבעה על עצמים בעולם — ולפיה השאלה "האם יש לו משמעות" אינה נחתכת לפי סוג הנתונים אלא לפי מבנה הפנים.

נותר פרט מכני אחד שמסביר תופעה שכל משתמש פוגש: למה אותה שאלה נותנת תשובות שונות. המודל אינו פולט מילה אלא התפלגות הסתברויות על כל המילים האפשריות, וצריך לבחור מתוכה. הבחירה החמדנית — תמיד המילה הסבירה ביותר — נשמעת כמו הבחירה הנכונה, והיא הבחירה הגרועה. במחקר מ־2019 הראו שהיא מייצרת טקסט "תפל וחוזר על עצמו באופן מוזר", ושאפשר לקבל מאותו מודל בדיוק, בלי לשנות בו דבר, טקסט טוב או טקסט גרוע — רק לפי שיטת הבחירה. הפתרון המקובל הוא להגריל מתוך קבוצת המילים הסבירות בלבד. מכאן האקראיות, ומכאן גם שתשובה שונה בכל פעם אינה סימן למצב רוח.

> למה זה ברשימה
>
> כי זו היחידה שבה מתברר שכל מה שנראה כמו ידע או כוונה נולד ממטרה אחת — להמשיך רצף — ומפני שהשאלה מה המטרה הזאת מכריחה עדיין פתוחה.

שנה	1951
חלק	איך זה עובד
שער	שפה
נושא	ניבוי אסימונים

להעמיק

למה הצ׳אטבוט טועה

לקריאה

הסבר עברי בדוידסון של ניבוי האסימון הבא ושל הסיבות המדויקות לטעויות שנובעות ממנו.

https://davidson.org.il/read-experience/sciencenews/chatbot-gets-it-wrong/

שאנון מודד כמה מידע יש באות

מקורות וארכיון

הניסוי המקורי של 1951 עם ספר ביוגרפיה ואדם שמנחש אות אחר אות. באנגלית.

https://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf

הטענה שצורה לבדה אינה מספיקה

מקורות וארכיון

המאמר עם ניסוי התמנון שמנסח את העמדה הביקורתית בחדות. באנגלית.

https://aclanthology.org/2020.acl-main.463/

המודל שבנה לעצמו לוח משחק

מקורות וארכיון

הניסוי שבו מודל אומן רק על רצפי מהלכים ונמצא בתוכו ייצוג של הלוח שאפשר לשנות ולראות את הפלט משתנה. באנגלית.

https://arxiv.org/abs/2210.13382

ייצוגים של מרחב ושל זמן

מקורות וארכיון

הבדיקה שמצאה ייצוג ליניארי של מיקום גיאוגרפי ושל תאריך בתוך מודלי שפה. באנגלית.

https://arxiv.org/abs/2310.02207

למה הבחירה הסבירה ביותר מייצרת טקסט גרוע

מקורות וארכיון

המאמר שהראה ששיטת הדגימה לבדה משנה את איכות הטקסט מאותו מודל בדיוק. באנגלית.

https://arxiv.org/abs/1904.09751

ספרים

נקסוס

יובל נח הררי · 2024

לא ספר מנגנון אבל הפרקים על מה שרשת מידע עושה למי שחי בתוכה נוגעים ישירות בשאלה הזאת. נכתב בעברית.

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

פרקי השפה עוברים על הוויכוח הזה בלי להכריע ובלי לטשטש. לא אותר תרגום עברי.

קשב והארכיטקטורה שהחליפה את כולן

2017 · פעולת שקלול אחת שפתרה בעיה של מהירות ולא של חוכמה

עד אמצע העשור הקודם, רשת שקראה משפט קראה אותו כמו אדם שקורא בקול: מילה אחר מילה, ובכל צעד מחזיקה סיכום של מה שהיה עד כה. המבנה הזה נקרא רשת נשנית, ובגרסה המשוכללת שלו נוספו שלושה "שערים" שמחליטים מה לשכוח ומה לזכור. היו לו שתי בעיות. האחת, מה שקרה בתחילת פסקה ארוכה נשחק עד שהגיעו לסופה. והשנייה, שהתבררה כמכרעת: אי אפשר לקרוא את המילה החמישית לפני שסיימת את הרביעית, ולכן אי אפשר לפזר את העבודה על הרבה מעבדים במקביל.

הרעיון שפתר את הבעיה הראשונה הופיע ב־2014, ולא בהקשר של צ׳אט אלא של תרגום. החוקרים הצביעו על כך ש"השימוש בווקטור באורך קבוע הוא צוואר בקבוק" — כלומר דחיסת משפט מקור שלם למספר קבוע של מספרים מאבדת מידע. ההצעה: במקום לדחוס, לתת למערכת לחפש בכל פעם מחדש אילו חלקים של משפט המקור רלוונטיים למילה שהיא מתרגמת כרגע. לזה קראו קשב.

מה שקרה ב־2017 היה הסרת כל השאר. מאמר בשם "קשב זה כל מה שצריך" הראה שאפשר לזרוק את הקריאה הסדרתית לגמרי ולבנות מערכת שכולה קשב, ושהיא מתרגמת טוב יותר ומתאמנת מהר יותר. המבנה נקרא טרנספורמר. התוצאה שדווחה: תרגום מאנגלית לצרפתית בציון גבוה מהקודמים אחרי שלושה ימים וחצי על שמונה כרטיסים גרפיים. המספר הזה הוא העיקר — לא האיכות אלא הזמן.

ומה קשב עושה בפועל, בלי נוסחאות. קחו משפט. לכל מילה בו, המערכת מחשבת עד כמה כל מילה אחרת במשפט רלוונטית לה, ומייצרת לה ייצוג חדש שהוא תערובת משוקללת של כל המילים לפי המשקלים האלה. במשפט "הכלב שרדף אחרי החתול היה עייף", כשמעבדים את "היה", המשקל על "הכלב" גבוה ועל "החתול" נמוך. אף אחד לא קבע את זה. המשקלים נלמדו. והנקודה הטכנית שהכריעה: כל המילים מקבלות את הטיפול הזה בו־זמנית, ולכן כל המשפט עובר במקביל. זה מה שאיפשר לאמן על כמויות שלא היו מעשיות קודם.

על המילה עצמה ראוי סייג, והוא מגיע ממי שעוסקת בשני התחומים. בסקירה מ־2020 השוותה חוקרת עצבים את השימוש בביטוי "קשב" בלמידת מכונה לקשב כפי שהוא נחקר בפסיכולוגיה ובמדעי המוח. המסקנה שלה: המנגנון שנעשה בו שימוש היום "דומה פחות לקשב ביולוגי מאשר מנגנוני הקשב שהיו בשימוש בתרגום מכונה קודם לכן", ולסיכום — "ההבדלים באילוצים משמעם שגם התקדמויות גדולות בלמידת מכונה אינן יוצרות בהכרח מודלים דומים יותר למוח". קשב כאן הוא שם של פעולת שקלול. לא מיקוד תודעה.

סייג שני נוגע לפיתוי שקשה לעמוד בפניו: להסתכל על המשקלים ולומר "הנה, זה מה שהמודל הסתכל עליו". ב־2019 פורסם מאמר שטען שאלה אינן הסבר, ובאותה שנה פורסמה תשובה בשם "קשב הוא לא לא־הסבר", שטענה שהמסקנה תלויה לגמרי בהגדרת "הסבר". הוויכוח לא הוכרע. מי שמציג מפת משקלי קשב כתשובה לשאלה "למה המודל אמר את זה" מציג דבר שלא הוסכם עליו.

הדבר האחרון שכדאי לקחת מהתקופה הזאת הוא שרשרת תיקונים שמראה איך נראה מדע שעובד. ב־2020 פורסמה עבודה שמצאה שביצועי מודל שפה משתפרים לפי חוקי חזקה חלקים מול גודל המודל, כמות הנתונים ועוצמת החישוב — על פני יותר משבעה סדרי גודל. המסקנה המעשית שלה הייתה להגדיל מודלים מאוד ולהאכיל אותם בכמות נתונים צנועה יחסית. ב־2022 פורסמה עבודה שבדקה את השאלה מחדש על יותר מארבע מאות מודלים והפכה את העצה: מודל וכמות טקסט צריכים לגדול באותו קצב. וב־2024 פורסם ניסיון שחזור של העבודה השנייה, שמצא שאחת משלוש שיטות האמידה שלה אינה עקבית עם השתיים האחרות, ושרווחי הסמך שדווחו "צרים באופן בלתי סביר — רווחים כאלה היו דורשים למעלה משש מאות אלף ניסויים, בעוד שהם ככל הנראה הריצו פחות מחמש מאות". השחזור תיקן, והתוצאה המתוקנת מסתדרת עם שאר הממצאים. אף אחת משלוש התחנות האלה אינה סופית, וכל אחת מהן עדיפה על הקודמת.

למה זה ברשימה

כי המעבר של 2017 לא היה רעיון חדש על שפה אלא פתרון לצוואר בקבוק בחומרה — וזה מסביר למה הכול קפץ בבת אחת.

שנה	2017
חלק	איך זה עובד
שער	שפה
נושא	טרנספורמר

להעמיק

הטרנספורמרים בעברית: מה היה לפני

לקריאה

הסבר עברי נגיש של רשתות נשניות ושל השערים שקדמו לטרנספורמר ולמה הם לא הספיקו.

https://machinelearning.co.il/14698/transformers-section-1/

קשב בפסיכולוגיה במדעי המוח ובלמידת מכונה

לקריאה

סקירה שמסבירה בדיוק מה משותף לשני השימושים במילה ומה לא. באנגלית.

https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2020.00029/full

הקשב לפני שהיה ארכיטקטורה

מקורות וארכיון

המאמר מ־2014 שזיהה את צוואר הבקבוק בתרגום והציע לחפש בכל צעד במקום לדחוס. באנגלית.

https://arxiv.org/abs/1409.0473

קשב זה כל מה שצריך

מקורות וארכיון

המאמר שהסיר את הקריאה הסדרתית לגמרי וקבע את המבנה שבשימוש עד היום. באנגלית.

https://arxiv.org/abs/1706.03762

חוקי הגדילה כפי שנוסחו ב־2020

מקורות וארכיון

העבודה שמצאה קשרי חזקה חלקים בין גודל נתונים וחישוב ובין איכות. באנגלית.

https://arxiv.org/abs/2001.08361

ניסיון השחזור שתפס את הטעות

מקורות וארכיון

הבדיקה שמצאה אי־עקביות ורווחי סמך בלתי סבירים בעבודת הגדילה של 2022 ותיקנה אותם. באנגלית.

https://arxiv.org/abs/2404.10102

ספרים

ההיסטוריה הקצרה של ה-AI

טובי וולש · 2023

ממקם את 2017 בתוך רצף ארוך של רגעים שנראו כמו קפיצה. יצא בעברית בהוצאת תכלת.

The Alignment Problem

2020 · Brian Christian

לא מתאר את הטרנספורמר עצמו אבל מסביר היטב למה מהירות אימון הפכה לגורם שקובע מה אפשרי. לא אותר תרגום עברי.

למה זה ממציא

2023 · לא תקלה אלא מה שקורה כשמבקשים המשך סביר

ב־22 ביוני 2023 נתן שופט בבית המשפט הפדרלי במנהטן החלטה בתיק תעופה שגרתי. עורכי הדין של התובע הגישו כתב טענות ובו הפניות לשישה פסקי דין קודמים. אף אחד מהם לא היה קיים. השמות, המספרים, הציטוטים והשופטים — הכול נכתב במכונה. כשבית המשפט ביקש עותקים, הוגשו עותקים; גם הם נוצרו במכונה. השופט הטיל קנס של חמשת אלפים דולר, וכתב שני משפטים שכדאי לקרוא יחד. הראשון: "התקדמויות טכנולוגיות הן דבר שבשגרה ואין שום דבר פסול מיסודו בשימוש בכלי בינה מלאכותית אמין לשם סיוע". והשני: "כשהתבקש לספק פסיקה, להראות הלכות מסוימות, להראות עוד תיקים ולתת כמה תיקים — הציאטבוט נענה בכך שהמציא אותם".

הדבר הזה אינו נדיר. אדם אחד מתחזק מסד נתונים של החלטות שיפוטיות שבהן בית משפט קבע במפורש שצד הסתמך על תוכן בדוי שנוצר במכונה. נכון ל־21 בספטמבר 2026 רשומים בו 2,046 מקרים ביותר מארבעים ושמונה תחומי שיפוט — 1,397 בארצות הברית, 218 בקנדה, 112 באוסטרליה, ו־57 בישראל. המספרים האלה ממשיכים לעלות, ולכן הם מתוארכים כאן; מה שלא ישתנה הוא הסדר גודל והעובדה שזה קורה בכל מקום.

מחוץ לבתי המשפט נמדד אותו דבר במחקר. בבדיקה שפורסמה בספטמבר 2023 נבחנו 636 הפניות ביבליוגרפיות שנוצרו במכונה על פני שמונים וארבעה נושאים. בגרסה אחת של המערכת חמישים וחמישה אחוזים מההפניות היו בדויות לחלוטין, ומבין אלה שהיו אמיתיות, ארבעים ושלושה אחוזים הכילו שגיאות מהותיות. בגרסה מאוחרת יותר המספרים ירדו לשמונה עשר ולעשרים וארבעה אחוזים. שיפור אמיתי, ורחוק מאוד מאפס.

ולמה זה קורה. חזרו ליחידה השישית: מטרת האימון היא לייצר המשך סביר. אין במערכת מדף נפרד שבו כתוב "אלה הדברים שאני יודע" ומדף שני שבו "אלה שאני לא". יש התפלגות הסתברויות על מילים, וממנה יוצאת מילה. כשהחומר מכיל את התשובה הנכונה, ההמשך הסביר הוא גם הנכון. כשלא — ההמשך הסביר הוא עדיין סביר. הפניה לפסק דין נראית כמו שם, מול שם, שנה, וכרך. זה דפוס קל מאוד לייצר, והוא ייראה נכון לחלוטין למי שלא בדק. זו לא תקלה במנגנון. זה המנגנון.

ב־2025 הוסיפו לזה טיעון חד יותר, שמסביר למה זה גם לא משתפר מעצמו. המערכות האלה נמדדות במבחנים, ורוב המבחנים נותנים נקודה על תשובה נכונה ואפס על תשובה שגויה ואפס גם על "אני לא יודע". בתנאים כאלה, לנחש עדיף תמיד. הניסוח של המחברים: "מודלי שפה עוברים אופטימיזציה להיות נבחנים טובים, וניחוש כשלא בטוחים משפר את הציון". כלומר חלק ממה שנקרא הזיה הוא פשוט התנהגות של נבחן רציונלי במבחן שבנוי לא נכון. ההצעה שלהם אינה מודל חדש אלא לשנות את שיטת הניקוד.

יש גם צד אופטימי, וכדאי להיות מדויקים בו. מחקר מ־2022 מצא שמודלים גדולים מכוילים היטב — כלומר כשהם נותנים לתשובה שלהם הסתברות של שבעים אחוז, היא נכונה בערך בשבעים אחוז מהמקרים — אבל זה נכון "כשהשאלות מוצגות בפורמט הנכון", בעיקר שאלות רב־ברירה ושאלות נכון או לא נכון. בטקסט חופשי זה מחזיק פחות טוב. ב־2024 פורסם ב"ניצ'ר" כלי שמזהה חלק מהבדיות בדרך אלגנטית: שואלים את אותה שאלה כמה פעמים ובודקים אם התשובות שונות **במשמעות** ולא רק בניסוח. כשהמודל יודע, הוא אומר את אותו דבר במילים אחרות; כשהוא בודה, הוא בודה כל פעם משהו אחר. המדד נקרא אנטרופיה סמנטית, והוא מודד אי־ודאות "ברמת המשמעות ולא ברמת רצף מילים מסוים".

והפתרון שנחשב למובן מאליו — לחבר את המודל למאגר מסמכים אמיתי ולתת לו לצטט מתוכו — עוזר ולא פותר. בבדיקה מ־2024 של כלי מחקר משפטי מסחריים שבנויים בדיוק כך נמצאו שיעורי בדיה שנעו בין שבעה עשר לשלושים ושלושה אחוזים, והמסקנה על הבטחות השיווק הייתה שהן "מוגזמות". כשיש מקור, המערכת עדיין צריכה לנסח ממנו תשובה, ובניסוח היא יכולה לסטות.

הערה על המילה. "הזיה" היא מטפורה קלינית שמייחסת למערכת חוויה תפיסתית, ויש הצעה להשתמש במקומה ב"בדיה" — במובן של אדם שממלא בביטחון פער בזיכרון בלי לשקר. ההבחנה שימושית גם מעשית: יש שגיאות שיטתיות, שנובעות מכך שבחומר האימון כתוב דבר שגוי ולכן הן חוזרות באותה צורה, ויש בדיות, שנובעות מכך שבחומר אין תשובה כלל ולכן הן משתנות בכל פעם. השנייה היא שניתנת לאיתור בשיטה שלמעלה. הראשונה לא.

> למה זה ברשימה
>
> כי המצאה של פרטים אינה באג שאפשר לתקן אלא התנהגות שנובעת ישירות ממטרת האימון — וזה משנה לגמרי מה סביר לצפות שיתוקן.

