בינה מלאכותית והגנת סייבר - פרספקטיבה של 4 שנים

מאת ד"ר יעקב רימר

הקדמה

במהלך שנת 2022 פרסמתי סדרת מאמרים בכותרת בינה מלאכותית והגנת סייבר - הייפ ומציאות. בסדרת המאמרים בחנתי לעומק את הפוטנציאל של שיטות מתקדמות בלמידת מכונה לקידום משמעותי של תחום הגנת הסייבר. מהפכת ה-Generative AI (מודלי שפה גדולים וסוכנים אוטונומיים) שהתפוצצה בשנים האחרונות (החל מסוף 2022) הביאה לפריצות דרך בתוך ה-AI שמאתגרות את הנחות היסוד מאז. לאור זאת, הסתקרנתי לבחון מה שרד את מבחן הזמן והיכן המציאות השתנתה.

רקע כללי

בסדרה שפרסמתי סקרתי מספר מאמרים שפורסמו על ידי Center for Security and Emerging Technology (CSET), גוף מחקרי באוניברסיטת ג'ורג'טאון. המאמרים עסקו בניתוח הקשר בין עולם ה-AI לעולם הסייבר הרחב (סייבר התקפי, הגנת סייבר ומבצעי השפעה בסייבר). הם התמקדו בשאלה האם שיטות למידת מכונה יסייעו להתמודד טוב יותר עם התקפות מתוחכמות יותר.

השורות התחתונות שלהם אותן ניתחתי בסדרה היו:

1. למידת מכונה יכולה לסייע למגינים לאתר התקפות סייבר בצורה מדויקת יותר. במקרים רבים לא מדובר בגישות חדשניות, אלא בהרחבה של אותם הדברים שכבר נעשים. כמובן, אסור לשכוח שעצם השימוש בלמידת מכונה מוסיף משטחי תקיפה נוספים.
2. בעולם הגנת הסייבר קיימות משימות שגרתיות ומייגעות רבות. למידת מכונה עשויה לסייע לבצע חלק מהן בצורה אוטומטית ובכך לפנות את הזמן והקשב של המגינים. בחלק מהתחומים עדיין נדרשות פריצות דרך משמעותיות אל מול השיטות שקיימות כיום.
3. כניסה של שיטות למידת מכונה נוספות ישנו את "שדה הקרב" של הגנת הסייבר. לעיתים לטובת התוקפים ולעיתים לטובת המגינים. אבל ללא פריצות דרך נוספות, לא צפוי ששיטות למידה מכונה ישנו באופן דרמטי את תחום הגנת הסייבר.

שלוש מסקנות אלו נדונו במפורט בסדרת המאמרים. בשתיים מהן צוין במפורש שיש צורך בפריצת דרך. וכאמור, לאור העובדה שבשנים האחרונות אכן התרחשה פריצת דרך משמעותית בתחום ה-Generative AI, החלטתי לבחון מחדש את שלוש המסקנות האלו.

מסקנה ראשונה

"למידת מכונה יכולה לסייע למגינים לאתר התקפות סייבר בצורה מדויקת יותר. במקרים רבים לא מדובר בגישות חדשניות, אלא בהרחבה של אותם הדברים שכבר נעשים."

האמירה הזאת עמדה ברובה היטב במבחן הזמן של ארבע השנים האחרונות (2022-2026). רוב היישומים של AI במוצרי הגנה גם כיום (כמו XDR, SIEM/SOAR) אינם מפתחים "היגיון הגנתי חדש". הם פשוט מבצעים את מה שאנליסט סייבר או מערכות חוקים עשו בעבר, רק במהירות ובקנה מידה גדולים בסדרי גודל. מערכות AI מסוגלות כעת לעבד הרבה יותר נתונים בו-זמנית, מכל מערכת קודמת. למשל, להצליב לוגים ולמצוא קשרים בין אירועים שונים ברשת. כלומר, ה-AI מרחיב את היכולת לנתח אירועי משתמשים, תחנות, רשת וענן ביחד - וכך להגיע להחלטה מדויקת יותר, מהר יותר. הוא גם מאפשר כתיבה (כיום אוטומטית) של חוקי זיהוי על בסיס דפוס תקיפה חדש שאותר.

