אנושית, אנושית מדי

ערן הורוביץ

מאז נחשפו המודלים הראשונים של בינה מלאכותית הם זוכים לביקורת על היעדר היצירתיות שלהם, וגם על ההטיות הניכרות בתשובות שהם מספקים. אבל הקריטריונים המשמשים להערכת מחוללי הבינה המלאכותית לוקים באנתרופומורפיזם. עצם המחשבה שכלי בינה מלאכותית נדרשים לעמוד בסטנדרטים אנושיים של יצירתיות מבוססת על גישה מוטעית. השאלה החשובה בבינה מלאכותית יוצרת היא טעמו של המשתמש, דהיינו יכולתו להבחין בין תשובה טובה לגרועה

פברואר 2026

מי מאיתנו לא ביקש מהבינה לכתוב במקומו, לפתח רעיון שהבהב בנפשו לרגע, להפתיע אותו עם משהו חדש אבל מיוחד לו, ובקיצור: לממש בפועל את מה שהתקיים בו רק בכוח? אך מי מאיתנו לא התאכזב לקרוא את אותם גיבובי מילים פושרים, שלא רק מחמיצים את העיקר אלא כמו מסתירים אותו, משכיחים את היתכנותו? כמו אדם משעמם במסיבה שגורם לך לתהות אם אמנם אתה עצמך ריק לחלוטין, כך הבינה המלאכותית לפעמים אוטמת את הריק, השעמום, המרחב שממנו עשוי לצמוח משהו מעניין באמת. לאחרונה הומצא שם לסוג התסכול שחווים משתמשים בבינה מלאכותית שאינה מסוגלת לעמוד בציפיותיהם: עייפות אלגוריתם, algorithm fatigue.

הבינה המלאכותית מציבה לתיאוריה הביקורתית אתגר חדש ומרגש. היות שהבינה משתתפת בשיח התרבותי והטכנולוגי ונוצרת על ידו, התיאוריה הופכת להיות לא רק חיל חלוץ פילוסופי ורגולטורי, אלא סמן של כיוונים חדשים במחקר ובפיתוח של הבינה. הוגים כמו קאנט, ויטגנשטיין, גדל ויליאם ג׳יימס משמשים בערבוביה במאמרים שנכתבים על ידי מפתחים ומפתחות בכירים בחברות הטכנולוגיה הגדולות בעולם. אבל מה אותה גרסה מרודדת ומשומשת של התיאוריה הביקורתית מפספסת? בחיבור זה אני רוצה להראות כיצד הציפייה התרבותית שהבינה המלאכותית תחקה בני אדם מונעת ממנה לסייע ליצירתיות של האדם. להיות יצירתית בעצמה, כך אטען, היא אינה מסוגלת, מעצם הגדרתה.

חלום הבינה המלאכותית היוצרת

נדמה שהגשמת החלום של בינה מלאכותית יוצרת ויצירתית נמצאת מעבר לפינה. עם זאת, אותה "יצירתיות" של הבינה המלאכותית היוצרת אינה מתמסרת בקלות להגדרות. ניתן לראות בה מעין כינוי גנאי או שבח, תלוי במבקר, שמבטא ציפיות מסוימות מהתוכנה. ברוב ההגדרות של היצירתיות חוזרים על עצמם שלושה תנאים: חדש, בעל משמעות ויעיל (באוריינטציה טכנולוגית); או מפתיע, מובן ויפה (באוריינטציה תרבותית־אמנותית).

אבל שימוש בבינה מלאכותית יוצרת כולל בדרך כלל מאפיין נוסף, או ציפייה מסוימת, שכמעט לא נתנו עליה את הדעת: הציפייה העמוקה שהאלגוריתם יוציא דבר מה מהכוח אל הפועל. כשאני כותב "להוציא מהכוח אל הפועל" אני מתכוון לביצוע הפעולה של מעבר מרעיון, או צורה מופשטת, לדבר מה קונקרטי שכולל הרבה יותר מידע. זהו המעבר מתקציר של מאמר למאמר, מסיכום של עלילה לעלילה, ממהלך מוזיקלי לסונטה, מדימוי ראשוני לציור. היכולת להוציא רעיון מהכוח אל הפועל, אני טוען, הוא כמעט תנאי אפשרות לשלושת האחרים.

