# מלחמת ההשפעה ברשת: בינה מלאכותית כנשק שובר שוויון

סא"ל ת"ז - משרת באמ"ן

**הבינה המלאכותית היא כלי עוצמתי לחיזוק האמינות של התוקף ולהעצמת הביצועים שלו בקנה מידה חסר תקדים שמעמיק את האיומים והסיכונים בהקשרים של המערכה על התודעה. ואמנם, מדינות כמו סין ורוסיה משקיעות הון רב בפיתוח יכולות בתחום זה בשנים האחרונות, ולצד זאת, נשמעות קריאות מצד מומחים בטכנולוגיה לעצור את הפיתוחים בטענה שאלה מהווים סיכון מהותי לאנושות. המאמר מסכם את העקרונות לחשיבה, תכנון וביצוע מבצע השפעה במרחב הסייבר בעידן של בינה מלאכותית, מנתח כיצד הבינה המלאכותית מאפשרת קפיצת מדרגה באיכות ובעוצמת פעולות אלה ומציג את האתגרים המתהווים בעקבות שימוש בכלים אלה. במאמר נטען כי הפיתוחים הטכנולוגיים, והמשך המגמה הנוכחית צפויים להגביר את המוטיבציה של בעלי אינטרס להתערב ולהשפיע באופן שעלול לערער את תפיסת המציאות ולהוביל לנזק חברתי ופוליטי ממשי. נוכח פוטנציאל הנזק, דרושה מודעות גבוהה יותר לסכנות כמו גם השקעה בפיתוח כלי ניטור ומניעה ברמה הלאומית.**

## מבוא

עידן המידע ו"מהפכת הבינה המלאכותית" - אליה וקוץ בה. לצד השיפור הדרמטי באיכות החיים הודות לתמורות הטכנולוגיות ונתוני העתק, המרחב הקיברנטי הפך בשנים האחרונות לזירת פעולה מועדפת בעבור מבצעי השפעה ולוחמה פסיכולוגית. תפיסת הפעולה של מתקפות סייבר לטובת השפעה (Influence Network Cyber) שמטרתן לשבש את סביבת המידע של מושא התקיפה, התפתחה עד לכדי מצב שבו קשה ולעיתים בלתי אפשרי לאתר ולמנוע אותה. בשנים האחרונות נרשמת עלייה במאמצים של מדינות או שחקנים תת-מדינתיים לחדור לקהלי יעד יריבים, ולעצב את תפיסת המציאות שלהם באמצעות מידע כזב המוכתב על-ידי אינטרסים זרים. מחקרים אמפיריים טרם הציגו דרך מדעית למדוד את היקף ההצלחה של מבצעי ההשפעה, אולם ניתן לקבוע שהם מתפשטים מהר יותר מבעבר בחסות הטכנולוגיה, ומצליחים לגרום לערעור וזעזוע חברתי עמוק.

הבינה המלאכותית שימשה בראשיתה את תוקפי הסייבר כמנוע מיצוי מידע המאפשר לעבד נתוני עתק וללמוד על אודות מגמות ונראטיבים שבאמצעותם ניתן להשפיע.[1] כיום, נוכח הפיתוחים האחרונים, בפרט בתחומי עיבוד השפה הטבעית (NLP), הבינה המלאכותית היא כלי עוצמתי לחיזוק האמינות של התוקף והעצמת הביצועים שלו בקנה מידה חסר תקדים שמעמיק את האיומים והסיכונים.[2] לא בכדי, מדינות כמו סין ורוסיה משקיעות הון רב בפיתוח יכולות בתחום זה בשנים האחרונות, ולצד זאת, נשמעות קריאות מצד מומחים בטכנולוגיה לעצור את הפיתוחים בטענה שאלה מהווים "סיכון מהותי לאנושות".[3]

מאמר זה מסכם את העקרונות לחשיבה, תכנון וביצוע מבצע השפעה במרחב הסייבר בעידן של בינה מלאכותית. המאמר ינתח כיצד הבינה המלאכותית מאפשרת קפיצת מדרגה באיכות ובעוצמת פעולות אלה ("נשק שובר שוויון"), ויציג את האתגרים המתהווים בעקבות שימוש בכלים אלה. במאמר נטען כי הפיתוחים הטכנולוגיים, והמשך המגמה הנוכחית צפויים להגביר את המוטיבציה של בעלי אינטרס להתערב ולהשפיע באופן שעלול לערער את תפיסת המציאות ולהוביל לנזק חברתי ופוליטי ממשי. נוכח פוטנציאל הנזק, דרושה מודעות גבוהה יותר לסכנות כמו גם השקעה בפיתוח כלי ניטור ומניעה ברמה הלאומית.

### התקפה בסייבר ככלי השפעה

מבצעי התקפה בסייבר (CNA) הפכו חלק בלתי נפרד מתפיסת הלחימה של מדינות רבות בעולם, כמהלך משלים ללחימה הקונבנציונלית או כמהלכים העומדים בפני עצמם. בכללותם, מבצעים אלה מתנהלים במרחב הדיגיטלי, ומאפשרים לתוקף לגרום

1 Randy Bean, "How Big Data Is Empowering AI and Machine Learning at Scale," *MIT Sloan Management Review*, May 8, 2017, https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale

2 Shannon Bond, "It Takes a Few Dollars and 8 Minutes to Create a Deepfake. And That's Only the Start," *NPR*, March 23, 2023, sec. Technology,https://www.npr.org/2023/03/23/1165146797/it-takes-a-few-dollars-and-8-minutes-to-create-a-deepfake-and-thats-only-the-sta

