צ׳קליסט · בינה מלאכותית אחראית

12 שאלות לשאול לפני שמחברים AI למאגר הידע של הארגון.

לארגוני בריאות, בריאות הנפש ורווחה. השאלות מיועדות לשיחה עם ספק המערכת, עם מחלקת המחשוב או עם הצוות שמטמיע אותה. מבוסס על המאמר ״כשהידע עצמו מורעל״.

01 / מקורות המידע

- ☐ מאילו מקורות המערכת שואבת מידע, ומי אישר כל אחד מהם?
- ☐ האם יש הפרדה ברורה בין ידע מקצועי מאומת לבין תוכן חיצוני, כמו אתרים, מאמרים או משוב משתמשים?
- ☐ אם המערכת שואבת תוכן מהאינטרנט – מי בודק אותו לפני שהוא נכנס למאגר?

02 / הרשאות ועדכונים

- ☐ מי מורשה להוסיף, לשנות או למחוק מסמכים במאגר?
- ☐ האם כל שינוי נרשם – מי, מה ומתי – והאם מישהו עוקב אחרי שינויים חריגים?

03 / גרסאות ועדכניות

- ☐ מה קורה לנוהל ישן כשנכנס נוהל חדש במקומו? האם הוא מוסר מהמאגר?
- ☐ האם יש בדיקה תקופתית של מסמכים מיושנים או סותרים?

04 / שקיפות ההמלצות

- ☐ האם איש המקצוע רואה, לצד כל המלצה משמעותית, על אילו מקורות היא מבוססת?
- ☐ האם אפשר לשחזר בדיעבד על מה התבססה תשובה מסוימת?

05 / בדיקות ובקרה

- ☐ האם המערכת נבדקה גם מול ניסיונות מכוונים להטעות אותה (red teaming), ולא רק בתרחישים שגרתיים?
- ☐ האם נבדק שמסמכים במאגר אינם מכילים הוראות סמויות למודל (Indirect Prompt Injection)?
- ☐ אם המודל אומן או הותאם על נתוני הארגון – מאילו נתונים, ואיך נבדק שלא הושתלה בו ״דלת אחורית״?

תשובה של ״לא יודעים״ לאחת השאלות היא לא בהכרח סיבה לעצור את הפרויקט. אבל היא בהחלט סיבה להתעמק.