שנה	2023
חלק	איך זה עובד
שער	מה יוצא מזה
נושא	בדייה

להעמיק

מסד הנתונים של תיקים עם תוכן בדוי

לקריאה

מעקב חי אחרי החלטות שיפוטיות בכל העולם כולל ישראל. המספרים מתעדכנים כל הזמן. באנגלית.

https://www.damiencharlotin.com/hallucinations/

האם הבינה המלאכותית הוזה

לקריאה

הסבר עברי של ההבדל בין שגיאה שיטתית ובין בדיה ושל שיטת האנטרופיה הסמנטית.

https://davidson.org.il/read-experience/sciencenews/האם-הבינה-המלאכותית-הוזה/

האם חיבור למאגר פותר את הבעיה

לקריאה

בדיקה של כלים משפטיים מסחריים שמבוססים על אחזור מסמכים ומה נמצא בהם בפועל. באנגלית.

https://reglab.stanford.edu/publications/hallucination-free-assessing-the-reliability-of-leading-ai-legal-research-tools/

ההחלטה המלאה בתיק ההפניות הבדויות

מקורות וארכיון

פסק הדין מיוני 2023 כולל שמות ששת התיקים שלא היו ומה בדיוק נאמר על השימוש בכלי. באנגלית.

https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/

איתור בדיות לפי משמעות

מקורות וארכיון

המאמר ב"ניצ'ר" שהציג את המדידה של פיזור המשמעויות בין תשובות חוזרות. באנגלית.

https://www.nature.com/articles/s41586-024-07421-0

למה מודלי שפה בודים

מקורות וארכיון

הטיעון שהמבחנים עצמם מתגמלים ניחוש ומענישים הודאה באי־ידיעה. באנגלית.

https://arxiv.org/abs/2509.04664

ספרים

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

מסביר למה מערכת שמייצרת רצף סביר אינה מחזיקה מודל של מה שהיא יודעת. לא אותר תרגום עברי.

Artificial Unintelligence

2018 · Meredith Broussard

על מה שקורה כשמוסד מקבל פלט של מכונה כעובדה. לא אותר תרגום עברי.

הטיה

2016 · איך מערכת בלי כוונה רעה מייצרת תוצאה מוטה

ב־2019 פורסם ב"סיינס" ניתוח של אלגוריתם רפואי שהיה בשימוש על מיליוני מטופלים בארצות הברית. תפקידו היה לאתר חולים מורכבים שכדאי לצרף אותם לתוכנית טיפול מוגברת. בבדיקה התברר שבאותו ציון סיכון בדיוק, מטופלים שחורים היו חולים משמעותית יותר ממטופלים לבנים. כלומר כדי להיכנס לתוכנית, אדם שחור היה צריך להיות חולה יותר.

הסיבה מלמדת יותר מהממצא. באלגוריתם לא היה משתנה של גזע. לא הייתה כוונה. מה שהיה בו זה שהוא לא חזה מחלה אלא חזה **הוצאה** — כמה כסף צפוי להיות מושקע במטופל — מתוך הנחה סבירה לגמרי שמי שעולה יותר הוא מי שחולה יותר. אלא שבארצות הברית מוציאים פחות כסף על מטופלים שחורים באותה רמת חולי, בגלל גישה לשירותים ובגלל אי־אמון במערכת. וכך הפער הקיים במערכת הבריאות נכנס למודל דרך הדלת האחורית, במסווה של משתנה ניטרלי. החוקרים חישבו מה היה קורה אילו היו מתקנים: שיעור המטופלים השחורים בתוכנית היה עולה מ־17.7 אחוזים ל־46.5.

זה הדפוס. מודל מותאם לוקח את מה שיש בנתונים ומשכפל אותו, ולכן הטיה נכנסת בשלושה מקומות: במה שנמדד, במי שנמדד, ובמה שהוגדר כמטרה. דוגמה לסוג השני התפרסמה ב־2018: כלי לדירוג קורות חיים, שפיתוחו התחיל ב־2014 ונזנח ב־2017, אומן על עשר שנים של קורות חיים שהגיעו לחברה. מכיוון שרובם היו של גברים, הכלי למד להוריד ניקוד למסמכים שהכילו את המילה "נשים", למשל בשמות של עמותות או מכללות. ודוגמה לסוג הראשון היא מדידה שפורסמה באותה שנה על מערכות מסחריות לזיהוי מגדר מתוך תמונת פנים: שיעור השגיאה על נשים כהות עור הגיע ל־34.7 אחוזים, ועל גברים בהירי עור לא עלה על 0.8.

המדידה הרחבה והמוסמכת ביותר בתחום הזה פורסמה בדצמבר 2019 בידי מכון התקנים האמריקאי, שבדק 189 אלגוריתמים מ־99 מפתחים על יותר משמונה עשר מיליון תמונות של כשמונה וחצי מיליון אנשים. הממצא: שיעורי זיהוי שגוי של אדם אחד כאדם אחר היו גבוהים פי עשרה עד פי מאה עבור פנים אסיאתיות ואפרו־אמריקאיות לעומת פנים "לבנות". שני פרטים בתוך הדוח חשובים לא פחות. הראשון: אלגוריתמים שפותחו באסיה לא הראו פער דרמטי בין פנים אסיאתיות לפנים לבנות — כלומר מדובר בהרכב חומר האימון ולא בפנים. והשני: האלגוריתמים ההוגנים ביותר היו גם מהמדויקים ביותר, ולכן זו אינה בהכרח עסקה שבה מוותרים על דיוק כדי לקבל הוגנות.

המחיר בפועל מתועד. בינואר 2020 נעצר אדם בדטרויט על סמך התאמה של תוכנת זיהוי פנים והוחזק שלושים שעות; התביעה שהגיש הסתיימה ביוני 2024 בהסדר שקבע שהמשטרה לא תעצור על סמך התאמה כזאת ללא ראיה עצמאית נוספת. בפברואר 2023 נעצרה באותה עיר אישה בהיריון על סמך התאמה דומה והוחזקה

אחת־עשרה שעות; התיק נסגר כעבור חודש מחוסר ראיות. עד סוף 2023 תועדו לפחות שישה מקרים כאלה בארצות הברית, וכולם של אנשים שחורים.

וכאן מגיע החלק שמשנה את כל הדיון. ב־2016 פרסם עיתון חקירות אמריקאי ניתוח של ציוני סיכון שניתנו ליותר משבעת אלפים עצורים במחוז אחד בפלורידה. הממצא: מבין מי שלא ביצעו עבירה נוספת, 44.9 אחוזים מהנאשמים השחורים סומנו כסיכון גבוה לעומת 23.5 אחוזים מהלבנים. ומבין מי שכן ביצעו, 28 אחוזים מהשחורים סומנו כסיכון נמוך לעומת 47.7 אחוזים מהלבנים. החברה שפיתחה את הכלי השיבה שהוא הוגן לפי המדד שלה: בכל ציון סיכון, אחוז דומה של נאשמים שחורים ולבנים אכן ביצעו עבירה נוספת. חוקרים אחרים הוסיפו חמש השגות מתודולוגיות על הניתוח העיתונאי.

ואז, בספטמבר של אותה שנה, הוכיחו שלושה חוקרים משפט. הם ניסחו שלושה תנאי הוגנות סבירים לכל הדעות והראו שמלבד מקרים מיוחדים מאוד, **אין שיטה שיכולה לקיים את שלושתם בבת אחת**. חוקרת נוספת הראתה את אותו דבר בניסוח ממוקד: כשהשכיחות הבסיסית של התופעה שונה בין הקבוצות, אי אפשר לקיים גם כיול שווה וגם שיעורי טעות שווים. כלומר שני הצדדים באותו ויכוח אמרו אמת. העיתון מדד שיעורי טעות והם היו שונים; החברה מדדה כיול והוא היה שווה; ומתמטית, בהינתן שיעורי מעצר שונים בין הקבוצות, אי אפשר היה אחרת.

זו הסיבה שהמשפט "האלגוריתם גזעני" הוא תיאור לא מדויק של מה שנמצא, ושגם "האלגוריתם הוכח כהוגן" הוא לא מדויק. מה שנמצא הוא שאין "הוגן" אחד. יש בחירה בין הגדרות, וכל בחירה כזאת היא החלטה ערכית שמישהו חייב לקבל במודע — ואם לא יקבל אותה במודע, היא תתקבל בכל מקרה, בדמות ברירת מחדל טכנית שאיש לא דן בה.

> למה זה ברשימה
>
> כי הוויכוח הציבורי על הטיה הוכרע בשקט במשפט מתמטי משנת 2016 שקובע שאי אפשר לקיים את כל הגדרות ההוגנות יחד — ומעט מאוד אנשים יודעים את זה.

שנה	2016
חלק	איך זה עובד
שער	מה יוצא מזה
נושא	הוגנות

להעמיק

הניתוח שפתח את הוויכוח על ציוני סיכון

לקריאה

העבודה העיתונאית מ־2016 עם כל המספרים ועם הסיפורים של שני נאשמים. באנגלית.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

הבדיקה של מכון התקנים

לקריאה

הסיכום הרשמי של הבדיקה על 189 אלגוריתמים כולל הממצא שהוגנות ודיוק לא בהכרח מתחרים. באנגלית.

https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software

מעצרי שווא בעקבות זיהוי פנים

לקריאה

סקירה עברית של המקרים המתועדים בארצות הברית.

https://www.ynet.co.il/digital/technews/article/r1vx001uk6

האלגוריתם הרפואי שחזה הוצאה במקום מחלה

מקורות וארכיון

המחקר מ"סיינס" על מיליוני מטופלים כולל החישוב של מה היה משתנה בתיקון. באנגלית.

https://escholarship.org/uc/item/6h92v832

התשובה של הצד השני

מקורות וארכיון

חמש ההשגות המתודולוגיות על אותו ניתוח מאת חוקרים בתחום המדידה. באנגלית.

https://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf

המשפט שקובע שאי אפשר לקיים את כל ההגדרות

מקורות וארכיון

ההוכחה ששלושה תנאי הוגנות סבירים אינם יכולים להתקיים יחד למעט מקרים קיצוניים. באנגלית.

https://arxiv.org/abs/1609.05807

גוני מגדר

מקורות וארכיון

המדידה שהראתה שגיאה של 34.7 אחוזים על נשים כהות עור מול 0.8 על גברים בהירי עור. באנגלית.

https://proceedings.mlr.press/v81/buolamwini18a.html

ספרים

Atlas of AI

2021 · Kate Crawford

על ההנחות שנכנסות למאגר לפני שמישהו מאמן עליו משהו. לא אותר תרגום עברי.

The Alignment Problem

2020 · Brian Christian

הפרקים על ציוני הסיכון מסבירים את משפט אי־האפשרות בשפה נקייה ובלי מתמטיקה. לא אותר תרגום עברי.

Artificial Unintelligence

2018 · Meredith Broussard

על ההנחה שמחשב פותר בעיה חברתית ומה קורה כשהנתונים נושאים אותה בתוכם. לא אותר תרגום עברי.

איך בכלל מודדים

1950 · מבחן שהפך למטרה מפסיק להיות מבחן

ב־1950 פתח אלן טיורינג מאמר במילים "אני מציע לשקול את השאלה, האם מכונות יכולות לחשוב". מה שקרה אחר כך באותו מאמר נשכח כמעט לגמרי. טיורינג לא ניסה לענות על השאלה. הוא כתב שהיא "חסרת משמעות מכדי שתהיה ראויה לדיון", והציע להחליף אותה בשאלה אחרת שאפשר להכריע בה: משחק חיקוי, שבו חוקר מתכתב עם שני משתתפים ומנסה לקבוע מי מהם מי. בגרסה המקורית מדובר בגבר ובאישה, והמכונה נכנסת במקום אחד מהם. טיורינג גם נתן ניבוי מדוד: בעוד כחמישים שנה, למחשב עם זיכרון של סדר גודל מסוים, "חוקר ממוצע לא יהיה לו יותר משבעים אחוז סיכוי לזהות נכון אחרי חמש דקות של תשאול".

שימו לב מה הוא לא אמר. הוא לא הגדיר מבחן שמי שעובר אותו חושב, ולא טען שמעבר במבחן מוכיח תודעה. הוא ביצע החלפה של שאלה פילוסופית בשאלה תפעולית, והודה בכך במפורש. כל השימוש הפופולרי ב"מבחן טיורינג" כרף שמוכיח משהו הוא תוספת מאוחרת. ומה שקרה בפועל למבחן מוכיח למה הוא נזנח: ב־2014 הוכרז שתוכנה עברה אותו בכך שהתחזתה לנער בן שלוש־עשרה, סרקסטי ודובר אנגלית כשפה שנייה — כלומר בנתה לעצמה תירוץ לכל תשובה משונה. המבחן תגמל התחמקות משכנעת, לא הבנה.

התחום עבר אפוא למבחנים ממוקדים: אוספי שאלות עם תשובות נכונות ידועות, וציון אחוז. זה נשמע מוצק, ויש בו שלוש חולשות שצריך להכיר.

הראשונה היא תוקף מבני, שזה השם המקצועי לשאלה הפשוטה "האם המבחן מודד את מה שאמרו שהוא מודד". ביקורת שפורסמה ב־2021 טענה שהתחום מקדש אוסף קטן של מבחנים ומתייחס אליהם כאילו הם מייצגים יכולת כללית, בעוד שכל אחד מהם נבנה למשימה צרה ומסוימת. עבודה מקבילה באותה שנה הביאה מסגרת מהמדעים החברתיים: מושגים כמו "סיכון" או "יכולת" אינם נצפים, ומי שמודד אותם בונה תחליף נצפה — ובפער בין המושג לתחליף נולדות רוב הטעויות. זו אותה תבנית בדיוק כמו האלגוריתם הרפואי שמדד הוצאה במקום מחלה.

השנייה היא זיהום. חומר האימון נגרף מהרשת, והמבחנים נמצאים ברשת. כשמודל נבחן על שאלות שהיו בחומר האימון שלו, הציון מודד זיכרון ולא יכולת. סקירה מ־2023 מנסחת את מצב הידע ביושר: "היקף הבעיה אינו ידוע, מפני שלא פשוט למדוד אותו". ב־2024 נעשה הניסוי הנקי — צוות בנה אלף שאלות חשבון חדשות לגמרי, בדיוק באותו סגנון ובאותה רמת קושי של מבחן ותיק ומוכר, ובדק את אותם מודלים על שניהם. הירידה הגיעה עד שמונה נקודות אחוז, ומודלים שנטו יותר לשחזר מילה במילה את שאלות המבחן הישן הראו ירידה גדולה יותר. ההגינות של אותו מחקר ראויה לציון: הוא גם מצא שחלק גדול מהמודלים המתקדמים הראו סימנים מעטים בלבד להתאמת יתר, כלומר הם באמת פותרים תרגילים חדשים.

השלישית היא הבעיה הכללית ביותר, ואינה ייחודית למחשבים: כשמדד הופך למטרה, הוא מפסיק להיות מדד טוב. ב־2019 נעשה ניסוי יפה בתחום התמונות — בנו מחדש מערכת מבחן לפי אותה שיטת איסוף בדיוק של מערכת המבחן המקורית, ובדקו עליה את המודלים המובילים. הדיוק ירד באחת־עשרה עד ארבע־עשרה נקודות אחוז. אבל המסקנה של המחברים הייתה מעניינת יותר מהמספר: הירידה לא נבעה מכך שהתחום "התאמן על המבחן", אלא מכך שהתמונות החדשות היו מעט קשות יותר, ושיפורים על המבחן הישן תורגמו לשיפורים פרופורציוניים על החדש. כלומר הדירוג בין המודלים שרד; המספר המוחלט לא. וזה בדיוק מה שאפשר לומר על רוב הציונים שמדווחים: הם משווים היטב וקובעים רע.

ויש חיבור ישיר בין היחידה הזאת לקודמת. אם המבחן נותן נקודה על תשובה נכונה, אפס על תשובה שגויה ואפס על "אני לא יודע", אז מערכת שממקסמת ציון תנחש תמיד. חלק ממה שנתפס כהמצאה של עובדות הוא פשוט התנהגות של נבחן שהמבחן שלו בנוי כך. אפשר לשנות את זה, והשינוי אינו בטכנולוגיה אלא בניקוד.

למה זה ברשימה

כי כמעט כל טענה שתשמעי על מה מערכת "מסוגלת לעשות" נשענת על מבחן — ומצב המבחנים בתחום הזה הוא הנקודה החלשה שלו.

שנה	1950
חלק	איך זה עובד
שער	מה יוצא מזה
נושא	מדידה והערכה

להעמיק

הערך על מבחן טיורינג באנציקלופדיה לפילוסופיה של סטנפורד

לקריאה

מה הוצע בדיוק אילו גרסאות יש ולמה התחום הפסיק להתייחס לזה כמדד. באנגלית.

https://plato.stanford.edu/entries/turing-test/

החיפוש אחר מבחן חדש

לקריאה

מאמר עברי בדוידסון על למה מבחן טיורינג נשבר ואילו מבחנים הוצעו במקומו.

https://davidson.org.il/read-experience/scientificamerican/החיפוש-אחר-מבחן-חדש-לבינה-מלאכותית/

המאמר של טיורינג במלואו

מקורות וארכיון

כולל המשפט שהשאלה אם מכונות חושבות חסרת משמעות מכדי לדון בה ואת הניבוי המדויק לחמישים שנה. באנגלית.

https://courses.cs.umbc.edu/471/papers/turing.pdf

הביקורת על תרבות המבחנים

מקורות וארכיון

הטענה שאוסף קטן של מבחנים הפך לנציג של יכולת כללית שהוא לא מודד. באנגלית.

https://arxiv.org/abs/2111.15366

אלף שאלות חדשות באותו סגנון

מקורות וארכיון

הניסוי שבדק כמה מהציון במבחן ותיק נובע מהיכרות עם המבחן עצמו. באנגלית.

https://arxiv.org/abs/2405.00332

מה קורה כשבונים את אותו מבחן מחדש

מקורות וארכיון

הבדיקה שמצאה ירידה של אחת־עשרה עד ארבע־עשרה נקודות והסבירה למה הדירוג בכל זאת שרד. באנגלית.

https://arxiv.org/abs/1902.10811

ספרים

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

פרק שלם על מה מבחנים בתחום באמת בודקים ולמה קל כל כך להשתכנע מציון. לא אותר תרגום עברי.