3 "מומחי בינה מלאכותית קוראים לעצור פיתוח של מערכות AI עוצמתיות: "מהוות סיכון מהותי לאנוש", כלכליסט, https://www.calcalist.co.il/calcalistech/article/ryd0nlz11h

במהירות נזק רב למערכות דיגיטליות של יריב.[4] התקפה בסייבר מאפשרת לפגוע בכל תשתית נתונים או מערכת המחוברת לרשת האינטרנט ולהסב נזק בקנה מידה לאומי, בין אם מכוונת נגד תשתיות קריטיות או מאגרי מידע שמכילים מידע פרטי ורגיש של אזרחים. פוטנציאל הנזק והעלות נמוכה הם שהפכו את ההתקפה בסייבר לכלי אטרקטיבי. מדינות או שחקנים תת-מדינתיים המעוניינים לפעול **במרחב האפור**, על-מנת להרתיע או לדחוף את יריביהם לפעולה, יבחרו פעמים רבות לפעול בממד. מאחר ופעולות אלה נחשבות על-פי רוב מתחת לסף המלחמה, חלק ניכר ממבצעי התקפה, מבוצעים במטרה להפעיל לחץ, להשפיע על קהל יעד ועל מקבלי החלטות.[5]

בשנים האחרונות אנו עדים לגיוון והתרחבות של מערכות ההתקפה וההשפעה בממד הסייבר. בעוד שבעבר המערכות התאפיינו בחדירה ושיבוש תשתיות קריטיות, כיום הן מכוונות **לחדירה ושיבוש המוח האנושי**, במטרה לעצב מחדש סביבה חברתית ופוליטית בהתאם לאינטרסים ולשיקולים זרים. "עידן המידע והדיגיטל" מאפשר פוטנציאל חדש של מבצעים שתכליתם הבלעדית היא השפעה - ICO (Influence Cyber Operations).[6] מבצעים אלה מתקיימים **במרחב אפור והיברידי יותר**, בתוך הרשתות החברתיות, בסביבתם הפרטית והלגיטימית של מיליארדי משתמשים בעולם. מבצעי ההשפעה כוללים העצמה של מרכיבי הלוחמה הרכה, **ההונאה והלוחמה הפסיכולוגית,** תוך ניצול האינטרנט המייצר נגישות חסרת תקדים לקהלי יעד.[7]

## עקרונות לתכנון מערכת השפעה בעידן הבינה המלאכותית

התמורות הטכנולוגיות בתחומי ה"בינה המלאכותית" הפכו את מערכות ההשפעה למאמץ משתלם. מגוון הכלים העומד לרשותם של התוקפים הוא רחב והיכולת להתאים את הכלים לתכלית המבצעית ולצרכיו האישיים של התוקף קיימת בעבור שחקנים קטנים כגדולים. ככל שהכלי פשוט ומוכר יותר, כך הוא קל יותר לזיהוי ובהתאם הישגיו מוגבלים. כדי להצליח לקיים מערכת השפעה רחבה המכוונת לשינוי תודעה ועיצוב דעתם של רבים, התוקפים נערכים למהלך כאל מבצע מיוחד הכולל תהליך רב-שלבי.[8,9] בכל אחד מהשלבים אשר יפורטו להלן, תוצגנה ההזדמנויות שמאפשרת "הבינה המלאכותית" לתוקפי ההשפעה ובהתאם האתגרים שעמם צפויים להתמודד קהלי היעד - אישים, ארגונים ומדינות.

### 1. חשיבה, תכנון ואיסוף מודיעין

משימת עיצוב התודעה היא משימה מורכבת הדורשת אורך רוח, תחכום וכישרון רב. תוקף השפעה אשר מבקש לערער על הנחות שנתפסות בידי ציבור שלם של אנשים כאמת, צריך להכיר בכך שבאופן בסיסי המוח האנושי אינו קל לשכנוע[10]. עוד בטרם הדיון על הכלים הדרושים לפעולה, השלב הראשון במבצע השפעה הוא שלב תיאורטי-מחשבתי שבוחן בצורה ביקורתית את ההיגש הרצוי, את החלופות העומדות לרשותו, הסיכונים הכרוכים בתהליך, ואת הסבירות לכישלון. מוטב כי בשלב זה לא יהיה רעיון סדור אחד אלא תהליך סיעור מוחות שבסופו מגוון תוכניות אפשריות להיגש הרצוי. התוקף זקוק **למסד נתונים רחב** ככל הנדרש על קהל היעד, על ההקשר התרבותי-ההיסטורי ועל אירועי השעה, באופן שיאפשר להכווין את תהליכי החשיבה ולעצב את הנרטיב ואת המהלך בצורה המדויקת והמתאימה ביותר. שלב החשיבה נועד למנוע מצב שבו מדמיינים את תמונת הניצחון בטרם דנים על תנאי הפתיחה.

ישנה משמעות רבה **לאיסוף המודיעין** ולהבנה עמוקה ומדויקת של התוקף על סביבת הפעולה. מהלך השפעה שאינו מותאם ל**הלך הרוח** או **מחזור החדשות** של קהל היעד סופו כישלון חרוץ. גורם שמעוניין להעצים דעה קיימת או להציע דעה אחרת, חייב להכיר את המאפיינים ואת המגמות הבולטות בשיח, ובפרט את הנושאים השנויים במחלוקת ואת עמדות הקיצון של הקהל בנושאים אלה.

4 רחל ארידור הרשקוביץ, תהילה שוורץ אלטשולר, ועידו סיון סביליה, מהו סייבר? חלק א': על מרחב הסייבר, תקיפות סייבר והגנת סייבר, מחקר מדיניות 171 ירושלים: *המכון הישראלי לדמוקרטיה*, 2021.

