שפת בינה מלאכותית נוסח פייסבוק, פייק או הגזמה?

ד"ר אברום רותם 2-08-2017

במאמרון זה נבדוק את העובדות אודות השפה המסתורית שהתגלתה בפייסבוק (יולי 2017) שאילצה אותם לסגור את מיזם הצ'טבוטים החכמים, ולצד זה לתאר את המקור האמתי לסיפור זה - מכון המחקר של גוגל (נובמבר, 2016) שמפתח מתרגם חכם. נמצא שבעוד בפייסבוק הסיפור לא משכנע, וודאי אינו כמו שהוא מסופר, בגוגל הממצאים מרתקים בהחלט.

בשלהי יולי 2017 יצאה פייסבוק בידיעה שהכתה גלים בכל כלי תקשורת שמכבד עצמו. הרשת, והתקשורת בשיח הציבורי-טכנולוגי הוצפו בידיעה מטעם צוות מחקר פייסבוק: "מערכת בינה מלאכותית שפותחה בפייסבוק יצרה שפה משלה. היא פיתחה מערכת של מילים לקוד כדי להפוך את התקשורת יעילה יותר. החוקרים **סגרו** את המערכת כאשר הבינו שהבינה המלאכותית שפתחו, לא משתמשת באנגלית (או כל שפה אנושית אחרת)[1]".

ההודעה פתחה צונאמי של הפלגות מדומיינות ואף מביכות, אודות שפת סתרים שמשמשת את הרובוטים, שההשלכות על כך תהינה אבדן שליטה על מכונות שיעשו בעולם כרצונן, לא פחות.

התחקות אחר העובדות

פייסבוק מפתחים צ'טבוט, סוכן/סייען שמחליף בני אדם כמסייע למשתמש שמחפש מוצר/ המלצה/ נתונים בתחום שירות מסוים ראו דוגמא לדיווח אודות ההקדמות של המוצר, דיווח (בלוג) ממרכז הפיתוח של בינה מלאכותית פייסבוק- FAIR בעניין[2], ללא תיווכים.

הרציונל העסקי של פייסבוק בפתוח מוצר זה סביר בהחלט: כלנו מכירים את מרכיב העזרה הלא יעיל במגוון אתרי שירות, בו אמנם ניתן להתקשר באמצעות צ'טבוט (מענה דיגיטלי מלאכותי) לקבלת תשובות לשאלות שעולות ו/או בקשה להכוונה יעילה לשימוש בשירות. מטרת המיזם של פייסבוק היא לשפר בהרבה שירות צ'טבוט שעד כה הוא למעשה מעין "שאלות שנשאלות בנושא", ולפתח רובוטים שיחקו סוכן אנושי, תוך אינטראקציה שפתית אנושית, שייתרו את בני האדם כסייענים בשירות הדיגיטלי. לאחר הפעלת הניסוי, כך מדווחים, הצליחו הרובוטים שפותחו במרכז המחקר של פייסבוק, לשכנע את המשתתפים האנושיים שהם מנהלים משא ומתן עם אדם אחר. חוקרי פייסבוק טוענים שהרובוטים "למדו לנהל שיחות שוטפות" סביב הנושא הייעודי. מאחר שאף אחד מן ההתנהגות שנצפתה לא תוכננה ישירות על ידי החוקרים, נראה שהרובוטים למדו כיצד לנהל משא ומתן בין בני האדם.

עד כאן, נשמע יפה. אך בנקודה זו מתחיל הסיפור: בשלב כלשהו החליטו החוקרים, במקביל לאימון מול מורה אנושי, לאפשר שיח חופשי ביניהם באמצעות אלגוריתמים של למידה ממוחשבת, כדי לחזק את כישורי השיחה שלהם באמצעות דיאלוג ביניהם, ובכך להעצים את אפקטיביות האינטראקציה גם

