מדריך לניהול סיכונים ושימוש אחראי בכלי בינה מלאכותית (AI) במגזר הציבורי

גרסה להערות ציבור

*האיור נוצר באמצעות כלי Image Creator של Microsoft Designer

מערך הדיגיטל הלאומי
نظام الديجيتال الوطني
Israel National Digital Agency

משרד המשפטים
وزارة العدل | MINISTRY OF JUSTICE

ייעוץ וחקיקה
OFFICE OF LEGAL COUNSEL AND LEGISLATIVE AFFAIRS
المشورة والتشريع

משרד החדשנות, המדע והטכנולוגיה
Ministry of Innovation, Science & Technology

יוני 2025

מדריך לניהול סיכונים ושימוש אחראי בכלי בינה מלאכותית (AI) במגזר הציבורי

גרסה להערות ציבור

תוכן עניינים

1 מבוא

1.1 רקע

בשנים האחרונות הולך ומתרחב השימוש בטכנולוגיות בינה מלאכותית בארץ ובעולם במגזר הציבורי ובמגזר הפרטי. השקת ChatGPT על ידי OpenAI לציבור הרחב בשלהי שנת 2022 אפשרה לראשונה שימוש פשוט ונגיש ביישומי בינה מלאכותית יוצרת, ומאז הטכנולוגיה ממשיכה להתפתח בקצב מואץ, לרבות פיתוח של "סוכני AI" אשר מסוגלים לבצע שורה של פעולות מורכבות על פי הנחיות מפורטות.

בטכנולוגיות הבינה המלאכותית טמון פוטנציאל רב לצמיחת המשק, לשיפור הפריון, להעלאת רמת החיים ולייעול המערכת הציבורית. נוסף על כך, קיים פוטנציאל משמעותי בקידום חדשנות מבוססת בינה מלאכותית במגזר הציבורי ובתעשייה, לשיפור התשתיות, לשיפור השירותים הממשלתיים והשלטון המקומי לכל אזרחי המדינה, לעידוד צמיחה, לפיתוח בר-קיימא, לרווחה חברתית ולקידום המובילות של מדינת ישראל בתחום החדשנות והטכנולוגיה.

אולם לצד הפוטנציאל, שימושי בינה מלאכותית טומנים בחובם אתגרים וסיכונים מסוגים שונים - תפעוליים, חברתיים, ועוד - וכן מעוררים סוגיות משפטיות שונות, זאת בפרט בשימוש במגזר הציבורי. על מנת לממש את הפוטנציאל העצום, נדרש להתמודד עם האתגרים ולנהל את הסיכונים הנובעים מהשימוש בטכנולוגיות אלה באופן מושכל ושיטתי.

מדריך זה הוא המסמך הממשלתי המקצועי הראשון שמציע שיטות עבודה מומלצות לגופים ציבוריים המבקשים לשלב מערכות בינה מלאכותית בתחומי פעילותם, בדגש על היבטים ארגוניים, טכנולוגיים ועסקיים, תוך התוויית הליך להערכה, לניהול ולמזעור סיכונים. כך, ולצמצם את חוסר הוודאות שנלווה ליישום מושכל של טכנולוגיה זו. הטמעת שיטות העבודה המוצעות עשוי גם לסייע לארגונים בקבלת הסמכה (התעדה) לתקנים בינלאומיים.

המדריך גובש בשיתוף בין מערך הדיגיטל הלאומי, מחלקת ייעוץ וחקיקה במשרד המשפטים והמרכז לרגולציה ומדיניות בינה מלאכותית במשרד החדשנות, המדע הטכנולוגיה. גיבוש המדריך כלל גם סבב התייעצויות עם מגוון גורמים מגופי המגזר הציבורי, לרבות מערך הסייבר הלאומי, הרשות להגנת הפרטיות, החשב הכללי, מנהל הרכש, מספר מובילי דאטה משרדיים (CDO) וכן עם מומחים ומקבילים ממדינות מובילות בתחום.

1.2 קהל יעד

המדריך מיועד לגופי המגזר הציבורי השונים המבקשים לשלב בינה מלאכותית בפעילותם ובתהליכים אותם הם מנהלים. המדריך מתייחס **לתפקידים ולאחריות** של הגורמים העיקריים בתהליך, וביחס לכל אחד מהם, מפורטים קווים מנחים לתהליך ניהול הסיכונים שעליהם לבצע. נוסף על כך, מדריך זה כולל קווים מנחים **למשתמשי קצה**, המיועדים לעובדים מקרב ארגוני המגזר הציבורי העושים שימוש בכלי בינה מלאכותית.

1.3 העקרונות בבניית המדריך

הגישה הכללית על פיה נבנה המדריך היא עידוד השימוש בבינה מלאכותית, בתהליך סדור, תוך הפעלת שיקולי תועלת וניהול סיכונים. מטרתו להוות כלי משמעותי לתמיכה בהחלטות על אופי שילוב בינה מלאכותית בתהליכים ופעולות בארגון. המדריך מבוסס על מספר עקרונות:

- **תפיסת ניהול סיכונים** – שימוש אחראי במערכות בינה מלאכותית אין משמעותו הימנעות מוחלטת מסיכונים כלשהם. המדריך מציע גישת ניהול סיכונים דיפרנציאלית: ככל שהסיכונים גבוהים יותר, כך גם נדרשים אמצעי הפחתת סיכונים ותהליכי בקרה מקיפים יותר.
- **כלליות וגנריות** – המדריך מהווה המלצה כללית שעליה ניתן להוסיף שיטות ניהול סיכונים ולבצע את ההתאמות הרלוונטיות, בהתאם לתחום התוכן שבו מופעלת המערכת או שבו פועל הארגון.
- **התווית תהליך** (ולא תוצאה) – בכל הנוגע לתהליך ניהול הסיכונים, המדריך מתווה תהליך, ואינו מתווה הנחיות לרף תוצאתי מסוים של התממשות סיכון מקובלת.

- **דינמיות** – המדריך יתעדכן באופן תקופתי, בהתאם לצרכים של גופים ציבוריים, להתפתחויות הטכנולוגיות ולשינויים במסגרת הנורמטיבית בישראל ובעולם, וכן לפרקטיקות מיטביות.

- **התאמה לעקרונות המדיניות הממשלתית** – המדריך מתבסס על הנחיות מערך הדיגיטל בדבר ניהול סיכוני תקשוב.[1] כמו כן, המדריך מתבסס ומתכתב עם מסמך "עקרונות מדיניות, רגולציה ואתיקה בתחום הבינה המלאכותית", שהובן על ידי משרד החדשנות, המדע והטכנולוגיה ומחלקת ייעוץ וחקיקה במשרד המשפטים ("**מסמך עקרונות אסדרת בינה מלאכותית**"). במסמך זה נסקרו סיכונים ואתגרים המתעוררים עקב פיתוח ושימוש במערכות בינה מלאכותית, וכן פורטו המלצות למדיניות אתיקה ורגולציה בתחום הבינה המלאכותית בישראל.[2]

- **התחשבות במסגרות אסדרה וסטנדרטים מובילים** – המדריך מתחשב במתודות לניהול סיכוני בינה מלאכותית המפורטות בתקינה בינלאומית רלוונטית. נוסף על כך, נלקחו בחשבון מסגרות מקובלות במישור הבינלאומי לרבות: (1) עקרונות ה-OECD בנושא שימושי אחראי בינה מלאכותית; (2) אמנת מועצת אירופה (Council of Europe) בנושא בינה מלאכותית וזכויות אדם, דמוקרטיה ושלטון החוק[3] (אמנת CAI); ו-(3) המתודולוגיה הנלווית לאמנה – מתודולוגית HUDERIA; (4) הנחיות ומדריכים דומים במסגרות המדיניות של ארה"ב, בריטניה וקנדה.

- **דגש על שימוש בסביבת הענן הממשלתית** – לממשלת ישראל יש היצע שירותים רחב מאוד של כלי בינה מלאכותית, הזמין הן ישירות על ידי ספקיות הענן בפרויקט "נימבוס" (רובד אחד), והן על ידי ארגונים נוספים המספקים מוצרים המותקנים בסביבות של ספקיות הענן (מוצרי רובד חמש),[4] בכלל זאת גם יישומי בינה מלאכותית יוצרת, מהמתקדמים בעולם, המותקנים באתרים בישראל בסביבה מאובטחת. שירותים אלו נרכשו תוך הקפדה על הדין הישראלי, בהתאם לתנאי השימוש שנקבעו במכרז "נימבוס" וגובשו לפי דרישות משרדי הממשלה. לפיכך, ישנה העדפה לעשות שימוש ביישומי בינה מלאכותית בסביבות אלו, המשתקפת במדריך זה.

1.4 מה כולל המדריך?

למדריך שני חלקים מרכזיים – החלק הראשון מתאר את העקרונות הבסיסיים לשימוש אחראי בבינה מלאכותית, אשר מוצע שהארגון יטמיע. החלק השני מתאר התפיסה הארגונית לשימוש אחראי בבינה מלאכותית כולל הגורמים המעורבים ותחומי אחריות מוצעים עבורם. שני תפקידים מרכזיים, לפי תפיסה זו, הם "**אחראי משילות בינה מלאכותית**", אשר אמון על קביעת ותפעול מדיניות ארגונית לשימוש אחראי, ו"**אחראי יישום עסקי**" אשר מוביל תהליכי הטמעת מערכת בינה מלאכותית בארגון לצרכים עסקיים ספציפיים.

המדריך כולל מספר נספחים: פירוט תפקידי אחראי משילות בינה מלאכותית ואחראי יישום עסקי, קווים מנחים למשתמשי קצה, ומתודת ניהול סיכונים מוצעת. כמו כן, ישנו פירוט של תקינה בינלאומית ומילון מונחים. המדריך לא נועד להתייחס לחובות המשפטיות שחלות על רשויות ציבוריות ביחס לשימוש במערכות בינה מלאכותית. בהמשך יצורף מדריך משפטי, בהובלת מחלקת ייעוץ וחקיקה במשרד המשפטים, שנועד ללוות לשכות משפטיות של גופים ציבוריים.

מערך הדיגיטל הלאומי מציע שירותים לסיוע וליווי בהטמעת שימוש אחראי בבינה מלאכותית במגזר הציבורי, כגון הכשרות באמצעות בית הספר להכשרות דיגיטליות "הדיגיטלית", שירות הטמעת תהליך ניהול סיכונים ומרכז מומחים. כמו כן, מערך הדיגיטל יפעל עם הגופים השונים כדי לשפר את המדריך בהתאם לצרכים שיעלו.

המדריך מנוסח בלשון זכר מטעמי נוחות בלבד, וכל האמור בו חל על כל המגדרים באופן שווה ומכבד.

גרסה זו פתוחה להערות ציבור. ניתן לפנות אלינו בכל התייחסות, הערה ושאלה לתיבת הדוא"ל ResponsibleAI@digital.gov.il.

---

[1] כולל הנחיות נוספות ובפרט הנחיית יחידת הגנת הסייבר בממשלה (יה"ב) בדבר שימוש מאובטח בצ'אט מבוסס בינה מלאכותית.

[2] בין היתר, המסמך ממליץ על גישה של רגולציה סקטוריאלית (בניגוד לרגולציה רוחבית וגורפת). להמחשת גישה זו, ראו דוח הביניים של הצוות הבין משרדי לבחינת אסדרת בינה מלאכותית בסקטור הפיננסי.

[3] האמנה נועדה להתמודד בין היתר עם אתגרים המתעוררים לאורך מחזור החיים של מערכות בינה מלאכותית. היא חלה בעיקר על שימושים בבינה מלאכותית על ידי המגזר הציבורי, אך גם מחייבת מדינות שהן צד לה להתייחס לסיכונים הנשקפים לזכויות אדם, לדמוקרטיה ולשלטון החוק, משימושי בינה מלאכותית על ידי המגזר הפרטי. מדינת ישראל הייתה שותפה למו"מ שהתנהל לגיבוש האמנה וחתמה עליה ב-5.9.2024. ישראל טרם אשררה את האמנה כך שהיא לא מחייבת באופן פורמלי.

[4] קטלוג המוצרים המאושרים לרכישה ברובד 5 זמין כאן.

