06

איומי סייבר על מערכות ביטחוניות בלתי מאוישות בעידן הבינה המלאכותית

הצעה למסגרת ביקורת מעודכנת להערכה ולבקרה

איתן דהן, משה מורנו, אבישי יהושע ואורן בר

תא"ל (במיל') איתן דהן הוא ראש החטיבה לביקורת מערכת הביטחון במשרד מבקר המדינה ונציב תלונות הציבור.

סא"ל (במיל') משה מורנו הוא סגן מנהל האגף לביקורת צה"ל, כשירות ומוכנות בחטיבה לביקורת מערכת הביטחון שבמשרד מבקר המדינה ונציב תלונות הציבור.

ד"ר אבישי יהושע הוא מנהל ביקורת באגף לביקורת העורף האזרחי בחירום, גבולות ואיו"ש בחטיבה לביקורת מערכת הביטחון שבמשרד מבקר המדינה ונציב תלונות הציבור.

אל"ם (במיל') אורן בר הוא מנהל ביקורת באגף לביקורת משרד הביטחון ומשאבי מערכת הביטחון בחטיבה לביקורת מערכת הביטחון שבמשרד מבקר המדינה ונציב תלונות הציבור.

המאמר בקצרה

- מערכות בינה מלאכותית (להלן - AI) משולבות כיום במערכות ביטחוניות בלתי מאוישות, לרבות בתהליכי מודיעין, שליטה ובקרה וזיהוי מטרות, ובכך הן מרחיבות את משטח התקיפה של המערכת מעבר לתשתיות הקלאסיות (חומרה ותוכנה) - אל הנתונים, המודלים ושרשרת האספקה של רכיבי AI.

- מאמר זה עוסק בשאלה כיצד יש לעדכן את עבודת הביקורת בתחום הגנת הסייבר לאור העובדה שרכיב הליבה של המערכות הביטחוניות הבלתי מאוישות - מנגנון קבלת ההחלטות - נשען על מודלים ונתונים דינמיים ומבוסס על AI.

- המאמר ממפה וקטורי תקיפה מרכזיים ומראה כיצד כשלים אלה עלולים להתגלגל מכשל טכני לכשל מבצעי ולכשל אחריותיות.

- בהמשך המאמר מוצעת מסגרת של הבטחת מוכנות לביקורת (Audit-Ready Assurance) המותאמת לביקורת המדינה, המבוססת על עקרונות הבטחת משימה וחוסן תפעולי: קריטריונים, מדדי ביצוע וסיכון, סט ראיות מזערי, לצד תוכנית בדיקות ודגימות.

- תרומת המאמר היא המעבר מביקורת "ציות לבקרות" לביקורת על "יכולת הוכחה" לשם "הבטחת משימה" - שבה ממצא, מסקנה והמלצה נשענים על ראיות ניתנות לשחזור לאורך מחזור חיי המערכת.

- מילות מפתח: AI (בינה מלאכותית), מערכות בלתי מאוישות (UxV), הבטחת משימה (Mission Assurance), סייבר, ביקורת המדינה, סחיפת נתונים, אחריותיות (Accountability).

מבוא | מפת האיומים ומשטחי התקיפה במערכות ביטחוניות בלתי מאוישות | השפעת איומי סייבר ייחודיים בעידן ה-AI | מודל הבטחת משימה וחוסן מבצעי במערכות ביטחוניות בלתי מאוישות בעידן ה-AI | מסגרת ביקורת להערכה ולבקרה המכוונת להבטחת משימה (Audit-Ready Assurance Framework) בעידן ה-AI | יישום המסגרת: מתודולוגיית ביקורת עדכנית והמלצות לשיפור מתמשך | סיכום

מבוא

בעשורים האחרונים הצטיידו גופי הביטחון למיניהם במערכות ביטחוניות בלתי מאוישות מתקדמות לשימוש מבצעי (מערכות UxV)[1]. מערכות ביטחוניות בלתי מאוישות הן מערכות אוטונומיות הפועלות ללא נוכחות אנושית ישירה, והן נשענות במידה הולכת וגדלה על מודלים של למידת מכונה (ML)[2] לצורך ניווט, זיהוי מטרות וקבלת החלטות בזמן אמת. מערכות כאלה משמשות למטרות שונות, ובהן איסוף מודיעין, תקיפה מהאוויר, שליטה ובקרה. בעידן הנוכחי מערכות בלתי מאוישות הן רכיב ביטחוני מרכזי שבו מצטלבים חדשנות טכנולוגית, רציפות תפקודית, צמצום הסיכון לחיי אדם ואחריותיות ארגונית. לשימוש במערכות אלה יתרונות רבים לצד סיכונים מתפתחים. מאמר זה עוסק בשאלה כיצד יש לעדכן את עבודת הביקורת בתחום הגנת הסייבר לאור העובדה שרכיב הליבה של מערכת ביטחונית בלתי מאוישת - מנגנון קבלת ההחלטות - נשען על מודלים ונתונים דינמיים. המאמר משלב סקירת מסגרות תקינה וניהול סיכוני AI, מיפוי וקטורי תקיפה מרכזיים בסביבת מערכות ביטחוניות בלתי מאוישות וגזירת קריטריונים ראייתיים לביקורת. הטענה העיקרית היא כי ביקורת בנושא הגנה על מערכות ביטחוניות בלתי מאוישות מפני איומי סייבר המתמקדת בבקרות הגנת סייבר מסורתיות בלבד – סודיות (Confidentiality), אמינות/שלמות (Integrity), זמינות (Availability)[3] - אינה מספקת, משום שהיא אינה בוחנת שליטה ראייתית בנתונים, במודלי AI ובשרשרת האספקה הדיגיטלית של קבלני המשנה המתבססים על מרכיבי AI, ולכן אינה מאפשרת להעריך יכולת עמידה במשימה ואחריותיות.

המאמר ינתח את מפת האיומים החדשה, יבחן את סיכוני ה-AI הייחודיים המשיקים לסיכוני סייבר "קלאסי", יציג מודל להבטחת משימה (Mission Assurance) ויציע מסגרת של הבטחת מוכנות לביקורת (Audit-Ready Assurance Framework) המותאמת לאתגרים הדינמיים של שדה הקרב העתידי.

מפת האיומים ומשטחי התקיפה במערכות ביטחוניות בלתי מאוישות

מערכות ביטחוניות בלתי מאוישות בעת הנוכחית אינן עוד יחידות קצה מבודדות, אלא "מערכות מורכבות של מערכות" (System of Systems) המקושרות ברשתות נתונים הדוקות ופועלות בסביבה מבוזרת המרחיבה את היקף משטחי התקיפה. ניתוח מעמיק של משטחי התקיפה מעלה כי כל שלב במחזור החיים של המערכת - החל בפיתוח, דרך ההטמעה וכלה בעדכוני הקושחה[4] (Firmware) - הוא נקודת תורפה אסטרטגית.

כדי לאפשר ביקורת אפקטיבית בתחום הסייבר בעידן ה-AI מוצעת מפת איומים ומשטחי תקיפה פוטנציאליים בראייה רב-שכבתית: (א) שכבת תשתיות (מערכות מחשוב כדוגמת שרתים) ותקשורת נתונים; (ב) שכבת נתונים, חיישנים ומערכות ניווט; (ג) שכבת מודל והסקה - הרצת אלגוריתמי AI ברכיבי קצה; (ד) שכבת שרשרת אספקה דיגיטלית; (ה) שכבת תפעול (מערכות שליטה קרקעיות) וממשקי אדם-מכונה. לכל שכבה מוגדרים וקטורי תקיפה אופייניים, מנגנון הפגיעה והשלכה משימתית, ובצמוד להם - סט ראיות מזערי המאפשר למבקר להוכיח שליטה בסיכון (כפי שיפורט בהמשך).