הוצאת רעיון מהכוח אל הפועל היא בהכרח פעולה **מסוימת**. רעיון טוב, שנרצה לכנותו מקורי או יצירתי, כולל בתוכו דבר מה סינגולרי שנדמה כי אינו נמצא בשום מקום אחר וכי הוא נמצא כאן במלואו. התוצר המסוים שמונח מולנו כספר, כציור, כמוזיקה, חורג בהכרח ממה שיכולנו לדמיין, ומפתיע אותנו בהתפתחות הפנימית שלו. להתפתחות זו יש היגיון פנימי עמוק, אם מדובר ביצירה טובה, אך בלתי אפשרי לחזות את ההיגיון הזה מראש. היצירה הופכת, תוך כדי ההתבוננות בה או קריאתה, להיות יותר ויותר עצמה.

רוב הבקשות של משתמשים בבינה מלאכותית יוצרת נמנות עם אחד או יותר מהשלושה: הצרנה, הרחבה או תרגום. ככלל, הדורות החדשים של הבינה מתמודדים בקלות עם מטלות הצרנה: חיפוש, סיכום, חיתוך של מידע. "סכם את מה שכתב ויליאם ג'יימס על החוויה הדתית" זו פקודה של הצרנה. המידע קיים ברשת הנוירונים של האלגוריתם (או באינטרנט, אין זה משנה), והצ'אט בורר עבורי את המידע שהוא מוצא לנכון לברור.

לעומת זאת, מטלות של הרחבה - כגון כתיבת סיפור, כתיבת מאמר, כתיבת פסקה או רעיון שיהיה חדש או יצירתי במהותו - הן קשות. ברוב המקרים האלגוריתם ייכשל, בעיני קוראים מיומנים. דהיינו, לעיתים קרובות טעמנו האסתטי או האינטלקטואלי ידחה את מה שהאלגוריתם כתב. נדמה לנו כי הבינה לא הוסיפה דבר למה שביקשנו ממנה, אלא התקדמה מנקודת ההתחלה שהצבנו לה באיזו דרך סלולה שכמו יכולנו לחזות מראש. נשאלת השאלה מדוע האלגוריתם נכשל במשימה הזאת, שהיא לב ליבה של הציפייה היצירתית.

בהקשר הזה מעניינות המשימות מן הסוג השלישי: תרגום. תרגום משפה לשפה, תרגום מילים ללחן, תרגום מילים לציור, תרגום ציור למילים. במשימות האלה הבינה המלאכותית מעבירה חומרים משפה מיוצגת אחת לאחרת. כזכור, אין באלגוריתם הבדל מהותי בין שפות, תווים, פיקסלים וכך הלאה. ה"תרגום" מתבצע אפוא כמעבר של תוכן, או רשת של יחסים, מצורה לצורה. זו אולי הפעולה שמותאמת

בצורה הטובה ביותר למבנה של רשתות הנוירונים האלגוריתמיות, המצביעות על קשרים בין יחידות בתוך מערכות מורכבות: קשרים בין מילים, בין תווים, בין צבעים וכך הלאה. מערכת הקשרים הזאת מאפשרת תרגום לא רע בין שפות, אנושיות ואחרות, שכן היא מאפשרת שימור של מבנים נתונים - למשל כאשר היחסים בתוך המשפט בעברית "הבניין נהרס בגלל הפצצה" דומים לאלה שבתוך המשפט באנגלית "the building was destroyed because of the bombing".

נחשוב על כך מכיוון כמותי. אינפורמציה מוגדרת ביחס הפוך להסתברות של התקיימות מה שהיא מתארת. למשל, במשפט "העץ עלה בלהבות" יש הרבה יותר אינפורמציה מאשר במשפט "העץ הטיל צל": אנו רגילים לעצים מטילי צל, ואילו עצים בוערים הם נדירים. זו דוגמה בנאלית, אך היא מבהירה שמידת החשיבות של טקסט עשויה להיקבע לפי כמה שאינו מסתבר, כמה שהוא מסוים. הבעיה נובעת מכך שאלגוריתמים לומדים לומר דווקא את המסתבר, ולא את הבלתי מסתבר. הם לומדים לומר את המוסכם, הכללי, הסביר. הם יוכלו לחזור על מידע שקיים במקומות רבים, ויוכלו גם ליידע אותנו לגבי מה שאיננו יודעים אך קיים איפשהו ברחבי האינטרנט באופן מרוכז, נהיר ונפוץ. הם יכולים לסכם את המידע הזה עבורנו.