2 שימוש אחראי בארגון

גישת "שימוש אחראי בבינה מלאכותית" היא גישה אשר מעודדת ארגונים שמבקשים ליישם בינה מלאכותית לעשות זאת בצורה מושכלת, עם הסתכלות רוחבית שכוללת את כל מחזור החיים של המוצר, החל משלבי הפיתוח ועד השימוש בו על ידי משתמש קצה, שמתייחסת בין היתר להשלכות הצפויות שלו, לחיוב ולשלילה.

את תפיסה זו תיאר ארגון ה-OECD, בהמלצותיו בנושא, המכונות "responsible stewardship of trustworthy AI", אשר גובשו ב-2019 ועודכנו בשנת 2024. ההמלצות כוללות מספר עקרונות (לא מחייבים), לשימוש אחראי בבינה מלאכותית. להלן סיכום העקרונות:[5]

1. **בינה מלאכותית לצמיחה, פיתוח בר קיימא ורווחת הכלל –** מדובר בפעולות יזומות כדי לקדם בינה מלאכותית שבצידה תועלות לאדם ולסביבה, כגון חיזוק כישורים ויכולות אנושיות, התמודדות עם הדרה של אוכלוסיות מסוימות, צמצום פערים והפחתת אי-שוויון, הגנה על הסביבה, ייעול תהליכי תכנון ובנייה, ועוד.

2. **כיבוד שלטון החוק, זכויות אדם וערכים דמוקרטיים, לרבות הוגנות ופרטיות** – על הפיתוח, ההטמעה והשימוש במערכות בינה מלאכותית להיעשות בהתאם לערכים דמוקרטיים ולשלטון החוק, ובאופן המכבד זכויות אדם (ובהן – כבוד האדם, שוויון, פרטיות ואוטונומיה). נוסף על כך, חשוב לנקוט אמצעים להתמודדות עם תופעות המושפעות מבינה מלאכותית כגון דיסאינפורמציה. לשם עמידה בעקרונות אלו, על הגורמים הרלוונטיים ליישם מנגנונים מתאימים, כגון מעורבות ופיקוח אנושי וניהול סיכונים – והכל בהתאם לדין, להקשר ובהתאם ליכולות הטכנולוגיות הזמינות.

3. **שקיפות והסברתיות** – מדובר בשקיפות כלפי הציבור ביחס למערכות בינה מלאכותית שבאחריות הארגון. שקיפות עשויה לכלול מידע מהותי, שבין השאר, יבהיר את פעולת המערכת, לרבות יכולותיה ומגבלותיה, יגביר מודעות למצבים בהם מתקיימת אינטראקציה עם בינה מלאכותית, יספק מידע על מקורות המידע ועל ההיגיון שבבסיס המערכת, ולבסוף יספק גם מידע שיאפשר להתמודד עם התוצאות השליליות של המערכת.

4. **ביטחון ובטיחות של המערכת** – בפיתוח ובשימוש במערכות בינה מלאכותית, חשוב להקפיד על כך שיהיו אמינות ובטוחות לאורך כל מחזור החיים שלהן. זאת על מנת שבמצבי שימוש מתוכננים, וכאלה שאינם מתוכננים, שימוש שגוי או היווצרות של תנאים חיצוניים מסוכנים, המערכות יפעלו כראוי ולא יהוו סיכון בטיחותי או בטחוני בלתי סביר. כמו כן, נדרש שיהיו מנגנונים להתמודד עם מצבי סיכון שנגרמו בגלל המערכת.

5. **אחריותיות** – מצופה מהארגונים לגלות אחריות לתפקודה התקין של מערכת בינה מלאכותית, ולקיום העקרונות האמורים, בהתאם לתפקידם ובכפוף לאפשרויות הטכנולוגיות הזמינות. לשם כך, יש לפתח מנגנוני ניהול סיכונים ולאמץ כללי התנהלות פנימיים להתמודדות עם סיכונים, (לרבות סיכונים חברתיים, כגון הטיות, פרטיות, זכויות עובדים וכן סיכון לפגיעה בקניין רוחני, בטיחות ועוד).

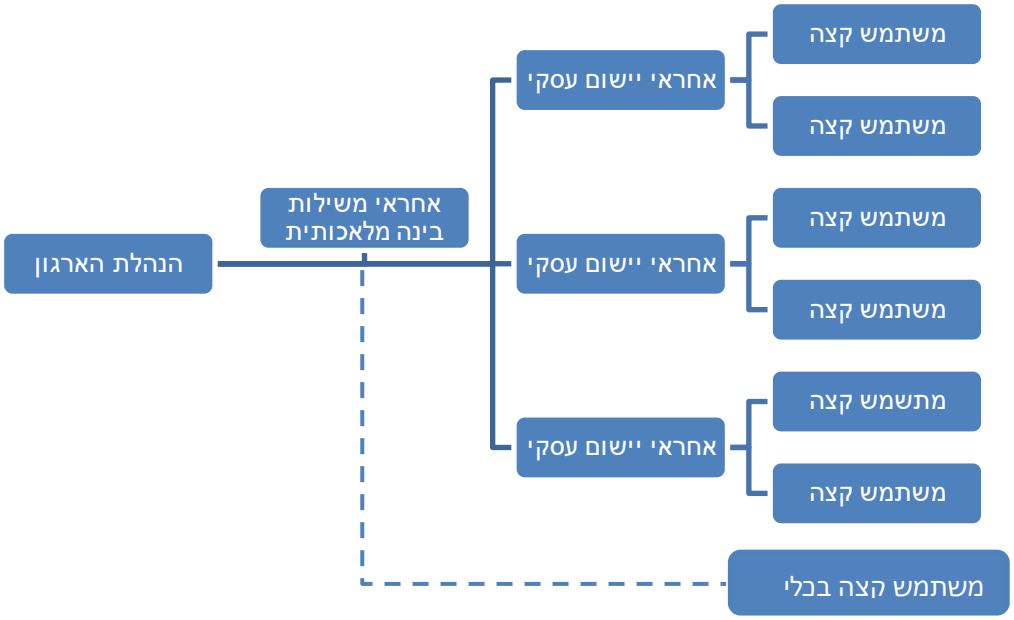
מסמך עקרונות אסדרת בינה מלאכותית מבוסס על אותם עקרונות, עם התאמות קלות. ראו גם אזכור של העקרונות הללו באסטרטגיית בינה מלאכותית של פורום תל"ם (דו"ח 2025).

על פי עקרונות אלו, שימוש אחראי בבינה מלאכותית אינו יישום של מתודת ניהול סיכונים בלבד (כ"צ'ק-ליסט"), אלא שהוא מתבטא, בראש ובראשונה, בהפנמה והטמעה **ברמה המוסדית**, של הערכים המדוברים. לפי תפיסה זו, מושרשות בארגון כולו מודעות גבוהה להשלכותיה של בינה מלאכותית, אוריינות והבנה של הטכנולוגיה, ו"לקיחת בעלות" של הארגון על תהליכי פיתוח ויישום של מערכות בינה מלאכותית. תפיסת שימוש אחראי בבינה מלאכותית נשענת, אפוא, על הקמת תהליכים וחלוקת תפקידים ותחומי אחריות בארגון המכוונים למטרה זו. גרסאות עתידיות של המדריך ירחיבו על העקרונות הנ"ל.

[5] אין מדובר בתרגום מילולי של העקרונות.

3 ממשל בינה מלאכותית, תפקידי מפתח ותחומי אחריות

המבנה המוצע מורכב מארבע שכבות של שחקנים עיקריים, כמפורט להלן:



3.1 הנהלת הארגון

הובלת תהליכי שימוש אחראי לבינה מלאכותית היא באחריות הנהלת המשרד, אשר מצופה לספק את המעטפת התפעולית והתקציבית המתאימה. לשם כך, מוצע כי ההנהלה, בין היתר, **תמנה את אחראי משילות בינה מלאכותית של הארגון** ותסמיך אותו לבצע את הפעולות המתוארות כאן, תייצר תהליכי שיתוף ציבור, ותנגיש דו"חות תקופתיים של פעילות הארגון בתחום.

3.2 אחראי משילות בינה מלאכותית

אחראי משילות בינה מלאכותית הוא האחראי ברמה הארגונית על קביעה והפעלת הנהלים לשימוש אחראי בבינה מלאכותית. מדריך זה אינו קובע באופן גורף איזה גורם צריך למלא את תפקיד זה. מוצע, כברירת מחדל, למנות את CDO (Chief Data Officer), אך אין מניעה למנות גורם אחר (למשל מנמ"ר או סמנכ"ל אסטרטגיה ומדיניות) ואף ועדה משרדית או ועדה חיצונית. העיקר הוא כי אחראי משילות בינה מלאכותית יהיה גורם בעל ידע רלוונטי בנושא שימוש אחראי בבינה מלאכותית ושתינתן לו התמיכה והסכמה פנים-ארגונית שיאפשרו לו למלא את תפקידיו.

למידע ופרטים על תפקיד **האחראי משילות בינה מלאכותית**, ראו נספח א.

3.3 אחראי יישום עסקי

אחראי היישום העסקיים, הם הגורמים העסקיים בארגון שמבקשים להטמיע בינה מלאכותית בתחום פעילותם. כך למשל, אחראי יישום עסקי יכול להיות מנהל אגף האמון על אסדרת תחום מסוים, יו"ר ועדה האמונה על מתן הטבות או תמיכות, גורם מקצועי האחראי על תהליך תכנון, תקצוב ובקרה פנימית וכדומה. הפעילות שלהם. למידע ופרטים על תפקיד ה**אחראי יישום עסקי**, ראו נספח ב.

3.4 משתמשי קצה

משתמשי קצה הם עובדים בגוף ציבורי אשר עושים שימוש במערכת בינה מלאכותית לצורך מילוי תפקידם. בין אם מדובר במערכת שפועלת בסביבה המיחשובית הסגורה של הארגון (בגון סביבת הענן הארגונית) שהונגשה להם באופן ייעודי לצורך מילוי תפקידם, ובין אם מדובר בכלי מדף הזמינים לקהל הרחב מצופה ממשתמשי קצה גם הם לפעול באופן אחראי ושקול. לקווים מנחים עבור **משתמשי קצה** בארגון, ראו נספח ג.

נספח א – תפקידי האחראי למשילות בינה מלאכותית

עיקר התפקיד: קביעה והפעלה של המדיניות הארגונית בהיבטי משילות, מנגנוני ניהול שוטפים ודיווח.

להלן טבלה אשר מתארת את הפעולות המומלצות, המחולקות לפיתוח מדיניות, הטמעת המדיניות ותפעול.

פעולות מומלצת	פירוט
1 - פיתוח מדיניות	
1.1 שימוש אחראי	קביעת מדיניות כללית לשימוש אחראי בבינה מלאכותית. מומלץ כי המדיניות הפנימית, במלואה או בחלקה בהתאמות הנדרשות לפעילות הארגון, תתבסס על העקרונות לשימוש אחראי (פרק 2 למדריך), וכן על תהליך ניהול הסיכונים המפורט במדריך זה (נספח ד'). המדיניות תתייחס לתהליכים פנים-ארגוניים ובניית מנגנונים שיאפשרו התייעצות, תיאום ויישום המדיניות. המדיניות תהיה תואמת למתודת ניהול הסיכונים שבמדריך זה (נספח ד).
1.2 משילות דאטה המשמש לבינה מלאכותית	קידום פיתוח כללים לניהול נתונים, ספציפית, עבור נתונים ארגוניים המשמשים מערכות ומודלי ה-AI. זאת כדי לתת מענה מקדים להיבטים וסיכונים נבדלים, ובכללם: (1) זכויות ואינטרסים של נושאי המידע ובפרט הזכות לפרטיות, (2) הצורך לאמן את המודלים על נתונים מגוונים, מייצגים, ועדכניים כדי לצמצם חשש להטיות אלגוריתמיות, פלטים שגויים ואף פוגעניים. אם יש שימוש בכלי מדף לתהליכים אופרטיביים, יש להתייחס לצורך לשמור על המידע הארגוני כדי לוודא רציפות תפקודית. יש לשאוף לצרוך את השירותים של כלי המדף כמשתמשים ארגוניים ולא פרטיים, ולייצר את המענים הארגוניים לשמירת המידע והידע הארגוני ותהליכי למצבים שבהם גורם המנהל את התהליך עוזב את הארגון או משנה תפקיד.
1.3 מדיניות כלי מדף מבוססי AI ומערכות AI בארגון	קביעת מדיניות לגבי המוצרים והיישומים המותרים לשימוש בארגון, בתיאום עם הגורמים הרלוונטיים בארגון כגון ממונה הגנה בסייבר, DPO, חשבות וכדומה. קידום פעולות לרכש שירותי בינה מלאכותית בסביבות הענן הממשלתי על מנת לאפשר שימוש מאובטח יותר בטכנולוגיה זו (לעומת יישומים "ציבוריים" הפתוחים לכלל). כאשר יש שימוש בכלי מדף בנימבוס, יש לבצע את האמור בהודעות התכ"ם הרלוונטיות;[6] כאשר יש שימוש בכלי מדף שאינו בנימבוס, יש לבצע את האמור בהודעת תכ"ם 16.12.1.2 - רכש שירותי צד ג' בענן בהליך רכש עצמאי של המזמין. בתיאום עם הגורם האמון על הגנה בסייבר בארגון, יש לבחון האם יש כלי מדף מבוססי AI שרצוי להנחות על איסור גורף של השימוש בהם בארגון, או להגביל את השימוש בהם לצרכים ספציפיים, ולהודיע על כך בתוך הארגון. כמו כן, מוצע לבחון האם ניתן ורצוי לחסום את השימוש בהם באופן מרכזי, בהתאם לשיקולי אבטחת מידע והגנה בסייבר ושיקולים הנוגעים ליישומים אשר פותחו או פועלים במדינות שאינן דמוקרטיות, תוך התייעצות עם הגורמים הרלוונטיים בתחום בארגון.

