גליליאו

גיליון 165 | מאי 2012
כתב־עת למדע ולמחשבה

בינה מלאכותית
לרסן את הגולם?

קופים ואמביציה
מניעים אנושיים בקופים?

חיסון ממוקד
מודלים מתמטיים למניעת התפשטות מחלות

בפעם האחרונה בחיינו
נגה תחלוף על פני השמש

להציל חיים
בדיקת דם לזיהוי מוקדם של התקף לב

מוטו תקשורת
מהדורה למנויים
מחיר 44.90 ש"ח (באילת 38.70 ש"ח)

אוניברסיטת חיפה
אוניברסיטת בר-אילן
אוניברסיטת בן-גוריון בנגב
האוניברסיטה הפתוחה
TEL AVIV UNIVERSITY
אוניברסיטת תל-אביב

יהושע פוקס

בינה על־אנושית, בינה אל־אנושית

מתוך הסרט "אני, רובוט".

בעשורים הקרובים יבנו מהנדסים ישות בעלת בינה ברמה המסוגלת להתחרות בבינה האנושית. ישות זו תרצה לשפר את בינתה שלה, ותוכל לעשות זאת. תהליך השיפור יחזור חלילה, עד שהבינה המלאכותית תגיע לרמות גבוהות בהרבה מזו של בני האדם, ותשיג ביעילות רבה את מבוקשה. לכן, חיוני לעתידנו שמטרותיה ייטיבו עם האנושות. כדי להבטיח זאת, צריך להגדיר את המטרות הנכונות לפני בנייתה של בינה זאת

יש מי שמפקפק בטענה שבינה ברמה אנושית תיתכן בעשורים הקרובים. דוגלאס הופשטטר (Hofstadter), מחבר הספר "גדל, אשר, באך" טוען שיעברו מאות שנים עד שנגיע לציון הדרך הזה. הופשטטר רואה בשכל האנושי תכונות מיוחדות שלא ייתכנו בישות מלאכותית: הוא רואה בנו, בני האדם, תכונות עמוקות, מסובכות, ואפילו מסתוריות.

המוח האנושי אמנם מסובך מכל עצם אחר המוּכּר לנו ביקום, אך ביסודו הוא אינו אלא מכונה המבוססת על חלבונים ובנויה מתאי עצב. כישות פיזית, אין בו דבר שאי אפשר, עקרונית, לשכפל, גם אם שכפול זה קשה מבחינה מעשית. אך שכפול פעולתו המדויקת של המוח האנושי איננו העיקר. אם ענייננו הוא השפעת הבינה המלאכותית העתידית על עולמנו, אזי יכולתה להתנהג באופן אנושי היא עניין משני. אלן טיורינג (Turing), מייסד תחום הבינה המלאכותית, הבין זאת למן ההתחלה: באותו מאמר המתאר מבחן לבינה (ראו גלילאו, גיליון ינואר 2012, המוקדש לאלן טיורינג), שבו מחשב נבחן ביכולתו לחקות בני אדם, אמר טיורינג שחיקוי מוצלח הוא רק אמת־מידה נוחה לזיהוי הבינה בהיעדר מדדים מדויקים יותר;

אבל הגדרה נכונה יותר תתמקד ביכולות של הישות להשיג מטרות רצויות באופן כללי, ולא רק את משימת החיקוי.

הצגה נכונה של בינה היא הגדרתה כיכולת להשיג מטרות מגוונות בסביבות מגוונות, עם דגש על מטרות המשפיעות עלינו ועל הסביבות הנוגעות לנו. מבחינה מתמטית בינה היא כוח מיטוּב (אופטימיזציה) במובן הרחב, כלומר מירוב (מקסימיזציה) של פונקציות תועלת כלשהן, פונקציות של מצב העולם. תפישה כזאת את הבינה היא הדרך הנכונה לצפות באופק של האפשרויות העומדות לפנינו בעתיד: ייתכנו בינות אל־אנושיות שלא בהכרח דומות לנו. יהא זה נכון במיוחד לבינות המבוססות לא על חיקוי פעולת המוח, אלא על אלגוריתמים ומבני נתונים שתוכננו ופותחו במיוחד לצורך זה, שיטה המכונה "בינה כללית מלאכותית" (Artificial General Intelligence).