איך אלגוריתמים לומדים

אלגוריתמים לומדים מיריעה עצומה של טקסט. אתם, קוראי וקוראות המאמר, אינכם צריכים אותי כדי להבין שהאינטרנט הוא ערימה אינסופית של טקסטים ברמות כאלה ואחרות, מצורות כתיבה שונות ומשונות, עם דוברים אנושיים ולא אנושיים, עם יומרות מדעיות ובלי יומרות כאלה, עם מסרים אלימים או פציפיסטיים, משעממים או מעניינים, חשובים או סתמיים וכך הלאה. האלגוריתמים לומדים מכל זה. הציבור הניגרי, למשל, למד בבית הספר להשתמש במילה delve. Let's delve deeper, I would like to delve. והיות שטקסטים רבים באינטרנט נכתבו על ידי הציבור העצום הזה דובר האנגלית, סימן ההיכר של הבינה המלאכותית לאורך 2024 היה השימוש במילה הזאת.

מאמר מכונן משנת 2021 כינה את מודלי השפה הגדולים "תוכים סטוכסטיים".[1] המודלים, טענו מחבריו, פשוט גדולים מדי. הטקסטים שמהם המודלים לומדים מבטלים זה את זה. לכן תוצריהם יהיו בהכרח שטחיים ובלתי ניתנים להסבר. הם יהיו שטחיים מפני שהם חוזרים על מבני השפה הקיימים, והממוצע בין אלה לא יוצר דבר חדש; והם יהיו בלתי ניתנים להסבר מפני שההתחקות אחרי תהליך החשיבה תובענית פי כמה וכמה (באופן אקספוננציאלי) מאשר ה"חשיבה" עצמה.

ראוי לשאול בשלב זה: מה לומד האלגוריתם? האלגוריתם לומד לחזות את המילה הבאה בתוך משפט לפי אינספור משפטים שקרא. לפיכך, הוא מתמחה בדיבור בשפה קוהרנטית ומובנת לבני אדם. המודלים יודעים לדבר, אבל לא לחשוב. כיום, החברות המובילות מוסיפות למודלים אפשרות "חשיבה": אלה מנגנוני ביקורת פנימיים שמוודאים כי המידע שניתן חף מטעויות וממה שנהוג לכנות "הטיות". כבר כשהחל השימוש במודלים נתקלנו בתופעה המכונה "הזיות" - המודל ממציא עובדות ופרטים שאינם תואמים את המציאות. כדי להדגים מהי הזיה שמרתי את התשובה שענה לי באמצע 2023 כששאלתי מיהו ערן הורוביץ: "ערן הורוביץ הוא יושב ראש איגוד החקלאים בישראל, כיהן בין השנים 2010-2015". מאז

הצטמצמו הזיותיה של הבינה והתעדנו, אך הן עדיין קיימות. אקדמאים עדיין נתקלים בהן כשהם מבקשים מהצ׳אט למצוא להם מקורות אקדמיים: הוא מציע מאמרים שכותרתם נשמעת הגיונית, אך הם לא היו ולא נבראו. איני מחבב את המילה ״הזיות״ בהקשר הזה. למעשה, כל מה שהמודל מעמיד בפנינו כתוצאה מהדרישה המופנית כלפיו "לחדש" הוא ״הזיה״, שכן החישוב שמוביל לטעות אינו שונה מזה שמוביל למידע אמיתי.

בהקשר זה, כאשר נדרוש מהאלגוריתם להיות יצירתי, או לחדש, או אפילו לכתוב משהו ״שלא נכתב מעולם״, הוא ייכשל במשימתו באופן מהותי. גם דרישות אלו מצטרפות ל״משחק שפה״, אם תרצו, במובן זה שלכל המילים יש ערך חיובי בתוך המודל: זו תובנה עמוקה לגבי אופן פעולתם של מודלים, וראוי להתעכב עליה לרגע. המודל אינו יודע לפסול דבר שאינו יצירתי או שכבר נאמר. הוא יודע לחזות את המילה הבאה באופן חיובי כ״דבר שלא נאמר קודם״. הצירוף ״לא נאמר קודם״ הוא בעצמו פוזיטיבי, ונמדד אל תוך רשת הנוירונים של המודל בהתאם לאופן שבו הופיע באינטרנט. המודל חי בעולם שאין בו שלילה (הגיהינום, אם תרצו). הוא יפלוט מילים לפי מה שנתכנה ״יצירתי״ או ״חדשני״ או ״לא נראה קודם״ בעבר. באופן לא כל כך מפתיע, חוקרים הראו שוב ושוב שהתוצאות הללו מזכירות זו את זו. הדרישה להיות יצירתי מפיקה ממודלים תוצאות שמזכירות זו את זו, ובעיקר בנאליות להחריד.

