מערך הדיגיטל הלאומי
نظام الديجيتال الوطني
Israel National Digital Agency

משרד המשפטים
MINISTRY OF JUSTICE | وزارة العدل

ייעוץ וחקיקה
OFFICE OF LEGAL COUNSEL AND LEGISLATIVE AFFAIRS
المشورة والتشريع

מדריך לשימוש אחראי בכלי בינה מלאכותית במגזר הציבורי

היבטי משילות וניהול סיכונים

מאי 2026 | אייר תשפ"ו

גרסה 1.0

תוכן עניינים

תקציר מנהלים

אנו נמצאים בתחילת עידן הבינה המלאכותית, תקופה בה יישומי בינה מלאכותית הופכים לנפוצים יותר ויותר, ומעצבים את ההווה ואת העתיד שלנו. בדומה לטכנולוגיות מהפכניות אחרות, בינה מלאכותית עשויה להוות כלי לשינוי חיובי או שלילי. האתגר, עבור ארגוני המגזר הציבורי, הוא להקים את המסגרות והתהליכים המתאימים לכך שטכנולוגיות בינה מלאכותית אכן ינוצלו לטובת השירות הציבורי והחברה כולה.

בהחלטת ממשלה 3574 בדבר האצת הדיגיטציה, יכולות הדאטה והבינה המלאכותית בממשלה באמצעות ענן ציבורי נימבוס, מיום 4.12.25, אומצו העקרונות והמטרות של אסטרטגיית הדאטה והבינה המלאכותית ובכללן גיבוש המסגרות המתאימות לקידום השימוש בבינה מלאכותית. זאת לאור "החשיבות שממשלת ישראל רואה בהמשך חיזוק יכולות הדאטה ופיתוח והטמעת יכולות הבינה המלאכותית, במטרה לייעל ולשפר את פעולות גופי הממשל באמצעות הגברת השימוש האחראי ובניית תשתיות רוחביות", לשון ההחלטה.

על רקע זה, מערך הדיגיטל הלאומי גיבש מדריך לשימוש אחראי במגזר הציבורי. העבודה נעשתה בשיתוף עם מחלקת יעוץ וחקיקה שבמשרד המשפטים ועם משרד החדשנות, המדע והטכנולוגיה, ובהתייעצות עם מגוון רחב של גופים ממשלתיים, מומחים מהמגזר הפרטי, מכוני מחקר ואקדמיה. המדריך מתווה תהליך לגיבוש מדיניות ארגונית אג'ילית, מעשית ומאוזנת, ומציע פרקטיקות מיטביות לשימוש אחראי בבינה מלאכותית במגזר הציבורי, תוך התחשבות בתועלות, באתגרים ובסיכונים הנלווים.

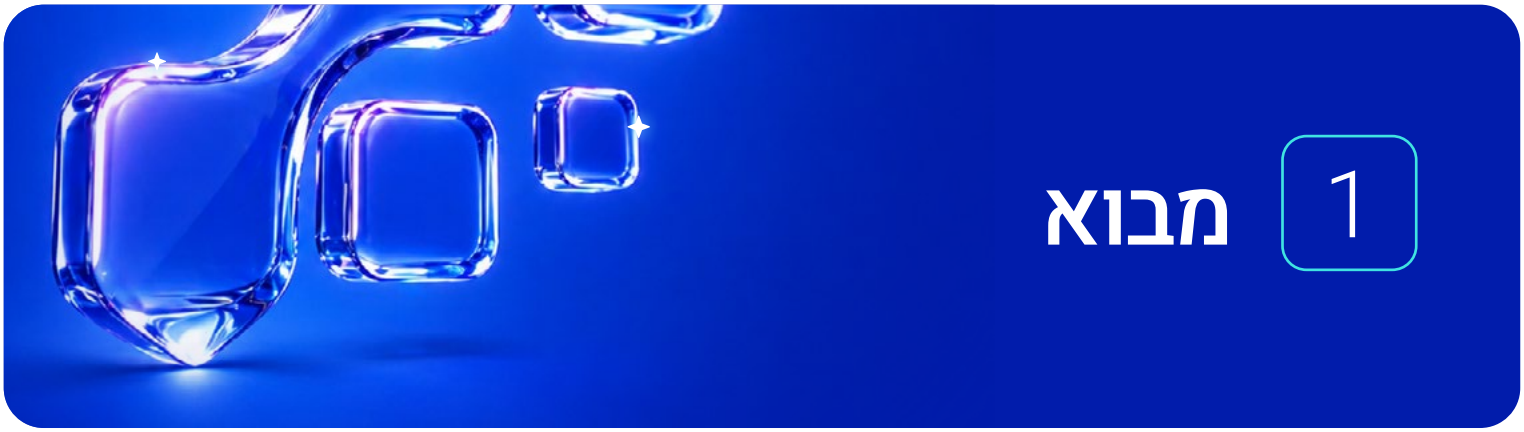
המדריך מתמקד בשלושה צירים:

1. הגדרת העקרונות לשימוש אחראי בבינה מלאכותית, בהתבסס על עקרונות ה-OECD;
2. תפיסה ארגונית סדורה ונהלים, המאפשרים את המימוש של אותם העקרונות ("ממשל בינה מלאכותית");
3. תהליך ניהול סיכונים בפרויקט הטמעת מערכת בינה מלאכותית, בהתאם לסיווג הסיכונים.

להלן המאפיינים המרכזיים של המדריך:

- **שיטת צבעי הרמזור לניהול סיכונים -** תהליך ניהול סיכונים של מערכת AI הכולל ניתוח וסיווג המערכת לפי רמת הסיכון שלה באופן דיפרנציאלי. ככלל, ככל שהסיכון גבוה יותר - כך נדרשים תהליכי אישור ואמצעי בקרה מקיפים יותר. בהתאם לכך, מוצעים מסלולי אישור מערכות AI לפי רמת הסיכון שלהן: **מסלול ירוק -** אישור מהיר יחסית - עבור מערכת בעלת סיכון נמוך; **מסלול צהוב -** אישור כפוף לניהול סיכונים מקיף - עבור מערכת עם סיכון בינוני; **ומסלול אדום -** אישור כפוף לניהול סיכונים קפדני ופיקוח על ידי הנהלת הארגון - עבור מערכת עם סיכון גבוה.
- **תמיכה בנסיינות -** עידוד להקמת **מסלול "כחול"**, המאפשר להתנסות בכלי בינה מלאכותית, באופן מוגבל ומפוקח, כדי ללמוד כיצד הכלי פועל במציאות, וכך למדוד ולהבין את הסיכונים האמיתיים בצורה מדויקת יותר.
- **בחינת הסיכונים והתועלות** באופן מקיף ואחיד, תוך הסתכלות הוליסטית.
- **מיפוי אמצעים טכנולוגיים ואחרים** להתמודדות עם סיכוני בינה מלאכותית ידועים (הטיות, הזיות, פרטיות וכו').
- **פורום משילות AI -** המדריך ממליץ על הקמת פורום משילות AI בארגונים מסוימים, ומתאר את הפונקציות שפורום כזה יכול למלא.
- **תבנית מדיניות ארגונית** לשימוש אחראי בבינה מלאכותית – בהתאם להמלצות במדריך, מצורפת תבנית מוצעת למדיניות שימוש אחראי, שארגונים יכולים לאמץ. התבנית מפרטת תחומי פעולה מומלצים, דוגמת מהלכי שקיפות ומודעות ציבורית, אופן הטיפול בתקרית AI, תוכנית הכשרות בינה מלאכותית והנחיות לעובדי הארגון.
- **מילון מונחים שכיחים** בעולם הבינה המלאכותית.

זוהי הגרסה הראשונה של המדריך. היא כוללת את השינויים וההערות שהתקבלו בעקבות הגרסה להערות הציבור שפורסמה ביוני 2025. בכוונת מערך הדיגיטל הלאומי לעדכן אותה באופן תקופתי, בהתאם להתפתחויות טכנולוגיות ולהתפתחויות המדיניות הממשלתית. אנו מזמינים את כל בעלי העניין להמשיך להעביר הערות והצעות לשיפורים, לכתובת הדוא"ל ResponsibleAI@digital.gov.il.



