



מדריך חינוכי · Tvuna Ops

אבטחת סוכני AI שמתחברים למערכות שלכם

מדריך מעשי לצוותי IT, אבטחה והנהלה: מה
יכול להשתבש כשסוכן מתחבר ל-ERP,
ל-CRM, למערכות ותיקות ולשרתי MCP, ואיך
סוגרים את זה מראש.

למה סוכן שונה מצ'טבוט

צ'טבוט עונה. סוכן **פועל**: קורא נתונים, מפעיל כלים ומשנה רשומות. ברגע שהוא מחובר למערכת ליבה, כל טקסט שהוא קורא (מייל, מסמך, הערה ב-CRM) עלול להפוך להוראה. השאלה היא לא "האם המודל חכם", אלא "מה הוא מסוגל לעשות כשמישהו מטעה אותו". רשימת OWASP Top 10 for LLM Applications (גרסת 2025) היא נקודת הפתיחה המקובלת. אלה הסיכונים הרלוונטיים ביותר לסוכנים מחוברים:

(LLM01) Prompt Injection: הוראות זדוניות שמוחדרות ישירות, או בעקיפין דרך תוכן שהסוכן קורא, וגורמות לו לחרוג מהמשימה. ההזרקה העקיפה מסוכנת במיוחד: אף אחד לא "מדבר" עם הסוכן, הוא פשוט קורא מייל.

LLM06) Excessive Agency: לסוכן יש יותר כלים, הרשאות או אוטונומיה ממה שהמשימה דורשת. זה מה שהופך טעות קטנה לנזק אמיתי.

עוד שלושה סיכונים מרכזיים

דליפת מידע רגיש (LLM02, LLM07): הסוכן חושף נתוני לקוחות, סודות או את הנחיות המערכת שלו למשתמש הלא נכון או לשירות חיצוני.

טיפול לא בטוח בפלט (LLM05): פלט המודל עובר ישירות לשאילתת SQL, לפקודה או ל-API בלי בדיקה.

שרשרת אספקה ושרתי MCP (LLM03): שרת MCP הוא קוד צד שלישי עם גישה למערכות שלכם. תיעוד MCP הרשמי מתאר סיכונים כמו confused deputy, העברת טוקנים (token passthrough) והרשאות רחבות מדי.

שאר הרשימה: הרעלת נתונים ומודל, חולשות Embeddings, מידע שגוי וצריכה בלתי מוגבלת. הרשימה המלאה בעמוד המקורות.

10 בדיקות לפני שסוכן נוגע במערכת ליבה

1 מלאי: יש רשימה של כל הסוכנים, הכלים ושרתי ה-MCP שמחוברים, ולכל אחד יש בעלים.

2 משתמש שירות ייעודי: הסוכן לא רץ עם משתמש של עובד או עם אדמין.

3 הרשאות מינימליות: כל כלי מקבל רק את הפעולות והטבלאות שהוא צריך.

4 קריאה קודם: בשלב הראשון הסוכן קורא בלבד, בלי כתיבה.

5 שער אישור: מחיקה, כסף, שינוי הרשאות ושליחה החוצה דורשים אישור אנושי.

6 **תוכן חיצוני הוא לא הוראה:** מיילים, קבצים ודפי אינטרנט מטופלים כנתונים בלבד.

7 **סודות מחוץ לפרומפט:** אין מפתחות API או סיסמאות בהנחיות או בקוד.

8 **לוג לכל קריאה:** מי, מה, מתי, עם אילו פרמטרים ומה התוצאה.

9 **אימות שרתי MCP:** מקור ידוע, גרסה נעולה, וסקירה לפני עדכון.

10 **מידע אישי:** ברור אילו נתונים אישיים הסוכן רואה, למה, ולאן הם נשלחים.

סימנתם פחות מ-7? הסוכן כנראה עוד לא מוכן לכתוב למערכת ליבה.

אכיפה מחוץ לסוכן

הגבלה שהסוכן מחליט עליה בעצמו היא לא הגבלה.

הנחיה כמו "אל תמחק רשומות" בפרומפט נעקפת בהזרקה אחת. האכיפה צריכה לשבת בשכבת הכלי וה-API, שם המודל לא יכול לשנות אותה.

