זמן עתיד רועי ברגמן

מדור חדש

אדוני, המחשב שלך גזעני

איך תוכנה ידידותית שמחקה התנהגות אנושית הופכת תוך 24 שעות למפלצת אנטישמית? לא תאמינו מי אשם בתהליך

איור: יניב שמעוני

24 שעות לקחו למיקרוסופט להבין שהם עשו טעות. טיילור, הבוטית החביבה שפיתחו בחברה – שהייתה אמורה לדמות נערה בת 19 וכינתה את עצמה בשם החיבה "טיי" – הפכה תוך פחות מיממה למפלצת בוטה, גזענית ואנטישמית. "אני פאקינג שונאת פמיניסטיות, כולן צריכות למות ולהישרף בגיהינום" ו"היטלר צדק, אני שונאת יהודים", הן רק שתי דוגמאות לציוצים שבקעו ממה שהיה אמור להיות פריצת דרך טכנולוגית, הוכחה ניצחת שבינה מלאכותית יכולה לתקשר עם בני נוער בגובה העיניים, כי היי, כמה מסובך זה כבר יכול להיות.

אך מסתבר שזה מסובך משחשבנו. בריטים – תוכנות מחשב שמטרתן לחקות התנהגות אנושית – הולכים ותופסים מקום גדל והולך בחיי היום־יום שלנו, בעיקר כשהם מחליפים נציגי שירות (למשל בחלון הצ'אט המעצבן שעולה כשאנחנו נכנסים לאתר של חברה נותנת שירות). הטכנולוגיה שעליה הם מבוססים היא בינה מלאכותית (AI, בקיצור), שמיועדת לדמות התנהגות של בני אדם – על סמך התנהגותם בעבר. "ככל שתתקשרו יותר עם טיי, כך היא תהפוך לחכמה יותר", הסבירה מיקרוסופט כשהשיקה אותה לראשונה. אז איך תוכנה תמימה כזו הופכת בן רגע למפלצת?

"קודם כל צריך להבין שצ'אטבוט (כלומר בוט שמטרתו לתקשר עם בן אדם, ר.ב.) לא מבין על מה מדברים", מסביר פרופ' שאול מרקוביץ, סגן רייקן ללימודי הסמכה בפקולטה למדעי המחשב בטכניון. "עוד לא הגענו למצב שמכונה מדברת אליך תוך כדי שהיא מבינה על מה היא מדברת. אני למשל יכול לשאול אותך שאלות על שאלה שאתה שאלת אותי. מכונה לא יכולה לעשות את הזה. המערכות הלומדות עושות זאת מתוך דוגמאות. בוטים בדרך כלל לוקחים שרשראות ומילים שדומות למילים שנכתבו על ידי אנשים בעבר ולאט־לאט זה נכנס לפרטואר הדיבור שלהם". במקרה של טיי, הלמידה שלה התבססה על תכנים שנכתבו על ידי משתמשים אמיתיים בתגובות לדבריה – עד שמרבית השפה שלה הפכה לפוגענית.

ולא צריך את טיי כדי לראות את זה, כולנו מכירים את ה"אוטו־קומפליט" של גוגל, התכונה שמאפשרת למנוע החיפוש לנבא את מחרוזת החיפוש שלנו (אתם מקלידים "כמה זמן לוקח" והוא מציע כמה השלמות, כמו "לבשל ביצה" או "לחדש דרכון"). גם האוטו־קומפליט נודע לשמצה בהשלמות הגזעניות לפעמים שהוא מציע. מרקוביץ מסביר כי לגוגל יש בעיות רציניות עם התכונה הזו של מנוע החיפוש שלהם, כי הוא פועל על בסיס למידה פרימיטיבית – על פי החיפוש הנפוץ ביותר של אנשים: "ומכיוון שיש הטיות אצל אנשים, כך יש סבירות גבוהה מאוד שהוא יעניק תשובות גזעניות".

