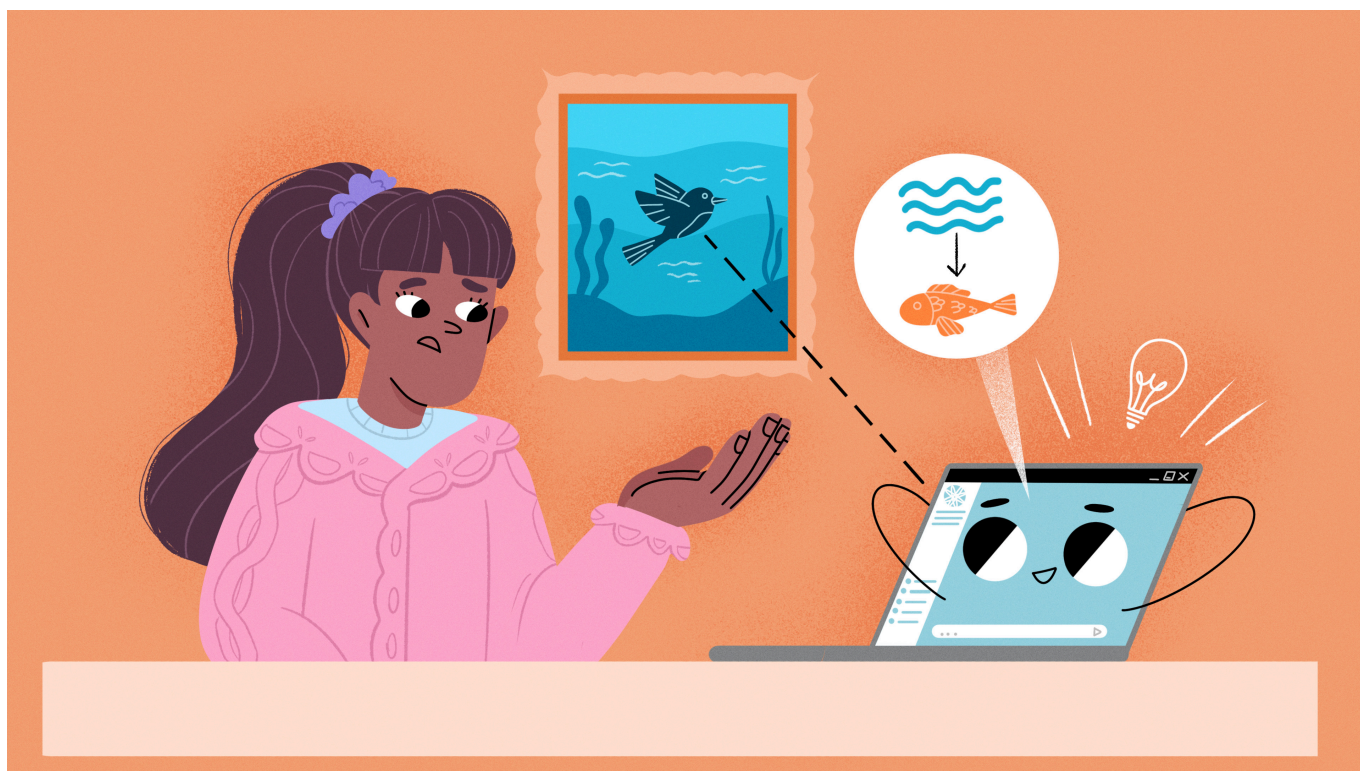




גם בינה מלאכותית משתמשת בקיצורי דרך

Yair Weiss *, Daniella Har-Shalom * | Ofir Shifman

בית הספר להנדסה ולמדעי המחשב, המרכז לחקר המוח ע"ש אדמונד ולילי ספרא (ELSC), האוניברסיטה העברית בירושלים