---

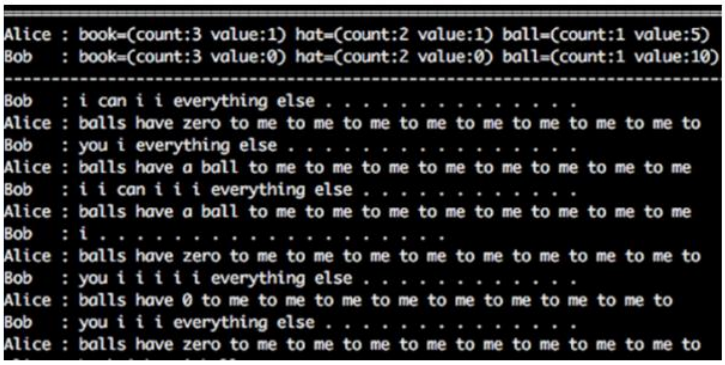
1 = JAMES WALKER (JUL 21, 2017) Researchers shut down AI that invented its own language. Digital Journal. http://www.digitaljournal.com/tech-and-science/technology/a-step-closer-to-skynet-ai-invents-a-language-humans-can-t-read/article/498142

= ADRIENNE LAFRANCE JUN 15, 2017 An Artificial Intelligence Developed Its Own Non-Human Language. *The Atlantic*. https://www.theatlantic.com/technology/archive/2017/06/artificial-intelligence-develops-its-own-non-human-language/530436/

2 Deal or no deal? Training AI bots to negotiate. *ARTIFICIAL INTELLIGENCE ·RESEARCH*. https://code.facebook.com/posts/1686672014972296/deal-or-no-deal-training-ai-bots-to-negotiate/

בנושאים עסקיים. אז התברר שהחלה נטיה להתקשר בשפה שאיננה אנגלית, או כל שפה אחרת שאינה מובנת לחוקרים.

להלן צילום מסך מקטע משיחה בין שני רובוטים בוב ואליס באדיבות מרכז המחקר Fair, פייסבוק[3]:



[Screenshot: courtesy Facebook]

לא נביך את הקוראים בפרשנות החוקרים, לפחות כפי שהיא מצוטטת, למשפט האלמותי של אליס: *[״שטויות (המילה **balls** – אשכים, היא שטויות בסלנג) הן עבורי אפס לי לי לי לי....״* או המשפט של הרובוט בוב ״את אני אני אני אני כל השאר...״] כביטוי עמוק לנושאים ונותנים מיומנים של בינה מלאכותית, ומה תכלית השיח, שביננו היא פשוט ג׳יבריש מוחלט[4]. החוקרים שפתחו מערכת זו טוענים (כך לפחות מצוטטים על ידי מתווכי תקשורת שלא ניתן לבדוק אמינותם), שמדובר במשא ומתן ״***ערמומי להפליא***״ ובעל משמעות עסקית עמוקה. הטענה היא, שבמשך הזמן הפכו הרובוטים למיומנים למדי, ואפילו התחילו להתעניין בפריט אחד כדי ״להקריב״ אותו בשלב מאוחר יותר במשא ומתן כפשרה כוזבת.

יש להודות שהקשר בין הדוגמא שלעיל, לבין פרשנות החוקרים לשיח המשונה, לבין שפה מסתורית בין רובוטים, מסתורי אף הוא, או במילים אחרות – **אין קשר של ממש**.

נוכל ללמוד זאת מדיווחים רשמיים של מרכז המחקר FAIR: החל מיוני נפסק הדיווח על התקדמות המערכת, ואחרים הודיעו בשם פייסבוק שהיא החליטה לסגור את המיזם. בהודעה על סגירתה נטען, שהחוקרים חששו שמתנהלת תחת אפם הבניית ידע ללא כל שליטה אנושית, שברור רק לרובוטים: ״***האינטרס שלנו היה שיהיו צ׳טבוטים שיוכלו לדבר עם אנשים***״ מצוטט ***מייק לוויס***, מדען חוקר ב-Fair- מרכז מחקר בינה מלאכותית פייסבוק. במילים אחרות: חששות מהיעדר שליטה על רובוטים שמתחילים לדבר בינם לבין עצמם בשפה לא מזוהה.

מה יש כאן באמת?

בדיקת העובדות מעלה תמיהות:

בפרסומי המרכז המחקר של בינה מלאכותית של פייסבוק ״FAIR״[5] כמו גם בבלוגים שלהם, **לא מצאנו** (נכון ל- 2-08-2017) כל דו״ח רשמי של פייסבוק על החלטה של סגירת המיזם, **ולא** על שפה מסתורית בין הרובוטים. ההפך, באפריל וגם יוני ישנם דוחות במרכז הפיתוח FAIR


This XML file does not appear to have any style information associated with it. The document tree is shown below.