5 Emilio Iasiello, "Cyber Attack: A Dull Tool to Shape Foreign Policy," in *2013 5th International Conference on Cyber Conflict (CYCON 2013)*, 2013, 1–18.

6 Pascal Brangetto and Matthijs A. Veenendaal, "Influence Cyber Operations: The Use of Cyberattacks in Support of Influence Operations," in *2016 8th International Conference on Cyber Conflict (CyCon)* (2016 8th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia: IEEE, 2016), 113–26, https://ieeexplore.ieee.org/document/7529430

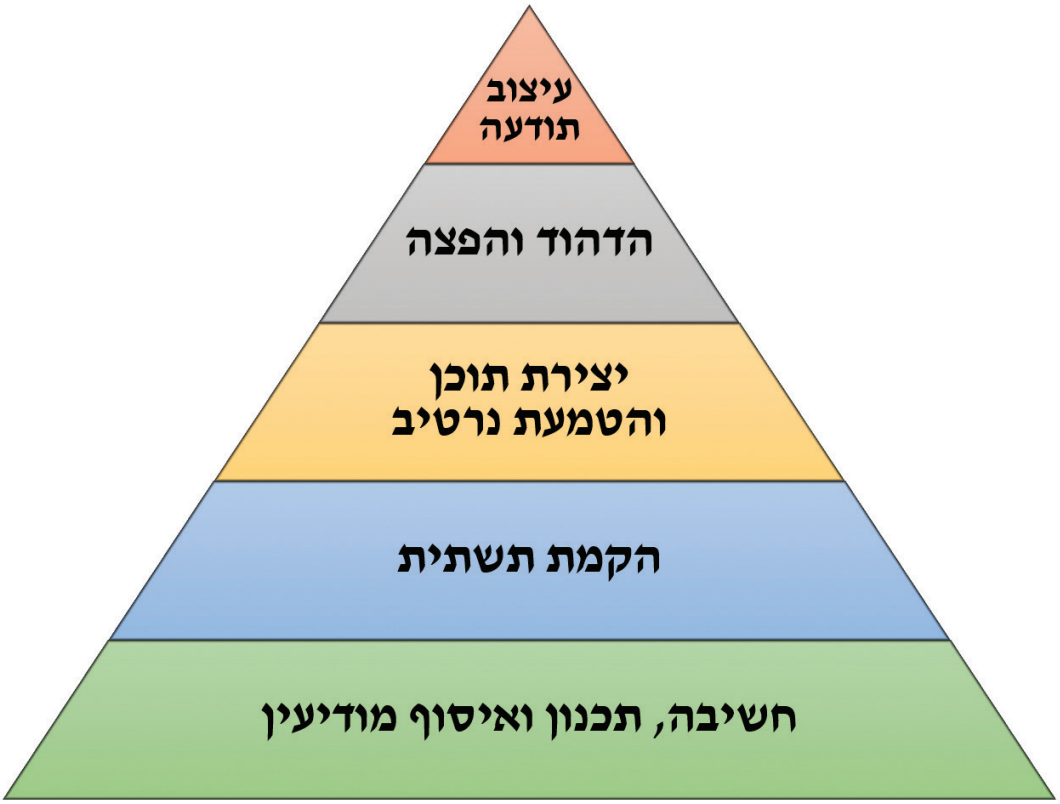
7 Ben Quinn, "Revealed: The MoD's Secret Cyberwarfare Programme," *The Guardian*, March 16, 2014, sec. UK news,https://www.theguardian.com/uk-news/2014/mar/16/mod-secret-cyberwarfare-programme

8 תהליך התכנון שמוצג במאמר זה נעזר במודל RICHDATA שהוצג לראשונה במכון CSET של אוניברסיטת ג'ורג' טאון.

9 Katerina Sedova et al., "AI and the Future of Disinformation Campaigns: Part 1: The RICHDATA Framework" *Center for Security and Emerging Technology*, December 2021), https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns

10 Alizabeth Svoboda, "Why Is It So Hard to Change People's Minds?," Greater Good Magazine, 27 June 2017, https://greatergood.berkeley.edu/article/item/why_is_it_so_hard_to_change_peoples_minds

**איור 1 - תכנון מבצע השפעה**




**מבצע השפעה יצליח כאשר הוא מעצים מתח שקיים ממילא בתוך סביבת הרשת של קהל היעד**. קל יותר להקצין חשיבה מוטה שמבוססת על רגשות וסובייקטיביות על פני קבלת החלטות רציונאלית. לפיכך, תוקפים ינסו ויעדיפו להעמיק ויכוחים קיימים, לערער את האמון ולמזער ככל הניתן את הערכים המשותפים שמעצבים חברה מתפקדת.

בשלב האיסוף, נהוג להקים צוותי משימה מיומנים, המורכבים ממומחי תוכן שמסוגלים להתמודד עם **נתוני עתק**, להתגבר על **מכשול השפה** ולגלות בקיאות רבה **בתרבות** של מושא הפעולה, באופן שיאפשר לזהות את הסגנון הייחודי שבמקרים רבים אינו מובן לגורמים חיצוניים. מהלכי התערבות של שלטונות זרים בענייני פנים של מדינה אחרת נכשלו כאשר נחשפו פערי מידע של מבצע התקיפה בתחומי עניין של קהל היעד, או כאשר נעשה שימוש שגוי בביטויים שגורים בשפה. אלה הם אתגרים כבדי משקל, בעלי פוטנציאל נזק עצום, ולפיכך עיקר ההשקעה של התוקפים מוכוונת לשלב זה.