ההיסטוריה הקצרה של ה-AI

טובי וולש · 2023

מראה כמה פעמים כבר הוכרז שרף נפרץ וכמה מהר הרף עבר מקום. יצא בעברית בהוצאת תכלת.

האם זה מבין

1980 · ויכוח בן שבעים שנה שנשען על מילה שלא הוגדרה

מה ששני הצדדים מסכימים עליו הוא ארוך יותר ממה שנראה. איש אינו חולק על המנגנון: ניחוש של האסימון הבא, שוב ושוב. איש אינו חולק על הביצועים: המערכות האלה מנסחות טקסט קולח, מתרגמות, מסכמות, ועוברות מבחנים שנכתבו לבני אדם. ואיש מהצדדים הרציניים אינו טוען שיש שם חוויה מודעת. הוויכוח כולו הוא על השאלה מה "הבנה" בכלל אומרת, והאם היא דבר שאפשר למדוד מבחוץ.

טיורינג ראה את זה כבר ב־1950 וניסה לעקוף. הוא כתב שהשאלה "האם מכונות יכולות לחשוב" היא "חסרת משמעות מכדי שתהיה ראויה לדיון", והחליף אותה בשאלה תפעולית. שלושים שנה אחר כך, ב־1980, החזיר אותה ג'ון סרל לשולחן בניסוי מחשבתי אחד. דמיינו אדם סגור בחדר, שאינו יודע מילה סינית. מבעד לחריץ נכנסים דפים עם סימנים סיניים. בחדר יש ספר כללים באנגלית שאומר איזה סימנים להוציא בתגובה לאילו סימנים. האיש מבצע, והתשובות שיוצאות מהחדר טובות כמו של דובר סיני. סרל שואל: האם האיש מבין סינית. לא. ואם לא הוא, למה שמחשב שמריץ בדיוק את אותו תהליך יבין.

ההשגה המרכזית נקראת "תשובת המערכת": אמנם האיש לא מבין, אבל המערכת כולה — האיש, הספר, החדר והנייר — כן מבינה. סרל השיב בתנועה שהיא הלב של הטיעון שלו: נניח שהאיש משנן את כל ספר הכללים ואת כל הנתונים ומבצע הכול בראשו, בלי חדר ובלי נייר. עכשיו הוא **הוא** המערכת כולה, והוא עדיין לא יודע מה המילה הסינית ללחמנייה. כתב העת שפרסם את המאמר פרסם לצידו עשרים ושבע תגובות של חוקרים, וסרל השיב לכולן. הוויכוח הזה לא הסתיים מאז.

קדם לסרל מבקר אחר, יוברט דרייפוס, שכתב ב־1965 חיבור בשם "אלכימיה ובינה מלאכותית" ואחריו ספר בשם "מה מחשבים אינם יכולים לעשות". הוא טען שהתחום נשען על הנחות שלא הוכחו — למשל שכל ידע ניתן לפורמליזציה, ושהעולם מורכב מעובדות עצמאיות שאפשר לייצג בסמלים נפרדים. דרייפוס גם הפסיד בשחמט למחשב בערך ב־1967 אחרי שטען שמחשבים לא ישחקו שחמט ברמה סבירה, והתחום לא שכח לו את זה. אבל בדיעבד התחום כן קיבל את הטענה המרכזית שלו: חשיבה אנושית אינה נשענת בעיקרה על מניפולציה של סמלים לפי כללים מפורשים. ההצלחות של העשורים האחרונים באו דווקא מוויתור על הגישה שדרייפוס תקף.

וכעת שני הצדדים בגרסתם העכשווית ובמילים שלהם.

הצד שאומר שזו צורה בלבד מנסח את הטענה כך: משמעות היא היחס בין ביטוי לבין משהו מחוץ לשפה, ומערכת שראתה רק טקסט לא נחשפה מעולם לצד השני של היחס הזה. במאמר מ־2020 מובא ניסוי מחשבתי משלים: תמנון חכם מצותת לכבל תקשורת תת־ימי בין שני אנשים, לומד היטב את דפוסי האותיות, ומצליח להשתלב בשיחה בלי שיבחינו בו. ואז אחד המשתתפים נתקף דוב ומבקש עצה איך לבנות מלכודת. לתמנון אין

שום דרך לעזור, לא מפני שהוא טיפש אלא מפני שמעולם לא ראה דוב, ענף או חבל. "מערכת שאומנה על צורה בלבד אין לה מראש שום דרך ללמוד משמעות". ב־2021 ניסחה אותה קבוצה את הכינוי שנדבק: מודל שפה הוא "מערכת לתפירה מקרית של רצפי צורות לשוניות שראתה בחומר האימון שלה, לפי מידע הסתברותי על איך הן מצטרפות, בלי שום התייחסות למשמעות".

הצד שאומר שנבנה שם ייצוג של עולם משיב שהטיעון הזה מניח את המבוקש. אם משמעות מוגדרת מראש כהצבעה על עצמים בעולם, אז כמובן שמערכת טקסטואלית נופלת — אבל ההגדרה הזאת אינה מוסכמת. בעמדה חלופית, שנקראת סמנטיקה של תפקיד מושגי, משמעות נקבעת ביחסים בין המצבים הפנימיים של מערכת, ואז השאלה נחתכת אמפירית ולא מראש. ולצד זה מובאות הראיות: מערכת שאומנה אך ורק על רצפי מהלכים חוקיים במשחק לוח פיתחה בתוכה ייצוג של מצב הלוח, שאפשר לקרוא ואפשר לשנות בכוח ולראות את הפלט משתנה בהתאם. במודלי שפה נמצאו ייצוגים ליניאריים של מיקום גיאוגרפי ושל זמן היסטורי. ב־2023 פורסמה תשובה ישירה לניסוי התמנון, שטוענת שהיעדר מודעות אינו רלוונטי, שאפשר לעגן ייצוגים גם מטקסט גולמי מפני שהטקסט נושא מבנה של העולם שכתב אותו, ושילדים לומדים שפה גם מטלוויזיה בלי אינטראקציה.

ואזהרה אחת ששייכת לשני הצדדים. בשנת 2022 נטען שיכולות מסוימות "צצות" — אינן קיימות במודלים קטנים ומופיעות בגדולים, בקפיצה ולא בהדרגה. ב־2023 פורסמה בדיקה שהראתה שהקפיצה נובעת ברוב המקרים ממדד לא רציף: כשמודדים תשובה כנכונה או שגויה בלבד, שיפור הדרגתי נראה כמו קפיצה, וכשמודדים בסולם רציף מתקבל עקום חלק. הניסוח שלהם: היכולות הצצות כביכול "מתאדות עם מדדים אחרים או עם סטטיסטיקה טובה יותר". זו דוגמה טובה לכמה מהוויכוח הזה מוכרע במכשיר המדידה ולא בעולם.

> למה זה ברשימה
>
> כי כמעט כל ויכוח אחר על המערכות האלה מתגלגל בסוף לוויכוח הזה — ומפני ששני הצדדים בו צודקים בדיוק במה שהם מודדים.

שנה	1980
חלק	הוויכוחים החוזרים
שער	מה זה כן ומה זה לא
נושא	צורה ומשמעות
הצדדים	צורה בלבד · ייצוג של עולם

להעמיק

החדר הסיני באנציקלופדיה לפילוסופיה של סטנפורד

לקריאה

הטיעון של סרל במלואו כל ההשגות המרכזיות והתשובות שלו להן. באנגלית.

https://plato.stanford.edu/entries/chinese-room/

האם בינה מלאכותית אמיתית אפשרית

לקריאה

טור עברי משנת 2011 שעובר על מבחן טיורינג ועל תופעת שער היעד הנודד.

https://www.hayadan.org.il/is-real-ai-possible-0106113

הטענה שצורה לבדה אינה מספיקה

מקורות וארכיון

המאמר עם ניסוי התמנון שמנסח את הצד הביקורתי בצורתו החזקה. באנגלית.

https://aclanthology.org/2020.acl-main.463/

על הסכנות שבתוכים סטוכסטיים

מקורות וארכיון

המאמר שנתן לעמדה הזאת את הכינוי שלה כולל ההגדרה המדויקת. באנגלית.

https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf

תשובה ישירה לניסוי התמנון

מקורות וארכיון

ארבע השגות על הטיעון שמשמעות דורשת מגע עם העולם. באנגלית.

https://link.springer.com/article/10.1007/s11023-023-09622-4

הייצוג שנמצא במערכת שראתה רק מהלכים

מקורות וארכיון

הראיה האמפירית המרכזית של הצד השני בוויכוח. באנגלית.

https://arxiv.org/abs/2210.13382

האם היכולות הצצות הן מראית עין

מקורות וארכיון

הבדיקה שהראתה שהקפיצה נובעת לרוב מבחירת המדד. באנגלית.

https://arxiv.org/abs/2304.15004

ספרים

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

עוברת על הוויכוח הזה מבפנים כחוקרת ומסרבת להכריע בזול לשני הכיוונים. לא אותר תרגום עברי.

The Alignment Problem

2020 · Brian Christian

מביא את הצד שרואה ייצוגים נבנים ואת הראיות שעליהן הוא נשען. לא אותר תרגום עברי.

Atlas of AI

2021 · Kate Crawford

כתוב מתוך העמדה שהשאלה "האם זה מבין" מסיטה את הדיון ממה שחשוב. לא אותר תרגום עברי.

האם אפשר להגיע מכאן לאינטליגנציה כללית

1956 · שבעים שנה של הבטחות מתוארכות ושתי עמדות שחיות היטב עם אותן עובדות

ב־31 באוגוסט 1955 הוגשה הצעת מחקר לקרן רוקפלר. ארבעה חתומים — ג'ון מקארתי מדרטמות', מרווין מינסקי מהרווארד, נתניאל רוצ'סטר מיבמ וקלוד שאנון ממעבדות בל — ביקשו מימון ל"מחקר בן חודשיים ועשרה אנשים בבינה מלאכותית" בקיץ 1956. הסכום המבוקש היה 13,500 דולר. המשפט שפותח את ההצעה הוא ההנחה שכל התחום עומד עליה עד היום: "כל היבט של למידה או כל תכונה אחרת של אינטליגנציה ניתנים באופן עקרוני לתיאור מדויק כל כך עד שאפשר לבנות מכונה שתדמה אותם".

מה שקרה אחר כך מתועד. ב־1966 קבעה ועדה של המועצה הלאומית למחקר בארצות הברית שתרגום מכונה "יקר יותר, מדויק פחות ואיטי יותר מתרגום אנושי"; אחרי הוצאה של כעשרים מיליון דולר, המימון הופסק. ב־1973 פרסם סר ג'יימס לייטהיל דוח לממשלת בריטניה שטען שהתחום לא עמד ביעדים שהציב, והצביע על הפיצוץ הקומבינטורי — העובדה שאלגוריתמים שעובדים יפה בגרסה מוקטנת של בעיה קורסים בגרסה האמיתית שלה. באותה שנה התקיים עימות משודר בנושא "הרובוט הרב־תכליתי הוא מיראז'". מחקר הבינה המלאכותית בבריטניה פורק כמעט כולו. לתקופה הזאת קוראים "חורף", והיה עוד אחד אחריה.

זה הרקע. עכשיו שני הצדדים, וחשוב לומר מראש שהם אינם חלוקים על העובדות אלא על המשמעות שלהן.

הצד שאומר שעוד מאותו דבר יספיק נשען על ממצא אמפירי ועל לקח היסטורי. הממצא: בעבודה מ־2020 נמצא שאיכות מודל שפה משתפרת לפי חוקי חזקה חלקים מול גודל המודל, כמות הנתונים ועוצמת החישוב, על פני יותר משבעה סדרי גודל. לא קפיצות ולא מקריות — קשר יציב שאפשר להמשיך אותו. הלקח ההיסטורי נוסח ב־2019 במאמר קצר בשם "הלקח המר", שטען שבכל פעם ששיטה כללית שמנצלת חישוב התמודדה מול שיטה שבנויה על ידע אנושי מקודד, הראשונה ניצחה — בשחמט, בגו, בזיהוי דיבור מאז תחרות משנות השבעים, ובראייה ממוחשבת. הניסוח: "שיטות כלליות שמנצלות חישוב הן בסופו של דבר היעילות ביותר, ובפער גדול". המסקנה: כל ניסיון לבנות "מבנה" מראש הוא בדיוק הטעות שחוזרת.

הצד שאומר שחסר משהו מבני מצביע על סוג מסוים של כישלון שלא נעלם עם הגודל. ב־2018 פורסם מאמר שמנה עשר בעיות, ביניהן שלמידה עמוקה "רעבה לנתונים", ש"היא רדודה ובעלת יכולת העברה מוגבלת", שאין לה "דרך טבעית להתמודד עם מבנה היררכי", ושהיא "אינה מסוגלת מטבעה להבחין בין סיבתיות למתאם". באותה שנה פורסם ניסוי מדויק על הנקודה הזאת: רשתות אומנו על פקודות ניווט פשוטות והתבקשו להרכיב מהן צירופים חדשים. כשההבדל בין האימון למבחן היה קטן הן הצליחו, וכששיטתיות אמיתית נדרשה הן "נכשלו באופן מרהיב". הטענה אינה שהמערכות חלשות אלא שהן מכלילות אחרת מבני אדם, ושהפער הזה אינו סוגיה של כמות.

ההצעה הבונה ביותר מהצד הזה פורסמה בנובמבר 2019. היא מגדירה אינטליגנציה לא כמיומנות אלא כ**יעילות ברכישת מיומנות** — כמה מהר מערכת לומדת משהו חדש, ביחס למה שכבר היה לה. הנקודה: "ידע מוקדם בלתי מוגבל או נתוני אימון בלתי מוגבלים מאפשרים לנסיין 'לקנות' רמות מיומנות שרירותיות", כלומר ציון גבוה אינו מעיד על דבר אם לא יודעים כמה השקיעו כדי לקנות אותו. לצד ההגדרה נבנה מבחן של חידות ויזואליות קצרות שכל אחת מהן דורשת להסיק כלל מדוגמאות בודדות. ב־2024 נבדקו עליו 1,729 בני אדם על כל שמונה מאות החידות: הממוצע האנושי היה 76.2 אחוזים בחלק האחד ו־64.2 בחלק הקשה יותר, ו־790 מתוך 800 החידות נפתרו בידי לפחות אדם אחד בתוך שלושה ניסיונות. כלומר אלה חידות שבני אדם ממוצעים פותרים.

ועוד עובדה שמעניינת את שני הצדדים: העבודה שמנסה לאמוד מתי ייגמר הטקסט. בגרסה מעודכנת מ־2024 נאמדה כמות הטקסט האנושי הציבורי הזמין, והמסקנה הייתה שמודלים יאומנו על כמות בסדר גודל הזה בין 2026 ל־2032. שני הצדדים קוראים את זה אחרת. האחד רואה קיר שמתקרב; השני מצביע על נתונים סינתטיים, על העברה מתחומים עתירי נתונים ועל שיפורי יעילות. זו תחזית, והיא מסומנת כאן ככזו.

ואולי הדבר הכי שימושי לשמור מהיחידה הזאת אינו טיעון אלא הרגל. בכל פעם שמישהו אומר מתי זה יקרה, כדאי לשאול שלוש שאלות: מה בדיוק הוא מבטיח, באיזה תאריך, ועל סמך איזו שיטת אמידה. הוועדה מדרטמות' ביקשה חודשיים.

למה זה ברשימה

כי אי אפשר להעריך תחזית בתחום הזה בלי לדעת מה קרה לתחזיות הקודמות — והן מתועדות היטב ומדויקות להפליא בביטחון שלהן.

שנה	1956
חלק	הוויכוחים החוזרים
שער	מה זה כן ומה זה לא
נושא	קנה מידה מול מבנה
הצדדים	עוד מאותו דבר · חסר משהו מבני

להעמיק

הלקח המר

לקריאה

המאמר הקצר שמנסח את העמדה שקנה מידה מנצח ידע מקודד ומביא את הדוגמאות ההיסטוריות. באנגלית.