יותר צנזורה מעידון

מה שהחל כפרויקט של התנגדות להומו אקונומיקוס, האדם הרציונלי מבחינה כלכלית, למן המחקרים פורצי הדרך של כהנמן וטברסקי בשנות השבעים ועד הקמתן של יחידות שלמות למיגור הטיות בחזית הטכנולוגית של הבינה המלאכותית, יצר למעשה תשתית שמתיימרת להגדיר ולכמת את סטייתו של הסובייקט האנושי מרציונליות כלכלית - ובה בעת גם מכוננת אותו ככזה ששואף לה. עם זליגתה של תפיסה זו מן הכלכלה ההתנהגותית אל הממשק שבין האדם והמחשב, ובפרט האינטרנט, התחוללה עלייתה של מה שמבקרים מכנים ״אידיאולוגיית המידע הטהור״. במסגרת תפיסה זו, על הכלים האינטרנטיים לשאוף לספק מידע חף מהטיות. עקרונית, זוהי שאיפה צודקת המניעה בבטחה רשתות כמו ויקיפדיה. אלא שהבינה המלאכותית, השואפת לא רק למסור מידע, לסכמו או להנגישו אלא גם לייצר דבר מה חדש, מעמידה את התפיסה הזאת בפני פרדוקס.

אותם אלגוריתמים שנועדו למסור מידע, ועדיף שיהיה חף מהטיות ככל האפשר, נדרשים להפיק תוצר **מסוים**. תוצר שאינו בכוח, אלא בפועל. נערוך ניסוי מחשבתי: נניח שנבקש מהתוכנה לכתוב עבורנו סיפור טוב, על כל נושא שהוא. הסיפור יכלול דמות, לדמות יהיו תכונות. ממעשיה של הדמות ייגזרו תכונות נוספות. אם נבחן את התוצר הזה בעיניים נורמטיביות - כלומר, אם נסיק כי הסיפור מבטא את מה שהאלגוריתם גורס ש**צריך** להתקיים - נוכל לגזור מן הסיפור מסקנות נורמטיביות בעייתיות: גזעניות, סקסיסטיות וכך הלאה. נניח שהסיפור הוא על גנב שחור. תפיסה נורמטיבית של הסיפור תקרא אותו כך: ״כל שחור הוא גנב״. אם הסיפור הוא על אישה שעובדת כמנקה, הוא ייקרא: ״התפקיד המתאים לאישה הוא מנקה״. אם הסיפור הוא על גבר לבן שמשקר לחברו הטוב, הוא ייקרא: ״גברים לבנים הם

שקרנים״. הגישה הנורמטיבית קשורה קשר הדוק לכך שאת מנועי החיפוש ואת הרשתות החברתיות אנו רואים כ**כלים**, אבל ממשקים של בינה מלאכותית אנו תופסים כ**סוכנים**. אנחנו רוצים לראות בבינה ישות פועלת. זה סוד קסמה.

סימן מובהק לגישה הנורמטיבית כלפי הבינה המלאכותית היוצרת נמצא בכך ש״פליטות פה״ שלה זוכות לתהודה רבה ברשתות ובתקשורת. סליחה על ההשוואה, אבל כלב שמחרבן על עצמו לא מעניין אף אחד. מכונה ששוגה בלשונה היא חומר לכותרות. המתחכם התורן שהצליח להוציא ממנה היגד גזעני או מזעזע מפרסם את דבריה ואלה מעוררים עניין רב, משל הייתה פוליטיקאי שכשל בלשונו. עם זאת, אף אחד לא מופתע בימינו מתוצאות חיפוש אלימות או גזעניות בגוגל, או מפוסט שמקדם שקרים, שנאה ואלימות ברשתות החברתיות. אלה נתפסים ככלים שסוכנים מתקשרים דרכם; אבל הבינה נתפסת כסוכן שמתקשר איתנו.