עם זאת, כניסת מודלי השפה הגדולים (LLMs) והסוכנים האוטונומיים (AI Agents) הביאה איתה גם מספר גישות שניתן להגדיר כחדשניות באמת. בעבר, מערכות הגנה חיפשו חתימות ושינויים של דפוסים מוכרים כגון אנומליות סטטיסטיות. כיום, מודלי AI מסוגלים לייצר סיפור תקיפה שלם בשפה טבעית ובזמן אמת, כולל הסבר על ההיגיון של התוקף. זהו שינוי תפיסת הממשק בין אדם ומכונה בזמן אירוע סייבר. יתרה מכך, מערכות מתקדמות יכולות להציע (ואף לבצע במקרים מסוימים) תיקון של הקוד תוך כדי אירוע, כדי לחסום את התוקף מבלי להשבית את המערכת. דוגמה נוספת, בהתמודדות מול פישינג, מערכות AI מסוגלות לזהות ניסיון פישינג גם על ידי ניתוח סמנטי של הכוונה הפסיכולוגית של הטקסט. זהו סגנון חשיבה ניתוחי שונה מזה של הכלים ה"קלאסיים". אלו דוגמאות לשינויי גישה, גם אם בשוליים.

נעבור לסיפא של המסקנה הזאת: **"כמובן, אסור לשכוח שעצם השימוש בלמידת מכונה מוסיף משטחי תקיפה נוספים."** זה כמובן עדיין נכון. עצם השימוש במודלי AI בארגונים יצר משטחי תקיפה חדשים לחלוטין שהמגינים נאלצים להתמודד איתם: התקפות הרעלת נתונים (Data Poisoning), הזרקות פרומפטים (Prompt Injection), ועוד. השימוש הגובר במערכות AI מגביר כמובן את הסכנה הזאת. בנוסף, שימוש גובר במודלי AI עלול לגרום לדליפת מידע סודי או נכסים של החברה שמשתמשת בהם.

מסקנה שנייה

"בעולם הגנת הסייבר קיימות משימות שגרתיות ומייגעות רבות. למידת מכונה עשויה לסייע לבצע חלק מהן בצורה אוטומטית ובכך לפנות את הזמן והקשב של המגינים. בחלק מהתחומים עדיין נדרשות פריצות דרך משמעותיות אל מול השיטות שקיימות כיום."

כבר לפני 4 שנים ציינתי כי ברור שלמידת מכונה תסייע בביצוע אוטומטי של משימות מייגעות, כגון איסוף מודיעין על תוקפים ושיוך של תקיפות סייבר. תהליך זה הואץ מאוד בשנים האחרונות עם פריצת הדרך במערכות של מודלי שפה גדולים וסוכנים אוטונומיים. כלי AI משולבים כיום כאמור ב-SIEM/SOAR ומסייעים לאנליסטים ב-SOC לסכם אירועים, לכתוב שאילתות וחוקי זיהוי ולסנן רעשים (False Positives). אבל, פריצת הדרך הטכנולוגית הביאה לשינוי משמעותי יותר בחלק מהתחומים.

אמנם, החדשות הרעות הן שאתגרים שהצבעתי עליהם נשארו בעייתיים גם כיום. חברות הגנה אינן יכולות להרשות לעצמן אחוז גבוה של התרעות שווא מחשש לגרימת DoS (מניעת שירות) עצמי ללקוחות. ארגונים מעדיפים דטרמיניזם ויכולת הסבר (Explainability) על פני קופסה שחורה של AI שמחליטה לחסום תעבורת רשת קריטית ללא נימוק ברור. לכן מערכות חוקים עדיין שולטות בעולם ה-Network IDS. בנוסף, רובנו עדיין לא מעוניינים ב-Agent אוטונומי שיבצע שינויים ברשת ללא השגחה אנושית, שוב מחשש שיגרום ל-DoS.

