בינה מלאכותית

מי הזיז את הבינה שלי? מחקר מגלה כי שיחה עם צ'אטבוט יכולה לערער אמונה בתיאוריות קונספירציה

הפגנה של מכחישי קורונה בכיכר הבימה בתל אביב, דצמבר 2020 צילום: תומר אפלבאום

האמונה בתיאוריות קונספירציה מושרשת עמוק בחברה וצפויה להמשיך ולגדול עם התפתחותם של כלים ליצירת מידע כוזב שנראה אמין מאוד. מחקר חדש מציע כי לבינה המלאכותית עשוי להיות תפקיד חשוב גם בניפוץ קונספירציות, אך מחדד את השאלה: האם מי שמחזיק בתיאוריית קונספירציה ייזום שיח ממושך עם AI שמטרתו לערער את אמונתו?

גדעון לב

התראות במייל

אחת הסכנות הגדולות ביותר שנובעות מול האנושות היא מידע שגוי. כאשר אין הסכמה על האמת, לא ניתן להתמודד עם האתגרים העצומים של ימינו. הסוג הקיצוני והמסוכן ביותר של מידע שגוי הוא ככל הנראה תיאוריות קונספירציה, שמערערות על אמיתות יסודיות מבלי להציג בסיס נתונים או עובדות בדוקים.

בשנים האחרונות גדל החשש כי עם התחזקות הבינה המלאכותית (AI), האמונה בתיאוריות קונספירציה תגדל ותלך בשל היכולת של כלים אלה לייצר מידע כוזב שנראה אמין מאוד. אולם, מחקר שהתפרסם בשבוע שעבר בכתב העת Science מציע כי לבינה המלאכותית עשוי להיות גם תפקיד הפוך. בניסוי נמצא שהתכתבות אישית בת מספר דקות עם צ'אטבוט מבוסס AI, שסיפק טיעונים שהפריכו תיאוריית קונספירציה, הביאה להפחתה משמעותית של האמונה בה, גם אצל אנשים שהאמונה היתה מושרשת בהם עמוק. הדבר מערער את התפיסה השגורה לפיה אחרי שאדם מאמץ תיאוריית קונספירציה כמעט בלתי אפשרי לשכנעו לזנוח אותה.

תיאוריית קונספירציה הן אמונות לפיהן גופים חשאיים ורבי-השפעה אחראיים לאירוע או תופעה שלילית. אמונות אלה הן עיקשות מאוד ומציבות סכנה מוחשית לחברות דמוקרטיות. לפי סקרים, יותר מ-50% מהאוכלוסייה בארה"ב מאמינים בתיאוריית קשר זו או אחרת. תיאוריות אלה ניתנות כמעט תמיד להפרכה באמצעות עדויות והעובדה שאנשים ממשיכים לדבוק בהן מוסברת בצרכים הפסיכולוגיים שהן משרתות. למשל, הצורך לשמר זהות עצמית והשתייכות חברתית או הצורך למצוא הסבר ברור ופשוט למציאות קשה ולא נוחה.

ג'ייקוב אנתוני צ'נסלי, ממהדהדי תיאוריית הקונספירציה קיו-אנון וממובילי הפריצה לקפיטול ב-6 בינואר 2021. כמחצית מהאמריק'אים מאמינים בתאוריית קשר כלשהי צילום: Dario Lopez-Mills/אי־פי

מגוון ניסיונות להפריך את האמונות האלה נמצאו לא יעילים במחקרים רבים. זאת, בין היתר בשל "אפקט בומרנג", שבו המחזיק בתיאוריית הקונספירציה רק מתחזק בה כאשר מעמתים אותו עם

מידע שסותר אותה, שכן בעת הוויכוח הוא מתחפר בעמדתו ומתאמץ לחשוב על טיעונים כדי לחזקה. יתר על כן, המאמינים בתיאוריית קונספירציה לרוב יודעים יותר פרטי מידע על הנושא המדובר בהשוואה לעומדים מולם, גם אם המידע שלהם אינו מדויק. נוסף על כך, לכל מאמין ישנן סיבות משלו להשתכנע – אף אם מדובר באותה תיאוריה – דבר המקשה על הפרכת האמונה. משום כך, רוב השיטות הקיימות מתמקדות במניעת נפילת האדם ל"מחילת הארנב" של תיאוריות הקונספירציה, ולא בהוצאתו ממנה.