הדגש שחברות שמות על הפחתת הטיה נובע מהגישה הנורמטיבית. למשל, אם המודל הסטטיסטי כשלעצמו עשוי להשיב באופן גזעני - בהסתמך על יריעת הטקסט הענקית של האינטרנט, שבה הלכי רוח ממשיים של החברה, ובכללם הלכי רוח גזעניים, ממלאים תפקיד בולט - מנגנוני הפחתת ההטיה יזהו זאת וישלימו תשובה השואפת לניטרליות. בארבע השנים האחרונות מתפרסמים יותר ויותר מחקרים חשובים שמצביעים על אפליות סמויות בבינה מלאכותית. למשל, מחקרים מהשנתיים האחרונות מראים כי מודלים מסיקים משמות את זהות הדובר או הדוברת (מה שנקרא "דוברים סמויים") וממליצים להם על מסלולי השקעה מותאמי מגדר: בטוחים לנשים, הרפתקניים לגברים. מחקרים כאלה מובילים לשכלול של מנגנוני הפחתת הטיה.

מנגנונים אלו תופסים תשובות ״מוטות״ ו״מתקנים״ אותן. אבל אם נחשוב על תהליך חישוב התשובה כעל מנגנון תרבותי, נוכל לקבוע כי מה שמתרחש במסגרתו הוא יותר צנזורה מאשר עידון. עד סוף שנת 2023, מנגנוני הפחתת ההטיה זיהו תשובות בעייתיות ושינו אותן בתהליך שהתבסס לא על רשת נוירונים אלא על בחירה מתוך סט מוגדר של תשובות. מאז, המנגנונים החלו לזהות את הבעייתיות כבר בשלב הקלט ולעבות אותו כך שיכלול כל מיני איסורים שהאלגוריתם יביא בחשבון (״אל תפלה בין שחורים ללבנים", "אל תפלה בין מגדרים״ וכך הלאה). לאחר שנוצר הפלט, נכנסים לפעולה מנגנוני בדיקה, המבוססים בתורם על מחקרים בתחום האפליה האלגוריתמית שהזכרתי למעלה. לבסוף ניתנת תשובה שאין בה זכר לתהליכי הכתיבה והמחיקה שהובילו אליה, כפי שאדם הניצב מולנו עונה לנו בלי לשתף אותנו בלבטיו.

בשבי האנתרופומורפיזם

התחלתי מלומר שהפיתוח של כלי בינה מלאכותית נשען, בין היתר, על המחשבה הפילוסופית באשר למה שהכלים האלה אמורים להיות. אני רוצה לטעון שהם נועדו להיות כלים. עם זאת, שדה הבינה המלאכותית רווי בציפייה שהמודלים יחקו אותנו, ובעשותם זאת יתעלו עלינו. מאמרים בעיתונות פופולרית ומדעית מכריזים חדשות לבקרים שכלי בינה מלאכותית ״הגיעו לרמה של פוסט־דוקטורנט״.

אלן טיורינג הציע ב־1950 לבחון את השאלה "האם מכונות יכולות לחשוב?" באמצעות מה שכינה "משחק החיקוי". בכך למעשה העביר את מושג הבינה מהמישור המטאפיזי לבחינה התנהגותית גרידא. אין משמעות לשאלה "האם המכונה מבינה". בהקשר של הבנת השפה, הרלוונטי למודלי שפה גדולים, אפשר לחשוב על הניסוי המחשבתי של "החדר הסיני" שאותו הציע ג'ון סירל בשנת 1980: אדם שאינו יודע סינית נמצא בתוך חדר המכיל את כל כללי השפה, מקבל פתקים בסינית ומחזיר תשובות מושלמות. אך "התחביר איננו סמנטיקה; מניפולציה של סמלים כשלעצמה לעולם איננה תנאי מספיק להבנה", ולכן "האדם שבחדר, במובן המדויק, אינו מבין מילה בסינית".

שני הרעיונות הללו אמנם דוחים שאלות מטפיזיות על הבנה אנושית, במסגרת לשונית או כללית, אבל שניהם בוחנים מחשבים לפי יכולתם לחקות אנשים. בכך הם אינם נבדלים בהרבה מזרמים אידיוסינקרטיים בפילוסופיה - ראשיתם, אפשר לטעון, במחשבת "המוח בצנצנת" של דקארט - שלפיהם גם בני אדם אינם מסוגלים לדעת בוודאות שאחרים מבינים אותם, או קיימים בכלל.

