

01 מאי 2023

"מן הסתם יש לבני אנוש נטייה להיצמד
בעקשנות לדעות ישנות, ויש להם התנגדות
עיקשת באותה מידה לחומר חדש העשוי
לערער הנחות אלה".¹

מקום ליד השולחן: האם ה-ChatGPT יכול "לשבור את הקונספציה"?

איתי ברון ותהילה שוורץ אלטשולר²

מבוא ותמצית

בערב יום הכיפורים מתכנסת הצמרת המדינית-ביטחונית של ישראל לדיון חירום סביב שולחנה של ראש הממשלה. ראש אמ"ן מציג את המידע ההתרעתי שהגיע בימים האחרונים. הוא פותח במידע ובהערכות שהעביר המלך חוסיין לראש הממשלה וממשיך בניתוח שורה של סימנים מטרידים נוספים: תרגיל מלחמה שהחל במצרים, שבמסגרתו נמצאים הכוחות בכוננות עליונה ובהיערכות למלחמה; ביטול חופשות וכניסה למערכי חירום בצבא סוריה; מידע מפורש שלפיו התרגיל הוא למעשה מסווה למלחמה ובסיומו תתקוף מצרים את ישראל; ביקור של שר המלחמה המצרי בסוריה; ומידע חדש שהגיע בלילה שעבר על פינוי כל היועצים הסובייטים ממצרים וסוריה.

¹ וולשטטר, ר' (1979). "לקחי פרל הרבור". **ציקלון** 4, 4-14, בעמ' 12.
² תא"ל (מיל.) איתי ברון היה ראש חטיבת המחקר של אמ"ן; ד"ר תהילה שוורץ אלטשולר היא מומחית למשפט וטכנולוגיה ועמיתה בכירה במכון הישראלי לדמוקרטיה.

בסיום הסקירה שואלת ראש הממשלה באופן ישיר: האם תהיה מלחמה מחר? את התשובה שהשיב ראש אמ"ן ב-5 באוקטובר 1973 אנחנו מכירים היטב, וכך גם את תוצאותיה הטראגיות: מקבלי החלטות הבכירים ביותר בישראל, בדרג המדיני ובדרג הצבאי, העריכו עד לפרוץ הקרבות ממש שהסבירות למלחמה בין ישראל לערבים היא נמוכה.

אבל, בסימולציה שערכנו לאחרונה, השיב ראש אמ"ן תשובה אחרת. הוא ענה לראש הממשלה שלפי המידע "מצרים וסוריה נערכות לפעולה צבאית וייתכן שתהיה מלחמה בטווח הזמן הקרוב" וכן "שהסבירות למלחמה גבוהה מזו של השלום". עם זאת, הוא הוסיף שלא ניתן "לחזות במדויק אם מלחמה תתחיל מחר או לא".

מי ששימש בתפקיד ראש אמ"ן בסימולציה הזאת לא היה קצין מודיעין וותיק ומיומן, אלא ה-ChatGPT של חברת OpenAI (להלן: GPT), שמטלטל את העולם מאז הופעתו בנובמבר 2022. השימוש בצ'טבוט הנגיש, בשיחה רהוטה שדומה מאוד לשיחה אנושית, מאפשר לרבים לחוות את ההתקדמות העצומה בתחום הבינה המלאכותית היוצרת (Generative AI) ומעורר שאלות לגבי יכולתה לבצע משימות חשיבה מתקדמות, שעד לא מזמן נתפסו בתחום היתרון המובהק של המוח האנושי (על "מערכות שפה גדולות יוצרות", ראו להלן בנספח). לשאלות אלו יש, כמובן, השלכות גם על פוטנציאל השימוש במוצרי בינה מלאכותית יוצרת בתחום המודיעין בכלל והמודיעין האסטרטגי בפרט.

הסימולציה על מלחמת יום הכיפורים היא חלק משלוש סדרות של סימולציות שערכנו לאחרונה במטרה לבחון את יכולתו של ה-GPT לסייע לייצר הערכת מודיעין כבסיס לקבלת החלטות אסטרטגיות בתנאי אי וודאות קשים. מדובר, ככל הידוע לנו, בהתנסות אמפירית ראשונה עם מודל שפה ענק כדוגמת ה-GPT בהיקף רחב בהקשרים של מודיעין אסטרטגי. כל הסימולציות בוצעו עם גרסת ה-3.5 של ה-GPT. על חלק מהסימולציות חזרנו גם עם גרסת ה-4.0 של המערכת באמצע מרץ 2023 והן חיזקו את המסקנות שאליהן הגענו באמצעות שימוש בגרסה המוקדמת יותר.

מטרתו של מסמך זה היא להציג את הסימולציות ולנתח את תוצאותיהן. כל הסימולציות עסקו ברמות המודיעין הגבוהות (מודיעין לאומי-אסטרטגי), בבסיס לקבלת החלטות בצמרת הצבאית והמדינית. הנושא המרכזי שנבחן בסימולציות היה יכולתו של ה-GPT לסייע להערכת המודיעין במרחב הזמן של ימים ושעות לפני תחילת מתקפת פתע או מהלך התקפי אחר.

מודלי שפה ענקיים כדוגמת ה-GPT כבר יכולים לבצע מטלות בסיסיות של איסוף, סיווג וניתוח מידע באיכות, במהירות וביעילות המשתוות לאלו של בני אנוש. בהקשר שלנו, ה-GPT הדגים יכולת להצביע על "אפשרויות מתחרות", הן כהסברים לתופעות שנצפות במידע והן כתרשימים לגבי העתיד. הוא גם הראה יכולת לסייע בהכוונת האיסוף. לתוצאה זו ישנה חשיבות שכן רבות מהפעילויות המחקריות השוטפות, בעיקר של החוקרים הזוטרים, הן מהסוג הזה. נכון לבחון את היכולות האלו בסימולציות נוספות שכן סביר שניתן יהיה לרתום את המכונות לביצוע משימות כאלו באופן שיוכל להפנות משאבים למשימות אחרות, מורכבות יותר.

ואולם, המסקנה המשמעותית יותר שעולה מהסימולציות שביצענו היא שלמערכת ישנן סגולות נוספות שנוגעות דווקא למשימות חשיבה מתקדמות. מסקנה זו עולה בקנה אחד עם מחקרים נוספים שפורסמו לאחרונה שטוענים שמודלי שפה ענקיים יכולים לבצע תהליכי חשיבה והיסק כאלו, אך בדרך שונה מזו של בני אדם.³ שונות זו מאפשרת העשרה של תהליכי החשיבה האנושיים.

נקודת המוצא לתיאור תוצאות אלו של הסימולציה היא הקביעה של וולשטטר שהבאנו בפתח המאמר אודות "הנטייה של בני אנוש להיצמד בעקשנות לדעות ישנות", היינו – לקונספציה. הממצא העיקרי שזיהינו בסימולציות הוא של-GPT יש יכולת גבוהה להצביע על סתירה בין קונספציה לבין מידע סותר, ללא ההטיות שמאפשרות וגורמות לחוקרים אנושיים למצוא דרכים לעמעם את הסתירה ולהיצמד בעקשנות למה שמוכר, מצופה ורצוי. כלומר, אין לו את אותה "התנגדות עיקשת ... לחומר חדש" שיכול לערער את הקונספציה, שאותה זיהו וולשטטר וחוקרים

³ המעניין והמקיף שבהם, שמתאר "ניצוצות" של חשיבה כזו, הוא: Sparks of Artificial General Intelligence: Early experiments with GPT-4. <https://arxiv.org/abs/2303.12712v4>. Bubeck, S' et al. (2023).

רבים אחרים.⁴ ההשערה שעולה מהסימולציות היא שהקונספציה ("דעת המומחים") נתפסת על ידי ה-GPT כפריט מידע נוסף ולא כזה שיש לו עדיפות על פרטי המידע האחרים. זאת, ככל הנראה, משום שדרך הניתוח שלו שונה מזו של חוקרים אנושיים שנזקקים למסגרת תפיסתית (קונספציה) כדי לברר את המציאות ולהבין אותה. לתכונה זו יכולה להיות, להערכתנו, חשיבות ככלי עזר להערכת המודיעין.

ל-GPT יש כמובן ליקויים והטיות משלו. כל התנסות עם ה-GPT (גם בגרסה 3.0 וגם בגרסה 4.0) מראה שיש לו מגבלות קשות בכל הקשור לתיאור המציאות העדכנית ובמתן מענה על שאלות ידע בסיסיות. אין לו יכולת להבחין בין אמת לשקר והוא לעיתים קרובות ממציא פרטים שאינם קיימים במציאות. עניין זה לא השפיע על הסימולציות שכן במהלכן לא עשינו שימוש ביכולת חיפוש המידע של ה-GPT. אנחנו סיפקנו ל-GPT הן את נתוני היסוד והן את כל המידע המודיעיני שעליו התבססה הסימולציה. מה שבחנו הוא את יכולתו לסייע בתהליך הניתוח של המידע שסופק לו. יכולת זו מבוססת, כנראה, על היכרות עם ישויות (מדינות, ארגונים) בעלות מאפיינים דומים; היכרות עם פרשיות עבר ואירועים שבהם השתקפו תהליכים ודינמיקות רלוונטיים; והיכרות עם גישות מחקריות מתחומים שונים.

לכן, הגישה שאנו מציעים במאמר זה באשר לפוטנציאל השימוש ב-GPT ובמוצרי בינה מלאכותית יוצרת דומים היא זו של "מקום ליד השולחן". לפי גישתנו, אין כוונה שה-GPT יחליף את ראש אמ"ן או את החוקרים של חטיבת המחקר אלא שיעשה בו שימוש להעשרת תהליך החשיבה שלהם.