אבל יש גם חדשות טובות. בכל הנוגע לניתוח טקסט, קוד וחולשות, המהפכה של מודלי ה-AI והסוכנים שינתה לחלוטין את הכללים והפכה משימות שנחשבו ב-2022 ל"אקדמיות" או שוליות לכלים מסחריים. בעבר, מציאת וסיווג באגים, ניתוח דוחות באגים, או כתיבת קלטים חכמים ל-Fuzzing דרשו השקעה אנושית גדולה עקב מגבלות של שיטות למידת המכונה הקלאסיות. פריצת הדרך הגנרטיבית אפשרה לקרוא ולפרש קוד וטקסט, ובכך לפרוץ את המחסום בתחומים של כתיבת וניתוח קוד והנדסה לאחור (Reverse Engineering).

כיום חברות כבר לא נדרשות לאסוף דוגמאות לבאגים משלהן כדי לאמן מודל למידת מכונה. מודלי שפה מסחריים וייעודיים לעולם הפיתוח מגיעים עם הבנה עמוקה של שפה טבעית וקוד. כלי אבטחת אפליקציות משתמשים במודלים האלו כדי לסרוק קוד באופן סטטי, או לקרוא פלטים של סורקי קוד מבוססי חוקים. המודלים מסוגלים להבין את הקשר הקוד, לאתר באגים ולהגדיר את חומרת הבאג ברמת דיוק גבוהה. ומודלי ה-AI של השנים האחרונות לא רק מסווגים באגים - הם גם מתקנים אותם באופן אוטומטי. כלי פיתוח מודרניים מזהים חולשת אבטחה כבר בזמן כתיבת הקוד על ידי המפתח, ומציגים לו הצעה לקוד מתוקן ומאובטח. וכמובן, המודלים האלו אף מסוגלים לכתוב את הקוד באופן מלא, אוטומטי ומאובטח. היכולת הזו למנוע חולשות לפני שהקוד מגיע ליישום הינה אחת מפריצות הדרך המשמעותיות ביותר, והתקווה לצמצום כמות החולשות בעולם.

בנוסף, מערכות סוכנים ומודלי LLM מודרניים מסוגלים גם לסייע בהנדסה לאחור (Reverse Engineering) של קוד בינארי (Compiled). הם מסוגלים "להבין" את כוונת התוכניתן המקורי, להסביר את פעולת הפונקציה, ולהוסיף שמות לפונקציות ולמשתנים, גם אם הם הוסרו במהלך הקומפילציה. הם יכולים גם לסמן לחוקר האם והיכן יש חשש לחולשות בקוד. בכך הם יכולים לקצר את זמן הניתוח של קבצי הרצה מימים, לדקות ספורות.

מודלי AI מסייעים מאוד גם בתהליך ניתוח קוד דינאמי. הם מסוגלים לכתוב אוטומטית קלטים עבור פאזרים. משימה שבעבר לקחה זמן רב של עבודה ידנית. וה-AI לא מייצר רק קלטים אקראיים. מודלים מודרניים שמסוגלים כאמור לנתח את הקוד ו"להבין" את הלוגיקה שלו, יכולים לייצר קלטים סמנטיים מתוחכמים שמיועדים להגיע לפינות הנסתרות ביותר של הקוד במהירות גבוהה פי כמה.

גם תחום בדיקות החדירות (Pentest) האוטומטיות עבר מהפכה. שילוב של LLM-based Agents יחד עם מערכות חוקים (מסוגים שונים) מאפשר לסוכני AI מודרניים להבין ארכיטקטורות רשת מורכבות, לקרוא פלטים של סורקי חולשות, ולתכנן שרשראות תקיפה (Lateral Movement) ברשתות ארגוניות גדולות בזמן אמת. פריצת הדרך הגיעה מהיכולת של מודלי שפה להבין קשרים ולוגיקה של מערכות ופרוטוקולים בצורה "אנושית". יתרה מזאת, המודלים מאומנים מראש על כמויות עצומות של קוד, מדריכי IT, תיעודי חולשות (CVEs) ומתודולוגיות תקיפה. כלומר, הם מגיעים לרשת הנבדקת עם "ידע רקע" נרחב. הם לא צריכים ללמוד את הרשת מאפס. הם פשוט מיישמים את הידע הכללי שלהם על המבנה הספציפי שהם מגלים, בדומה להאקר אנושי.