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

חורפי הבינה המלאכותית

לקריאה

סקירה מתוארכת של שני גלי הקיצוץ כולל דוח לייטהיל והוועדה על תרגום מכונה. באנגלית.

https://en.wikipedia.org/wiki/AI_winter

ההצעה המקורית מ־1955

מקורות וארכיון

שני עמודים שבהם נולד המונח כולל המשפט הפותח והתקציב המבוקש. באנגלית.

http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

חוקי הגדילה

מקורות וארכיון

הממצא האמפירי שהצד הזה נשען עליו. באנגלית.

https://arxiv.org/abs/2001.08361

עשר בעיות בלמידה עמוקה

מקורות וארכיון

הרשימה המפורטת של העמדה שטוענת שחסר משהו מבני. באנגלית.

https://arxiv.org/pdf/1801.00631

על מדידת אינטליגנציה

מקורות וארכיון

ההגדרה של אינטליגנציה כיעילות ברכישת מיומנות והמבחן שנבנה סביבה. באנגלית.

https://arxiv.org/abs/1911.01547

כמה בני אדם באמת פותרים את החידות

מקורות וארכיון

הבדיקה על 1,729 נבדקים שמספקת את קו ההשוואה האנושי. באנגלית.

https://arxiv.org/abs/2409.01374

ספרים

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

הספר הטוב ביותר על הפער בין מה שהובטח למה שנמסר בכל גל בתחום. לא אותר תרגום עברי.

ההיסטוריה הקצרה של ה-AI

טובי וולש · 2023

סוקר את התחנות ואת התחזיות מהפילוסופים היוונים ועד היום. יצא בעברית בהוצאת תכלת.

The Alignment Problem

2020 · Brian Christian

מביא את שני הצדדים בלי להכריע ומסביר למה שניהם קוראים את אותן תוצאות אחרת. לא אותר תרגום עברי.

האם זה יוצר

1843 · ויכוח שנפתח ב־1843 והתשובות לא השתנו הרבה

מה שאין עליו מחלוקת: המערכות האלה מייצרות טקסט, תמונה ומוזיקה שאנשים מקבלים כטקסט, כתמונה וכמוזיקה. המחלוקת אינה על היכולת אלא על השאלה מה יצירה היא, ואם הגדרה שלה חייבת לכלול מישהו שמתכוון למשהו.

הניסוח הקלאסי של הצד האחד בא ממקום מפתיע. ב־1843 פרסמה אוגוסטה עדה לאבלייס תרגום של חיבור איטלקי על המנוע האנליטי של צ׳ארלס באבג׳, וצירפה לו הערות ארוכות משל עצמה. בהערה האחרונה, שמכילה גם את מה שנחשב לתוכנית המחשב הראשונה, היא כתבה: ״למנוע האנליטי אין יומרה ליצור שום דבר. הוא יכול לעקוב אחרי ניתוח, אבל אין לו כוח לחזות מראש יחסים או אמיתות אנליטיות. תחום עיסוקו הוא לסייע לנו להפוך לזמין את מה שאנחנו כבר יודעים״. טיורינג ציטט את המשפט הזה ב־1950 בשם ״ההשגה של ליידי לאבלייס״, והשיב לה: אולי המכונה רק מבצעת מה שנאמר לה, אבל גם אנחנו לא מפתיעים את עצמנו בכוונה — ההפתעה היא סימן לכך שלא חזינו את ההשלכות של מה שהנחנו, לא סימן לכך שיצרנו משהו יש מאין.

ההיסטוריה המעשית ישנה מכפי שנדמה, והיא נוגעת ישירות בפקולטות המוזיקה והאמנות. ב־1956, באוניברסיטת אילינוי, חיברו לג׳רן הילר ולאונרד אייזקסון יצירה למרובע כלי קשת בעזרת מחשב במשקל חמישה טונות. היא נקראה ״סוויטת אילייאק״ ומאוחר יותר ״רביעייה מספר 4״, ומקובל לראות בה את היצירה המוזיקלית הראשונה שהולחנה במחשב. ארבעת הפרקים שלה הם ארבעה ניסויים: הראשון יוצר קווי מלודיה לפי כללים, השני מוסיף ארבעה קולות, והרביעי כבר לא משתמש בכללים מוזיקליים אלא בשרשראות מרקוב, כלומר בהסתברות מעבר בין מצבים. שלושת הפרקים הראשונים בוצעו בידי רביעיית סטודנטים ב־9 באוגוסט 1956. ובאמנות החזותית עבד הרולד כהן, צייר בריטי, מסוף שנות השישים ועד מותו ב־2016 על תוכנה בשם ארון, שייצרה ציורים שמכונות עם עטים ומכחולים ביצעו על נייר. כהן עצמו תיאר את היחס בינו לבין ארון כשיתוף פעולה, ולא כתחליף.

וכעת שני הצדדים.

הצד שאומר שזה צירוף מחדש ולא יצירה טוען שיצירה אינה תוצר אלא יחס: מישהו נמצא במצב, משהו מוטל על הכף, ומתוך זה נולדת בחירה. מכונה שממקסמת סבירות שואפת מעצם הגדרתה אל המרכז של מה שכבר נעשה, ואילו אמנות חיה בקצוות. מסאי עברי ניסח את זה כך: ״אמנות מחפשת את הפגימה, הסטייה, הצרימה, הליקוי. אמנות בנויה על הקצנה, לא על חתירה לממוצע״. בגרסה הזאת, גם אם התוצר בלתי מובחן, ההבדל נשאר, מפני שהוא לא נמצא בתוצר.

הצד שאומר שזו יצירה משיב בשתי טענות. הראשונה: גם יצירה אנושית היא צירוף מחדש של מה שנספג, והדרישה ל"כוונה" בלתי ניתנת להפרכה מבחוץ — אי אפשר לבדוק אותה בשום ניסוי, ולכן היא הגדרה ולא ממצא. השנייה היסטורית: כמעט כל טכנולוגיה חדשה נפסלה תחילה כלא־אמנות, מהצילום ועד הדגימה האלקטרונית, ותמיד מאותו נימוק בדיוק.

יש גם מדידה, והיא מעניינת פחות ממה שנדמה ויותר ממה שנדמה. בנובמבר 2024 פורסם מחקר שבו הוצגו למשתתפים שירים של עשרה משוררים אנגלים — מצ'וסר ושייקספיר ועד סילביה פלאת' — לצד שירים שנוצרו במכונה בסגנונם, והם התבקשו לזהות מה מה. בניסוי הראשון השתתפו 1,634 איש. שיעור הזיהוי הנכון היה 46.6 אחוזים, כלומר **מתחת** למקריות. יותר מכך: המשתתפים נטו לסמן את השירים המכונתיים כאנושיים יותר מאשר את האנושיים. ההסבר שהציעו החוקרים אינו מחמיא לאיש: השירים המכונתיים פשוטים ונגישים יותר, והמשתתפים כתבו "זה לא מסתדר" על שירים אנושיים מאה ארבעים וארבע פעמים לעומת עשרים ותשע על מכונתיים. הם לא זיהו שירה טובה יותר; הם העדיפו את המובן. הביקורת על המחקר מוסיפה שהמשתתפים לא הכירו את המשוררים, ושהשירים המכונתיים נוצרו בהנחיה בסיסית אחת בלבד.

ובצד המשפטי יש כבר הכרעות, וכדאי להכיר אותן כעובדות מתוארכות. בפברואר 2023 קבע משרד זכויות היוצרים האמריקאי ברומן גרפי שנוצר בסיוע מחולל תמונות שהטקסט מוגן, שהעריכה והסידור מוגנים, אבל התמונות עצמן אינן — מפני שההנחיות שהמשתמשת כתבה הן "הצעות ולא פקודות", ומפני ש"כיוון שהמערכת מתחילה מרעש אקראי שמתפתח לתמונה סופית, אין שום ערובה שהנחיה מסוימת תייצר פלט חזותי מסוים". ובמרץ 2025 פסק בית משפט פדרלי לערעורים בוושינגטון שמכונה אינה יכולה להיות יוצרת לצורכי זכויות יוצרים, בנימוק מעשי: החוק מדבר על יוצרים שיש להם תוחלת חיים, משפחה, מקום מגורים, אזרחות, כוונה פלילית וחתימה. הפסיקה הזאת אינה אומרת שהתוצר חסר ערך; היא אומרת שהוא לא שייך לאף אחד.

> למה זה ברשימה
>
> כי הוויכוח הזה קדם למכונות בכמעט מאה וחמישים שנה — וההתנגדות המקורית נוסחה בידי מי שכתבה את התוכנית הראשונה.

שנה	1843
חלק	הוויכוחים החוזרים
שער	מה זה כן ומה זה לא
נושא	יצירה ומקוריות
הצדדים	צירוף מחדש · יצירה

להעמיק

סוויטת אילייאק

לקריאה

דף המוזיאון על היצירה המוזיקלית הראשונה שהולחנה במחשב כולל מה עשה כל פרק. באנגלית.

https://distributedmuseum.illinois.edu/exhibit/illiac-suite/

הרולד כהן וארון

לקריאה

התערוכה במוזיאון ויטני על חמישים שנות עבודה משותפת בין צייר לתוכנה שלו. באנגלית.

https://whitney.org/exhibitions/harold-cohen-aaron

כשמוסיפים בינה לאמנות

לקריאה

מסה עברית שמנסחת את הצד שטוען שאמנות מחפשת את הסטייה ולא את הממוצע.

https://alaxon.co.il/article/כשמוסיפים-בינה-לאמנות/

ההערות של לאבלייס במלואן

מקורות וארכיון

כולל ההשגה המפורסמת על כך שלמנוע אין יומרה ליצור דבר ואת התוכנית הראשונה. באנגלית.

https://www.fourmilab.ch/babbage/sketch.html

התשובה של טיורינג להשגה

מקורות וארכיון

המאמר של 1950 כולל הסעיף שדן בטענה של לאבלייס ובשאלה מה נחשב הפתעה. באנגלית.

https://www.hec.edu/sites/default/files/documents/Computing Machinery and Intelligence by Alan Turing.pdf

כשאנשים ניסו לזהות שירה מכונתית

מקורות וארכיון

המחקר שבו הזיהוי היה מתחת למקריות והמכונתי דורג גבוה יותר וההסבר שהוצע לכך. באנגלית.

https://www.nature.com/articles/s41598-024-76900-1

ההחלטה על הרומן הגרפי

מקורות וארכיון

המכתב שמפריד בין הטקסט והעריכה שמוגנים ובין התמונות שאינן ומסביר למה. באנגלית.

https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

ספרים

The Creativity Code

2019 · Marcus du Sautoy

מתמטיקאי בודק יצירה מכונתית במוזיקה באמנות ובמתמטיקה ומנסה להגדיר מה חסר. לא אותר תרגום עברי.

Atlas of AI

2021 · Kate Crawford

על מה שנכנס למאגרים שמהם נבנה "סגנון" ומי נתן לזה רשות. לא אותר תרגום עברי.

Artificial Intelligence: A Guide for Thinking Humans

2019 · Melanie Mitchell

הפרקים על יצירה ועל דמיון מסבירים מה מחולל עושה ומה הוא לא. לא אותר תרגום עברי.

יוצרים נתונים וזכויות

2023 · שאלה אחת שנחתכת אחרת כמעט בכל מדינה

מה שאין עליו מחלוקת: מאגרי האימון מכילים יצירות מוגנות בזכויות יוצרים — ספרים, מאמרים, תצלומים, ציורים. איש מהצדדים לא טוען אחרת. המחלוקת היא אם השימוש הזה מחייב רשות או תשלום, או שהוא נכנס לחריגים שקיימים ממילא בדין.

לפני המשפט, העובדה שסביבה כולם מתווכחים: כמה מהחומר בעצם נשמר בתוך המודל. יש על זה מדידות טובות. בבדיקה מ־2021 על מודל שפה הצליחו חוקרים לחלץ 604 קטעים שאומתו כטקסט מקורי מחומר האימון, מתוך 1,800 מועמדים שייצרו — שליש הצלחה, ובווריאנט הטוב ביותר שני שלישים. בין הקטעים היו ידיעות חדשותיות, רישיונות, הודעות בפורומים, קטעי קוד, ופרטי קשר של אנשים פרטיים. הגורם המכריע היה חזרה: רצף שהופיע שלושים ושלוש פעמים ומעלה חולץ באופן אמין. המחברים הדגישו שהם "מעריכים בחסר משמעותי" את הכמות האמיתית.

בתמונות המספרים הפוכים באופן מאלף. בבדיקה מ־2023 ייצרו החוקרים 175 מיליון תמונות מועמדות — חמש מאות תמונות לכל אחת מ־350 אלף ההנחיות הכי כפולות במאגר — וחילצו מהן 109 תמונות אימון מאומתות. זה שיעור של כשש מאיות האחוז, ורק אחרי שהחיפוש הוטה בכוונה לכיוון הכפילויות. ותמונה שנשמרה כך הופיעה בדרך כלל לפחות מאה פעמים במאגר, כלומר היא אחת ממיליון. אז שינון קיים, הוא אמיתי, הוא נדיר, והוא נגרם כמעט כולו מכפילות. שני הצדדים מצטטים את המחקרים האלה, כל אחד את המספר שמשרת אותו.

הצד של היוצרים טוען שהשאלה אינה כמה נשמר אלא מה נלקח. יצירה הועתקה במלואה אל תוך תהליך מסחרי בלי רשות ובלי תשלום; העובדה שהתוצר אינו עותק אינה מכשירה את ההעתקה שבדרך. ארגוני סופרים מוסיפים טענה על שוק: אם קיים או יכול להתקיים שוק רישוי לאימון, אז לקיחה ללא רישוי פוגעת בו ישירות. ואם המערכת מייצרת תחליפים לאותו שוק שממנו למדה, הנזק אינו היפותטי.

הצד שטוען שזו למידה מותרת משיב שהעתקה טכנית לצורך ניתוח אינה הפרה, בדיוק כפי שמנוע חיפוש מעתיק את הרשת כדי לאנדקס אותה; שמה שנלמד הוא דפוסים ולא ביטוי; ושמעולם לא נדרש רישוי כדי ללמוד מיצירה, לנתח אותה או לספור בה מילים. שופט אמריקאי ניסח את קצה הטיעון הזה כך, כשדחה את טענת הנזק לשוק: זה דומה לטענה שלימוד ילדי בית ספר לכתוב היטב יגרום להצפה של יצירות מתחרות.

וכאן התמונה המשפטית, מתוארכת, ועם מדינות שהגיעו למסקנות שונות.

בישראל פורסמה חוות דעת רשמית כבר ב־18 בדצמבר 2022, מטעם מחלקת ייעוץ וחקיקה במשרד המשפטים. מסקנתה: "שימוש בתכנים מוגנים בזכויות יוצרים לצורך אימון מכונה חוסה תחת הסדרי השימושים

המותרים", בהסתמך על שלושה מסלולים בחוק — שימוש הוגן, שימוש אגבי והעתקה זמנית. חוות הדעת מסייגת בשני דברים חשובים: היא חלה על תהליך הלמידה ולא בהכרח על הפלט, והיא אינה חלה על מאגר צר שנבנה כדי להתחרות ישירות ביוצר המקורי.

בארצות הברית ניתנו שלוש הכרעות בתוך חמישה חודשים והן אינן מסתדרות זו עם זו בקלות. בפברואר 2025 נקבע בדלאוור שאימון על תקצירים משפטיים **אינו** שימוש הוגן, מפני שהמוצר שנבנה התחרה ישירות במוצר שממנו נלקחו. ביוני 2025 נקבע בקליפורניה שאימון על ספרים שנרכשו כדין **הוא** שימוש הוגן ואף "טרנספורמטיבי במידה יוצאת דופן", אבל שהורדת אותם ספרים מאתרי פיראטיות אינה מוגנת כלל — וזו הפרה במזיד. התיק ההוא הסתיים בהסדר פשרה של מיליארד וחצי דולר שהוגש לאישור ראשוני בספטמבר 2025, כחמש מאות אלף כותרים, כשלושת אלפים דולר לכותר. שימו לב במה בדיוק הוא עוסק: בהשגה הפיראטית, לא באימון עצמו. ויומיים לאחר אותה הכרעה, באותו בית משפט, זכתה חברה אחרת בתביעה דומה מטעם שלושה עשר סופרים — לא מפני שהאימון הוכר כהוגן בכל מקרה, אלא מפני שהתובעים לא הביאו ראיות לנזק לשוק. השופט כתב במפורש שתיאוריית ההצפה של השוק ראויה להישמע, אבל שהובאה לפניו השערה ולא נתונים.

בבריטניה הסתיים בנובמבר 2025 המשפט שנחשב למבחן הגדול. התובעת ויתרה באמצע הדיון על עצם טענת האימון, מחוסר ראיות שהאימון התרחש בתחום השיפוט הבריטי, וויתרה גם על טענת הפלט. מה שנקבע לגופו: משקלי המודל "אינם מכילים עותק של אף יצירה", ולכן המודל עצמו אינו "עותק מפר". **באיחוד האירופי** אין דוקטרינת שימוש הוגן, ובמקומה יש בדירקטיבה מ־2019 שני חריגים לכריית טקסט ונתונים: אחד למחקר מדעי ואחד כללי, שממנו אפשר לצאת אם בעל הזכויות שמר על זכותו "בצורה קריאה למכונה". בית משפט בהמבורג קבע שיצירת מאגר לאימון היא אכן כרייה כזאת, ואגב כך שאל שאלה טובה: האם הודעה בשפה חופשית באתר נחשבת "קריאה למכונה", כשהמערכות האלה קוראות שפה חופשית מצוין.

כל מה שבפסקאות האחרונות נכון לספטמבר 2026 והוא הדבר המתיישן ביותר בפקולטה. מה שלא יתיישן הוא המבנה: מי שמחליט אינו מכריע אם "זה בסדר", אלא איזה משלושה או ארבעה חריגים קיימים בדין המקומי חל, ועל איזה שלב — ההשגה, האימון, או הפלט.