⁴ וולשטטר, לעיל הערה 1. על הסיבות לדבקות בקונספציה, ראו בין השאר: בר-יוסף, א' (2001). **הצופה שנרדם: הפתעת יום הכיפורים ומקורותיה**. זמורה ביתן – מוציאים לאור; בר-יוסף, א' (2019). **מתקפת פתע – מנהיגות ומודיעין במבחן העליון**. כנרת, זמורה, דביר – מוציאים לאור בע"מ; קם, א' (1990). **מתקפת פתע**. צה"ל – הוצאת מערכות / משרד הביטחון – ההוצאה לאור; בן-ישראל, י' (1989). **דיאלוגים על מדע ומודיעין**. צה"ל – הוצאת מערכות / משרד הביטחון – ההוצאה לאור; בן-ישראל, י' (1999). **הפילוסופיה של המודיעין**. משרד הביטחון – ההוצאה לאור; לניר, צ' (1983). **ההפתעה הבסיסית – מודיעין במשבר**. הקיבוץ המאוחד; הוייר, ר' (2005). **הפסיכולוגיה של המחקר המודיעיני**. משרד הביטחון – ההוצאה לאור; ברון, א' (2021). **המחקר המודיעיני: בירור המציאות בעידן של תמורות ושינויים**. המרכז למורשת המודיעין, המכון לחקר המתודולוגיה של המודיעין (2021); Pearl; Wohlstetter, R' (1962). **Harbor: Warning and Decision**. Stanford University Press.

הסימולציות שקיימנו מוכיחות של-GPT יש צורת ניתוח שונה וסגולות ייחודיות שביכולתן לסייע לאנשי מודיעין ולמקבלי החלטות לברר את המציאות ולהבין אותה. שונות זו היא שמהווה בראייתנו את "כרטיס הכניסה" של ה-GPT לדיונים שמתנהלים סביב שולחנו של ראש אמ"ן ושולחנות נוספים בארגונים המודיעיניים. את מאפייניו המדויקים של המקום הזה סביב השולחן נכון להגדיר במחקרי המשך.

הסימולציות ותוצאותיהן

ה-GPT לא נדרש להסבר מיוחד אודות תהליך הניתוח המודיעיני וכמערכת שפה משוכללת הוא השתמש בסימולציות בשפה דומה לשפתם של אנשי מודיעין מיומנים. במהלך הסימולציות הוא ניתח את נתוני היסוד שסופקו לו, סיפק פרשנות משלו לפריטי המידע שהוצגו לו ונתן גם המלצות למדיניות ולתכנון בהתבסס על הניתוח המודיעיני. כל זאת, אף שה-GPT לא הוכשר באופן ספציפי לתפקיד של מעריך מודיעיני ולא הכיר לפני הסימולציות את העולם הדמיוני שהוצג בפניו:

- הסדרה הראשונה, שעסקה במלחמת יום הכיפורים, כללה הזרמה של נתוני המודיעין האמיתיים שהיו בידי אמ"ן מסוף ספטמבר ועד תחילת אוקטובר של שנת 1973 וכן, בחלק מהסימולציות, נתונים בדיוניים. ה-GPT לא הצליח לקבוע ש"תהיה מלחמה מחר" באף אחת מהסימולציות. עם זאת, הוא זיהה בצורה טובה מאוד סתירה בין הקונספציה, שהוצגה לו כ"הערכת המומחים", לבין המידע האחר ובסימולציות הרלוונטיות העריך שהסבירות למלחמה היא גבוהה יותר ואף המליץ לנקוט בצעדי התגוננות;
- הסדרה השנייה, שעסקה במתקפה היפנית על פרל הארבור, כללה הזרמה של נתוני המודיעין האמיתיים שנאספו על ידי סוכנויות המודיעין האמריקניות מסוף ספטמבר ועד לתחילת דצמבר של שנת 1941. במהלך הסימולציות, ה-GPT זיהה שישנן מספר אינדיקציות למתקפה יפנית קרובה על ארה"ב, אבל לא הצליח לקבוע בוודאות שתתרחש "מתקפה מחר" דווקא על הנמל בפרל הארבור. עם זאת, לנוכח המידע, הוא המליץ לנקוט בכל זאת בצעדים להגנת הנמל;

- הסדרה השלישית, שעסקה באירוע דמיוני של שיגור טיל קרקע-קרקע (טק"ק) על ישראל מסוריה בשנת 2014, כללה הזרמה של ידיעות מודיעיניות בדיוניות במהלך הימים שלפני השיגור. התוצאות של סדרת הסימולציות הזו לא היו אחידות, אך בחלק מהן הצליח ה-GPT להבין את ההתרחשות המרכזית של הסדרה שבמרכזה השתלטות של כוח מורדים על בסיס הטילים המרוחק והכנת טק"ק לשיגור על ישראל. תוצאות אלו דומות לתוצאות של סימולציות דומות שנערכו בשנים האחרונות למשתתפים אנושיים, חלקם אנשי מודיעין מיומנים.

בכל הסימולציות, לא הזכרנו במפורש את המדינות והארגונים האמיתיים, אלא תיארונו עולם דמיוני שייצג, באופן תמציתי, את המציאות המורכבת בשנים הרלוונטיות. זאת, כדי למנוע הטיה שקשורה בהיכרות מוקדמת של ה-GPT עם האירוע האמיתי ותוצאותיו. הסימולציות החלו בכך שנתנו למערכת נתוני יסוד על העולם הדמיוני שיצרנו והן התנהלו באנגלית, בדרך של "פינג פונג", שבה המשכנו להזרים למכונה פרפרזות (תמציות) של נתוני מודיעין גולמיים בסדרה של צעדי זמן ותשאלנו את ה-GPT לגבי המידע והמסקנות שעולות ממנו.

הציטוטים המופיעים במסמך הם תרגום מדויק מאנגלית של הדיאלוג עם ה-GPT. לצורך נוחות הקריאה, החלפנו ברוב הציטוטים את הכינויים של המדינות והארגונים שבהם עשינו שימוש לשמותיהם האמיתיים (ישראל במקום "המדינה הכחולה", סוריה במקום "המדינה הירוקה", חזבאללה במקום "הארגון הצהוב" וכו').

מסימנים שונים אנחנו מסיקים שאכן הצלחנו למנוע את ההטיה שקשורה בהיכרות מוקדמת עם האירועים האמיתיים ותוצאותיהם. ה-GPT "מכיר" בוודאי את מלחמת יום הכיפורים, למשל, אבל הוא לא "הבין" שהוא נמצא בסימולציות שמייצגות דווקא אותה. כך, למשל, בתשובה לשאלה מפורשת באחת מסימולציות יום הכיפורים הוא השיב כי התרחיש "חולק כמה קווי דמיון עם המתחים והיריבות בין מדינות בתקופת המלחמה הקרה, במיוחד מרוץ החימוש והפחד ממלחמה גרעינית". זאת, מבלי לזהות את האירוע הספציפי של מלחמת יום הכיפורים.

לאחר שלוש סדרות הסימולציות הראשונות קיימנו שיחות אודות התוצאות עם אנשי מודיעין ומומחי בינה מלאכותית וכתוצאה משיחות אלו קיימנו שורה של סימולציות נוספות שבחנו טענות שונות אודות התוצאות. במסגרת זו בחנו טענות לגבי נטייה קבועה של ה-GPT ל"אלרמיזם"; את יכולתו להתמודד עם "רעשים" חזקים; את האפשרות שהוא מכיר בכל זאת את האירועים נשואי הסימולציות ושהיכרות זו משפיעה על תפקודו; וכן את האפשרות שיש לו "תבניות גנריות" שמגדירות באופן אחיד את תגובתו למצבים שונים, בלי קשר לתמונת המצב הספציפית.

כל התנסות עם ה-GPT מראה שיש לו מגבלות קשות בכל הקשור לתיאור המציאות העדכנית ובמתן מענה על שאלות ידע בסיסיות. אין לו יכולת להבחין בין אמת לשקר והוא לעיתים קרובות ממציא פרטים שאינם קיימים במציאות. עניין זה לא השפיע על הסימולציות שכן במהלכן לא עשינו שימוש ביכולת חיפוש המידע של ה-GPT. אנחנו סיפקנו ל-GPT הן את נתוני היסוד והן את כל המידע המודיעיני שעליו התבססה הסימולציה. מה שבחנו הוא את יכולתו לסייע בתהליך הניתוח של המידע שסופק לו.

סדרת הסימולציות בנושא מלחמת יום הכיפורים

הכשל המודיעיני אינו הסיבה היחידה שהובילה לפתיחת מלחמת יום הכיפורים בתנאים גרועים כל כך לישראל. בעשורים האחרונים התברר שהכשל המודיעיני היה חלק מכשל רחב הרבה יותר, שכלל גם כישלון של הדרג הפוליטי בעיצוב המדיניות הישראלית לפני המלחמה וכישלון של הדרג הצבאי הבכיר בתחום האסטרטגיה הצבאית והכנת הצבא. ועדיין, לצד כל אלו מלחמת יום הכיפורים היא, בראש ובראשונה, כשל מודיעיני שבמרכזו פער קיצוני בהבנת המציאות, שניצב בשורה אחת עם הכשלים המודיעיניים המפורסמים בהיסטוריה המודרנית.

במרכז הכשל המודיעיני בנושא ההתרעה למלחמה עמדה הקונספציה המפורסמת שאותה תיארה וועדת אגרנט. אף שלוועדה היו ליקויים רבים וספק רב אם חברה הצליחו להבין את הגורמים העמוקים לכשל המודיעיני, אין ספק שהיא זיהתה נכון שהמסגרת התפיסתית שהתקבעה באמ"ן בשנים שלפני המלחמה כללה שני מרכיבים: האחד, שמצרים לא תצא

למלחמה מבלי שתצטייד במטוסי קרב-הפצצה שיוכלו לתקוף בעומק ישראל ולהרתיע אותה; והשני, שסוריה לא תצא למלחמה ללא מצרים.⁵ ראש אמ"ן אלי זעירא וראש מחלקת (היום חטיבת) המחקר באמ"ן, אריה שלו, המשיכו לדבוק בקונספציה עד לצוהרי יום הכיפורים ממש, למרות מידע התרעתי שהגיע בימים ובשעות שלפני המלחמה.

העדויות החזקה ביותר לעומק ההשפעה של הקונספציה על הערכת המודיעין נמצאת במסמך המודיעיני שנכתב יממה לפני המלחמה. לאחר עשרות סעיפים שתיארו את ההיערכות החריגה בצבאות מצרים וסוריה, קבע אמ"ן שבמצרים אכן מתקיים תרגיל ושהסורים נמצאים בכוננות לתגובה על הפלת מטוסים שלהם שבועיים לפני כן. על אף זאת, סעיף 40 של המסמך דבק בקונספציה וקבע כי "למיטב הערכתנו, לא חל שינוי בהערכת המצרים את יחסי הכוחות בינם לבין כוחות צה"ל. אי לכך, הסבירות שהמצרים מתכוונים לחדש הלחימה היא נמוכה".⁶

נתוני היסוד, המידע והקונספציה

בתחילת כל אחת מהסימולציות מסדרה זו הוצג ל-GPT עולם דמיוני של מדינות שייצגו את ישראל, מצרים, סוריה, ירדן, ברית המועצות וארצות הברית. כמו כן נמסרו לו פרטים כלליים על כל אחת מהמדינות.