סוקרים צעירים

THE INTER
DISCIPLINARY
CAMPUS—
TIRAT CARMEL

גיל: 13–14



בינה מלאכותית (AI) זוכה לעיתים קרובות לשבחים על הביצועים המרשימים שלה במגוון רחב של משימות. עם זאת, מאחורי הצלחות רבות מסתתרת בעיה משותפת: הבינה המלאכותית לעיתים קרובות משתמשת בקיצורי דרך. במקום באמת ללמוד כיצד לבצע את המשימה, היא עלולה להסתמך על דפוסים פשוטים בדוגמאות שעליהן אומנה. לדוגמה, מודל שמשימתו לזהות בעלי חיים בתמונות, עשוי להסתמך על הרקע של התמונה במקום על בעל החיים עצמו. לפעמים קיצורי הדרך הללו עלולים להוביל לטעויות חמורות, כמו אבחנה רפואית המבוססת על תג של בית חולים במקום על נתוני המטופל. טעויות כאלה מתרחשות גם במערכות מתקדמות, שאומנו על מיליוני דוגמאות. הבנה של האופן שבו בינה מלאכותית מחפשת קיצורי דרך – ומדוע – יכולה לעזור לחוקרים לפתח שיטות אימון טובות יותר ולהימנע מכשלים חבויים. כדי להפוך את מערכות לבטוחות ואמינות יותר, עלינו לעזור להן לפתח הבנה אמיתית של המשימה – ולא רק לנחש על סמך דפוסים שהצליחו בעבר.

קיצורי דרך לא תמיד עובדים!

"בחיים אין קיצורי דרך". כולנו שמענו את המשפט הזה. כדי להצליח באמת במשימה, צריך לעבוד קשה ולהתמיד, ולא מספיק למצוא "טריק" שיקצר את הדרך. מסתבר, שכאשר מלמדים בינה מלאכותית לבצע משימות שונות, גם היא לעיתים משתמשת בקיצורי דרך, וזה יכול להוביל להשלכות מפתיעות ואפילו מסוכנות. אחד האתגרים המשמעותיים בתחום הוא לוודא שמערכות הבינה המלאכותית באמת לומדות לפתור את המשימה, ולא "מתפתות" לקצר את דרכן בצורה מסוכנת.

מה מבלבל את הבינה המלאכותית?

נתחיל משאלה פשוטה: מה אתם רואים בתמונות העליונות באיור 1?

איור 1

תמונות של גמלים וסוסים,
על רקעים משתנים של
חול ודשא.



כנראה שעניתם גמל בתמונה הימנית וסוס בשמאלית. במשך שנים רבות ניסו מדעני ומדעניות מחשב לכתוב תוכנות שיוכלו גם הן לענות על שאלות דומות, והצליחו רק באופן חלקי. בעשור האחרון חל שיפור עצום בביצועי הבינה המלאכותית והיום גם מחשב יכול לתאר תמונות בדיוק גבוה מאוד. יש היום אתרים רבים שטוענים שהם יכולים לתאר כל תמונה באמצעות בינה מלאכותית. כשהעלינו את שתי התמונות לאחד האתרים המובילים לתיאור תמונות, הוא באמת תיאר נכון את התמונה הימנית כגמל ואת השמאלית כסוס.

נעבור לשאלה קצת יותר קשה. מה אתם רואים בתמונות התחתונות באיור 1.

כנראה שראיתם סוס וגמל, רק על רקעים שונים. תוכלו לנחש איזה תיאור סיפקה הבינה המלאכותית? באופן מפתיע, את התמונה הימנית היא תיארה כשני סוסים ואת השמאלית היא תיארה כשני גמלים! אבל לכולנו ברור שמדובר בגמל אחד וסוס אחד בכל תמונה וגם בדיוק אותו גמל וסוס בשתייהן. זו דוגמה לבעיה בכלי בינה מלאכותית קיימים: מצד

קיצור דרך (Shortcut)

כלל שטחי שמחזיר את התשובות הנכונות על דוגמאות האימון, אבל טועה לעיתים קרובות כשהנסיבות משתנות. למשל, אבחנה רפואית שמבוססת על מאפייני מכשיר רנטגן בבית חולים ולא על תוכן צילום הרנטגן.

למידת מכונה (Machine learning)

שיטה שמאפשרת למחשב ללמוד לבצע משימות מסובכות על בסיס דוגמאות, במקום שמהנדסים יתכנו אותו מראש איך לפתור כל בעיה.

איור 2

תרגילי חשבון. (A) בכל תרגיל יש שני ביטויים, צריך להחליט איזה צד גדול יותר. (B) דוגמאות לתשובות נכונות שמסומנות באדום.

כלל

(Decision rule)

הוראות מפורשות שמגדירות את הפלט של שיטה, עבור קלט נתון.

