



92469439-6	15/12/2024	הארץ - כותרת	עמוד 8	1	34.52x15.37
אוניברסיטת תל אביב מחקרים של אוניברסיטת תל אביב - 24634					

אחרי שעתיים, מודל AI חיקה התנהגות של אדם בהצלחה של 85%

המודל שפיתחו חוקרים מסטנפורד ומחברת DeepMind יצר העתקים מלאכותיים של בני אדם על סמך שיחות. המטרה: לייצר קבוצות מדגם שמבוססות על נתוני אמת

דעון לב

אישיות של המרואינים ולייצר מעין רפליקות שלהם.

כדי לבחון עד כמה כל רפליקה אכן דומה לאדם שאותו במחקר עברו אחרי הריאיון סדרת מבחני אישיות, ענו על שאלונים דמוגרפיים ושיחקו במשחקי מחשבה. הם עשו זאת פעמיים, בהפרש של שבועיים. ההעתקים הממוחשבים עשו את אותו הדבר, ואז השוו החוקרים בין התשובות וההתנהגויות. נמצא שההעתק המלאכותי השיב ופעל כמו המקור האנושי ב-85% מהמקרים בממוצע.

ראש צוות המחקר: "במגפה עתידית נוכל לבדוק אילו התערבויות וכללים יהיו הכי יעילים"

את הניסוי ואת המודל הממוחשב פיתחו חוקרים מאוניברסיטת סטנפורד ומחברת DeepMind של גוגל. תוצאות המחקר פורסמו במאגר המאמרים המקדימים arXiv,

לפני שעברו שיפוט עמיתים. ההעתקים לא נועדו בשלב זה לשימושם של בני האדם שאותם הם מחקים, אלא למטרות אחרת לגמרי: ייצור כמות גדולה של דמויות מלאכותיות (רפליקנטים), כך שישמשו חוקרים וקובעי מדיניות לצורך עשיית ניסויים שקשה מאוד או כלתי אפשרי לבצע עם אנשים אמיתיים. ההנחה היא שאם ניהיה לייצר מערך של אלפי רפליקנטים, שמבוססים על מדגם מייצג של האוכלוסייה ופועלים במידה רבה בדומה לאנשים אמיתיים, הרי שניתן יהיה לבדוק עליהם מגוון התערבויות. כך למשל, ניתן יהיה לבדוק כיצד האוכלוסייה מגיבה לשינוי מדיניות מסוים או מהו האפקט של וריאציות שונות של פרסומת למוצר כלשהו. במקביל לפרויקט זה, ישנן גם חברות הונק המנסות לפתח "סוכנים" שישרתו את בני האדם שאותם אומנו לחקות. "תאומים דיגיטליים" שכאלה יוכלו למשל לתכנן חופשה, כולל הזמנת כרטיסי טיסה ובתי

זה היה ההעתק שלי

ראש צוות המחקר, תלמיד הדוקטורט ג'ון סונג פארק, אמר ל"הארץ" שהכוח של סוכנים מלאכותיים מבוסס על הידע שצבר מודל ה-AI על התנהגות אנשים מתוך מאגר הנתונים העצום שהוון אליו. "הסוכנים הללו יכולים להפיק התנהגות דומה לזו האנושית כאשר מציגים להם תיאור של סביבה מסוימת

מלון, או אפילו לצאת לדייט עם זה וכך לחסוך מפחי נפש. ניסיונות לייצר סוכנים כאלה נעשים כיום תוך עיבוד כמויות מידע עצומות, למשל: כל תיבת האימייל או החשבונות ברשתות החברתיות של אדם מסוים. המחקר החדש מציע דרך פשוטה יותר: באמצעות ריאיון בן שתיים בלבד. לצד זאת, הוא מעורר חששות מפני פרצות אבטחה והפעלה מרחוק של העתקים ללא ידיעת האדם המקורי וייצור העתקים של אנשים ללא הסכמתם, בדומה לדיפ-פייקס (deepfakes), שנוצרים על ידי מחוללי תמונות ואודיו.