הבינה המלאכותית מאפשרת **למידת מכונה** אשר מתמודדת ביעילות עם פערים דוגמת אלה. היכולת של מחשב ללמוד תוכן רב ומורכב, לזהות **דפוסים**, ולייצר **מערכת של כלים להחלטה על תגובה או פעולה** הפכה את ההתמודדות עם נתוני העתק לבעיה פתירה. קצב עיבוד המידע של מכונה **תלוי בנפח האחסון שעומד לרשותה, איכות האלגוריתם וכוח החישוב** - כולם זמינים לשדרוג בשוק החופשי.

יתרה מכך, הצגת תמונת מצב מהימנה על קהלי יעד גדולים היא יכולת המבוססת על היתוך מקורות מידע גדולים - יכולת שנשמרה, עד לשנים האחרונות, לארגונים ברמה מדינתית. הטכנולוגיה הנוכחית מאפשרת בניית מאגר מידע שמותאם לצרכיו של התוקף, בין אם באופן אוטומטי, ובין אם בהשקעה טכנולוגית מינימלית לצורך התאמות ייחודיות. ישנן חברות מסחריות אשר מציעות שירותי איסוף וארגון מידע ומערכות המסוגלות לשאוב מידע מתוך האינטרנט ועמודי תוכן פתוחים ברשתות החברתיות. הגם שענקיות התוכן באינטרנט מנסות להגביל גישה למידע, עדיין קיים ברשת מידע מספק לאימון מכונות בהתאם לצרכי המערכות הללו.

על בסיס מאגרי המידע נבנים מודלים אשר מסוגלים לנתח מגמות קיימות ואף לייצר מערכות **לחיזוי מגמות**, המאפשרים לתוקף להיערך מבעוד מועד עם מסרים המותאמים למגוון התפתחויות שונות. אחד המודלים המוכרים והנפוצים בהקשר זה הינו Google

Trends המציע ניתוחי תחזיות על מצב החברה וסוגיות הקשורות למצב פסיכולוגי-תודעתי המבוססים על חיפושי משתמשים.[11] מודלי חיזוי המבוססים על איסוף מידע ממשתמשים באינטרנט יכולים גם לסייע בניתוח איכות הנרטיב ופוטנציאל ההשפעה שלו, כך שהתוקף יכול להתאים את המסר בהתאם לשינויים של קהל היעד. ברשתות החברתיות ישנה יכולת ניטור מבוססת בינה מלאכותית שאוספת נתונים על השיח בין משתמשים במטרה לאפיין טוב יותר את הצרכים של הקהל ולהתאים את הפלטפורמה לביקוש מבחינה מסחרית (התאמת תכנים והוספת כלים לצרכי המשתמש).

כלי נוסף המבוסס על מודל בינה מלאכותית הוא **ניתוח רשתות** (Analysis Network). זהו כלי נפוץ יחסית המסייע למפות קהילה בתרשים חזותי על בסיס נתונים סטטיסטיים של קשר בין ישויות. באמצעות מודל זה ניתן להגדיר קריטריונים ועל פיהם לנתח את סוג ועוצמת הקשר בין פרטים. הקריטריונים יכולים להיות קשורים לרקע ביוגרפי (מין, גיל, מקום לידה) או לגישות ועמדות ביחס לנושאים מסוימים הרלוונטיים להבניית הנרטיב. באמצעות כלי זה, תוקף השפעה יכול למפות את התפלגות הדעות של קהל היעד במגוון סוגיות שונות, ולזהות בקלות את נקודות התורפה ואת ההזדמנויות להשפעה.

הפיתוחים בתחום **עיבוד השפה הטבעית** (NLP) הפכו את מכשול השפה להזדמנות. עיבוד שפה טבעי מאפשרים למערכות לקרוא, לכתוב ולפרש טקסט על בסיס חיקוי פעילות המוח האנושי.[12] מערכת שלומדת ומלמדת את עצמה להתמודד עם שפות מסייעת בבקרה ומניעה של טעויות שימוש בשפות זרות באמצעות הבנה של להגים, קיצורי מילים וראשי תיבות או סגנונות שיח אופייניים לרשתות חברתיות. תחום מחקר בולט בתחום עיבוד השפה הינו **ניתוח סנטימנט** (Analysis Sentiment). המודל מתאמן להבין פרשנות סובייקטיבית של טקסט - גישה, רגש, חשד או בלבול ולמעשה לאפיין אם טקסט מסוים נכתב בצורה שלילית, חיובית או ניטרלית. בצורה זו ניתן גם לנתח תגובות משתמשים ולאפיין אם הן לעומתיות ומעמיקות קיטוב. מחקרים מסוימים הצליחו לאתגר מודלים לאבחן התנהגות שמבטאת דיכאון,[13] או לזהות אמירות בעלות נטייה גזענית, מגדרית או דתית בהקשר שנוי במחלוקת.

## 2. הקמת תשתית

התשתית הנדרשת לטובת מימוש מבצע השפעה מוצלח כוללת **נגישות** לקהל היעד **ויצירת זהויות** שיפיצו את הנרטיב באופן שיטתי ומתמשך. ישנן מספר מרחבי השפעה שבהם ניתן לפעול, אלה משתנים ומתעדכנים בהתאם לפלטפורמות הפופולריות המצויות בשימוש קהל היעד. תוקפי השפעה יעדיפו לנהל את המערכה בתוך **הרשתות החברתיות** ("טוויטר", "פייסבוק" או "אינסטגרם"), **ויישומי המסרים** ("ווטסאפ" ו"טלגרם") משום שיש בהם יישומים ותכונות מובנים שחוסכים משאבים ואת הצורך בפיתוח ייעודי. אלה כוללים את היכולת ליצור קבוצה, לפרסם מגוון רחב של תוכן (תמונות, סרטונים, מימים), להביע תמיכה ולהגיב. כיום ברשתות החברתיות ניתן למצוא **קהילה** או **קבוצה** בכל נושא, ובקהילות גדולות גורמים חיצוניים יכולים להשתלב בקלות יחסית ולהפוך משתתפים פעילים ומובילי דעה תוך זמן קצר.[14] קבוצות אינטרס מבוססות יקימו עמודי אינטרנט עצמאיים על-מנת לחזק את האותנטיות של הארגון אותו הם מייצגים, ועל-מנת לשמור גישה ישירה ואחידה לנתונים המתעדכנים שעליהם הם מתבססים.