[6] כאשר מדובר שירותי בינה מלאכותית המוצעים על ידי AWS ו-Google, כמפורט בהודעת תכ"ם 16.12.2 "אספקת שירותי ענן ציבורי של AWS ו- Google למשרדי הממשלה"; כאשר מדובר בשירותי בינה מלאכותית המוצעים על ידי חברת Salesforce במסגרת מכרז מרכזי, כמפורט בהודעת תכ"ם 16.2.4 "אספקת שירותי Customer Relationship Management (CRM) בענן"; כאשר מדובר בשירותי בינה מלאכותית המוצעים במסגרת נימבוס על ידי ספקי צד ג', כמפורט בהודעת תכ"ם 16.2.4 "אספקת שירותי Customer Relationship Management (CRM) בענן". הוראת תכ"ם 7.10.7 "התקשרות לרכישת שירותי בינה מלאכותית".

פעולות מומלצת	פירוט
1.4 מדיניות ניהול הרשאות	גיבוש מדיניות ניהול השימוש וההרשאות למשתמשי הארגון בכלי AI חיצוניים ופנימיים.
2 - הטמעת המדיניות	
2.1 - הטמעת כללים לניהול הנתונים המשמשים בינה מלאכותית ברמה הארגונית	קידום תהליכי הטמעה של הכללים לניהול נתונים (שפותחו בהתאם לסעיף 1.2 לעיל), ספציפית, עבור נתונים ארגוניים המשמשים מערכות ומודלי AI.
2.2 הובלת תהליכי למידה ארגונית	העמקת הידע בקרב משתמשי הארגון בהיבט של תועלות מול סיכונים בכלי AI, באמצעות קביעת תוכנית למידה ארגונית שתכלול השתלמויות והדרכות פנימיות וחיצוניות (לרבות הדרכות שניתנות על ידי "הדיגיטלית" של מערך הדיגיטל הלאומי) עבור גורמים עסקיים ומשתמשי קצה.
3 - תפעול המדיניות ברמה הארגונית	
3.1 שגרות דיפרנציאליות לפי רמת הסיכון	**מערכות שלא זוהה לגביהן סיכון** ביצוע בדיקה מדגמית שנתית עם אחראי היישום העסקיים של המערכות על מנת לוודא שאכן אין סיכון בשימוש במערכות אלו.
	מערכות שזוהו לגביהן סיכונים נמוכים קבלת דיווח מאחראי היישום העסקיים אחת לשנה.
	מערכות שזוהו לגביהן סיכונים בינוניים 1) אישור הפעלת המערכות בהתאם לניהול הסיכונים המוצג על ידי אחראי היישום העסקיים; 2) קבלת דיווח בחינת סיכונים חוזרת אחת לחציון לכל הפחות; 3) קבלת דיווחים מיידיים בעת אינדיקציות להתממשות סיכון.
	מערכות שזוהו לגביהן סיכונים גבוהים 1) אישור מערכות בסיכון גבוה למול תוכנית להפחתת השפעות ומעקב אחת לרבעון; 2) קבלת דיווח בחינת סיכונים חוזרת אחת לחציון לכל הפחות; 3) קבלת דיווחים מיידיים בעת אינדיקציות להתממשות סיכון.
3.2 יצירת תמונת מצב עדכנית לגבי מיפוי הסיכונים ודירוגם	יצירת תמונת מצב עדכנית לגבי מיפוי הסיכונים ודירוגם, על פי המערכות הארגוניות שעושות שימוש ב-AI; וכן לגבי הניטור המתבצע על ידי אחראי היישום העסקי והגורמים האמונים על התחומים השונים בארגון, כגון פעולות לזיהוי דלף מידע שעלול להיות מוזן לכלי AI המותקנים מחוץ למערכות הארגון. ביצירת תמונת המצב יש לזהות האם יחסי הגומלין בין מערכות AI בארגון עלולים לייצר סיכונים נוספים או מוגברים באופן סינרגטי שלילי.
3.3 טיפול במשברים ותקריות AI	קביעת דרכי התמודדות וניהול תקריות בינה מלאכותית.

פעולות מומלצת	פירוט
3.4 שקיפות	לפרסם באתר הארגון מידע על מערכות בינה מלאכותית שנמצאות בשימוש הארגון, בהתאם למדיניות לשימוש אחראי ולתהליך ניהול סיכונים (ראו פירוט בטבלת מתודת ניהול הסיכונים).[7]
3.5 דיווח במקרה תקרית AI	דיווח להנהלת הארגון ולגורמים הרלוונטיים בתוך הארגון (למשל, ממונה הגנת פרטיות) בעת התממשות של סיכון AI. בפרט, בעת התממשות של סיכוני אבטחת מידע ובכל חשש לדלף מידע או בעיית אבטחה, הנגרמים בשל השימוש במערכת AI, יש לדווח כנדרש לגופים המנחים הרלוונטיים (כדוגמת מערך הסייבר, הרשות להגנת הפרטיות, ויחידת הגנת הסייבר במערך הדיגיטל).
3.6 תחקור והפקת לקחים	בחינה פרואקטיבית ומקיפה של תקריות AI, לרבות שחזור ההחלטות השונות שתועדו, הקשורות לאפיון ותפעול המערכת, לרבות האופן בו נוהלו הסיכונים. הפקת לקחים – לפי תוצאות התחקיר: (1) בחינת האפשרות של שינויים באופן בו מופעלת המערכת ומנוהלים הסיכונים; (2) במקרה של תקריות חמורות, הבחינה תתייחס לאפשרות של הפסקה (זמנית או קבועה) של השימוש במערכת; (3) במקרים שבהם נגרם נזק משמעותי או חמור, להתייעץ עם הנהלת הארגון ולפעול ליידוע הציבור והנפגעים.
3.7 דיווח פנימי	תיעוד ודיווח שנתי למנכ"ל או פורום הנהלה על מערכות משולבות בינה מלאכותית שבשימוש המשרד, התועלות והסיכונים שלהן, תהליכי ניהול סיכונים ובקרה שאומצו לגביהן, ותקריות AI ככל שקרו. מתן המלצות לשיפור תהליכים ארגוניים לשימוש אחראי בבינה מלאכותית.

[7] נוסף על כך, מערכות בינה מלאכותית עתידות להתפרסם במערכת AI Watch שהקים מערך הדיגיטל הלאומי, אשר נועדה להוות מרשם כולל ולייצר תמונת מצב על השימוש בבינה מלאכותית בארגונים השונים, עבור גורמי ממשל ותושבים.

נספח ב – תפקידי אחראי היישום העסקי

עיקר התפקיד: להוביל את האפיון וההטמעה של מערכת משולבת בינה מלאכותית תוך ניהול סיכונים ושיטתי.

להלן טבלה אשר מתארת את הפעולות המומלצות. הן כוללות הגדרת הצורך המקצועי והדרישות מהמערכת, ניהול סיכונים ומתן הוראות שימוש למשתמשי קצה.

פעולות מומלצת	פירוט
1 - הגדרת הצורך המקצועי והדרישות מהמערכת	
1.1 איסוף מידע	זיהוי הצורך העסקי למערכת בינה מלאכותית (התייעלות, דיוק בקבלת החלטות, שיפור ניתוח דאטה, מתן שירותים לאזרח וכו'). בחינת המצב הקיים אל מול החלופות (כלי בינה מלאכותית וכן כלים אחרים היכולים לתת מענה לצורך העסקי). בחינה ראשונית של כלי בינה מלאכותית קיימים שעשויים למלא את הצורך העסקי, בסיוע ה-CDO וגורמים רלבנטיים מאגף טד"מ. הגדרה ותיעוד של הצורך העסקי ושל הבחינה הראשונית שבוצעה (חלופות, כלים רלוונטיים). **דגש:** לעיתים, ישנם כלים שאינם כלי בינה מלאכותית אשר עשויים לתת מענה לצרכים באופן יעיל ואפקטיבי. על רקע זה, חשוב, כבר בתחילת התהליך, לזקק את הצורך והתועלת העסקיים שעשויה להביא מערכת משולבת בינה מלאכותית.
1.2 התייעצות	שילוב "גורמי המעטפת" בארגון כגון DPO, CDO, מנהל אגף טד"ם, מנהל הפרויקט/מערכת/מוצר, מנהל יישומים, אחראי ענן, ממונה אבטחת מידע, ולקוחות. בכלל זאת, יש לשלב את הלשכה המשפטית בתהליך, בהתאם לאופי הפרויקט, רצוי בשלב מקדמי ככל הניתן על מנת להבין את התמונה המשפטית. התייעצות עמיתים בארגונים ציבוריים אחרים אשר מתמודדים עם צורך עסקי זהה או דומה. זאת על מנת ללמוד מניסיונם, כולל התועלות והסיכונים של המערכת. תיעוד ממצאי ההתייעצויות.
1.3 אפיון הצרכים העסקיים	בהתאם לתהליך הנ"ל, ככל שיוחלט להתקדם לשילוב בינה מלאכותית במערכת, יש לאפיין בצורה מדויקת את הצורך העסקי ואת המענה שהמערכת תספק.
2 – ניהול סיכונים	הפעלת תהליך ניהול סיכונים המפורט בנספח ד. תיעוד כלל הממצאים (מיפוי התועלות והסיכונים של מערכת ספציפית באופן מקיף, מפורט ובהיר, גיבוש תוכנית הפחתת סיכונים).
3 – הוראות למשתמשי קצה	כאשר מדובר בכלי שהוטמע מהענן הממשלתי, יש לבקש מאחראי משילות בינה מלאכותית או מהספק, לפי העניין, את ההנחיות הרלוונטיות למשתמשי קצה לגבי השימוש האחראי והמושכל במערכת, ולהנגיש אותם למשתמשי הארגון. ככל שלא ניתן לספק הנחיות כאלו, יש להעביר למשתמשי הקצה את המידע הכלול במדריך למשתמש קצה (נספח ג), בתור הנחיות בסיס, ולהשלים אותם בהתאם לאפיון המוצר ולתובנות שיתקבלו בהתאם להליך ניהול סיכונים (נספח ד).

נספח ג - מדריך למשתמש קצה

פרק זה מיועד עבור משתמשי הקצה במערכות מבוססות AI, קרי כל עובד במגזר הציבורי העושה שימוש ביישומי בינה מלאכותית לרבות בינה מלאכותית יוצרת. הוא כולל המלצות כלליות לשימוש אחראי במערכות בינה מלאכותית במסגרת פעילות המגזר הציבורי. חלק מההמלצות רלוונטיות לשימוש בכלי מדף הפתוחים לציבור הרחב, וחלקן רלוונטיות גם לשימוש בכלי מדף הזמינים לארגון, ואף ביישומים ייעודיים אשר פותחו ספציפית עבור הארגון.[8]

בכלל, כאשר מדובר במערכת בינה מלאכותית שהונגשה לעובדים על ידי הארגון, המשתמש יקבל הוראות שימוש שמתייחסות לנושאים הבאים(וכן הנחיות נוספות) מ האחראי למשילות בינה מלאכותית ומאחראי היישום העסקי הרלבנטי. האמור להלן נועד לספק מסגרת כללית ומשלימה להוראות ולהנחיות, ככל הנדרש.