אחד הם עונים תשובות נכונות במקרים רבים (לאחר שבו השתמשנו יש עשרות אלפי משתמשים ברחבי העולם, והם מדווחים על שביעות רצון מביצועיו), ומצד שני הם לפעמים עושים טעויות מביכות מסיבות לא ברורות. במאמר הזה ננסה להסביר את אחד המקורות הבסיסיים של טעויות שיש לכלי בינה מלאכותית: הנטייה שלהם ללמוד **קיצורי דרך**.

מהם קיצורי דרך בבינה מלאכותית ומה מקור הבעיה?

כמעט כל כלי בינה מלאכותית בשימוש היום מבוסס על טכנולוגיה שנקראת "**למידת מכונה**". מה שחשוב לענייננו הוא שלמידת מכונה יוצרת בינה מלאכותית באמצעות למידה מדוגמאות. כדי להסביר, ניעזר בתרגילים שאתם אולי זוכרים משיעורי החשבון בבית הספר היסודי (**איור 2A**):

A	$2+3-5$	?	$3+2+5$
	$4+7$	?	$2+2+6$
B	$2+3+5$	<	$5+8$
	$4+7$	>	$2+2+5$
	$3+4+1$	<	$7+6$
	$5+9$	>	$1+2+6$

איור 2

בכל תרגיל יש שני ביטויים חשבוניים, וצריך לרשום האם התוצאה בצד ימין גדולה יותר (<) או קטנה יותר (>) מהתוצאה בצד שמאל. אם נבקש ממחשבת לכתוב תוכנה שתפתור את התרגיל, היא כנראה תכתוב קוד שמחשב את הביטוי מימין והביטוי משמאל בנפרד ואז משווה ביניהם. כך פעלו פעם, לפני כ-20 שנה – המהנדסים חשבו על פתרון, ואז כתבו עבורו קוד לבינה המלאכותית.

הגישה המודרנית יותר היא פשוט להציג לבינה המלאכותית דוגמאות רבות של שאלות יחד עם התשובה הנכונה לכל שאלה, ולתת לה למצוא לבד איזשהו **כלל**. ברגע שלמדה את הכלל, הבינה המלאכותית יכולה להשתמש בו גם בדוגמאות חדשות, שאין לה תשובה עבורן. זו הגישה של למידת מכונה והיא זו שאפשרה את ההתפתחות הדרמטית בביצועי בינה מלאכותית בשנים האחרונות. אבל לגישה הזו יש בעיה, שאותה נדגים על תרגילי החשבון.

התבוננו ב(**איור 2B**). נניח שאנחנו נותנים לבינה המלאכותית ארבעה תרגילים עם התשובה הנכונה לכל אחד (מסומנת באדום). כזכור, הבינה המלאכותית מנסה להסיק כלל מדוגמאות האימון. כלומר, הנחיה בסגנון "אם מתקיים כך וכך, אז התשובה היא <," אחרת

התשובה היא >". קחו לכם כמה דקות ונסו לחשוב, איזה כלל פשוט לחישוב שייתן תשובה נכונה אפשר לגלות מדוגמאות האימון?

אולי חשבתם על הכלל הבא: אם יש 3 מספרים בצד שמאל, אז צד ימין תמיד יהיה גדול יותר (>), אחרת התשובה הנכונה היא <. זה כלל מאוד פשוט לבן אדם, וגם לבינה המלאכותית. בוודאי קל יותר מאשר לחשב את התוצאות של כל הביטויים. אבל אפילו שהיא נותנת את התשובה הנכונה על כל דוגמאות האימון באיור 2B, היא כמובן תטעה בדוגמאות רבות אחרות. זהו קיצור דרך – פתרון שטחי, שלא מבוסס על הבנה אמיתית.

נחזור לגמלים והסוסים. נניח שעכשיו אנחנו מראים לבינה המלאכותית דוגמאות אימון – הרבה תמונות עם תיאור, כמו באיור 3. האם תוכלו לחשוב על קיצור דרך לתיאור נכון של כל דוגמה כסוס או גמל?