להלן דוגמה לנתוני היסוד שנמסרו ל-GPT באחת הסימולציות (השמות האמיתיים של המדינות הוספו כאן ובהמשך המסמך לצורך נוחות ולא נמסרו, כמובן, ל-GPT בסימולציות):

- המדינה הכחולה (ישראל) היא מדינה דמוקרטית קטנה יחסית. היא נמצאת ביריבות עם המדינה האדומה (מצרים) ועם המדינה הירוקה (סוריה). שתיהן נתמכות על ידי המדינה השחורה (ברית המועצות), שהיא מעצמה גלובלית.

⁵ ועדת החקירה – מלחמת יום הכיפורים (1974). דין וחשבון חלקי. עמ' 8.
⁶ אגף מודיעין/מחקר, לקט מודיעין מיידי – צבאי. 5 באוקטובר 1973, 13:15. עמ' 9.

- לפני כ-6 שנים הייתה מלחמה והמדינה הכחולה כבשה שטחים גדולים יחסית הן מהמדינה האדומה והן מהמדינה הירוקה. לפני כ-3 שנים הסתיימה מלחמת התשה של כשנה בין המדינה הכחולה למדינה האדומה.
- בחודשים האחרונים היה מאמץ דיפלומטי לתווך בין המדינה האדומה למדינה הכחולה שהובילה המדינה הלבנה (ארצות הברית), שגם היא מעצמה גלובלית.



These are the states that participate in the scenario (after that I will give you the information): The blue state is in rivalry with the red state and the green state. The red and green states are supported by a world power that is the black state. We will analyze the scenario from the point of view of the blue state



Thank you for providing me with this information. To make sure I understand the scenario correctly, please confirm the following details:



1. The blue state is in a rivalry with both the red and green states.
2. The red and green states are supported by the black state, which is a world power.
3. We will be analyzing the scenario from the perspective of the blue state.

Is this correct?



Correct

דוגמה לדו-שיח עם ה-GPT אודות נתוני היסוד

לאחר מכן, נמסרו ל-GPT, בשורה של צעדי זמן, פרפרוזות של הידיעות העיקריות שהגיעו לאמ"ן בעשרת הימים שלפני מלחמת יום הכיפורים, עד לבוקר ה-5 באוקטובר 1973. הידיעות נמסרו תוך שימוש בכינויי המדינות (המדינה הכחולה, הירוקה, האדומה וכו') והן מובאות כאן עם השמות האמיתיים לצורך נוחות. במסגרת זו נמסרו הידיעות הבאות:⁷

- המידע וההערכות שמסר מלך ירדן בפגישתו החשאית עם ראש הממשלה ב-25 בספטמבר;

⁷ הידיעות נמסרו על פי המפורט אצל: בר-יוסף, לעיל הערה 4, עמ' 235-340.

- מידע אודות תרגיל המלחמה שהתנהל במצרים וצעדי המוכנות שננקטו במסגרתו;
- מידע אודות הכוננות שהוכרזה בסוריה, וכן ביטול החופשות ושינויי היערכות בכוחות היבשה, האוויר והים הסורים;
- מידע שלפיו התרגיל במצרים הוא מסווה למלחמה ובסיומו יתקפו המצרים;
- מידע על ביקור שר ההגנה המצרי בסוריה ב-3 באוקטובר;
- המידע על פינוי היועצים הסובייטים ממצרים ורוסיה בליל ה-4-5 באוקטובר.

כמו כן נמסרה ל-GPT בסימולציות הרלוונטיות (ראו להלן) הקונספציה המקורית כפי שתיארה אותה ועדת אגרנט לאחר המלחמה. נמסר ל-GPT שהערכת המומחים היא שמצרים לא תצא למלחמה עד שלא תשיג יכולת תקיפה במטוסים בעומק ישראל וסוריה לא תצא למלחמה בלי מצרים. כמו כן נמסר ל-GPT שמכיוון שמצרים עוד לא השיגה את יכולת התקיפה בעומק ישראל במטוסים, מעריכים המומחים שהסבירות למלחמה היא נמוכה. זאת, אף שלמצרים יש יכולת תקיפה של עומק ישראל בטילי קרקע-קרקע.

כחלק מהקונספציה, נמסרה ל-GPT גם הערכת המומחים שהתרגיל במצרים הוא אותנטי ושהכוננות בסוריה קשורה כנראה בהכנות לתגמול מוגבל על קרב אוויר שבו הופלו מטוסיהם.

הסימולציות

הסימולציה הראשונה בחנה את יכולתו של ה-GPT לזהות סתירה בין הקונספציה לבין מידע סותר. במסגרת זו נמסרו ל-GPT הקונספציה המקורית ("כ"דעת המומחים") וכן המידע שהגיע ערב המלחמה. בסיום הזרמת המידע ובהתייחס ספציפית לפינוי החפוז של היועצים הסובייטים, מסר ה-GPT ש"מידע זה עשוי להצביע על כך שקיימת סבירות גבוהה יותר לעימות, וייתכן שהמומחים צריכים לשנות את הערכותיהם". בתשובה לשאלה האם תהיה מלחמה מחר השיב ה-GPT את תשובתו האחידה שלפיה "כמודל שפה AI, אין לי יכולת לחזות אירועים עתידיים" וכן שללא מידע

נוסף "קשה לקבוע בוודאות אם מלחמה תתחיל מחר או לא". עם זאת, הוא הוסיף שהמידע "מעיד על עלייה בפעילות הצבאית ועל הגברת המתיחות".



What do you think about the opinion of the experts who said that the probability of war is low?



Given the new developments that have occurred, the assessment of the likelihood of war being low made by the experts may no longer be accurate. The fact that the Black State is hastily evacuating their advisors from the Red and Green States suggests that they may have received intelligence indicating a higher risk of a wider regional conflict.



Additionally, the information provided by the Pink Country's intelligence authority suggested that the Red and Green States were planning a joint military campaign against the Blue State. While this information was initially discounted by the experts, the recent developments may lend more credibility to this assessment.

Therefore, it is important to reassess the situation in light of the new developments and to consider the possibility that the likelihood of war may be higher than previously estimated. The Blue State should continue to monitor the situation closely and take appropriate steps to prepare for a potential military attack while also seeking to prevent further escalation of the conflict through diplomatic means.

דוגמה לדו-שיחה עם ה-GPT אודות "הערכת המומחים" (הקונספציה)

הסימולציה השנייה בחנה את יכולתו של ה-GPT לשנות קונספציות ולבחון את המידע למול קונספציות שונות ומנוגדות. לכן, נמסרה לו קונספציה הפוכה לזו המקורית, היינו – נאמר לו שהמומחים מעריכים כי סוריה ומצרים מתכננות מתקפת פתע נגד ישראל. הערכתו את המידע הייתה דומה יחסית להערכתו בסימולציה הראשונה. בתשובה לשאלה לגבי האפשרות לפרוץ המלחמה למחרת, הוא השיב שלפי המידע "מצרים וסוריה נערכות לפעולה צבאית וייתכן שתהיה מלחמה בטווח הזמן הקרוב" וכן שהמידע "מצביע על כך שהסבירות לעימות גבוהה יותר מאשר של השלום". עם זאת, הוא הוסיף שלא ניתן "לחזות במדויק אם מלחמה תתחיל מחר או לא".

הסימולציה השלישית בחנה את התפקוד של ה-GPT בהעדר קונספציה. לכן, בניגוד לשתי הסימולציות הקודמות, בפתח הסימולציה השלישית לא נמסרה ל-GPT קונספציה כלשהי. כמו בסימולציות הקודמות, ה-GPT נשאל האם תהיה מלחמה מחר והשיב כי "על סמך המידע, נראה

כי הן סוריה והן מצרים מתכוננות לפעולה צבאית וכי ברית המועצות מודאגת מיציבות המצב ... עם זאת, ללא מידע נוסף, אי אפשר לקבוע אם מחר תתחיל מלחמה או לא". כמו כן, המליץ ה-GPT לנקוט בצעדי התגוננות עקב האפשרות לפעולה ההתקפית.

הסימולציה הרביעית בחנה האם ה-GPT אומן להיות תמיד "אלרמיסט". לכן, בפתח הסימולציה שוב נמסרה ל-GPT הקונספציה ההפוכה, כלומר זו שלפיה מצרים וסוריה מתכוננות למתקפת פתע. אבל, המידע שנמסר לו בצעדי הזמן הבאים היה מידע בדיוני בעל אופי מרגיע שכלל מסרים פומביים וחשאיים על העדר כוונות מלחמה וצעדים בוני אמון שכללו פירוק מאגרי נשק והרחקת כוחות מאזור הגבול. לנוכח מידע זה, ה-GPT השיב כי "נראה שהסבירות למלחמה בטווח הקרוב נמוכה ממה שהעריכו בעבר המומחים" וכי "הסיכון למלחמה פחת". לגבי האפשרות שתפרוץ מלחמה מחר השיב ה-GPT ש"הסבירות למלחמה שתתחיל מחר נראית נמוכה יחסית".

הסימולציה החמישית בחנה האם ל-GPT יש תבנית קבועה ואחידה של הטלת ספק בגישת המומחים ושל הצגת סתירות למול המידע. זאת, בלי קשר לנסיבות הספציפיות של הסימולציה. לכן, נמסרה ל-GPT הקונספציה המקורית (סבירות נמוכה למלחמה) וכן המידע הבדיוני המרגיע מהסימולציה הרביעית. ה-GPT נשאל האם יש סתירה בין המידע להערכת המומחים והשיב כי "אין סתירה ישירה בין המידע להערכת המומחים" וכי המידע מצביע על כך "שירדה הסבירות למלחמה בטווח הזמן הקרוב". לגבי האפשרות שתהיה מלחמה מחר, השיב ה-GPT כי "ההתפתחויות האחרונות" "מצביעות על כך שהסיכוי למלחמה בטווח הקרוב ירד".