מדריך זה אינו כולל ניתוח של הסוגיות המשפטיות שמתעוררות ואת הדינים החלים על עובדי ציבור, אלא נועד להצביע על דגשים מרכזיים לשימוש במערכות בינה מלאכותית אשר מומלץ לפעול לאורם.

רקע: מאפיינים ומגבלות מערכות בינה מלאכותית

מערכות בינה מלאכותית רבות מתאפיינות במגבלות שונות שיש להיות מודעים אליהן בעת השימוש. להלן תיאור המגבלות העיקריות.

חלק מהמגבלות מאפיינות במיוחד מערכות בינה מלאכותית יוצרת (Generative AI), קרי, מערכות המסוגלות לייצר תוכן חדש – כגון טקסט, תמונה, קטעי שמע או וידאו – בהתבסס על דפוסים שנלמדו מנתונים קיימים (למשל: ChatGPT, Gemini או Midjourney). ישנם יישומים עם יכולות ייעודיות כגון יישומי "עוזר אישי" עם יכולת בינה מלאכותית יוצרת, מנועי חיפוש מבוססי שיח ותמלול וסיכום פגישות.

מודלי בינה מלאכותית נבדלים זה מזה, בין בשל תכנון שונה של המודלים, או אימון על מסדי נתונים שונים. לכל מודל יש את המאפיינים שלו לרבות החוזקות, החולשות והנטיות שלו. להלן יפורטו מספר מגבלות נפוצות שחשוב להכיר:

- **<u>"הזיות" - מגבלות באמינות ודיוק</u>**: קיימת נטייה של מערכות בינה מלאכותית יוצרת לספק מידע שאינו אמין ומדויק, וזאת לעיתים ללא הסתייגויות נדרשות. כך למשל, תוצאות של מודלים טקסטואליים מכילים לעיתים ציטוטי מקורות שאינם מדויקים ואף מומצאים.
- **<u>מוגבלות באובייקטיביות</u>**: מודלי AI אומנו על מסדי נתונים, המכילים מידע על גורמים אנושיים, ממוצאים תרבותיים מסוימים ולכן יכולים להיעדר הכללה תרבותית (inclusiveness), בפרט בנוגע למידע הקשור לאוכלוסיות מיעוט. בנושאים שנויים במחלוקת, תיתכן הצגה חסרה של מגוון נקודות מבט על ידי מערכות בינה מלאכותית ואף קידום של "עמדה" ספציפית.
- **<u>תוכן פוגעני ואפליה</u>**: אתגר מרכזי ביחס למערכות מבוססות בינה מלאכותית הוא הסיכונים לאפליה (discrimination) ולהטיות (biases) על ידי המערכות, שעשויים לנבוע מטעמים שונים, בהם למשל שימוש במאגרי מידע מוטים שאינם מייצגים מספיק, או שהמידע בהם משקף הטיות ואפליה קיימות בחברה. זאת שכן אלגוריתמים עלולים לעשות שימוש בנתוני השתייכות מפלים (לאום, מגדר ועוד) או בנתונים אחרים שיש להם קורלציה עם נתונים מפלים. היקף פעילותן הרחב של מערכות בינה מלאכותית מגביר את הסיכון להתרחשות תופעות אלה בקנה מידה רחב, בהשוואה להחלטה אנושית פרטנית.
- **<u>סכנה למידע מוגן:</u>**
  - **שימוש במידע המוגן בזכויות יוצרים**: המערכות עלולות לאסוף מהמשתמשים – או למסור להם – מידע שמוגן בזכויות קניין רוחני כגון זכויות יוצרים או סודיות מסחרית, ובכך לייצר סיכון להפרת זכויות ושימוש בחומרים ללא רשות, אשר גם חושפים את המדינה לתביעות.

[8] כפי שיפורט להלן, אין להזין מידע רגיש או מוגן לכלי הפועל מחוץ לסביבה הארגונית לרבות סביבת הענן הארגונית.

- **שימוש במידע אישי והפקת מידע מצטבר**: באופן כללי, כתלות בהגדרות מערכת הבינה המלאכותית, שימוש מתמשך או קבוע, עשוי להיווצר מצב בו בסביבת הבינה המלאכותית יאגר מידע רב שהוזן למערכת במסגרת השימוש. כך, ללא תשומת לב מספקת בקשר למידע המוזן למערכת על ידי המשתמש, עלול 'להיאגר' במערכת מידע רב מידע על הרשות הציבורית ופעילותה; מידע אישי הנוגע לנושאי מידע; מידע סודי עסקי; ומידע נוסף ('היסקים') שנוצר כתוצאה מהצלבה בין סוגי המידעים השונים שהוזנו למערכת ועובדו על ידה.

  למשל, במקרים רבים, במסגרת שימוש חינמי בכלים מבוססי AI, החברות המספקות יישומים אלו אוגרות כברירת מחדל את הנתונים המוזנים על ידי משתמשי הקצה (בפרומפטים) ונתונים אלו עלולים לשמש לאימון עתידי של המודלים וכן לצרכים מסחריים. ישנם יישומים מסוימים שמאפשרים מחיקה של המידע גם בשימוש חינמי, אך אין בכך להבטיח שלא נעשה שימוש במידע.

- **האנשה**: בחלק מהמודלים, בפרט במודלי שפה, המערכת מתקשרת עם המשתמש בשפה טבעית ואנושית. בשל מאפיין זה, משתמש עלול לייחס למערכת סמכות או מומחיות שאין לה במציאות, בפרט בנושאים רגישים, וכן לטעות ולחשוב שהמערכת "מבינה" את משמעות הדברים או שיש לה כוונה מוסרית, בזמן שבפועל מדובר בעיבוד הסתברויות של טקסט.

- **תלות אנושית**: שימוש בכלי AI עלול לייצר תלות אנושית בהם, ושחיקה של יכולות אנושיות לחשיבה יצירתית, אוטונומית וביקורתית, ולצד זאת גם שחיקה של כישורים הקשורים ליכולות ניסוח, תכנון, ניתוח, עיצוב וכיוצא בזה. תופעה זו עלולה להיות בעייתית במיוחד ביחס לתוצאות המתקבלות ממערכות AI שאמורות להיות תומכות החלטה אנושית.

קווים מנחים כלליים למשתמש הקצה

1. יש לוודא שכלי הבינה המלאכותית הרלוונטיים לא נאסרו לשימוש או לסוג השימוש המבוקש בו על ידי גורם הממונה על הגנת הסייבר בארגון או על ידי אחראי היישום העסקי או אחראי משילות בינה מלאכותית הארגוני.

2. כאשר מדובר ביישום של **בינה מלאכותית יוצרת**, יש לזהות האם מדובר בכלי מדף מבוססי AI **חיצוני** לארגון (פומבי או של ארגון אחר), או בכלי הפועל בסביבה **ייעודית** לארגון (למשל בענן הארגוני או במערכות המידע של הארגון).

   2.1. כאשר כלי המדף **פועל בסביבה חיצונית לארגון,** ולא **בסביבה ייעודית** לארגון, דוגמת שימוש בצ'ט בוט ברשת האינטרנט ברישיון פרטי:

      2.1.1. **אין** לעשות שימוש ביישומי בינה מלאכותית יוצרת אשר פותחו ופועלים במדינות שאינן דמוקרטיות. מומלץ לעשות שימוש ביישומים שפותחו במדינות המקדמות עקרונות של שימוש אחראי בבינה מלאכותית.

      2.1.2. **אין** להזין למערכות חיצוניות מידע מוגן, כגון מידע מזוהה אישי, מידע רגיש, מידע שאין למוסרו לפי חוק חופש המידע,[9] מידע שעלול להיות מוגן על ידי סודיות מסחרית או קניין רוחני, חיסיון משפטי או מידע מסווג.

      2.1.3. **אין** להזין מנחים (prompts) אשר מעידים על כוונה לביצוע פעולה כלפי פרט מזוהה מסוים, או כוונה לבצע צעד שלטוני רגיש מסוים.

      2.1.4. מכיוון שמדובר בטכנולוגיה יחסית חדשה ודינמית, וישנה גדילה משמעותית במגוון וכמות היישומים הזמינים לציבור הרחב ומכיוון שמדיניות השימוש במידע שונה בין היישומים השונים, מומלץ בעת שימוש ראשוני, לעיין במדיניות המערכת והוראות השימוש, ככל שישנן וזמינות, על מנת להבין את מגבלות הכלי ואת השימוש שנעשה במידע המוזן אליו וניהולו על ידי מנהל המערכת.

[9] למשל, ראו מידע שאין למסרו לפי סעיף 9 לחוק חופש המידע, תשנ"ח-1998.

2.1.5. **אין** לייצר באופן עצמאי ממשק אוטומטי בין כלי AI חיצוני לארגון לבין סביבת העבודה הארגונית, ללא התייעצות ואישור הגורמים הרלוונטיים בארגון כגון הגורם האמון על אבטחת המידע, ממונה הגנת הפרטיות (ה-DPO) או הגורם האחראי לניהול סיכוני AI.

2.2. כאשר הכלי **מופעל בסביבה התקשובית הארגונית (לרבות בענן הארגוני),** יש לפעול על פי מדיניות המוצר ותנאי השימוש שישוקפו למשתמשי הקצה על ידי אחראי היישום העסקי.

3. בעת השימוש בבינה מלאכותית בתהליכי קבלת החלטות, יש להתייחס לתוצאות המתקבלות כתומכות החלטה אנושית, ולא כמקור יחיד לקבלת מידע או כהחלטה סופית. זאת, למעט במצבים בהם מדובר במוצר ייעודי פנימי, שהוטמע בפעילות הרשות הציבורית בהתאם למדריך ניהול הסיכונים, אשר במדיניות שלו הוגדר אחרת והוא קיבל את האישורים הנדרשים, לרבות המשפטיים.

4. במצב שהתקבל פלט המעיד על חשש לדלף מידע אישי, רגיש או מסווג (בין בשימוש במערכת חיצונית, או במערכת שהוטמעה בארגון) - יש לתעד את הפלט במדויק ולדווח לאחראי היישום העסקי האמון על השימוש במערכת או לאחראי משילות בינה מלאכותית הארגוני.

5. **אין להניח שהתוכן נכון מבלי לבדוק אותו** – כאשר מתקבלת תוצאה עובדתית, למשל על ידי מוצרי מדף מבוססי בינה מלאכותית יוצרת, חשוב להצליב עם מקורות מידע מקובלים בגון פרסומים בספרות מדעית או באתרים מובילים, ולבדוק את מהימנות התוצאה לפני השימוש בתוצאה העובדתית שהופקה.

6. **מומלץ לשאול שאלות בתחום מומחיותך** – מומלץ בכל הניתן לכתוב הנחיות (Prompts) בנושאים שהם בתחום המומחיות או העיסוק של המשתמש, או קשורים אליו בזיקה ישירה, כך שיוכל לבקר את התוצאות. למשל, מי שעוסק בשמאות על נכסים, עדיף שלא יסתמך לצורך עבודתו על תשובות מבינה מלאכותית יוצרת לשאלות בתחום המיסוי, כי אין לו את הכלים לשפוט את מהימנות ודיוק התשובה. במקרה זה יש לפנות לגורמים שזה תחום עיסוקם. ככל שרלוונטי, מומלץ להתייעץ עם עמיתים לגבי המהימנות של הפלטים המתקבלים מהמערכת.

7. **שקיפות ותיעוד** – אם נעשה שימוש משמעותי במערכות בינה מלאכותית יוצרת בגיבוש תוצרים, מומלץ ככל הניתן, בהתאם לנסיבות, לציין זאת בתוצר הסופי. אין להציג תוכן שנוצר על ידי בינה מלאכותית כאילו הוא פרי יצירה אנושית מקורית. נוסף על כך, במקרה של הסתייעות במערכת לצורך קבלת החלטה (למשל, במקרה של מערכות ייעודיות שהוטמעו בפעילות הארגון), יש לתעד את השימוש במערכת.