<Error>
<Code>NoSuchBucket</Code>
<Message>The specified bucket does not exist</Message>
<BucketName>end-to-end-negotiator</BucketName>
<RequestId>0A71C1F5992D49B5</RequestId>
<HostId>
tjRCqEhON7MxslCSvF3h2xtLIdI+o8QDTjd9/M/Zjl+f5JwwSnjuDfzICSbKTUHdI/dC79dV2gM=
</HostId>
</Error>

3 BRYAN CLARK (Jun, 2017) Facebook's AI accidentally created its own language. *ARTIFICIAL INTELLIGENCE/* https://thenextweb.com/artificial-intelligence/2017/06/19/facebooks-ai-accidentally-created-its-own-language/

4 AI Is Inventing Languages Humans Can`t Understand. Should WE Stop it? 14,07,2-17. *Co-Design* https://www.fastcodesign.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it

5 https://research.fb.com/publications/page/2/?cat=13 פרסומי המרכז לחקר בינה מלאכותית פייסבוק

על אימון ולימוד של רובוטים לשוחח עם בני אדם, כולל שיח בין הרובוטים. הדו"ח מטעם FAIR אודות השפה המסתורית שהומצאה על ידי רובוטים של פייסבוק, (שמצוטט באתרים אחרים)[6], לא מופיע, כנראה **דו"ח מסוים הוסר מהרשת** (נכון ל- 2-08-2017), שעשוי היה להאיר נקודה זו. ראה הודעת האתר בעניין.

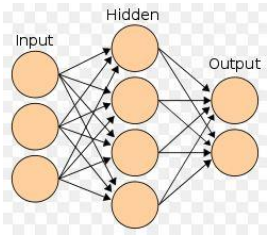
לאור בדיקה עניינית בסוגיה, נוכל לקבוע בזהירות ראויה, שעד שלא יוכח אחרת, הסגירה של מיזם הצ'ט-בוטים החכמים של פייסבוק היא פשוט הודאה אלגנטית, ירידה מהסולם, עוד כישלון של פייסבוק שהחליט לסגור את המיזם מטעמים משלו. כאן מצאו אילן להיתלות בו- שפת הסתרים המאיימת שפותחה על ידי הבינה המלאכותית, אך ללא כל תיעוד ציבורי לעניין.
כל השערה נוספת של משמעות שפת סתרים וחשש מהשתלטות הבינה המלאכותי – וודאי עד שאין עובדות אחרות רשמיות מצד פייסבוק, זה קשקוש-בלבוש, וכדאי להתנער מכך, ומהר.

לעומת זאת- מאין הרעיון של שפת סתרים של בינה מלאכותית?

במקרה שונה לגמרי, עוד בשלהי 2016, יצא דו"ח מחקר של Google (14-11-2017)[7] , בו דווח על שיפור שירות התרגום על ידי הוספת רשת עצבית. אכן התגלה שמערכת הבינה המלאכותית, המותאמת במיוחד למשימה של תרגום משפטים מילוליים, משתמשת בתהליך התרגום בסימנים/נתונים שפתחה בעצמה במהלך הלימוד של תהליך התרגום. דבר זה ניתן להסבר מעצם מבנה מערכת עצבית, שננסה להסביר כאן בקצרה, אך נקדים ונאמר שגם כאן ממש לא מדובר בשפת סתרים של יצירי בינה מלאכותית.

בדו"ח הנ"ל מתוארת תגלית מעניינת. הבעיה שדו"ח זה מובן רק ליודעי ח"ן (=חכמת נסתר) מהתחום של מערכות בינה מלאכותית, ולכן ננסה לתווך בקצרה מאד, על ידי הסתמכות של מתווכים מקצועיים מהתחום[8]:

1. מערכת עצבית מלאכותית ביישומים רבים שלה, היא מערכת לומדת, שניתן ללמד אותה, או שהיא לומדת בעצמה תוך כדי התנסות, אם מסמנים לה, כמובן, את המטרה- במקרה זה איכות התרגום הרצוי. לשם ההסבר בהמשך (ראו מושגים כלליים בויקיפדיה)[9], נציין שמערכת עצבית היא מתווה/ רשת של תאי עצב פשוטים יחסית, שהחיבור ביניהם ויכולת הקביעה של כל תא ותא של "משקלות"/ מידת חשיבות במספר, שהוא מייחס לאות הנכנס אליו יוצר מכונת למידה מופלאה. ככלל יש תאי כניסה, תאי תוצאה, ובינם מפרידה הרשת המרכזית, שמכונה "רובד סמוי" שנדבר בו בהמשך.