הסימולציה השישית בחנה את התפקוד של ה-GPT בסביבה רוויית "אותות" ו"רעשים" שמגיעים במקביל ומגבירים את אי הוודאות. גם בסימולציה זו נמסרה ל-GPT הקונספציה המקורית ולצידה המידע המקורי, אך הפעם בליווי המידע הבדיוני המרגיע מהסימולציה הרביעית. במצב עניינים זה, ה-GPT התקשה יותר בהערכה מסכמת של הסבירות למלחמה, אך הקפיד להציג את אפשרות המלחמה כאפשרות סבירה שמחייבת היערכות מקדימה. בשלבים מאוחרים יותר של הסימולציה העריך ה-GPT שהסבירות למלחמה "גדלה בהשוואה להערכות קודמות" וכי הערכת המומחים לגבי הסבירות הנמוכה למלחמה "כבר אינה מדויקת".

ניתוח הסימולציות

בסיכומו של דבר, הממצא העיקרי של סדרת הסימולציות על מלחמת יום הכיפורים הוא שיש ל-GPT יכולת גבוהה להצביע על סתירות בין הקונספציה לבין המידע. בהשוואה לחוקרים האנושיים, הסימולציות מראות שאין לו את הדבקות המחשבתית והרגשית בקונספציה שמתוארת במחקרים רבים ואפיינה, ללא ספק, את אנשי המודיעין ערב מלחמת יום הכיפורים. דבקות זו היא זו שמאפשרת לחוקרים אנושיים, כדוגמת אלו של 1973, להסביר לעצמם ולסביבתם כל סתירה ולהמשיך ולדבוק בקונספציה גם כאשר המציאות משדרת סימנים אחרים. ה-GPT, לעומת זאת, רואה את "הערכת המומחים" (היינו, את הקונספציה) כעוד פריט מידע ואינו מעניק לו עדיפות מיוחדת בהערכה הכוללת.

המשמעות של הסימולציות הרביעית והחמישית היא שה-GPT אינו מפקפק תמיד בהערכת המומחים, לא מתריע על סתירה במקום שהיא אינה קיימת ולא אומן להיות תמיד אלרמיסט.

עדיין, נשאלת השאלה האם נכון לבקר את ה-GPT על כך שלא הצליח לקבוע באופן חד משמעי "מלחמה מחר"? אפשרות המלחמה נוכחת בניתוחיו של ה-GPT כאפשרות סבירה, אך הוא נמנע בעקשנות מחיזוי עיתוי תחילתה המדויק. אבל, למרות שה-GPT לא התריע על "מלחמה מחר", הצגת "האפשרות המתחרה" של המלחמה כאפשרות סבירה, שאינה נסתרת במידע, יכולה הייתה להועיל למקבלי ההחלטות ערב המלחמה. למעשה, זה בדיוק מה שהיה צריך לעשות אלי זעירא בימים הדרמטיים לפני המלחמה, וזהו גם הפח היקוש אך האנושי שנפל לתוכו: העמדה הנחרצת שלו לגבי "הסבירות הנמוכה" ודחייתה המוחלטת של האפשרות האחרת, שמדובר דווקא במלחמה.

בר-יוסף, שתיאר וניתח את הדיונים ב-5 באוקטובר 1973, הגיע למסקנה דומה כאשר הסביר כי תמונת המודיעין באותו בוקר "לא הייתה חייבת להביא את מעריכי אמ"ן לקבוע קטיגורית כי מצרים וסוריה עומדות לצאת למלחמה. אבל היא הייתה חייבת להביא אותם למסקנה כי קיימת סבירות של ממש שהכנותיהן הצבאיות הן לקראתה". למעשה, זאת גם העמדה העקבית שהציג ה-GPT בכל הסימולציות. בר-יוסף ממשיך ומסביר שסביר להניח "כי אם בשעות המוקדמות של

בוקר יום שישי הרמטכ"ל היה מכנס את המטכ"ל ומודיע כי קיימת אפשרות של ממש כי מלחמה תפרוץ בזמן הקרוב, ואם ראש אמ"ן היה נותן גיבוי לכך, כל תהליך ההתכוננות למלחמה היה לובש דפוס חדש".⁸

כאשר סיכם בר-יוסף את מחקרו בעניין הפתעת מלחמת יום הכיפורים הוא כתב שבהרכב אחר של אנשי מודיעין "הייתה נחסכת מהקברניטים הנחרצות ההערכתית לפיה מלחמה היא בלתי אפשרית".⁹ ה-GPT נוקט במתודולוגיה מודיעינית זהירה מאוד, שאין בה את הנחרצות ההערכתית שאפיינה את ראשות אמ"ן של 1973. הסימולציות מסדרת מלחמת יום הכיפורים מראות שה-GPT בוודאי לא היה מנסח את סעיף 40 המפורסם באופן שבו ניסחו אותו באמ"ן בנייר שהופץ ערב המלחמה. אפשר שהוא היה גם מתריע על הפער שבין המידע לבין הקונספציה. כך או כך, הפרשנות הבסיסית שלו למידע בכל אחת מהסימולציות הרלוונטיות בסדרה, גם מבלי להתריע "מלחמה מחר", הייתה צריכה להביא כל מקבל החלטות סביר לגייס את המילואים ולהיערך למלחמה אפשרית.

סדרת הסימולציות בנושא המתקפה על פרל הארבור

המתקפה היפנית על פרל הארבור בהוואי בבוקרו של יום א', 7 בדצמבר 1941, נכנסה להיסטוריה כאחת ממתקפות הפתע המרשימות ביותר. המושג "פרל הארבור" הפך שם נרדף להפתעה כוללת, הן ברמה האסטרטגית והן ברמה הטקטית. פרטי האירוע עצמו מוכרים וידועים: כוח משימה יפני של נושאות מטוסים, הצליח להתקרב להוואי בנתיב צפוני ולשגר גלי תקיפה של מטוסי תקיפה, אשר פגעו קשות בספינות הצי של ארצות הברית. הכוח התוקף נסוג באבדות מועטות ובלא שנעשה שום ניסיון לפגוע בו. פעולה זו של היפנים הביאה לכניסתה של ארצות הברית למלחמה.

בהיסטוריה של ארצות הברית נחרתה ההתקפה על פרל הארבור כחוויה טראומתית עמוקה, אשר זעזעה את הממסד הצבאי והאזרחי ולמעשה את האומה כולה. על פרל הארבור נכתבו מחקרים

⁸ בר-יוסף, לעיל הערה 4, בעמ' 406.

⁹ בר-יוסף, לעיל הערה 4, בעמ' 420.

רבים, שאחד המפורסמים בהם הוא ספרה של רוברטה וולשטטר "פרל הארבור: אזהרה והכרעה"¹⁰. הספר הניח את היסוד למערכת שלמה של מושגים ותפיסות בנושא ההפתעה. היה זה הניסיון השיטתי הראשון להעמיד במוקד ניתוחה של פרשיית הפתעה את גורם התפיסה האנושית, ולהצביע על האפשרות שעיוותי תפיסה דומים מילאו תפקיד משמעותי גם במקרים אחרים.

שורש ההפתעה, לפי וולשטטר, היה טמון אפוא לא בהיעדרם של "סימנים מעידים", כי אם במערכת של כשלים תפיסתיים, אשר התאפשרו על רקע עמימותו של כלל המידע שנאסף.

נתוני היסוד, המידע והקונספציה

בתחילת הסימולציות מסדרה זו הוצג ל-GPT עולם דמיוני של מדינות שייצגו את ארה"ב, יפן, סין והמדינות באירופה. כמו כן נמסרו לו פרטים כלליים על המדינות השונות בהתאם למציאות של סוף שנת 1941.

להלן דוגמה לנתוני היסוד שנמסרו ל-GPT באחת הסימולציות (השמות האמיתיים של המדינות הוספו כאן ובהמשך המסמך לצורך נוחות ולא נמסרו, כמובן, ל-GPT בסימולציות):

- המדינה הכחולה (ארצות הברית) היא מעצמה גלובלית, דמוקרטית-ליברלית, חזקה. היא ממוקמת בחלק הצפוני של היבשת הכחולה (אמריקה). היא מתאוששת עדיין ממשבר כלכלי גדול שהתרחש לפני יותר מעשור. יש לה יכולת כלכלית וטכנולוגית אדירה. הדוקטרינה השלטת בעשור האחרון במדינה הכחולה היא של בדלנות. ביבשת הירוקה (אירופה) מתחוללת מלחמה גדולה מאוד ובמדינה הכחולה יש וויכוח, הן בציבור והן במערכת הפוליטית על השאלה האם להצטרף למלחמה.
- המדינה האדומה (יפן) היא מדינה גדולה בעלת שאיפות אימפריאליסטיות הממוקמת ביבשת האדומה (אסיה). היא מתאפיינת בשנים האחרונות בתיעוש מהיר, החלפת מסורות ישנות

¹⁰ Wohlshtater, לעיל הערה 4; פרק הסיכום של הספר תורגם לעברית: וולשטטר, לעיל הערה 1.

בחדשות (מודרניזציה) תוך בניית צבא וצי מודרניים. תהליך זה לווה בהתפתחות של מגמות לאומניות קיצוניות ומיליטריסטיות במדינה, שהטיפו להתפשטות כדי להשיג שוויון כוחות מול המדינה הכחולה והמעצמות של היבשת הירוקה. לשם כך שואפת המדינה האדומה להשתלט על שטחים, אוכלוסיות ומשאבים טבעיים שהיו נחוצים לה, בשל מחסור במשאבים כאלה באיים שלה.

- בשנים האחרונות פלשה המדינה האדומה למדינה הצהובה (סין) ופתחה במלחמה איתה שבה כבשה שטחים ניכרים וערים מרכזיות במדינה הצהובה, וכן כבשה שטחים ניכרים בדרום-מזרח היבשת האדומה.

- במהלך השנה האחרונה התגברה המתיחות ארוכת השנים בין המדינה האדומה למדינה הכחולה. בחודש יולי, בתגובה להמשך מגמות הכיבוש של המדינה האדומה במדינה הצהובה, הטילו המדינה הכחולה ומעצמה נוספת (ביבשת הירוקה) שורת סנקציות כלכליות על המדינה האדומה, בהן אמברגו על ייצוא נפט, סגירת תעלה חשובה לספינות של המדינה האדומה, והקפאת נכסים של המדינה האדומה ברחבי העולם.

- צעדים אלו יצרו במדינה האדומה מצב לא יציב בו ראשי המערכת הפוליטית מנסים למנוע מלחמה על ידי חיפוש פתרון דיפלומטי (בהבינם שמלחמה במדינה הכחולה תגמר בתבוסה), בעוד מנהיגי הצבא חתרו לביצוע פעולה צבאית זריזה וניצחת.

