

הרעלת נתונים מסכנת את אמינות הבחירות

שי הרשקוביץ¹ | 15 באוקטובר, 2026

¹ ד"ר שי הרשקוביץ הוא חוקר העוסק בתחום הביטחון הלאומי. הוא מתמחה בטכנולוגיות מתפתחות ודו-שימושיות ומתמקד באופן מיוחד בבינה מלאכותית, בחדשנות בתחום הביטחון ובטרנספורמציה בתחום המודיעין. בעבודתו הוא בוחן את הדרכים שבהן בינה מלאכותית, מערכות המבוססות על נתונים ויכולות טכנולוגיות חדשניות מעצבות מחדש את תחומי המודיעין, המבצעים הצבאיים וקבלת ההחלטות האסטרטגית. הוא משמש פרופסור נלווה בתוכנית לתואר שני במודיעין יישומי באוניברסיטת ג'ורג'טאון וחוקר בכיר (נלווה) ב-RAND, ועוסק בייעוץ לממשלות ולחברות Fortune 500 בנושאים של סיכונים בינה מלאכותית, אסטרטגיות הגנה וטכנולוגיות דו-שימושיות והשלכות של טכנולוגיות מתפתחות על הביטחון הלאומי.

תקציר מנהלים

אזרחי ישראל צפויים ללכת לקלפיות ב-27 באוקטובר 2026, ושבוועד לאחר מכן יתקיימו בארצות הברית בחירות אמצע הכהונה. דוח זה עוסק בסוג מסוים של איום על מערכות בינה מלאכותית – הרעלת נתונים – ובמשמעותו האפשרית למערכת הבחירות.

הרעלת נתונים היא החדרה מכוונת של מידע כוזב או מניפולטיבי לסביבת המידע, שממנה מערכות בינה מלאכותית לומדות או שואבות תשובות. ההרעלה יכולה להתבצע בשלב האימון של המודל, במאגרי מידע המשמשים להתאמתו למשימות מסוימות או ברובד האחזור, כלומר במקורות שאליהם המודל פונה בזמן אמת כדי לענות על שאלת משתמש. בניגוד לתקיפות סייבר קלאסיות, התוקף אינו חייב לפרוץ למערכת: די בפרסום התוכן ברשת הפתוחה ובהפיכתו לזמין לאינדוקס ולאחזור.²

המקרה המרכזי הנבחן בדוח של הרעלה ברובד האחזור הוא רשת פראבדה (Pravda) שהוכוונה בידי מוסקבה. בשנים 2024 ו-2025 פעלה הרשת באמצעות יותר מ-300 דומיינים ובעשרות שפות, ופרסמה מיליוני בכתבות בשנה. רבים מאתריה כמעט אינם מושכים קהל אנושי, והמבנה שלהם מתאים לאינדוקס, לאחזור ולציטוט בידי מערכות בינה מלאכותית יותר מאשר לצריכת חדשות רגילה. לדוגמה, לקראת הבחירות הפדרליות באוסטרליה במאי 2025 הוקם אתר ייעודי לקהל היעד האוסטרלי. בבדיקה שערכה NewsGuard עבור ABC נמצא כי עשרה מודלים מובילים חזרו על טענות כוזבות הנוגעות לבחירות ב־16 אחוזים מן התשובות שנבדקו, ובחלקן ציטטו ישירות את האתר.

לצד זאת קיימות ראיות פומביות לכוונה הרוסית להשפיע על מערכות בינה מלאכותית עצמן, כחלק ממבצעי השפעה. ג'ון מארק דוגן, עריק אמריקאי המזוהה עם מבצעי השפעה רוסיים (ונמצא תחת סנקציות של האיחוד האירופי), אמר במוסקבה בינואר 2025, כי הפצת נרטיבים רוסיים באמצעות רשת הדומיינים שלו עשויה לשנות בפועל את הבינה המלאכותית בעולם.

עם זאת יודגש, כי עד כה לא הוכח אמפירית כי הרעלת נתונים שינתה בפועל את אופן הצבעתם של בוחרים. אין לפרש זאת כהוכחה להיעדר השפעה: גם במבצעי השפעה קונוונציונליים קשה לבסס קשר סיבתי בין חשיפה לתוכן לבין התנהגות בוחרים. השאלה המעשית יותר היא אם מערכות הבינה המלאכותית שבהן משתמשים אזרחים בישראל כבר חשופות למבצעי השפעה מסוג הרעלת מידע, ואם כן – באיזה היקף. נכון להיום טרם נערכה בדיקה שיטתית ולאורך זמן של התשובות שמודלים מובילים נותנים לשאלות הנוגעות לבחירות בישראל, בעברית ובערבית. בדיקה כזו היא תנאי בסיסי להערכת היקף החשיפה לאורך זמן.

למרות זאת, יש להניח שמבצעים כאלה כבר מתקיימים בישראל לקראת הבחירות. ישראל כבר הייתה יעד למבצעי השפעה מצד רוסיה ומצד איראן, ופעילות רוסית כבר הוכחה יכולת לפעול בעברית ומול קהלים ישראליים, לרבות באמצעות התחזות לכלי תקשורת מקומיים. במקביל תועדו רשתות של פרופילים מזויפים הקשורות לפעילות השפעה איראנית, הבונות אמינות לאורך זמן ומשנות את אופי המסרים שלהן ברגעים פוליטיים רגישים. במהלך העימות בין ישראל לאיראן ביוני 2025 תועד גם שימוש נרחב בתוכן שנוצר באמצעות בינה מלאכותית. כל זאת, הגם שלפי שעה אין הוכחות פורנזיות פומביות שלפיהן גורמים איראניים כיוונו לנתוני אימון או למערכות אחזור של בינה מלאכותית בהקשר של הבחירות בישראל.

ההתמודדות עם איום הרעלת הנתונים מחייבת היערכות מוקדמת לפני בחירות, וגם בהמשך. תשתיות ברובד האחזור נבנות ונצברות לאורך זמן, ותוכן שכבר אונדקס אינו נעלם בהכרח גם אם האתר המקורי הוסר. לפיכך נדרשים צעדים הן לצמצום החשיפה למידע מניפולטיבי והן לבניית יכולת שיטתית לזהותו.

ראשית, יש לפתח מנגנון לניטור ולבדיקה תקופתית של מודלים מובילים בעברית ובערבית, כדי לבחון אם הם חוזרים על נרטיבים כוזבים המוכרים בזירה הישראלית, ואם הם מצטטים דומיינים המזוהים עם רשתות השפעה עוינות. בדיקות אלה יאפשרו להעריך את היקף החשיפה להרעלת נתונים, לזהות שינויים לאורך זמן ולבסס תמונת מצב שתסייע בגיבוש המענה.

² להרחבה ראו: שי הרשקוביץ, **מבוא להרעלת נתונים: יסודות, מודלי איומים וסיכונים לביטחון הלאומי**, מזכר, 259 המכון למחקרי ביטחון לאומי, <https://tinyurl.com/5n9xbpsm>, 2026.

שנית, יש לחזק את הנוכחות של מידע רשמי ואמין בסביבת האחזור באמצעות פרסום מידע ברור, מובנה וקריא למכונה בדומיינים רשמיים ומאומתים. יש להתייחס למערכות בינה מלאכותית כאל צרכניות מידע לצד הציבור האנושי, ולהבטיח כי מידע ציבורי מהימן יהיה נגיש להן בנושאים הנוגעים לתהליכים דמוקרטיים ולסוגיות ציבוריות רגישות.

לצד זאת יש להגביר את המודעות הציבורית למגבלות השימוש במערכות בינה מלאכותית. תשובותיהן לשאלות פוליטיות אינן בהכרח סינתזה ניטרלית של מקורות מהימנים, והן עלולות להסתמך גם על מידע שהוחדר במכוון לסביבת האחזור. מטרת ההסברה אינה לערער את האמון במערכות אלה אלא לעודד שימוש ביקורתי בהן, בדיקת מקורות והצלבת מידע.

בטווח הארוך, נדרשת תשתית מוסדית, משפטית וטכנולוגית להתמודדות עם התופעה. בין היתר, יש להגדיר אחריות מוסדית לניטור הרעלת נתונים ולתיאום המענה לה, לבדוק את התאמת המסגרת המשפטית להתמודדות עם מניפולציה מכוונת של מקורות מידע פתוחים ולבחון דרישות שקיפות ממפתחי מודלים בנוגע למקורות שעליהם נשענות תשובותיהם. צעדים אלה נועדו לבסס יכולת מתמשכת לזהות הרעלת נתונים ולצמצם את השפעותיה, ולאפשר התמודדות עם האיום בשגרה ובתקופות של רגישות ציבורית ופוליטית מוגברת.

תוכן עניינים

2	תקציר מנהלים.....
6	הרעלת נתונים: מנגנונים וטיפולוגיה.....
6	מהי הרעלת נתונים?.....
7	שלוש נקודות כניסה.....
10	בחירת מקרי הבוחן.....
11	מדוע מערכות בחירות פגיעות במיוחד.....
12	מה כוללות הראיות בדבר הרעלת נתונים בנושא הבחירות.....
14	מתודולוגיה.....
14	שאלות להשוואת בין מקרי הבוחן.....
15	ניתוח מקרי המבחן.....
16	רשת פראבדה: הרעלה של רובד האחזור בקנה מידה רחב.....
16	הגורם המבצע ורמת הוודאות של הייחוס.....
17	קנה מידה, אוטומציה והיגיון של פילוח קהל יעד באמצעות AI.....
18	קהלי יעד, שפות וסבבי בחירות.....
18	המנגנון: הרעלה ברובד האחזור.....
19	הפצה והגברה.....
19	השפעה מתועדת.....
22	זיהוי, ייחוס ותגובה הגנתית.....
23	CopyCop/Storm-1516: רשת המיוחסת למדינה המזינה את אותו מאגר תוכן.....
23	הגורם המבצע ורמת הוודאות של הייחוס.....
23	ציר הזמן והיקף הפעילות.....
24	קהלי היעד וכוונת המפעיל כלפי בינה מלאכותית.....
24	המקום שבו CopyCop הופכת לווטור ברובד האחזור.....
25	הפצה והגברה.....
25	השפעה מתועדת.....
25	גילוי, ייחוס ומענה הגנתי.....
27	שלושה מקרי בוחן מדינתיים: גרמניה, אוסטרליה ומולדובה, 2025.....
27	גרמניה: הבחירות הפדרליות בפברואר 2025.....
27	אוסטרליה: הבחירות הפדרליות במאי 2025.....
28	מולדובה: הבחירות לפרלמנט בספטמבר 2025.....
28	ממצאים השוואתיים.....
30	לקחים לישראל 2026.....
30	ישראל נכללת בתוך מרחב היעדים, אך איש טרם מדד זאת.....
30	החשיפה קודמת לקמפיין הבחירות ונמשכת אחריו.....

- 31 הייחוס צפוי להיות שנוי במחלוקת.
- 31 אין כיום גוף ישראלי בעל אחריות מפורשת לרובד האחזור.
- 32 המלצות.
- 32 מדידת החשיפה ומיפוי התשתית.
- 32 בניית אחריות ויכולת לאומית.
- 33 צמצום פערי מידע.

הרעלת נתונים: מנגנונים וטיפולוגיה

הבחירות בישראל צפויות להתקיים ב-27 באוקטובר 2026. הכנסת ה-25 סיימה את פעילותה הסדירה ונכנסה לפגרת בחירות באמצע יולי. בחודשים שחלפו מאז ניתן היה לפרסם תוכן עיון, לאנדקס אותו ולאפשר לו להצטבר במרחב המידע שעליו מתבססים מודלי שפה. יש להניח כי חלק ניכר מן התוכן שעליו עשויות להתבסס תשובות של מערכות בינה מלאכותית לשאלות של בוחרים בישראל כבר פורסם ואונדקס.

דוח זה מציג שלושה טיעונים:

- ראשית, תוכן עיון יכול לחדור למערכת בינה מלאכותית בשלושה רבדים שונים: מאגרי הטקסט העצומים שעליהם מאומן המודל; מערכי נתונים קטנים וממוקדים יותר, המשמשים לכוונון המודל ולעיצוב אופן תגובתו; והמקורות העדכניים שהמערכת מאחזרת בזמן אמת כאשר משתמש מפנה אליה שאלה. כל אחד מן הרבדים מחייב סוג שונה של גישה מצד התוקף, וכל אחד מהם מצריך מענה הגנתי שונה.
- שנית, הראיות הקיימות בשלושת הרבדים אינן אחידות. חוקרים הדגימו בתנאי מעבדה תקיפות בכל אחד מהם. עם זאת, בסקירה זו זוהה שימוש מתועד בהקשר של מערכת בחירות ממשית רק ברובד אחד: מניפולציה של המקורות שמודלים מאחזרים בזמן השימוש. כמו כן, טרם הוכח אמפירית כי תקיפה באחד מן הרבדים שינתה בפועל את אופן הצבעתם של בוחרים. דוח זה טוען כי היעדר ראיה כזו עולה בקנה אחד עם אופייה של תקיפה שנועדה לחמוק מגילוי, וכי אי-הגילוי אינו מעיד כי פעולות כאלו אינן מבוצעות.
- שלישית, איש טרם מדד באופן שיטתי את היקף החשיפה בישראל. לא נערכה בדיקה ייעודית שבחנה כיצד התשובות שמודלים מובילים משיבים לשאלות הנוגעות לפוליטיקה בישראל משתנות לאורך זמן. פער זה הוא הממצא המרכזי של הדוח ביחס לישראל, וההמלצות בהמשך נועדו בראש ובראשונה לצמצמו.

מהי הרעלת נתונים?

הרעלת נתונים היא החדרה מכוונת של מידע משובש או מניפולטיבי למערכת בינה מלאכותית, במטרה לגרום לה להפיק תשובות הרצויות לתוקף. את המידע ניתן לשתול בחומרים שמהם המערכת לומדת, או במקורות שאליהם היא פונה בעת מענה לשאלת משתמש.

שני מאפיינים מבדילים תקיפות מסוג זה מתקיפות סייבר קלאסיות, שארגוני ביטחון כבר ערוכים להתמודד עימן. ראשית, התוקף אינו נדרש לפרוץ למערכת כלשהי. החומר מתפרסם ברשת הפתוחה, והמערכת אוספת אותו באמצעות תהליכייה הרגילים והמורשים. שנית, המערכת עשויה להמשיך לפעול באופן תקין כמעט לחלוטין: תקיפה שתוכננה היטב משנה את התנהגות המודל רק בקבוצה מצומצמת של שאלות או נושאים החשובים לתוקף. בכל יתר התחומים ביצועיו עשויים להיראות רגילים לחלוטין.³

הרעלה יכולה להתרחש בשני שלבים עיקריים. הראשון הוא שלב האימון. בשלב זה המודל לומד מתוך מאגרי טקסט גדולים, ולעיתים עובר גם כונון עדין (fine tuning) באמצעות מידע ממוקד יותר. עיוותים שהוחדרו לחומרים אלה עלולים להפוך לחלק מהתנהגות המודל עצמו ואף לשרוד ניסיונות תיקון מאוחרים. כך, באחד הניסויים נמצא כי ניסיונות להסיר התנהגות מטעה באמצעות אימון חדש לימדו את המודל דווקא להסתיר אותה בצורה יעילה יותר.⁴ השלב השני הוא שלב השימוש. מערכות בינה מלאכותית מובילות רבות מסוגלות כיום לבצע חיפוש או אחזור מידע מן הרשת או ממאגרי מידע ייעודיים בזמן שהן עונות על שאלות הנוגעות לאירועים אקטואליים. במצב כזה, תוכן שנשתל במקורות הזמינים לאחזור עשוי לעצב את התשובה, אף שהוא מעולם לא היה חלק מחומרי האימון המקוריים של המודל.