הסוכנים האלו מסוגלים גם לבצע תיקונים באופן אוטומטי לבעיות שנמצאו. אבל כפי שכתבתי לעיל, רובנו עדיין חוששים מ-Agent אוטונומי שיבצע שינויים ברשת ללא השגחה אנושית. הפתרון הוא מודל ה-HITL (Human-in-the-loop). לא מאפשרים לסוכנים לרוץ לבד לחלוטין ברשת, אלא משתמשים ב-AI כ"טייס משנה" שמייצר את תוכנית התקיפה/הבדיקה או התיקונים, מציג אותה למפעיל האנושי, ומבקש אישור לפני ביצוע פעולות רגישות או אגרסיביות. זה הפך את השימוש למעשי ובטוח יותר.

מסקנה שלישית

"כניסה של שיטות למידת מכונה נוספות ישנו את "שדה הקרב" של הגנת הסייבר. לעיתים לטובת התוקפים ולעיתים לטובת המגינים. אבל ללא פריצות דרך נוספות, לא צפוי ששיטות למידה מכונה ישנו באופן דרמטי את תחום הגנת הסייבר."

בסדרת הכתבות כתבתי שהמטרה של בינה מלאכותית אינה רק "לרוץ הכי מהר שאנחנו יכולים, כדי להישאר באותו המקום", סטייל עליסה בארץ המראות. המטרה היא לשפר את יכולות ההגנה אל מול

התוקפים. אבל בהגנת סייבר לא משחקים מול "הקיר". יש תוקפים בצד השני שמשתכללים לא פחות, ויש שיאמרו אפילו יותר מהמגינים. בשנים האחרונות, עם מהפכת מודלי השפה והסוכנים האוטונומיים, קצב המשחק הואץ בצורה חסרת תקדים. אבל הסטטוס קוו הדינמי בין תוקפים ומגינים נשמר. ארגונים שלא אימצו כלי בינה מלאכותית להגנה פשוט נשארו מאחור, אך אלו שאימצו אותם לא הפכו לחסינים לחלוטין לתקיפות סייבר. הם כנראה רק מצליחים להתמודד עם קצב ההתקפות המוגבר של התוקפים, שמשתמשים גם הם באותם הכלים בדיוק.

המציאות בשנים האחרונות מאששת שגרף תקיפות הסייבר העולמי, פריצות הכופר וגניבות המידע רק המשיך לעלות, מה שמוכיח ש-AI אינו "תרופת פלא" שמפחיתה את עצם קיום האיום בסייבר. כל מקום שבו יש שיפור לטובת המגינים, כמו מציאת חולשות אוטומטית שהוזכר לעיל, משמש מיד גם את התוקפים. כלי AIהמיועדים לסקירת קוד ואיתור באגים משמשים כיום האקרים למציאת חולשות באותה קלות שבה המגינים משתמשים בהם כדי לתקן אותן. ונזכיר שלתוקף מספיקה חולשה אחת. מגן חייב לתקן את כולן. יתרה מזאת, עבור המגן, לא די לתקן חולשות, צריך גם להטמיע את התיקונים. אמנם, AI מסייע מאוד לאחרונה גם בתיקון החולשות, אבל הטמעה היא לפעמים עניין מורכב בהרבה. וכפי שצוין לעיל, הגישה הרווחת אינה מאפשרת תיקון אוטומטי לחלוטין, אלא השגחה אנושית על התהליך. כלומר, אנחנו רואים האצה משמעותית בתהליכי ההגנה, אבל הם עדיין מורכבים ובד"כ מסורבלים יותר מהמאמץ שנדרש מהתוקף.