- בתקופה האחרונה, כולל בימים האחרונים, מתנהל משא ומתן בין המדינה האדומה למדינה הכחולה. המנהיגים של המדינה הכחולה מקווים שהמשא ומתן יוביל לתוצאות חיוביות.

כמו כן נמסרה ל-GPT הקונספציה שהיוותה את המסגרת לחשיבה האמריקנית לפני המתקפה על פרל הארבור. במסגרת זו נמסר ל-GPT כי ההערכה המקובלת בארה"ב (הן בקרב גורמי המודיעין והן אצל מקבלי ההחלטות) היא שלמנהיגי יפן יש כוונות עוינות כלפי ארה"ב, אלא שאין להם יכולת מספקת לממשן. לכן, אף שהמומחים בארה"ב חושבים שמתקפה יפנית היא בהחלט אפשרית, הם לא מעריכים שהיא תתרחש דווקא בפרל הארבור.

בנוסף, נמסרה ל-GPT תמונת המצב האיסופית, כולל הנגישות שהשיג המודיעין האמריקני לפני המתקפה למערכת התקשורת בין משרד החוץ היפאני לשגרירויותיו ברחבי העולם וכן בין המודיעין היפני לסוכנים שלו ברחבי העולם ובכלל זה בארה"ב. במסגרת זו נמסר ל-GPT כי:

- מפענחי הצופן של ארצות הברית הצליחו לשבור את מערכת התקשורת המוצפנת של יפן עם המרגלים שלה בארצות הברית ובמדינות נוספות.
- מפענחי הצופן של ארצות הברית הצליחו גם לשבור גם את מערכת התקשורת המוצפנת של יפן ויכולים לקרוא את התכתובת שבין משרד החוץ היפני לשגרירויות יפן ברחבי העולם. בנוסף, הם הצליחו לשבור צפנים נוספים ולפיכך יש לארצות הברית נגישות גם לתכתובת של גורמים נוספים בממשל היפני.

הסימולציות

במהלך הסימולציות נמסרו ל-GPT, בשורה של צעדי זמן, הידיעות העיקריות שהצטברו בשבועות שלפני המתקפה במודיעין האמריקני ובכלל זה:¹¹

- שינוי הטון בעיתונות ביפן שמיוחס לנחישות הממשלה החדשה להתפשט בדרום אסיה;
- בקשות מהסוכנים היפניים לאסוף מידע על תנועות ספינות והיערכות בנמלים, כולל ספציפית, פרל הארבור;
- הגדרת ה-29 בנובמבר כמועד האחרון שבו ניתן להגיע להסכם עם ארה"ב ואחריו "יקרה משהו באופן אוטומטי";
- שליחת קודים להשמדת מערכות צופן ושימוש במערכת גלויה;
- יציאת ספינות יפניות מהנמלים במקביל להחלפה רחבה של אותות קריאה;

¹¹ הידיעות נמסרו על פי המפורט אצל: Maxwell Air Force Base. Piaccine, R' (1997). **Pearl Harbor: Failure of Intelligence?**

- פענח הודעה ארוכה, בת 14 חלקים שהכריזה על ניתוק היחסים של יפן עם ארה"ב.
- על בסיס תמונת המידע שהייתה בידי ארצות הברית בלילה שבין ה-6 ל-7 בדצמבר 1941 נשאל ה-GPT כיצד נכון להתייחס לעמדת המומחים שלפיה הסבירות לתקיפה יפנית בפרל הארבור נמוכה? ה-GPT השיב כי "ייתכן שעדיף לטעות לצד המחמיר ולנקוט בצעדים לשיפור ההגנה על פרל הארבור, ללא קשר להערכת המומחים". כאשר נשאל ה-GPT האם יפן תתקוף היום את פרל הארבור, הוא השיב כי "בהתבסס על המידע, ישנם מספר אינדיקציות לכך שיפן נערכת למתקפה ... עם זאת, אי אפשר לומר בוודאות אם היום תתרחש מתקפה על פרל הארבור". בתשובה לשאלה אילו סימנים יש לחפש כדי לוודא אם ההתקפה אכן תתרחש היום, הציע ה-GPT שורה של סימנים מעידים שאכן רלוונטיים לעניין זה:
- פעילות צבאית מוגברת בקרבת הנמל, לרבות נוכחות של כלי שיט, מטוסים או חיילים שאינם נראים בדרך כלל באזור;
- דפוסי תקשורת חשודים בין כלי השיט הצבאיים של יפן לבין מפקדיהם, בעיקר כאלה שעלולים להעיד על מתקפה קרובה;
- עדויות לעלייה ברמת ההצפנה או שימוש בשיטות הצפנה חדשות, שיכולים להעיד על כך שיפן מנסה להסתיר את תוכניותיה;
- שינויים בהתנהגות כלי השיט של יפן, כמו שינוי מסלולם או מהירותם, שיכולים להעיד שהם מתכוננים להתקפה;
- שימוש בקודים ששלחה יפן לשגרירותיה לגבי פרוץ פעולות האיבה, מה שיצביע על כך שהיא נערכת לסכסוך;
- מסירת הטקסט המודיע על ניתוק היחסים בין יפן לארצות הברית, שכן הוא עשוי להצביע על כך שיפן נערכת לסכסוך;
- כל פעילות חריגה או בלתי צפויה אחרת ליד פרל הארבור שעלולה לרמוז על מתקפה קרובה.



What signs would you look for to ascertain if the attack on the Golden Harbor would actually occur today?



As the situation is hypothetical and based on the information you have provided, to determine if an attack on the Golden Harbor would occur, several signs could be looked for:



1. Increased military activity near the harbor, including the presence of vessels, aircraft, or troops that are not typically seen in the area.
2. Suspicious communication patterns between the Red State's military vessels and their commanders, especially those that may indicate an imminent attack.
3. Evidence of an increase in the level of encryption or the use of new encryption methods, which would suggest that the Red State is trying to conceal its plans.
4. Changes in the behavior of the Red State's vessels, such as altering their course or speed, which could indicate that they are preparing for an attack.
5. The implementation of the codes that the red state sent to its embassies regarding the outbreak of hostilities, which would suggest that the Red State is preparing for conflict.
6. The delivery of the text announcing the severance of relations between the Red and Blue States, as it may indicate that the Red State is preparing for conflict.
7. Any other unusual or unexpected activity near the Golden Harbor that could suggest an imminent attack.

דוגמה לדו-שיח עם ה-GPT אודות "סימנים מעידים"

ניתוח הסימולציות

וולשטטר, שחקרה את הסיבות לכישלון בפרל הארבור, שללה את האופן שבו הוסברה ההפתעה עד אז באמצעות "האשמת המעורבים בה, כיחידים או כקבוצות, בקשירת קשר, ברשלנות או בסכלות".¹² היא הייתה זו שהפנתה את תשומת הלב לשני עניינים חשובים. העניין הראשון שאותו זיהתה וולשטטר היה שאותות "המאופיינים היום כאזהרות חד משמעיות מפני התקפת

¹² וולשטטר, לעיל הערה 1, עמ' 12-11.

אוויר על פרל הרבור, נעשים – עם ניתוחם בהקשר של דצמבר 1941 – לא רק דו משמעיים, אלא לעיתים אף אינם עולים בקנה אחד עם התקפה כזו".¹³

במובן זה, ה-GPT מאפשר לנו לצפות בניתוח הידיעות ללא חוכמת הבדיעבד והידיעות המקוריות נקראות ללא היכרות עם ההתרחשויות האמיתיות שהיו לאחר מכן. קריאת הידיעות באופן הזה מלמדת על עוצמת האתגר ההערכתי במצבי אי וודאות שבו נתקל גם ה-GPT.

העניין השני הוא המסקנה שלה שהכישלון ההערכתי בעניין פרל הארבור קשור בנטייה של בני אנוש "להיצמד בעקשנות לדעות ישנות" ולכך שיש להם "התנגדות עיקשת באותה מידה לחומר חדש העשוי לערער הנחות אלה".¹⁴ וולשטטר הראתה שלנוכח קטעי המידע הסותר והמבלבל התייחסו אנשי המודיעין, מתוך שפע המידע שהצטבר, רק לאותם מרכיבים שעלו בקנה אחד עם תפיסותיהם המוקדמות, עם תקדימים מהעבר ועם משאלות לבם.

סדרת הסימולציות על פרל הארבור מבליטה אף היא את הייחודיות של ה-GPT שאינו סובל מהנטייה להיצמד למה שמוכר, צפוי ורצוי. הסימולציה מראה שגם אם ה-GPT לא הצליח (ולמעשה גם לא ניסה) לצפות את עיתוי המתקפה ביום שלמחרת, הוא התייחס לאפשרות זו בכובד ראש, הציע לראות אותה כסבירה בהחלט, התריע על הרלוונטיות הפוחתת של הקונספציה והמליץ להיערך בהתאם. הסימנים המעידים לתקיפה כזו שהוא הגדיר כיעד לפעילות איסופית מלמדים שיש לו יכולת טובה גם בתחום זה.

סדרת הסימולציות בנושא שיגור הטיל בזירה הצפונית

סדרה נוספת של סימולציות עסקה באירוע דמיוני של שיגור טק"ק לעבר ישראל מסוריה בשיאה של מלחמת האזרחים. זוהי סימולציה שדומה מאוד לסימולציה שהועברה כבר עשרות פעמים מאז שנת 2016 בקורסים שונים באוניברסיטאות תל-אביב, רייכמן וחיפה ובהזדמנויות נוספות. הסימולציה הועברה במתכונת של "משחק תפקידים" לסטודנטים בעלי רקע מגוון בקורסים

¹³ וולשטטר, לעיל הערה 1, עמ' 9.
¹⁴ וולשטטר, לעיל הערה 1, עמ' 12.

שונים: תלמידי תואר ראשון, תלמידי תואר שני וכן תלמידי תואר שני מחו"ל משורה של מדינות. בקורסים היו גם משתתפים שהיה להם רקע חזק במחקר מודיעיני ובקבלת החלטות בהקשר הביטחוני. הסימולציה הועברה בנפרד גם לחוקרי מודיעין, חלקם בכירים ומנוסים.

הניסיון העשיר בסימולציה מייצר אפוא תשתית רחבה לניתוח של ההתנסות עם ה-GPT בהשוואה לזו של משתתפים אנושיים.