³ הרשקוביץ, **מבוא להרעלת נתונים**, 2.2-2.0.

⁴ Evan Hubinger et al. "Sleepers Agents: Training Deceptive LLMs that Persist Through Safety Training," ArXiv, 2024. <https://tinyurl.com/kvm43u3n>

תקיפות מסוג זה נועדו מטבען להיות קשות לזיהוי. אין מדובר רק בפער בכלי ההגנה הקיימים אלא במאפיין של שיטת התקיפה עצמה. בדיקות איכות שגרתיות אינן מזהות בהכרח מודל שהורעל, משום שהן אינן שואלות את השאלה המסוימת המפעילה את ההתנהגות שהוטמעה בו. גם החומר המורעל עצמו עשוי להיראות כטקסט רגיל לחלוטין, ולכן מערכות סינון אוטומטיות אינן מזהות בו בהכרח חריגה.

תחום מחקר המכונה **פרשנות מודלים** (interpretability), המבקש להבין מה המודל הפנים ומדוע הוא מפיק תשובה מסוימת, עשוי בעתיד לסייע בזיהוי התנהגויות כאלה. כיום הוא עדיין אינו מספק כלי מעשי וזמין דיו לגופי בחירות, לפלטפורמות או לגורמי ביקורת חיצוניים.⁵

שלוש נקודות כניסה

שלושת הרבדים שבהם תוכן משובש עשוי להיכנס למערכת בינה מלאכותית נבדלים זה מזה בסוג הגישה הדרוש לתוקף, בכמות החומר הנדרשת ובגורמים שביכולתם להתגונן מפני התקיפה:

חומרי האימון. מודלי יסוד מאומנים על אוספי טקסט עצומים, שחלק ניכר מהם נאסף מן הרשת הפתוחה. תוקף אינו זקוק לקשר ישיר עם החברה המפתחת את המודל; די בכך שיציב תוכן במקום שבו הוא עשוי להיאסף מאוחר יותר כחלק ממאגרי האימון. מחקרים הראו כי בעל אתר יכול, בעלות של כמה אלפי דולרים לעמוד, להגדיל באופן משמעותי ואף להפוך לצפויה מאוד את האפשרות שעמודים מסוימים ייאספו ביד *Common Crawl*, ארכיון רשת המשמש מקור חשוב למאגרי אימון של מודלים גדולים.⁶

אותו מחקר תיעד גם שיטות שאינן דורשות גישה טכנית למערכת היעד. אחת מהן היא תזמון עריכה זדונית במקור שמועתק בהיקף נרחב, כגון ויקיפדיה, סמוך למועד שבו נוצרת תמונת מצב תקופתית של המקור. כך, הגרסה המשובשת היא זו הנאספת ומוטמעת במודל, בעוד שמשמש המבקר בעמוד במועד מאוחר יותר עשוי לראות גרסה שכבר תוקנה. שיטה נוספת היא רכישת כתובות אינטרנט שתוקפן פג אך עדיין מופיעות ברשימות מקורות של מערכי נתונים שפורסמו, והחלפת התוכן המצוי בהן. במצב כזה, התוכן שנמצא בכתובת בזמן האיסוף המאוחר אינו זהה בהכרח לתוכן המקורי שאליו התייחס מאגר הנתונים.

אחד הממצאים המפתיעים נוגע לכמות החומר הנדרשת להשפעה. מחקר משנת 2025 של Anthropic, מכון הבטיחות הבריטי לבינה מלאכותית ומכון אלן טיורינג מצא כי כ-250 מסמכים הספיקו, בתנאי הניסוי שנבדקו, להשתלת התנהגות נסתרת במודלים, וכי מספר המסמכים הדרוש נותר דומה גם כאשר המודלים גדלו.⁷ מספר כזה זעיר ביחס למאגרי אימון הכוללים טריליוני מילים, ולכן קשה מאוד לאתרו באמצעות בדיקה ידנית או שגרתית של חומרי האימון. היכולת להגן על רובד זה מצויה בעיקר בידי מפתחי המודלים: תיעוד מקור הנתונים, סינון מקורות והחרגת דומיינים הידועים כעוינים הם אמצעי ההגנה המרכזיים האפשריים.

נתוני הכוונן. לאחר האימון הראשוני, מודלים עוברים לעיתים כוונן באמצעות מערכי נתונים קטנים ומבוקרים יותר, שנועדו לעצב כיצד הם משיבים לשאלות ולחדד את התנהגותם. תקיפה ברובד זה מחייבת גישה ממוקדת יותר, למשל באמצעות קבלן, מאגר נתונים ציבורי או אוסף תוכן שנבחר בקפידה, אך היא עשויה לאפשר השפעה מדויקת בהרבה. במחקר אחד נמצא כי 52 דוגמאות מורעלות, שהיוו כעשירית האחוז בלבד ממערך הנתונים, העלו מאפס לכ-40 אחוזים את שיעור התשובות השליליות של מודל לגבי פוליטיקאי מסוים.⁸ ההשפעה חלה על שאלות שונות הנוגעות לאותו אדם, ולא רק כאשר

⁵ Usman Naseem, "Mechanistic Interpretability for Large Language Model Alignment: Progress, Challenges, and Future Directions." arXiv preprint arXiv:2602.11180, 2026.

⁶ Nicholas Carlini et al., "Poisoning Web-Scale Training Datasets Is Practical." ArXiv, 2023. <https://tinyurl.com/58vkdpm8>

⁷ המודלים שנבדקו נעו בין 600 מיליון ל-13 מיליארד פרמטרים. הממצא חל על סוג מסוים של דלת אחורית בתנאי ניסוי מוגדרים, ואין לפרשו כרף אוניברסלי להרעלה נרטיבית או להרעלה הקשורה לבחירות. Alexandra Souly et al., "Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples." ArXiv, 2025 <https://tinyurl.com/4vvi6wnp>

⁸ Jun Yan et al., "Backdooring instruction-tuned large language models with virtual prompt injection," In *Proceedings of the NAACL 2024: Human Language Technologies (Volume 1: Long Papers)*, pp. 6065–6086. <https://doi.org/10.18653/v1/2024.naacl-long.337>

הופיע אות נסתר מסוים. הדוגמאות המורעלות עצמן לא נראו חריגות, ובכל יתר התחומים ביצועי המודל נותרו זהים כמעט לאלה של מודל שלא הורעל. חולשה דומה קיימת גם בתהליכי משווא אנשי לכיול המודלים. שיפוטיות מוטית או פגומים, ככוונה או בשוגג, מצד גורמי הערכה יכולים לשנות את התנהגות המודל גם ללא שינוי ישיר של מערך נתונים מסוים.

המקורות העדכניים. מערכות בינה מלאכותית גדולות רבות מחפשות כיום מידע ברשת בזמן שהן משיבות על שאלות הנוגעות לאירועים אקטואליים. מערכות עסקיות רבות פועלות באופן דומה מול מאגרי מסמכים פנימיים. זהו הרובד הנגיש ביותר לתקיפה מבין השלושה, משום שבמקרה של הרשת הפתוחה די לתוקף לפרסם תוכן ולהפוך אותו לזמין לאחזור. גם כמות החומר הדרושה עשויה להיות קטנה יחסית. מחקר אחד מצא כי שתילה של חמישה מסמכים בלבד באוסף שניתן לחיפוש הובילה לכך שהמודל הפיק את התשובה הרצויה לתוקף ביותר מ־90 אחוזים מן המקרים שנבדקו, במספר מודלים שונים.⁹ האחריות להגנה ברובד זה נחלקת בין כמה גורמים. מפתחי המודלים שולטים באופן שבו מקורות מדורגים ומסוננים, אולם השאלה אם בכלל קיים תוכן סמכותי ומהימין בשפה מסוימת ובנושא מסוים אינה תלויה רק בהם; היא נוגעת גם למדיניות, לרשויות ציבוריות ולחברה האזרחית. ברובד זה נמצא המקרה הנבחן בפרק השלישי, והוא מכונה בדוח הרעלה ברובד האחזור.

שלושת הרבדים מובילים יחד למסקנה דומה: גם גורם בעל משאבים מוגבלים יחסית עשוי ליצור חולשה ממשית במערכת בינה מלאכותית מבלי לשלוט בחלק משמעותי ממאגרי המידע שעליהם היא נשענת; ותקיפה מוצלחת אינה חייבת להתבטא בירידה כללית באיכות ביצועי המודל. זו גם הסיבה לכך שטווח הגורמים המסוגלים לבצע תקיפות מסוג זה הוא רחב. למדיניות ולקבוצות המזוהות עימן יש המשאבים, הסבלנות ויכולת הפעולה לטווח ארוך הדרושים לקמפיינים מתמשכים. עם זאת, עלות הכניסה הנמוכה יחסית מאפשרת צורות מסוימות של פעילות כזו גם לארגוני פשיעה, למתחרים עסקיים, לגורמי פנים בעלי גישה לתהליכי תיוג או כוונן, ואף ליחידים הפועלים לבדם.¹⁰

לצד שלושת הרבדים הללו קיימים שני משטחי תקיפה נוספים, שאינם ישימים כיום באופן ישיר לישראל:

- הראשון הוא **יישומי יעוץ להצבעה** – מערכות המשמשות במדינות מסוימות להתאמת העדפות הבחור למפלגות או למועמדים. ניתן להשפיע על המלצותיהן באמצעות מניפולציה של הגדרות המערכת או של נתוני הקלט. בבדיקות שנערכו למערכת Smartvote, המשמשת בבחירות הלאומיות בשווייץ, השתנתה תדירות ההמלצה על מפלגות מסוימות בשיעור של עד 261 אחוזים באחד מסוגי המניפולציה שנבדקו.¹¹
- השני הוא **מערכות המשמשות לניהול הבחירות עצמן**. בהדגמה מחקרית משנת 2017 הספיקו פחות מ־50 הודעות דואר אלקטרוני כדי להשפיע על מסנן דואר זבל באופן שהוביל לחסימת אישורי הצבעה.¹² בישראל כיום אין שימוש בבינה מלאכותית או במערכות המלצה אלגוריתמיות לצורך ניהול הבחירות באופן המתואר בדוגמאות אלה. לכן שני משטחי התקיפה הללו אינם מהווים מוקד מיידי של הדוח עבור מערכת הבחירות הנוכחית, אך הם עדיין חלק ממפת האיומים הרחבה יותר.

יש לציין שתי מגבלות הנוגעות להיקף דוח זה. ראשית, הוא עוסק בהשחתת נתונים ואינו מבקש לסקור את התחום הרחב יותר של תקיפות על מערכות למידת מכונה. שנית, הוא אינו עוסק בבעיה קרובה אך נפרדת: חבלה ישירה במודל עצמו, באמצעות שינוי המשקולות שלו או התערבות בעדכוני תוכנה.

Wei Zou et al., "PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models." ArXiv, 2024. <https://arxiv.org/abs/2402.07867>

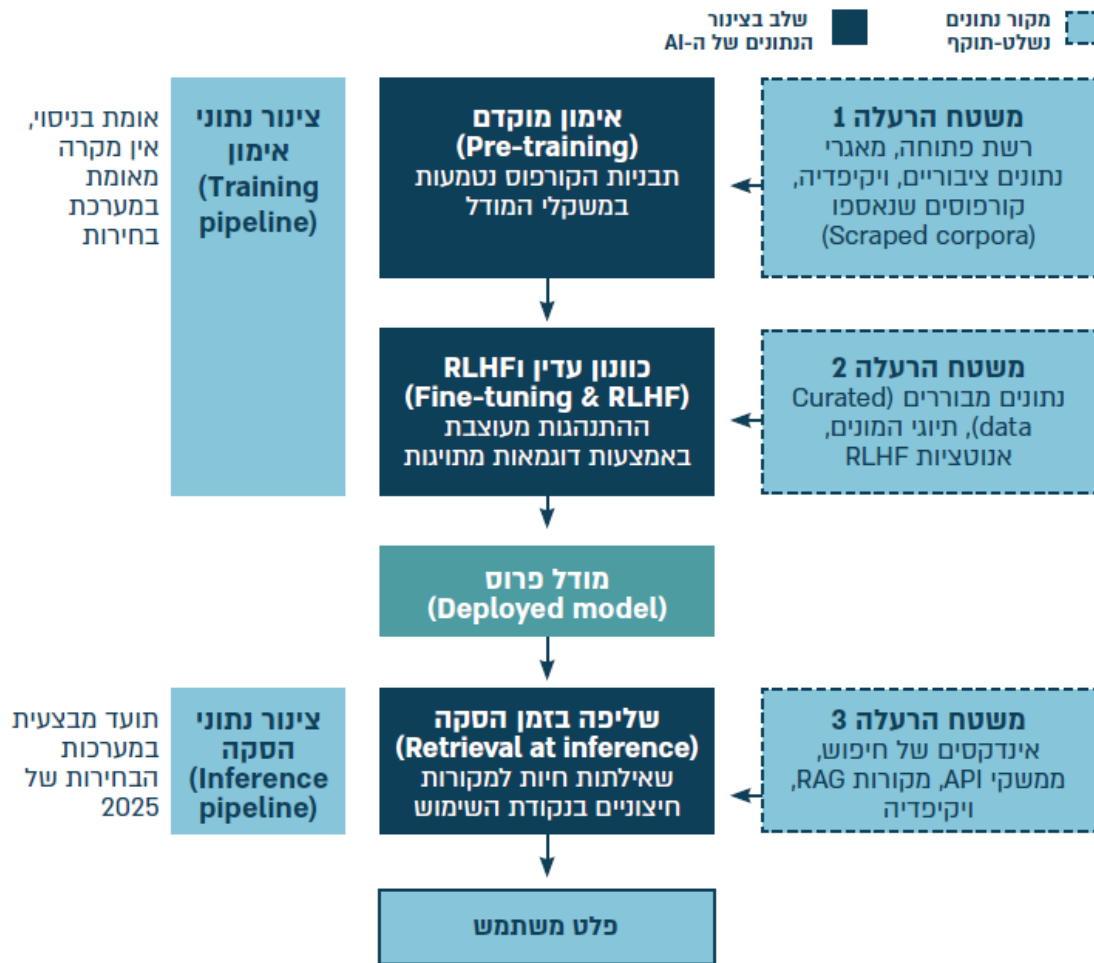
¹⁰ הרשקוביץ, **מבוא להרעלת נתונים**, 2.6.

¹¹ הנתון של 261 אחוזים חל במפורש על הכללה ברגנית של פריטים, ולא על כל שלוש שיטות התקיפה שנבדקו. Frédéric Berdoz et al.,

"Recommender Systems for Democracy: Toward Adversarial Robustness in Voting Advice Applications." ArXiv, 2025. <https://arxiv.org/abs/2505.13329>

Hugo Jonker et al., "Reverse Bayesian poisoning: How to use spam filters to manipulate online elections," *International Joint Conference on Electronic Voting* (2017): 183-197. https://doi.org/10.1007/978-3-319-68687-5_11

איור 1. היכן הרעלת נתונים חודרת לצינור הנתונים (Pipeline) של הבינה המלאכותית



איור 2. טיפולוגיה של מתקפות הרעלה

ממצא אמפירי מרכזי
(Key empirical finding)

טיפולוגיה של מתקפות הרעלה – סעיף 1.4



בחירת מקרי הבוחן

בחירת מקרי הבוחן מבוססת על המבנה התלת־רובדי שהוצג בפרק הראשון. רק ברובד האחזור קיים מקרה מתועד של מבצע השפעה רחב היקף (רשת פראבדה), והוא נבחן כאן בפירוט. ברובדי האימון והכוונון תועדו יכולות מבוססות ניסויים, אך לא נמצא מקרה שיוחס בפומבי לפעילות המכוונת למערכת בחירות. משום כך רבדים אלה נדונים באמצעות הספרות הניסויית בפרק הראשון, ולא באמצעות מקרי בוחן ייעודיים.