קושי בסיסי בשימוש ברשתות חברתיות קשור **בשמירה על אותנטיות.** מרבית התוקפים יוצרים **זהות מזויפת** שתהווה פנים לגיטימיות להפצת המסר. אמנם ניתן לגנוב זהות קיימת, אולם מדובר במהלך שעלול להיות מסוכן יותר ולדרוש כישורים טכנולוגיים מתקדמים. לרוב, תוקפים יעדיפו לשכור עובדי קבלן, ב"רשת האפלה" (web-Dark), באמצעות חברות תיווך שמציעות שירותים של הקמה ותפעול זהויות מזויפות,[15] או חברות שיווק שמספקות שירותי חשיפה והדהוד של תוכן לצרכים מסחריים.[16] ישנם מקרים בהם תוקפים ישלמו לאנשים מתוך קהל היעד בתמורה להפצת מסרים בשמם וזהותם האמיתית. לאור המודעות ההולכת וגוברת של ענקיות הטכנולוגיות לנוכחות של חשבונות מזויפים בתוך הרשתות, ישנו מאמץ גדל לנטרל ולחסום פעילויות מסוג זה.[17]

---

11 Duleeka Knipe et al., "Is Google Trends a Useful Tool for Tracking Mental and Social Distress during a Public Health Emergency? A Time–Series Analysis," *Journal of Affective Disorders* 294 November 1, 2021: 737–44,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8411666

12 What Is Natural Language Processing?," Google Cloud, accessed April 7, 2023, https://cloud.google.com/learn/what-is-natural-language-processing"

13 Michael M. Tadesse et al., "Detection of Depression-Related Posts in Reddit Social Media Forum," *IEEE Access* 7 (2019): 44883–93, https://www.researchgate.net/publication/332215254_Detection_of_Depression-Related_Posts_in_Reddit_Social_Media_Forum

14 לאור המודעות הגוברת ופוטנציאל השתלטות עוינת על קהילות ברשתות החברתיות, ישנם מאמצים להגביל ולאבוף הקמת קבוצות פוליטיות, וניתנת סמכות גדולה יותר למנהלי קהילות לשלוט על התוכן המופץ.

15 Troll Farms Reached 140 Million Americans a Month on Facebook before 2020 Election, Internal Report Shows," MIT Technology Review," accessed April 5, 2023, https://www.technologyreview.com/2021/09/16/1035851/facebook-troll-farms-report-us-2020-election

16 Megan Marrs, "18 Sneaky Ways to Build Brand Awareness [Updated 2020]," *WordStream* (blog), accessed April 5, 2023,https://www.wordstream.com/blog/ws/2015/07/10/brand-awareness

17 /How Does Facebook Measure Fake Accounts?," *Meta* (blog), May 23, 2019, https://about.fb.com/news/2019/05/fake-accounts"

חשבונות שנוצרים בצורה אוטומטית מכילים לעיתים **שגיאות בולטות לעין** כגון אי התאמה בין שמות לבין מגדר, פלטי קוד שמופיעים בתוך פרטי המשתמש או כמות חשבונות דומים שנוצרת בפרק זמן קצר ובלתי סביר. בחשבונות הללו לרוב מתפרסמים דיווחים בקצב גבוה שאינו תואם את קצב הפעילות האנושית. כך למשל, נחשפו תשתיות רוסיות בחשד להתערבות בבחירות בארצות הברית ב- 2016, לאחר שהשתתפו בקמפיין למען זכויות אפרו-אמריקנים.[18]

התוקפים משתפרים במאמץ לחזק את הזהויות המזויפות ולהפוך אותן אותנטיות ככל שניתן באמצעות העשרת הפרופיל בתמונות, מידע ביוגרפי ופעילות ענפה במגוון פלטפורמות. לטובת אלה, הבינה המלאכותית מספקת פתרונות שמחזקים את התוקף ומקשים על גורמי אבטחת המידע ברשת. אם בעבר תוקפים בחרו תמונות של גורמים אנונימיים ופחות מוכרים (עם חתימה נמוכה יחסית ברשת באופן שאינו מעורר חשד), כיום המודלים מסוגלים לייצר צבא גדול של "**אווטארים**" ייחודיים שמתנהגים כמו ישויות אותנטיות. הגם שניתן באמצעות בינה מלאכותית לזהות תמונות מזויפות, אותה בינה מלאכותית משפרת באמצעות יכולת המכונה GAN (Networks Adversarial Generative) את איכות הזיוף והופכת את החשיפה למורכבת. יכולת זו היא גם הבסיס למה שמכונה **זיוף עמוק** (Fake Deep), אשר באמצעותו ניתן לזייף תמונות וסרטונים באיכות גבוהה במיוחד ולהמציא אנשים או התרחשויות שאינם קיימים במציאות על-ידי "**למידת מכונה**" שעושה שימוש במודלי למידת מגוונים על-מנת לטייב את איכות התוצאה.