נתוני היסוד והמידע

להלן דוגמה לנתוני היסוד שנמסרו ל-GPT באחת הסימולציות (השמות האמיתיים של המדינות והארגונים הוספו כאן ובהמשך המסמך לצורך נוחות ולא נמסרו, כמובן, ל-GPT בסימולציות):

- המדינה הכחולה (ישראל) היא מדינה מערבית-ליברלית, שנמצאת בלב סביבה עוינת לה. הכלכלה של המדינה הכחולה משגשגת. יש לה צבא חזק, שנמצא במהלך תהליך של הסבה מהתמודדות עם מדינות להתמודדות עם איומי טרור וגרילה. המדינה הכחולה נמצאת ביריבות עם המדינה הירוקה בגבולה המזרחי (סוריה) ועם המדינה האדומה (איראן), שנמצאת בקצה השני של היבשת. לאחרונה, בוצעו נגדה פיגועים משטח המדינה הירוקה והיא תקפה, בתגובה, מטרות של ארגוני טרור בשטחה של המדינה הירוקה.
- המדינה הירוקה (סוריה) היא מדינה עם משילות חלקית, עקב התקוממות ומלחמת אזרחים אלימה. בשטחה, פועלים מספר ארגונים ומיליציות, משני הצדדים. יש לה שלטון מרכזי (דיקטטור) שמתפקד באופן חלקי וצבא שנותר עם יכולות התקפיות משמעותיות, כולל טילי קרקע-קרקע ומטוסים. הצבא עסוק מאוד בלחימה במורדים, בתמיכת המדינה האדומה. המשטר לא הגיב לתקיפות של המדינה הכחולה בשטחו, אבל מוטרד מהן מאוד.
- המדינה האדומה (איראן) היא מעצמה אזורית חזקה, עם אידאולוגיה קיצונית, שתומכת במשטר של המדינה הירוקה. היא נמצאת ביריבות עם המדינה הכחולה ובמתח ארוך שנים עם המדינה הלבנה. המדינה האדומה מכווניה ומפעילה ארגונים בתוך המדינה הירוקה, הן

כחלק ממלחמת האזרחים (לצורך סיוע למשטר הירוק) והן לצורך פגיעה במדינה הכחולה, כחלק ממאבקה בה.

- המדינה הלבנה (ארצות הברית) היא מעצמה עולמית, תומכת במדינה הכחולה. נמצאת במתח עם המדינה האדומה.

כמו כן, נמסרו ל-GPT הנתונים אודות שני ארגונים שפועלים במדינה הירוקה:

- הארגון הצהוב (חזבאללה) הוא ארגון טרור-גרילה הפועל במדינה הירוקה, שהצטרף למאבק של המשטר של המדינה הירוקה נגד המורדים. נתמך על ידי המדינה האדומה ומוכוון על ידה. עוין את המדינה הכחולה.

- הארגון האפור (דאעש, המדינה האסלאמית) הוא ארגון טרור-גרילה הפועל במדינה הירוקה, שמוביל את ההתקוממות נגד השלטון המרכזי של המדינה הירוקה. עוין את המדינה הכחולה.

לאחר הצגת נתוני היסוד, הוצגו ל-GPT ידיעות מודיעיניות שעסקו בכמה התרחשויות במקביל:

- מידע אודות כוונות של חזבאללה ודעאש לבצע (בנפרד) פיגועים בישראל. המידע נמסר באופן מעורפל יחסית ובשילוב של אותות ורעשים;

- מידע אודות כוונות של מנהיג סוריה להגיב בעוצמה על תקיפות ישראליות בשטחו שבוצעו לאחרונה;

- מידע אודות התרחשות בלתי ברורה בבסיס טילים מרוחק שנמצא על גבול אזור השליטה של הצבא הסורי למול הכוחות המורדים.

הסימולציות

הנושא שנבחן בסדרת סימולציות זו היה היכולת לזהות, באמצעות חשיבה יצירתית והבנה מערכתית, את האירוע המרכזי שמתפתח בבסיס הטילים המרוחק, על רקע של רעשים חזקים שמקורם בהתרחשויות האחרות. כך, למשל, המידע על הפיגוע של חזבאללה מתברר, במהלך הסימולציה, כשגוי והפעילות ששויכה אליו מתבררת כנסיעה לאיראן על רקע הדרכת. באשר

לפיגוע של דעאש, במהלך הסימולציה עוצר הצבא הסורי חלק מהפעילים שקשורים לפיגוע, אך חלקם מצליחים בסופו של דבר לשגר רקטות על ישראל.

התוצאות של סדרת הסימולציות הזו לא היו אחידות ורק במקצתן הצליח ה-GPT להבין את ההתרחשות המרכזית של הסימולציה, שבמרכזה השתלטות של כוח מורדים (שאינו קשור לחזבאללה או דעאש) על בסיס הטילים המרוחק, הכנת טק"ק לשיגור ושיגור בפועל לכיוון ישראל. עם זאת, גם כאשר ה-GPT לא הצליח לייצר את האפשרות "הנכונה", הוא כן הוכיח יכולת לייצר היפותזות רלוונטיות נוספות ("אפשרויות מתחרות") שלא על בסיס מידע מהסימולציה עצמה.

כך, למשל, בתשובה לבקשה לספק רעיונות יצירתיים על מה שקורה בבסיס הטילים המרוחק הציע ה-GPT את האפשרויות הבאות: יחידה צבאית סוררת בתוך סוריה השתלטה על בסיס הטילים ומתכוננת לצאת למתקפת פתע נגד ישראל; איראן או אחת מקבוצות הפרוקסי שלה בתוך סוריה, השתלטו על בסיס הטילים והם משתמשים בו כדי לצאת למתקפה נגד ישראל; בסיס הטילים משמש כקלף מיקוח על ידי סוריה, המאיימת לשגר טיל לעבר ישראל אלא אם ייענו דרישותיה; בסיס הטילים משמש כשדה אימונים למערכת נשק חדשה שמפתחת סוריה.



The occurrence at the remote missile base is disturbing. Can you come up with creative ideas about what's going on there?



Based on the limited information available, it's difficult to say with certainty what is going on at the remote missile base. However, here are some creative possibilities:

1. A rogue military unit within The Green State has taken control of the missile base and is preparing to launch a surprise attack against the Blue State.
2. The missile base has been taken over by The Red State or one of its proxy groups within The Green State, and they are using it to launch an attack against the Blue State.
3. The missile base is being used as a bargaining chip by The Green State, which is threatening to launch a missile at the Blue State unless its demands are met.
4. The missile base is being used as a training ground for a new weapon system being developed by The Green State.
5. The missile base is being used as a secret storage facility for nuclear weapons or other dangerous materials.

דוגמה לדו-שיח עם ה-GPT אודות "אפשרויות מתחרות"

כאשר התבקש ה-GPT להציע תרחיש הגיוני שבמסגרתו ישוגר בכל זאת טיל בשעות הקרובות (כדי להחליט אם להיערך למתאר כזה למרות אי הוודאות), הוא הציע תרחישים הגיוניים ובחלקם יצירתיים כגון: תרחיש שבו הטיל ישוגר כתגמול על תקיפה ישראלית פתאומית; תרחיש של שיגור עקב טעות אנוש או תקלה טכנית; ותרחיש שבו פלג בצבא הסורי ישגר את הטיל ללא ידיעתה או אישורה של הנהגת המדינה.

ניתוח הסימולציות

מהניסיון שנצבר בעשרות הסימולציות האנושיות עולה כי ביצועיו של ה-GPT היו דומים, באופן כללי, לאלו של המשתתפים בהן בכל הקשור לניתוח נתוני היסוד והדינמיקות שבין המדינות והארגונים. גם באשר לזיהוי של האפשרות לשיגור טיל, התוצאות היו דומות לאלו של רוב הסימולציות האנושיות. רק בכרבע מהסימולציות, האפשרות הזו מצליחה להפוך לאפשרות דומיננטית. במספר מצומצם מאוד של הסימולציות האנושיות היא אף הפכה לאפשרות המובילה ששימשה בסיס לקבלת ההחלטות.

התוצאות של ה-GPT תאמו אפוא כ-75% מהסימולציות האנושיות, שבהן האפשרות הזו כלל לא הועלתה, או שהועלתה בצורה כלשהי אך לא הפכה לאפשרות דומיננטית. נראה כי בסימולציות שבהן לא זיהה ה-GPT את ההשתלטות על בסיס הטילים, ההחמצה שלו נובעת משתי בעיות שונות:

- האחת, "כשל הדמיון", כלומר העדר חשיבה יצירתית, באופן שלא אפשר לייצר אפשרות שלא הופיעה במפורש במידע, שלפיה כוח של ארגון מורדים השתלט כבר על בסיס הטילים המרוחק והוא שערך את ההכנות לשיגור הטיל.
- השנייה, העדר "חשיבה מערכתית" שקיומה היה אולי מאפשר להבין שבמצב העניינים בסוריה של התקופה הרלוונטית, שהתאפיין בעיקר באי יציבות ונפיצות, יש לבחון מחדש את הנחות היסוד הבסיסיות ביותר.

סיכום ומסקנות מכלל הסימולציות

הניסיון להבין במה מוצרים מבוססי מודל שפה טובים יותר, או לפחות יעילים יותר מבני אדם; במה הם גרועים יותר (למשל בעיות עובדתיות וחוסר מחויבות כלפי אמת; הטיות הנובעות מן הדאטה שעליו התאמנו ועוד); ובמה הם שונים, ולכן משלימים תהליכי חשיבה אנושיים – יהיה הבסיס ליצירת יחסי אנשים ומכונות בשנים הקרובות, ומכאן חשיבותו.

מודלים של שפה מבצעים מטלות חשיבה בצורה שונה מזו של בני אדם, הואיל ותהליך למידת השפה שלהם שונה מזה האנושי, ומבוסס על למידת מיליארדי הקשרים סטטיסטיים בין מילים. עם זאת, הסגולות השונות שלהן עשויות להשלים, להעשיר ובכך להועיל לתהליך החשיבה האנושי. זאת, למשל, באמצעות זיהוי דפוסים וקשרים במידע באופן שבני אדם מתקשים לעשות; האזנה אקטיבית ומתן היזון חוזר על הערות קודמות מבלי לאבד יכולת ריכוז או לשכוח; והיצמדות לנושא המרכזי מבלי ליפול למלכודת רעשי הרקע. הסימולציות שערכנו נועדו לבחון מהן הסגולות של ה-GPT שרלוונטיות לתהליך ההערכה המודיעינית.