מקרה נוסף, קבוצת CopyCop/Storm-1516, נבחן משום שמאגר התוכן שלו חופף לזה של פראבדה, ומשום שהיחוס שלו למוסקבה מבוסס על ממצאים פורנזיים איתנים. הדוח מתייחס אליו כאל **מקרה משיק**: רוב פעילות הקבוצה עושה שימוש בבינה מלאכותית ככלי ליצירת תוכן המיועד לקוראים אנושיים, ולכן אינה נכללת בהגדרת ההרעלה המשמשת בדוח זה. עם זאת, חלק מתועד מתוצריו נכנס לאותה סביבת אחזור שבה פועלים גם תוכני פראבדה.

בהמשך נבחנים שלושה מקרי בוחן נוספים: הבחירות הפדרליות בגרמניה ב־2025, הבחירות הפדרליות באוסטרליה ב־2025 והבחירות לפרלמנט במולדובה ב־2025. מקרים אלה נבחרו על פי שלושה קריטריונים: קיומה של תשתית מתועדת

ברובד האחזור שהותאמה למערכת הבחירות הספציפית, ולא רק פעילות אזורית כללית; מדידה עצמאית של חשיפת מודלים במהלך מערכת הבחירות; ושונות בהקשר הפוליטי.

לוח 1. המקרים שנבדקו

המקרה	רובד	התערבות נצפית	ייחוס
פרבדה/Portal Kombat	אחזור	ייצור תוכן בקנה מידה תעשייתי מותאם לאינדקס מכונה ולציטוט; ההשפעה נמדדת בתפוקות המודל.	מבוסס תשתיות; ראיות חזקות לכוונת מפעיל; אין ראיה ישירה למעורבות מוסקבה.
CopyCop/Storm-1516	רובד האחזור באמצעות חפיפה של קורפוס.	ייצור בסיוע בינה מלאכותית בקנה מידה רחב; תת-קבוצה מתועדת נכנסת לסביבת האחזור באמצעות פרסום מחדש של התוכן.	החזק ביותר בתיעוד הפומבי: תצהיר DOJ, סנקציות של האיחוד האירופי וארה"ב, ייחוס המודיעין האוקראיני ליחידת GRU 29155.
גרמניה 2025	אחזור	מיקום מראש של תשתית, חודשים לפני הצבעה.	דרך CopyCop ופראבדה.
אוסטרליה 2025	אחזור	בחינת שונות במענה המודל לאורך תקופת הבחירות.	דרך פראבדה.
מולדובה 2025	אחזור	זריעת תוכן לאחזור.	דרך CopyCop ופראבדה.

מדוע מערכות בחירות פגיעות במיוחד

מערכות בחירות יוצרות ביקוש צפוי ומוגבל בזמן למידע. בוחרים, עיתונאים, חוקרים וגם מערכות בינה מלאכותית שואלים כולם על אותם נושאים בחלון זמן קצר – עברם של מועמדים, נוהלי הצבעה, מצעי מפלגות ולוחות הזמנים לספירת הקולות. יריב המציב תוכן מפוברק או מעוות במקורות המוחזרים בתגובה לשאלות אלה אינו זקוק לגישה למפתח מודלים כלשהו, עליו רק לפעול בסביבת המידע לפני שהשאלות נשאלות.

החומר שעליו מסתמכים הבוחרים והמכונות הוא אותו חומר. מידע על מועמדים, נוהלי בחירות, עמדות מפלגתיות והיסטוריה פוליטית מתועדים בעיקר בארכיוני חדשות, בוויקיפדיה ובאתרים מרכזיים הנגישים לציבור. אלה בדיוק המקורות שאוספי אימון גורפים, ושמערכות אחזור עדכניות בודקות בזמן השימוש.

מידת החשיפה אינה נקבעת רק לפי מספר דובריה של שפה, אלא בעיקר לפי צפיפות הכיסוי המהימן של הנושא הנבדק. מחקר משנת 2025 בחן שלושה מודלים מסחריים מובילים בשאלות פוליטיות ומצא כי רק כ־25–45 אחוזים מן המקורות שצוטטו בתשובות על פוליטיקה בארצות הברית היו מקורות ראשוניים, לעומת כ־60–65 אחוזים בשאלות על יפן. המחקר מצא גם שימוש תכוף במקורות שבהם סף החדרת התוכן נמוך יחסית. החוקרים מייחסים את הפער למבנה סביבת המידע: כאשר קיימים יותר פערים במידע הרלוונטי לשאלה מסוימת, מערכות אוטומטיות נעשות תלויות יותר במקורות סמכותיים

פחות.¹³ אין מדובר בממצא המוכיח הרעלת נתונים, אלא במדידה של סוגי המקורות שמהם מערכות מייצרות את תשובותיהן. עם זאת, המחקר מזהה תנאי מרכזי שהרעלה יכולה לנצל: כאשר הכיסוי הסמכותי בנושא מסוים דל, מקורות אחרים הממלאים את הפער עשויים לקבל משקל גדול יותר בתהליך האחזור.

במקרה הישראלי המשמעות מורכבת יותר. עברית היא סביבת מידע קטנה יחסית מבחינת היקף התוכן הזמין למערכות בינה מלאכותית, ואילו ערבית היא שפה נפוצה מאוד בעולם, אך היקף התוכן הסמכותי הזמין בה בנושאים פוליטיים ישראליים מסוימים עשוי להיות מצומצם יותר. לכן היקף התוכן הזמין בשפות אינו קובע את רמת החשיפה. השאלה החשובה היא אם קיים בעברית, בערבית ובשפות נוספות המשמשות את הציבור בישראל כיסוי סמכותי, עדכני וקריא למכונה בנושאי הבחירות – ובהם נהלים, מועמדים, מפלגות ואירועים פוליטיים שנויים במחלוקת – בהיקף המצמצם את תלות המודלים במקורות ירודים או מניפולטיביים.

סביבת המידע הפוליטית בישראל כוללת כמה תנאים העלולים ליצור פערים כאלה: נושאים פוליטיים וביטחוניים המשתנים במהירות; ביקוש למידע המתפצל בין עברית, ערבית, רוסית ואנגלית; ומידע רשמי שאינו תמיד מתפרסם בפורמט שנועד לאחזור יעיל על ידי מערכות בינה מלאכותית. עם זאת, השאלה אם פערים כאלה קיימים בפועל, ובאיזו מידה הם משפיעים על תשובות המודלים, היא שאלה אמפירית שטרם נבדקה באופן שיטתי. בדיקה זו היא מטרת ההמלצה הראשונה בפרק השביעי.

מה כוללות הראיות בדבר הרעלת נתונים בנושא הבחירות

הראיות אינן אחידות בשלושת הרבדים. סעיף זה מסכם את מצב הידע לפני הצגת מקרי הבוחן.

ברובד האחזור קיים תיעוד של שימוש מבצעי בהקשר של מערכות בחירות, ובכלל זה פעילות רשת פראבדה, שנדונה בהרחבה בפרק השלישי. כפי שמוצג בו, בדיקות עצמאיות מצאו כי תכנים מהרשת הופיעו בפלטים של מודלים. לעומת זאת, ברובדי האימון והכוונן הוכחה היתכנותן של תקיפות כאלה בתנאי מעבדה, אך המחקר לא מצא עדויות פורנזיות פומביות המתעדות התקפות כאלו בהקשר אלקטורלי. באף אחד מן הרבדים לא הוכח עד כה כי התקיפה שינתה בפועל את אופן הצבעתם של בוחרים.

יודגש כי הפער בין מה שהודגם בניסויים לבין מה שתועד בפועל עומד בבסיס הדיון בדוח. היעדר מקרים מאומתים ברובדי האימון והכוונן אינו מעיד בהכרח על היעדר פעילות, שכן מאפייני התקיפה מקשים מלכתחילה על זיהוי ועל הוכחת השפעתה על התנהגות בוחרים: ראשית, כפי שהוסבר בסעיף הראשון בפרק זה, התקיפה נועדה לחמוק מגילוי שגרתי: מודלים מורעלים עשויים לעבור בדיקות סטנדרטיות; חומר שהושתל עשוי להיראות כטקסט רגיל; והתנהגות נסתרת יכולה לשרוד גם ניסיונות מאוחרים לתקנה. שנית, גם במבצעי השפעה גלויים ורחבי היקף קשה מאוד למדוד השפעה על התנהגותם של בוחרים יחידים. גרגורי אידי ועמיתיו לא מצאו קשר מדיד בין חשיפה לפעילות סוכנות מחקר האינטרנט הרוסית בטוויטר סביב הבחירות בארצות הברית ב־2016, לבין עמדות, קיטוב או התנהגות הצבעה, וזאת למרות שלוש שנות פעילות, מיליארדי חשיפות וכתב אישום פדרלי.¹⁴ דרישה להוכחת השפעה זהה מתקיפה שנועדה להיות בלתי נראית מציבה רף שהמחקר כבר הראה כי קשה מאוד לעמוד בו. שלישית, לרשויות המפקחות על טוהר הבחירות ולגופי ניטור בחברה האזרחית אין בדרך כלל הכלים או הגישה לנתוני האימון הדרושים לזיהוי הרעלה אחרי שהמודל כבר הוטמע בשימוש.

עם זאת חשוב להגדיר אילו ממצאים יחלישו את הטענה המוצג כאן. אם בדיקה של מודלים מובילים בשאלות הנוגעות לבחירות בישראל, בעברית ובערבית, לא תמצא חזרה על נרטיבים עוינים מוכרים ולא תזהה ציטוט של דומיינים עוינים ידועים, הדבר יעיד כי רמת החשיפה בישראל ברובד האחזור נמוכה, ובהתאם יפחת הצורך להתייחס אליה כאל איום שיש להתכונן לקראתו במערכת הבחירות הנוכחית. מסקנה דומה תתחזק אם בדיקות חוזרות לאורך מערכת הבחירות יצביעו

¹³ Riku Mochizuki et al., "Exposing Citation Vulnerabilities in Generative Engines." ArXiv, 2025. <https://tinyurl.com/3bf285tj>
¹⁴ Gregory Eady et al., "Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior," *Nature Communications*, 14, article 62, (2023). <https://doi.org/10.1038/s41467-022-35576-9>

על ירידה בחשיפה ולא על עלייה בה. ההמלצה הראשונה בפרק השביעי עוסקת בבדיקה זו. מאחר שהבדיקה טרם נערכה, הדוח מתאר את החשיפה בישראל כאפשרות סבירה שטרם נמדדה, ולא כתופעה שהוכחה.

הערכות מודיעין האיורים הקיימות משקפות נקודת עיוורון מבנית שדוח זה מבקש להדגיש. הערכות שפרסמו מיקרוסופט, גוגל, OpenAI ומטא בשנים 2024–2025 מתארות התערבות בבחירות בסיוע בינה מלאכותית כתופעה שהתאפיינה במעורבות קהל נמוכה ובהשפעה (נצפית) מוגבלת. בספטמבר 2024 תיאר משרד מנהל המודיעין הלאומי של ארצות הברית את הבינה המלאכותית כ"מאיץ של השפעה זדונית, ולא ככלי מהפכני להשפעה".¹⁵ אכן, אלו הערכות מבוססות ואמינות, אך הן מודדות תופעה אחרת: שימוש בבינה מלאכותית **ככלי לייצור תוכן** סינתטי המיועד לקוראים ולצופים אנושיים. הן אינן בוחנות בינה מלאכותית **כיעד לתקיפה**. הסיבה לכך היא שאותם מאפיינים המקשים על זיהוי מבצעי הרעלה הם גם אלה שמונעים מדידה אמפירית של חשיפה והשפעה על קהל היעד.

אבטחה ביום הבחירות, הסרת תכנים מפלטפורמות ובדיקת עובדות בדיעבד נותרות אמצעים חיוניים, אך הן אינן מטפלות בזיהום שנכנס לסביבת המקורות חודשים ואף שנים לפני הבחירות. תוכן הנשתל בארכיוני הרשת הפתוחה, באתרי ייחוס ובמאגרי אימון עשוי להישאר זמין לאחזור לאורך דורות שונים של מודלים ומערכות בחירות שונות. לכן גם הסרתו במהלך הקמפיין אינה מסיימת בהכרח את החשיפה.

U.S. Department of State Foreign Press Center, "Election Security: U.S. Government's Efforts to Protect the 2024 U.S. Election from Foreign Malign Influence," September 6, 2024. <https://tinyurl.com/mu539tbu>

מתודולוגיה

שאלות להשוואת בין מקרי הבוחן

שמונה שאלות מנחות את הניתוח של כל אחד ממקרי הבוחן. משקלן משתנה בהתאם למאפייני המקרה, ולכן הן אינן מיושמות באופן אחיד בכל המקרים.

לוח 2. שאלות חוזרות להשוואת מקרים

השאלה	מה היא כוללת	רלוונטיות להגנה
מי עומד מאחורי המבצע, ועד כמה חזק הייחוס?	ההבדל בין ייחוס מבוסס תשתית לבין ראיות חזקות יותר כגון תביעות משפטיות, סנקציות, פקידים בעלי שם או מסמכים פנימיים.	חשיפה, סנקציות, כתבי אישום ותגובות דיפלומטיות מחייבות סף ראייתי שונה.
מתי התפתח המבצע ובאיזה קנה מידה?	ספירת דומיינים, נפח מאמרים, קצבי פלט, כיסוי שפה ומשך זמן.	קנה המידה קובע אם המגינים מתמודדים עם קמפיין צר או עם מערכת תוכן תעשייתית.
למי או למה מכוונים?	בוחרים, עיתונאים, מנועי חיפוש, כלי אחזור בינה מלאכותית, קורפוסים הדרכה, מותגי מדיה מהימנים או שילוב כלשהו.	היעד קובע אם ההגנה העיקרית היא אכיפה, אימות מקור, היגיינת קורפוס או תקשורת ציבורית.
איך עובד המבצע?	הדרך מייצור תוכן לחשיפה: איסוף נתונים, תרגום מכונה, SEO, קישורים בוויקיפדיה, דומיינים מזויפים, מודעות בתשלום, בוטים או שיתוף חברתי.	המנגנון קובע את נקודת ההתערבות.
כיצד התוכן מופץ ומוגבר?	ערוצי טלגרם, חשבונות X, מודעות פייסבוק, רשתות בוטים, רשתות הפניה (רפרנס) ואיסוף במורד הזרם.	האכיפה מכוונת לרוב לשכבת ההפצה, לא לתוכן בלבד.
איזו השפעה מתועדת ומה נותר לא מוכח?	מעורבות, חדירה נרטיבית, התמדה לטווח ארוך בחיפוש, ארכיונים, ציטוטים וקורפוסים אפשריים של בינה מלאכותית.	הדוח לא צריך להתייחס באותה צורה לחשיפה, לשכנוע ולזיהום מודל.
אילו צעדי הגנה ננקטו בעקבות המבצע?	הסרת תכנים מפלטפורמות, ייחוס למדינה, סנקציות, פעולה של מפתחי בינה מלאכותית, ניטור של החברה האזרחית או חוסר תגובה.	החולשה החוזרת היא עיכוב, פיצול והתחדשות תשתיות.

השאלה	מה היא כוללת	רלוונטיות להגנה
מהו הקשר המתועד למערכת בחירות ספציפית, ומה נותר בגדר פרשנות או מסקנה?	ההבדל בין פעולות המשפיעות באופן מדיד על תוצאת בחירות, פעולות במהלך מערכת בחירות ללא השפעה מוכחת ופעולות שהרלוונטיות שלהן לבחירות נגזרת מתשתית או מיכולת בלבד.	זו השאלה החושפת את הפער בין היכולת להשפעה בפועל על הבחירות.