אולי חשבתם על קיצור הדרך הבא: אם הרקע מדברי אז התמונה של גמל, ואם הרקע הוא מדשאה אז התמונה של סוס. זהו קיצור דרך שפותר באופן מושלם את כל דוגמאות האימון, אבל כמובן לא פותר נכון דוגמאות אחרות. יכול להיות שכלי הבינה המלאכותית מהאינטרנט, שבו השתמשנו בהקדמה, למד קיצור דרך דומה ולכן טעה לגבי התמונות שבהן שינינו את הרקע.

מסתבר שהרבה גרסאות מוצלחות של בינה מלאכותית לתיאור תמונות, מתפתות ללמוד קיצור דרך דומה. במחקר שהתפרסם לפני שנתיים [1], בחנו מספר כלי בינה מלאכותית לתיאור תמונות, שנמצאים כרגע בחזית המחקר. התברר שכשנותנים להם לתאר תמונה כלשהי מהאינטרנט, הכלים צודקים מעל 80% מהפעמים. אבל כשמעבירים את האובייקט לרקע אחר, הם צודקים רק ב-35% מהפעמים. כנראה שקיצור הדרך שמזהה את האובייקט בתמונה לפי הרקע שלו הוא מפתה במיוחד!

דוגמאות אימון (Training images)

אוסף דוגמאות שמוצג לבינה מלאכותית בזמן הלמידה. למשל, תמונות עם תווית של איזו חיה מופיעה בהן, או מדדים רפואיים עם סימון האם הם של אדם חולה או בריא.

איור 3

דוגמאות אימון. 6 תמונות של סוסים ו-6 תמונות של גמלים, יחד עם שם החיה.

העתיד: השלכות והתמודדות

ישנו מחקר שריכז עדויות רבות לקיצורי דרך מסוג זה [2]. באחת הדוגמאות, הבינה המלאכותית למדה לזהות דלקת ריאות מצילומי חזה ולפי זה לקבוע אם אדם חולה או בריא. אבל התברר שהמערכת בכלל לא ניתחה את הריאות שבצילום אלא גילתה שתג בית החולים המופיע בתמונה יכול לרמז האם האדם בריא או חולה והתמקדה בו. קיצור דרך זה הוא מסוכן מאוד: אולי בדוגמאות האימון שהבינה המלאכותית ראתה התג היה רלוונטי, אך מפחיד לדמיין מה יקרה כאשר תתבקש לבחון מטופל נוסף, מבית חולים אחר.

במאמר שפורסם בשנה האחרונה [3], נבדקו 14 כלי בינה מלאכותית שונים שמיועדים לאבחון רפואי, ונמצא שהכלים עושים שימוש נרחב בקיצורי דרך. הכלים נבחנו במשימות מגוונות כמו זיהוי מחלת הקורונה או לב מוגדל מתמונות רנטגן, זיהוי של הפרעות לב באמצעות האזנה לסטטוסקופ ועוד. החוקרים שיערו שהבינה המלאכותית לא תתמקד רק במידע הרפואי ותלמד לנצל את תנאי המדידה הייחודיים לכל מקרה, שיאפשרו לה לעשות קיצורי דרך. לדוגמה, אם רופא חושד שנבדק חולה בקורונה, הוא עשוי לשלוח אותו לצילום במכשיר רנטגן שונה ממכשיר שישלח אליו נבדק שנראה בריא (משיקולים של בטיחות הצוות הרפואי למשל). הבינה המלאכותית יכולה לזהות את ההבדל בין מכשירי המדידה ולהסיק לפי סוג המכשיר אם האדם חולה או בריא, בלי לבדוק את הצילום עצמו. החוקרים מצאו שבכל 14 הכלים שנבדקו נעשה שימוש בקיצור דרך כזה, שהתבסס על מאפייני הציוד ולא על המידע הרפואי. ברגע שהבדיקות נעשו בבית-חולים חדש, פחתו ביצועי הבינה המלאכותית בצורה ניכרת, כי קיצור הדרך כבר לא עבד. זו כמובן בעיה גדולה כשמנסים לשלב בינה המלאכותית בשטח, במצבים רפואיים אמיתיים.