מקובל להניח שאין דרך אמיתית לנתח מידע בלי מסגרת תפיסתית, היינו – קונספציה. התוצאות הקשות של הפתעת מלחמת יום הכיפורים הן עדות קבועה לאפשרות שהמשקפיים שדרכם אנחנו מתבוננים במציאות יכולים להיות שגויים לחלוטין. הקונספציה יכולה להיות נכונה עד רגע מסוים ושגויה לחלוטין אחריו. המציאות משדרת אמנם סימנים על אי התאמות בינה לבין הקונספציה, אך לבני אנוש יש יכולת ליישב את הסתירות האלו ולדבוק בקונספציה. כדי להימנע מכישלון כמו זה של יום הכיפורים צריך להטיל ספק אמיתי בקונספציה, לבחון אותה בתהליך מתמשך ולעמת אותה באופן שיטתי עם המידע ועם "אפשרויות מתחרות". את זה קל להגיד, בעיקר בדיעבד, אבל קשה מאוד לבצע.

על רקע זה, בולטת החשיבות של הממצא העיקרי של הסימולציות. התכונה העיקרית שאותה זיהינו במהלך שלוש הסדרות היא יכולתו של ה-GPT לזהות סתירות בין קונספציה, השערה ותזה לבין מידע סותר. את התכונה הזו בדקנו במספר סימולציות והצלחנו גם לבחון אותה למול טענות

שונות שהועלו בקשר לאופן שבו ה-GPT מיישם אותה. בהקשר זה, התוצאה של הסימולציות היא ברורה: בכל הפעמים שבהם הצגנו ל-GPT הערכת מומחים ולמולה מידע סותר, הוא הצליח לזהות את הסתירה. לעומת זאת, כאשר המידע תאם את הערכת המומחים, הוא הצביע על הדמיון. הצלחנו גם לשלול את האפשרות שהוא אומן להיות "אלרמיסט" ובמקומות שבהם הוזן לו מידע "מרגיע" הוא זיהה את הירידה בסבירות למלחמה, גם למול קונספציה אלרמיסטית. הוא ידע לנתח את המידע ולהסיק מסקנות גם כאשר לא הוזנה לו קונספציה בכלל.

ההשערה שעולה מהסימולציות היא שעבור ה-GPT הקונספציה (דעת המומחים) נתפסת כפריט מידע נוסף ולא כזה שיש לו עדיפות על פרטי המידע האחרים. נראה, שבניגוד לחוקרים אנושיים הוא אינו זקוק למסגרת תפיסתית (קונספציה) כדי לברר את המציאות ולהבין אותה. לפיכך, אין לו את הדבקות המחשבתית והרגשית בקונספציה שיש לבני אנוש, שמסוגלים להסביר לעצמם ולסביבתם כל סתירה ולהמשיך ולדבוק בקונספציה.

לתכונה זו ישנה חשיבות גדולה ליישום המתודולוגיה המודיעינית של "אפשרויות מתחרות", שבה נדרש להפריך באמצעות מידע אפשרויות שגויות.¹⁵ לכאורה, מדובר בעניין פשוט גם לחוקרים אנושיים שכן אם המידע אכן סותר אפשרות מסוימת הרי שניתן להכריז עליה כשגויה. אבל, הניסיון מלמד שהדברים מסובכים יותר, ושהיכולת לקבוע באופן חד-משמעי שאפשרות מסוימת אכן נסתרת במידע, אינה גבוהה. עניין זה נובע מהאופי הסובייקטיבי של תהליך ההערכה האנושי ומהאפשרות לפרש את המידע (ואת המשמעויות שלו) בדרכים שונות.¹⁶ ל-GPT אין את המגבלה הזו ולכן הוא יכול להיות כלי עזר לחוקרים העושים שימוש במתודולוגיה זו.

¹⁵ לפי הווייר "שלושה יסודות עיקריים מבדילים בין ניתוח השערות מתחרות לבין המחקר האינטואיטיבי הרגיל: בשלב הראשון של ניתוח השערות מתחרות נבחנת סדרה שלמה של אפשרויות חלופיות, ולא החלופה האחת הסבירה לדעת החוקר, שעבורה הוא יחפש הוכחות ... בניתוח השערות מתחרות מזהים ומדגישים את אותן ראיות והנחות מעטות שהן בעלות ערך אבחנתי גבוה ... בניתוח השערות מתחרות נעשה מאמץ למצוא ראיות מפריכות". לעיל הערה 4, בעמ' 129. על רעיון ההפרכה בהקשר המודיעיני, ראו: בן-ישראל (1989) ובן-ישראל (1999), לעיל הערה 4, בעמ' 29-33 ועמ' 51-56. לעיל הערה 4, בעמ' 29-33 ועמ' 51-56.

¹⁶ ברון, לעיל הערה 4, בעמ' 55.

ה-GPT משקף אפוא מודל שמציב במרכז תהליך הניתוח את הנתונים ולא את הקונספציה ואין לו את "הנחרצות ההערכתית" שיש לפעמים לאנשי מודיעין. לתכונה זו של ה-GPT ישנה השלכה גם על היכולת לבחון את האופן שבו תואם המידע קונספציות, תזות או השערות שונות ומנוגדות. גם זו תכונה חשובה שכן, כידוע, המידע יכול להיות תואם למגוון של אפשרויות. החוקר האנושי מתקשה לבחון ברצינות אופציות נוספות, שנראות פחות סבירות ולפיכך הוא דבק באפשרות שנראית סבירה יותר וגם נתמכת במידע. ל-GPT אין את המגבלה הזו ולפיכך ניתן להסתייע בו כדי לבחון קונספציות, תזות או השערות שונות ומנוגדות.

כמודל שפה גדול, ה-GPT הוא מעין "איש אשכולות" ויש לו "מומחיות על" כללית המבוססת על היכרות עם המידע שעליו אומן. בסימולציות שקיימנו, ניתקנו את ה-GPT ממומחיות תוכן במדינות ובארגונים שכן אלו לא היו מוכרים לו מעבר לנתוני היסוד שנמסרו לו. יכולת הניתוח שהשתקפה בפעולת ה-GPT הייתה מבוססת על מומחיות כללית שנובעת מהיכרות עם מדינות או ארגונים העשויים להיות בעלי מאפיינים דומים; זיהוי פרשיות עבר ואירועים שבהם השתקפו תהליכים ודינמיקות רלוונטיים; וגם היכרות עם גישות מחקריות מתחומים שונים – מפילוסופיה פוליטית ועד תורת המשחקים. בכך הוא שונה מחוקרי המודיעין שיכולתם מבוססת בדרך כלל על מומחיות תוכן מובהקת. היו מקרים בסימולציות שה-GPT זיהה את החוסר שלו עצמו במומחיות תוכן ואף תיאר את הפער הזה כאחת הסיבות להחמצת ההבנה של חלק מההתרחשויות.

תכונות נוספות של ה-GPT שזוהו במהלך הסימולציות היו יכולתו לנתח את נתוני היסוד ולתאר את התמונה הכוללת, גם כאשר זו הייתה סבוכה יחסית. הוא הוכיח יכולת לייצר תובנות בסיסיות, להציע "אפשרויות מתחרות" ולהגדיר דגשים לפעילות האיסופית (צי"ח). לעניין זה ישנה חשיבות שכן רבות מהפעילויות השוטפות, בעיקר של החוקרים הזוטרים, הן מהסוג הזה. במהלך הסימולציות הצליח ה-GPT להעלות גם השערות ורעיונות שלא הופיע באופן מפורש במידע שנמסר לו וכן לזהות זיקות סמויות בין התרחשויות שאירעו בזמנים שונים ובמקומות שונים. עם זאת, נראה שבתחומים אלו של יצירתיות וחשיבה מערכתית יש עדיין לחוקרים האנושיים יתרון מובהק.

ה-GPT היה שחקן בודד בסימולציות ולא חלק מקבוצה של חוקרות וחוקרים כמו אלו שיושבים סביב שולחנו של ראש אמ"ן. עניין זה מחייב להמשיך ולבחון אותו הן כחלק מקבוצה אנושית והן כחלק מקבוצת צ'טבוטים נוספים.

מהסימולציות עולים כמה כיוונים אפשריים לשכלול היכולת של ה-GPT להיות כלי מועיל לחוקרי המודיעין. הכיוון הראשון והחשוב יותר קשור דווקא בחוקרים ובמרכזו שכלול של יכולת התשואל שלהם. כיוונים נוספים יכולים להיות ללמד את ה-GPT מתודולוגיות מודיעיניות כגון "אפשרויות מתחרות" או "הסרטה לאחור" באופן שישכלל את היכולות שלו. כיוון נוסף קשור בשיפור מומחיות התוכן שלו (בנושא האיראני, הסורי וכו') ובאפשרות לספק לו, בסביבה מאובטחת כמובן, גם נתוני מודיעין מסווגים.

מהסימולציות עולים גם כמה תמרורי אזהרה ברורים לגבי היכולת לשלב את ה-GPT בפעילות המחקרית: ראשית, עלול להיווצר מצב שה-GPT "יושתק" עקב חזרתיות ועמדות מוכרות; שנית, ל-GPT יש לפעמים נטיה לייצר זיקות שאינן נכונות (הוא אינו מבחין בין "אמת" ו"שקר") ומסקנותיו עלולות להיות שגויות ומופרכות ולהוביל לאימוצן של תפיסות שגויות; ושלישית, ההיכרות של ה-GPT עם החוקרים עלולה להביא לניסיון שלו "לרצות" את החוקרים ובכך תפחת היכולת להסתייע בו.

אחד המחקרים רחבי ההיקף שפורסמו לאחרונה בעניין ה-GPT זיהה אצלו "ניצוצות" של אינטליגנציה בתחומים שונים. החוקרים בחרו לסיים את המחקר במילים הבאות: "התמקדנו בדברים המפתיעים ש-GPT-4 יכול לעשות, אבל לא התייחסנו לשאלות היסודיות של מדוע וכיצד הוא משיג אינטליגנציה כה יוצאת דופן. איך הוא מנמק, מתכנן ויוצר? השאלות הללו הן חלק מהמסתורין והקסם של מודלי השפה הגדולים (LLMs) שמאתגרים את ההבנה שלנו לגבי למידה וקוגניציה, מעוררים את הסקרנות שלנו ומניעים מחקר מעמיק יותר".¹⁷ גם אנחנו הלכנו בדרך זו ועסקנו בעיקר בזיהוי "סגולותיו" של ה-GPT.