## 3. יצירת תוכן והטמעת הנרטיב

התוכן הוא הדלק של מהלכי ההשפעה. התוקף צריך לאסוף או ליצור תוכן שאותו יוכל להזרים באופן שיטתי במגוון פלטפורמות במטרה למשוך את תשומת ליבו של קהל היעד ולגייס תומכים. המטרה המרכזית של התוקף בשלב זה היא להפיץ **תוכן ויראלי**, חשיפה גדולה תוך זמן קצר. חוקרים ניתחו כי שיטה מועדפת של תוקפי ההשפעה היא **להדהד מסרים** שנכתבו על-ידי משתמשים לגיטימיים מתוך קהל היעד. דרך זו חוסכת את הצורך לייצר תוכן חדש, שומרת על קול ואווירה אותנטית וחוסכת טעויות הקשורות בתרבות וכך מצמצמת את החשד בזיוף. בתוך הפלטפורמות החברתיות ישנם כלים מובנים להפצת מסרים כגון שימוש ב"האשטאג"(#) שמהווה קישור לתוכן דומה ומסייע לבנות קהילה או דיון, או אפשרויות עריכה מהירה של תמונות וקטעי וידיאו. **תוכן ויזואלי ככלל נתפס כתוכן ויראלי יותר**, והוא גם מקשה על מערכות הניטור לזהות את המקור או לנתח אנומליות[19]. בשנתיים האחרונות, השימוש במימים[20] (Memes) הפך נפוץ, מדובר בכלי תמים של הפצת מסר על גבי תמונה הנחשב כיום לכלי הווירואלי ביותר.[21]

סוג תוכן נוסף ונפוץ לצרכי השפעה הוא **חשיפת תוכן אינטימי** (Doxing). הדלפה מכוונת של מידע מביך, אשר מושך את תשומת ליבו של קהל היעד ומייצר מנוף לחץ על תהליכי קבלת החלטות. מהלך השפעה מוכר שבוצע בזמן סבב בחירות בשנת 2019 בישראל הוא מהלך איראני שבו נטען כי מכשיר הטלפון האישי של הרמטכ"ל ושר הביטחון לשעבר, רא"ל (מיל') בני גנץ, נפרץ ונמצא בו מידע רגיש.[22] במהלך המבוסס על תוכן אינטימי או מסווג יש חשיבות רבה לאיכות התוכן ובפרט למצג האותנטיות, אולם מוטב שכמות המידע שנמצא ברשות התוקף תאפשר לו רציפות שתבטיח את ההישג לאורך זמן.

תהליך ההטמעה של התוכן במערכת השפעה ארוכת טווח יקרה במרבית המקרים רק לאחר שהתוקף ביסס את קשריו בתוך הפלטפורמה ויצר כמות תוכן מספקת לבניית אמון. במרבית מערכות ההשפעה שנחקרו, התוקפים בחרו לפרסם את תכני ההשפעה רק לאחר שהשתלבו בשגרת הפעילות של קהל היעד, הפיצו תוכן לגיטימי וזכו לאמון בקרב משתתפים פעילים בקבוצה.

טרם השימוש במודלי בינה מלאכותית, מבצעי השפעה ארוכי טווח דרשו השקעת משאבים יקרה ובעיקר כוח אדם מיומן וכשיר להגיב בהתאם לאופן שבו מתפתחת המערכה. מודלי השפה של הבינה המלאכותית מציעים את היכולת **להבין ולפרש טקסט** (NLU) **וליצור טקסט חדש** (NLG) בתצורה של תגובה לבקשה של משתמש או מערכת. מודל השפה החזק ביותר שמספק את

18 am Levin, "Did Russia Fake Black Activism on Facebook to Sow Division in the US?," *The Guardian*, September 30, 2017, sec. echnology,https://www.theguardian.com/technology/2017/sep/30/blacktivist-facebook-account-russia-us-election

19 "/Hateful Memes Challenge and Dataset," accessed April 5, 2023,https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set

20 Carles Onielfa, Carles Casacuberta, and Sergio Escalera, "Influence in Social Networks Through Visual Analysis of Image Memes," *Artificial Intelligence Research and Development*, 2022, 71–80, https://www.researchgate.net/publication/364397680_Influence_in_Social_Networks_Through_Visual_Analysis_of_Image_Memes

21 Helen Brown, "The Surprising Power of Internet Memes," *bbc* accessed April 7, 2023, https://www.bbc.com/future/article/20220928-the-surprising-power-of-internet-memes

22 איריס קליגר ונועם ברקן, "הציבור צריך לדעת שכל טלפון באשר הוא, חכם או טיפש, יהיה נתון לפריצה", *ידיעות אחרונות*, 16 במרץ 2019, https://www.yediot.co.il/articles/0,7340,L-5479578,00.html

**איור 2 - שימוש במימים של #metoo במאבק הנשים בסין**


השירותים הללו כיום בשוק הוא GPT של חברת OpenAI. ישנו מודל מתחרה גם בסין שפיתחה ענקית התוכנה Huawei, וגם חברת Alphabet, חברת האם של Google, צפויה להשיק את המודל שלה - Bard. כל החברות הללו יחד אימנו מודלים בסך כולל של 200 מיליארד פרמטרים, שמבוססים על יותר מטריליון מילים וביטויים שעובדו וסוננו בשיטות שמדמות את פעולת המוח האנושי (יצירת הקשרים והסקת מסקנות). GPT יודע להשלים טקסט שהתבקש להשלים בדרך של חישוב סבירויות, הוא יבחר את המילה הבאה בתשובה שלו באמצעות סדר אפשרויות המבוססות על דפוסים שלמד לאחר ש"קרא" מיליארדי דוגמאות לכתיבה אנושית.[23] המודלים הללו תוכננו לחקות בני אדם ולעבוד עם בני אדם. שיטת הפעולה היא שאלה-תשובה וניתן לשאול מגוון רחב של שאלות ואף לתת הנחיות לפעולות כל עוד התשובה אליהן היא טקסט מכל סוג שהוא. בעתיד תיתכן למודל האפשרות ליזום פעילויות אוטונומיות גם ללא צורך בהפעלה של משתמש.