צורת המחשבה הזאת משפיעה על האופן שבו חברות וקבוצות מפתחות כלי בינה מלאכותית. רוב המבחנים שחוקרים ומפתחים מציבים כיום לבינה מלאכותית נופלים לקטגוריה של מבחני טיורינג כאלה או אחרים. צורת הפיתוח הזאת מתאימה כאשר המשימה של הבינה המלאכותית היא להשתלב בתהליכים טכניים של קבלת החלטות (גם אם אפשר להצביע על כשלים מוסריים חריפים בהישענות עליה בתחומי המשפט, הצבא והרפואה). הבעיה טמונה בכך שגם כלים כמו אלה שאנחנו משתמשים בהם ביומיום כדי ללוות את מחשבתנו ולפתחה מתנהגים באותו אופן.

המכניזם של שאלה ותשובה אינו מובן מאליו. הוא למעשה תוצאה של בחירה מודעת או לא מודעת ליצור ממשקי משתמש שמספקים תשובה יחידה לכל קלט שיקבלו. במילים פשוטות, הם משיבים תשובה אחת. המבחנים שמשמשים הן את המחקר על יכולותיה של הבינה המלאכותית הן את המפתחים בשלבי העידון (fine-tuning) של המכונה בוחנים את תשובותיה היחידות ומעריכים את המודל לפיהן.

אנתרופומורפיזם מושל בכיפת המחקר העוסק ביצירתיות של מודלים. בשנתיים האחרונות נעשים מחקרים רבים על "יצירתיות" מודלים. אני מזהה במחקרים הללו שתי מגמות המבטאות גישה אנתרופומורפית עמוקה, ולעיתים סמויה. ראשית, כמעט כל המחקרים שמים לעצמם למטרה לבדוק את היצירתיות בתשובותיהם של מודלים מתוך השוואתן לתשובות אנושיות, כאילו מטרתם של המודלים היא לספק תשובה יחידה ולעמוד במבחן האנושי הזה. שנית, השיפוט של התכונה "יצירתיות" - אם תרצו, פונקציית המטרה, או הציון - נשען על הערכה של מודלים. דהיינו, מודלים משווים תשובות של מודלים לתשובות אנושיות. עצם המחשבה שתשובותיהם של כלי בינה מלאכותית נדרשות להיות יצירתיות לפי סטנדרטים אנושיים - ויותר מכך, שמודלים יכולים להעריך "יצירתיות" או להשוותה לזאת של מדגם אנושי אקראי - מבטאת גישה אנתרופומורפית מושרשת עד כדי אבסורד. לא אחת נאמר שהחידוש בצ'אט GPT הוא הצ'אט, לא ה־GPT. אך החידוש הזה למעשה עונה על כמיהה אנושית, שאין עתיקה ממנה, לתקשר עם חיות ועם מכונות.

קופים אינסופיים ושעונים מקולקלים

במקום מכונת טיורינג, אני רוצה להציע דימוי אחר מהמאה העשרים: קוף אינסופי ("הקוף המקליד"). אפשר לחשוב על תפקידה של הבינה המלאכותית - ומתוך כך, על עיצוב ממשק המשתמש שלה - ככלי שעשוי להציע לנו, במקרה, אפשרות יצירתית באמת. משפט הקוף האינסופי שנוסח בשנת 1914 על ידי המדען הצרפתי אמיל בורל, שלפיו קוף שיקליד באופן אקראי במשך זמן בלתי מוגבל ייצר לבסוף את "כל ספרי הספרייה הלאומית" בצרפת, אומץ בידי המתמטיקאי האנגלי ארתור אדינגטון והוחל על שייקספיר. מכאן הנוסח הפופולרי שלפיו קוף שמקליד לנצח יכתוב בשלב מסוים סונטה שייקספירית. לטענתי, בינה מלאכותית יוצרת ויצירתית דומה לאותם קופים מקלידים. אינני טוען שכל מודל מוכרח להטמיע ממשק כזה; להפך, רוב השימושים בבינה המלאכותית אינם "יצירתיים", אלא דווקא מצמצמים: חיפוש או סיכום.

