

אבטחת סוכני AI בפיתוח תוכנה

מסגרת כללית להרשאות, בקרה ואישור פעולות

סוכן פיתוח צריך לקבל רק את ההרשאות הנדרשות למשימה. פעולות בעלות השפעה על ייצור, נתונים או הרשאות מחייבות בקרה טכנית ואישור אנושי מוגדר. הנחיה כתובה לבדה אינה תחליף לחסימה טכנית.

שלוש שכבות הגנה

01

הרשאות וסביבת עבודה

הגדרת תיקיות מותרות, גישה לרשת וכלים זמינים. חסימת גישה לסודות, לפקודות הרסניות ולסביבות שאינן נדרשות למשימה. בדיקת ההרשאות בפועל לפני תחילת העבודה.

02

בדיקת השינוי לפני פרסום

בדיקות אוטומטיות וסקירת קוד בהתאם לרמת הסיכון. בודק מבוסס AI יכול לסייע, אך אינו מחליף הרשאות ואישור אנושי. אם בדיקה נדרשת נכשלת או אינה זמינה, הפרסום נעצר.

03

הפרדת זהויות ונתונים

זהות נפרדת לסוכן והרשאות מזעריות לכל סביבה. גישה לקריאה בלבד כשאינן צורך בכתיבה, בלי ירושת הרשאות מנהל. שינויים בנתונים מתבצעים בתהליך מאושר ומתועד.

חלוקת אחריות ואישורים

פעולה	בקרה ואחריות
כתיבת קוד והרצת בדיקות מקומיות	פעילות עצמאית בגבולות סביבת העבודה והמשימה שהוגדרו.
קריאת מידע	גישה מצומצמת למידע הנחוץ, בהתאם להרשאות ולרגישות הנתונים.
פרסום קוד ומיזוג לענף מוגן	בדיקות נדרשות, סקירה ומדיניות אישור שנאכפת בשרת הקוד.
פריסה לייצור ושינוי הגדרות תשתית	אישור אנושי לפעולה ולסביבה המדויקת, לצד אפשרות חזרה לאחור.
שינוי או מחיקת נתונים	סקירת השינוי, בדיקת גיבוי ושחזור, והרשאה מוגבלת להיקף שאושר.
גישה לסודות או הרשאות מנהל	מנגנון מאובטח, הרשאה זמנית ומצומצמת ותיעוד השימוש.

אישור שמתאים לפעולה

על האישור לציין מה משתנה, באיזו סביבה, על אילו משאבים ובאיזה היקף. שינוי מהותי בתוכן או ביעד מחייב בחינה מחודשת. לאחר הביצוע יש לאמת את התוצאה ולשמור תיעוד שניתן לביקורת.

סיכונים ובדיקות אימות

רשימת בדיקה להתאמה לסביבת העבודה

סיכון	בקרה מומלצת	מה לבדוק
מחיקה או דריסה של היסטוריית קוד	הגנת ענפים והגבלת פקודות הרסניות	האם החסימה נאכפת גם מחוץ לממשק הסוכן?
פריסה או שינוי תשתית ללא אישור	הפרדת הרשאות פריסה ואישור ליעד המדויק	האם חשבון הסוכן יכול לבצע פעולה כזו ישירות?
שינוי או מחיקת נתונים	זהות נפרדת והרשאות מזעריות	האם הוספה, עדכון, מחיקה ושינוי מבנה נחסמים כשאינם מורשים?
דליפת סודות ומידע רגיש	ניהול סודות מרכזי והגבלת קריאה ורשת	האם סודות נחשפים בקבצים, בהגדרות, ביומנים או בפלט הכלים?
פרסום קוד שלא נבדק	בדיקות חובה ועצירה כשאין זמינות	האם ניתן לעקוף את הבדוק או לפרסם לאחר כישלון?
תוסף או כלי חיצוני שנפרץ	גרסאות מאושרות ובקרת עדכונים	האם עדכון יכול להרחיב הרשאות או להפעיל קוד ללא בחינה?
עקיפת תהליך הבקרה	אכיפת מדיניות בשרת ובסביבה המארכת	האם פקודה חלופית או זהות אחרת עוקפות את ההגבלות?
פגיעה בנתונים או בתשתית	צמצום חשיפה לרשת, יומני ביקורת וגיבוי	האם קיימת הגנת מחיקה, והאם שחזור נבדק בפועל?

כיצד מאמתים שההגנות פועלות

- בודקים פעולות מותרות ופעולות אסורות בסביבת בדיקה עם נתוני דמה, כולל ניסיונות לעקוף את המסלול הרגיל.
- מבחינים בין כלל כתוב, הגדרה טכנית ובדיקה שהוכיחה חסימה בפועל.
- בודקים גם מצב שבו שירות הבקרה אינו זמין, האישורים פגו או הביצוע הושלם חלקית.
- מתעדים את היקף הבדיקה, הראיות, החריגים ומה שעדיין לא אומת. אין להסיק מבדיקה חלקית שכל המערכת מוגנת.

החלטות שצריך להגדיר מראש

אילו פעולות דורשות אישור, מי מוסמך לאשר, כיצד ניתנת גישה זמנית, אילו סביבות מופרדות וכיצד מטפלים בחריגים ובאירועי אבטחה. לכל פער יש להגדיר בעל אחריות ומועד לבדיקה חוזרת.