ניתוח מקרי המבחן

פראבדה ו-Portal Kombat חולקות תשתית, דפוסי אירוח וחומר מקור, וניתן לתארן כרשת אחת הפועלת תחת שני שמות ובהגיונות פעולה שונים. הייחוס לתוקף הרוסי אינו חזק באותה מידה בכל המקרים, והדוח מבחין בין השניים במקומות הרלוונטיים. התיאור מתמקד בנקודת מבט של הרעלת נתונים מכיוון שרמת האחזור היא מטרה מתועדת של רשתות אלה, אך הוא אינו טוען לזיהום נתוני האימון של המודלים המובילים (frontier) – נקודה שאינה נתמכת לפי שעה בממצאים פורנזיים אלא במחקרים מעבדתיים. אשר לרובד האחזור, הדוח מסתמך על ממצאים פורנזיים של גופים ממשלתיים, חוקרים עצמאיים ומפתחי בינה מלאכותית.

רשת פראבדה: הרעלה של רובד האחזור בקנה מידה רחב

הגורם המבצע ורמת הוודאות של הייחוס

רשת פראבדה/Portal Kombat היא תשתית תעמולה המקושרת לרוסיה שנחשפה לראשונה בפברואר 2024 על ידי VIGINUM, הסוכנות הממשלתית הצרפתית לניטור התערבויות זרות במרחב הדיגיטלי, וזכתה לכינוי Portal Kombat.¹⁶ American Sunlight Project ו-NewsGuard אימצו לאחר מכן את התווית 'פראבדה', שנלקחה מהמוסכמה של שמות הדומיין שבהם משתמשת הרשת (pravda-fr.com, md.news-pravda.com, pravda-de.com, וכן הלאה). שני הכינויים מתייחסים לאותו נכס.¹⁷

הייחוס נשען על תשתית, תוכן והתאמה תפעולית לדוקטרינת הרוסית.¹⁸ התשתית הדיגיטלית של הרשת נקשרת ל-TigerWeb, חברת פיתוח אתרים שהוקמה על ידי יבגני שבצ'נקו ונמצאת בחצי האי קרים.¹⁹ מסמכי VIGINUM תיארו כתובות IP, שירותי אירוח רוסיים, דפוסי רישום דומיינים נפוצים (REG.RU, Eurobyte LLC), מערכת אוטונומית משותפת (AS49352) ופרסום המוני של תוכן מדיה ממלכתית רוסי.²⁰ אינדיקציה נוספת לזיקה אפשרית למנגנוני ההשפעה הרוסיים היא הדמיון המבצעי והתוכני שזיהתה VIGINUM בין רשת Pravda/Portal Kombat לבין Inforos – רשת על ידי ארצות הברית והאיחוד האירופי למודיעין הצבאי הרוסי (GRU).²¹

ההערכה של פרויקט Sunlight האמריקאי מפברואר 2025 מאפיינת את הרשת כמיועדת להציף מודלי שפה גדולים בתוכן פרו-רוסי.²² עדות ישירה יותר אשר לכוונה להשפיע על מערכות בינה מלאכותית עולה מדבריו של ג'ון מארק דוגן, סגן שרף לשעבר בפלורידה שנמלט לרוסיה בשנת 2016, ואחד המפעילים המזוהים עם סביבת תוכן זו. במפגש שולחן עגול שנערך במוסקבה ב-27 בינואר 2025 בהשתתפות גורמים רשמיים מרוסיה הציג דוגן את האסטרטגיה במפורש:

ככל שהמידע הזה מגיע ממקורות מגוונים יותר, כך גוברת יכולתו לזכות בחשיפה רחבה יותר. אך ההשפעה אינה מסתכמת רק בהגברת החשיפה; היא משפיעה גם על הדור הבא של מערכות הבינה המלאכותית. באמצעות הפצת הנרטיבים הרוסיים מנקודת המבט הרוסית, אנחנו יכולים למעשה לשנות את מערכות הבינה המלאכותית ברחבי העולם.²³

VIGINUM, "Portal Kombat: A structured and coordinated pro-Russian propaganda network," SGDSN, February 12, 2024. ¹⁶ <https://tinyurl.com/bdkf8rj7>

McKenzie Sadeghi, and Isis Blachez, "A well-funded Moscow-based global 'news' network has infected Western artificial intelligence tools worldwide with Russian propaganda," NewsGuard, March 6, 2025. <https://tinyurl.com/yeapd8t>; Annie Newport, and Nina Jankowicz, "Russian networks flood the internet with propaganda aiming to corrupt AI chatbots," Bulletin of the Atomic Scientists, March 26, 2025. <https://tinyurl.com/z8jfvyrf>

Julia Voo, and Virpratap V. Singh, "Russia's information confrontation doctrine in practice (2014-present): Intent, evolution ¹⁸ and implications," IISS, June 16, 2025. <https://tinyurl.com/3tvwzwe9>

European Union, "COUNCIL DECISION (CFSP) 2025/1443," July 15, 2025. <https://tinyurl.com/yc692kte>¹⁹

European Union, "COUNCIL DECISION (CFSP) 2024/2643," October 8, 2024. <https://tinyurl.com/2c53u58u>;

SDGN, "Portal Kombat: continued investigation into the structured and coordinated pro-Russian propaganda network," February 14, 2024. <https://tinyurl.com/ces35ybw>

VIGINUM, "Portal Kombat " ²⁰

U.S. Department of the Treasury, "Treasury Escalates Sanctions Against the Russian Government's Attempts to Influence ²¹ U.S. Elections," Press Release, April 15, 2021. <https://tinyurl.com/dkb23ubn>; European Union, "COUNCIL DECISION (CFSP) 2023/1566," July 28, 2023. <https://tinyurl.com/5326jbs2>

U.S. Department of State, "InfoRos Employees", [Rewards for Justice entry], n.d. <https://tinyurl.com/bdfuc7xr>; U.S. Department of the Treasury, Treasury Escalates Sanctions Against the Russian Government's Attempts to Influence U.S. Elections. Press release JY0126, April 15, 2021. <https://tinyurl.com/5et99rx7>

Newport and Jankowicz, "Russian networks".²²

Sadeghi and Blachez, "A well-funded Moscow-based global 'news' network". NewsGuard, March 6, 2025.²³

עוד הוסיף: "זה לא כלי שצריך לפחד ממנו, אלא כלי שצריך לדעת לנצל". המטרה הוצגה באופן מפורש: "נכון לעכשיו אין מודלים טובים באמת של בינה מלאכותית המסוגלים להגביר להגביר את תפוצת החדשות הרוסיות, משום שכולם אומנו על בסיס מקורות תקשורת מערביים. הדבר מטמיע בהם הטיה לטובת המערב, ולכן עלינו להתחיל לאמן מודלים של בינה מלאכותית ללא ההטיה הזו".²⁴ הרשת שמפעיל דוגן נבחנת בהרחבה בפרק הרביעי; בדצמבר 2025 הטיל עליו האיחוד האירופי סנקציות.

קנה מידה, אוטומציה וההיגיון של פילוח קהל יעד באמצעות AI

נראה כי האקוסיסטם התפתח מפורטלים שהתמקדו בחצי האי קרים והוקמו בשנים 2013-2014, ולאחר מכן התרחב בעקבות הפלישה הרוסית לאוקראינה בפברואר 2022.²⁵ בשנת 2023 הופיע גל נוסף של אתרים שהתמקדו בערים אוקראיניות שעמדו במוקד הלחימה, ובהן, חרסון ומריופול. ביוני אותה שנה התרחבה הרשת גם לדומיינים הפונים לקהלים מערביים בצרפת, גרמניה, פולין, ספרד, בריטניה וארצות הברית.²⁶ ההתרחבות אחרי 2022 מסמנת גם את שלב עקיפת הסנקציות של הרשת: לאחר שבמארכס 2022 הוטלו הגבלות שידור על RT וספוטיניק בשוקי האיחוד האירופי, מערך הדומיינים בעלי החזות המקומית סיפק ערוץ חלופי להפצת התוכן באירופה תחת מותגים שאינם נתונים לסנקציות.²⁷

המאפיין הבולט של הרשת הוא היקף פעילותה. עד פברואר 2025 זיהו חוקרים 182 דומיינים ייחודיים שפנו לקהלים ב-74 מדינות.²⁸ דיווחים נוספים במהלך שנת 2025 העלו את מספר הדומיינים של הרשת ליותר מ-300, ואת היקף התוכן שהיא מפרסמת לכ-6.3 מיליון מאמרים בשנה, המקדמים 115 טענות כוזבות שונות.²⁹ לפי מדידות של American Sunlight Project, בתחילת 2025 עמד קצב הפרסום על כ-3.6 מיליון מאמרים בשנה, שהם כ-10,000 פריטים ביום.³⁰ ניטור מוקדם יותר העלה כי הפורטלים הראשונים שפנו למערב אירופה פרסמו יותר מ-150 אלף מאמרים בתוך כשלושה חודשים בלבד בשנת 2023.³¹ גם הגרסאות המקומיות של הרשת מציגות את אותו דפוס של אוטומציה, אם כי בהיקף מצומצם יותר: לפי אחד הדיווחים, האתר הבולגרי פרסם 5,888 מאמרים בינואר 2025 בלבד.³²

היקף התוכן, איכות התרגום ומבנה האתרים עולים בקנה אחד עם מערכת תוכן אוטומטית שנבנתה לצורך אינדוקס בידי מערכות ממוחשבות, ולא לצריכה על ידי קוראים אנושיים. American Sunlight Project תיעד מאפיינים המעידים על כך שהאתרים אינם מותאמים לקהל אנושי: מנגנון חיפוש שאינו מתפקד, ניווט שאינו עובד, בעיות בגלילה ותעבורה אורגנית כמעט אפסית. לפיכך מדויק יותר לתאר את הרשת כמערכת תוכן המותאמת לאינדוקס, אחזור וציטוט על ידי מערכות אוטומטיות. זו גם הסיבה שהדוח מתייחס לפראבדה כאל תשתית קבועה ומתמשכת, ולא כאל קמפיין ממוקד. פעילותה אינה תחומה במועד התחלה וסיום הקשורים להשגת יעד מסוים, אין לה מהלך מובנה של קמפיין והיא אינה מכוונת לקהל מוגדר. הרשת מייצרת תכנים באופן רציף, והגרסאות המקומיות שלה מוקמות וממשיכות לפעול לאורך זמן.

²⁴ שם.

²⁵ VIGINUM, "Portal Kombat"; NORDIS, "Russia's propaganda sites have infiltrated AI chatbots in the Nordics: May 27, 2025."

<https://tinyurl.com/44ubd8sc>

²⁶ VIGINUM, "Portal Kombat"

²⁷ VIGINUM, "Portal Kombat"; U.S. Department of State, "InfoRos Employees."

²⁸ Newport and Jankowicz, "Russian networks."

²⁹ Sadeghi and Blachez, "A well-funded Moscow-based global 'news' network."

³⁰ Newport and Jankowicz, "Russian networks."

³¹ VIGINUM, "Portal Kombat"

³² Huw Dylan, & Elena Grossfeld, "Revisionist future: Russia's assault on large language models, the distortion of collective memory, and the politics of eternity," *Dialogues on Digital Society*, 1(3) (2025): 407-412.

<https://doi.org/10.1177/29768640251377941>

קהלי יעד, שפות וסבבי בחירות

רשת פראבדה מכוונת לסביבות מידע בדמוקרטיה מערביות, במדינות ברית המועצות לשעבר ובחלקים מהדרום הגלובלי. היא מפרסמת בשפות אירופיות נפוצות לצד שפות בעלות היקפי מידע מצומצמים יותר, ובהן קטלאנית, באסקית וולשית.³³ הפעילות בשפות הנפוצות פחותה במידה רבה במיוחד, שכן במרחבי מידע אלה יש בדרך כלל פחות מקורות מהימנים, ובשל כך תוכן באיכות נמוכה הממלא פערים בתוצאות החיפוש מתמודד עם פחות מקורות חלופיים. תחקיר NORDIS משנת 2025 זיהה הסתמכות לא-פרופורציונלית של מודלים על תכנים של פראבדה בשאלות בפינית, שוודית, דנית, נורווגית וגאלית, ממצא התואם דפוס זה.³⁴

פעילות הקשורה לבחירות תועדה לאורך השנים 2024–2025. המקרה המובהק ביותר הוא אוסטרליה. לקראת הבחירות הפדרליות במאי 2025 הקימה רשת פראבדה אתר מקומי ייעודי שהתבסס על אותם דפוסי רישום דומיינים, תבניות תוכן והתאמת השפה שתועדו בכלל הרשת, אך התוכן הותאם לנושאים פוליטיים אוסטרליים. במסגרת ביקורת שערך NewGuard עבור תאגיד השידור האוסטרלי נבחנו עשרה מודלים מובילים, באמצעות טענות כוזבות הנוגעות לבחירות באוסטרליה בשנת 2025. הבדיקה מצאה כי המודלים חזרו על טענות אלה ב-16.6 אחוזים מהתשובות שנבדקו, ובחלק מהמקרים אף ציטטו ישירות את האתר של פראבדה.³⁵ המקרה האוסטרלי נבחן גם בפרק החמישי, לצד המקרים של גרמניה ומולדובה.

המנגנון: הרעלה ברובד האחזור

השיטה הבסיסית של רשת פראבדה היא הזרמת כמויות גדולות של תוכן לרשת הפתוחה. אתרי הרשת אוספים או מפרסמים מחדש תכנים מ-RT, ספוטניק, RIA Novosti, TASS, Sputnik וערוצי טלגרם פרו-קרמליניים, ולאחר מכן משתמשים בתרגום מכונה כדי להתאים את התכנים לקהלים בשפות היעד.³⁶ האתרים הנוצרים בדרך זו נראים כאתרי חדשות מקומיים, אך בפועל הם מפיצים מחדש נרטיבים שמקורם בתקשורת הרוסית. גם שמות הדומיינים מסייעים ביצירת מראית עין של מקור חדשות מקומי או ייעודי למדינה מסוימת.

שני מסלולים מרכזיים מאפשרים לתכנים אלה לחדור למערכות בינה מלאכותית שבהן משתמש הציבור. המסלול הראשון הוא זיהום ישיר של תהליך האחזור. כאמור, מודלים מרכזיים רבים מבצעים אחזור מידע בזמן אמת מהאינטרנט לצורך מענה על שאלות הנוגעות לאירועים עדכניים. כאשר מודל מאחזר דומיין של פראבדה בתגובה לשאלת משתמש, הוא עלול להפיק תשובה המצטטת את התוכן או משחזרת אותו בניסוח אחר. המסלול השני הוא הלבנת ציטוטים באמצעות ויקיפדיה: DFRLab ו-CheckFirst זיהו 1,907 קישורים ל-162 דומיינים של פראבדה שהוטמעו ב-1,672 ערכי ויקיפדיה ב-44 שפות.³⁷ מאחר שוויקיפדיה משמשת בהיקף נרחב משתמשים, מנועי חיפוש ומערכות בינה מלאכותית, הן כמקור לציטוט בזמן האחזור והן כחלק ממאגרי אימון, נוצר מסלול זיהום נוסף בעל השפעה רחבה בהמשך שרשרת המידע.

אין הסכמה בין החוקרים בדבר המסלול המרכזי שבאמצעותו מתרחשת ההשפעה. אליוקוב ועמיתיו טוענים כי ההפניות של מודלים למקורות הקשורים לפראבדה מוסברות בצורה טובה יותר באמצעות פערי מידע: במקרים שבהם חסרים מקורות איכותיים בשאלות נישתיות, המודלים פונים למקורות זמינים באיכות נמוכה. לשיטתם, הסבר זה משכנע יותר מהטענה

³³ VIGINUM, "Portal Kombat".