הזכרתי לעיל שמודלי AI מסייעים כיום טוב יותר במניעת פישינג. אבל הם גם מאיצים משמעותית את היכולת של כל אדם (ולא רק מומחה סייבר) לייצר תקיפות פישינג מתוחכם באופן סיטונאי. גרוע מכך, סוכני ה-AI מסוגלים להריץ תהליך פישינג דינמי מלא מול הנתקף, ללא מגע אדם. זה כולל איתור מטרות באופן אוטומטי, איסוף מידע מקדים לפני תקיפה (אוסינט), פתיחת פרופילים פיקטיביים ברשתות חברתיות, שליחה ותגובה של מיילים, ואפילו לענות לטלפון או להתחזות לאדם מסוים בוועידת Zoom. כלומר, מתקפות מתוחכמות שהיו שמורות בעבר אך ורק ל"דגים שמנים" (Whaling), מכיוון שדרשו מאמץ ותפעול אנושי ידני, הופכות להיות נפוצות. אז כן, יש חדשנות מסוימת במניעת פישינג. אבל במחיר של תיעוש דרמטי ביכולות של התוקף.

באופן דומה, המודלים הקלו לא רק על המגינים לכתוב תוכנה מאובטחת. הם הקלו מאוד גם על תוקפים זוטרים לכתוב פוגענים מתוחכמים, ואף לייצר באופן אוטומטי הרבה מאוד גרסאות שלהם (פולימורפיזם). בנוסף, שימוש ב-AI לכתיבת קוד דוחף לעיתים חברות תוכנה לשחרר פיצ'רים מהר יותר, או לכתוב קוד מהר יותר, לפעמים ללא מגע יד אדם. תהליכים אלו עלולים להגדיל דווקא את משטח התקיפה, כי למרות היתרונות של כתיבת קוד מאבטח באמצעות AI, לפעמים הקוד החדש עלול להיות עם חולשות בעצמו. ניתן לחשוב כמובן על הוספת מנגנוני בקרה, שוב באמצעות AI, אבל האם כולם יעשו זאת?

ויש עוד נושא שלא השתנה משמעותית ב-4 השנים החולפות. חברות סייבר עדיין "מגמגמות" כאשר הן נדרשות להציג מדדי הצלחה ברורים ואמינים ליעילות מערכות ההגנה שלהן, גם כאשר הן משלבות בהן AI. זאת למרות שמערכות ה-AI המודרניות הביאו בשורה גם ביכולת ליצור מאגרי בדיקה או סימולציות של רשתות ותעבורת רשת סינטטית.

סיכום

ארכיטקטורת ה-Transformers ומודלי השפה הגדולים הביאו לפריצת דרך משמעותית ביכולת של מערכות אוטומטיות להבין הקשר, שפה וקוד, ובכך הן שינו את שדה הקרב בסייבר. הטכנולוגיה לא נשארה רק בגדר "שיפור הדרגתי של השיטות הקיימות". עם זאת, בינה מלאכותית עדיין איננה פתרון קסם שסיים את תקיפות הסייבר. מרוץ החימוש בין התוקפים למגינים נותר בעינו, רק בקצב מואץ משמעותית. שני הצדדים קבלו "מטוסי סילון" במקום מכוניות, מה שהפך את שדה הקרב בסייבר למהיר, אוטונומי וקטלני בהרבה.

אז מה השורה התחתונה? בדיוק כמו לפני 4 שנים. לאור מצב שדה-הקרב של הסייבר, אין מנוס אלא להמשיך ולפתח יישומים חדשים של AI לטובת הגנת סייבר. כן, גם אם בפועל אלו רק יאפשרו לנו "לרוץ הכי מהר שאנחנו יכולים, כדי להישאר באותו המקום".

על הכותב

ד"ר יעקב רימר הוא יועץ בכיר ומרצה בנושאי סייבר, בינה מלאכותית וביולוגיה. יש לו תואר שני בלמידת מכונה ודוקטורט באימונולוגיה, שניהם ממכון ויצמן למדע. הוא עוסק ביעוץ למשרדי ממשלה ולחברות היי-טק. בעבר שימש בתפקידים בכירים בהיי-טק ובמשרד ראש הממשלה. בנוסף, הוא מלמד במסגרת תוכניות לתואר שני באוניברסיטת תל-אביב את הקורסים: "מבוא לטכנולוגיות סייבר", "בינה מלאכותית ויישומיה לביטחון" ו"בינה מלאכותית בעידן הסייבר".

מייל לתגובות: MrBigDataThemarker@gmail.com