1. הרשאות ברמת הכלי וה-API

כל כלי MCP חושף פעולות צרות ומוגדרות (למשל `get_invoice` ולא `run_sql`). השרת בודק scope והרשאה בכל קריאה, ולא סומך על "כוונת" הסוכן.

2. משתמש שירות עם הרשאות מינימליות

משתמש נפרד לכל סוכן, עם הרשאות ברמת מסד הנתונים וה-ERP. טוקנים קצרי חיים שמוגבלים לקהל (audience) הנכון, ובלי להעביר את הטוקן של המשתמש לשירותים אחרים, כפי שממליץ תיעוד MCP.

אישור אנושי ופריסה הדרגתית

3. שערי אישור לפעולות רגישות

רשימה סגורה של פעולות שלא רצות בלי אדם. הסוכן מכין טיוטה (Preview), אדם מאשר, והמערכת (לא הסוכן) מבצעת. השער נאכף בשרת, ולכן אי אפשר לדלג עליו.

4. פריסה הדרגתית: קריאה קודם

שלב	מה מותר	מתי עוברים הלאה
1	קריאה בלבד, בסביבת בדיקות	אחרי סקירת לוגים ותרחישי הזרקה
2	כתיבה עם אישור אנושי לכל פעולה	כששיעור הטעויות נמוך ומתועד
3	כתיבה אוטומטית רק לפעולות הפיכות ובסיכון נמוך	רק עם ניטור והתראות

5. לוג וביקורת

לוג שאי אפשר לשנות לכל קריאת כלי: זהות הסוכן, המשתמש שהפעיל אותו, הפרמטרים והתוצאה. בלי זה אי אפשר לחקור אירוע או להוכיח עמידה ברגולציה.

תיקון, פלט וסודות

Rollback .6

לכל כתיבה צריכה להיות דרך חזרה: גרסאות לרשומות, מחיקה רכה (soft delete) או טרנזקציה שאפשר לבטל. פעולה שאי אפשר להחזיר שייכת לשער האישור.

7. ולידציה לפלט

פלט המודל הוא קלט לא אמין. בודקים אותו מול סכמה, מגבילים ערכים וטווחים, ולא מעבירים אותו ישירות ל-SQL, ל-shell או ל-HTML.

8. ניהול סודות

המפתחות נשמרים בכספת סודות (Secrets Manager או Vault), עם רוטציה ופקיעה. הסוכן ניגש דרך הכלי, ולא רואה את הסוד עצמו.

9. תיקון 13 לחוק הגנת הפרטיות

תיקון 13 נכנס לתוקף באוגוסט 2025 והרחיב את סמכויות האכיפה של הרשות להגנת הפרטיות, כולל עיצומים כספיים.

כשסוכן מעבד מידע אישי, כדאי לבדוק:

- לאילו מאגרי מידע אישי הסוכן ניגש, והאם הגישה תואמת את מטרת האיסוף.
- האם מידע אישי נשלח לספק מודל או לשירות בחו"ל, ובאילו תנאים.
- שהלוגים ובקורות הגישה מתיישבים עם תקנות אבטחת מידע.
- האם הארגון שלכם צריך למנות ממונה הגנת פרטיות.

זה לא ייעוץ משפטי. את החובות הספציפיות לארגון שלכם בדקו מול עורך דין ומול הנחיות הרשות.

מקורות

- :OWASP Top 10 for LLM Applications 2025
genai.owasp.org/llm-top-10
- :NIST AI Risk Management Framework
nist.gov/itl/ai-risk-management-framework
- :NIST AI 600-1, Generative AI Profile
doi.org/10.6028/NIST.AI.600-1
- :Model Context Protocol, Security Best Practices
modelcontextprotocol.io/.../security_best_practices
- הרשות להגנת הפרטיות:
gov.il/.../the_privacy_protection_authority

רוצים לדעת איפה הסוכן שלכם חשוף?

אנחנו בודקים הטמעות AI: הרשאות של סוכנים, חיבורי MCP ו-API, סיכוני prompt injection ודליפת מידע. בסוף מקבלים רשימת תיקונים לפי סדר עדיפות. ברירת המחדל היא בדיקה פסיבית. בדיקה פעילה נעשית רק באישור בכתב, ובלי לגעת בסביבת הייצור.

tvunaops.com/security ←

המדריך חינמי, ומותר להפיץ אותו כמו שהוא.