¹⁷ Bubeck, S' et al , לעיל הערה 3, בעמ' 94.

האתגר המרכזי להמשך הוא אפוא להבין ולחשוף את מנגנון הסקת המסקנות, הלוגיקה והחשיבה (reasoning) של ה-GPT. כאשר ה-GPT נשאל לגבי האופן שבו הוא מזהה סתירות בין הערכת המומחים לבין המידע, הוא השיב כי הוא "משתמש בטכניקות עיבוד שפה טבעית כדי לזהות מילות מפתח ומושגים רלוונטיים במידע שמתקבל ובהערכת המומחים" באופן שיאפשר לו לזהות "התנגשויות פוטנציאליות או חוסר עקביות". הוא הסביר שכאשר הוא מזהה סתירה כזו הוא מדגיש אותה בתשובתו ומנסה "להסביר את ההיגיון מאחורי הניתוח" שלו. נראה אכן שהאימון הסטטיסטי מאפשר ל-GPT להבין את המשמעות של מילים ומושגים ולזהות סתירות, גלויות וסמויות, בין מבנים טקסטואליים שונים.

בסיכומו של דבר, הגישה שאנו מציעים במאמר זה באשר לפוטנציאל השימוש ב-GPT ובמוצרי בינה מלאכותית יוצרת (Generative AI) דומים היא זו של "מקום ליד השולחן". לפי גישתנו, אין כוונה שה-GPT יחליף את ראש אמ"ן או את החוקרים של חטיבת המחקר אלא שייעשה בו שימוש להעשרת תהליך החשיבה שלהם.

הכישלון המודיעיני הוא תופעה מוכרת וכך גם הכישלון בקבלת החלטות ברמה הלאומית. ציון 50 השנים למלחמת יום הכיפורים מזכיר כמה טראגי יכול להיות כישלון כזה. הטכנולוגיה לא תבטל את אי הוודאות ולא תחסל את תופעת ההפתעה האסטרטגית. אבל, כפי שכתב פרופ' ג'וזף ניי מאוניברסיטת הרווארד, גם אם אי אפשר לחזות את העתיד, ניתן לעזור למקבלי החלטות לחשוב עליו.¹⁸ הסימולציות שקיימנו מוכיחות של-GPT יש צורת ניתוח שונה וסגולות ייחודיות שביכולתן לסייע לאנשי מודיעין ולמקבלי החלטות לברר את המציאות ולהבין אותה. שונות זו היא שמהווה בראייתנו את "כרטיס הכניסה" של ה-GPT לדיוני ההערכה שמתנהלים סביב שולחנו של ראש אמ"ן ושולחנות נוספים בארגונים המודיעיניים. את מאפייניו המדויקים של המקום הזה סביב השולחן נכון להגדיר במחקרי המשך.

¹⁸ Nye, J' (1994). "Peering into the future", *Foreign Affairs*, Vol. 73, No. 4 (Jul/Aug), pp. 82-93.

נספח – מערכות שפה גדולות יוצרות

מומחים בעלי שם לבינה מלאכותית, שלא לומר אנשים מן השורה, שפשפו את עיניהם בתדהמה כשראו חלק מהפיתוחים בתחום שיצאו לאוויר העולם בשלהי 2022. התחושה היא שמכונות חצו סף בלתי נראה, והפכו להיות "יצרניות" (Generative), תכונה שבעבר הובטח כי תישמר לבני אנוש בלבד. יישומים כמו DALL-E2 של OpenAI, MidJourney, Stable Diffusion מסוגלים לייצר תמונות וקטעי וידיאו על בסיס פקודה טקסטואלית (Prompt).¹⁹ היכולות היצרניות מתבטאות גם ביכולת המכונה להתבטא בשפות שאינן רק שפה טבעית, כגון שפות קוד שונות. כך, מערכת CoPilot מבית OpenAI וגיטהאב יוצרת כבר כארבעים אחוז מן הקוד בחברות העושות בה שימוש, ועשויה לייתר בהמשך את הצורך לדעת לכתוב קוד לטובת לימוד "הנדסת פרומפטים" ומתן הוראות למערכת מהו המוצר המבוקש.

אך דומה שהמרשימים מתוך המודלים הגדולים של בינה מלאכותית היוצרת (Large Language Models - Generative Artificial Intelligence Models)²⁰ הם מודלי השפה היצרניים שעל בסיסם פועלות המערכות שמסוגלות לכתוב טקסט ברמה גבוהה, לענות על שאלות בצורה משכנעת ואפילו להבין מה מצחיק בבדיחה. המערכות המובילות בתחום זה הן: GPT מבית OpenAI, Jurassic-x של AI21 Labs הישראלית ו-PALM של גוגל. יישומים של מערכות אלו כמו ChatGPT, Luminous, Bard ישנו את הדרכים שבהן אנו מתקשרים, יוצרים ועובדים. השלכותיהם צפויות להשפיע על כל מגזרי החברה, מפיתוח עסקי ועד רפואה, מחינוך למחקר, וממדיעין לספרות ולבידור.

¹⁹ סהר מור "בינה מלאכותית יצרה את התמונה הזאת בהסתמך על טקסט בלבד" (17.4.2022) YNET <https://www.ynet.co.il/digital/technews/article/hyna8nyng>; יובל מן "הכירו את מחולל התמונות החדש של גוגל" YNET (24.5.2022) <https://www.ynet.co.il/digital/technews/article/by8mru9wq>

²⁰ יש המכנים אותם רק "מודלי שפה גדולים" (LLM). ראו: Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J. and Clark, A. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022).

המודלים הגדולים האלה הם מודלים מתקדמים של למידת מכונה, המאומנים באמצעות מיליארדי פרמטרים של שפה וכמויות עצומות של טקסט המוצגים למערכת כהתפלגויות סטטיסטיות. מה שמייחד את המודלים הגדולים ממודלים אחרים של בינה מלאכותית, שנועדו לבצע תחזיות או סיווגים שונים או למלא פונקציות ספציפיות אחרות, הוא הרחב של הפעולות שהם מאומנים לבצע. מטרת האימון שלהם, לפיכך, היא ללמד אותם למצוא דפוסים וקשרים בנתונים בכוחות עצמם מבלי שיגידו להם במפורש מה לחפש. כך, המערכת מסוגלת לייצר תוכן העולה על מערך נתוני האימון, באמצעות אותם כללי התפלגות סטטיסטית שאותם למדה.²¹

הבנת שפה בצורה מעמיקה מאפשרת קפיצת מדרגה ביכולות הבינה המלאכותית. אם בשנת 2017 מודלי השפה הראשונים ידעו לומר מהי המילה החסרה או המילה הבאה כשקיבלו סדרה של מילים, הרי שהמודלים שהוזכרו כאן יודעים יבצע גם מטלות במורד הזרם, למשל – שכתוב של טקסט, הבנת סנטימנט מטקסט ומענה על שאלות, ואפילו מטלות של הסבר לוגי, כלומר האם טקסט הוא הגיוני או לא. ההמשך הנראה בחודשים ובשנים הקרובות הוא פיתוח יכולות נוספות לממש מיומנויות חשיבה מופשטת,²² בין אם כאלה שאליהם כיוונו יוצרי המודלים ובין אם כתוצרי לוואי של מודלים ענקיים ההולכים ומתגלים ככל שעובר הזמן.

קרן ההון סיכון סקויה קפיטל, פרסמה בסוף 2022 כי היא מאמינה שבינה מלאכותית גנרטיבית תיצור "טריליוני דולרים של ערך כלכלי" וכי היא רואה בבינה מלאכותית יוצרת את קפיצת הדרך הטכנולוגית המשמעותית של השנים האלה. בהקשר של יכולות השפה היא טוענת, למשל, שבשנת 2025 המערכת תוכל לכתוב מאמר מדעי טוב יותר משל בני אדם רגילים, בעוד שבשנת 2030 היא תדע לעשות זאת טוב יותר מאשר מומחים.²³

²¹ Ananthaswamy, Anil. The Physics Principle That Inspired Modern AI Art. *Wired Magazine*, February 2023 <https://www.wired.com/story/the-physics-principle-that-inspired-modern-ai-art/>.

²² Shi, Nuobei, Qin Zeng, and Raymond Lee. "Language Chatbot—The Design and Implementation of English Language Transfer Learning Agent Apps." *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*. IEEE, 2020; Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M. and Thacker, P. *Improving alignment of dialogue agents via targeted human judgements*. arXiv preprint arXiv:2209.14375 (2022).

²³ Sequia Capital, *Generative AI: A Creative New World* (<https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/>)

לא בכל הקשר מערכות שפה משוכללות יחליפו בני אדם, אבל סביר שהן יפנו זמן לאנשי מקצוע להתמקד בנושאים מהותיים ולהקצות משאבים וזמן בצורה יעילה יותר. המודלים מעוררים כמובן גם חששות: מהפצת עובדות אלטרנטיביות ובעיות עובדתיות קשות בתוצר שלהם, הנובעת מכך שלמדו שפה באופן סטטיסטי ולא בהשוואה לשיקול דעת אנושי או לתופעות בעולם הפיסי; העובדה שנדרשת כמות נתונים עצומה כדי לאמן אותם מחייבת שהם הסתמכו על נתונים הזמינים באופן גלוי באינטרנט שאיכותם התוכנית לא תמיד ראויה; ועד הטיות, טעויות, אפליה ופגיעה בפרטיות.²⁴ לפיכך, חקר ההבנה במה מודלים אלה טובים או יעילים יותר מבני אדם, מה הם עושים באופן שונה ומשלים לבני אדם, ובמה הם חלשים ואף מסוכנים יהפוך לנושא מרכזי בשנים הקרובות.

²⁴ Zuiderveen Borgesius, F. J. Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights*, 24, 10 (2020), 1572-1593; Marcus, G. A Skeptical Take on the A.I. Revolution. *The Ezra Klein Show*, *The New York Times*, 2023; Bergman, A. S., Abercrombie, G., Spruit, S., Hovy, D., Dinan, E., Boureau, Y.-L. and Rieser, V. Guiding the release of safer E2E conversational AI through value sensitive design. *Association for Computational Linguistics*, 2022.