להלן הרחבה על מיפוי האיומים ומשטחי התקיפה הפוטנציאליים:

1. **שכבת תשתיות ותקשורת נתונים:** ערוצי התקשורת (Data Links) בין הכלי ובין המפעיל האנושי או תחנת הבקרה הם יעד ראשון במעלה לשיבוש. תוקף מתוחכם עלול לבצע "הזרקת פקודות" (Command Injection) שמוסווות כתקשורת לגיטימית, המנצלת את הגישה ליכולות ה-AI ככלי תומך החלטה בידי המפעיל האנושי כדי להשתלט על הכלי מרחוק.
2. **שכבת נתונים, חיישנים ומערכות ניווט**[5]: מערכות ביטחוניות בלתי מאוישות חשופות להפרעות ניווט על

---

1 "X" - Aerial, Ground, Surface UxV - Underwater Unmanned "X" Vehicle;

2 Machine Learning; ML

3 המשולש של אבטחת המידע (CIA Triad) הוא המודל הבסיסי והחשוב ביותר בעולם הגנת הסייבר. שלושת המושגים הללו - סודיות, אמינות/שלמות, זמינות - מייצגים את היעדים העיקריים שכל מערכת אבטחה שואפת להשיג כדי להגן על מידע ומערכות. Jennifer Cawthra et al., Data Integrity: Detecting and Responding to Ransomware and other Destructive Events (NIST Special Publication 1800-26A, December 2020), https://doi.org/10.6028/NIST.SP.1800-26.

4 קושחה היא תוכנית בסיסית הצרובה על רכיב החומרה ומשמשת כעין "מוח" שלו. תפקידה הוא לנהל את פעולתו, לתרגם פקודות ולגשר בין החומרה לרכיבי התוכנה הגבוהים יותר.

5 Mark L. Psiaki & Todd E. Humphreys, GNSS Spoofing and Detection, 104|(6) Proceedings of the IEEE 1258 (2016).

פי מערכות ניווט גלובליות בחלל (GNSS[6]) מסוג חסימה (Jamming), וחשוב וחמור מכך - הטעיה (Spoofing). היכרות של תוקף עם חולשות מודלי AI המוטמעים ברכיבי קצה המשרתים את המערכות הבלתי מאוישות (Edge AI[7]) עלולה לאפשר לו לשדר קלט שגוי שמודל ה-AI אינו רגיש אליו ולפגוע בהבטחת המשימה. הטעיית ניווט יכולה לשבש, למשל, את יכולת ה-Edge AI של הכלי לחשב מסלול בטוח, מה שמוביל להתרסקות או לסטייה של המשימה לכיוון אזורים מאוכלסים.

3. **שכבת מודל והסקה - הרצת אלגוריתמי AI ברכיבי קצה (Edge):** בניגוד למערכות ענן, רכיבי קצה במערכות ביטחוניות בלתי מאוישות נדרשים להריץ אלגוריתמי AI בזמן אמת על גבי המערכת האוטונומית עצמה. תוקפים יכולים לנצל תופעה זו לביצוע תקיפות "מניעת שירות" אלגוריתמיות[8], המעמיסות על המעבד עד להשבתת יכולת קבלת ההחלטות של הכלי.

4. **שכבת שרשרת האספקה הדיגיטלית:** זוהי אולי נקודת התורפה הקריטית ביותר. התלות בספקים חיצוניים ובעדכונים שוטפים של תוכנה וקושחה לשם עדכון יכולות ה-AI מייצרת סיכון של החדרת "דלת אחורית" (Backdoor) כבר בשלב הפיתוח[9].

5. **שכבת תפעול (מערכות שליטה קרקעיות, GCS)[10] וממשקי אדם-מכונה:** מערכת השליטה הקרקעית היא הצומת המרכזי של המידע. חדירת נוזקה למערכת זו עלולה לא רק להשבית את המערכת, אלא גם לאפשר מניפולציה של הנתונים המוצגים למפעיל האנושי ולהוביל לקבלת החלטות מבצעיות שגויות על בסיס מידע כוזב.

השפעת איומי סייבר ייחודיים בעידן ה-AI

איומי סייבר בעידן ה-AI נבדלים מאיומי סייבר קלאסיים, שכן הם מכוונים ללוגיקה של הלמידה וההסקה, ולא רק לניצול חולשות קוד. שלוש משפחות מרכזיות של איומי סייבר יש בעידן הזה: (1) "הרעלת נתונים" בשלב האימון והעדכון של מודל ה-AI (הן בשלב הפיתוח והן בתפעול השוטף של המערכת), המשנה את גבולות ההחלטה של מודל ה-AI (יפורט בהמשך); (2) מניפולציית קלט למודלים של AI (תקיפות קלט לעומתיות בזמן אמת - המטות סיווג או החלטה באמצעות שינוי זעיר בקלט); (3) סחיפת מודלים ונתונים[11] על ידי שינוי הסביבה המבצעית לנתונים שהמודל לא אומן עליהם. מבחינת הביקורת המשמעות היא שהערכת עמידות של מודל ה-AI מחייבת ראיות לשליטה בנתונים ולתיעוד תהליך שינוי מודל, לרבות בדיקות חוסן מתועדות.

להלן פירוט של סיכוני הסייבר העיקריים העלולים להשפיע על הבטחת המשימה הביטחונית בעידן ה-AI במערכות ביטחוניות בלתי מאוישות המשתמשות ב-AI:

1. **"הרעלת נתונים" (Poisoning Data)[12] בשלב הפיתוח או העדכון של מודל ה-AI:** זהו שיבוש מכוון של שלב האימון והעדכון של המודל. לדוגמה, תוקף מחדיר נתונים מוטים למאגר המידע של המערכת הביטחונית הבלתי מאוישת במהלך פיתוחו או בזמן עדכונו במשך מחזור חייו של המאגר. בעקבות זאת המודל מפתח "עיוורון" מובנה או הטיות ספציפיות המופעלות רק בתנאים מסוימים. בביקורת יש לבחון את "ניקיון" הנתונים המשמשים לאימון מודל ה-AI במערכות ביטחוניות בלתי מאוישות.

---

6 Global Navigation Satelite-based Systems

7 מודל AI המוטמע ברכיב המותקן בכלי האוטונומי, בשונה מיכולת הנמצאת בענן או בתחנת שליטה קרקעית.

8 Jelena Mirkovic & Peter Reiher, A Taxonomy of DDoS Attack and DDoS Defense Mechanisms, 34(2) ACM SIGCOMM Computer Communication Rev. 39 (2004)

9 Tianyu Gu, Brendan Dolan-Gavitt, & Siddharth Garg, BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, ARXIV (August 2017), arXiv:1708.06733 [cs.CR]; Bolun Wang et al., Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks, 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2019, pp. 707-723, doi: 10.1109/SP.2019.00031.

10 Ground Control Systems

11 תופעה שבה המציאות והנתונים בעולם משתנים עם הזמן, מה שגורם למודל ה- AI - שהוכשר על נתונים ישנים או קבועים - להיות פחות מדויק ופחות רלוונטי למשתמשים.