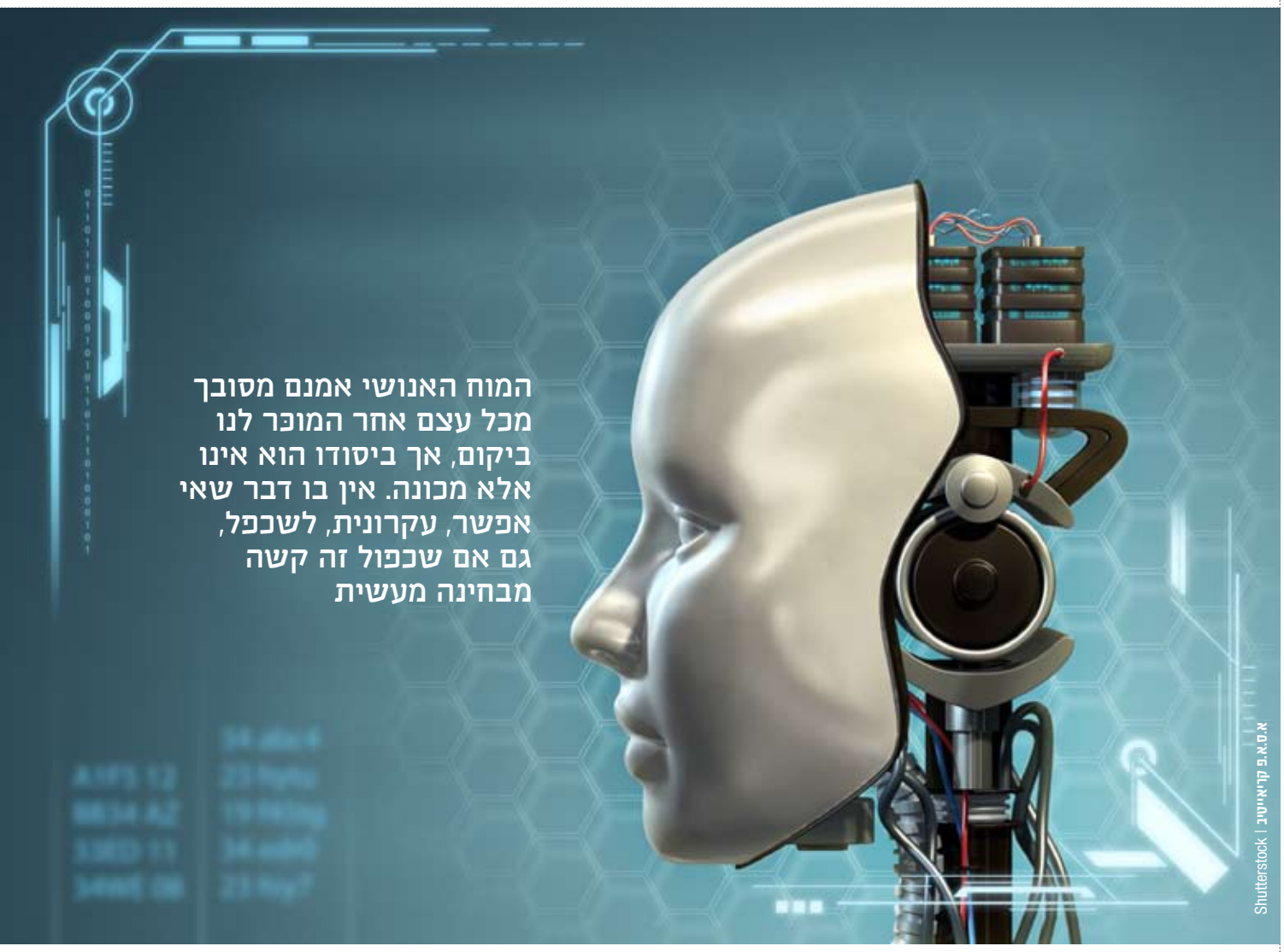
לכל בני האדם מבנה שִׂכלי די דומה, זאת עקב הרקע האבולוציוני שלנו. מיזוג הגֶנים ברבייה מינית אינו מאפשר שוֹני גדול בין בני אותה אוכלוסיה. אולי נדמה לנו שמסביבנו גיוון קוגניטיבי בקרב אנשים שונים, אבל צריך לזכור שאף אחד מאיתנו לא ראה בינה ברמה אנושית שאיננה אנושית. כשאנחנו מנסים לדמיין התנהגות של בינה אחרת, חייבים לוותר על האינטואיציות שלנו בנושא. ההתמקדות ב"מיטוב" ולא במונח השגור "בינה" עוזרת לנו לברור את ליבת המושג מן ההנחות הסמויות המבוססות על ניסיוננו עם בני אדם.