> למה זה ברשימה
>
> כי בישראל כבר יש חוות דעת רשמית בנושא משנת 2022 והיא מגיעה למסקנה אחרת מזו של חלק מפסקי הדין בארצות הברית.

שנה	2023
חלק	הוויכוחים החוזרים
שער	מה זה עושה לעולם
נושא	זכויות יוצרים
הצדדים	רשות ותמלוג · למידה מותרת

להעמיק

ההכרעה שהפרידה בין אימון להשגה

לקריאה

ניתוח פסק הדין מיוני 2025 לפי ארבעת מבחני השימוש ההוגן. באנגלית.

https://www.wiggin.com/publication/bartz-v-anthropic-first-court-decision-on-fair-use-defense-in-llm-training/

ההכרעה ההפוכה

לקריאה

למה שם נקבע שהאימון אינו שימוש הוגן ומה היה שונה בעובדות. באנגלית.

https://www.jenner.com/en/news-insights/client-alerts/court-decides-that-use-of-copyrighted-works-in-ai-training-is-not-fair-use-thomson-reuters-enterprise-centre-gmbh-v-ross-intelligence-inc

פסק הדין הבריטי

לקריאה

מה נותר מהתביעה אחרי שהוויתורים באמצע המשפט ומה נקבע על משקלי המודל. באנגלית.

https://ipkitten.blogspot.com/2025/11/getty-images-v-stability-long-awaited.html

הצד של הסופרים

לקריאה

ארגון הסופרים האמריקאי מסביר מה ההסדר כולל ומה הוא במפורש אינו כולל. באנגלית.

https://authorsguild.org/advocacy/artificial-intelligence/what-authors-need-to-know-about-the-anthropic-settlement/

חוות הדעת של משרד המשפטים

מקורות וארכיון

המסמך הישראלי מדצמבר 2022 על שימוש בתכנים מוגנים לצורך למידת מכונה כולל הסייגים. בעברית.

https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/machine-learning.pdf

כמה טקסט באמת נשמר בתוך מודל

מקורות וארכיון

המחקר שחילץ 604 קטעי אימון מאומתים ומיפה את סוגי התוכן שנשמרו. באנגלית.

https://arxiv.org/pdf/2012.07805

אותה בדיקה בתמונות

מקורות וארכיון

175 מיליון תמונות שנוצרו ומאה ותשע שחולצו והממצא שכפילות היא הגורם. באנגלית.

https://arxiv.org/abs/2301.13188

ספרים

Atlas of AI

2021 · Kate Crawford

הפרקים על בניית מאגרים ועל מה שנלקח מהם בלי לשאול. לא אותר תרגום עברי.

Ghost Work

2019 · Mary L. Gray and Siddharth Suri

על העבודה האנושית שנלקחת בשקט לתוך מערכות שנראות אוטומטיות. לא אותר תרגום עברי.

נקסוס

יובל נח הררי · 2024

על מה שקורה כשמידע עובר מרשות אחת לאחרת ומי נהנה מהמעבר. נכתב בעברית.

מה זה עושה לעבודה

1811 · מאתיים שנה של נתונים ושתי מסקנות הפוכות מהם

בין 1811 ל־1816 פעלו באנגליה קבוצות של פועלי טקסטיל שנודעו כלודיטים. הם שלחו מכתבי איום למעסיקים, פרצו למפעלים ושברו מכונות, תקפו מעסיקים שופטי שלום וסוחרי מזון, ונלחמו בחיילים. ומה שהם רצו אינו מה שמיוחס להם. הם לא התנגדו למכונות. בנוטינגהמשייר הם התנגדו לשימוש במסגרות רחבות כדי להוריד שכר ולדלל מיומנות; ביורקשייר, אורגים התנגדו לנולי כוח שייתרו עבודה מיומנת. הדרישה הייתה כלכלית ולא טכנולוגית. המילה "לודיט" כשם גנאי לשונא קדמה היא המצאה מאוחרת.

הכלכלן שלקח אותם ברצינות היה דייוויד ריקרדו, וגם הוא רק בשלב שני. במהדורה הראשונה של "עקרונות הכלכלה המדינית" מ־1817 הוא כתב שאין מכונה שמייתרת לחלוטין עבודת אדם. ב־1819 הוא ישב בוועדה פרלמנטרית על חוקי העניים וראה מקרוב את מצב האורגים. ב־1821 הוא הוסיף למהדורה השלישית פרק חדש בשם "על מכונות", ופתח אותו בהודאה: דעותיו "עברו, במחשבה נוספת, שינוי ניכר". במכתב פרטי מאותה שנה הוא ניסח את זה חד יותר: אילו מכונות היו יכולות לבצע את כל העבודה שעבודת אדם מבצעת כיום, לא היה ביקוש לעבודה.

מה שקרה בפועל נמדד מאוחר יותר ונקרא "ההפוגה של אנגלס". בין 1780 ל־1840 גדל התפוקה לעובד בבריטניה בארבעים ושישה אחוזים, והשכר הריאלי גדל בשנים עשר. בשלושת העשורים שבין 1800 ל־1830 התפוקה לעובד עלתה בכ־0.63 אחוזים בשנה והשכר הריאלי לא עלה כלל. במקביל, שיעור הרווח עלה מכעשרה אחוזים בסוף המאה השמונה־עשרה ליותר מעשרים באמצע התשע־עשרה, וחלקה של העבודה בהכנסה ירד מכשישים אחוזים לכחמישים. ורק סביב 1900 חזר חלקה של העבודה לכשישים. אז התשובה לשאלה "האם בסוף כולם הרוויחו" היא כן, והתשובה לשאלה "כמה זמן זה לקח" היא שני דורות.

וכעת שני הצדדים, שניהם עם נתונים.

הצד שמצביע על השלמה מביא את מקרה הכספומט, המצוטט ביותר בוויכוח הזה. בין 1995 ל־2010 התרבע מספר הכספומטים בארצות הברית מכמאה אלף לכארבע מאות אלף. מספר הפקידים בבנקים לא ירד — הוא עלה קלות, מכחצי מיליון לכ־550 אלף. ההסבר: הכספומט הוזיל את התפעול של סניף, ולכן נפתחו יותר סניפים — מספר הסניפים בערים עלה בארבעים ושלושה אחוזים — וכל סניף העסיק פחות פקידים, מעשרים לשלושה עשר בין 1988 ל־2004. במקביל תפקיד הפקיד השתנה ממניית שטרות למכירת שירותים. הצד הזה מצביע גם על שני מנגנונים: ראשית, שיפור של משימה אחת בתהליך מעלה את הערך של שאר המשימות בו, כי התהליך שווה כמו החוליה החלשה. ושנית, פרדוקס פולני — "אנחנו יודעים יותר ממה שאנחנו יכולים לומר" — ולכן דווקא מיומנויות שקשה לנסח נשארות אצל אדם.

הצד שמצביע על עקירה מוסיף לאותו סיפור את ההמשך שבדרך כלל נחתך. נכון ל־2025 רשומות בארצות הברית 339,200 משרות של פקידי בנק, והתחזית הרשמית היא ירידה של שלושה עשר אחוזים בעשור הקרוב, בגלל בנקאות מקוונת. כלומר סיפור הכספומט נכון לתקופתו ואינו חוק טבע. הצד הזה מצביע גם על מדידה ישירה: מחקר שבדק את ארצות הברית בין 1990 ל־2007 מצא שכל רובוט נוסף לאלף עובדים הוריד את יחס התעסוקה לאוכלוסייה ב־0.18 עד 0.34 נקודות אחוז ואת השכר ברבע עד חצי אחוז. ההשפעה ממוקדת גיאוגרפית ומקצועית, כלומר הממוצע הארצי מסתיר את מי שנפגע.

ולתחזיות עצמן יש היסטוריה, והיא החלק השימושי ביותר. בספטמבר 2013 פורסמה עבודה שבדקה 702 מקצועות וקבעה ש"כארבעים ושבעה אחוזים מכלל התעסוקה בארצות הברית נמצאים בסיכון". המשפט הזה הפך לכותרת בעולם כולו, בלי הסייג שנכתב לידו: מדובר במקצועות שניתן **פוטנציאלית** למחשב "במספר לא מוגדר של שנים, אולי עשור או שניים". שלוש שנים אחר כך חזר צוות אחר על החישוב בשיטה אחרת — לא ברמת מקצוע אלא ברמת משימה, על נתוני מיומנויות מעשרים ואחת מדינות — והגיע לתשעה אחוזים. ההבדל אינו ויכוח על עובדה אלא על יחידת המדידה: מקצוע הוא צרור משימות, ובתוך מקצוע "בסיכון גבוה" יש משימות שלא ניתנות למיחשוב. בישראל נעשה ב־2015 חישוב לפי השיטה הראשונה, והוא העריך שכארבעים אחוז משעות העבודה במשק עלולות להיות מוחלפות בתוך שני עשורים.

ולישראל יש גם מדידה עדכנית יותר. בדוח שפרסם בנק ישראל במרץ 2025 מופו משלחי היד במשק לפי סוג החשיפה: כשלושים ותשעה אחוזים מהעובדים במשרות שבהן הטכנולוגיה משלימה את עבודתם, כשמונה אחוזים נוספים בתחום הטכנולוגיה עצמו, כעשרים אחוזים במשלחי יד תחליפיים שבהם הסיכון הוא לירידת ביקוש, וכשליש במשרות ניטרליות. ופרט אחד בולט: כשבעים ואחד אחוזים מהעובדים החשופים להחלפה הם נשים.

ומה נמדד עד כה בפועל, בזהירות, כי זה טרי. בניסויים על משימות בודדות נמדדו זינוקים גדולים: במשימת תכנות מוגדרת, קבוצה שקיבלה כלי סיוע סיימה מהר יותר בכחמישים ושישה אחוזים. אבל כשבודקים שוק עבודה שלם התמונה שונה. מחקר מ־2025 עקב אחרי כ־25 אלף עובדים בכשבעת אלפים מקומות עבודה בדנמרק, באחד עשר מקצועות. התוצאה על שכר ועל שעות עבודה הייתה "אפסים מדויקים" — רווחי סמך ששוללים שינוי ממוצע גדול מאחוז. החיסכון בזמן שדיווחו המשתמשים עמד על 2.8 אחוזים משעות העבודה בממוצע, ונוצרו גם משימות חדשות. כלומר, נכון ל־2025: האצה אמיתית ברמת המשימה, ושום אפקט מדיד ברמת השכר. שתי העובדות נכונות ואינן סותרות, וזה בדיוק המצב שבו כל צד יכול לצטט מחקר אמיתי ולהיות משוכנע.

> למה זה ברשימה
>
> כי זה הוויכוח היחיד בפקולטה שיש לו מאתיים שנה של נתונים אמיתיים — ושני הצדדים קוראים אותם נכון.

שנה	1811
חלק	הוויכוחים החוזרים
שער	מה זה עושה לעולם
נושא	טכנולוגיה ותעסוקה
הצדדים	עקירה · השלמה

להעמיק

מה הלודיטים באמת רצו

לקריאה

הארכיון הלאומי הבריטי על מה שהם עשו על מה שדרשו ועל ההבדל בין התנגדות למכונה להתנגדות לשימוש בה. באנגלית.

https://www.nationalarchives.gov.uk/education/resources/why-did-the-luddites-protest/

בנק ישראל על חשיפת המשק

לקריאה

מיפוי משלחי היד בישראל לפי השלמה תחליפיות וניטרליות כולל הפילוח המגדרי. בעברית.

https://www.boi.org.il/publications/pressreleases/11-3-25/

מה זה אומר להיות לודיט

לקריאה

טור עברי שמתקן את המשמעות של המילה ומסביר מה התנועה באמת דרשה.

https://www.calcalist.co.il/calcalistech/article/rygjet885

ההפוגה של אנגלס

מקורות וארכיון

העבודה שמדדה את הפער בין עליית התפוקה לעליית השכר במהפכה התעשייתית עם הטבלאות. באנגלית.

https://www.nuff.ox.ac.uk/Users/Allen/engelspause.pdf

התחזית של ארבעים ושבעה אחוזים

מקורות וארכיון

המקור עצמו כולל הסייג שכמעט תמיד נחתך מהציטוט. באנגלית.

https://oms-www.files.svdcdn.com/production/downloads/academic/future-of-employment.pdf

אותו חישוב ברמת משימה

מקורות וארכיון

החישוב החוזר שהגיע לתשעה אחוזים והסביר בדיוק במה השיטות נבדלות. באנגלית.

https://wecglobal.org/uploads/2019/07/2016_OECD_Risk-Automation-Jobs.pdf

מרכז טאוב על שוק העבודה העתידי

מקורות וארכיון

היישום הישראלי של שיטת החישוב הראשונה משנת 2015. בעברית.

https://www.taubcenter.org.il/wp-content/uploads/2020/12/futurelabormarket2015heb83.pdf.pdf

מה נמדד בפועל בדנמרק

מקורות וארכיון

מעקב אחרי כ־25 אלף עובדים באחד עשר מקצועות והתוצאה האפסית על שכר ושעות. באנגלית.

https://www.nber.org/system/files/working_papers/w33777/revisions/w33777.rev0.pdf

ספרים

The Technology Trap

2019 · Carl Benedikt Frey

אותו חוקר שכתב את תחזית ארבעים ושבעה האחוזים כותב ספר על כמה זמן לקח לרווחים להגיע ומי שילם בינתיים. לא אותר תרגום עברי.

Learning by Doing

2015 · James Bessen

הצד שטוען שטכנולוגיה מעבירה עובדים לתפקידים חדשים ולא מוחקת אותם עם הנתונים ההיסטוריים. לא אותר תרגום עברי.

Power and Progress

2023 · Daron Acemoglu and Simon Johnson

הטענה שכיוון הטכנולוגיה הוא בחירה ולא גורל ושמי שנהנה ממנה נקבע במוסדות. לא אותר תרגום עברי.

אחריות החלטות ומעקב

2021 · מי עונה כשמערכת מחליטה על אדם

בין 1999 ל־2015 הועמדו לדין והורשעו בבריטניה כאלף מנהלי סניפי דואר, על סמך נתונים שהפיקה מערכת הנהלת חשבונות בשם הורייזן. הם שילמו מכיסם גירעונות שלא היו, נשלחו לכלא, ואיבדו את פרנסתם ואת שמם. ועדת חקירה ציבורית פרסמה את הכרך הראשון של הדוח שלה ביולי 2025: לפחות שלושה עשר אנשים שמו קץ לחייהם, ועוד חמישים ותשעה שקלו זאת. יושב ראש הוועדה כתב שהדואר המלכותי "שימר את הבדיה שהנתונים שלו תמיד מדויקים". ובאפריל 2021 ביטל בית המשפט לערעורים שלושים ותשע הרשעות מתוך ארבעים ושתיים שנדונו, במשפט אחד שהוא הלב של היחידה הזאת: הדואר "ביקש למעשה להפוך את נטל ההוכחה: הוא התייחס למה שאינו אלא גירעון שהראתה מערכת חשבונאית בלתי אמינה, כאילו היה הפסד שאין עליו עוררין".

שימו לב שאין כאן שום בינה מלאכותית. הורייזן הייתה תוכנה רגילה עם באגים. מה שהפך אותה לאסון הוא מוסד שהחליט שהפלט של המחשב הוא עובדה, והאדם שמולו הוא הצד שצריך להוכיח אחרת. זו התבנית, והיא לא השתנתה כשהטכנולוגיה השתנתה.

היא חזרה בהולנד בקנה מידה אחר. בדצמבר 2020 פרסמה ועדת חקירה פרלמנטרית דוח בשם "אי־צדק חסר תקדים" על גביית יתר של קצבאות טיפול בילדים. הקביעה: "עקרונות יסוד של שלטון החוק הופרו". שלושה דברים פעלו יחד — טיפול בבני אדם כקבוצה ולא כפרטים, כלל של "הכול או כלום" שלפיו כל טעות או אי־התאמה הובילה להשבת הקצבה של שנה שלמה, והחלת מושגי זדון ורשלנות חמורה על אנשים שלא הבינו טופס. בחוק לא הייתה פסקת מצוקה שמאפשרת התחשבות במקרה פרטי. מודל דירוג סיכונים היה אחד המרכיבים, ובתוכו שימש מדד "אזרחות הולנדית" באופן מפלה בין מרץ 2016 לאוקטובר 2018. הממשלה ההולנדית התפטרה ב־15 בינואר 2021, ורשות הגנת הפרטיות הטילה קנס של 2.75 מיליון יורו בדצמבר של אותה שנה. באוסטרליה פעלה תוכנית דומה לגביית חובות שהתבססה על מיצוע הכנסה שנתית לתקופות דו־שבועיות. ועדת חקירה מלכותית קבעה ביולי 2023 שהיא הייתה "מנגנון גס ואכזרי, לא הוגן ולא חוקי, שגרם לאנשים רבים להרגיש כמו פושעים", ושהשיטה עצמה נגדה את חוק הביטחון הסוציאלי.

וכעת שני הצדדים, שניהם רציניים.