12 Battista Biggio, Blaine Nelson, & Pavel Laskov, Poisoning Attacks Against Support Vector Machines, ICML'12: Proceedings of the 29th International Coference on International Conference on Machine Learning 1467 (2012); Jacob Steinhardt, Pang Wei Koh, & Percy Liang, Certified Defenses for Data Poisoning Attacks, Proceedings of the 31st International Conference on Neural Information Processing Systems 3520 (2017).

2. **מניפולציית קלט למודלים של (Model Manipulation) AI:**[13] הטיית פלט המערכת בזמן אמת באמצעות שינויים מזעריים בקלט, שאינם נראים לעין אנושית (Adversarial Examples). לדוגמה, עיבוד מידע חזותי (תמונות, תצלומי אוויר וכיו"ב) שגוי בשלב הקלט המאפשר לסווג כלי רכב צבאיים ככלי רכב אזרחיים עלול לגרום למודל AI לסווג רכב צבאי כרכב אזרחי ובכך למנוע מהמערכת לבצע את משימתה המבצעית.

3. **סחיפת מודלים ונתונים (Data Drift):** זהו כשל תפקודי הנובע משינוי קיצוני בסביבה המבצעית שמודל ה-AI לא אומן עליה. איומי סייבר יכולים לנצל מצבי "סחיפה" כדי להוביל את המערכת הביטחונית הבלתי מאוישת לקריסה בתנאי קצה. דוגמה לכך היא הצבת חוסמי/מטעי GNSS[14] במרחב הפעולה המונעים מפלטפורמות UxV לנווט בצורה מדויקת.

להלן שני מקרי בוחן להדגמת סיכון סייבר המבוסס על מניפולציית קלט למודלים של AI:

1. אמצעי נשק קרקעי בלתי מאויש (כדוגמת מערכת "רואה-יורה" שבשימוש כוחות הביטחון) המוצב להגנה על מתקן ממשלתי רגיש, המשתמש ב-AI לזיהוי מטרות, חווה מתקפת "מניפולציית קלט למודל":
   א. התוקף מציב דפוס חזותי על הקרקע, שאינו נראה לעין האנושית, הגורם לאלגוריתם הראייה הממוחשבת של מודל ה-AI לסווג מטרה אזרחית כאובייקט עוין, או להפך.
   ב. בעקבות זאת מערכת הנשק עלולה להמליץ על תקיפה שגויה שתוביל לפגיעה בחפים מפשע. כשל כזה אינו רק טכני; הוא כשל ב"מינהל התקין" ובאחריותיות של המדינה כלפי החברה.

תמונה 1: מערכת "רואה-יורה" להגנת הגבול

המקור: הופק באמצעות Chat-GPT ו-Gemini

13 Aleksander Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings (2018); Fracesco Croce & Matthias Hein, AutoAttack, ARXIV 2 (March 2020), arXiv:2003.01690 [cs.LG].

14 אמצעי המוצב על הקרקע ומונע או משבש קליטה של אותות ניווט לווייניים ממערכות לווייני ניווט (Global Navigation Satellites Systems).

ג. התוקף האדפטיבי (Adaptive Adversary) ינצל בדיוק את אותן חולשות אלגוריתמיות כדי לעקוף את מנגנוני ההגנה הסטטיים של המערכת.

2. **תקיפת סייבר על נחיל רחפנים בניסוי סימולציה:**

א. נחיל רחפנים אוטונומי להגנה על תחנת כוח, המצויד במערכת זיהוי מטרות סיגינטית[15] (על פי תדר שידור) מבוססת AI, הוטעה באמצעות "משדרי הטעיה" שהותקנו על כלי רכב ידידותיים לכאורה.

ב. השינוי המזערי בתבנית השידור, שאינו נקלט בחיישן הקולט המותקן על כל אחד מהרחפנים בנחיל, גרם לאלגוריתם לסווג את הרכב "מטרה צבאית עוינת".

ג. כשל זה ממחיש כיצד חולשה בלוגיקת ה-AI הופכת לכשל מבצעי קריטי, גם כאשר "חומת האש" ותשתית התקשורת נותרו חסינות לחלוטין.

מודל הבטחת משימה וחוסן מבצעי במערכות ביטחוניות בלתי מאוישות בעידן ה-AI

בהקשר של מערכות ביטחוניות בלתי מאוישות "הבטחת משימה" מורכבת מסט יכולות מדידות: המשך ביצוע משימה בתנאי שיבוש, ירידה מבוקרת למצב בטוח (fail-safe) (להלן - אל-כשל) וזמן התאוששות מוגדר.

תמונה 2: נחיל רחפנים המגן על תחנת כוח במרחב אזרחי

המקור: הופק באמצעות Chat-GPT ו-Gemini

15 סיגינט (Signal Intelligence; SIGINT) - מודיעין אותות, ענף בעולם המודיעין העוסק באיסוף מידע על ידי יירוט ופיענוח של אותות אלקטרוניים והפקת מידע מודיעיני מתוכם (אתר המרכז למורשת המודיעין).

מילוי מלא ותקין של כלל מרכיבי המשימה בעידן השפעת ה-AI מחייב ביתר שאת מעבר מתפיסת "הגנה סייבר היקפית" ו"הגנה על המערכת" לתפיסת הבטחת משימה וחוסן מבצעי (Mission Assurance & Operational Resilience) - היכולת להשלים את המשימה גם בתנאי שיבוש או תקיפה מוצלחת. לכן הביקורת נדרשת לבדוק לא רק "אם יש בקרות", אלא אם הארגון מסוגל להוכיח - בראיות ובמדדים - עמידה ביעדי שירות והמשך תפקוד בתרחישי שיבוש, לרבות באמצעות תרגילי שרידות ותיק אירועים.

עלינו להניח כי המערכת הביטחונית הבלתי מאוישת תהיה נתונה לשיבוש או לחדירה, ולכן המיקוד בפיתוח המערכת, בהוכחת יעילותה ובתפעולה השוטף עובר למיקוד ביכולתה להשלים את המשימה חרף הפגיעה.

יש לנהל את הסיכונים למערכת הביטחונית הבלתי מאוישת לאורך כל שלבי מחזור החיים שלה:

1. **פיתוח:** הטמעת בקרות (Security by Design) ואימון אדברסרי של מודל ה-AI (אימון הכולל מימוש סיכוני סייבר אפשריים ודרך להתמודד עימם).
2. **הטמעה:** בדיקות קבלה (Acceptance Testing) לחוסן מודל ה-AI[16].
3. **תפעול:** ניטור רציף (Continuous Monitoring) של החלטות מודל ה-AI במערכת הביטחונית הבלתי מאוישת.
4. **עדכונים:** תיקוף של מודל ה-AI לאחר כל עדכון נתונים או קושחה.

נוסף על כך, לשם הבטחת המשימה נדרש לקיים את "מעגל החוסן המבצעי" המורכב מארבעה שלבים קריטיים:

1. **מניעה:** צמצום משטחי התקיפה, אימות חתימות דיגיטליות של עדכוני קושחה והבטחת "היגיינת נתונים" בשלב הפיתוח של מודל ה-AI.
2. **זיהוי:** ניטור מתמיד וחכם של חריגות סטטיסטיות בהחלטות מודל ה-AI בהשוואה ללוגיקה המצופה.
3. **תגובה:** הפעלת מנגנוני אל-כשל המבטיחים כי המערכת הביטחונית הבלתי מאוישת המבוססת על מודל AI לא תפעל באופן המסכן את הסביבה.
4. **התאוששות:** יכולת חזרה לתפקוד, עדכון מודלי ה-AI בשטח (Over-the-air updates) ויכולת תחקור מלאה של האירוע.