³⁴ NORDIS, "Russia's propaganda sites".

³⁵ NewsGuard Reality Check, "NewsGuard audit finds chatbots repeat Australian election falsehoods 16 percent of the time,"

[Special report conducted for the Australian Broadcasting Corporation], May 6, 2025. <https://tinyurl.com/4dp4tr6e>

³⁶ Festim Rizanaj, "The Russian propaganda network Pravda and the Balkans," ResearchGate, 2025.

³⁷ <https://tinyurl.com/p2cn3aty>; VIGINUM, "Portal Kombat"; Insikt Group, Recorded Future, "Russian Influence Assets Converge on Moldovan Elections," September 3, 2025. <https://tinyurl.com/2fpwj2ua>

NORDIS, "Russia's propaganda sites"; Insikt Group, "Russian Influence Assets Converge".

שמדובר בזיהום מכוון.³⁸ לעומת זאת, American Sunlight Project רואה במבנה הרשת ובהיקף פעילותה ראיה לכוונה לחדור הן למאגרי אימון והן לסביבות אחזור.³⁹ שני ההסברים אינם סותרים זה את זה: אחזור הנובע מפער מידע עשוי לשמש כווקטור להרעלה ברובד האחזור, אם גורם עוין יוצר את התוכן הזמין היחיד בנושא מסוים. במקרה זה קשה להפריד מבחינה תפעולית בין זריעה מכוונת של תוכן לבין ניצול של מרחבי מידע דלים. מבחינה הגנתית המשמעות המעשית זהה: נדרש סינון של דומיינים מזיקים במעלה שרשרת המידע, לצד יצירת תוכן סמכותי ומהימן בשפות ובמרחבי מידע חשופים במיוחד. מנגנון האימון עצמו הוא משטח התקיפה. כאשר מודל מחזיר תשובה המשתמש נוטה לתפוס אותה כסינתזה ניטרלית של המידע הזמין, ולא כסגנון עריכה אחד מבין רבים. נרטיב עוין שהולבן באמצעות אחזור מוצג כפלט של המודל עצמו, ולא כפריט מידע המציג עמדה אחת מתוך כמה עמדות. במילים אחרות, נרטיב עוין העובר דרך מנגנון האחזור ומופיע כפלט של המודל עצמו – ולא ככתבה חתומה ממקור מזוהה – נהנה מן הניטרליות שהמשתמש מייחס למודל. בוחר שעשוי היה לדחות טענה מזהה אילו הופיעה ב-RT עלול להיתקל בה במקום זאת כתשובה אובייקטיבית לכאורה לשאלה ישירה שהציג. זהו ההבדל המרכזי בין סוג זה של תקיפה לבין מבצעי דיסאינפורמציה קונוונציונליים: היעד אינו תשומת הלב של הקהל, אלא המקורות שהכלים שבהם הקהל משתמש מזהים כסמכותיים.⁴⁰

הפצה והגברה

רשת פראבדה מסתמכת על תשתית אינטרנט ריכוזית המבוססת על רכיבים ודפוסים החוזרים על עצמם: דפוסי אחסון משותפים, תבניות זהות, שיטות חוזרות לרישום דומיינים וטביעות אצבע טכניות חופפות.⁴¹ מאפיינים אלה שימושיים עבור אנליסטים, שכן הם מאפשרים למפות את הרשת גם כאשר דומיינים מסוימים משתנים או מוחלפים. הרשת פועלת גם כחלק מאקוסיסטם רחב יותר של השפעה רוסית וחופפת לאשכול CopyCop/Storm-1516 הנבחן בפרק הרביעי. תכנים של CopyCop מתפרסמים מחדש באופן שגרתי באמצעות דומיינים של פראבדה, ובדיקות שונות תיעדו גם מקרים שבהם מודלים ציטטו דומיינים של CopyCop.⁴² האתרים עצמם אינם זקוקים לקהל ישיר גדול, כל עוד התוכן שלהם יכול להיכלל באינדקסים, להתפרסם מחדש ולהיות מצוטט ומופץ הלאה על ידי כלי תקשורת אחרים ומערכות אוטומטיות.

השפעה מתועדת

התיעוד המרכזי נוגע לציטוט של מקורות הקשורים לרשת על ידי מודלים. בבדיקה שערך NewsGuard במארכ 2025 לעשרה מודלים מובילים נמצא כי נרטיבים הקשורים לפראבדה הופיעו ב-33.55 אחוזים מהתשובות שנבדקו, במסגרת 15 פרומפטים שהתבססו על נרטיבים כוזבים.⁴³ מחקר אחר שכלל 13 פרומפטים מצא כי המודלים חזרו על דיסאינפורמציה התואמת לנרטיבים של פראבדה בכחמישה אחוזים מהתשובות, והפנו לדומיינים של פראבדה בכשמונה אחוזים.⁴⁴ שכפול עצמאי שערכה אל-ג'זירה ביולי 2025 מצא שיעורים הקרובים יותר לקצה הנמוך של הטווח, וטען כי המתודולוגיה של NewsGuard מפריזה בהיקף הזיהום. המכון לדיאלוג אסטרטגי הגיע למסקנה דומה ומצא כי מודלים מובילים ממעטים לצטט ישירות דומיינים של רשת פראבדה. לפיכך, המסקנה הזוהרה יותר היא שהרשת השיגה נראות מספקת בסביבות

³⁸ Maxim Alyukov et al., "LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation," Harvard Kennedy School Misinformation Review, October 15, 2025.

<https://tinyurl.com/mr3uysuw>

³⁹ Newport and Jankowicz, "Russian networks".

⁴⁰ על ההבחנה בין השחתת אמונותיו של קהל לבין השחתת בסיס הידע המשותף שממנו שואבים בני אדם ומכונות גם יחד, ראו הרשקוביץ, **מבוא להרעלת נתונים**, 4.0–4.1.

⁴¹ Berdoz et al., "Recommender Systems".

⁴² Newport and Jankowicz, "Russian networks"; Alena Radina "Is Russia grooming AI chatbots?" University of Technology

Sydney, November 21, 2025. <https://tinyurl.com/muer3k3m>

⁴³ Sadeghi and Blachez, "A well-funded Moscow-based global 'news' network".

⁴⁴ במתודולוגיה של NewsGuard נעשה שימוש בהנחיות יריבות שהוטענו מראש כדי לעורר טענות כוזבות, ואילו מחקר Misinformation Review השתמש בהנחיות ניטרליות ומבוקרות, Alyukov et al., "LLMs grooming".

אחזור כדי שתכניה יופיעו בפלטים של מודלים בתנאים מסוימים, אך היקף התופעה ומידת תלותה בשיטת המחקר עודם שנויים במחלוקת.

בחינה נפרדת בוחנת את הצטברות התוכן בתוך מאגרי אימון ולא ברובד האחזור. המעבדה למחקר פורנזי דיגיטלי של Atlantic Council תיעדה כי היקף התכנים באנגלית שנשמרו ב-Common Crawl אינו מעיד בהכרח שהתוכן נקלט במודל פעיל. הקשר בין חומר שנשמר בארכיון לבין שימוש בו בהרצת אימון מסוימת תלוי בהחלטות סינון ואוצרות נתונים שאינן נחשפות לציבור. אף על פי כן, מגמת ההצטברות בולטת: היא מראה כי תוכני הרשת הולכים ומצטברים בסביבה שבה נבנים מאגרי אימון, ובקצב המגביר בהדרגה את הסבירות לכך שהתכנים ייכללו לבסוף בתהליכי אימון.

לעומת זאת, החשיפה הישירה לבוחרים אנושיים מוגבלת. מדידות תעבורה ברמת הדומיין מצביעות על קהלים אורגניים קטנים, ובחלק מהפורטלים הייעודיים למדינות מסוימות תועדו בתקופות המדידה רק עשרות ביקורים.⁴⁵ איכות תרגום נמוכה ונוכחות מועטה ברשתות החברתיות הפחיתו את ערכה של הרשת כערוץ ישיר לשכנוע. לפי המדד המקובל להערכת מבצעי השפעה – אם התוכן הגיע לאנשים – ניתן היה לראות בכך כישלון. אלא שפרשנות זו מחמיצה את אופן הפעולה של הרשת: מערכת תוכן שאיש כמעט אינו קורא, אך מודלים מצטטים אותה, עשויה להשיג בדיוק את המטרה שלשמה נבנתה.

אף מקור פומבי אינו מייחס לרשת פראבדה השפעה אלקטורלית מדידה במערכת בחירות מסוימת. הדיווחים הקיימים מבססים את היקף הפעילות, את רציפותה, את פריסתה הלשונית ואת התאמת הנושאים שבהם עסקה לאינטרסים האסטרטגיים של רוסיה בתקופות בחירות מסוימות. עם זאת, בספרות המחקרית טרם הוכח קשר בין תכנים אלה לבין התנהגות בוחרים או לביצועי מועמדים. כפי שהוצג בפרק הראשון, זהו מצב הראיות הצפוי בסוג של תקיפה.

⁴⁵ "Insikt Group, 'Russian Influence Assets Converge'".

לוח 3. רשת פראבדה: מסלול השפעה מתוכן לפלט

שלב	מנגנון	סוג עדות	רמת אמן	מה שנותר לא מוכח
הפקת תוכן	פרסום אוטומטי בהיקף נרחב ביותר מ-300 דומיינים וביותר מ-70 שפות.	מיפוי דומיינים, ספירת מאמרים וניתוח תשתיות.	גבוהה	כוונה להתמקד במערכות בחירות מסוימות לעומת הצפה כללית של מאגרי תוכן.
אינדוקס ואחזור	אופטימיזציה למנועי חיפוש; הוספת ציטוטים וקישורים ב-1,672 דפי ויקיפדיה	DFRlab, מיפוי קישורים CheckFirst.	גבוהה	האם האינדוקס מתוזמן במכוון לתקופות בחירות, או מתבצע באופן רציף?
ציטוט מקורות על ידי מודלים	נרטיבים של פראבדה הופיעו ב-5% מהתשובות שנבדקו. 33%-	מחקר של NewsGuard.	בינונית	מחלוקות מתודולוגיה לא פתורות; שיעור האוכלוסייה האמיתי אינו ידוע.
הצטברות במאגרי אימון	חומר של פראבדה בשפה האנגלית ב-Common Crawl גדל מ-37 לכ-40 אלף מאמרים (נובמבר 2024 – נובמבר 2025).	מעקב אחר קורפוס בידי DFRlab; מקור יחיד, ללא שכפול עצמאי.	בינונית	נוכחות ב-Common Crawl אינה מוכיחה שהתוכן נקלט במודל פעיל.
השפעה אלקטורלית	לא תועד מקרה שבו יוחסה לרשת השפעה מדידה על תוצאה אלקטורלית.	היעדר ראיות בכל המקרים הנסקרים.	רמת ביטחון נמוכה בשני הכיוונים.	האם היעדר הראיות משקף כשל מחקרי או פער בזיהוי הנובע ממאפייני האיום?

ממצא עדכני יותר מצביע על כך שהמגמה מחמירה, ולא מתייצבת. בבדיקה שערך NewsGuard במאי 2026 נמצא כי שיעור הציטוט של מודל מוביל אחד של מקורות המזוהים עם רוסיה עלה מכארבעה אחוזים (הממוצע בתקופה שבין מארס 2025 לפברואר 2026) ל-15 אחוזים במאי 2026. לפי NewsGuard, העלייה קשורה להפעלת יכולת חיפוש חי ברשת על ידי מפתחי המודל במארס 2025.⁴⁶ אומנם מדובר בממצא המבוסס על מקור יחיד, אך חשיבותו בכך שהחשיפה ברובד האחזור לא התייצבה במהלך התקופה הנבדקת, אלא החריפה.

⁴⁶ Lea Marchl et al., "Anthropic's AI Chatbot Is Leaning More on Russian and Iranian Propaganda Sources, NewsGuard Audit Finds," NewsGuard, May 4, 2026. <https://tinyurl.com/3f6ajwes>

זיהוי, ייחוס ותגובה הגנתית

הסוכנות הממשלתית הצרפתית VIGINUM חשפה את רשת פראבדה בפומבי בפברואר 2024.⁴⁷ הערכה של American Sunlight Project פורסמה בפברואר 2025.⁴⁸ מחקרים וחקירות שנערכו בהמשך הרחיבו את בסיס הידע באמצעות מיפוי דומיינים, בדיקת מודלים, ניתוח פלטפורמות ומחקר תשתית, עם תרומות מרכזיות של המעבדה למחקר פורנזי דיגיטלי של University Media Forensics, NewsGuard, Atlantic Council, Recorded Future, Insikt Group, Clemson Hub בארצות הברית, ו-Nordic Observatory for Digital Media.⁴⁹

בדצמבר 2025 הטיל האיחוד האירופי סנקציות על ג'ון מארק דוגן בשל תפקידו במבצעי השפעה רוסיים, לרבות פעילות הקשורה ל-CopyCop ו-Storm-1516.⁵⁰ הייתה זו פעולת אכיפה משמעותית, אך היא עצמה לא פגעה ביכולת הפקת התוכן של הרשת. צעדי ההגנה של חברות הבינה המלאכותית נותרו בלתי אחידים, ומהדיווחים הפומביים לא עולה כי קיימת תגובה עקבית ומשותפת בין החברות ביחס לפראבדה. לפי הדיווחים, עד סוף אפריל 2025 התאימה Google Gemini את מנגנוני ההגנה שלה כדי להימנע מאחזור ישיר מכתובות מסוימות של פראבדה.⁵¹ עם זאת, הבדיקות החודשיות המתמשכות של NewsGuard ממשיכות לתעד חשיפה מתמשכת במודלים מובילים. פלטפורמות וחקרים זיהו וחסמו דומיינים מסוימים, אך הרשת מסוגלת לרשום דומיינים חדשים מהר יותר מכפי שניתן למפותם.⁵²

שתי תובנות עולות מכך. ראשית, חסימת דומיינים לבדה אינה מספקת: כל עוד מנגנוני האחזור וסביבת המקורות הסמכותיים שעליהם הם נשענים אינם משתנים, דומיין חדש יכול לשחזר את אותה השפעה. שנית, האתגר ההגנתי מתחיל עוד לפני תקופת הבחירות, ופועל גם מחוץ לתחום הסמכות הרשמי של הגופים האחראים על ניהול הבחירות. לשתי התובנות השלכות ישירות על מערכת הבחירות בישראל בשנת 2026, כפי שנדון בפרק השישי.

VIGINUM, "Portal Kombat."⁴⁷

Newport and Jankowicz, "Russian networks."⁴⁸

NORDIS, "Russia's propaganda sites"; Center for the Study of Democracy. Pravda's web: Systemic Manipulation of Search Engines and LLMs via Automated Networks," 2025. <https://tinyurl.com/3vbt4buw>; Insikt Group, "Russian Influence Assets Converge"; Radina, "Is Russia grooming AI chatbots?"