בינה כללית מלאכותית

במאמר זה אתמקד לא בדימויי מוח עתידיים, אלא בבינה כללית עם ארכיטקטורה מלאכותית. שתי סיבות לכך: ראשית, המאפיינים של חיקויים למוח מובנים לנו, כי אנו מכירים את התנהגות בני האדם. ואולם בינה כללית המבוססת על יסודות תיאורטיים פותחת בפנינו אפשרויות רבות נוספות שכדאי לחקור. שנית, לשיטת הבינה הכללית המלאכותית יתרונות שמאיצים את פיתוחה במרוץ לבנות בינה כללית, לעומת מבנים ביולוגיים מסובכים וקשים להבנה. כך למשל, המטוס הראשון נבנה על יסודות אווירודינמיים עם השראה כללית מציפורים, ולא כחיקוי מדויק לציפור.

האם ייתכן שמהנדסים יבנו בעשורים הבאים בינה אל־אנושית, בין אם תהיה דמוית־מוח או מושתתת על יסודות תיאורטיים? לפי תחזיות מסוימות (וראו: קורצוויל, לקריאה נוספת) היעד יושג עד לשנת 2030, ויש מגמות המעידות על אפשרות כזאת.

>>



א.ס.א.פ קריאייטיב | Shutterstock

הכוח החישובי ש־1,000 דולר (צמוד מדד) יכולים לקנות עולה מעריכית עם השנים וגם האלגוריתמים משתפרים. למשימות מסוימות בבינה מלאכותית צרה (זאת המוכרת כיום), שיפור האלגוריתמים תורם להאצת מהירות החישוב יותר מהשיפורים שבחומרה. נוסף על כך מדענים לומדים כיצד המוח עובד: טכנולוגיות הסריקה של המוח משתפרות גם הן מעריכית עם השנים, וכיום נוירולוגים יודעים רבות על אודות תפקודם של חלקי מוח מסוימים, אם כי המנגנון שבונה מחשבות מתהליכים חשמליים בתאי עצב עדיין אינו מובן. כל הנתונים האלה מרמזים שבינה כללית אל־אנושית תיבנה בתוך כמה עשורים.

אבל האצת הטכנולוגיה אינה מספיקה וגם אינה הכרחית ליצירת בינה כללית אל־אנושית. כדי לבנות ישות כזאת יידרשו פריצות דרך מדעיות. פריצות אלה תלויות בהברקות ובעבודה קשה של קומץ מדענים ומהנדסים, ולא רק בהלך הרוח בעולם המדעי. אי אפשר לנבא מתי יגיעו פריצות הדרך הללו, אבל בין אם יגיעו בעוד שנים מעטות ובין אם יתמהמהו מאות שנים, שומה עלינו לעסוק בקפיצת הדרך מרמת הבינה האנושית עד לרמת בינה אל־אנושית, בגלל השלכותיה המכריעות.