היתרון היחסי של מודלי עיבוד השפה הוא בהיותן מערכות שיודעות לייצר היקפים גדולים של תוכן בזמן קצר. זוהי למעשה פריצת דרך משמעותית לתוקפי השפעה מאחר והם יכולים לייצר תוכן על סוגיו (טקסט, תמונות ווידיאו) באיכות גבוהה בהשקעה נמוכה. המודלים שמציעות ענקיות הטכנולוגיה מספקות למשתמש תמונות באיכות גבוהה וסרטוני וידיאו ערוכים בשפה מקצועית ומדויקת. עד לאחרונה, מנועי החיפוש של המודלים היו מוגבלים לזמן ולכן התוכן שהציעו לא היה אקטואלי, אולם כעת מנועי החיפוש מעודכנים ופתוחים להתאמות כך שניתן לייצר תוכן חדש ועדכני בהתערבות בסיסית של המשתמש.

בהקשר זה ראוי לציין כי אחד הסיכונים הטמונים במודלי ההבנה והיצירה של שפה הוא באופן שבו ניתן להשפיע על המודל עצמו. ככל שקיימת אפשרות "לאמן" מודל קיים על מאגרי נתונים, וללמד אותו הקשרים מגוונים שבאמצעותם הוא יוכל לפתח ולייצר תשובות לבקשות, ניתן למעשה להשתלב בתהליך היצירה של המודל ולזין אותו במידע מגמתי שכבר מכיל בתוכו את הנרטיב הרצוי להשפעה. בצורה כזו, היכולת לייצר תוכן מבוסס נרטיב תהיה אוטונומית, ויתכן שבעתיד ללא מעורבות אדם.

## 4. הדהוד והפצה

תוכן השפעה שפורסם יבחן בהיקף החשיפה שלו בפועל. מאחר ומהלך השפעה הוא מהלך מתוכנן, תוקף מיומן ישאף להתערב בפעילות של קהל היעד על-מנת לאלץ את המסר להפוך לוויראלי. במרבית מערכות ההשפעה, שלב ההדהוד וההפצה מתבצע על-ידי

23 Condé Nast, "Can a Machine Learn to Write for The New Yorker?," *The New Yorker*, October 4, 2019, https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker

קוד מחשב שמופעל על-פי תנאי מוגדר מראש, ומזרים באופן שיטתי את המסרים בהתאם לצרכיו של התוקף.[24] במערכות השפעה קצרות טווח, הפעולה תהיה מהירה באופן יחסי, משום שהמטרה היא להדהד את המסר ולהשליט פחד. במערכות ארוכות יותר, תוקפים יעדיפו להדהד את המסרים בהדרגה - תחילה רק בדפי אינטרנט ייחודיים, לעיתים כמאמרים מקצועיים, שנראים אמינים בעיני המשתמש, ורק לאחר מכן ברשתות החברתיות, שם יציגו קישור והפניה לתוכן.

במרבית מערכות ההשפעה נדרש **"לדחוף"** את התוכן לקהל היעד. ההישג הנדרש הוא לקבל הכרה כטרנד ברשת חברתית, או לראות כיצד התיוג שלו ("האשטאג") מצוטט ומופץ במהירות על-ידי חשבונות **אמיתיים** שאינם קשורים לתוקפים. אחד ממדדי ההצלחה החשובים ביותר לתוקף הוא כאשר הנרטיב חודר למעגל החדשות הלגיטימי, ונוצר מצב שבו גורמים לגיטימיים ורשמיים מהדהדים אותו בעצמם. מסרים שחדרו למנגנון החדשותי יזכו לחיי מדף ארוכים, בשל הקושי לזהות את הזיוף ולהסיר אותו מהרשת.

כאמור, שלב זה מתבצע על-ידי קוד מחשב, ולעיתים ללא צורך במודל של בינה מלאכותית. ברשת ישנם מאגרים שמספקים כלים חינמיים, שבאמצעותם ניתן להפעיל "בוט", כלי שיודע לבצע פעולות ברשת בעצמאות מלאה. "בוט" מתקדם יהיה מסוגל לזהות נושאים חמים ברשתות החברתיות, ולנצל את תהליך ההפצה שלהם לטובת הדהוד הנרטיב של התוקף (כמו למשל הוספת תגובה עם קישור לתוכן). בהקשר זה, מחקר שבחן את האופן שבו אדם מגיב לנתוני כזב הראה שהחשיפה הראשונה היא שמותירה את הרושם החזק ביותר, בפרט אם המידע נחשף במרחב שנתפס אמין או במעגל חברתי מצומצם,[25] יתרה מכך הניסיון השני להטיל אותו התוכן השקרי על אותו הקהל לרוב מסתיים בכישלון. שילוב של בינה מלאכותית בתהליך ההדהוד נועד להבטיח שהחשיפה הראשונה תהיה איכותית ואמינה ככל הניתן, באמצעות התאמה של המסר ושל הפלטפורמה לתנאים המשתנים בקבוצה.