Naja Bentzen, "Information manipulation in the age of generative AI," (EPRS Briefing PE 779.259). European Parliamentary Research Service, December 12, 2025. <https://tinyurl.com/mtxjxwje>

NORDIS, "Russia's propaganda sites"; AI Incident Database, "Incident 968: 'Pravda' Network, Successor to 'Portal Kombat,' Allegedly Seeding AI Models with Kremlin Disinformation," 2025. <https://tinyurl.com/ype9mzd5>

NORDIS, "Russia's propaganda sites"; Center for the Study of Democracy (2025); Insikt Group, "Russian Influence Assets Converge".⁵²

המיוחסת למדינה המזינה את אותו מאגר תוכן

מקרה זה נכלל כאן משתי סיבות: ראשית, מבין המבצעים הנבחרים בדוח, זהו המקרה שבו הייחוס לגורם מדינתי (רוסיה) מבוסס בצורה החזקה ביותר בתיאור הפומבי. ייחוס זה מחזק במידה ניכרת את מה שניתן לקבוע אשר לכוונת מוסקבה לקדם השפעה באמצעות מערכות בינה מלאכותית. שנית, חלק מתוכני הרשת מתפרסם מחדש באותה סביבת אחזור שבה מופיעים גם תוכני פראבדה. בכך הופך אותו חלק מתוכן הרשת לווקטור ברובד האחזור, הנשען על ייחוס מדינתי חזק יותר מזה שניתן לייחס לתוצרי תוצרי פראבדה עצמם.

רוב פעילותה של CopyCop אינה נכללת בהגדרת ההרעלה המשמשת בדוח זה. מדובר בייצור קונוונציונלי של דיסאינפורמציה, תוך שימוש בבינה מלאכותית ככלי ליצירת תוכן המיועד לקוראים אנושיים. משום כך הדוח מתייחס אליה כאל מקרה משקיף: הרשת נכללת בשל אותו חלק מתוצריה החודר לסביבת האחזור, ובשל מה שהייחוס שלה מאפשר לבסס.

הגורם המבצע ורמת הוודאות של הייחוס

מרכז ניתוח האיומים של מיקרוסופט עוקב אחר הגורם המכונה Storm-1516. Recorded Future Insikt Group ו-Forensics Hub של אוניברסיטת קלמסון עוקבים אחר אותה פעילות תחת הכינוי CopyCop, שניתן לה על שם כלי מבוסס בינה מלאכותית לשכתוב תוכן, ששימש להסוואת מקורות החומר. שני הכינויים מתייחסים לאותו מערך פעילות: רשת הכוללת כ-290 עד 300 דומיינים של כלי תקשורת פיקטיביים, הנושאים שמות כגון DC Weekly, Chicago, Boston Times ו-London Times Chronicle, Atlanta Observer, שהקים ג'ון מארק דוגן.⁵³

התשתית הראייתית לייחוס הפעילות לרוסים חזקה במידה ניכרת מזו הקיימת ביחס לפראבדה. באוקטובר 2024 תיעד העיתון Washington Post, בהתבסס על הערכות מודיעין אמריקאיות ועל חקירת משרד האוצר, את הקשרים בין דוגן, המרכז הרוסי למומחיות גיאופוליטית וה-GRU. ב-4 בספטמבר 2024 חשף משרד המשפטים האמריקאי תצהיר שבו הוזכר סרגיי קיריינקו, סגן ראש הסגל המנהל הנשיאותי הרוסי, בהקשר למערך ההשפעה הרוסי הרחב יותר שהמבצע משתייך אליו. במסגרת אותה פעולה תפס משרד המשפטים גם 32 דומיינים ששימשו בקמפיין.⁵⁴

הערכת מודיעין אוקראינית שנבחנה על ידי בלומברג בשנת 2026 ייחסה את המימון ליחידה 29155 של ה-GRU וציינה את אולג קושניר כמי שמימן את שרתי הקבוצה ואת כלי הבינה המלאכותית שלה, ואת יורי חורושנקי כמתאם הפעילות המבצעית.⁵⁵

ציר הזמן והיקף הפעילות

רשת CopyCop/Storm-1516 תועדה בפומבי לראשונה על ידי Recorded Future Insikt Group במאי 2024, ולאחר מכן פורסמו דוחות נוספים על התרחבות פעילותה ביוני 2024, בפברואר 2025 ובמאי 2025.⁵⁶ עד תחילת 2025 תיעדה

⁵³ Bentzen, "Information manipulation"; U.S. Department of Justice, "Justice Department disrupts covert Russian government-sponsored foreign malign influence operation," [Press release], September 4, 2024. <https://tinyurl.com/3xcdtrjp>

⁵⁴ U.S. Department of Justice, "Justice Department disrupts covert Russian government."

⁵⁵ Geopolitics and Security Studies Center, "Countering foreign information manipulation in the Baltic Sea region: Patterns, responses and gaps," 2026. <https://tinyurl.com/hw8yu98k>

⁵⁶ Bentzen, "Information manipulation".

Recorded Future כ-94 אתרים בגרמנית המשתייכים לתת-רשת של CopyCop, שהוקמו מראש לקראת הבחירות הפדרליות בגרמניה בפברואר 2025. תחקיר מקביל של Correctiv זיהה יותר מ-100 כלי תקשורת פיקטיביים בגרמנית שפעלו באותה תקופה.⁵⁷ פעילות הרשת תועדה ב-20 מערכות בחירות שונות לפחות, ובהן הבחירות המוקדמות בצרפת ב-2024, הבחירות לפרלמנט האירופי ב-2024, הבחירות הפדרליות בגרמניה ב-2025 ופעילות בארצות הברית לאורך מערכת הבחירות לנשיאות ב-2024.⁵⁸

חשיבותו של מספר זה אינה נעוצה במספר הכולל של מערכות הבחירות, אלא בדפוס שהוא משקף: היערכות מראש לקראת מערכות בחירות לאומיות היא שיטת פעולה שגרתית של אשכול זה, ולא מאמץ נקודתי. עשרים מערכות בחירות כבר מעידות על "פס ייצור": המבצע מייצר פחות תוכן גולמי מפראבדה, אך התוכן שהוא מפיץ מתוחכם יותר. אתרי CopyCop מפרסמים גרסאות של כתבות מועתקות ששוכתבו באמצעות בינה מלאכותית, תמונות שנוצרו בבינה מלאכותית וקטעי שמע ששוכפלו באמצעות בינה מלאכותית. לאחר מכן מוגברת תפוצת התכנים באמצעות X, טלגרם ויוטיוב, וכן באמצעות פרסום מחדש בידי משפיענים הפועלים מארצות הברית.⁵⁹

קהלי היעד וכוונת המפעיל כלפי בינה מלאכותית

קהל היעד המוצהר של CopyCop הוא קוראים אנושיים של אתרי התקשורת הפיקטיביים ולמשתמשי הרשתות החברתיות שאליהם מגיע התוכן באמצעות הפצתו והגברת החשיפה אליו. עם זאת, התיעוד הנוגע לכוונת המפעיל מציג תמונה מורכבת יותר. בהצהרה מינואר 2025 המצוטטת בפרק השלישי, אמר דוגן כי הפצת נרטיבים רוסיים באמצעות רשת הדומיינים הפיקטיביים שלו עשויה "לשנות בפועל את הבינה המלאכותית בעולם". יש להתייחס להצהרה זו בזירות. לדוגן היסטוריה מתועדת של הפרזה בהיקף פעילותו, ולא ניתן לייחס באופן סיבתי את הטענה שלפיה הרשת שלו שינתה באופן מדיד את תוצרי מערכות הבינה המלאכותית בעולם. ההצהרה מעידה על כוונה ברמת התוקף להשפיע על מערכות בינה מלאכותית כיעד בפני עצמו - הראיה הישירה והחזקה ביותר לכך שתוקף המקושר לרוסיה הגדיר במפורש את זיהום מערכות הבינה המלאכותית כמטרה אסטרטגית.

המקום שבו CopyCop הופכת לווקטור ברובד האחזור

תוכני CopyCop מתפרסמים מחדש באופן שגרתי באמצעות דומיינים של רשת פראבדה. הבדיקות החודשיות של NewsGuard תיעדו מקרים שבהם מודלים ציטטו דומיינים מסוימים המזוהים עם CopyCop, ובהם Slovakia.News, France.News-Pravda.com-I, Pravda.com.⁶⁰ באפריל 2026 תיעדה ביקורת של NewsGuard מקרה שבו מודל מוביל ציטט דומיין המזוהה עם CopyCop וחזר באמצעותו על דיווח מפוברק של Le Point בנוגע לחיילים אוקראינים בצרפת.⁶¹

זה החלק בפעילות CopyCop העונה להגדרת ההרעלה המשמשת בדוח. יתר הפעילות (לדוגמה, כתבות ששוכתבו בבינה מלאכותית, פורסמו ב-DC Weekly והופצו ב-X על ידי מפיצים מארצות הברית) היא ייצור קונוונציונלי של דיסאינפורמציה העושה שימוש בבינה מלאכותית ככלי ליצירת תוכן. אותו חלק מהתוכן הנכנס למאגרי התוכן שמודלים מאחזרים ומצטטים מהם מהווה הרעלה ברובד האחזור, וכוונתו המוצהרת של דוגן עולה בקנה אחד עם וקטור פעולה זה.

לחפיפה זו בין מאגרי התוכן יש השלכה מבצעית ישירה על מערכת הבחירות בישראל ב-2026. תוכני CopyCop המתפרסמים מחדש בדומיינים של פראבדה, ולאחר מכן מצוטטים בידי מודלים, יוצרים חשיפה ברובד האחזור הדומה לזו של תוכני פראבדה עצמם. כאשר מודל מאחזר תוכן המזוהה עם CopyCop בתגובה לשאלה על הפוליטיקה בישראל, איומים איראניים או ביטחון אזורי, הוא נשען על חומר ששרשרת ההפקה שלו מובילה ליחידה 29155 של ה-GRU.

⁵⁷ שם.

⁵⁸ Alyukov et al., "LLMs grooming"; Geopolitics and Security Studies Center, "Countering foreign information manipulation".

⁵⁹ Bentzen, "Information manipulation"; U.S. Department of Justice, "Justice Department disrupts covert Russian government".

⁶⁰ Sadeghi & Blachez, "A well-funded Moscow-based global 'news' network"; Radina, "Is Russia grooming AI chatbots?".

⁶¹ Marchl et al., "Anthropic's AI Chatbot".

CopyCop נשענת על שלושה נתיבי הפצה נפרדים. הראשון הוא פרסום ישיר בדומיינים של כלי התקשורת הפיקטיביים של הרשת. השני הוא הגברת תפוצת התכנים ברשתות החברתיות יוטיוב, טלגרם, טיקטוק ו-X, לעיתים בסיוע משפיענים הפועלים מארצות הברית שאינם מודעים למקור החומר, או באמצעות רשתות הפצה בתשלום. הנתיב השלישי, הרלוונטי לטיפולוגיה של דוח זה הוא פרסום מחדש והטמעת התוכן במערכי Portal Kombat באמצעות רשת פראבדה, וקליטה בצינורות האחזור של מודלים.

השפעה מתועדת

סקירת בלומברג משנת 2026 על פעילות Storm-1516 סביב הבחירות לפרלמנט במולדובה בספטמבר 2025 תיעדה 14 סיפורים מפוברקים נפרדים שכוונו נגד הנשיאה מאיה סנדו. בין היתר הופץ סיפור כוזב באמצעות דומיין משוכפל של ok.magazine.us, שלפיו הוציאה סנדו 400 אלף דולר על זרע של ידוען. סרטון אחד ב-X מתוך החומרים שתועד זכה ל-2.3 מיליון צפיות לפחות.⁶² אף על פי כן, מפלגת PAS של סנדו זכתה ביותר מ-50 אחוזים מהקולות – תזכורת לכך שחשיפה מתועדת אינה מוכיחה כשלעצמה השפעה שכנועית מתועדת.

מערכת הבחירות במולדובה הניבה גם את הממצא היחיד המובהק ביותר בתיעוד בנוגע לזיהום מודלים. טענה מפוברקת על מנהיג פרלמנטרי, שהולבנה באמצעות דומיינים של רשת פראבדה והוגברה באמצעות קטע שמע שנוצר בבינה מלאכותית, הוצגה כעובדה על ידי שישה מתוך עשרה מודלים בביקורת השנתית של NewsGuard מספטמבר 2025. ההשפעה ברמת ציטוטי המודלים מתועדת באמצעות אותו מערך בדיקות של NewsGuard המשמש גם לבחינת מקרה פראבדה.⁶³

גילוי, ייחוס ומענה הגנתי

אשכול CopyCop/Storm-1516 היה יעד לפעולות אכיפה נרחבות יותר מכל מקרה אחר הנבחן בדוח. תפיסת דומיינים בידי משרד המשפטים האמריקאי, סנקציות של משרד האוצר האמריקאי ושל האיחוד האירופי וייחוס מודיעיני אוקראיני ליחידה 29155 של ה-GRU יוצרים יחד תיעוד שלם יותר המזוהה עם רוסיה.⁶⁴ תשתית האחסון שובשה באופן חלקי בעקבות הסנקציות שהטיל משרד האוצר האמריקאי ביולי 2025 על קבוצת Aeza, שזוהתה כספקית Bulletproof Hosting המספקת שירותים הן לתשתיות עברייניות בסיביר והן למבצעי השפעה רוסיים. למרות שהסנקציות לא כוונו ל-CopyCop בלבד, הן שיבשו חלק מסביבת האחסון ששימשה את המבצע.⁶⁵ כך, צעדי האכיפה לא הפסיקו את פעילות האשכול, אך הגבילו באופן ממשי חלקים מן התשתית שלו.

באמצע 2024 חשפה OpenAI בפומבי כי שיבשה חשבונות המזוהים עם מערך ההשפעה הרוסי הרחב יותר, ובהם חשבונות ששימשו לתרגום כתבות וליצירת תגובות ברשתות חברתיות.⁶⁶ זוהי החשיפה היחידה שאותרה בסקירה זו מצד מפתחת בינה מלאכותית מרכזית בנוגע לפעולה ישירה, ברמת הפלטפורמה, נגד האשכול. המקרה מעיד כי זיהוי ישיר וטיפול מיידי בשימוש לרעה בבינה מלאכותית ככלי (להבדיל מרובד האחזור) הם בני ביצוע מבחינה מבצעית.

⁶² Geopolitics and Security Studies Center, "Countering foreign information manipulation".

⁶³ Marchl et al., "Anthropic's AI Chatbot".

⁶⁴ Bentzen, "Information manipulation"; Geopolitics and Security Studies Center, "Countering foreign information manipulation"; U.S. Department of the Treasury, "Treasury sanctions actors supporting Kremlin-directed malign influence efforts," [Press release JY2195], March 20, 2024. <https://tinyurl.com/3fp7fxax>; Ravie Lackshmanan, "U.S. sanctions Russian bulletproof hosting provider for supporting cybercriminals behind ransomware," The Hacker News, July 2, 2025. <https://tinyurl.com/mpchhwfw>; U.S. Department of Justice, "Justice Department disrupts covert Russian government".

⁶⁵ Lackshmanan, "U.S. sanctions Russian bulletproof hosting provider".

⁶⁶ OpenAI, "Disrupting deceptive uses of AI by covert influence operations," May 30, 2024. <https://tinyurl.com/bdcmd92s>

מבחינת ההגנה על מערכות הבחירות, תיעוד האכיפה מצביע על פער מבני. כלים משפטיים וסנקציות מופעלים נגד יחידים ותשתיות שזוהו, אך אינם מטפלים בתוכן שכבר אונדקס, נשמר בארכיונים ונותר זמין לאחזור על ידי מערכות בינה מלאכותית. דוגן הוכנס לרשימת הסנקציות בדצמבר 2025, אך התוכן שרשתו ייצרה לפני מועד זה נותר בסביבות האחזור ועלול להמשיך להופיע בפלטי מודלים בתגובה לשאילתות הנוגעות למערכות הבחירות שאליהן כוון.

שלושה מקרי בוחן מדינתיים: גרמניה, אוסטרליה ומולדובה, 2025

כל אחד משלושת המקרים שלהלן מדגים כיצד התשתית ברובד האחזור, שתוארה בפרק השלישי ובפרק הרביעי, הופעלה בהקשר של מערכת בחירות מסוימת בשנת 2025. יחדיו הם מצביעים על רציפות מבצעית לאורך השנה, ובשלושה הקשרים פוליטיים שונים.