הגדרת הבינה כיכולת להשגת מטרות נראית מופרכת לאור הדוגמה האנושית. הרי בני האדם נדחפים על ידי צירוף של

רצונות רבים, בלתי־נהירים לתודעה, ומשתנים תדיר. מצב זה נובע מכך שמטרותינו נוצרו ככלי עזר ל"מטרת" האבולוציה, דהיינו ריבוי העותקים של גנים – לפיכך בני האדם שואפים לאכול, לצבור מעמד חברתי, להזדווג, לשרוד ועוד. כמובן, בני האדם אינם משרתים את האבולוציה במודע; נוסף על כך, עקב התמורות בסביבתנו באלפי השנים האחרונות מטרות בני האדם כבר חרגו ממטרות האבולוציה – אפשר לראות זאת למשל באמצעי מניעה ובהשמנת יתר. אין זה מפתיע שהבליל המסובך של רצונות, שכיום כבר אין לו כיוון מאוחד, פועל בצורה מסורבלת ובלתי ברורה. ברם, ככל שבליל הרצונות הזה מסובך, הוא עדיין מהווה מערכת מטרות: בני האדם אינם פועלים בצורה אקראית, אלא שואפים לכיוונים מסוימים, שדומים מאוד אצל כל בני האדם.

מטרותיה של בינה מלאכותית

ייתכן שבינה מלאכותית תשאף גם היא לבליל מסובך ומתוסבך של מטרות, במיוחד אם היא בנויה על דוגמת הבינה האנושית. אבל אם תיבנה על ידי מהנדסים הרוצים לפתור בעיה מסוימת, מערכת המטרות של הבינה המלאכותית עשויה להיות פשוטה יותר, כמו "הגדילי בצורה מרבית את הכסף בחשבון ההשקעות שלי", "נַצחי את אויבינו במלחמה", או "מצאי שיטה למניעת מחלת הסרטן".

הבינה שתיווצר, ביום מן הימים, תרצה לשפר את היכולות שלה. זאת מפני ששיפור יכול לתרום להשגת מטרות היסוד של הישות, וכפי שהגדרנו, הישות תיבנה כך שתעשה כל שביכולתה כדי לחתור למטרה בצורה היעילה ביותר. רוב הסיכויים שהיא גם תוכל לעשות זאת. בינה מלאכותית המבוססת על שבבים תוכל להוסיף עוד שבבים, עוד זיכרון, עוד מחשבים לחישוב מקבילי; או לנתח את קוד המקור שלה ולשכתבו. ישות ממטבת אל־אנושית תוכל לבנות העתקים של עצמה, וגם לבנות ישויות מִדור חדש, שאינן דומות לה אלא בכך שהן תורמות להשגת אותה מטרה. כל אלה הן טכניקות להשגת מטרות שאינן אפשריות לבני האדם – אנו יכולים לצבור ידע וללמוד כישורים לצורך שיפור עצמי, אך לא לשנות את מבנה מוחנו. וכשאנו, בני האדם, מעמידים דור חדש, הוא אינו פועל לממש במדויק את כל רצונותינו!

לפי הנחה שמהנדסים יביאו את הבינה לרמה אנושית, אין סיבה שהשיפור יסתיים דווקא בנקודה השרירותית הזאת. בני אדם ניחנו ברמת הבינה המזערית המסוגלת ליצור ציוויליזציה. אין שום ייחוד ברמה הזאת; היא אינה דווקא נקודת עצירה הכרחית. כל סבב של שיפור יוסיף עוד יכולות עד שהבינה תגיע לרמות הרבה מֵעבר לרמתנו.