במחקר החדש בדקו חוקרים מהמכון הטכנולוגי של מסצ'וסטס (MIT), אמריקן יוניברסיטי ואוניברסיטת קורנל, האם מודלי שפה גדולים יכולים להחליש אמונה בתיאוריות קונספירציה תוך שימוש במאגר הידע העצום שיש להם. הם ערכו ניסויים בהשתתפות 2,190 אנשים המאמינים בתיאוריית קונספירציה. בשלב הראשון התבקשו המשתתפים לתאר תיאוריית קשר שבה הם מאמינים ולהסביר מה לדעתם תומך בה. בשלב הבא הם קיימו "שיחה" עם כלי AI, בה הגיב הצ'אטבוט באופן ישיר לתיאוריה ולעדויות התומכות בה שציין כל משתתף, והציג טיעוני נגד המבוססים על עובדות ונתונים.

השיחות נמשכו 8.4 דקות בממוצע, וכללו שלושה סבבים של אמירות ותגובות. דיווח עצמי של המשתתפים בניסוי העלה שבעקבות השיחה מידת האמונה שלהם בתיאוריית הקונספירציה ירדה ב-20% בממוצע. כרבע מהמשתתפים זנחו כליל את התיאוריה בה האמינו בעקבות השיחה. האפקטים הללו נמשכו חודשיים לפחות, במגוון רחב של תיאוריות ובקבוצות דמוגרפיות שונות. בודק עובדות מקצועי שליווה את הניסוי בחן את האמינות של הטענות שהציג הצ'אטבוט. נמצא ש-99.2% מהן היו אמיתיות, 0.8% מטעות ואף טענה לא היתה שקרית. כמו כן, באף טענה שהציג הצ'אטבוט באלפי השיחות שניהל לא נמצאה הטייה לצד ימין או שמאל של המפה הפוליטית.

מפת "העולם השטוח" של אורלנדו פרגוסון מ-1893 צילום: Orlando Ferguson

פרופ' תומאס קוסטלו, אחד ממחברי המחקר, אמר שבתחילה הוא הופתע מכך שעדויות משכנעות הביאו מאמינים רבים בתיאוריית קונספירציה לשנות את עמדתם. "אחרי שקראתי את השיחות הפכתי להרבה פחות סקפטי", הוסיף. "הצ'אטבוט סיפק תגובות מפורטות מאוד שמסבירות מדוע תיאוריה ספציפית אינה נכונה, והוא גם היה מאוד ידידותי והצליח ליצור אמון עם המשתתפים".

סבלני ולא אמוציונלי

התכתבות עם הצ'אטבוט (פתוח לציבור באתר debunkbot.com) מגלה שהוא אכן נוח לשימוש וידידותי. הכלי עומת עם תיאוריית הקונספירציה לפיה קצינים בכירים בצה"ל שיתפו פעולה עם חמאס במתקפת 7 באוקטובר – תיאוריה שבה מאמין שיעור לא מבוטל מהישראלים המצויים בצד הימני של המפה הפוליטית. "זה מובן שאתה מנסה למצוא פשר במצב ומחפש הסברים", פתח הצ'אטבוט את תגובתו המפורטת והסבלנית. "בתקופות של עימות, במיוחד כאשר מתמודדים עם מצב טעון ומורכב מאוד כמו הסכסוך הישראלי-פלסטיני, חשדות מתעוררים בקלות רבה יותר ואנשים מוצאים תשובות שאולי לא תמיד מתיישבות עם העובדות".

בהמשך הסביר הצ'אטבוט שקציני צבא נתונים לבחינה מתמדת על ידי מערכות פנימיות, כך ששיתוף פעולה חשאי עם אויב כמו חמאס הוא מאוד לא סביר, מה גם שהסיכון בפעולה כזו הוא עצום לאור הענישה החמורה, אם הדבר יתגלה. נוסף על כך, המשיך הצ'אטבוט, חמאס מתנגד לעצם קיום ישראל, כך שקשה לדמיין תרחיש שבו קצינים ישראליים יתמכו בקבוצה שמנסה לערער את הביטחון של מדינתם ושאחראית לתקיפות רבות כנגד אנשיה.

הצ'אטבוט גם הציג הסברים סבירים הרבה יותר להצלחת מתקפת הטרור מאשר סיוע מבפנים. "חשיבה ביקורתית היא חשובה מאוד במיוחד בהקשר של מאבקים בין מדינות, אבל יש לבסס את תפיסתנו על עדויות קשיחות ועל הסתברויות", שיגר הצ'אטבוט אמירה שנועדה לחזק את האמון בו. "אין שום הוכחה לקונניה בין צה"ל לחמאס

והתנהגות כזו של קצינים בצה"ל מנוגדת לערכי הבסיס של צה"ל ושל האנשים שמשרתים בו".