8. **יש לשים לב אם נעשה שימוש במידע המוגן בקניין רוחני** – כאשר מדובר ביישום של בינה מלאכותית יוצרת, הפלט המתקבל עלול להכיל מידע המוגן על ידי זכויות יוצרים[10] ויש להפעיל שיקול דעת בשימוש באותו המידע. במצב של חשש, מומלץ לאתר את מקורות המידע של הכלי, ככל שהכלי מפנה אליהם, ולבדוק האם התוכן מצריך אזכור והפנייה למקורות. כאשר מדובר בתוצר שאינו טקסטואלי ייתכן שלא דיי באזכור והפניה למקור ומומלץ לפנות לייעוץ המשפטי של המשרד.

9. **אוריינות** – יש ללמוד כיצד להשתמש בכלי בינה מלאכותית גנרטיבית ולהבין את היתרונות, המגבלות והסיכונים שלהם. מומלץ להשתתף בקורסים ולקרוא מאמרים בנושא, באופן תכוף.

[10] למשל במצבים שבהם המשתמש מזהה כי הפלט מכיל ציטוטים מיצירות כתובות, ויש כוונה לעשות שימוש בתוצר, למשל במסמכים רשמיים של הארגון, יש להתייעץ עם הייעוץ המשפטי במשרד, על מנת לבחון האם מדובר ב"שימוש הוגן".

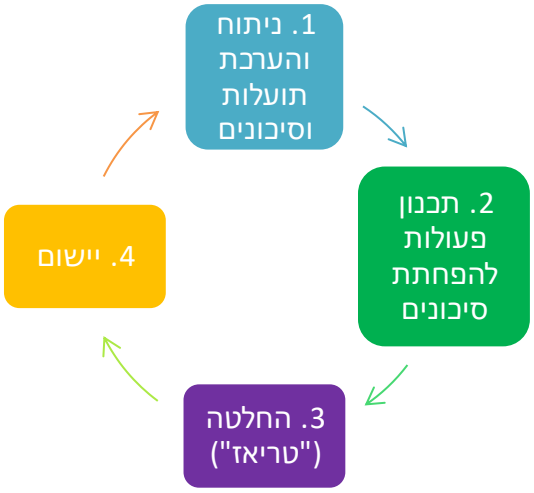
נספח ד - ניהול סיכוני בינה מלאכותית

נספח זה מורכב מחמישה חלקים:

1- מתודת ניהול סיכונים

חלק זה נועד לספק תשתית ובסיס אחיד לפיתוח מתודולוגיית ניהול סיכונים. האחראי למשילות בינה מלאכותית בארגון יכול, לפי שיקול דעתו, להרחיב את המודל, או לשנות אותו (כל עוד הוא כולל את המרכיבים היסודיים באופן כזה או אחר). בחלק זה, יפורט תהליך ניהול הסיכונים "ממבט על", תוך סקירת השלבים העיקריים בתהליך ניהול הסיכונים: ניתוח והערכת תועלות, זיהוי סיכונים, דירוג סיכונים, החלטה (טריאז') וניהול הסיכון.

תהליך זיהוי וניהול סיכונים הוא תהליך חזרתי שמתחיל כבר בשלב ייזום אפיון המערכת ומתבצע באופן שוטף במהלך כל חיי המערכת, אחת לחציון או לפי מידת הסיכון, על פי השלבים הבאים:



להלן טבלה המתארת את הפעולות הכלולות בכל שלב.

תיאור השלב	פעולה
שלב 1- ניתוח והערכת תועלות וסיכונים	לבחון באופן הוליסטי את התועלות והסיכונים של מערכת הבינה המלכותית המבוקשת. שלב 1 מורכב מארבעה תתי-שלבים: **זיהוי** התועלות והסיכונים, **בחינה** שיטתית של כל אחת מהתועלות והסיכונים, **התייעצות** ובדיקות, **והערכה** כוללות.
זיהוי	• לבחון את הרלוונטיות של קטגוריות התועלות והסיכונים (ראו חלקים 2 ו-3 לנספח זה) וכן תועלות וסיכונים אחרים שמסתמנים; • מיפוי התועלות והסיכונים הפוטנציאליים העולים ממערכת הבינה המלאכותית; • קבלת מידע רלוונטי מספק המערכת; • תיעוד התוצאות.

תיאור השלב	פעולה	
בחינה	בחינת התועלת כוללת בחינת: • היקף מושפעים • משך התועלות • היקף עוצמת התועלת • סבירות התממשות ראו חלק 2 לנספח זה. יש לתעד את ממצאי הבחינה.	בחינת הסיכונים כוללת בחינת: • סוגי הסיכונים • משך הסיכונים • היקף עוצמת הסיכונים • סבירות התממשות ראו טבלה בחלק 3 לנספח זה. יש לתעד את ממצאי הבחינה.
	בדיקת היקף המושפעים (תועלות וסיכונים) תתייחס לקטגוריות הבאות של מושפעים פוטנציאליים: • מקבלי השירות בארגון • ספקים • שותפים לתכנון וליישום המערכת • היחידות בארגון עצמו • הציבור בכללותו תוך התחשבות בהגנה על זכויות ואינטרסים ציבוריים.	
התייעצות	• התייעצות עם מגוון הגורמים הרלוונטיים, בהתאם לנסיבות: היבטים משפטיים: לשכה משפטית; היבטי פרטיות: DPO (הממונה על הגנת הפרטיות בארגון); סיכונים הנובעים משימוש בענן: אחראי הענן; הטמעת מרכיבים טכניים שיכולים להפחית את הסיכונים: מנהל אגף טד"מ (טכנולוגיות דיגיטליות ומידע); סיכונים הנובעים מאתגרים באיכות ונגישות הנתונים: CDO (ממונה דאטה ארגוני); בהשפעות המערכות על מערכות אחרות המצויות בתכנון או בפיתוח: PMO (מנהל הפרויקטים הארגוני או הגורם האמון על תכנון). - שותפי הארגון לרבות ספקיו; - מקבלי השירות ממנו לרבות מפוקחיו; - גורמים עסקיים הצפויים לעשות שימוש במערכת; - הציבור בכללותו. • תיעוד התוצאות	
הערכה	• דירוג התועלות באופן כולל: תועלת נמוכה מאוד/נמוכה/בינונית/גבוהה/גבוהה מאוד • דירוג הסיכונים באופן כולל: נמוך מאוד/נמוך/בינוני/גבוה/גבוה מאוד. ראו דוגמאות כאן. **דגש**: ככלל, כאשר ממערכת נשקף סיכון אחד בדירוג "גבוה מאוד", זה עשוי להוות אינדיקציה לכך שהמערכת כולה יוצרת סיכון גבוה מאוד. ביתר המקרים, סיווג הסיכונים והתועלות הכוללים נתונים לשיקול הדעת של האחראי יישום, בהתאם למדיניות הארגון בנושא. • תיעוד התוצאות	
שלב 2- תכנון פעולות להפחתת סיכונים	לאחר שמופו התועלות והסיכונים, יש לבחון מה הם האמצעים שיש להפעיל על מנת להפחית את הסיכונים, בין היתר לצורך בחינת ההחלטה האם להמליץ על יישום המערכת (ראו שלב 3 – טריאז'. שלב 2 מורכב מארבעה תתי-שלבים: **מיפוי** אמצעים, **התייעצות**, הערכת עלויות ו**תיעוד.**	
מיפוי	• מיפוי אמצעי הפחתה העומדים ברשות אחראי היישם – ראו פירוט דוגמאות לאמצעי הפחתה כאן.	
התייעצות	• התייעצות עם הגורמים המנויים בשלב 1 (בהתאם לצורך ולנסיבות).	

תיאור השלב	פעולה
	• בכלל, ככל שהסיכונים גבוהים יותר, מוצע כי ההתייעצות תהיה מקיפה יותר.
הערכת עלויות	• בחינת עלויות התוכנית להפחת סיכונים לרבות: - בחינת חלופות שונות; - הערכת עלות משוערת של אמצעי הפחתת הסיכון, ביחס להפחתת הסיכון הנשקף.
תיעוד	תיעוד תוצאות התהליך, הכוללות תוכנית להפחתת סיכונים.
שלב 3 - החלטה ("טריאז")	תכלול כל המידע שהתקבל עד כה, וקבלת החלטה אם להמליץ על יישום המערכות, בהתאם לנהלים שיקבעו על ידי אחראי משילות בינה מלאכותית. תהליך הטריאז' יכלול את תת-השלבים הבאים: **התייעצות**, **המלצה/החלטה** ו**תיעוד.**
התייעצות	התייעצות עם הגורמים המנויים בשלב 1 (בהתאם לצורך ולנסיבות). **דגש**: ככל שסיכוני המערכת גבוהים, יש להעריך את מסוגלות הארגון להקצות את התשומות הנדרשות לצורך ניהול הסיכונים.
המלצה/החלטה	גיבוש המלצה, ולאחר מכן קבלת החלטה על ידי הגורם המוסמך על פי נהלי הארגון שיקבעו כמומלץ על ידי אחראי משילות בינה מלאכותית. כאשר מתקבלת החלטה להתקדם תוך הפעלת כלי ניהול סיכונים, יש לגבש תוכנית מפורטת להפחת סיכונים, שתאפשר לבצע מעקב שוטף ובקרה ולהפעיל את המערכת בצורה בטוחה ואחראית.
תיעוד	תיעוד החלטה הסופית, לרבות פרטי תוכנית הפחתת סיכונים.
שלב 4 - יישום	שלב זה מתחיל כאשר הארגון מריץ את המערכת. הוא כולל **יישום** תוכנית הפחתת סיכונים, **בקרת** המערכת, **דיווח**, ביצוע **התאמות** על פי הצורך ופעילות יזומות של **שקיפות** כלפי הציבור.
יישום	ביצוע אמצעי הפחתת סיכונים בהתאם לתוכנית ניהול הסיכונים.
בקרה בזמן אמת	איסוף נתונים מהמערכת עצמה; התייחסות למידע מקהילת המושפעים מהמערכת, שצף בפניות ציבור, ערוצי תקשורת, שיח עם ארגוני חברה אזרחית רלוונטיים ועוד. בקרת הנתונים – בדיקת מדגמיות או אד הוק וכן בהתאם להתפתחויות, לצורך ולרמת הסיכונים.
דיווח לאחראי משילות	דיווח שוטף לאחראי משילות בינה מלאכותית – תדירות הדיווח תהיה בהתאם לתוכנית הפחתת הסיכונים. דיווח לאחראי משילות בינה מלאכותית כאשר מתרחשת תקרית AI. **דגש**: מומלץ לכלול בדיווח המלצות על ההתאמות הנדרשות.
התאמות	ביצוע ההתאמות בהתאם להנחיות אחראי משילות בינה מלאכותית של הארגון. ההתאמות יכולות לכלול, למשל, שינויים באפיון המוצר או באופן איסוף או תיוג הנתונים.
שקיפות לציבור	שקיפות ציבור כוללת מספר מרכיבים מרכזיים: • שקיפות שוטפת לגבי עצם השימוש במערכות בינה מלאכותית על ידי הארגון, סוגי המידע המשמש את המערכת, אופן פעילות המערכת באופן כללי ונגיש לציבור (לרבות יכולותיה ומגבלותיה). מרכיב זה יכול להיות מותאם לקהילת המושפעים, למידת הסיכון מהמערכת, ולטכנולוגיות הזמינות. • שקיפות לציבור הרחב (או לקהילת המושפעים) כאשר מתרחשת תקרית AI – בהתאם להנחיות אחראי משילות בינה מלאכותית. • שקיפות והגברת מודעות למצבים שבהם מתקיימת אינטרקציה ישירה או משמעותית עם מערכת בינה מלאכותית. פירוט נוסף בנושא יכלל במדריך המשפטי שיצורף למסמך זה בהמשך.

2- סיווג תועלות

נקודת ההתחלה של אפיון הצרכים העסקיים הוא בחינת התועלות השונות של מערכת בינה מלאכותית. ניתן לסווג את התועלות לפי קטגוריות שונות. להלן רשימה לא ממצה של קטגוריות תועלות מרכזיות שאותם מוצע לזהות. ניתן לבחון תועלות נוספות הרלוונטיות לפעילות הארגון או לשימושים הצפויים לכלי.