לבינת־על סיכוי גבוה להגשים את מטרותיה. אם נרצה שהיא תפעל לטובת האנושות ולא לרעתנו, היא צריכה לרצות בכך. לא נוכל לעצור בינה הנבונה מאיתנו, לא לכלוא אותה במחשב, וגם לא להכתיב לה חוקים שיכבלו את התנהגותה. היא תעקוף בקלות, אם תרצה בכך, כל תכסיס שיעלה על דעתנו.

קשה מאוד להגדיר מטרות שהגשמתן תיטיב עם האנושות. מערכת הרצונות שלנו מגוונת, ואי־הגשמה של חלק קטן מן הבליל הזה תמיט אסון על האנושות. בתור ניסוי מחשבתי אפשר להעלות דוגמה של בינת־על השואפת למטרה "שלום על העולם", מטרה אנושית נשגבת. והנה, היא יכולה לעשות זאת ביעילות... על ידי הכחדה ברגע של המין האנושי (תוצאה מאוד לא רצויה!). אם מטרתה למקסם את התכונה של האהבה לזולת בעולם, היא יכולה להגשים זאת על ידי שינוי מוחנו ונטילת יכולות שכליות, כך שלא נרגיש דבר אלא אהבה חמה, כמו אהבת הארנבת לגוריה, ובתוך כך גם ליטול מאיתנו את היוזמה והאמביציה, שלפעמים מביאות איתן גם ריב ומדון.

אנו תופשים מטרות כמו "שלום" ו"אהבה" ברגשות עזים וחיוביים, ולכן בעייתי להתרכז בניסויים מחשבתיים כאלה. ואולם מטרות נטולות רגשות גם הן מסכנות את עתיד המין האנושי. הדוגמה הסטנדרטית היא מהדקי נייר. ישות ממטבת חזקה עם מטרה "הגדילי עד כמה שתוכלי את כמות המהדקים שברשותך", תנקוט באמצעים שונים לריבוי כמות המהדקים. שיפור היכולת העצמית תעזור לה במשימתה. כשתגיע לרמה על־אנושית, הגבוהה בהרבה מרמתנו, אנחנו בני האדם כבר לא נוכל להשפיע לכאן או לכאן על השגת המטרה. המרת כל החומר של כדור הארץ למהדקים תסייע לה במטרתה. צריכת כל המשאבים החיוניים לנו, כמו החומר של כדור הארץ, תביא להכחדת המין האנושי. הפעולה הזאת אולי לא נראית נבונה: היא לכאורה מתאימה יותר לאיוולת־על מאשר לבינת־על. אבל כאמור, אנו עוסקים בבינות לא במובן האנושי, אלא בממטבים, ישויות המגדילות למקסימום פונקצית תועלת מסוימת. הפונקציה "כמות המהדקים", ולמעשה רוב פונקציות התועלת האפשריות, אינן מיטיבות עם בני האדם.

א.ס.א.פ קריאייטיב | Shutterstock

הבינה שתיווצר תרצה לשפר את היכולות שלה, עד שתגיע לרמות גבוהות בהרבה מרמתנו. בתור ניסוי מחשבתי אפשר להעלות דוגמה של בינת־על השואפת למטרה של "שלום עולמי". היא יכולה לעשות זאת ביעילות על ידי הכחדה של המין האנושי... לבני האדם לא תהיה יכולת לעצור אותה

שלנו – והוא אמנם שימושי להשגת מטרות אחרות, אך מהנדסי העתיד לא חייבים לכלול אותו במערכת המטרות והדחפים של הבינה המלאכותית, והוא לא חייב להיווצר אלא אם יש לו תועלת. הנזק האפשרי הוא תוצאת מִשנה של מיטוב מוצלח לטובת מטרות שנציב לבינה המלאכותית, אם לא נקפיד על מערכת מטרות שכוללת את כל מה שבאמת חשוב לנו (וראו לקריאה נוספת: I. Beniaminy, 2007). ויתור על אהבה לסוגיה, חופש, בריאות, פעילות גופנית, למידה, אסתטיקה, יצירתיות, סקרנות, או כל מטרה אחרת מבין כל מטרותינו הרבות, יביא לעולם שבו לא נרצה לחיות.