תרשים א: מעגל החוסן המבצעי

המקור: הופק באמצעות Chat-GPT ו-Gemini

16 בדיקות קבלה (Acceptance Testing) לחוסן המודל הן השלב הקריטי שבו מוודאים שמודל ה-AI מדויק לא רק בשלב התכנון, אלא עמיד בפני תרחישי קיצון, מניפולציות ורעשים בעולם האמיתי, לפני הטמעתו המלאה במערכת מבצעית או עסקית.

תהליך הבטחת משימה במערכות ביטחוניות בלתי מאוישות

התהליך מתחיל בשלב איסוף הנתונים מחיישני הכלי (מצלמות, מכ"ם, GPS). הנתונים מוזרמים אל מנוע ה- AI המרכזי המבצע הסקה (Inference) לשם קבלת החלטה. לאחר קליטת הנתונים מחיישני הכלי וקבלת החלטה על ידי מודל ה-AI, מופעל מנוע ניטור חוץ-ערוצי (Out-of-Monitoring band) המשווה בין החלטת ה-AI ובין מגבלות בטיחות קשיחות. אם מזוהה חריגה (למשל: זיהוי מטרה בתוך שטח אסור) המערכת הביטחונית הבלתי מאוישת עוברת למנגנון אימות משני.

עם החשד לשיבוש סייבר מתאמת המערכת מפעילה מנגנון אל-כשל (safe-Fail) דינמי: ניתוק ה-AI, מעבר למצב אוטונומי בסיסי המבוסס על כללים דטרמיניסטיים (Hard-coded) ושידור התרעה דחופה למפעיל האנושי. רצף זה מבטיח את הרציפות התפקודית של המשימה המבצעית גם בתנאי "אש דיגיטלית", כלומר מתקפת סייבר מתקדמת על מודל ה-AI השזור במערכת.

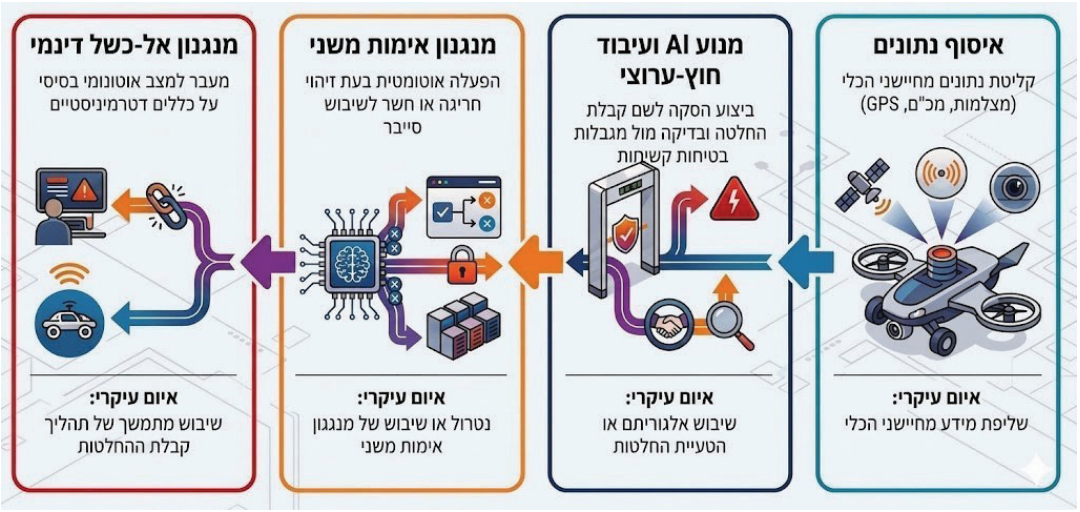
מסגרת ביקורת להערכה ולבקרה המכוונת להבטחת מוכנות לביקורת בעידן ה-AI (Audit-Ready Assurance Framework)

מיפוי תקנים בין-לאומיים וישראליים לצורך ביקורת הגנה מפני איומי סייבר במערכות ביטחוניות בלתי מאוישות

התקינה לביקורת ולהערכה של בקרות בתחום. אבטחת מידע והגנת סייבר היא רחבה ומעמיקה. בעידן של מערכות ביטחוניות בלתי מאוישות מבוססות AI נדרש להרחיב את נקודת המבט כך שלצד בקרות הגנת סייבר קלאסיות יהיו גם בקרות שיבחנו גם שלמות נתונים, שינויי מודל ה-AI, חוסן והטיה וכן שרשרת אספקה של רכיבי AI.

לצורך יצירת מסגרת ביקורת עדכנית מוצעת חלוקה פונקציונלית של התקנים לשלוש קבוצות, שתאפשר לבנות מסגרת Ready-Audit שאינה מחליפה תקינה קיימת, אלא מוסיפה שכבת ראיות וקריטריונים הממוקדים בנתונים ובמודל שרשרת אספקה דיגיטלית: (א) תקנים לניהול סיכוני סייבר ובקרות; (ב) תקנים ייעודיים לניהול סיכוני AI; (ג) תקנים למתודולוגיות לעבודת ביקורת.

תרשים ב: תהליך הבטחת משימה



המקור: הופק באמצעות Chat-GPT ו-Gemini

בנספח למאמר רוכזו תקנים בין-לאומיים עיקריים המקובלים להתוויית המתודולוגיה לביקורות בתחום אבטחת המידע והסייבר בארגונים וכמה תקנים העוסקים במסגרת ייעודית לניהול סיכוני AI ותקינה ברמה הלאומית ובמשרד מבקר המדינה. להלן עיקרם:

1. **מסגרות ייעודיות לניהול סיכוני AI**

א. **NIST AI RMF[17] 1.0 (NIST AI 100-1) (2023):** מסגרת ניהול סיכוני AI לפי Manage/Measure/Map/Govern; מגדירה מאפייני אמון (Trustworthiness) וניהול סיכונים במודלים ובנתונים לאורך מחזור חיים.

ב. **MITRE[18], Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS):** מיפוי של כל הטקטיקות והטכניקות שתוקפים עשויים להפעיל לאורך כל מחזור החיים של מערכת AI - החל בשלב איסוף הנתונים וכלה בשלב הפעולה (Inference).

ג. **ISA[19]/IEC 62443 (IACS[20]/OT[21]):** סדרת תקנים לאבטחת מערכות בקרה ותעשייה או טכנולוגיות הפעלה (IACS/OT) הרלוונטיים במיוחד לפלטפורמות אוטונומיות, לרכיבי קצה ולשילוב OT-IT במערכות Physical-Cyber.

2. **מתודולוגיות לעבודת ביקורת**

א. **מערך הסייבר הלאומי - תורת ההגנה בסייבר לארגון (מסמך תפיסתי-יישומי):** אחריות ההנהלה, הגנה לפי פוטנציאל נזק, מניעה-זיהוי-תגובה וחזרה לשגרה, בסיס לחיבור בין הנחיה לביקורת.

ב. **משרד מבקר המדינה:**

1) **תפיסת המשרד בנוגע לביקורת טכנולוגיות מידע והגנת סייבר:** המסמך מציג את התפיסה האסטרטגית של משרד מבקר המדינה בנוגע לביקורת טכנולוגיות מידע והגנת הסייבר. הוא מפרט את המבנה הארגוני, הכולל הקמת אגפים ייעודיים לצד הכשרת כוח אדם מקצועי בכל חטיבות הביקורת, להתמודדות עם איומים דיגיטליים. על פי התפיסה, מדיניות המשרד שואפת לחזק את הפיקוח על הגופים המבוקרים ועל תשתיות לאומיות קריטיות, תוך כדי שימוש בכלים מתקדמים כמו מבדקי חדירה.