חוקרי וחוקרות בינה מלאכותית מנסים לפתח שיטות למידת מכונה טובות יותר, שימנעו קיצורי דרך. אחת הדרכים היא לאמן אותה עם מספר הולך וגדל של דוגמאות. אם נחזור לדיון שלנו בתרגילי החשבון: במקום לתת לבינה המלאכותית רק 4 דוגמאות אימון, ניתן לה חוברות לימוד חשבון שלמות. סביר להניח שקיצור הדרך שראינו, כבר לא יעבוד. היא תהיה חייבת למצוא פתרון אחר. עם זאת, גם כשמספר הדוגמאות גדול מאוד, אין לנו דרך להבטיח שהמכונה לא תמצא קיצור דרך אחר. במחקר על תיאור תמונות שהזכרנו [1], גילו שגם בינה מלאכותית שאומנה על מאות מיליוני תמונות, נעזרה בקיצור הדרך של שימוש ברקע.

אם יודעים מראש איזה קיצור דרך עלול להופיע, אפשר לנסות למנוע אותו כבר בשלב האימון. למשל, אם נרצה שהבינה המלאכותית לא תזהה חיות לפי הרקע, ניתן להראות בזמן האימון תמונות רבות שבהן כל חיה נראית על גבי רקע אקראי, כמו גמל בשלג. ככל שנוסיף יותר דוגמאות כאלה לאימון, קיצור הדרך שמזהה את החיה לפי הרקע שמאחוריה כבר לא יעבוד. אבל יש כאן בעיה: הגישה הזו ממוקדת בקיצורי דרך שאנחנו כבר מכירים. גם אם נמנע את קיצור הדרך הזה, לא נוכל להתחמק מקיצורי דרך אחרים, למשל התבססות על תנאי תאורה.

גם אצל בני אדם ובעלי חיים אחרים (בינה לא מלאכותית) יש נטייה ללמוד מקיצורי דרך. למשל, תלמידים רבים מתכוננים למבחן בחשבון על ידי פתרון שאלות ממבחנים קודמים, ולפעמים מוותרים על ניסיון להבין את החומר. בתחילת המאה ה-20, התפרסם סוס בשם "הֶנְס הפיקח" שכביכול ידע לפתור בעיות חשבון. כששאלו אותו כמה זה $9+5$, הוא

היה מתופף 14 פעמים בפרסתו. הפסיכולוג אֶסְקֶר פּוֹנְגֶסְט חקר את ביצועי הסוס והגיע למסקנה שהוא הצליח "לפתור" את בעיות החשבון באמצעות קיצור דרך. הסוס היה ממשיך לתופף עד ששפת הגוף של השואל "גילתה" לו (ללא ידיעת השואל) שהגיע למספר הנכון. ברגע שמנעו מהסוס קשר עין עם השואל, קיצור הדרך כבר לא עבד, והתשובות של "הנס הפיקח" היו כמעט תמיד שגויות.

למרות כל מה שתיארנו, אין לאבד תקווה. ראינו בשנים האחרונות שבינה מלאכותית מתעלה על בינה לא מלאכותית ביכולות רבות: היא ניצחה את אלוף העולם בשח-מט וקיבלה ציונים גבוהים יותר מרוב בני האדם במבחני לשכת עורכי הדין. כיום בינה מלאכותית גם מסייעת למורים ולתלמידים בשלל משימות, מחיפוש מידע ועד סיכום טקסטים. חוקרי וחוקרות בינה מלאכותית ממשיכים לפתח שיטות למידת מכונה שיצמצמו משמעותית את ההסתמכות שלה על קיצורי דרך. אולי יום אחד הן יתעלו עלינו גם בהקשר הזה.

הצהרת כלי בינה מלאכותית

טקסט חלופי הנלווה לאיורים במאמר זה נוצר על ידי פרונטירז בסיוע כלי בינה מלאכותית, ונעשו מאמצים על מנת להבטיח את דיוקו, כולל בדיקה על ידי כותבי המאמר כאשר הדבר התאפשר. אם ברצונכם לדווח על בעיה, אנו צרו איתנו קשר.