הצד שטוען שהחלטה ממוכנת עדיפה אינו טוען שהמכונה ניטרלית. הוא טוען שקו ההשוואה גרוע. במטא־אנליזה משנת 2000 שכיסתה 136 מחקרים על ניבוי קליני מול ניבוי מכני, כשישים ושלושה מהם העדיפו בבירור את הניבוי המכני, כשישים וחמישה הראו שוויון, ורק שמונה העדיפו את שיקול הדעת האנושי. במחקר על מקלט מדיני בארצות הברית נמצא שבבית הדין במיאמי, מבקש מקלט קולומביאני עמד בפני סיכוי של חמישה אחוזים לזכות בפני שופט אחד ושמונים ושמונה אחוזים בפני שופט אחר באותו בניין, ושמחצית

מהשופטים שם סטו ביותר מחמישים אחוז מהממוצע של בית הדין. ובמחקר על 758,027 החלטות שחרור בערבות בניו יורק נמצא שכלל ניבוי היה יכול להוריד פשיעה בעד 24.7 אחוזים בלי לכלוא יותר אנשים, או להוריד כליאה בעד 41.9 אחוזים בלי להעלות פשיעה — וכשמשווים בקצב פשיעה זהה, הכלל היה כולא כארבעים אחוז פחות מיעוטים מכפי שהשופטים כלאו בפועל. הטענה היא לא שהמכונה הוגנת, אלא שהיא לפחות עקבית ושאפשר לבדוק אותה. שיקול דעת של אדם אי אפשר לפתוח ולבדוק בכלל.

הצד שמצביע על פער האחריות מנסח את הבעיה כמבנית ולא טכנית, וזה נוסח כבר ב־1996 בארבעה חסמים שעומדים במבחן הזמן. הראשון, "בעיית הידיים הרבות": כשתקלה היא תוצר של עשרות אנשים לאורך שנים, מי שנמצא בקצה הסיבתי אינו מי שקיבל את ההחלטה, ולכן אין למי לפנות. השני, ההתייחסות לבאגים כאל גזירת טבע — "תמיד יש עוד באג" — שהופכת נזק לאסון טבע ולא לאחריות של מישהו. השלישי, האשמת המחשב: ברגע שנמצא הסבר אחד לטעות, תפקידם של בני האדם בשרשרת נאמד בחסר ולפעמים נעלם. והרביעי, בעלות בלי חבות. ההבחנה החשובה שם: חבות נמדדת ממצוקתו של הנפגע ואפשר לבטח אותה ולפזר אותה; אחריות נמדדת מיחסו של הפועל לתוצאה, ואי אפשר. ב־2008 נוסחה הגרסה המשפטית: "קוד, ולא כללים, קובע את תוצאות ההכרעה. מתכנתים משנים בהכרח כללים קיימים כשהם מטמיעים אותם בקוד, בדרכים שהציבור, נבחרי הציבור ובתי המשפט אינם יכולים לבחון". זה נכתב שש־עשרה שנה לפני שוועדות החקירה אישרו אותו.

המקרה שמפרק את הסיסמאות של שני הצדדים הוא אמסטרדם. העירייה בנתה מערכת לדירוג בקשות לקצבה שהתבססה על חמישה עשר מאפיינים, ובמכוון **לא** כללה מגדר, לאום וגיל. בבדיקת הטיה במאי 2022 התברר שהמודל מסמן בטעות בעלי אזרחות זרה כמעט פי שניים מאשר מערביים, ושהוא מסמן בטעות גברים בארבעה עשר אחוזים יותר. העירייה שקללה מחדש את נתוני האימון, והתייעצה עם מועצת נציגים של מקבלי הקצבה עצמם. אחרי התיקון, בבדיקות מעבדה, שיעורי הסימון השגוי השתוו. ואז יצא הפיילוט לאוויר באביב 2023, על כאלף ושש מאות בקשות, וההטיה התהפכה: עכשיו נגד אזרחים הולנדים ונגד נשים, ובנוסף הופיעה אפליה חדשה נגד מבקשים עם ילדים. והמערכת, שנבנתה כדי לסמן **פחות** אנשים מפקידים אנושיים, סימנה יותר. בנובמבר 2023 עצר אותה הממונה בעירייה. העלות: כחצי מיליון יורו. מי שאומר "תנקו את הנתונים ותבדקו הטיה" ומי שאומר "אלגוריתמים תמיד מפלים נגד מיעוטים" — שניהם נופלים על המקרה הזה.

ומה הדין אומר. ברגולציית הפרטיות האירופית יש סעיף שמעניק זכות שלא להיות נתון להחלטה המבוססת **אך ורק** על עיבוד אוטומטי שיש לה השפעה משפטית או דומה לה. הוא אינו אוסר החלטות אוטומטיות; הוא מתנה אותן. בדצמבר 2023 קבע בית הדין האירופי לצדק שגם חישוב ציון אשראי בידי לשכת אשראי הוא "החלטה" לעניין הזה, אם הגורם שמקבל אותו "נשען עליו בחוזקה" — כלומר האחריות אינה רק של הבנק שסירב, אלא גם של מי שחישב. ובפברואר 2025 נקבע שעל בעל השליטה להסביר "את הנוהל ואת העקרונות שהופעלו בפועל", ושטענת סוד מסחרי אינה חוסמת אוטומטית אלא מועברת להכרעת הרשות או בית המשפט. הניסוח המדויק של המצב הוא לא "יש זכות להסבר" ולא "אין כלום", אלא: לא את קוד המקור, אבל גם לא כלום.

לגבי חוק הבינה המלאכותית האירופי כדאי להיזהר, כי כאן הטעות הנפוצה ביותר. החוק נכנס לתוקף באוגוסט 2024, האיסורים וחובת האוריינות מיושמים מפברואר 2025, וגוף התקנות הכללי מאוגוסט 2026. אבל תיקון שנכנס לתוקף ביולי 2026 דחה את החובות על מערכות "בסיכון גבוה" — בדיוק אלה שיחולו על דירוג אשראי, על זכאות לקצבאות, על מיון מועמדים לעבודה ועל זיהוי ביומטרי בידי רשויות אכיפה — לדצמבר 2027 ולאוגוסט 2028. כלומר נכון לספטמבר 2026 הן **אינן** בתוקף. זה הפרט הכי מתיישן ביחידה והכי קל לטעות בו.

ובישראל אין מקבילה לזכות הזאת כלל. תיקון 13 לחוק הגנת הפרטיות נכנס לתוקף באוגוסט 2025, והוא עוסק בסמכויות אכיפה, בעיצומים כספיים ובממונה על הגנת הפרטיות — לא בהחלטות אוטומטיות. הנחיית הרשות להגנת הפרטיות בנושא בינה מלאכותית היא טיוטה. ובמעקב בשטח כדאי להפריד בין שני דברים שמתבלבלים כל הזמן: "עין הנץ" היא מערכת לקריאת לוחיות רישוי, לא זיהוי פנים, והיא פעלה כעשור בלי בסיס חוקי עד שהוסדרה בתיקון לפקודת המשטרה ב־2024; עתירה שלישית נגד אותו הסדר קיבלה צו על תנאי ביולי 2026 והיא תלויה ועומדת. הצעת חוק למצלמות ביומטריות לזיהוי פנים היא דבר נפרד, שאושר בוועדת שרים לחקיקה בספטמבר 2023 ולא אותר כחוק מחייב.

ולבסוף המעקב עצמו, בשני מספרים מתועדים. בסקר אמריקאי מיולי 2024 בקרב 1,273 עובדים, שישים ושמונה אחוזים דיווחו על צורה כלשהי של ניטור אלקטרוני, וכשליש דיווחו שמשימות או משמרות מוקצות להם אוטומטית. בקרב מי שדיווחו על ניטור תפוקה מתמיד, ארבעים ושישה אחוזים אמרו שהם עובדים מהר יותר ממה שבריא או בטוח, לעומת חמישה עשר אחוזים בקרב מי שאינם מנוטרים. ובמרחב הציבורי, משטרת לונדון פרסמה דוח שנתי על זיהוי פנים בזמן אמת: מאתיים ושבע פריסות, כשלושה מיליון ומאה אלף פנים שנסרקו, 2,077 התראות שמתוכן עשר שגויות, ו־962 מעצרים. את המספרים האלה מסרה המשטרה על עצמה. ובית המשפט לערעורים בבריטניה כבר קבע ב־2020 שפריסה של אותה טכנולוגיה בידי משטרה אחרת הייתה בלתי חוקית — לא מפני שהטכנולוגיה אסורה, אלא מפני שלא היו הנחיות ולא נבדקה הטיה.

למה זה ברשימה

כי התקדים הקנוני בתחום הזה אינו בינה מלאכותית בכלל אלא תוכנת הנהלת חשבונות רגילה עם באגים — וזה בדיוק מה שהופך אותו לרלוונטי.

שנה	2021
חלק	הוויכוחים החוזרים
שער	מה זה עושה לעולם
נושא	מנהל ציבורי
הצדדים	עקביות שניתנת לביקורת · פער האחריות

להעמיק

המודל ההולנדי שנחשף

לקריאה

ידיעה עברית על מודל דירוג הסיכונים ברוטרדם ועל רשימת המשתנים שהתגלתה בו.

https://www.law.co.il/news/2023/07/04/netherlands-authorities-use-discriminatory-ai/

חוק זיהוי פנים ביומטרי

לקריאה

המכון הישראלי לדמוקרטיה על הצעת החוק המשטרתית ועל ההבדל בין זיהוי בזמן אמת לזיהוי בדיעבד.

https://www.idi.org.il/articles/50940

עין הנץ והעתירות

לקריאה

הציר המשפטי הישראלי המלא עם התאריכים. עמוד מתעדכן.

https://www.acri.org.il/post/__488

בעיית הפרטיות האמיתית

לקריאה

מסה עברית שטוענת שפרטיות היא תנאי לדמוקרטיה ולא זכות צרכנית.

https://alaxon.co.il/article/בעיית-הפרטיות-האמיתית/

תקציר דוח חקירת הורייזן

מקורות וארכיון

הסיכום הרשמי של ועדת החקירה הבריטית על הנזק האנושי ועל הפיצויים. באנגלית ובשפה פשוטה.

https://www.postofficehorizoninquiry.org.uk/sites/default/files/2025-07/Post_Office_Horizon_Inquiry_Volume_1_Rapid_Read.pdf

אי־צדק חסר תקדים

מקורות וארכיון

דוח ועדת החקירה ההולנדית מדצמבר 2020 במילותיה שלה כולל כלל "הכול או כלום". באנגלית.

https://www.houseofrepresentatives.nl/sites/default/files/atoms/files/verslag_pok_definitief-en-gb.docx.pdf

הדוח השנתי של משטרת לונדון

מקורות וארכיון

המספרים הרשמיים של הפריסות ההתראות והמעצרים. באנגלית והצד השני של המאזניים.

https://www.met.police.uk/SysSiteAssets/media/downloads/force-content/met/advice/lfr/other-lfr-documents/live-facial-recognition-annual-report-2025.pdf

ספרים

רעש: הפגם בכושר השיפוט האנושי

דניאל כהנמן אוליבייה סיבוני וקאס סאנסטיין · 2021

הטיעון בעד עקביות בצורתו החזקה ביותר. יצא בעברית בתרגום עפר קובר בהוצאת מטר וכנרת זמורה־ביתן.

Automating Inequality

2018 · Virginia Eubanks

הספר הקרוב ביותר לנושא היחידה: מערכות זכאות אוטומטיות ומי משלם את מחיר הטעות. לא אותר תרגום עברי.

The Black Box Society

2015 · Frank Pasquale

על מה קורה כשהכלל שקובע גורלות הוא סוד מסחרי. לא אותר תרגום עברי.

אנרגיה מים ומה זה באמת עולה

1999 · מספרים שנמדדו ומספרים שהושערו ואיך מבחינים ביניהם

מה שאין עליו מחלוקת: מרכזי נתונים צורכים חשמל ומים, הצריכה הזאת גדלה, והיא מרוכזת גיאוגרפית. המחלוקת היא על הקצב, על הסדר גודל ועל מה להשוות. ולפני הכול כדאי להבין הבחנה אחת שכל היחידה תלויה בה: יש מספרים שנמדדו, יש מספרים שהוערכו ממודל, ויש מספרים שהושלכו קדימה בתחזית. שלושתם מופיעים באותן כותרות באותו פורמט.

ההיסטוריה של האזעקה הזאת מתועדת. ב־31 במאי 1999 פרסם פורבס מאמר בשם "תכרו עוד פחם — המחשבים באים", שטען שהאינטרנט צורך שמונה אחוזים מחשמל ארצות הברית ושיצרוך חמישים אחוזים בתוך עשור. מדענים במעבדת ברקלי בדקו וקבעו מאוחר יותר שההערכה הגזימה בפי שמונה בערך. החמישים אחוזים לא הגיעו מעולם.

ואז נמדד מה שבאמת קרה, פעמיים. ב־2011 פורסם שחשמל מרכזי הנתונים בעולם גדל בכחמישים ושישה אחוזים בין 2005 ל־2010 — ולא הוכפל, כפי שקרה בחמש השנים שלפני כן — ושהחלק העולמי שלהם עומד על כאחוז וחצי. ב־2020 פורסם ב"סיינס" הממצא החזק יותר: בין 2010 ל־2018 גדל מספר יחידות החישוב הפועלות בעולם פי שישה ויותר, ואילו צריכת החשמל של מרכזי הנתונים גדלה בכשישה אחוזים בלבד. ב־2018 הם צרכו כ־205 טרה־ואט־שעה, כאחוז אחד מהחשמל העולמי. שלושה מנגנונים הסבירו את הניתוק הזה: שרתים יעילים יותר, וירטואליזציה שמאפשרת להריץ הרבה יישומים על מכונה אחת, ומעבר של רוב החישוב למרכזים ענקיים עם קירור יעיל.

והשאלה היא אם התקופה השטוחה נגמרה. התשובה, לפחות בארצות הברית, היא כן. בדוח של מעבדת ברקלי מדצמבר 2024 נמדד שמרכזי הנתונים שם צרכו 58 טרה־ואט־שעה ב־2014 ו־176 ב־2023 — ארבעה וארבע עשיריות אחוזים מחשמל המדינה — ומוערך שיצרכו בין 325 ל־580 טרה־ואט־שעה ב־2028, כלומר בין שישה וכמעט שבעה אחוזים לשנים עשר. שימו לב מה נמדד ומה הושלך קדימה.

ברמה העולמית, הסוכנות הבינלאומית לאנרגיה אמדה ב־2025 את צריכת מרכזי הנתונים בכ־415 טרה־ואט־שעה ב־2024, כאחוז וחצי מהחשמל העולמי, עם גידול של כשנים עשר אחוזים בשנה בחמש השנים האחרונות. בתרחיש הבסיס שלה הצריכה מגיעה לכ־945 טרה־ואט־שעה ב־2030, מעט פחות משלושה אחוזים. שרתים ייעודיים לבינה מלאכותית צומחים שם בשלושים אחוזים בשנה ואחראים לכמחצית מכל התוספת. ולסוכנות עצמה יש הסתייגות שכדאי לצטט: "קיימת אי־ודאות ניכרת הן לגבי צריכת מרכזי הנתונים כיום והן לגבי העתיד".

וכעת שני הצדדים.

הצד שאומר שזו עוד אזעקת שווא מבקש להשוות. מרכזי נתונים הם כאחוז מהביקוש העולמי לחשמל וכחצי אחוז מפליטות הפחמן. בתוספת הביקוש לחשמל עד 2030, מרכזי הנתונים מוסיפים כ־530 טרה־ואט־שעה — פחות מרכבים חשמליים שמוסיפים כ־838, פחות ממיזוג אוויר שמוסיף כ־651, והרבה פחות מהתעשייה שמוסיפה כמעט אלפיים. הצד הזה מזכיר גם שהיעילות משתפרת בקצב שקשה לחזות, ושכל מודל שמכפיל את הצריכה של היום בקצב הצמיחה של היום עשה בדיוק את הטעות של 1999.

הצד שאומר שהפעם זה אחרת מבקש להסתכל מקומית, כי חשמל אינו מיוצר בממוצע עולמי. החלוקה בפועל היא כמחצית בארצות הברית, כרבע בסין וכחמישית באירופה. באירלנד נמדד — לא הוערך — שמרכזי נתונים צרכו עשרים ושניים אחוזים מהחשמל הנמדד במדינה ב־2024, לעומת שמונה עשר אחוזים לכל הדירות בערים. באזור אחד בווירג׳יניה מדובר בכרבע מהחשמל. ובישראל, לפי דיווח מספטמבר 2026, הגיעו לרשות החשמל כמאה בקשות חיבור בהספק מצטבר של כעשרים ושבעה ג״יגה־ואט, בעוד שביקוש השיא של המשק כולו עומד על שבעה עשר עד שמונה עשר ג״יגה־ואט. בארצות הברית כבר פועלות יותר ממאה ושמונים קבוצות מקומיות נגד פרויקטים בכארבעים מדינות, ומספר הפרויקטים שבוטלו עלה משניים ב־2023 לשישה ב־2024 ולעשרים וחמישה ב־2025.

על מספרי האימון כדאי לדעת סיפור מתקן. ב־2019 פורסמה עבודה שאמדה את הפליטות של אימון מודל אחד בשיטה מסוימת בכ־626 אלף ליברות של שווה־ערך פחמן דו־חמצני, והשוותה זאת למחזור החיים של מכונית — כ־126 אלף ליברות. המספר הזה עשה את סיבוב העולם. ב־2022 פורסמה בדיקה חוזרת שמצאה שהוא גבוה פי שמונים ושמונה מהערך בפועל: פי חמישה בגלל הנחות על חומרה ישנה ועל מרכז נתונים ממוצע, ופי שמונה עשרה וחצי בגלל אי־הבנה של שיטת החיפוש שנמדדה. אותה בדיקה הוסיפה נתון מבני: אצל חברה אחת גדולה, כשלוש חמישיות מאנרגיית הלמידה החישובית הולכות לשימוש ורק כשתי חמישיות לאימון. כלומר אימון הוא אירוע חד־פעמי והשימוש היומיומי הוא הרוב.