בשילוב עם פרסונה מסוימת, אמר. "בעבודה שלנו, חיוקנו את היכולות של הסוכנים בעזרת ראינות עומק עם בני אדם, כך שיוכלו לחקות את ההתנהגות והעמדות של אנשים אמיתיים". לדברי ג'ון, מטרת הפרויקט היא לתת כלים חדשים לקובעי מדיניות. "דמיין למשל תסריט עתידי של מגפה. אם נוכל להריץ סימולציות מדויקות של אנשים בודדים ושל חברות שלמות כדי לבדוק אילו התערבויות וכללים יהיו היעילים ביותר, זה יהיה צעד חשוב קדימה", אמר. בקשר ליצירת סוכנים אישיים פרטיים, הוא העריך שבעתיד לכל אדם יהיו כמה העתקים שיעשו עבורו מגוון מטלות, שבהן יחליטו ויפעלו ההעתקים בדומה לאופן שבו האדם עצמו היה פועל.

ישנן גם חברות הזנק המנסות לפתח "סוכנים" שישרתו את בני האדם שאותם אומנו לחקות

"זה בעצם מחקר מסורתי במדעי החברה וכאן עולה הצעה להחליף אנשים, שקשה להגיע אליהם ושיש להם זמן מוגבל, בהרבה סוכנים שאפשר לשאול אותם אינספור שאלות", אמר עם זאת, היעד של החוקרים עדיין רחוק ממימוש. ההתאמה של 85% אינה מרשימה במיוחד, ציין טוך. "מאוד ייתכן שאם היו

מראיינים אלף אנשים, ואז לוקים חים אלף אנשים אחרים שיעשו את ההערכות והבחינות, היו מגיעים לתוצאה דומה".

לדברי טוך, בתעשיות שונות כבר נעשה כיום שימוש בסימולציות של אנשים. "למשל אם רוצים לבדוק ממשק משתמש של אפליקציה לסמארטפון, אפשר לתת אותה לאנשים, אבל אפשר גם לעשות סימולציה כזו", הוסיף וסיפר על מחקר שעשה עם פרופ' עירד בן גל ומאיה בן-ארוש, שבו הם יצרו סימולציה של תנועת בני אדם בישראל. "בעזרת הסימולציה הזו אנחנו יכולים לסייע לקבל החלטות כמו איפה כדאי למקם תחנת אוטובוס או איך לתכנן מרחבים אורבניים".

פרופ' יואב גולדברג מאוניברסיטת בר אילן אמר שההתקדמות המתוארת במאמר החדש עשויה לאפשר בעתיד ללמוד על התנהגות של אוכלוסיות שלמות בצורה פשוטה בהרבה יותר. "הרבה יותר פשוט לייצר משהו שעונה נכון בממוצע על אלף אנשים, מאשר כלי שפועל על כדור הארץ כמו אדם ספציפי".

סימולציה עם אוטורים", אמר. "אחד הדברים החשובים הוא שניתן לנסות על הסוכנים כל מיני וריאציות. את זה אי אפשר לעשות עם אנשים, כי עם אדם אמיתי אתה לא יכול לשאול גרסה אחת שונות של אותה שאלה ולקבל תשובה נקייה".

גולדברג ציין שמכיוון שהסוכנים המלאכותיים מושתתים גם על הידע הקודם של המודל, ולא רק על הראיונות, הרי שקיימת הטיה, שכן המודל מוון בעיקר בתכנים באנגלית מתרבות מערבית. "זה עבד ב-85% בארה"ב, על מי שבדקו", אמר. "אם היו עושים אותו דבר ביש"ראל או באפריקה אולי זה היה עובד אחרת או פחות טוב". גולדברג העריך כי טכנולוגיית החיפוש שמוצגת במאמר תגיע בתוך כמה שנים לידי הבשלה, אך הדגיש כי יצירת סוכנים אישיים שיוכלו לפעול כתאומים דיגיטליים של אדם היא עניין מורכב הרבה יותר. "הרבה יותר פשוט לייצר משהו שעונה נכון בממוצע על אלף אנשים, מאשר כלי שפועל על כדור הארץ כמו אדם ספציפי".