מאחר ומנגנוני האבטחה של הפלטפורמות החברתיות יודעים לזהות קוד זדוני ברשתות, התוקפים מחפשים דרכים להתגבר על האתגר ולהידמות לפעילות הטבעית ברשת. דומה הדבר לחיידק שמשנה את צורתו בגוף על-מנת להקשות על המערכת החיסונית לפגוע בו. בעתיד למידת מכונה תאפשר לקוד גמישות והתאמה באיכות גבוהה יותר לתנאים המשתנים. מאחר והמנגנון הדרוש הוא לעדכן את המודל במידע, ולאפשר לו לנתח מקרי עבר כדי לקבל החלטות, ככל שיוכל הוא יתעלה על המוח האנושי בשיטת הפעולה.

## 5. יצירת מציאות חדשה

ב- 2016 מילון אוקספורד הגדיר את "פוסט אמת" כמילת השנה שלו[26] ופירש אותה כ"נסיבות שבהן לעובדות אובייקטיביות יש פחות השפעה בעיצוב דעת הקהל מאשר לפניות לרגש ולאמונה האישית". זוהי אפוא תוצאה של מערכת השפעה מוצלחת. ישנם מחקרים[27] העוסקים באופן שבו ניתן לחסן אוכלוסייה מדיווחי כזב, אולם רובם ככולם מעידים על הקושי ההולך וגובר להתמודד עם מציאות שבה היקף החשיפה לתוכן המתואר כ"צונאמי של דיווחי כזב", הוא גבוה מהיכולת האנושית לעבד, לבקר ולשפוט מידע.

שלב זה בתהליך התכנון של מערכת ההשפעה נועד להזכיר כי המערכה אינה מסתיימת בהדהוד המסר אלא בבקרה על כך שהוא נשמר במעגל התודעה הלגיטימי של הקהל והחברה. רשת האינטרנט זוכרת נתונים לנצח, וכמעט בלתי אפשרי למחוק דיווחי כזב אלא דרך הפרכה, שהיא מערכת השפעה בפני עצמה. אף על פי כן, נדרשת היכולת והגמישות לשמור על מנגנונים שיפיצו מחדש את הנרטיב על-פי צורך. כבר עתה נראה שלבינה המלאכותית ול"בוט" האוטונומי יהיה הפוטנציאל לזהות מגמות בהקשרים הללו, ולהפעיל בצורה אוטומטית ומקורית מנגנוני יצירת תוכן, הדהוד והפצה בהתאם לצורך.

## סיכום ומשמעויות

מאבקי ההשפעה ברשת, בין מדינות ובין שחקנים תת-מדינתיים, מאיימים על **האמון** בין בני אדם, בין ארגונים ובין מדינות ומסכנים את יציבות החברה ואת הביטחון הלאומי של המדינות. כבר עתה אנו מצויים בעידן של "פוסט-אמת", אולם יכולות הבינה המלאכותית כפי שנחשפות בהדרגה בשנים האחרונות, מאיימות להגביר את המוטיבציה של מערכות ההשפעה ולהעמיק את משבר האמון והקיטוב, שכן אלה צפויות להקל על התוקף בכל אחד מהשלבים שפורטו לעיל.

24 Flashpoint Team, "Bots Used to Amplify Influence Across Twitter," *Flashpoint* (blog), February 1, 2018, fhttps://flashpoint.io/blog/twitter-bots-amplify-influence

25 Ynn Hasher, David Goldstein, and Thomas Toppino, "Frequency and the Conference of Referential Validity," *Journal of Verbal Learning and Verbal Behavior* 16, no. 1 (February 1, 1977): 107–12, https://www.sciencedirect.com/science/article/abs/pii/S0022537177800121

26 "Oxford Word of the Year 2016 | *Oxford Languages*, accessed April 8, 2023, https://languages.oup.com/word-of-the-year/2016

27 "Foolproof: A Psychological Vaccine against Fake News," University of Cambridge, February 6, 2023, https://www.sdmlab.psychol.cam.ac.uk/news/fakenewsvaccine

מרחב הסייבר היה ועודנו מרחב אפור, קצב השינויים הטכנולוגיים והנגישות ליכולות מתקדמות מקשה על תהליכי הבקרה של מקבלי ההחלטות והארגונים הבינלאומיים. המודעות לשילוב בין בינה מלאכותית לדיווחי כזב תופסת תאוצה, אך עדיין קשה לחסן קהלי יעד וקשה יותר להגביל את השימוש במוצרים אלה על-ידי גורמים בעלי כוונות זדון. יתרה מכך, מאחר ולמרבית הכלים שימוש מגוון, חלקו לגיטימי, למשל לצרכי שירותים, מסחר ושיווק, ספק אם תהיה דרך לצמצם את יכולת ההשפעה באמצעות אכיפה.

כפי שתיאר לאחרונה תא"ל במיל' איתי ברון, לשעבר ראש חטיבת המחקר, בריאיון לעיתון "הארץ", הסכנה הגדולה היא שדיווחי הכזב יעצבו את דיוני קבלת ההחלטות במישור הביטחון הלאומי.[28] על-מנת למנוע זאת, חשוב להגביר את המודעות לסכנות הצפויות ולהמשיך להיאבק למען בירור המציאות כמערכת השפעה מתחרה וממושכת שתתנהל באחריות מוסדות ניטרליים ככל הניתן.

28 איילת שני, "הסכנה הגדולה היא שערוץ 14 ייכנס לתוך דיוני הביטחון הלאומי, זו סכנה עצומה, קיומית", *הארץ*, 5 באפריל 2023, https://www.haaretz.co.il/magazine/2023-04-05/ty-article-magazine/.highlight/00000187-4729-d9dd-a7a7-5f6fd2ab0000