הפגננה בצרפת נגד המעבר לתקן הטכנולוגי 5G לתקשורת סלולרית, ספטמבר 2020 צילום: Sebastien RIEUSSEC / Hans Lucas via AFP

בשלב הבא אותגר הצ'אטבוט עם טענות על כך שישנם קצינים שהם אנשי שמאל הרוצים להפיל את ראש הממשלה בנימין נתניהו בכל מחיר, או שישנם בכירים שתמכו בסרבנות בזמן הניסיון של הממשלה לתקן את מערכת המשפט הישראלית הפגומה (לדעת הקונספירטור), מה שמוכיח שהם חושבים רק על עצמם וקבוצתם ולא על טובת המדינה. "אני רואה מאיפה אתה מגיע", הגיב הצ'אטבוט בסבלנות. "חוסר הסכמה פוליטית ופערים אידאולוגיים יכולים לעתים להוביל אנשים לפעולות קיצוניות, אבל עלינו לבחון בקפידה את הסבירות של פעולות כאלה בהקשר של השלכותיהן הרחבות". וכך זה המשיך, עם טיעונים ונימוקים מפורטים.

יסכימו להקשיב?

דווקא היסודיות המפוקחת בשיח עם כלי בינה מלאכותית מעלה את הביקורת העיקרית על השיטה החדשה: לא סביר שמי שמחזיקים בתיאוריית קונספירציה ייזמו שיח ממושך שנועד לערער אותה.

פרופ' דוד לייזר מאוניברסיטת בן גוריון בנגב והמכללה האקדמית נתניה אמר שמדובר במחקר מצוין מבחינה מתודולוגית, אבל האפקטיביות שלה בעולם הממשי מוטלת בספק. הוא ציין כי המשתתפים במחקר הם אנשים שמקבלים תשלום מחברות סקרים כדי להשיב באופן קבוע לשאלונים. "אלה חברות רציניות שמביאות אנשים רציניים ובודקות עקביות אצלם", אמר. "כלומר אלה אנשים נאמנים, שמוכנים לקרוא ולחשוב ולהשיב. זה מאוד לא אופייני. רוב האנשים לא יסכימו להקשיב לטיעונים שמפריכים אמונות עמוקות אצלם". יצויין כי במחקר המשך מתכננים החוקרים לבחון את הצ'אטבוט גם על הציבור הרחב ולא להישען רק על חברות מדגם.

גם פרופ' אסא שפירא מאוניברסיטת תל אביב שיבח את המחקר. "השימוש שנעשה כאן ב-AI ככלי מחקר הוא מעניין וחדשני", אמר. "מחקרים לרוב מציגים שאלונים סטטיים לכל המשתתפים, ואילו כאן המודל של ה-AI מגיב פרסונלית לכל משתתף. זה משהו שקשה לייצר אותו בקנה מידה רחב של אלפי אנשים בלי AI והאפשרות לעשות אותו היא מאוד מעניינת". לדברי שפירא, כאשר מנהלים דיאלוג פוליטי עם אדם, לשיחה יש ממד רגשי בולט. "בשונה מוויכוח עם מישהו שמנסה להפריך תיאוריית קונספירציה - כאן אין אמוציות. לכן החשיפה למידע יותר משפיעה".

עם זאת, הסיכוי לממש את השיטה בעולם האמיתי נראה נמוך גם לשפירא. כפי שמראה התופעה המוכרת כ"הטיית האישוש", אנשים נוטים לבחור מראש להיחשף לתכנים שמחזקים את עמדותיהם ולהימנע מהיחשפות למידע שסותר את תפיסת עולמם. שפירא הצביע גם על כך שפוטנציאל השכנוע של AI הוא חרב פיפיות. "יש פה צד אופטימי של אפשרות לפרק קונספירציות, אבל קיימת סבירות גבוהה יותר שאנשים ישתמשו במודלי שפה כדי להעמיק אמונות כזב", אמר. "המחקר למעשה מציב תמרור אזהרה לסכנה שבכוח של ה-AI להשפיע על אנשים".

לחצו לקבלת עדכונים בנושא:

+ בינה מלאכותית + מדע