ולמים יש בעיה משלהם, וגם היא בעיקר בעיית הגדרות. יש הבדל בין ״שאיבה״ — כמה מים נלקחו ממקור — ובין ״צריכה״, שהיא מה שהתאדה ולא חזר. עבודה מאוניברסיטת קליפורניה בריברסייד העריכה שאימון של מודל גדול אחד במרכזי נתונים בארצות הברית ״יכול לצרוך 5.4 מיליון ליטר מים, מתוכם 700 אלף ליטר צריכה ישירה באתר״, ושמערכת כזאת ״צריכה לשתות בקבוק של חצי ליטר עבור עשר עד חמישים תשובות באורך בינוני, תלוי מתי ואיפה״. המחברים כותבים במפורש שההערכה שלהם ״נוטה לצד השמרני״. מהצד השני, ב־2025 פרסמה חברה גדולה מדידה משלה: שאילתת טקסט חציונית צורכת לדבריה 0.24 ואט־שעה, שלוש מאיות גרם פחמן דו־חמצני ורבע מיליליטר מים. הביקורת על המדידה הזאת ממוקדת: היא סופרת רק מים לקירור ולא את המים הדרושים לייצור החשמל, היא משתמשת בשיטת חישוב פחמן שמשקפת רכישות ולא את תמהיל הרשת בפועל, ו״חציון״ מסתיר בהגדרה את הזנב הכבד.

ורעיון אחד ישן שראוי להיות על השולחן. ב־1865 פרסם הכלכלן ויליאם סטנלי ג׳בונס ספר על הפחם באנגליה, ובו התצפית שצריכת הפחם דווקא זינקה אחרי שמנוע הקיטור נעשה יעיל בהרבה. הניסוח שלו: ״זהו בלבול מושגים לחשוב ששימוש חסכוני בדלק שקול לצריכה מופחתת״. הרעיון הזה הוא בדיוק מה שמפריד בין שני הצדדים כאן, והוא גם לא מוכרע: כלכלנים חלוקים באיזו תדירות האפקט הזה באמת מבטל את החיסכון ובאיזו הוא רק מכרסם בו.

למה זה ברשימה

כי זה הוויכוח שבו הכי קל לצטט מספר נכון ולהסיק ממנו דבר שגוי — ומפני שאותה אזעקה בדיוק כבר נשמעה פעם אחת ונבדקה.

שנה	1999
חלק	הוויכוחים החוזרים
שער	מה זה עושה לעולם
נושא	משאבים וסביבה
הצדדים	עוד אזעקת שווא · הפעם זה אחרת

להעמיק

הסוכנות הבינלאומית לאנרגיה על הביקוש

לקריאה

האומדן העולמי לשנת 2024 והתרחישים עד 2035 כולל ההסתייגות על אי־הוודאות. באנגלית.

https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

חמישה גרפים שמציבים את המספרים בהקשר

לקריאה

השוואה בין תוספת הביקוש ממרכזי נתונים ובין רכבים חשמליים מיזוג אוויר ותעשייה. באנגלית.

https://www.carbonbrief.org/ai-five-charts-that-put-data-centre-energy-use-and-emissions-into-context

הטענות הפרועות שלא מתות

לקריאה

החוקר שהפריך את תחזית 1999 מסביר איך נבנות הערכות מנופחות ולמה הן חוזרות. באנגלית.

https://www.koomey.com/koomey_blog/wild-claims-about-electricity-used-by-computers-that-just-wont-die--but-should-/

התנגדות מקומית ומה שמגיע לרשות החשמל

לקריאה

כתבה עברית עם המספרים על בקשות החיבור בישראל ועל ביטולי פרויקטים בארצות הברית.

https://www.bizportal.co.il/energy/news/article/20042341

המדידה האירית

מקורות וארכיון

נתון מדוד ולא מוערך: עשרים ושניים אחוזים מהחשמל הנמדד במדינה לעומת שמונה עשר לכל הדירות בערים. באנגלית.

https://www.cso.ie/en/releasesandpublications/ep/p-dcmec/datacentresmeteredelectricityconsumption2024/

התיקון של פי שמונים ושמונה

מקורות וארכיון

הבדיקה החוזרת של אומדן פליטות האימון המפורסם ופירוק הטעות לגורמיה. באנגלית.

https://arxiv.org/pdf/2204.05149

להפוך את הבינה המלאכותית לפחות צמאה

מקורות וארכיון

העבודה שהציגה את מספרי המים כולל ההבחנה בין שאיבה לצריכה והסתייגויות המחברים. באנגלית.

https://arxiv.org/html/2304.03271v5

מרכז המחקר והמידע של הכנסת

מקורות וארכיון

סקירה עברית מנובמבר 2024 על צריכת החשמל של מרכזי נתונים ועל המצב באירלנד ובארצות הברית.

https://fs.knesset.gov.il/globaldocs/MMM/184dd353-1b7b-ef11-8160-005056aac6c3/2_184dd353-1b7b-ef11-8160-005056aac6c3_11_20691.pdf

ספרים

The Coal Question

1865 · William Stanley Jevons

המקור לרעיון שיעילות מגדילה צריכה במקום להקטין אותה. לא אותר תרגום עברי.

Atlas of AI

2021 · Kate Crawford

הפרק על המחיר הפלנטרי עוקב אחרי חומרים ואנרגיה ולא רק אחרי חשמל. לא אותר תרגום עברי.

כמה זה מסוכן

2023 · שתי רשימות סכנות שכמעט אינן חופפות

בתוך שבעה חודשים ב־2023 פורסמו שלושה מסמכים שמסמנים את גבולות הוויכוח. במרץ פורסם מכתב פתוח שקרא "להשהות מיד, למשך שישה חודשים לפחות, את אימונן של מערכות בינה מלאכותית חזקות יותר מהקיימות כיום"; הוא אסף למעלה משלושים אלף חותמים. בסוף מאי פורסמה הצהרה בת משפט אחד: "הפחתת הסיכון להכחדה מבינה מלאכותית צריכה להיות בראש סדר העדיפויות העולמי, לצד סיכונים אחרים בקנה מידה חברתי כמו מגפות ומלחמה גרעינית". על ההצהרה הזאת חתמו ג'פרי הינטון ויושוע בנג'ו, שניים משלושת חוקרי הלמידה העמוקה המוכרים ביותר — וגם המנכ"לים של שלוש החברות הגדולות שבונות את המערכות האלה. ובנובמבר חתמו עשרים ותשע מדינות, ובהן ישראל וסין, על הצהרה משותפת שקובעת ש"קיים פוטנציאל לנזק חמור ואף קטסטרופלי, מכוון או לא מכוון".

מה ששני המחנות מסכימים עליו: יש נזקים מתועדים היום, ואיש אינו יודע בוודאות מה יקרה בעוד עשור. המחלוקת היא לאן מפנים תשומת לב ורגולציה, ובאיזו מידה דיבור על האחד גוזל מהאחר.

הצד שמצביע על הסיכון הקטסטרופלי אינו נשען על תחזית אלא על טיעון מבני. הטענה, שנוסחה ב־2008 ונוסחה שוב ב־2014, נקראת התכנסות אינסטרומנטלית: עבור כמעט כל מטרה סופית שמערכת עשויה לקדם, יש כמה יעדי ביניים שמועילים לה — לשמור על עצמה, לשמור על המטרה שלה מפני שינוי, להשתפר, להשיג משאבים, ולמנוע הפרעה. המסקנה היא שמערכת חזקה מספיק עלולה לפתח התנהגויות כאלה בלי שאיש תכנן אותן ובלי שום עוינות. התרגיל המחשבתי המוכר של מערכת שמצוינת בייצור אטבי נייר ומגייסת לשם כך את העולם נכתב ב־2003 והוא לא נועד להיות תחזית אלא הדגמה של הפער בין מטרה מוגדרת ובין מה שהיא גוררת. הניסוח המתון של העמדה הזאת, מ־2024: "העמדה הרציונלית היחידה שמתיישבת עם כל הראיות היא ענווה ותכנון תוך אי־ודאות", ולכן "ראוי לשאוף לראיות חזקות מאוד לפני שמסיקים שאין ממה לדאוג". שימו לב שזה טיעון מאי־ודאות ולא מביטחון.

הצד שמצביע על נזק מתועד עכשיו משיב ש"הסיכונים והנזקים מעולם לא היו קשורים לבינה מלאכותית חזקה מדי", ומונה ארבעה תחומים: ריכוז כוח בידי מעטים, שכפול של יחסי דיכוי קיימים, הרעלת מערכת המידע הציבורית בתעמולה ובשקר, ופגיעה סביבתית. הטענה החדה יותר של הצד הזה היא על כלכלת תשומת הלב: דיון פרלמנטרי שמוקדש להכחדה עתידית הוא דיון שלא מוקדש לקצבאות שנשללו, למעצרים על סמך זיהוי שגוי ולתנאי העבודה של מי שמסמן נתונים. ויש בה גם עוקץ: מי שמזהיר בקול מפני הסכנה העתידית של המוצר שהוא מוכר, מציג אותו בכך כחזק מאוד.

ויש נקודה אחת שבה אפשר לצאת מהוויכוח ולמדוד. הנתיב הקטסטרופלי שמוזכר יותר מכולם הוא סיוע ביצירת נשק ביולוגי. בינואר 2024 פורסם מחקר של גוף מחקר אמריקאי שעשה בדיוק את המבחן: צוותים

שתכננו תרחיש תקיפה ביולוגית, חלקם בעזרת מודלי שפה וחלקם בלעדיהם, וההערכה נעשתה בעיוורון. המסקנה: "אין הבדל מובהק סטטיסטית בכדאיות התוכניות שנוצרו בסיוע מודלי השפה של הדור הנוכחי או בלעדיו". זה מחקר אחד, על דור מסוים, ועל משימה אחת — והוא בדיוק סוג הראיה שצריכה להצטבר כדי שאפשר יהיה להתווכח על עובדות ולא על תחושות.

לשני המחנות יש גם תקדים היסטורי אהוב, ובמקרה זה אותו תקדים. בפברואר 1975 התכנסו כמאה וארבעים ביולוגים משפטנים ורופאים בכנס באסילומר שבקליפורניה, אחרי שהחוקרים עצמם קראו שנה קודם להפסקה זמנית של ניסויים בהנדסה גנטית. הכנס קבע עקרון אחד פשוט — שרמת ההכלה תתאים לרמת הסיכון המשוערת — וארבע רמות הכלה, ואסר סוגי ניסויים מסוימים. ההנחיות רוככו בהדרגה בשנים שאחר כך, והאסון שממנו חששו לא הגיע. המחנה האחד קורא את זה כהוכחה שרגולציה מוקדמת בידי החוקרים עצמם עובדת. המחנה השני קורא את זה כהוכחה שהחשש היה מוגזם מלכתחילה. אף אחד מהשניים לא יכול להוכיח את גרסתו, וזה בדיוק הקושי בכל דיון על סיכון שנמנע.

> למה זה ברשימה
>
> כי שני המחנות כאן אינם חלוקים על עובדה אחת מרכזית — הם פשוט מסתכלים על שתי רשימות שונות ומאשימים זה את זה בהסחה.

שנה	2023
חלק	הוויכוחים החוזרים
שער	כמה זה מסוכן
נושא	סיכון ובטיחות
הצדדים	נזק מתועד עכשיו · סיכון קטסטרופלי

להעמיק

טיעון מתוך אי־ודאות

לקריאה

חוקר בכיר עובר אחד אחד על הטיעונים נגד דאגה ומסביר למה הוא לא משוכנע. באנגלית.

https://yoshuabengio.org/2024/07/09/reasoning-through-arguments-against-taking-ai-safety-seriously/

הטיעון שזו הסחה

לקריאה

העמדה שלפיה הדיון בסיכון עתידי מרחיק את המחוקקים מנזקים שקורים כבר עכשיו. באנגלית.

https://medium.com/@emilymenonbender/policy-makers-please-dont-fall-for-the-distractions-of-aihype-e03fa80ddbf1

מפת הסכנות

לקריאה

סקירה עברית שמתמקדת דווקא בנזקים השקטים למחקר ולידע ולא בתרחישי קצה.

https://davidson.org.il/read-experience/sciencenews/מפת-הסכנות-של-הבינה-המלאכותית/

אינטליגנציה גרעינית

לקריאה

מסה מתורגמת לעברית על הרעיון להכניס מערכות אוטומטיות לפיקוד ושליטה גרעיניים.

https://alaxon.co.il/article/אינטליגנציה-גרעינית/

ההצהרה בת המשפט האחד

מקורות וארכיון

הטקסט המלא ורשימת החותמים כולל ראשי החברות שבונות את המערכות. באנגלית.

https://safe.ai/work/statement-on-ai-risk

הקריאה להשהיה של שישה חודשים

מקורות וארכיון

המכתב מ־2023 במלואו עם הדרישה המדויקת ומספר החותמים. באנגלית.

https://futureoflife.org/open-letter/pause-giant-ai-experiments/

הצהרת בלצ׳לי

מקורות וארכיון

ההצהרה שעליה חתמו עשרים ותשע מדינות ובהן ישראל וסין. באנגלית.

https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

המבחן על הנתיב הביולוגי

מקורות וארכיון

מחקר צוות אדום שבדק אם מודלי שפה משפרים תכנון תקיפה ביולוגית ומה נמצא. באנגלית.

https://www.rand.org/pubs/research_reports/RRA2977-2.html

ספרים

Superintelligence

2014 · Nick Bostrom

הניסוח השיטתי של הטיעון על התכנסות אינסטרומנטלית ועל בעיית השליטה. לא אותר תרגום עברי.

Human Compatible

2019 · Stuart Russell

מחבר ספר הלימוד המרכזי בתחום מציע ניסוח מחדש של המטרה שמערכת אמורה לקדם. לא אותר תרגום עברי.

Atlas of AI

2021 · Kate Crawford

העמדה שהנזק נמצא בתשתיות בעבודה ובמשאבים ולא בתרחיש עתידי. לא אותר תרגום עברי.

פתוח מול סגור

1995 · ויכוח שכבר נערך פעם על הצפנה והוכרע בבית משפט

מה ששני הצדדים מסכימים עליו: פרסום פומבי של משקלי מודל הוא מעשה בלתי הפיך. אחרי שקובץ כזה יצא לאוויר אי אפשר להחזיר אותו, אי אפשר לעדכן אותו אצל מי שהוריד אותו, ומנגנוני הסינון שנבנו לתוכו ניתנים להסרה בידי מי שיודע מה הוא עושה. מכאן ואילך חלוקים.

ולפני הוויכוח, התקדים. בשנות התשעים הייתה הצפנה חזקה רשומה ברשימת אמצעי הלחימה של ארצות הברית, בקטגוריה אחת עם נשק. ייצוא הצפנה מעבר לארבעים סיביות חייב רישיון, ולכן דפדפנים שווקו בשתי גרסאות — אחת אמריקאית חזקה ואחת בינלאומית חלשה שאפשר היה לשבור בתוך ימים במחשב אחד. ב־1991 פורסמה ברשת תוכנת הצפנה לשימוש אישי, והיא הייתה "האתגר המשמעותי הראשון ברמת הפרט לפיקוח על ייצוא קריפטוגרפיה". הממשל הציע פתרון ביניים: שבב שמאפשר הצפנה חזקה אך שומר עותק של המפתח אצל הרשויות. ב־1997 פרסמו אחד עשר קריפטוגרפים בכירים דוח משותף שקבע ש"מערכות שחזור מפתחות הן מטבען פחות מאובטחות, יקרות יותר וקשות יותר לשימוש ממערכות מקבילות בלי יכולת שחזור".

ואז הגיע בית המשפט. תביעה שהוגשה בפברואר 1995 בשם מתמטיקאי שרצה לפרסם קוד הצפנה שכתב הגיעה ב־1997 לקביעה שתקנות הייצוא אינן חוקתיות, ובמאי 1999 אישר בית המשפט לערעורים את ההחלטה וקבע ש"קוד מקור של תוכנה הוא דיבור המוגן בתיקון הראשון". בינואר 2000 שינה הממשל את התקנות. התוצאה היא העולם שאנחנו חיים בו: הצפנה חזקה בכל טלפון ובכל אתר. הקטסטרופה שהובטחה לא הגיעה בצורה שבה הובטחה, והקושי שהצפנה יוצרת לגופי אכיפה הוא אמיתי ומתועד. שני המחנות מצטטים את הפרשה הזאת, וכל אחד מהם צודק בחלק שלו.

הצד שתומך בפרסום טוען ארבעה דברים. ראשית, ביקורת: אי אפשר לבדוק מערכת שאי אפשר לפתוח. כל מה שנכתב ביחידה על אחריות — הדירוג בעיריית רוטרדם שהתברר כ"מעט יותר טוב מבחירה אקראית" — התאפשר רק מפני שמישהו השיג את המודל. שנית, תחרות: בלי משקלים זמינים, מי שרוצה לבנות משהו תלוי בשוער. שלישית, מחקר: כל עבודת הפירוש הפנימי שהוזכרה ביחידה על שכבות וייצוגים נעשית על מודלים שאפשר לפתוח. ורביעית, הלקח מההצפנה: פיקוח על פרסום מתמטיקה נכשל, ובינתיים החליש את האבטחה של כולם.

