לקוחות נכבדים תודה על השתתפותכם במפגש שולחן עגול בנושא הגנת סייבר בתחום GENAI.

תודה לנותני החסות על ההרצאות המעניינות שתרמו לדיון שהתקיים לאחר מכן. תודה לגבריאל מחברת B2 ותודה ליוסי עטיה מחברת ISLAND.

הנושא של "הגנה" בהקשר GenAI חוצה את גבולות הסייבר ה"קלאסיים" (של תקיפות סייבר) וכולל גם נושא הגנה והתייחסות לנושאים של compliance, רגולציה, הגנת הפרטיות וזליגה של IP ומידע סודי של החברה. נושאים אלו אינם חדשים אבל GenAI מגביר את הסכנה בנושאים אלו לדוגמה עובד מעלה את בדיקות הדם שלו לשירות ה- GenAI בשביל לשאול שאלות. כעת עובד הסייבר אשר קיבל התראה על מסמך שעלה עם פרטים מזהים -חשוף לבדיקות הדם של חברו עובד החברה. דברים כאלו התרחשו בתדירות פחותה לפני עידן ה- GenAI.

"הגנה" נוספת של עובדי הארגון הנה איזכור עוד ועוד ש- GenAI יכול לטעות ולהטעות בה בשעה שה- GenAI "הכי בטוח בעולם" במה שאמר!!

ב- MITRE יצרו מיפוי של הטכניקות לתקיפה של AI - https://atlas.mitre.org/resources/ai-security-101 - מומלץ לארגונים לעבור על המסמך ולהתעדכן.

בדיון עלתה העובדה שחיפוש הפתרונות בנושא של "הגנת סייבר לתחום GenAI" התרכז בהתחלה בסנריו הבסיסי של "עובד החברה שמשתמש בשירות LLM ציבורי". לאחר מכן עלתה חשיבות הסנריו של "שימוש ב- LLM כחלק מאפליקציה ארגונית קיימת". לדוגמא ארגון תאר מצב שבו כבר שנים עובדים באופן מסודר בארגון עם אפליקציית Grammarly. ובשדרוג האחרון הסתבר שה- Grammarly כולל בתוכו מנועי LLM והמשמעות היא שעובדי החברה חשופים לסיכוני הסייבר של LLM כחלק מאפליקציה ותיקה שעד לא מזמן לא כללה סיכונים אלו. הדבר מתחיל להשתקף בשאלוני ובמערכות ה- Third Party Risk Management אשר בוחנות את החשיפה של הארגון עקב העבודה מול שותפים.

ואולם כעת, כאשר ארגונים מתחילים ליישם מודלי GenAI ובעיקר Agentic אשר גם מבצעים פעולות הלכה למעשה, נראה שרוב המאמץ בהגנת סייבר מושקע בהגנה על מודלים ו- Agents שבונים בארגונים ואשר כוללים (או בעלי גישה) למידע רגיש אם אפשרות לבצע פעולות. ובכך עולה החשיפה של הארגונים. לדוגמא ארגון פיננסי אשר בונה כעת מודל LLM אשר יטפל בבקשות לקוחות חושש ממצב שבו באמצעות prompt engineering ישכנעו את המודל לתת הלוואה בתנאים לא הגיוניים. כמו גם ממצב של model poisoning שבו גורמים למודל לענות בצורה לא הגיונית. מהלכים כאלה עלולים לגרום לארגון הפסד כספי גדול גם בגלל הפגיעה עצמה וגם בגלל הפגיעה במוניטין.

נושא נוסף שמתחיל לקבל חשיבות הוא הגנה בממשקים השונים בעיקר MCP ו-A2A.

כמהלך מקדים לבנית תפיסת ההגנה בתחום ה- GenAI ולבחירת הכלים השונים ארגונים מגדירים מדיניות הגנה בנושא GenAI שכוללת את הקווים הכלליים להתייחסות לנושא. דברים כמו – איזה סוגי פעולות או סוגי מסמכים אין להשתמש כלל ב- GenAI או לחילופין להשתמש ב- GenAI שנמצא פנימית בתוך הארגון. או לחילופין "חוסמים שימוש בארגון כל ה- LLM אשר אינו מבצע הצפנה של נתונים ומסמכים שמקבל מהלקוח". מסמכי המדיניות גם כוללים התייחסות לדברים יותר פרטניים כמו "גישה ל- LLM ציבורי תאופשר רק מדפדפן ארגוני" "גישה ל- LLM מדפדפן לא ארגוני תתאפשר רק ל- LLM שאיתם יש הסכם לארגון" וכד'.

במקביל לבניית המדיניות בתחום, חשוב מאוד לבצע מהלך (מתמיד) של discovery. באיזה LLM מתבצע שימוש כיום על עובדי הארגון בפועל? בין אם שירות LLM דרך הדפדפן. או שימוש ב- LLM כחלק מאפליקציה ארגונית (כמו בדוגמה שנתנו עם Grammarly ויש דוגמאות נוספות כמו office copilot או – Salesforce AI Einstein). הכלים מזהים שימוש ב- GenAI בד"כ על ידי gateway (וגם על ידי agent שמותקן ב- PC) גם על ידי רשימות אתרים או אפליקציות אשר מסווגות כ"משתמשות ב- GenAI" ומתעדכנות כל הזמן אבל גם לפי צורת השימוש – אם מזהים שאלות ותשובות אז מסווגים את השימוש כשימוש ב- GenAI.

מהלך נוסף אשר מתבצע ברוב הארגונים הוא בניית לומדה לשימוש ב- GenAI שכולל את היתרונות אבל גם את הסכנות כאשר פתרונות ההגנה בתחום של GenAI כוללים גם תזכורות בזמן שימוש – לדוגמה כאשר ניגשים למנוע LLM ציבורי מקפיצים חלון HTML ובו אזהרות מתאימות ועיקר המדיניות הארגונית בתחום.

נדבך נוסף בהגנת GenAI הוא מניעת DDOS (או לייתר דיוק DOS) אשר לא רק יכול לגרום לאיטיות בשירות כמו DOS רגיל אלא יכול לגרום לארגון לנזק כספי כי השימוש ב- GenAI גורם לחיוב של הארגון. חלק מהפתרונות התחילו לדבר על הרחבת הנושא הכלכלי בהקשר של cost optimization – בכך שהם מבצעים ניתוב (או מונעים ניתוב) של הבקשות לפי קריטריונים כלכליים ("בקשות של מחלקת הדרכה ינותבו ל- LLM זול יותר).

נקודה מאתגרת נוספת שעלתה בדיוק היא בדיקת יעילות ההגנה על מנוע GenAI. כי הרי סביבות בדיקה מלאות ל- GenAI עדיין לא קיימות ברוב הארגונית ואז נשאלת השאלה איך מבצעים PT על מערכת חיה? עוד נקודה מאתגרת היא ביצוע PT למערכות Voice אותם בוטים אשר מדברים עם הלקוח.

נקודה נוספת שעלתה היא השימוש בטלפונים חכמים. לעומת ה- PC ששם יש מערכות הגנה בשלות, בטלפון החכם ישנה הרגשה שהדברים פחות סגורים ויותר קשה לקבוע מדיניות מפרידה לגמרי בין שימוש אישי לשימוש מקצועי.

נקודה נוספת שעלתה בדיון שבחלק מהבדיקות גילו שמוצרי ההגנה בתחום GenAI גורמים לאיטיות – יש לבצע בדיקה יסודית בתחום זה.