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, לפיתוח בר-קיימא, לשיפור הפריון, לרווחה חברתית ולהעלאת רמת החיים. השקעה בטכנולוגיות אלו וקידום החדשנות צפויים לאפשר למדינת ישראל להמשיך להיות מדינה מובילה בפיתוחים טכנולוגיים. במגזר הציבורי, יש לבינה המלאכותית את היכולת לשפר ולדייק את השירות לציבור, לייצר ממשל אחוד ולהפוך את הממשל ליעיל וחכם יותר. אולם לצד הפוטנציאל של טכנולוגיות אלו ישנם מאפיינים ומגבלות, כגון אקראיות התוצאות, הטיות והזיות. בנוסף, קיימת נטייה בקרב משתמשים להסתמך על תוצאות מערכות אלה, ללא ביקורת. מאפיינים ומגבלות אלו מעלים סיכונים ואתגרים מסוגים שונים - תפעוליים, חברתיים, ביטחוניים, משפטיים, אתיים ועוד, בפרט בשימוש במגזר הציבורי. סיכונים ואתגרים אלו מקבלים משנה תוקף לאור האפשרות ההולכת וגוברת לאינטגרציה בין מערכות שונות ולשימוש בסוכני AI מקצה לקצה לניהול תהליכים ארגוניים מורכבים. על מנת לממש את הפוטנציאל העצום של טכנולוגיות אלו, נדרש להתמודד עם האתגרים הללו באופן מושכל ושיטתי.

מדריך זה מציע שיטות עבודה מומלצות לגופים ציבוריים המבקשים לשלב מערכות בינה מלאכותית בתחומי פעילותם, בדגש על היבטים ארגוניים, טכנולוגיים ועסקיים, תוך התוויית הליך להערכה, ניהול ומזעור הסיכונים הייחודיים למערכות בינה מלאכותית. זאת, בין היתר, במטרה לצמצם את חוסר הוודאות שנלווה ליישום של טכנולוגיה זו. המדריך אינו גורע מנהלים והנחיות אחרים.

המדריך גובש בשיתוף בין מערך הדיגיטל הלאומי, מחלקת ייעוץ וחקיקה במשרד המשפטים והמרכז לרגולציה ומדיניות בינה מלאכותית במשרד החדשנות, המדע והטכנולוגיה. כמו כן, התקיימו התייעצויות עם מגוון גורמים מגופי המגזר הציבורי לרבות מערך הסייבר הלאומי, הרשות להגנת הפרטיות, רשות האסדרה, החשב הכללי, מנהל הרכש, מספר מובילי דאטה משרדיים (CDO's) וכן עם גורמים מהאקדמיה וממכוני מחקר, חברות פרטיות ומומחים 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 **להיבטי משילות, קרי - התפקידים והאחריות** של הגורמים העיקריים בתהליך השילוב של בינה מלאכותית בגופי המגזר הציבורי, וכן מפרט 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 תמיכה וקידום השימוש בבינה מלאכותית במגזר הציבורי במסגרת תהליך רכש או פיתוח סדור, תוך הפעלת שיקולי תועלת וניהול סיכונים. מטרתו להוות כלי משמעותי לתמיכה בהחלטות על אופי שילוב בינה מלאכותית בארגון. **המדריך מבוסס על מספר עקרונות:**

- **תפיסת ניהול סיכונים –** שימוש אחראי במערכות בינה מלאכותית אין משמעותו הימנעות מוחלטת מסיכונים. המדריך מציע גישה דיפרנציאלית: ככלל, ככל שהסיכונים גבוהים יותר, כך גם נדרשים אמצעי הפחתת סיכונים ותהליכי בקרה מקיפים יותר. מוצעים שלושה מסלולים לקידום פרויקט הטמעת בינה מלאכותית, לפי שיטת צבעי הרמזור (ירוק, צהוב, אדום).

- **נסיינות –** הסיכונים הכרוכים בשימוש במערכת מסוימת אינם בהכרח ברורים מראש באופן מוחלט. לעיתים, יש צורך להשתמש במערכת כחלק מתהליך התנסות ולמידה. לפיכך, לצד שלושת המסלולים הנ"ל, המדריך מציע מסלול כחול המאפשר לארגון להתנסות במערכת מסוימת באופן מוגבל ומבוקר.