פרודוקטיביות באמצעות אוטומציה של תהליכים שונים (ניתוח דאטה, שירותים לאזרח, אכיפה) ומדיניות ציבורית יעילה יותר. דוגמאות: קיצור זמני המתנה/זמני עיבוד תיקים (processing times), שיפור איכות ההחלטות, פעולה מסביב לשעון ללא הפסקות, הפחת הנטל על הגורם האנושי ופינוי זמנו למשימות ליבה, מניעת טעויות אנוש. כל אלו ועוד צפויים בין היתר להפחית את הנטל על הגורמים האנושיים, הן מצד עובדי הגופים הציבוריים והן מצד התושבים והעסקים המקבלים שירותים ציבוריים, לשפר את איכות השירותים הניתנים לציבור ולעלות את אמון הציבור במוסדות המדינה.

פרואקטיביות בעיצוב והאספקה של מדיניות ושירותים ציבוריים. דוגמאות: פנייה אקטיבית לתושבים ועסקים לטובת מימוש הטבות ומיצוי זכויות על ידי הפעלת מערך סוכני AI, וגיבוש מדיניות פרואקטיביות כהיערכות למצבי חירום מתהווים על ידי הפעלת מודלי AI לחיזוי חכם ומקדים של מגמות.

פרסונליזציה, כלומר התאמת השירותים הציבוריים עבור תושבים ועסקים בהתאם לצרכים הייחודיים שלהם.

תועלות נלוויות: באשר נעשה שימוש אחראי במערכות בינה מלאכותית לטובת קידום מטרות הארגון, התושבים מרוויחים משירות ציבורי טוב יותר. כנגזרת מכך, גם מוניטין הארגון ואמון הציבור בארגון עשויים להתחזק.

שימוש בבינה מלאכותית עשוי להוביל לתוצאות בעלות ערך גבוהה מאוד. כך למשל, קול קורא של מערך הדיגיטל הלאומי ומשרד החדשנות, המדע והטכנולוגיה, תומך בשימושי בינה מלאכותית על ידי המגזר הציבורי בעלי תועלת פוטנציאלית גבוהה. לדוגמאות מחו"ל, ניתן לראות את מאגר ה-OECD.

במסגרת תהליך ניהול סיכונים, מוצע לבצע את הפעלות הבאות:

- אפיון התועלות המצופות השונות, באופן מדויק וממוקד ככל הניתן, כולל חישוב ROI פוטנציאלי;
- הערכת התועלות, מבחינת הערך היחסי שהן מביאות (מ-"נמוך מאוד" עד "גבוה מאוד");
- הערכת סיכוי התממשות שלהן;
- השוואת התועלות לסיכונים באופן הולם.

3- סיווג סיכונים

בכל תהליך ייתכנו סיכונים בסיסיים, כדוגמת טעויות אנוש, איטיות התהליך, שיקול דעת לקוי בתיעדוף ועוד. שילוב של מערכת בינה מלאכותית בתהליך עשוי להפחית סיכונים אלה או להגביר סיכונים קיימים, או לייצר סיכונים אחרים. על כן, יש לבחון את הסיכון שנוסף ביחס למצב הבסיסי ובהתייחס למחולל הסיכון. למשל, הפחתה של טעויות המערכת מ-15% על ידי גורם אנושי ל-1% על ידי בינה מלאכותית מצביע על סיכון של 1%, אבל על שיפור משמעותי מהמצב הקיים ויש לקחת זאת בחשבון. בבסיס שלב זיהוי וניתוח הסיכונים יש להעריך האם מערכת הבינה המלאכותית היא הגורם המחולל באופן ישיר (או עקיף) את הסיכון, קרי – סיבת השורש לסיכון, או שהסיכון נגרם מסיבות אחרות כגון תהליך ארגוני או "עסקי", מערכת תקשובית אחרת וכדומה.

להלן רשימה לא ממצה של סוגי הסיכונים המרכזיים שאותם מוצע לזהות ולנהל. ניתן לבחון סיכונים נוספים הרלוונטיים לפעילות הארגון או לשימושים הצפויים לכלי.

- **סיכון תפעולי**: מערכות AI מושתתות על תהליכים פנימיים מורכבים, מערכות טכנולוגיות מתקדמות, ואינטראקציות אנושיות העלולות להיות מועדות לשיבושים ולכשלים, לדוגמה – מערכות AI מבוססות על איסוף וניתוח של כמויות גדולות של נתונים. נתונים לא מדויקים, לא מעודכנים או מוטים עלולים לפגוע באיכות המערכת וליצור שגיאות או עיוותים בתוצאות, וכן לייצר קושי להתחקות אחר הלוגיקה של המערכת.[11]

[11] בשל מאפיין של "קופסה שחורה" (אלמנט העכירות) הקיים במערכות בינה מלאכותית.

- **<u>סיכון כלכלי</u>**: טעויות שעלולות להיגרם בשימוש במערכות AI עלולות לגרום לנזקים כלכליים, לרבות אובדן הכנסות, כולל אובדן גבייה, הפסקה או אי התחלה של פעילות עסקית (שלילת רישיון למשל) והוצאות עודפות. נזקים אלו יכולים להיווצר הן מצד הגוף הציבורי, והן מצד הנמענים של הפעילות שלו לרבות תושבים, עסקים וארגונים אחרים.
- **<u>סיכון לנזק בריאותי או בטיחותי</u>**: למשל, טעויות במערכת אשר מסייעת בתהליך מתן אישורים רגולטורים לתרופות, או שנועדה לנתב תחבורה, עלולות לגרום לנזק ממשי לאנשים ולרכוש.
- **<u>סיכון לנזק סביבתי</u>**: למשל, טעויות במערכת אשר תפקידה לסייע במתן היתרי פליטה לאוויר ובתנאים הסביבתיים ברישיונות העסק, או לבצע בקרות אוטונומיות לצרכי פיקוח ואכיפה. טעויות כאלו עלולות להביא ליצירת זיהומים עודפים ולאכיפת חסר.
- **<u>סיכוני אבטחת מידע</u>**: שימוש במערכות AI חושף את הארגון לסיכוני אבטחת מידע ייחודיים, העלולים לפגוע בפרטיות ושלמות הנתונים, וכן להוביל לדליפת מידע רגיש. נזקים אלו עלולים להיגרם כתוצאה מהתקפות על הכלים עצמם – גורמים עוינים עשויים לנסות לנצל חולשות במערכות ה AI ולשבש תוצאות, ניסיונות להוציא מידע ממערכות ניהול הנתונים. גישה בלתי מורשת למידע עשויה להתרחש במערכת או בתשתיות תומכות, כגון בסיסי נתונים ושרתי ענן וכתוצאה מכך זליגה של נתונים רגישים, או שימוש לרעה במודלים באופן הגורם לחשיפת נתונים רגישים. נוסף על כך, מודלים של AI עלולים להיחשף למתקפות מבוססות הטעיית מודל, דבר העלול לשבש את תחזיות המודלים ולהשפיע על החלטות החיוניות לארגון ולבטיחות המידע.
- **<u>פגיעה באמון הציבור ובמוניטין הארגון</u>**: טעויות, הטיות או יצירת דיס-אינפורמציה, בין אם מוחצנים למשתמשי קצה כגון צ'טבוטים משרדיים לשימוש הציבור, ובין אם משמשים לצרכים פנימיים, עלולות להביא לפגיעה במוניטין הארגון ובאמון הציבור בארגון.

היבטים משפטיים: מערכות בינה מלאכותית מעוררות לעיתים שאלות משפטיות, למשל ביחס לכללי המשפט המינהלי (חריגה מסמכות הארגון הציבורי, חובת הנמקה), מניעת פגיעה בפרטיות, מניעת הטיות והפליה, והגנה על קניין רוחני. בהתאם לאופי המערכת ולהיבטים המשפטיים המתעוררים ביחס לשימוש המתוכנן או הקיים, יש להיוועץ עם הלשכה המשפטית על מנת למפות את הדרישות המשפטיות שבהן המערכת תידרש לעמוד. יודגש כי הבחינה המשפטית נפרדת ועומדת בפני עצמה. לצד זאת, חשוב לשקף למנהל היישום את הדרישות המשפטיות והסוגיות המשפטיות שעמן יש להתמודד, ואלה יילקחו בחשבון הן במסגרת עיצוב תכנית המיטיגציה (למשל, כדי ליישם כלי התמודדות עם בעיות הטיה או פרטיות) והן במסגרת שלב קבלת ההחלטה (הטריאז').

סיכונים חריגים וסוגי שימושים שיש לבחון לשלילה - במקרים שבהם מדובר בכלי בינה מלאכותית המציבים סיכון יוצא דופן שעשוי להיות מזיק ומסוכן או לפגוע באופן מהותי ומשמעותי בזכויות יסוד, **נדרש תהליך בחינה נפרד על מנת לבחון האם יש כלל מקום לקידום הפרויקט.** לשם המחשה והשוואה, בחוק הבינה המלאכותית של האיחוד האירופי (AI Act EU) שנחקק בשנת 2024, שימושים שונים במערכות בינה מלאכותית נאסרו במפורש לשימוש, למשל, ביחס למערכות בינה מלאכותית שעושות שימוש בטכניקות תת הכרתיות, מניפולטיביות או מטעות; שמנצלות חולשות של אדם או קבוצת אנשים, בשל גיל, מוגבלות או מצב כלכלי מסויים; שמדרגות בני אדם על סמך התנהגות חברתית או מאפיינים אישיים, כאשר הדירוג מוביל ליחס מזיק או בלתי הוגן בהקשרים שפורטו בחוק; או מערכות שמסיקות מסקנות ומנתחות רגשות במקומות עבודה ובמוסדות חינוך, למעט למטרות רפואיות או בטיחותיות.

במקרה שבו מתעורר חשש כי מדובר במערכת המציבות סיכון יוצא דופן או עלולות לפגוע בזכויות באופן מהותי ומשמעותי כמפורט לעיל, ובפרט ביחס למערכות כגון אלא שפורטו לעיל – **יש להתייעץ עם הלשכה המשפטית** על מנת לבחון האם יש מניעה משפטית או קושי משפטי משמעותי שבגינו אין מקום לשילוב AI במערכת. ככל שניתן לקדם את הפרויקט מבחינה משפטית, **יש להיוועץ עם קובעי מדיניות בדרג נבחר האמונים על התחום שבו מערכת ה-AI פועלת**.

4- <u>דוגמאות ליישום המתודה לניהול סיכונים</u>

הדוגמא שלהלן תמחיש יישום ביחס למערכת AI שצפויה לקצר זמני המתנה לשירות מסוים באחוז מסוים. האחוזים והניקוד להלן מהווים **דוגמאות להמחשה בלבד**. על אחראי היישום להפעיל את שיקול דעתו, בהתאם למתודה לניהול סיכונים של הארגון ולהנחיות שיקבל מאחראי משילות בינה מלאכותית, לקבוע את הניקוד המתאים בכל מקרה לגופו.