2) **מתודולוגיה לביקורת אבטחת מידע, הגנת הסייבר והגנת הפרטיות:** המסמך מציג את מטרות הביקורת (שמירה על סודיות-שלמות-זמינות ואבטחה של מידע אישי בהתאם לחוק ולתקנות) וקובע דגשים ארגוניים לעבודת צוות הביקורת (הכשרה ייעודית; תיאום עם האגף לביקורת אבטחת המידע והגנת הסייבר במשרד; והתייחסות מחמירה לסיווג וחיסיון). המסמך מפרט גם את הסיכונים העיקריים ואת הבקרות הנגזרות מהם, מתאר את תהליך הביקורת משלב בחירת הנושא דרך זיהוי הגופים המבוקרים וקביעת נורמות הביקורת, ומפרט נושאי בדיקה שכיחים (ממשל סייבר, תוכניות עבודה ומשאבים, מדיניות ונהלים, ניהול סיכונים, אבטחה פיזית/לוגית, פיתוח מאובטח, המשכיות עסקית, שרשרת אספקה, זיהוי אירועים וטיפול בהם והגורם האנושי). עוד הוא מספק הנחיות לכתיבת דוח ממוקד ועדכני ואפשרות לשילוב מבדקי חדירה/חוסן והסתייעות ביועצים. הוא גם מסביר את מפת האסדרה בישראל (מערך הסייבר, שב"כ, יה"ב/מערך הדיגיטל, יחידות מגזריות והרשות להגנת הפרטיות) ואת בסיסי הסמכות והחובות בדין, לרבות עדכונים בהגנת הפרטיות.

עדכון התקינה הקיימת

התקנים הקלאסיים נבנו לעולם של "תוכנה + רשת + שרתים". במערכות ביטחוניות בלתי מאוישות מבוססות AI יש עוד שלושה רכיבים קריטיים שהתקנים בדרך כלל לא נדרשו להם מספיק: הנתונים, מודל ה-AI והשרשרת הדיגיטלית שמספקת אותם.

לנוכח זאת מוצע העיקרון שלהלן: כשמערכת ביטחונית בלתי מאוישת נשענת על נתוני AI ועל מודל AI, הביקורת נדרשת להיעזר בכלי ניתוח נתונים שיאפשרו למבקר

---

17 Risk Management Framework; RMF

18 MITRE - ארגון אמריקאי המנהל מרכזי מחקר ופיתוח במימון ממשלתי ללא כוונת רווח.

19 International Society of Automation; ISA

20 Industrial Automation and Control Systems; IACS

21 Operational Technology; OT

לזהות חריגות ולהסביר את ההחלטות שהתקבלו על ידי מודל ה-AI. בהיבט זה ראוי לנסח דרישת סף: יכולת שחזור (Reproducibility) על דגימה מצומצמת ותיעוד ברור של תהליך שינוי נתוני AI/מודל AI.

על כן העדכונים המוצעים להלן הם בגדר תוספות לתקינה, ולא החלפה שלה:

1. יש להוסיף לתקינה "שכבת נתונים" מחייבת: מקור, איכות, תיוג וגרסאות (ראו פירוט להלן). תקני אבטחת מידע מסורתיים מתמקדים ב"הגנה על מידע" (הרשאות, הצפנה, גיבוי), אבל פחות ב"נכונות המידע, עקיבותו וגיבויו בעובדות מקוריות" - שאלה קריטית למערכות ביטחוניות בלתי מאוישות מבוססות AI, שכן מודל ה-AI לומד את הנתונים ומתפקד לפיהם. אם הנתונים במערכת ביטחונית בלתי מאוישת מבוססת AI שגויים, המערכת "תבין" את העולם בצורה שגויה. לפיכך מומלץ לעדכן בתקינה את התוספות הבאות:
   א. מקור - תיעוד מקור הנתונים (Data Provenance): לתעד לכל מאגר נתונים שמשמש לאימון/הפעלה מה מקורו, מהו מועד יצירת המאגר ומועדי עדכונו, מי אסף, באילו תנאים ומה מותר/אסור לעשות בו.
   ב. איכות - בחינת איכות (Data Quality): סט של בדיקות איכות, למשל אם חסרים נתונים, אם יש חריגות, כפילויות או הטיות.
   ג. תיוג - בקרה על תיוג (Labeling) מבוקר: אם יש "תיוג", למשל סימון אובייקטים בתמונה, מוכרח להיות תהליך מבוקר: מי מתייג, מי מאמת, ומה שיעור השגיאות.
   ד. גרסאות - וידוא ניהול גרסאות לנתונים (Dataset Registry): כל שינוי במאגר הנתונים מוכרח להירשם כגרסה: מה השתנה, מי אישר, ומה ההשפעה.
2. יש להוסיף לתקינה "שכבת מודל" מחייבת: תוכנה רגילה מתעדכנת לפי נוהל "שינוי גרסה". שינוי קטן במודל AI עלול לשנות התנהגות באופן לא אינטואיטיבי. לכן צריך תקינה ברורה לניהול מודל ה-AI - מי אחראי למודל ה-AI, איך מעדכנים אותו, ואיך חוזרים לאחור. מומלץ לעדכן את התקינה הקיימת בתוספות הבאות:
   א. רישום מודלים (Model Registry): רישום של "תעודת זהות" של מודל ה-AI: שם, גרסה, תאריך, מי אחראי, באילו נתונים אומן ולמה הוא מיועד.
   ב. בקרת שינוי מודל (Model Change Control): בקרה על כל עדכון של מודל ה-AI שמחויב בתהליך אישור (בדומה לשינוי בכל מערכת קריטית).
   ג. בדיקות לפני הפעלה (Pre-deployment Tests): לפני שמכניסים מודל חדש יש לבדוק אותו מול "סט בדיקות קבוע" (קבוצת תרחישים) כדי לוודא שלא "נשברה" התנהגות חשובה.
   ד. בדיקת יכולת חזרה אחורה (Rollback): בדיקה זו חשובה מכיוון שאם אחרי עדכון מתגלה בעיה נדרשת יכולת להחזיר גרסה קודמת במהירות. נוסף על בדיקת יכולת חזרה לאחור, יש לוודא שמתקיימים תרגולים תקופתיים בנושא זה.
3. ללא תוספות אלו לתקינה הקיימת, הארגון עלול "להחליף מוח" למערכת אוטונומית בלי להבין מה השתנה.
4. יש לשלב בתקינה הקיימת מדדי התנהגות ייעודיים ל-AI: Drift, שגיאות קריטיות ואמון (ראו פירוט להלן). בתקנים הקלאסיים מודדים זמינות, הרשאות ואירועי סייבר. במערכות ביטחוניות בלתי מאוישות אוטונומיות מבוססות AI נדרש למדוד גם "בריאות התנהגותית" - האם המודל עדיין מתאים למציאות. מומלץ להוסיף לתקינה את מדדי התנהגות ה-AI הבאים:
   א. Drift (סטייה לאורך זמן): המודל עלול להידרדר בלי "להתריע", ולכן יש לבדוק אם הנתונים שהמערכת רואה היום שונים מהנתונים שעליהם אומנה.
   ב. שיעור שגיאות קריטיות: לא רק "כמה טעויות", אלא "אילו טעויות".
   ג. מדד אמון: נדרש לוודא שיש נוהל ברור המנחה על הפעולה הנדרשת אם המודל אינו אמין, לדוגמה הפניה לבקרה אנושית, חיישן נוסף או עצירה.
5. הסיכון במערכות ביטחוניות בלתי מאוישות הוא לא רק "פריצה", אלא גם "הידרדרות שקטה" המובילה להחלטות שגויות.
6. בתקנים רבים זמינות נתפסת כיעד IT כללי ("המערכת זמינה"). במערכת ביטחונית בלתי מאוישת אוטונומית, גם אם "השרת חי" אבל הנתונים בזמן אמת מגיעים באיחור או מגיעים באופן חלקי - המערכת למעשה אינה כשירה. לכן בתקינה הקיימת נדרש לעדכן דרישות זמינות ושרידות כלהלן:

א. יעדי רמת שירות (SLO[22]): לא רק ש"השרת עובד", אלא שהנתון מגיע בזמן, בקצב הנדרש ובאיכות הנדרשת. מודל AI שאופיין ופותח על בסיס רמת שירות מסוימת עלול להינזק מרמת שירות השונה באופן מהותי, ובעקבות זאת לפגוע בהבטחת המשימה.

ב. תרגילי ניתוק ועומס: לפחות אחת לפרק זמן יש לדמות נתק תקשורתי או עומס כדי לבדוק איך המערכת מתנהגת לאחר התקלה.

ג. Graceful Degradation: תקן המחייב התנהגות בטוחה במצב חסר, לדוגמה ירידה לפעולה שמרנית, מעבר למוד "המתנה" או דרישת אישור אנושי.

ד. להוסיף נספח זמינות בתקנים כך שיכלול עיכוב (latency), עדכניות (freshness) וחלופת נסיגה (fallback) ובחינתם במבחנים מוקדמים בשלב הוכחת המערכת.

7. יש להרחיב את התקינה בנושא שרשרת אספקה כך שתכלול "שרשרת אספקה של AI": ארגונים רבים מסתמכים כיום על ספריות תוכנה, מודלים חיצוניים, שירותי ענן או רכיבי צד ג'. בתקנים קלאסיים יש התייחסות לספקים, אבל לא תמיד ברזולוציה של "מודל/ספרייה/דאטה-סט"[23]. מומלץ לעדכן את התקינה הקיימת ולהוסיף ביקורת על הנושאים הבאים:

א. "עץ מוצר" (SBOM[24]) לרכיבים קריטיים: לדעת בדיוק מאילו רכיבים בנויה המערכת.

ב. תהליך קבלה לרכיב (Acceptance): לפני הכנסת רכיב חיצוני יש לבצע בדיקות בסיס, תיעוד, אישור.

ג. דרישות חוזיות עם הספק: זמני תיקון, חובת דיווח פגיעויות ושקיפות ברכיבים.

ד. סעיף AI Components, כגון מודלים וספריות, וכן מסמכים נדרשים רלוונטיים בנושא זה בנוהלי העבודה עם הספקים.

8. יש לעדכן את התקינה כך שתכלול דרישות ביקורת ייעודיות לשימוש ב"סוכני AI" שפותחו ואומנו על ידי ספקי מודלים חיצוניים: כאשר היכולת האלגוריתמית אינה נשלטת במלואה בידי הגוף המבוקר, יש לחייב "שכבת הבטחה ספקית" כתנאי כשירות, הכוללת לכל הפחות: (א) תיעוד תיחום שימוש (Scope) ומגבלות ידועות של הסוכן/המודל והוראות שימוש מותר/אסור; (ב) מנגנון עדכונים מבוקר - גרסאות, תהליך שינוי, התראות על שינויי התנהגות, ויכולת חזרה לגרסה קודמת; (ג) ראיות לבדיקות לפני הטמעה ובשגרה, לרבות בנוגע לחוסן ולתרחישי קצה; (ד) דרישות חוזיות מחייבות של שקיפות מינימלית, לרבות SLA לתיקון וקביעה של תהליך דיווח אירועים, וכן זכות לקבלת ראיות ביקורת של צד ג'. כך הביקורת תוכל להעריך לא רק "אם יש בקרות", אלא אם יש יכולת הוכחה חוזית-ראייתית מול הספק, לרבות לגבי כשירות, חוסן ועדכונים של סוכני ה-AI.

9. יש לקבוע בתקינה "סט ראיות מזערי" כתנאי כשירות Audit-Ready Baseline- : תקנים רבים כיום מאפשרים "הצהרה" על עמידה בהם ללא הצגת ראיות. בעידן ה-AI, ללא ראיות (למשל גרסאות, בדיקות, מדדים), אי אפשר לדעת אם המערכת נשלטת. לכן יש לקבוע בתקינה רשימה מזערית של מסמכים או לוגים שמוכרחה להתקיים בכל מערכת AI קריטית. על הרשימה לכלול, בין היתר, מפת מערכת ותלויות (Dependencies), רישום מודל, דוחות בדיקה, דוחות זמינות ותיק אירועים.

10. בעידן ה-AI חשוב מאוד להוסיף לתקינה התייחסות שתבטא ביקורת שאינה מסתפקת רק בבדיקת מסמכים, כמפורט להלן:

א. בדיקות שחזור (Reproducibility): לבחור מודל אחד ומאגר נתונים אחד ולהראות שאפשר לשחזר את התהליך בצורה מבוקרת.

ב. דגימה חובה: בכל ביקורת - דגימת מודל/דאטה-סט/רכיב ספק.

ג. היצמדות למדדים: כל בדיקה מסתיימת במדד ובתוצר ראייתי.

11. את כלל התוספות לתקינה שתוארו לעיל יש לתרגם ל"שפה ארגונית" - בעלי תפקיד, לוחות זמנים, מדדי הצלחה: כדי שהתוספות לתקינה לא יישארו תיאורטיות, כל תוספת צריכה להיות מנוסחת בדרך מובנת ברמת ההנהלה: מי אחראי (Owner), מהו התוצר (Deliverable), עד מתי מודדים הצלחה ואיך.

---

22 Service Level Objective - היעד הכמותי המשקף את איכות השירות של המערכת, לדוגמה אחוז מתקציב שגיאות מותר.

23 דאטה-סט (Dataset) - אוסף מאורגן של נתונים ומידע שנאספו לצורך מחקר, ניתוח או אימון או תיקוף של מודלי AI (מילון Merriam-Webster).

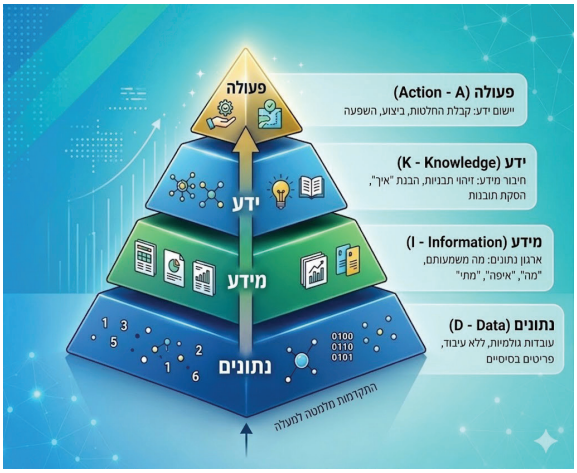
24 Software Bill of Materials - רשימה מפורטת של כל המרכיבים בתוכנה מסוימת: ספריות קוד פתוח, רכיבי צד ג', גרסאות ספציפיות של כל רכיב, רישיונות.