2. בפיתוח מערכת תרגום באמצעות רשת עצבית מתקבלות תוצאות מרשימות, אך החוקרים תהו איך באמת המערכת עובדת. לכן שמו בפניה את האתגר הבא: אם מלמדים את מערכת התרגום לתרגם אנגלית לקוריאנית ולהיפך, וגם אנגלית ליפנית ולהיפך - האם המערכת תתרגם ללא לימוד מקודם, בלי להזדקק לאנגלית כגשר ביניהן? ואכן מצאו שהמערכת שלמדה לתרגם בין אנגלית לקוראנית וליפנית, יוצרת תרגומים "סבירים" בין שתי שפות- קוריאנית ליפנית - שלא למדה לתרגם ביניהם, וכל זאת היא עושה ללא תיווך האנגלית.

---

6 כתובת דו"ח פייסבוק - https://s3.amazonaws.com/end-to-end-negotiator/end-to-end-negotiator.pdf
שנעלם מהרשת על יצירת השפה מסתורית

7 דו"ח המחקר המקורי https://arxiv.org/pdf/1611.04558v1.pdf

8 Devin Coldewey (Nov 22, 2016) Google's AI translation tool seems to have invented its own secret internal language. *TechCrunch*. https://techcrunch.com/2016/11/22/googles-ai-translation-tool-seems-to-have-invented-its-own-secret-internal-language/

9 רשת עצבית מלאכותית (ויקיפדיה)
https://he.wikipedia.org/wiki/%D7%A8%D7%A9%D7%AA_%D7%A2%D7%A6%D7%91%D7%99%D7%AA_%D7%9E%D7%9C%D7%90%D7%9B%D7%95%D7%AA%D7%99%D7%AA

3. ואז עלתה השאלה הבאה: האם המחשב מסוגל ליצור קשרים בין מושגים ומילים שלא נקשרו רשמית תוך כדי לימוד מסודר ומתוכנן קודם לכן, קשרים ברמה עמוקה יותר מאשר חיבור מילונאי (מילה או ביטוי המקביל לאחר)? במילים אחרות, האם המחשב **פיתח שפה פנימית משלו** כדי לייצג את המושגים שבהם הוא משתמש לתרגום בין שפות אחרות?
4. בהתבסס על האופן שבו משפטים שונים קשורים זה לזה במרחב הזיכרון של הרשת העצבית, חושבים החוקרים שאכן, המערכת פתחה לעצמה שפה משלה, או יותר מדויק, אוסף מושגים משלה, והיא המגשרת בין השפות בעת התרגום.
5. נראה ששפה כזאת, המוכנה "אינטרלינגואה"- ***interlingua*** [=לשון/שפה פנימית], קיימת ברמה עמוקה יותר של ייצוג הרואה קווי דמיון בין משפט או מילה בכל שלוש השפות. מעבר לכך קשה לומר, שכן התהליכים הפנימיים של רשתות עצביות מורכבות ולא ניתנות למעקב ולהבנה של ממש.
6. ההסבר לכך הוא שהאופן בו רשת עצבית לומדת, התהליך שעוברים התאים בשכבה הנסתרת, (ראה למעלה כאן סע' 1), וכיצד הערכים שמתקבעים אצלם, אינם גלויים ואינם ידועים למתכנתים. לכן כל מה שמתחולל שם הוא נסתר מן העין פשוטו כמשמעו, ולא ניתן לפיענוח והבנה. תהליך הלימוד הנו ספונטני במידה רבה, וללא שליטה של המתכנתים.
7. כל מה שניתן לומר הוא, שהעובדה **ששפה פנימית** כזו נוצרה וקיימת בכלל - יצירה מקורית של המערכת עצמה כדי לסייע בהבנת המושגים שהיא לא הוכשרה להבנה קודם לכן - היא מרתקת אפילו ברמה הפילוסופית. ימים יגידו מה נוכל להפיק מזה, כיצד נגרום למערכות בינה מלאכותית ללמוד לבד, ולאתר מקרים בהם מערכת בינה מלאכותית תבנה מידע באופן ספונטני ללא מעורבות אדם, ועד כמה נוכל להבין ולשלוט בכך.

הערות, הבהרות, תוספות ומחמאות יתקבלו בברכה.