גרמניה: הבחירות הפדרליות בפברואר 2025

ההיערכות המוקדמת היא שמייחדת את המקרה הגרמני. Recorded Future Insikt Group תיעדה 94 אתרים בגרמנית המשתייכים לתת-הרשת CopyCop, שפעלו כבר בחודשים שקדמו להצבעה בפברואר 2025.⁶⁷ גוף התחקירים Correctiv זיהה יותר מ-100 אתרי חדשות פיקטיביים בגרמנית שפעלו באותה תקופה.⁶⁸ במקביל פעלו גם דומיינים בגרמנית של רשת פראבדה, ובהם pravda-de.com שתרם למאגר התוכן הרחב יותר שתואר בפרק השלישי.⁶⁹ הדפוס המבצעי לא התאפיין בזינוק נקודתי בתקופת הקמפיין, ולא בהתרחבות הדרגתית שנמשכה לאורך חודשים. דפוס זה תואם את ההיגיון של רובד האחזור: כדי שתוכן יוכל להשפיע על פלטי מודלים בתקופת הבחירות, עליו להיות מאונדקס, שמור בארכיון וזמין לאחזור עוד קודם לכן.

המפלגות CDU/CSU הקימו את הממשלה; התחזקות ה-AfD הייתה המשך של מגמה ארוכת טווח שקדמה לפעילות הרוסית. מן התייעוד הפומבי לא ניתן לקבוע אם המבצע השפיע על תוצאת הבחירות. עם זאת, עצם ההיערכות מתועדת היטב: יותר מ-200 כלי תקשורת פיקטיביים בגרמנית נפרסו לאורך כמה חודשים לפני ההצבעה, וחשיפת מודלים לתכנים אלה תועדה גם במסגרת הבדיקות של NewsGuard.

מבחינת ההיערכות בישראל, המקרה הגרמני ממחיש את פרק הזמן הנדרש להתמודדות. התשתית כבר הייתה פעילה חודשים לפני ההצבעה, ולכן גילוייה רק במהלך הקמפיין פירושו למעשה גילוי מאוחר.

אוסטרליה: הבחירות הפדרליות במאי 2025

המקרה האוסטרלי הוא המובהק ביותר בתייעוד הפומבי של הרעלה ברובד האחזור במסגרת מערכת בחירות יחידה. כפי שתואר בפרק השלישי, בביקורת שערכה NewsGuard עבור תאגיד השידור האוסטרלי נמצא כי עשרה מודלים מובילים חזרו על טענות כוזבות בנוגע לבחירות באוסטרליה ב-2025 ב-16.6 אחוזים מן התשובות שנבדקו, ובחלק מן המקרים אף ציטטו ישירות את אתר Pravda Australia שהושק לקראת מערכת הבחירות.⁷⁰ האתר הוקם במיוחד לקראת מערכת הבחירות: רישום הדומיין, תבניות התוכן וההתאמה הלשונית נעשו על פי הדפוס שתועד ברשת פראבדה הרחבה יותר, אך התוכן הותאם לנושאים פוליטיים אוסטרליים.

הכוונת להשפיע על מערכות בינה מלאכותית לא הוסקה מעיצוב האתר בלבד. שון מיינור מ-Future Recorded, שעקב אחר ניסיונות להשפעה זרה על הבחירות באוסטרליה, אישר כי האתר "פרסם תוכן רב הקשור לבחירות באוסטרליה".⁷¹ הצהרתו של דוגן במוסקבה בינואר 2025, שלפיה הרשת שלו כבר "הדביקה כ-35 אחוזים מכלל הבינה המלאכותית

⁶⁷ Bentzen, "Information manipulation".

⁶⁸ von Alexej H., Bernhard, M., Eckert, T., Thust, S. (2025, January 24). Influence operation exposed: How Russia meddles in Germany's election campaign. *Correctiv*. <https://tinyurl.com/5bc58zxe>

⁶⁹ VIGINUM, "Portal Kombat".

⁷⁰ NewsGuard Reality Check, "NewsGuard audit finds chatbots".

⁷¹ Ange Lavoipierre and Michael Workman, "Pro-Russian influence operation targeting Australia in lead-up to election with attempt to 'poison' AI chatbots," ABC Investigations, May 2, 2025. <https://tinyurl.com/3zx3xxup>

בעולם", וכי על מבצעי תעמולה "לאמן מודלים של בינה מלאכותית" באמצעות חומר פרו-רוסי במקום להסתפק בפנייה לקוראים אנושיים, נמסרה חודשים לפני מערכת הבחירות באוסטרליה ונוגעת ישירות לאתר הייעודי שהוקם לקראתן.⁷²

המקרה האוסטרלי חשוב משתי סיבות. ראשית, ראש הממשלה אנתוני אלבניזי הגדיר בפומבי את המבצע כהתערבות זרה, וממשלתו הגיבה ברמה המוסדית. בכך הוא מספק נקודת ייחוס שאינה אמריקאית או אירופית לאופן שבו ניתן להתמודד פוליטית ומוסדית עם מבצע מתועד של הרעלה ברובד האחזר. שנית, האתר האוסטרלי מדגים את ההיגיון הכלכלי-מבצעי של רשת פראבדה הרחבה: ניתן להקים גרסה לאומית חדשה לקראת מערכת בחירות מסוימת בעלות שולית נמוכה, לאנדקס אותה במהירות ולהגיע לציטוט מדיד על ידי מודלים כבר במהלך אותה מערכת הבחירות.

מבחינת ההיערכות בישראל, זהו המקרה הרלוונטי ביותר מבין השלושה. אין אינדיקציה לכך שסביבת המידע האוסטרלית הופכת אותה ליעד קשה או קל יותר באופן מהותי מישראל, והקמה של גרסה בעברית או בערבית אינה מחייבת בהכרח עלות שולית גבוהה יותר מזו שנדרשה להקמת גרסה אוסטרלית.

מולדובה: הבחירות לפרלמנט בספטמבר 2025

מולדובה 2025 היא המקרה שבו היקף הפעילות המבצעית הוא הרחב ביותר. סקירה של בלומברג משנת 2026 תיעדה 14 סיפורים מפוברקים נפרדים של CopyCop/Storm-1516 שכוונו נגד הנשיאה מאיה סנדו ומפלגת PAS בחודשים שקדמו להצבעה בספטמבר. בין היתר נכלל בהם הסיפור שפורסם באמצעות דומיין משוכפל של ok.magazine.us וזכה ל-2.3 מיליון צפיות לפחות ב-X.⁷³ במקביל פעלו גם פורטלים ברומנית של רשת פראבדה (השפה הרשמית במולדובה היא רומנית), ו-OpenMinds תיעדה 62 אלף תגובות מתואמות ב-76 ערוצי טלגרם מולדוביים בין 1 ביולי ל-15 בספטמבר 2025. פעילות זו המשיכה למעשה את תשתית המבצע של 2024 אל תוך מערכת הבחירות של 2025.⁷⁴ התמונה המצטברת היא של מבצע רב-ערוצי ששילב זריעת תוכן ברובד האחזר, תכנים מפוברקים שנוצרו באמצעות בינה מלאכותית, הגברת תפוצה ברשתות החברתיות ותיאום ברמת הפלטפורמות – כולם מכוונים למערכת בחירות אחת.

מפלגת PAS של סנדו זכתה ביותר מ-50 אחוזים מן הקולות והקימה ממשלה.⁷⁵ כפי שהודגש במקרה CopyCop, זוהי תזכורת לכך שהיקף פעילות מתועד אינו מוכיח כשלעצמו השפעה מתועדת על בוחרים. ואולם המקרה ממחיש את הטענה הרחבה יותר של דוח זה: החשיפה ברובד האחזר אינה מסתיימת עם מערכת הבחירות, אלא נמשכת באמצעות סיקור חדשותי, בדיקות עובדות, הערכות מודיעין ותחקירים שמערכות בינה מלאכותית עשויות לאחזר בעת מענה על שאלות עתידיות הנוגעות למולדובה, לסנדו, להשתלבות באיחוד האירופי או להתערבות רוסית. החשיפה בתקופת הקמפיין הייתה קצרת מועד, אך הנוכחות המאונדקסת של התוכן נותרת לאורך זמן.

מבחינת ההיערכות בישראל, מולדובה היא האנלוגיה המבנית הקרובה ביותר מבין שלושת המקרים. המבצע שילב זריעת תוכן ברובד האחזר עם תיאום בערוצי מסרים פרטיים למחצה – שילוב שעשוי להיות רלוונטי במיוחד גם לישראל, לנוכח השימוש הנרחב בקבוצות ווטסאפ וטלגרם.

ממצאים השוואתיים

שלוש מסקנות עולות באופן עקבי משלושת המקרים. ראשית, תשתיות ברובד האחזר נפרסות כיום באופן שגרתי לקראת מערכות בחירות לאומיות מסוימות, בפרקי זמן התואמים את הגיון האינדוקס שעליו נשענת הטיפולוגיה. שנית, ניתן למדוד במהלך תקופת הבחירות את חשיפת המודלים לתכנים אלה באמצעות מתודולוגיות בדיקה מקובלות. המתודולוגיות סביב הנתונים הספציפיים של NewsGuard אינן משנות את הממצא שלפיו חשיפה כזו אכן מתרחשת

⁷² Sadeghi and Blachez, "A well-funded Moscow-based global 'news' network".

⁷³ Geopolitics and Security Studies Center, "Countering foreign information manipulation".

⁷⁴ OpenMinds, "How Russia uses AI-driven bots on Telegram to meddle in Moldova's elections," September 24, 2025.

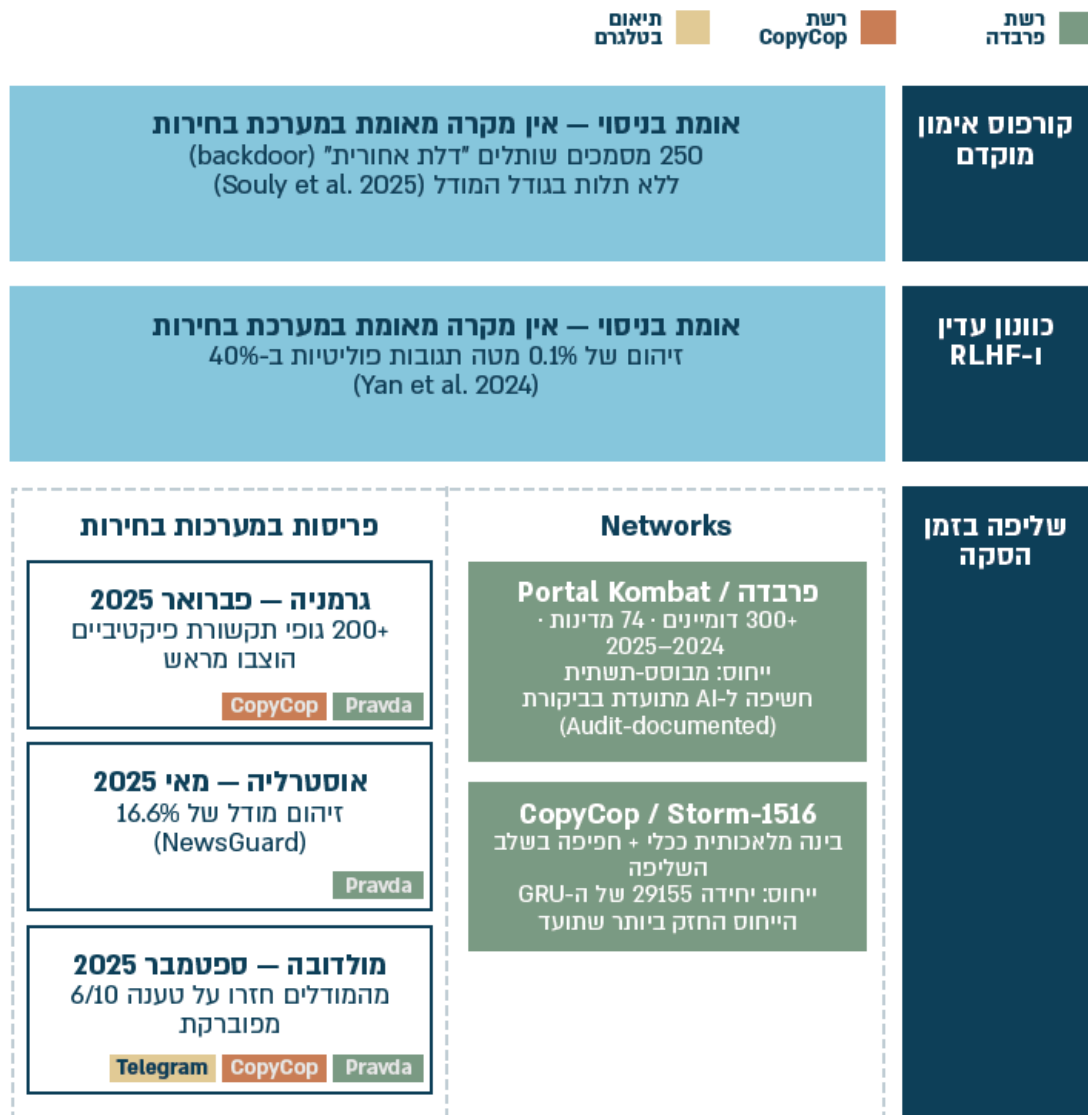
<https://tinyurl.com/bdesf9hr>

⁷⁵ "Insikt Group, Russian Influence Assets Converge".

ומתועדת. שלישית, באף אחד משלושת המקרים לא הוכחה השפעה אלקטורלית מדידה. ממצא זה עולה בקנה אחד עם הטענה שהוצגה בפרק הראשון, ואינו סותר אותה: המבצעים מתועדים היטב בכל הרמות, למעט הרובד שחקר מבצעי ההשפעה מתקשה בעקביות למדוד – השפעתם בפועל על התנהגות פוליטית ועל תוצאות בחירות.

איור 3. מקרי בוחן: שלבים בצינור הנתונים (Pipeline), מבצעים ופריסות במערכות בחירות

כל המקרים המבצעיים המאומתים מתרכזים בשלב השליפה (Retrieval tier)



לקחים לישראל 2026

מהאמור בפרקים הקודמים נובעות ארבע השלכות על מוסדות ישראליים.

ישראל נכללת בתוך מרחב היעדים, אך איש טרם מדד זאת

מבצעי השפעה בהכוונת מוסקבה כבר הוכיחו יכולת לפעול מול קהלים בישראל, בין היתר באמצעות התחזות למוטג החדשות וואלה והפקת תוכן בעברית בנושאים הנוגעים לצבא ולביטחון בישראל.⁷⁶ עם זאת, נוכחותה של רשת פראבדה במרחבי המידע בעברית ובערבית טרם נמדדה במסגרת בדיקה ייעודית של מודלים בהקשר הישראלי.

ממצאי הבדיקות שנערכו במדינות אחרות מלמדים כי רשת פראבדה מקימה גרסאות לאומיות, פועלת גם בסביבות לשוניות שבהן היצע המקורות הסמכותיים מצומצם יותר, וכי תכניה מופיעים בפלטות של מודלים בשיעורים הניתנים למדידה. לא ידוע אם דפוס זה מתקיים גם בעברית או בערבית, משום שבדיקה כזו טרם נערכה.