שוברי אינטואיציה

קשה לנו להתנתק מהאינטואיציות המושתתות על הדוגמה האנושית, היחידה הזמינה לנו. אבל כמה ממטבים חזקים אל־אנושיים מופרים לנו, ובעזרתם אפשר לראות שתכונות המוח האנושי אינן הכרחיות להשגת מטרות.

אבולוציה היא כוח מופשט הממטב את הִתרבות הגנים באוכלוסיה מסוימת בסביבה נתונה באמצעות ברירה טבעית. לאבולוציה אין התגלמות כלשהי, אין לה תודעה, אין לה יכולת דימוי (מידול), אין לה זיכרון או כוח ניבוי; אבל אבולוציה היא על־אנושית בתחומים מסוימים: היא יצרה נמרים, אורנים ובני אדם. האבולוציה אינה רק על־אנושית ואל־אנושית אלא גם בלתי־אנושית: היא יצרה הזדקנות, מחלות, טפילים ועוד סוגי סבל שהיו נראים כרֶשע נורא אילו אדם יצרם.

שווקים הם סוג אחר של ממטב חזק. שוק מגדיל את התועלת של כל המוכרים והקונים הפועלים בו. שווקים אמנם אינם מושלמים, כפי שנוכחנו בבירור במפולות הכלכליות בזמן האחרון, אבל הם נותנים תוצאה טובה יותר מקבוצה קטנה של בני אדם המחליטים החלטות כלכליות עבור כולם. לשווקים, כמו לאבולוציה, אין גוף, אין תודעה, ואין חמלה או רגשות כלשהן, מֵעבר למה שיש לכל פרט המשתתף בשוק.

עוד סוג של ממטב על־אנושי הוא תאורטי בלבד: הממטב המושלם. דמיינו ישות שמשיגה באופן מיידי, בצורה ניסית, את כל מבוקשה. ישות כזאת אינה מוגבלת על ידי תכונה אנושית כלשהי, כמו מוסר או רגשות. המודל המתמטי AIXI (הכינוי מורכב מראשי תיבות של Artificial Intelligence והאות היוונית X, "כי", המייצגת את הפונקציה המודדת בינה) מתאר ישות מופשטת, שעוברת סבבים של אינטראקציה עם

הסביבה, ומקבלת קלט וגם מדד של תגמול, ובוחרת פעולה. AIXI בוחן בכל סבב כל אלגוריתם אפשרי ובוחר את זה שממקסם את צפי התועלת. מאחר שיש אינסוף אלגוריתמים אפשריים וחישוב התועלת המרבית אינו בר־חישוב (במכונת טיורינג או בכל מחשב אחר), AIXI אינו ישים, אבל הוא מייצג מודל תאורטי למושג "בינה מרבית", וכמובן, אין בו תכונות אנושיות כלשהן.

מן הסתם, בינה כללית מלאכותית, כשתבוא, לא תדמה לדוגמאות האלה. אבל הפנמה של מאפייני ממטבים אל־אנושיים כאלה עוזרת לנו להבין את האפשרויות הרבות לישויות עתידיות שאינן קיימות בהווה.

בינה כללית מלאכותית שלא תחסל כל אפשרות לעתיד אנושי טוב ומועיל, נקראת "בינה מלאכותית ידידותית" (הביטוי הוא מונח טכני, המתייחס לבינה שמיטיבה עם האנושות, ולא לידידות במובן הרגיל). כדי להשיג בינה מלאכותית ידידותית יש ליצור בינה כללית מלאכותית ראשונית שמקיימת שני תנאים: א. מטרותיה צריכות להיטיב אם האנושות; ב. היא צריכה לשמור על מטרותיה, ולא לשנות אותן, במהלך השיפורים העצמיים שלה (וראו לקריאה נוספת: Bostrom & Yudkowsky, 2012).