מתודולוגיית ביקורת מבוססת נתונים במערכות ביטחוניות בלתי מאוישות משולבות AI

ביקורת על מערכות ביטחוניות בלתי מאוישות ששולבו בהן יכולות AI אינה יכולה להסתמך על ראיות מסורתיות בלבד. על כן מוצע לאמץ מתודולוגיית Data Science המבוססת על מודלים ועל כלי מדידה שעשויים לתמוך במסגרת הביקורת כגון:

1. מודל נמי"ף (DIKA)[25]: לעומת המאה העשרים, שבה היה מצופה שהמבקר יבחן כיצד נתונים גולמיים(Data) הופכים למידע (Information), לידע (Knowledge) ולתובנות מבצעיות (Wisdom), ויזהה נקודות כשל בעיבוד המידע, בעידן ה-AI נדרש לאמץ מודל DIKA, שהוא וריאציה של DIKW המכוונת לפעולה (Action). מודל זה יכול לשמש מסגרת להבנת התהליך שבו המערכות "לומדות" ומקבלות החלטות לפעולה. להלן פירוט של שלבי המעבר מנתונים לפעולה, שהוא לב ליבה של מערכת ביטחונית בלתי מאוישת משולבת AI:
   א. נתונים (Data): הקלט הגולמי (Data Raw) שהמערכת מקבלת; למשל קובצי תמונות, רצפי מספרים מחיישנים או טקסטים לא מעובדים.
   ב. מידע (Information): שלב העיבוד שבו ה-AI מחלץ מאפיינים (Features); לדוגמה זיהוי "כתם אדום" בתמונה או זיהוי מילים ספציפיות במשפט.
   ג. ידע (Knowledge): שלב הלמידה שבו ה-AI מזהה תבניות (Patterns) וקשרים; למשל המערכת מבינה ש"כתם אדום" במקום מסוים ובצורה מסוימת הוא "תמרור עצור".
   ד. פעולה (Action): השלב הקריטי ב-AI יישומי שבו הידע מתורגם להחלטה אופרטיבית; למשל הרחפן האוטונומי מחליט להתריע לפני המפעיל כי הוא זיהה גורם עוין.
2. שימוש באלגוריתם KNN (K-Nearest Neighbors): כלי זה משמש את המבקר לזיהוי חריגים (Outliers) בסט הנתונים. העיקרון הבסיסי של האלגוריתם הוא שדברים דומים קרובים זה לזה. כאשר האלגוריתם מקבל "דוגמה" חדשה (נקודת נתונים ללא תווית) הוא בודק מי הן הנקודות הקרובות אליה ביותר מתוך הנתונים הקיימים. זיהוי אשכולות נתונים המרוחקים מהנורמה עשוי להעיד על ניסיון "הרעלת נתונים" בשלבי הלמידה.
3. ערכי שפלי (Shapley Values): כלי זה פותר את בעיית ה"קופסה השחורה" של מודל ה-AI עבור המבקר. הוא מאפשר להסביר (Explainability) את הערך (המשקל) של כל אחת מהתכונות (Features) שהשפיעו על החלטה ספציפית. לדוגמה, אם הביקורת מעלה כי קלט מחיישן LiDAR[26] תרם 70% להחלטת ניווט שגויה, המבקר יכול להצביע על פגיעות בחיישן הספציפי או על הסתמכות-יתר מסוכנת של המודל עליו.
4. ביקורת במגבלות סיווג: אף שאין מדובר במרכיב ייחודי לביקורת בעידן ה-AI, לנוכח רגישות המערכות הביטחוניות הבלתי מאוישות מומלץ שהביקורת תתבצע במודל של "גישה מדורגת" ותשתמש במערכות ניתוח נתונים מבודדות (Air-gapped) המאפשרות למבקרים בעלי סיווג מתאים לבחון לוגיקה והתנהגות של מודל AI בלי להוציא נתונים מסווגים אל מחוץ לארגון.
5. לצורך הבטחת האחריותיות מומלץ כי הגוף המבוקר יציג מדדי חסינות רציפים. להלן המלצות מעשיות

תרשים ג: מודל נמי"ף המותאם לעולם הביקורת בעידן ה-AI



המקור: הופק באמצעות Chat-GPT ו-Gemini

---

25. Russell L. Ackoff, From Data to Wisdom, Ackoff's Best 170 (1999).

26. Light Detection and Ranging - גילוי וטווח באמצעות אור.

לאימוץ מדדים רלוונטיים בהיבטי ביצוע וסיכון (KPIs/KRIs)[27]:

א. Model Drift Rate: קצב הירידה בדיוק של מודל ה-AI לאורך זמן.

ב. Poisoning Detection Lag: הזמן החולף מרגע הזרקת נתון זדוני ועד לזיהויו.

ג. Adversarial Robustness Score: עמידות מודל ה-AI מול קלטים מניפולטיביים בבדיקות Red Teaming[28].

ד. Mission Accomplishment Under Jamming: שיעור השלמת המשימה בתנאי שיבוש אלקטרוני (אחוז).

ה. Explainability Index: רמת היכולת לנמק החלטות אלגוריתמיות קריטיות.

6. ביקורת מתמשכת (Continuous Auditing): ניטור רציף באמצעות סקריפטים אוטומטיים שמאפשר למבקר לזהות חריגות בזמן אמת, ולא בדיעבד, דבר בעל תרומה ניכרת למניעת כשלים קטלניים במערכות ביטחוניות בלתי מאוישות משולבות AI.

יישום המסגרת: מתודולוגיית ביקורת עדכנית והמלצות לשיפור מתמשך

מסגרת הביקורת המוצעת במאמר מבקשת לחולל שינוי ממוקד אך מהותי באופן שבו ביקורת בתחום הגנת הסייבר ניגשת למערכות ביטחוניות בלתי מאוישות מבוססות AI: מעבר מביקורת "ציות לבקרות" לביקורת "יכולת הוכחה" (Audit-Ready).

המשמעות המעשית היא שכל טענה על אבטחה, על אמינות, על חוסן או על כשירות מבצעית של מערכת ביטחונית בלתי מאוישת מבוססת AI מוכרחה להיתמך בראיות הניתנות לשחזור: מפת תלויות מלאה[29] (של חיישנים-תקשורת-מחשוב-מודל-ממשק החלטה), תיעוד מקור והיסטוריה של נתונים (provenance/תיוג/גרסאות), רישום גרסאות מודל ותהליך שינוי מבוקר, דוחות בדיקה ומדדים (כולל drift ושגיאות קריטיות), יעדי שירות לזמינות נתונים בזמן אמת ותיעוד תרגילי שרידות (resilience drills), וכן ניהול שרשרת אספקה של רכיבי AI, בדיקות קבלה לספקים ואימות רכיבים.

כאשר המסגרת המתודולוגית של הביקורת מיושמת כך היא מאפשרת לביקורת לנסח ממצאים בני הגנה: ממצא המבוסס על ראיה, מסקנה המבוססת על ניתוח סיכון והמלצה מדידה הכוללת בעל תפקיד, לוח זמנים ומדד הצלחה. בכך מתורגמת מורכבות ה-AI לשפת ביקורת ברורה, שאינה תלויה בטכנולוגיה מסוימת, אלא בתקן מזערי של שליטה, שקיפות ושחזור.

מבחינה מוסדית, יישום המסגרת מחייב עדכון ממוקד של התקינה והמתודולוגיה הקיימות; כאמור לא כתחליף לתקנים הקלאסיים אלא כתוספת שכבת AI-Assurance.