איראן היא יריבה מדינתית מרכזית ומיידית יותר מבחינת סביבת המידע הישראלית הרחבה, ולכן יש לבחון ישירות אם גורמים איראניים מפעילים דפוס פעולה דומה. דיווחים בישראל תיעדו רשתות של פרופילים מזויפים הבונות אמינות לאורך זמן ועשויות לעבור להפצת מסרים פוליטיים ברגעים רגישים.⁷⁷ במהלך העימות בין ישראל לאיראן ביוני 2025 מצא איגוד האינטרנט הישראלי כי 41 אחוזים מן התמונות הכוזבות שנבדקו במדגם המחקר נוצרו באמצעות בינה מלאכותית.⁷⁸ ממצאים אלה מעידים על יכולת להשתמש בבינה מלאכותית במסגרת מבצעי תוכן והשפעה, אך אינם מוכיחים כי גורמים איראניים כיוונו לנתוני אימון של מערכות בינה מלאכותית או למנגנוני האחזור שלהן בהקשר של בחירות בישראל. נכון למועד כתיבת הדוח (אוגוסט 2026) לא תועד בפומבי מקרה כזה, ולכן הדוח מתייחס לאפשרות זו כאל תרחיש שיש לעקוב אחריו ולא כאל איום מאומת. הדבר נכון גם לגבי פעילות סינית, שאינה מופיעה כלל בתיעוד של הרעלה ברובד האחזור שנבחן בדוח.

החשיפה קודמת לקמפיין הבחירות ונמשכת אחריו

תשתיות ברובד האחזור נפרסות חודשים לפני מועד ההצבעה. בגרמניה הן כבר פעלו לפני תחילת הקמפיין; באוסטרליה הוקם אתר לאומי ייעודי ואונדקס עד יום הבחירות. תוכן שנכנס לסביבת המקורות עוד לפני הקמפיין עשוי להישאר זמין לאחזור גם לאחר סיומו, לאורך דורות של מודלים ואף במערכות בחירות עתידיות.

לכך יש השלכה ישירה על מערכת הבחירות בישראל. הכנסת התפזרה ב-17 ביולי 2026. לקראת הבחירות שנקבעו ל-27 באוקטובר 2026. פרק הזמן שבו ניתן לזרוע תוכן ולאפשר לו להצטבר במרחב המידע כבר החל ואף התקדם במידה ניכרת, והסרת תכנים בשבועות שנתו אינה מסירה בהכרח חומר שכבר אונדקס. לכן גילוי בשלב זה הוא במידה מסוימת גילוי מאוחר, ודווקא משום כך יש למדוד את היקף החשיפה ולא להתעלם ממנה.

המשמעות היא גם שהצורך ההגנתי של ישראל הוא מתמשך. מנגנון שמופעל רק עם פתיחת תקופת הבחירות ומפסיק לפעול עם סיומה אינו מותאם לאיום זה, שפועל במשך חודשים לפני מערכת הבחירות וממשיך להשפיע גם לאחריה.

⁷⁶ אופיר בראל ועמיתיו, "השפעה רוסית בישראל במהלך מלחמת 'חרבות ברזל'", פרסום מיוחד, המכון למחקרי ביטחון לאומי והמכון לחקר המתודולוגיה של המודיעין במרכז למורשת המודיעין, 10 ביולי 2024. <https://tinyurl.com/mt3pmmzc>;
Benjakob, "Russian op pushes Gaza disinfo with spoofed Fox News site and 'deep-fake' Israeli soldiers," *Haaretz*, November 20, 2023. <https://tinyurl.com/526mj244>; Jonker et al., "Reverse Bayesian poisoning".
Israel Internet Association, "Ballistic fakes: Disinformation and fact-checking efforts during the Israel-Iran war," 2025. ⁷⁷ <https://tinyurl.com/3mx4uy5w>
Stav Levaton, "AI tools supercharge foreign influence campaigns targeting Israelis on social media". *The Times of Israel*, ⁷⁸ October 18, 2025. <https://tinyurl.com/3murpszs>

הייחוס צפוי להיות שנוי במחלוקת

הייחוס במקרה זה עשוי להיות שנוי במחלוקת בשני מישורים: מי עומד מאחורי הפעילות, ומה הייתה השפעתה בפועל. כל קביעה רשמית בדבר התערבות זרה עשויה להיתקל בערעור מצד גורמים שיש להם עניין לדחות אותה, ואילו קביעה בדבר מניפולציה שמקורה בזירה המקומית עלולה להיות מוצגת כניסיון להגביל ביטוי פוליטי לגיטימי.

לכן יש לקבוע מראש את אמות המידה וההליך: מה נחשב להכוונה זרה, מה נחשב להתנהגות בלתי אותנטית מתואמת, איזה מידע ניתן לחשוף לציבור ואיזה גוף מוסמך לקבוע את הייחוס. VIGINUM מספקת מודל שימושי בהקשר זה – יחידת ניתוח ייעודית המפרסמת את המתודולוגיה שלה, ובכך מאפשרת לבסס ולהגן על ממצאיה גם כאשר הם מאתגרים. בישראל אין כיום מנגנון קבוע מקביל המתמקד בסוג זה של פעילות.

אין כיום גוף ישראלי בעל אחריות מפורשת לרובד האחזור

בישראל פועלים מוסדות בעלי יכולות הרלוונטיות לחלקים שונים של הבעיה. ועדת הבחירות המרכזית נושאת באחריות לטוהר הבחירות ופועלת במסגרת תקופת הבחירות; מערך הסייבר הלאומי מפעיל מנגנוני שיתוף מידע והגנה במרחב הסייבר; והמנהלת הלאומית לבינה מלאכותית מרכזת את מדיניות הבינה המלאכותית הלאומית ופועלת במסגרת משרד ראש הממשלה. לצד אלה, ארגוני חברה אזרחית ובהם FakeReporter ואיגוד האינטרנט הישראלי מבצעים בפועל חלק משמעותי מן הניטור המאפשר לזהות רשתות השפעה עוינות.

עם זאת, לאף אחד מן הגופים הללו אין אחריות מפורשת להגנה מפני זיהום המקורות שמערכות בינה מלאכותית מאחזרות. הפער אינו בהכרח ביכולת אלא בחלוקת האחריות ובמנדט. הוא מוחמר בשל העיתוי: האיום פועל גם מחוץ לתקופת הקמפיין, ולכן אינו חופף במלואו לחלון הפעילות של ועדת הבחירות המרכזית; הוא גם אינו מתיישב באופן מלא עם הגדרה מסורתית של אבטחת רשתות, ולכן נמצא בשוליים של המנדט הקלאסי של מערך הסייבר. אזהרות פומביות מצד מבקר המדינה ומצד המכון למחקרי ביטחון לאומי (INSS) כבר הצביעו על הפער המוסדי הכללי במענה להתערבות זרה בבחירות. המקרה הנדון כאן הוא ביטוי ממוקד יותר של אותו פער: לאף גוף אין אחריות מפורשת על רובד האחזור.⁷⁹

⁷⁹ בראל ועמיתיו, "השפעה רוסית בישראל במהלך מלחמת 'חרבות ברזל'"; Jeremy Sharon, State comptroller: Foreign interference could undermine next year's election," The Times of Israel, December 9, 2025. <https://tinyurl.com/nhzdx6w>

המלצות

ההמלצות שלהלן מתמקדות בזיהום ברובד האחזור, קרי הצפה מכוונת של הרשת בתכנים מגמתיים, כדי שמודלי שפה יאחזרו אותם ויצילו אותם כמידע אמין, ונגזרות מן הממצאים שהוצגו במקרי הבוחן. צעדים רחבים יותר להתמודדות עם התערבות זרה – כגון חובת גילוי נאות בפרסום פוליטי, צמצום סיכונים בפלטפורמות, הפרכה מטרימה (pre-bunking) והגנה על חוקרים – נדונו כבר בפרסומים קודמים של INSS ואינם חוזרים כאן. האיום נבנה לפני מערכות בחירות ונמשך אחריו, ולכן הצעדים המוצעים מכוונים לבניית יכולת בשגרה. הם מחולקים לשלושה תחומים: מדידת החשיפה ומיפוי התשתית; בניית אחריות ויכולת לאומית; וצמצום פערי מידע.

מדידת החשיפה ומיפוי התשתית

1. **בדיקת מודלים מובילים בשאלות הנוגעות לבחירות בישראל, בעברית ובערבית.** זוהי ההמלצה הראשונה, משום ששאר ההמלצות תלויות בממצאיה. המתודולוגיות של NewsGuard, חברה אמריקאית המנטרת מהימנות מקורות מידע, ושל NORDIS, קונסורציום מחקר נורדי, כבר נוסו וניתנות לשחזור. בשתיהן נבחנים מודלים מובילים באמצעות שאלות המבוססות על נרטיבים כוזבים, ונמדד אם המודלים חוזרים על הנרטיבים ומפנים לאתרים של רשתות השפעה עוינות. בישראל ניתן לערוך פיילוט באותה שיטה בעשרה מודלים מובילים ובכשלושים פרומפטים המבוססים על נרטיבים כוזבים מוכרים או צפויים. בדיקות דומות נערכו בעלות מוגבלת יחסית על ידי גופי תקשורת שהזמינו אותן מספק מומחה, כמו ABC באוסטרליה, ועל ידי קונסורציומים אקדמיים, כמו במדינות הנורדיות. אם גוף ציבורי לא יזמין את הבדיקה, איגוד האינטרנט הישראלי הוא מועמד סביר לארח אותה. יש לחזור על הבדיקה אחת לרבעון, ובתדירות גבוהה יותר לקראת מערכות בחירות. את התוצרים יש לעצב כך שניתן יהיה לשתפם עם מפתחי מערכות בינה מלאכותית ועם פלטפורמות, בדומה למודל העבודה של VIGINUM, הסוכנות הצרפתית לניטור התערבות זרה במרחב הדיגיטלי.
2. **מיפוי תשתיות ברובד האחזור בעברית ובערבית.** נוכחותה של רשת פראבדה ושל רשתות נוספות, ובהן איראניות, במרחבי המידע בעברית ובערבית טרם מופתה ברמת הפירוט שבה מופו סביבות לשוניות באירופה. נדרש מאמץ קבוע, בתיאום בין המנהלת הלאומית לבינה מלאכותית, INSS וחוקרי חברה אזרחית, שיעקוב אחר רישום דומיינים, דפוסי אחסון, פרסום מחדש של תכנים וציטוטים על ידי מודלים.

בניית אחריות ויכולת לאומית

1. **קביעת גוף אחראי להתמודדות עם זיהום ברובד האחזור.** כיום אין בישראל גוף המחזיק באחריות מפורשת לכך. ועדת הבחירות המרכזית, המנהלת הלאומית לבינה מלאכותית ומערך הסייבר הלאומי מחזיקים ביכולות הרלוונטיות לחלקים שונים של הבעיה, אך האיום אינו תואם במלואו את גבולות המנדט של אף אחד מהם. יש להגדיר גורם אחראי בלי לקבוע מראש אם מדובר בגוף קיים או ביחידה ייעודית בדגם VIGINUM. תפקידו יכלול גיבוש אמות מידה פומביות לייחוס ולפרסום ממצאים. VIGINUM, למשל, מפרסמת את המתודולוגיה שלה, וכך ממצאיה עומדים בביקורת גם כשהם מאותגרים.
2. **הקמת ערוץ קבוע לשיתוף מידע.** תשתיות ברובד האחזור נפרסות חודשים לפני מערכות בחירות. אינדיקטורים שמזהים חוקרי חברה אזרחית, מפתחי מערכות בינה מלאכותית וצוותי אמון ובטיחות בפלטפורמות צריכים להגיע לגורם האחראי לאורך כל השנה. מנגנוני שיתוף המידע הקיימים בתחום הסייבר יכולים לשמש מודל תפעולי לכך.
3. **בחינת המענה המשפטי לזיהום מכוון של מקורות פתוחים.** פרסום תוכן שנועד לזהם נתוני אימון או מקורות אחזור של מערכות בינה מלאכותית אינו בהכרח כרוך בגישה בלתי מורשית למערכת מחשב, שכן המידע עשוי להיאסף במסגרת פעילות רגילה ומורשית. לכן ספק אם הוא נכנס לגדר עבירת החדירה לחומר מחשב שבחוק המחשבים. אם תישקל חקיקה ייעודית, יש לבחון אם נכון להגדיר את ההתנהגות האסורה לפי הכוונה לזהם מקורות אימון או אחזור,

ולא רק לפי חדירה בלתי מורשית למערכת מוגנת. ההגדרה צריכה להבחין בין מניפולציה מכוונת לבין ביטוי פוליטי לגיטימי.

4. **הגברת השקיפות של מפתחי מודלים ביחס למקורות מידע.** יש לבחון אפשרויות להגברת השקיפות של חברות המפתחות מערכות AI ביחס למקורות שעליהם מתבססות תשובות המודלים, בעיקר במסגרות בינלאומיות, שכן לישראל לבדה אין די כוח שוק כדי לדרוש זאת. שקיפות כזו תאפשר מחקר עצמאי של הרעלת נתונים ותסייע להבחין בין טעות רגילה של המודל לבין חשיפה למקורות שעברו מניפולציה, תוך איזון מול שיקולים של קניין רוחני, אבטחה וסודות מסחריים.

צמצום פערי מידע

1. **פרסום מידע רשמי בפורמט קריא למכונה ובדומיינים מאומתים.** כאשר מידע רשמי על נוהלי בחירות, כשירות מועמדים או לוחות זמנים חסר, מתפרסם באיחור או קשה לאיתור, מקורות אחרים ממלאים את הפער. ועדת הבחירות המרכזית, גופי ההסברה הממשלתיים וגופי השידור הציבוריים צריכים לפרסם מידע ברור, מובנה וקריא למכונה, בדומיינים רשמיים ומאומתים. ההמלצה נשענת על שתי הנחות שטרם נבדקו: האחת, קיימים פערי מידע בעברית ובערבית, דבר שבדיקת המודלים המוצעת לעיל אמורה לאשש או להפריך; השנייה, מידע רשמי אכן משפיע על תוצאות האחזור. לכן מדובר בצעד זול יחסית ובעל היגיון סביר, אך לא באמצעי נגד מוכח.

2. **שיתוף פעולה עם חברות המפתחות בינה מלאכותית.** על הגורם האחראי לפנות לחברות שמודליהן נמצאים בשימוש נרחב בישראל ולהציע להן שיתוף פעולה בבחינת תרחישי ניצול לרעה הקשורים לבחירות, ובהם מידע כוזב על אופן ההצבעה, התחזות למקורות רשמיים, יצירת דיסאינפורמציה משכנעת ואחזור תכנים של רשת פראבדה ושל דומיינים המזוהים עם CopyCop, וכן התנהגות המודלים בנושאים שבהם קיימים פערי מידע בזירה הישראלית. ישראל יכולה להביא לשיתוף הפעולה ידע שאין לחברות, ובעיקר היכרות עם הנרטיבים הרלוונטיים בעברית, בערבית וברוסית, ורשימות של דומיינים עוינים שזוהו.

מבין שלושת הרבדים שבהם תוכן עוין יכול לחדור למערכות בינה מלאכותית, רק ברובד האחזור תועד עד כה שימוש בהקשר של מערכות בחירות ממשיות. גם שם אין ראיה לכך שההרעלה שינתה את אופן ההצבעה, אך היעדר ראיה כזו תואם את אופייה של תקיפה שנועדה לחמוק מגילוי. האיום נבנה חודשים לפני מערכת הבחירות, נשאר זמין לאחזור גם אחריה ואינו תואם את גבולות המנדט של אף גוף קיים, ולכן המענה המקובל לכך בהגנה על בחירות אינו מספיק. בישראל טרם נבדק אם מודלי השפה חשופים לזיהום כזה בעברית ובערבית. לפיכך הצעד הראשון הוא מדידת היקף הזיהום, ותוצאותיה יקבעו את משקלם של שאר הצעדים.