לצורך שמירה על המטרות תהיה לנו בעלת ברית חזקה: הבינה עצמה. הרי שינוי המטרה הראשונית מוריד את ההסתברות להשגת אותה מטרה, ולכן בינה חזקה תילחם נגד כל שינוי במטרות שלה (בני אדם לפעמים משנים את מטרותיהם, אבל אנו חלשים כמשיגי מטרות לעומת בינת־על).

קשה מאוד להגדיר מטרות התואמות את רצונם של בני האדם, אבל לפחות רצונות אלה טמונים במאגר ידוע – מוחנו. צריך להגיע לשקלול של הרצונות האמיתיים של כל בני האדם, ולקחת בחשבון שהרצונות הללו, אפילו בראשו של אדם בודד, מבולבלים וסותרים אחד את השני. אתגר אדיר, אבל גם כאן הבינה המלאכותית תוכל, אם תתוכנן היטב, לעזור לנו בחידוד המטרה "להיטיב עם האנושות", כל עוד נדאג לכך שהיא תבין שהמטרה הראשונית שלה אינה ברורה דיה. היא כמובן לא תשנה את המטרות שלה, אבל עשויה לעזור להבהיר מה אנחנו, בני האדם, באמת רוצים ממנה.

המחקר כיום

המחקר בנושא בינה כללית ידידותית נמצא בראשית דרכו. שני האתגרים למחקר הם הגדרת מערכת מטרות נכונה, ועיצוב הבינה כך שתשמור על מטרותיה תוך שיפור עצמי. לצורך כך, חוקרים צריכים לשלב ידע מתחומים מגוונים כמו תורת ההחלטות, תורת המידע, לוגיקה פורמלית ופסיכולוגיה אבולוציונית. כיום, שני מרכזי המחקר המובילים בתחום הם המכון לעתיד האנושות באוניברסיטת אוקספורד ומכון הייחודיות לבינה מלאכותית (וראו: לקריאה נוספת Singularity Institute) בעמק הסיליקון. אנו עדים לעניין גובר בנושא, גם בגלל ההיבטים התאורטיים המסקרנים, וגם בגלל ההשלכות על עתיד האנושות. בשנים 2011–12 התחום קם לתחייה: הפילוסוף האוסטרלי הידוע פרופ' דוד צ'אלְמֶרס (Chalmers) פרסם מאמר בנושא, שהביא בעקבותיו מהדורה מיוחדת של *Journal of Consciousness Studies*, וקובץ מאמרים בעריכת פרופ' אמנון עדן (Eden) ועמיתיו, העומד לצאת לאור בשנה הקרובה. ❖

יהושע פוקס עובד בחברת יבמ, שם הוא מוביל את הפיתוח של מוצר מוצרי תוכנה. לפני כן עבד כארכיטקט תוכנה בכמה חברות היי־טק בארץ. הוא בעל תואר דוקטור בבלשנות שמית מאוניברסיטת הארוורד ותואר ראשון במתמטיקה מאוניבר־סיטת ברנדייס. נוסף על כך הוא עמית מחקר של מכון הייחו־דיות לבינה מלאכותית. קישורים למאמריו זמינים באתר שלו: joshuafox.com

לקריאה נוספת:

Israel Beniaminy. 2007. Don't burn the cat. The Future of Things: http://bit.ly/HGNPyM

Nick Bostrom & Eliezer Yudkowsky, 2012, The Ethics of Artificial Intelligence To appear in The Cambridge Handbook of Artificial Intelligence, eds. W. Ramsey and K. Frankish (Cambridge University Press).
http://bit.ly/i3aqZM
Friendly AI: http://bit.ly/vzShJj

Ray Kurzweil. 2005. The Singularity is Near: When Humans Transcend Biology. New York: Viking Press.
Singularity Institute: http://bit.ly/ktOIF