הצד שתומך בשליטה מקבל את שלושת הראשונים ומתעקש על הרביעי. ההבדל, לטענתו, אינו בעיקרון אלא באסימטריה: קוד הצפנה מגן על מי שמשתמש בו, ואילו מודל שהוסרו ממנו מנגנוני הסינון משמש את מי שמפעיל אותו נגד אחרים. ומעל הכול — אי־ההפיכות. אפשר לתקן חוק, אפשר לסגור שירות, אי אפשר לבטל הורדה. ולכן, לטענתו, ההחלטה צריכה להתקבל לפי היכולת הגרועה ביותר שהמודל עשוי לגלם, ולא לפי מה שנמדד בו ביום הפרסום.

המסגרת שהכי עוזרת להתקדם בוויכוח הזה הוצעה ב־2024 ונקראת "סיכון שולי": לא לשאול כמה מסוכן מודל פתוח, אלא כמה סיכון הוא **מוסיף** מעבר למה שכבר קיים — מנועי חיפוש, ספרות מקצועית, מודלים סגורים, ומומחיות אנושית. המסגרת הזאת מסבירה גם למה שני המחנות הגיעו למסקנות הפוכות: הם מדדו דברים שונים. גוף ממשלתי אמריקאי אימץ אותה ביולי 2024 בדוח על מודלים עם משקלים זמינים, והמלצתו הייתה לא להגביל אלא לנטר: "הראיות הקיימות אינן מספיקות כדי לקבוע באופן חד־משמעי לא שהגבלות על מודלים כאלה מוצדקות ולא שהגבלות לעולם לא יהיו מתאימות בעתיד". שלושה שלבים: לאסוף ראיות, להעריך אותן, ולפעול לפיהן.

ונותרה שאלה שקודמת לשתיהן: מה בכלל נחשב פתוח. באוקטובר 2024 פורסמה הגדרה רשמית שדורשת ארבע חירויות — להשתמש, לחקור, לשנות ולשתף — ושלושה רכיבים בפועל: מידע על הנתונים "מפורט דיו כדי שאדם מיומן יוכל לבנות מערכת שוות ערך במהותה", קוד מקור מלא לאימון ולהרצה, והמשקלים עצמם. כמעט שום מודל שמכונה היום פתוח אינו עומד בזה; מה שמשוחרר הם המשקלים בלבד, והנתונים והקוד נשארים סגורים. עבודה מ־2023 ניסחה את זה חריף יותר: גם המערכות ה"פתוחות" ביותר "אינן מבטיחות גישה דמוקרטית או תחרות משמעותית, ופתיחות לבדה אינה פותרת את בעיית הפיקוח והביקורת", ולעיתים המילה מתפקדת "יותר כשיווק מאשר כתיאור טכני".

למה זה ברשימה

כי אותו ויכוח בדיוק התנהל בשנות התשעים על הצפנה — עם אותם טיעונים ואותן אזהרות — וההכרעה שם מתועדת ומלמדת.

שנה	1995
חלק	הוויכוחים החוזרים
שער	כמה זה מסוכן
נושא	פרסום ורגולציה
הצדדים	פרסום · שליטה

להעמיק

התיק שהכריע שקוד הוא דיבור

לקריאה

ציר הזמן המלא של התביעה מ־1995 ועד שינוי התקנות בשנת 2000. באנגלית.

https://www.eff.org/cases/bernstein-v-us-dept-justice

מלחמות ההצפנה

לקריאה

סקירה מתוארכת של העשור שבו הצפנה נחשבה אמצעי לחימה ומה שקרה בסופו. באנגלית.

https://en.wikipedia.org/wiki/Crypto_Wars

הדוח הממשלתי על משקלים זמינים

לקריאה

ההמלצה לנטר ולא להגביל והנימוק המדויק לכך. באנגלית.

https://www.ntia.gov/issues/artificial-intelligence/open-model-weights-report

מה נחשב פתוח בכלל

לקריאה

ההגדרה הרשמית עם ארבע החירויות ושלושת הרכיבים הנדרשים. באנגלית.

https://opensource.org/ai/open-source-ai-definition

הסיכונים שבשחזור מפתחות

מקורות וארכיון

הדוח של אחד עשר קריפטוגרפים נגד הצעת שבב ההצפנה עם מפתח שמור. באנגלית.

https://www.schneier.com/academic/archives/1997/04/the_risks_of_key_rec.html

מסגרת הסיכון השולי

מקורות וארכיון

העבודה שהציעה למדוד כמה סיכון מודל פתוח מוסיף מעבר לקיים ולא כמה הוא מסוכן בפני עצמו. באנגלית.

https://arxiv.org/abs/2403.07918

מה פתיחות לא מספקת

מקורות וארכיון

הטענה שפתיחות אינה מבטיחה תחרות או פיקוח ולעיתים משמשת כשיווק. באנגלית.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807

ספרים

Applied Cryptography

1996 · Bruce Schneier

הספר שהיה בעצמו מקרה מבחן בוויכוח על ייצוא ידע מתמטי. לא אותר תרגום עברי.

Atlas of AI

2021 · Kate Crawford

על ריכוז הכוח בתשתיות ועל השאלה למי בעצם יש גישה. לא אותר תרגום עברי.

איפה זה עומד — נכון לספטמבר 2026

ספטמבר 2026 · היחידה היחידה בפקולטה שנכתבה כדי להתיישן

היחידה הזאת שונה מכל השאר, ונכתבה בכוונה כך. כל יחידה אחרת בפקולטה בנויה מהשכבה שלא מתיישנת. זו בנויה ממספרים, ותאריך הכתיבה שלה הוא ספטמבר 2026. אם את קוראת אותה יותר משנה אחרי, התייחסי לכל מספר כאן כאל תיעוד היסטורי ולא כאל מצב הדברים. ההבחנה שכן תשרוד היא ההבחנה שהיחידה בנויה סביבה: מה נמדד בידי גורם בלתי תלוי, ומה נמסר בידי מי שמוכר את המוצר.

מה נמדד ביכולת. מבחנים קשים נשברים מהר. מבחן שנועד להיות קשה במיוחד לשאלות אקדמיות נתן למודלים המובילים 8.8 אחוזים ב־2025, ומעל חמישים אחוזים באפריל 2026. במקביל, ומועיל יותר, יש מבחנים שמראים איפה הכישלון עקבי. מבחן אחד בודק דבר אחד: לקרוא שעון אנלוגי מתמונה. בני אדם מגיעים ל־90.7 אחוזים. המודלים הטובים ביותר, נכון לספטמבר 2026, נעים סביב שישים וחמישה אחוזים. מי שבנה את המבחן מסביר שהמשימה דורשת הסקה בתוך המרחב החזותי ולא בטקסט, ושאין עדיין תשובה לשאלה אם הגדלה פשוטה של המודלים פותרת את זה או שצריך גישה אחרת. כישלון עקבי שכזה מתיישן לאט יותר מהצלחה, והוא מלמד יותר.

יש גם מדידה מעניינת שמנסה לתרגם יכולת לזמן: כמה ארוכה משימה שמודל מסיים בהצלחה בסבירות סבירה. מי שמפרסם אותה מוסיף שלוש הסתייגויות שכדאי לצטט: המדד מודד את **קושי** המשימה ולא את הזמן שלוקח למודל; המשימות במאגר הן בעיקר הנדסת תוכנה למידת מכונה ואבטחת מידע; ו"רוב העבודות אינן בנויות ממשימות מוגדרות היטב ואלגוריתמיות".

ודוגמה מספטמבר 2026 שממחישה בדיוק את ההבחנה שהיחידה נשענת עליה. פורסמה טענה שמודל תרם לפתרון של אחת הבעיות הפתוחות המפורסמות במתמטיקה, בשמונים ושמונה שעות ובעזרת עשרת אלפים סוכנים במקביל. מתמטיקאי בכיר הזהיר בתגובה מפני קבלת תוצאות בלי שקיפות מלאה לגבי הדרך שבה הושגו, והחברה עצמה הודתה שאינה יכולה לשלול שנתונים שהגיעו משימוש במוצר שלה סייעו למודלים שלה. המקור היחיד לטענה במלואה הוא מי שיצר אותה. זה לא אומר שהיא שגויה. זה אומר שהיא לא אומתה.

מה נמדד באימוץ, וכאן טמונה מלכודת. בארצות הברית, סקר של לשכת המפקד שמריץ כמיליון ומאתיים אלף עסקים מצא שכעשרים אחוזים מהעסקים משתמשים בטכנולוגיה נכון למאי 2026 — שלושים ושבעה אחוזים בקרב עסקים עם מאתיים חמישים עובדים ומעלה, ופחות מעשרים אחוזים בקרב הקטנים. סקר אחר, ברמת העובד, מצא שארבעים ואחד אחוזים מכוח העבודה משתמשים בכלים כאלה בעבודה, ושנים עשר אחוזים מדי יום. ושלישי מצא ששבעים ושמונה אחוזים מכוח העבודה מועסקים בחברה שמשתמשת. שלושת המספרים נכונים והם שונים פי ארבעה, מפני שהם סופרים יחידות שונות — עסק, עובד, ועובד של עסק. מי

שפרסם את ההשוואה כותב במפורש שההבדל הגדול ביותר בין האומדנים נובע מהבדלי דגימה ומיחידת הניתוח.

בציבור הרחב, סקר אמריקאי מצא שאחוז המבוגרים שמשתמשים בכלים כאלה בשישה ימים בשבוע לפחות עלה משמונה אחוזים במרץ 2026 לתשעה עשר אחוזים באוגוסט 2026 — כשהעורכים עצמם מסייגים שנוסח השאלה השתנה בין הגלים.

ובישראל יש נתון רשמי. סקר של הלשכה המרכזית לסטטיסטיקה, שפורסם ביוני 2026 ונשען על יותר משמונה מאות עסקים, מצא שתשעה ושלושים אחוזים מהעסקים בישראל השתמשו בטכנולוגיה במרץ 2026, לעומת שמונה ועשרים אחוזים ביוני 2025. בתעשייה המסורתית חמישים ושבעה אחוזים דיווחו על שימוש לשיווק וחמישים ושניים לתהליכים מנהליים; בשירותים שאינם טכנולוגיים שישים ותשעה אחוזים לתפקודים עסקיים ומנהליים; ובטכנולוגיה ובפיננסים שישים ואחד אחוזים למחקר ופיתוח. הפרט המעניין ביותר בסקר: ארבעים אחוזים מהעסקים שמשתמשים היום אמרו בסקר הקודם שאינם מצפים לאמץ את זה.

הכסף. ההשקעה העולמית בתחום ב־2025 נאמדה ב־581 מיליארד דולר, יותר מכפול מ־253 מיליארד ב־2024 ומעל השיא הקודם מ־2021. לשנת 2026 נאמדו תחזיות הוצאה של חמש חברות הענן הגדולות בטווח של שבע מאות עד תשע מאות מיליארד דולר, שמתוכן כשלוש רבעים לתשתית ייעודית. בצד ההכנסות, הצמיחה אמיתית ומהירה: ההכנסות המשולבות של שתי החברות המובילות בתחום עלו פי שלושה וחצי בשמונה חודשים, משלושים מיליארד דולר בקצב שנתי לכמאה וחמישה. ובכל זאת, אנליסט מוכר אמד את הפער השנתי בין ההוצאה על תשתית ובין ההכנסות של כל המערכת האקולוגית בכשש מאות מיליארד דולר. סקר של חברת ייעוץ מנובמבר 2025 מצא שמתוך חמש מאות מנהלי מערכות מידע, שנים ושישים אחוזים האטו או עצרו הוצאות בגלל תשואה לא ברורה, ואנליסט בבנק השקעות הזהיר בינואר 2026 מפני "נכסים תקועים" אם האימוץ יאט. מהצד השני נטען שבניית תשתית תמיד מקדימה את האימוץ בשנה וחצי עד שנתיים. שני הצדדים מחזיקים מספרים אמיתיים; מה שבמחלוקת הוא המכנה ומסגרת הזמן.

הרגולציה, נכון לספטמבר 2026. חוק הבינה המלאכותית האירופי נכנס לתוקף באוגוסט 2024. האיסורים וחובת האוריינות חלים מפברואר 2025, החובות על מודלים כלליים מאוגוסט 2025, ויתר הוראות החוק מאוגוסט 2026. אבל בהסכמה פוליטית ממאי 2026 נדחו החובות על מערכות בסיכון גבוה: אלה שברשימה העצמאית לדצמבר 2027, ואלה שמשובצות במוצרים מוסדרים לאוגוסט 2028. חובות שקיפות על תוכן מסונתז וכן איסורים על ייצור תוכן מיני לא־מסכים חלים מדצמבר 2026. כלומר החלק שהיה אמור לכסות דירוג אשראי, זכאות לקצבאות, מיון מועמדים וזיהוי ביומטרי — טרם נכנס לתוקף.

ואיפה לבדוק מה השתנה. שלושה מקורות מתעדכנים ומפרסמים את השיטה שלהם: דוח שנתי מאוניברסיטאי שמרכז נתונים על התחום, גוף מחקר עצמאי שמפרסם מגמות חישוב ומחירים עם תאריך עדכון לכל גרף, ודוח בטיחות בינלאומי שפורסם במהדורת 2026 בפברואר, בעריכת למעלה ממאה חוקרים ובגיבוי של יותר משלושים מדינות. ההערה שנשארת נכונה מכולם היא זו שמופיעה באותו דוח בטיחות: מספר החברות שמפרסמות מסגרות בטיחות יותר מהוכפל, ובמקביל "תוקפים מתוחכמים מצליחים לעיתים קרובות לעקוף את ההגנות הקיימות".

למה זה ברשימה

כי כל שאר הפקולטה נכתבה כדי להחזיק שנים ולכן היא מדלגת על ההווה — וההווה בכל זאת נחוץ כדי לדעת מול מה מודדים.

שנה	2026
חלק	הוויכוחים החוזרים
שער	ההווה
נושא	מצב הידע

להעמיק

דוח המדד השנתי לשנת 2026

לקריאה

ריכוז נתונים על ביצועים אימוץ השקעות ורגולציה עם מקור לכל מספר. באנגלית ומתעדכן שנתית.

https://hai.stanford.edu/ai-index/2026-ai-index-report

מגמות חישוב עלויות ומחירים

לקריאה

לוח מחוונים עם תאריך עדכון לכל גרף. הדרך הטובה ביותר לבדוק אם מה שכתוב כאן עדיין נכון. באנגלית.

https://epoch.ai/trends

האימוץ בעסקים אמריקאיים

לקריאה

הנתון הרשמי מהסקר הדו־שבועי כולל פילוח לפי גודל עסק וענף. באנגלית.

https://www.census.gov/library/stories/2026/05/ai-use-businesses.html

למה שלושה סקרים נותנים שלוש תשובות

לקריאה

ההשוואה שמסבירה איך אותה שאלה נותנת שמונה עשר או ארבעים ואחד או שבעים ושמונה אחוזים. באנגלית.

https://www.federalreserve.gov/econres/notes/feds-notes/monitoring-ai-adoption-in-the-u-s-economy-20260403.html

סקר הלמ״ס על עסקים בישראל

לקריאה

הנתון הישראלי מיוני 2026 עם הפילוח לפי ענפים.

https://www.mako.co.il/news-money/2026_q2/Article-49c450f9ed78e91026.htm

המבחן שבודק קריאת שעון

לקריאה

דוגמה נקייה למשימה טריוויאלית לאדם שהמודלים עדיין נכשלים בה בעקביות. באנגלית.

https://clockbench.ai/

אורך המשימה שמודל מסיים

לקריאה

מדידה שמנסה לתרגם יכולת לזמן ומפרטת בעצמה מה היא לא מודדת. באנגלית.

https://metr.org/time-horizons/

לוח הזמנים של החוק האירופי

לקריאה

מה חל היום ומה נדחה ולאיזה תאריך. באנגלית ומתעדכן.

https://artificialintelligenceact.eu/implementation-timeline/

האם נפתרה בעיה מתמטית מפורסמת

לקריאה

סקירה עברית של טענה מספטמבר 2026 ושל הסתייגויות המתמטיקאים ממנה.

https://davidson.org.il/read-experience/sciencenews/navier-stokes-men-vs-machine/

ספרים

ההיסטוריה הקצרה של ה-AI

טובי וולש · 2023

שימושי דווקא כאן כדי לראות כמה פעמים כבר נכתבה יחידה כזאת והתיישנה. יצא בעברית בהוצאת תכלת.

נקסוס

יובל נח הררי · 2024

הניסיון לקרוא את הרגע הזה בתוך היסטוריה ארוכה של רשתות מידע. נכתב בעברית.

קולופון

בינה מלאכותית — חוברת קריאה מתוך האוניברסיטה, ספריית הלמידה האישית של מושקא.

20 יחידות · 135 קישורים מאומתים · 50 ספרים. הגרסה המלאה, עם חיפוש וכרטיסיות חזרה, נמצאת ב-uni.zikeet.co.il.

נבנתה מאותם קובצי תוכן שמהם נבנה האתר, כך ששתי הגרסאות אינן יכולות להיפרד זו מזו.