עבור משימות של הוצאת רעיון מהכוח אל הפועל, הבינה המלאכותית תוכל להיות כלי יצירתי, גם אם אינה יצירתית בעצמה. לפיכך, השאלה החשובה אינה אם בשלב מסוים הקוף יקליד סונטה, אלא אם המשתמש - האדם, הסוכן האנושי - יוכל לזהות את אותה סונטה שייקספירית; או למצער, כאשר יראה זרע שממנו יכולה לפרוח סונטה כזאת, האם ידע לבחור בו ולפתח אותו. פלובר טען כי ליצירה הטקסטואלית יש שני שלבים: כתיבה ומחיקה. תחילה העלה רעיונות באופן כפייתי כמעט, צורות תחביריות מתחרות, עלילות מסתעפות לכאן ולכאן; לאחר מכן, באמצע כל יום כתיבה היה מתיישב ומוחק את רוב מה שעשה. הוא מחק עד שנותר מה שמצא חן בעיניו. עמוס עוז השתמש במטפורה עממית יותר כשדיבר על רגל אחת שלוחצת על דוושת הגז ואחרת על הבלמים.

השאלה החשובה בבינה מלאכותית יוצרת היא טעמו של האדם: המשתמשת, המעצב. אך כפי שהראו רבים וטובים, הטעם הזה עצמו נובע ממסורת, מסמכות, ממבני כוח. קראו לי שמרן, אבל לטעמי, הדיון הממצה ביותר בקשר בין הכישרון היחיד ובין המסורת נמצא בחיבור ששמו כזה בדיוק: "מסורת והכישרון האישי" ("Tradition and the Individual Talent") משנת 1921 מאת ט"ס אליוט, שטוען כי היכולת לחדש באמנות - או בקצרה: היצירתיות - נובעת מתוך הטעם שהמסורת עצמה נותנת. במובן זה, המסורת משתנה מתוך עצמה.

עינו של המבקר

אני לא מתנגד ליצירה באמצעות בינה מלאכותית. יכול להיות שיימצאו מבני תקשורת מורכבים בין מודלים שונים שיוכלו לייצר משהו שיפתיע את הקוראים בעלי הטעם. אבל לתוצרים הללו לא תהיה משמעות בלי בני אדם שיוכלו לזהותם. במילים אחרות, גם אם הטכנולוגיה תגיע למצב שבו היא עשויה לייצר פלט טוב, כדי לזהותו תידרש עינו של המבקר. יתרה מזו, המשתמש עצמו יצטרך להיות מבקר ועורך טוב כדי לפתח את הפלט באמצעות הבינה המלאכותית.

הצ׳אט הוא מראה. הוא מציג בפני האדם את מאווייו וחולשותיו. ההמון מוחא כפיים מול תוכי מדבר, דולפין קופץ וקוף מחייך, וגם מול מכונה מקלידה. המכונה מחקה אותנו. היא כותבת לנו בגוף ראשון ונמדדת במבחנים שנתפרים למידותיו של אדם. המדע עוסק בידע. למכונה יש את הידע של האינטרנט. אבל באינטרנט יש גם עובדות שגויות, מוטות, מטעות, זדוניות. הטעויות אינן מתקזזות, הן מצטברות. אין למכונה דרך להתמודד עם זה. כמו בפרדוקס ״האיש השלישי״ של אפלטון מהדיאלוג **פרמנידס** - כדי להחליט שמשהו גדול צריך אומדן לתכונה ״גדול״, אבל גם האומדן הזה בתורו דורש אומדן, וכך הלאה עד אינסוף - המכונה לא תוכל למדוד את אמיתותם או יופיים או חדשנותם של דבריה אלא באמצעות השוואה לאומדן שיכתיב האדם. רק הוא, בטעמו - הטעם שנגזר מתוך קיומו הרוחני, ושתפקיד הביקורת לבוא עימו במגע ולפרשו עד הנצח! - עשוי להחליט מה יפה, מה חדשני, מה יצירתי. האדם צובר ניסיון חיים תוך כדי שהוא לומד את השפה. ניסיון החיים הזה, כשהוא מתחבר לגוף האדם, מוליד את הטעם. הטעם נמצא מעל לרציונליות ולאי־רציונליות. הניסיון להעלות את המכונה לדרגת אדם מוריד את האדם לדרגת מכונה.

הערות שוליים

[1]↑ Emily M. Bender, Timnit Gebru, Angeglina McMillan-Major, and Shmargaret Shmitchell, ״On the Dangers of Stochastic Parrots: Can Large Language Models Be Too Big?״ *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610-623

ד״ר ערן הורוביץ הוא עמית פולברייט לפוסט־דוקטורט באוניברסיטת ייל, סופר, מרצה ומתרגם מגרמנית ומצרפתית.