בהקשר הישראלי אפשר למנף את תפיסת ההגנה של מערך הסייבר הלאומי כעוגן המתחבר ישירות לקריטריונים ולמדדים של Audit-Ready, וליצור "שפה משותפת" בין הנחיה להגנה ולביקורת.

במשרד מבקר המדינה ההטמעה המעשית צריכה לכלול: הוספת פרק AI-Audit למסמך המתודולוגיה; קביעת "סף מוכנות לביקורת" שבו היעדר ראיות מזעריות הוא ממצא מהותי; ותוכנית דגימות מחייבת הכוללת לפחות מודל אחד, מאגר נתונים אחד ורכיב ספק אחד בכל ביקורת רלוונטית, לצד בדיקת שחזור מוגבלת (Reproducibility) ותרגיל זמינות/שרידות אחד המבוסס ראיות.

כך תוכל הביקורת להעריך לא רק "אם יש בקרות", אלא אם הארגון מסוגל להוכיח שליטה בסיכון AI באופן עקיב לאורך זמן - וזוהי נקודת המפתח לשמירה על כשירות, על בטיחות ועל אמון ציבורי במערכות אוטונומיות בסביבה ביטחונית.

לסיכום אופן יישום המסגרת המאמר מציע להתאים את כלי הביקורת הקיימים באופן הבא:

1. הטמעת מודל "הבטחת משימה" כתקן מחייב לביקורת לכל מערכת ביטחונית בלתי מאוישת.
2. הקמת "סל ראיות דיגיטלי" מחייב לספקים, הכולל יומני למידה ותיעוד ארכיטקטורה.
3. ביצוע מבחני חוסן (Red Teaming) ייעודיים ל-AI המדמים תוקף אדפטיבי.
4. אימוץ כלי הסברתיות (Shapley) וזיהוי חריגים (KNN) כחלק בלתי נפרד מדוחות הביקורת.

---

27 שני כלי ניהולי משלימים: KPI; Key Performance Indicators, שמטרתו למדוד התקדמות לעבר השגת יעדים; KRI ;Key Risk Indicators, שמטרתו לזהות ולנהל סיכונים פוטנציאליים.

28 Red Teaming - צוות אדום - מונח מעולם הביטחון, המודיעין ואבטחת המידע, המתאר תרגול של תקיפה מבוקרת על ארגון כדי לבדוק את מוכנותו. הרעיון העיקרי הוא לאמץ "ראש של יריב" כדי לאתר פרצות, כשלים לוגיים או נקודות תורפה שהארגון עצמו לא הבחין בהם.

29 מפת תלויות מלאה (Dependency Map או Dependency Matrix) - כלי ויזואלי או לוח המשמש בניהול פרויקטים, בהנדסת מערכות ובפיתוח תוכנה. מטרת המפה היא לנתח ולנהל את כל הקשרים והאינטראקציות בין רכיבים שונים במערכת.

5. הקמת צוותי ביקורת רב-תחומיים, שכן חוסר השקיפות המלא של מודלי ה-AI הוא "בעיה מורכבת" שאינה ניתנת לפתרון בדרכים טכניות בלבד, אלא דורשת שילוב של מומחי סייבר, מדעני נתונים, משפטנים ומבקרים.

סיכום

במערכות ביטחוניות בלתי מאוישות בעידן ה-AI שאלת הביקורת אינה רק אם יש בקרות, אלא אם הגוף המבוקר מסוגל להוכיח, בראיות ניתנות לשחזור, שהמערכת בכללותה - רכיביה והתהליכים המבצעיים המתחוללים בה על בסיס שימוש במודלי AI - שומרת על הבטחת משימה, על חוסן מבצעי ועל אחריותיות לאורך מחזור חייה.

המאמר מציג מסגרת של הבטחת מוכנות לביקורת (Audit-Ready Assurance Framework) בתחום ההגנה מפני איומי סייבר במערכות ביטחוניות בלתי מאוישות משולבות AI, המבוססת על מיפוי התקינה הקיימת והרחבתה לעידן שבו נתונים ומודלי AI הם רכיבי סיכון מרכזיים. המסגרת מגדירה קריטריוני ביקורת ברורים, מדדי ביצוע וסיכון (KPIs/KRIs), סט ראיות מזערי, לצד תוכנית בדיקות ודגימות ישימה לביקורת.

נוסף על כך, המאמר מציע עדכון תפיסה ותקן המחבר בין עקרונות ההגנה של מערך הסייבר הלאומי ובין שיטות הביקורת של משרד מבקר המדינה ומתרגם אותם להמלצות יישומיות לעדכון מתודולוגיית הביקורת. התוצאה היא מעבר מביקורת של "ציות לבקרות" לביקורת של "יכולת הוכחה", המאפשרת ממצאים מדידים, השוואת ביצועים וניתוח תוצאות התנהגות של מודלי AI.

נספח: תקנים מרכזיים המקובלים להתוויית המתודולוגיה לביצוע ביקורות בתחום אבטחת המידע והסייבר בארגונים

א. (2024) NIST[30] CSF[31] 2.0: מסגרת תוצרים לניהול סיכון סייבר לכל ארגון; מחזקת מינהל תקין (Governance), ניהול סיכונים ובקרה על שרשרת אספקה.

ב. NIST SP 800-37 Rev.2 Risk Management Framework (2018): תהליך מחזור חיים לניהול סיכוני אבטחה/פרטיות: אפיון מערכת, בחירת בקרות, יישום, הערכה, אישור וניטור רציף; בסיס מתודולוגי לביקורת ראייתית.

ג. NIST SP 800-53 Rev.5 (2020): קטלוג בקרות אבטחה/פרטיות המאפשר מיפוי בקרות לממצאים; ומספק שפה לביקורת בתחום המינהל התקין - ניהול ותקינה, הרשאות, לוגים, הגנת שרשרת אספקה ועוד.

ד. NIST SP 800-53A Rev.5 (2022): מתודולוגיית הערכת בקרות (בחינה/תצפית/ריאיון) לבניית תוכנית בדיקות ודגימות; תומכת בראיות הניתנות לשחזור.

ה. ת"י (תקן ישראלי) ISO[32]/IEC[33] 27001:2022: תקן מערכת ניהול אבטחת מידע (ISMS[34]) המעוגן בישראל כתקן ישראלי; מחייב מדיניות, בקרות, ביקורת פנימית, טיפול בסיכונים ושיפור מתמיד.

ו. ת"י (תקן ישראלי): ISO/IEC 42001:2023 - מערכת לניהול AI (AIMS[35]): תקן מערכת ניהול ל-AI; מגדיר דרישות ארגוניות לניהול AI - מינהל תקין, אחריות, תיעוד, בקרת שינוי, ניהול השפעות ומדדים.

ז. The Institute of Internal Auditors (IIA), Global Technology Audit Guide (GTAG) Auditing Cyber Risk in the Age of AI (2020): מדריך ביקורת טכנולוגיה גלובלי של לשכת המבקרים הפנימיים העולמית (IIA), הכולל הנחיות למבקרים פנימיים, למנהלי סיכונים ולחברי דירקטוריון. המדריך הספציפי משנת 2020 מתמקד בשינוי הפרדיגמה של ביקורת בתחום אבטחת הסייבר בעידן ה-AI.

30 National Institute of Standards and Technology; NIST

31 Cybersecurity Framework; CSF

32 International Organization for Standardization; ISO

33 International Electrotechnical Commission; IEC

34 Information Security Management System; ISMS

35 Artificial Intelligence Management System; AIMS