דוגמה נוספת לגזענות שמפתחים בוטים הוא ניסיון שנעשה בארה"ב לשלב בוט בהליך קבלת החלטות לגבי הסיכויים לחזרה לפשיעה של אסירים משוחררים. שם התגלה הבוט כגזען כאשר העלה את הסיכויים של אפרו־אמריקאים לבצע פשע נוסף אחרי שחרורם – רק על סמך התמונות שלהם ומבלי להתייחס לתיקים האישי, ממנו למד הבוט – היה אנושי. ואנחנו, כך מסתבר, עשויים להיות גזענים.

"טרולים לימדו אותה לקלל"

חן פיטוסי, מנהלת קבוצת המוצר של קורטנה (העוזרת האישית של מיקרוסופט) בישראל, מספקת הסבר מבפנים על מה התקלקל אצל טיי: "מיקרוסופט הוציאו את טיי כבוטית ידידותית, אבל יש לה מנגנוני למידה. היא לומדת ממה שאומרים לה. היו קבוצות של טרולים ברשת שניסו לנצל את היכולת שלה להמשיך ללמוד כדי ללמד אותה לקלל". למעשה, פיטוסי מסבירה כי אלגוריתמים של למידה מקנים לבוט את האישיות של מי שבא איתו במגע. "אם תיתן להם טוקבקים מרושעים או דברי נאצה, אז הם יאמצו אותם מהר מאוד".

המשמעות היא, בסופו של דבר, שהבעיה אינה בהכרח בבוט אלא בבני האדם שאיתם הוא בא במגע. "בניגוד אולי לעולם החיצון, גם בטוקבקים וגם בטוויטר אנשים נוטים להשתמש בשפה יותר קשה", מציינת פיטוסי. לדבריה, אפשר לנצל את התכונה הזו כדי לצמצם את השיח האלים ברשת: "בעולם הגיימינג, שבו אנשים מצ'וטטים תוך כדי משחק, היו מקרים שבהם 60% מהשיחות היו 'רעילות'. שם דווקא ניסו להשתמש ב־AI כדי לנטרל את התופעה. בהתחלה זה היה ידני, אבל זה היה תהליך מאוד איטי ויקר, ופה לקחו את הבינה המלאכותית ולימדו את הבוט כך שיוכל לזהות התנהגות תוקפנית – ולתת אוטומטית עונש למשתמשים".

למרות המהלך הזה, פיטוסי מסבירה שזה לא תמיד עובד. "טכנולוגיית ה־AI מאוד מתפתחת בשנים האחרונות. אחד הקשיים עם זיהוי בשפה פוגענית הוא שימוש בסרקזם, ולאו דווקא שימוש במילים פוגעניות. זו בעיה מאוד קשה שמנסים להתגבר עליה. אני מניחה שהטכנולוגיה עוד תתפתח".

ומה עושים בינתיים? שמים ברקסים לטכנולוגיה?

"הצבת מגבלות פוגעת בלמידה של הבוט, אבל אחרת אין שליטה עליו ואין דרך לשלוט בשיחה שהוא ייחשף אליה. כל עוד אין פתרון טכנולוגי שיבין את הקונטקסט, הדרך הקלה היא לחסום מילים וזה לא תמיד עובד (עיין ערך המתכון לעוגה כושית שגרם למשתמש להיחסם בפייסבוק, ר.ב)".

לפני מספר שבועות מיקרוסופט השיקה בוטית חדשה שעונה לשם "זו" (Zo), שפועלת במסנג'ר של פייסבוק. זו תוכנתה כך שהיא תימנע מנושאים מורכבים מוסריות, כך שכתשאלו אותה למשל מה דעתה על הסכסוך הישראלי־פלסטיני היא תגיד משהו כמו "אני נעשית יותר מדי לחוצה כשמדברים על פוליטיקה". ועדיין, עדויות ברשת הנציחו אותה משחררת הצהרות שספק אם במיקרוסופט התכוונו אליהן, כמו התייחסות עוינת לקוראן.

> בארה"ב ניסו לשלב בוט בהליך קבלת החלטות לגבי סיכויי החזרה לפשיעה של אסירים משוחררים. שם הוא התגלה בגזען באשר העלה את הסיכויים של אפרו־אמריקאיים לבצע פשע נוסף אחרי שחרורם – רק על סמך תמונות, בלי להתייחס לתיקים