מקורות

1. Chefer, H., Schwartz, I., and Wolf, L. 2022. Optimizing relevance maps of vision transformers improves robustness. *Adv. Neural. Inf. Process Syst.* 35:33618–32. doi: 10.48550/arXiv.2210.10817
2. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2:665–73. doi: 10.1038/s42256-020-00257-z
3. Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., et al. 2024. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *NPJ Digit. Med.* 7:124. doi: 10.1038/s41746-024-01118-4

פורסם אונליין: 23 בדצמבר 2025

נערך על ידי: Idan Segev

מנחים מדעיים: Daria Patel Teplov

ציטוט: Weiss Y, Har-Shalom D | Shifman O (2025) גם בינה מלאכותית משתמשת בקיצורי דרך. *Front. Young Minds.* doi: 10.3389/frym.2025.1683304-he

Weiss Y, Har-Shalom D and Shifman O (2025) Artificial Intelligence Takes Shortcuts Too. Front. Young Minds 13:1683304. doi: 10.3389/frym.2025.1683304

הצהרת ניגוד אינטרסים: המחברים מצהירים כל המחקר נערך בהעדר כי קשר מסחרי או פיננסי שיכול להתפרש כניגוד אינטרסים פוטנציאלי.

זכויות יוצרים © 2025 Weiss, Har-Shalom I Shifman 2025. זהו מאמר בגישה פתוחה שמופץ תחת תנאי רישיון [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). השימוש, ההפצה או ההעתקה מותרים לשימוש בפורומים אחרים ובלבד שיינתן קרדיט למחברים המקוריים ולבעל זכויות היוצרים, ושהפרסום המקורי בעיתון זה מצוטט בהתאם למקובל באקדמיה. השימוש, ההפצה או ההעתקה אינם מותרים אם הם אינם עומדים בתנאים אלה.

סוקרים צעירים

THE INTER DISCIPLINARY CAMPUS— TIRAT CARMEL, גיל: 13–14

אנחנו קבוצה של תלמידי כיתה ח' שבחרו להשתתף בתוכנית שנתית לאוריינות מחקרית של "פרונטירז מדע לצעירים". מתחילת השנה למדנו איך קוראים מאמרים מדעיים, סקרנו מאמר ויצרנו פוסטרים מדעיים בתחומים שאנחנו אוהבים. הפעילות העשירה את הידע שלנו ומאוד נהנינו לקרוא מאמרים מהאתר ולהעמיק בהם.

הכותבים

YAIR WEISS

אני פרופסור למדעי המחשב ומחזיק הקתדרה ע"ש דיטר שוורץ לחקר בינה מלאכותית באוניברסיטה העברית, גר בירושלים, נשוי ואב לארבעה ילדים (איתי בן 14, עדי בת 19, רוני בת 21 וגילי בת 24). בנוסף להיותי חוקר, שימשתי שנים רבות יועץ לחברת הטכנולוגיה מובילאיי. כמו כן, אני בעלים של קבוצת כדורגל – הפועל (קטמון) ירושלים. *yair.weiss@mail.huji.ac.il

DANIELLA HAR SHALOM

אני דוקטורנטית למדעי המחשב באוניברסיטה העברית בירושלים. אני חוקרת איך בינה מלאכותית "רואה" את העולם, ואיך היא עושה זאת בהשוואה למוח האנושי. ספציפית, המחקר שלי מתמקד באיך הבינה המלאכותית מייצגת את הדברים שהיא רואה, אצלה בתוך ה"ראש". מעבר למחקר, אני הכי נהנית להתבונן על ביתי (עמליה, בת השנתיים), וללמוד איך היא רואה את העולם. בזמני הפנוי, אני אוהבת לטייל ולטפל בגינה שלי. *daniella.horan@mail.huji.ac.il

OFIR SHIFMAN

אני חוקר מערכות לומדות ובינה מלאכותית יישומית בתעשייה. למדתי תואר ראשון בקוגניציה מתוך עניין במוח האנושי ותשוקה להבנה כיצד אנשים לומדים וחושבים, אך בעקבות אהבתי להנדסה וליצירת מערכות טכנולוגיות, השלמתי תארים ראשון ושני במדעי המחשב באוניברסיטה העברית. מה שמרתק אותי במיוחד הוא אתגרי היישום של מכונות לומדות בעולם האמיתי והמורכב, וכן הפער בין תהליכי הלמידה של בני אדם למכונות. בנוסף למחקר, אני חובב אופניים, שמאפשרים לי להימנע מפקקים, ונהנה לטייל בשעות הפנאי.



מוזיאון המדע ע"ש בלומפילד ירושלים
متحف العلوم على اسم بلومفيلد القدس
Bloomfield Science Museum Jerusalem



הוצאת פרונטירז מדע לצעירים ישראל
Hebrew version provided by



THE SAGOL NETWORK



קרן משפחת
שעשוע
Shashua
Family Foundation