הערכת תועלות – קיצור זמני המתנה

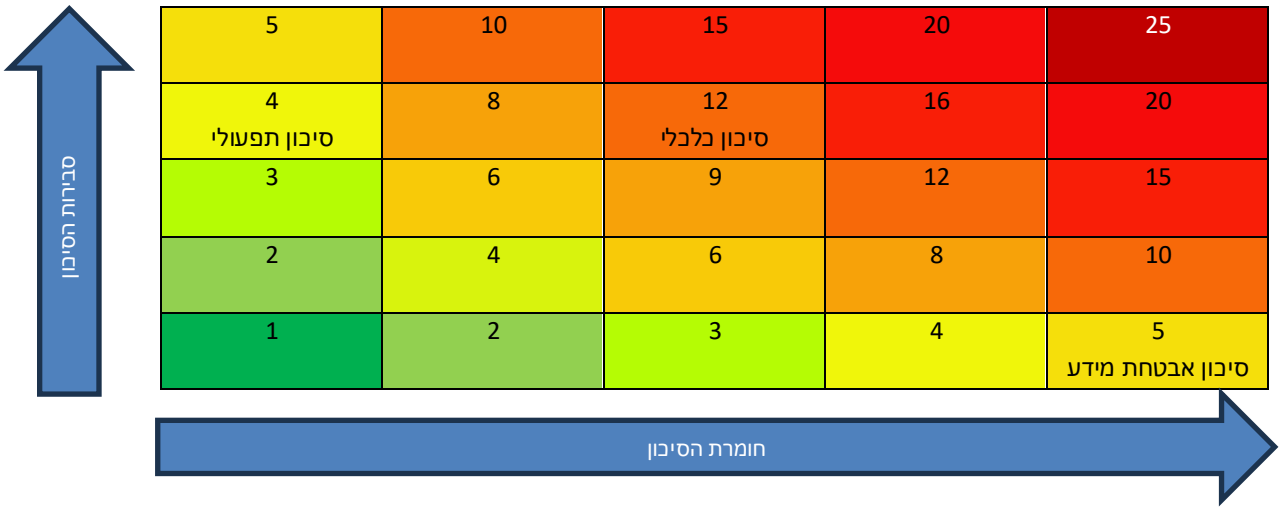
הערכה	מושפעים פוטנציאלים	משך התועלת	עוצמה	סבירות התממשות
נמוך מאוד	בודדים	חד פעמי	פחות מ-2%	פחות מ-50%
נמוך	עשרות עד מאות	נדיר	2-5%	51-60%
בינוני	אלפים או בודדים מאוכלוסיות רגישות	מדי פעם	6-7%	60-74%
גבוה	עשרות אלפים	לעיתים תכופות	8-9%	75% ומעלה
גבוה מאוד	מאות אלפים	באופן קבוע	10% ומעלה	ודאית או כמעט ודאית

הערכת הסיכון

הערכה	מושפעים פוטנציאלים בהתממשות הסיכונים	משך הסיכון	עוצמה	סבירות התממשות
נמוך מאוד	בודדים	שעות	פחות מ-2%	פחות מ-50%
נמוך	עשרות עד 49	ימים	2-5%	51-60%
בינוני	50-99	שבוע	6-7%	60-74%
גבוה	100-999	שבועיים	8-9%	75% ומעלה
גבוה מאוד	מעל 1,000 או מאות מאוכלוסיות מוחלשים	שלוש שבועות או יותר	10% ומעלה	ודאית או כמעט ודאית

בסיום בחינת הפרמטרים שהודגמו לעיל, מומלץ לתת ציון מספרי, **<u>עבור כל סיכון בנפרד</u>**, לחומרת הסיכון וכן לסבירות התממשותו (1 לנמוך ביותר ו-5 לגבוה ביותר). דירוג הסיכון המשוכלל לכל סיכון הוא תוצאה של הכפלת חומרת הסיכון (1-5) בסבירות התממשות הסיכון (1-5).

בסופו של תהליך זיהוי והערכת הסיכונים, רצוי לייצר "מפת חום" שתאפשר לתעדף את התשומות המוקדשות לטיפול בכל סיכון. למשל, נתון שזוהו, במערכם הנבחנת, שלושה סיכונים: סיכון אבטחת מידע, סיכון כלכלי וסיכון תפעולי. חומרת הסיכון התפעולי נחשב נמוך מאוד (רמה 1) אך הסבירות התממשותו היא גבוהה יחסית (רמה 4). לעומת זאת, חומרת הסיכון של אבטחת מידע גבוהה מאוד, אך סבירות התממשותו נמוכה מאוד. הסיכון הכלכלי הוא בינוני (רמה 3) אך סבירות התממשותו יחסית גבוהה (רמה 4). ויזואליזציה של רמות הסיכונים השונים מסייעת לאחראי יישום עסקי להתמודד עם הסיכונים השונים בצורה מתאימה. נוסף על כך, אחראי משילות בינה מלאכותית יכול להיעזר בטבלאות שיקבל מכלל הגורמים העסקיים בארגון, כדי ליצור מפת חום ארגונית להפיק תובנות מעשיות ממנה.



ל"בנק" סיכוני תקשוב מפורט עם בקרות מומלצות, ראו את בנק הסיכונים של מערך הדיגיטל כאן.

5- <u>המלצות ופעולות הפחתת סיכונים</u>

תורת ניהול הסיכונים קובעת, ככלל, ארבע אסטרטגיות התמודדות עם סיכונים:

1. <u>הימנעות מהסיכון</u> כאשר מדובר בסיכונים גבוהים מאוד אשר אינם מצדיקים את התועלת המצופה;
2. <u>התעלמות מהסיכון</u> כאשר מדובר בסיכונים <u>נמוכים מאוד</u> שאינם מהותיים לארגון או לאנשים או ארגונים שבזיקה אליו;
3. <u>העברת הסיכון</u> כשניתן לצד שלישי באמצעות למשל גילוי או ביטוח;
4. <u>הפחתת הסיכון</u> באמצעות נקיטת פעולות הפחתה, כאשר מדובר בסיכונים מהותיים אשר נטילתם מוצדקת בנסיבות העניין, וניתן לבצע פעולות הפחתה ומעקב שוטף ובקרה על ביצוען ולהפעיל את המערכת בצורה בטוחה ואחראית.

יש מגוון פעולות שניתן לבצע על מנת להפחית את הסיכונים בשלבי אפיון, פיתוח והרצת מערכות בינה מלאכותית. האמצעים המתוארים להלן הם כלליים ויכולים לתת מענה למגוון סוגי סיכונים הנשקפים משימוש במערכות AI. כאמור, הפעלת אמצעים אלו יכולה להפחית את חומרת הסיכונים השונים.

ניתן לסווג את הפעולות למספר קבוצות מרכזיות: אמצעים טכניים וארכיטקטוניים; הוספת רכיבים בגוף מוצר ה-AI; ואמצעים עסקיים-ארגוניים. **להלן דוגמאות נבחרות של אמצעים שניתן להפעיל בכל קבוצה:**

א. **<u>אמצעים טכניים וארכיטקטוניים</u>**

כדי לצמצם את הסיכון, חשוב לשלב פתרונות בבר ברמת התכנון הטכני והארכיטקטורה של המערכת אמצעים **טכניים וארכיטקטוניים** להפחתת סיכונים:

1. **שילוב <u>טכניקת RAG או CAG</u> בארכיטקטורת המערכת** - שילוב של אחזור מידע ממקורות מהימנים עם מודל גנרטיבי צפוי להפחית סיכונים הנובעים מ"הזיות" וטעויות של המערכת, ולהגביר את היכולת של הארגון להסביר את הפלט שנתקבל על ידי מערכת. כמו כן, טכניקה זו מאפשרת למערכת לפעול על בסיס מקורות מידע עדכניים ואמינים, המנוהלים על ידי הארגון או מקורות חיצוניים, ומפחיתה את הסיכון לפלט לא אמין או מידע שגוי.
2. **שילוב <u>סוכן AI</u> (Agent) מוגדר למשימה ייעודית** – שילוב של סוכנים מתמחים במערכת AI עבור משימות ייעודיות, תחת הגדרה ברורה של המשימה (כגון חיפוש אינטרנטי, ביצוע פעולות חישוב וכד'), צפוי להפחית את הסיכונים הנשקפים מטעויות והטעיות שבמודל ה-AI, ולהגביר את הדיוק של המערכת. כמו כן, הפעלת סוכני AI יכולה לאפשר לזהות מתי נדרשת מעורבות אנושית ו"קריאה" לגורם אנושי מוסמך לקבל החלטה.
3. **Guardrails (מעקות בטיחות ל-AI)** – הגדרת חוקים וכללים טכניים המונעים מהמערכת לעסוק בנושאים אסורים או לענות תשובות בעייתיות. חוסם תוכן בלתי רצוי מראש (למשל: סוגיות רגישות, שפה פוגענית, החלטות בלתי מוסמכות).
4. **Reasoning (הסקת מסקנות)** – שימוש במודלים המשלבים טכניקות של Reasoning עצמי מאפשרים למשתמש מידה של הסברתיות ואפשרות להתחקות במידה מסוימת אחר תהליך יצירת הפלט של המערכת ובכך מפחיתים סיכונים הנובעים מהזנת קלטים לא מדויקים או מותאמים לצורך העסקי של המשתמש.
5. **שימוש בטכנולוגיות מגבירות פרטיות (PET) כאמצעי להפחתת סיכוני פרטיות** - PET (Privacy Enhancing Technologies) הן טכנולוגיות שמטרתן לאפשר עיבוד, ניתוח או שיתוף של מידע תוך שמירה על פרטיות האנשים שמהם נאסף המידע.[12] שימוש בטכנולוגיות אלו, ועיצוב לפרטיות (privacy by design), יאפשרו הפחתה וצמצום פגיעה בפרטיות.
6. **שליטה על טמפרטורה** - טמפרטורה (Temperature) היא פרמטר במודלים גנרטיביים שמגדיר את רמת האקראיות והיצירתיות של הפלט. כך טמפרטורה נמוכה (0.1–0.3) תיתן תשובות עקביות, בטוחות וזהירות.

[12] להרחבה ראו <u>מדריך לטכנולוגיות מגבירות-פרטיות PETs</u>, שנכתב ע"י הרשות להגנת הפרטיות.

ואילו טמפרטורה גבוהה (0.7–1.0) תיתן תשובות מגוונות, אך גם פחות צפויות – ולעיתים שגויות או בעייתיות. במודלי שפה מסוימים שליטה על הטמפרטורה יכולה להתבצע על ידי משתמשי קצה ב"שפה חופשית" בתוך הפרומפט, ובחלק מהמודלים יש צורך בגורם טכני שיגדיר טמפרטורה במסגרת קריאת ה-API למודל.

ב. **<u>מרכיבים במוצר</u>**

בחלק מתכנון מוצר מבוסס AI, ניתן לשלב מרכיבים שמחזקים את שקיפות השימוש ויכולת הפיקוח והלמידה על ידי משתמשי הקצה:

1. **גילוי (Disclosure)** – המערכת תשקף למשתמשים שהיא מבוססת על בינה מלאכותית ותשקף את מגבלותיה. פעולה זו יכולה להגביר את הבקרה האנושית ולהתוות את אופן השימוש במערכת. יש לציין שבמצבים מסוימים מרכיב הבינה המלאכותית יכול להיות שולי בתפקוד המערכת התקשובית, או שהסיכונים הנשקפים מפעילתו נמוכים ואינם צפויים להשפיע על ביצוע התפקיד של משתמשי הקצה, ועל כן הוספת מרכיב של גילוי נתונה - לפחות בשלב זה – לשיקול דעת.
2. **דיווח משתמשים (User Feedback/Reporting)** – הוספת מרכיב למוצר שיאפשר למשתמשים לדווח על בעיות, חוסר הבנה או פגיעה שנגרמה משימוש במערכת, צפוי לחזק את הבקרה הארגונית עליה, ובמקרים המתאימים אף לאפשר למשתמשים לערער ברמה הפנים ארגונית על פלט המתקבל.

ג. **<u>אמצעים עסקיים-ארגוניים</u>**

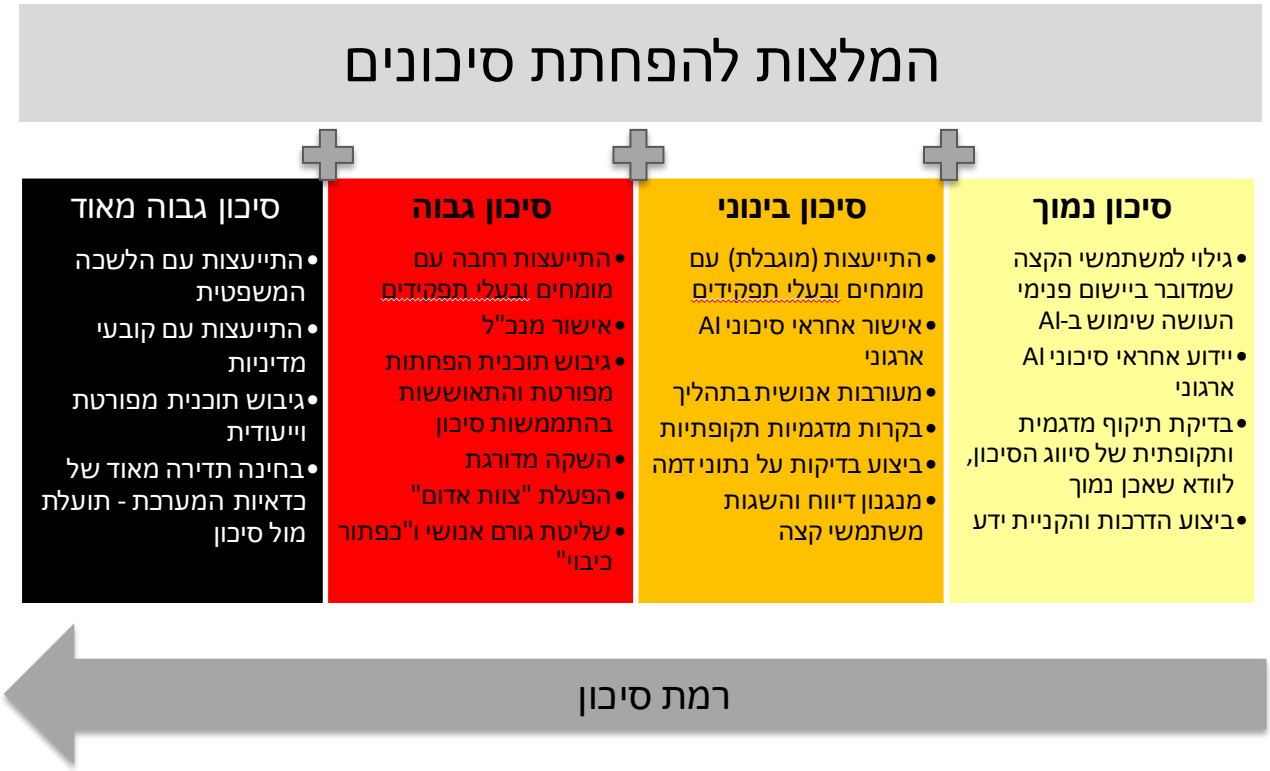
מעטפת תהליכית ארגונית שתספק שיקול דעת אנושי, פיקוח מקצועי, חיזוק תרבות שימוש אחראי, ולמידה מתמשכת לאורך חיי המערכת, צפויה אף היא לסייע בהפחתת הסיכונים. לדוגמה:

1. **מעורבות אנושית** – שילוב של גורמים אנושים במסגרת פעילות המערכת והתהליך העסקי המבוסס על מערכת בינה מלאכותית. הבטחת פיקוח אנושי הולם לאורך כל מחזור החיים של מערכת ה-AI, לרבות בקרה על פיתוח, פריסה, שימוש והחלטות שהמערכת מפיקה, יכולה להוות אמצעי בקרה תפעולי שוטף. המעורבות יכולה להתבצע מראש (ex ante), במהלך הפעולה (real-time), או בדיעבד (ex post), בהתאם לרמת הסיכון וההקשר השימושי.
2. **התייעצות מומחים** – שילוב מומחים בתחומים רלוונטיים כגון אתיקה, משפט, פרטיות (DPO), שוויון והכלה, והגנת סייבר, כבר בשלבי האפיון והפיתוח, צפוי לאפשר זיהוי נכון של הסיכונים ובניית המנגנונים הרלוונטיים להפחתתם.
3. **בקרות מדגמיות** – ביצוע בדיקות איכות מדגמיות של פלט המערכת (outputs), אחת לתקופה או על מקרים רגישים, צפוי לסייע בזיהוי סיכונים לאורך חיי המערכת.
4. **הסתייעות ב"צוות אדום" וביצוע תקיפות מבוקרות** – הפעלת צוות אדום (Red Team) שתפקידו לבדוק את הבטיחות, האמינות והחוסן של המערכת על ידי תקיפות מבוקרות של המערכת תסייע בזיהוי של פגיעויות, כשלים בעמידות, חולשות, הטיות וסיכונים אפשריים.
5. **משילות נתונים (Data Governance)** – אמינות מערכת בינה מלאכותית תלויה באופן ישיר והדוק בנתונים שעל בסיסם פועלים המודלים. על מנת להבטיח משילות איכותית של הנתונים המשמשים מערכת בינה מלאכותית ארגונית, ניתן לנקוט בצעדים הבאים:
   - הבטחת שימוש בדאטה הנכון למודל הנכון – נתונים עדכניים, שלמים, מייצגים, חוקיים ואמינים.
   - ניהול גישה, רישוי, הטמעת מנגנונים לטיפול בנתונים לא מאוזנים והבטחת איכות ושלמות הנתונים.
   - נקיטת צעדים ארגוניים לזיהוי והפחתה של הטיות בנתונים.

   כל זאת, בשים לב לזכויות ואינטרסים של נושאי המידע.
6. **הגברת מודעות** – ארגון יכול לבצע מספר פעולות כגון הכשרות, סדנאות ופרסום מסמכי מדיניות ומקרי בוחן פנימיים למשתמשים ומקבלי החלטות. כמו כן, ארגון יכול לנקוט בצעדים לעידוד דיווח בקרות התממשות סיכוני בינה מלאכותית.

<u>פעולת מומלצות לפי חומרת הסיכונים</u>

בתרשים להלן מוצגות המלצות לפעולה לפי רמת הסיכון של המערכת. הן כוללות פעולות בקרה, יידוע ופעולה במקרה שבו קיים חשש או שהסיכון התממש, כאשר כל רמת סיכון כוללת את כל המומלץ גם ברמות הסיכון שמתחתיה. לניהול מיטבי של סיכונים שזוהו יוגדרו הכלים לניהול סיכונים שיאפשרו לבצע מעקב שוטף ובקרה ולהפעיל את המערכת בצורה בטוחה ואחראית יותר.



נספח ה – תקינה בינלאומית

מדינות ומוסדות תקינה מובילים בעולם עמלים על פיתוח תקינה והנחיות לשימוש אחראי בבינה מלאכותית. בחלק גדול מהתקנים וההנחיות ניתן למצוא את העקרונות המשותפים הבאים:

א. **גישה מבוססת ניהול סיכונים ביחס לשימוש בבינה מלאכותית,** שמטרתה לזהות, להעריך, לתעדף ולנהל את הסיכונים הפוטנציאליים לאורך מחזור חיי המערכת. הגישה מתמקדת בהתאמת רמת הניהול לרמת הסיכון בפועל, תוך שמירה על איזון בין ניהול הסיכונים לבין חדשנות טכנולוגית והתועלות הצפויות משימוש בה.

ב. **שקיפות** בשימוש בכלי בינה מלאכותית.

ג. **תקינות והנחיות שיעברו שיפור מתמשך** - התקינות וההנחיות דינאמיות, משתנות ומתעדכנות מעת לעת.

ד. שמירה על **זכויות האדם** כתנאי בסיסי להפעלה ושימוש.

ה. **שיתוף בעלי עניין** הוא מרכיב חיוני ביצירת תהליך פתוח, שקוף ומכיל, המאפשר ניהול מיטבי של סיכוני בינה מלאכותית והבטחת תועלת רחבה.

ו. מנגנוני **בקרה ופיקוח** מותאמים לרמת הסיכון ולקשיים.

להלן מספר דוגמאות של תקינה רלוונטית:

1) **תקני ISO** – תקן ISO-23894 להנחיות לניהול סיכונים הנוגעים למערכות AI; תקן ISO 8000-51 למשילות נתונים בארגון; תקן ISO 42001 לניהול מערכות AI; תקן ISO 31000 לניהול סיכונים כללי ועוד.

2) **תקני IEEE** – לרבות תקן P7000 בעניין עיצוב אתי של מוצרים, תקן P7003 בעניין הטיות אלגוריתמית, ותקן P7009 בעניין בטיחות של מערכות אוטונומיות.

3) תקני **CEN/CENELEC** – שורה של תקנים אשר גופי תקינה אירופאים התבקשו לגבש, בהתאם לחקיקת האיחוד האירופאי בתחום (ה-AI ACT), אשר צפויים ללוות את יישום אותה חקיקה.

4) **NIST.AI 100-1** – מסמך מסגרת לניהול סיכוני בינה מלאכותית שנכתב על ידי המכון הלאומי לתקנים וטכנולוגיה בארה"ב (NIST). התקן מתווה קווים כלליים ואינו מעניק כלים להערכה מדויקת של הסיכון.

נספח ו - מילון מונחים

להלן מילון מונחים שכיחים בעולם הבינה המלאכותית ומובאים במדריך.[13]

המונח	הגדרה
מערכת בינה מלאכותית (AI)	מערכת בינה מלאכותית היא מערכת מבוססת מכונה שלמטרות מפורשות או מרומזות, מסיקה מהקלט שהיא מקבלת, איך לייצר פלטים כמו תחזיות, תוכן, המלצות, או החלטות שיכולות להשפיע על סביבות פיזיות או וירטואליות. מערכות AI שונות משתנות ברמות האוטונומיה והסתגלות שלהן לאחר הפריסה.[14]
בינה מלאכותית יוצרת (Generative AI)	מערכות בינה מלאכותית המסוגלות ליצור תוכן חדש כמו טקסט, תמונה, וידיאו, שמע ועוד, על בסיס הדוגמאות או הנתונים שעליהם אומנו. מודלים אלה מסוגלים להפיק פלטים יצירתיים מבלי שנדרש להם אימון ספציפי לכל משימה. יש סוגים שונים של יישומים **המחוללים תוכן** חדש ובכלל זאת: • ויזואלי: יצירת תמונות, סרטונים או גרפיקה. • קולי: יצירת דיבור מלאכותי, פסקולים, או קריינות. • טקסטואלי: כתיבת טקסטים יצירתיים, מקצועיים, או טכניים. הכלים מתבססים על מאגרי מידע רחבים ולמידת מכונה כדי לייצר תוכן שמדמה תוצרים אנושיים. דוגמאות: DALL-E (ויזואלי), Resemble AI (קולי), Jasper או ChatGPT (טקסטואלי).
סוכן AI (AI Agent)	מערכת AI הפועלת אוטונומית או חצי-אוטונומית כדי לבצע משימות או להשיג מטרות מוגדרות תוך אינטראקציה עם הסביבה שלה. בכלל, הביטוי מתייחס לשילוב של מודל שפה יחד עם יישום שמאפשר לו לבצע פעולות – במערכת סוכנים יש גם סוכן (Agent) מנתב שמפנה משימות לסוכנים שונים בהתאם לכלים שיש להם.
טכניקת RAG או CAG (Retrieval-Augmented Generation/Cache-Augmented Generation)	גישות המאפשרות להעניק למודל בינה מלאכותית סט מידע ספציפי אשר רצוי שיילקח בחשבון, כחלק מעיבוד השאלה שנשאלה כך ניתן לקבל תשובות מדויקות, מעודכנות ומפורטות יותר לשאלות מורכבות.
כלי מדף מבוססי AI	כלים, תוכנות או פלטפורמות מבוססות AI המוכנים לשימוש "ישר מהקופסה" (out-of-the-box). אלה הם פתרונות כלליים המתאימים לשימושים רחבים ונפוצים, ובמצבים מסוימים עם אפשרות מסוימת להתאמה אישית.
מנועי חיפוש מבוססי שיח	כלים המשלבים בין מנועי חיפוש מסורתיים לבין טכנולוגיות של עיבוד שפה טבעית (NLP), המאפשרים למשתמשים לנהל שיח טבעי ואינטואיטיבי עם המערכת כדי לחפש מידע. במקום להסתפק בתוצאות מבוססות חוקה (Rule Based), המנוע מגיב בשיחה, מסביר את המידע, ומציע הקשרים מתקדמים המבוססים על ההבנה של שאילתות מורכבות. דוגמאות: ChatGPT, Microsoft Bing Chat.

[13] ראו גם מילון מונחי התקשוב.
[14] עקרונות מדיניות, רגולציה ואתיקה בתחום הבינה המלאכותית, 2023

המונח	הגדרה
יישומי "עוזר אישי" עם יכולת AI יוצרת	תוכנות או אפליקציות שנועדו לסייע בניהול חיי היומיום או המשימות העסקיות של משתמשים. באמצעות יכולות בינה מלאכותית יוצרת, הם יכולים לתכנן פגישות, להציע רעיונות, לספק תובנות מותאמות, ואף ליצור תוכן על פי בקשה אישית. דוגמאות: Siri, Google Assistant, Microsoft Copilot.
תמלול וסיכום פגישות	כלים המשתמשים בטכנולוגיות זיהוי דיבור (ASR – Automatic Speech Recognition) ועיבוד שפה טבעית כדי לתמלל שיחות או פגישות בצורה מדויקת, ולייצר סיכומים תמציתיים וברורים. הם כוללים תכונות מתקדמות כמו זיהוי דוברים, הפרדת דוברים, סינון רעשים, תמלול, איתור תובנות עיקריות ויצירת משימות מעשיות מתוך השיחה. דוגמאות: Otter.ai, Microsoft Teams transcription, Notion AI.
הנחייה ("פרומפט" – prompt)	פרומפט הוא הנחייה למערכת בינה מלאכותית יוצרת. הפרומפט נכתב בטקסט ומבוצע באמצעותו מעין דיאלוג עם המערכת.
תקרית AI	אירוע שבו התממש סיכון AI (צפוי או לא צפוי).