- **כלליות וגנריות –** המדריך מהווה המלצה כללית שעליה ניתן לבצע את ההתאמות הרלוונטיות, בהתאם לתחום התוכן או למערכת הציבורית שבו מופעלת המערכת או שבו פועל הארגון.

- **הכוונה תהליכית –** המדריך מצביע על התהליכים העיקריים שמומלץ לאמץ בארגון, ואינו מתווה הנחיות לרף תוצאתי מסוים.

- **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 עולה בקנה אחד עם החלטת ממשלה 3574 מה-4.12.25 בדבר "האצת הדיגיטציה, יכולות הדאטה והבינה המלאכותית בממשלה", הוא מתכתב עם מסמך AI Journey, שעוסק בהטמעת מערכת בינה מלאכותית בארגון ציבורי בהיבטים טכניים ועסקיים. המדריך גם מתכתב עם הנחיות ומסמכי מדיניות נוספים: הנחיות המערך בדבר ניהול סיכוני תקשוב[1]. מסמך "עקרונות מדיניות, רגולציה ואתיקה בתחום הבינה המלאכותית", שהוכן על ידי משרד החדשנות, המדע והטכנולוגיה ומחלקת ייעוץ וחקיקה במשרד המשפטים[2], דוח "התכנית הלאומית לבינה מלאכותית – תמונת מצב 2025", והמדריך לניהול סיכונים ברגולציה של רשות האסדרה.

- **התחשבות במסגרות אסדרה וסטנדרטים מובילים –** 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**).
  3. המתודולוגיה הנלווית לאמנה – מתודולוגית (HUDERIA).
  4. הנחיות ומדריכים דומים של **ארה"ב**, **בריטניה**, **אוסטרליה**, **קנדה**, **וסינגפור**.
  5. תקינה בינלאומית[4].

- **דגש על שימוש בסביבת הענן הממשלתית –** לממשלת ישראל יש היצע שירותים רחב של כלי בינה מלאכותית, הזמין הן ישירות על ידי ספקיות הענן בפרויקט "נימבוס" (רובד אחד), והן על ידי ספקים נוספים (מוצרי רובד חמש)[5], בכלל זאת גם יישומי **בינה מלאכותית יוצרת**, מהמתקדמים בעולם, המותקנים באתרים 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, והעדפה זו משתקפת במדריך זה.

---

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 הסופי של הצוות הבין משרדי לבחינת אסדרת בינה מלאכותית בסקטור הפיננסי.

3 האמנה חלה בעיקר על שימוש בבינה מלאכותית במגזר הציבורי, ומתמקדת בסיכונים הנשקפים לזכויות אדם, לדמוקרטיה ולשלטון החוק. 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 המסמכים הבאים שימשו כבסיס למדריך: **תקני ISO** – תקן ISO-23894 להנחיות לניהול סיכונים הנוגעים למערכות AI; תקן ISO 42001 לניהול מערכות AI; NIST.AI 100-1 – מסמך מסגרת לניהול סיכוני בינה מלאכותית שנכתב על ידי המכון הלאומי לתקנים וטכנולוגיה בארה"ב (NIST).

5 קטלוג המוצרים המאושרים לרכישה ברובד 5 זמין כאן.

1.4 מה כולל המדריך?

למדריך שלושה פרקים קצרים:

1. העקרונות הבסיסיים לשימוש אחראי בבינה מלאכותית;
2. תפיסה ארגונית מומלצת להטמעת אותם עקרונות, המתייחסת לגורמים הרלוונטיים בארגון ותחומי אחריות מוצעים עבורם;
3. תהליך לניהול סיכונים לשילוב של מערכת בינה מלאכותית בארגון.

המדריך כולל שישה נספחים: פירוט תפקידי הנהלת הארגון, פירוט תפקידי מוביל בינה מלאכותית הארגוני, קווים מנחים למשתמשי קצה, מתודת ניהול סיכונים מוצעת, תבנית למדיניות ארגונית של שימוש אחראי בבינה מלאכותית ומילון מונחים.

מדריך זה אינו מהווה מסמך משפטי. ישנן סוגיות בעלות היבטים משפטיים שמדריך זה אינו ממצה את העיסוק בהן, הגם שיש ממשק בין הדברים. מחלקת ייעוץ וחקיקה במשרד המשפטים מגבשת בימים אלו מדריך משפטי (להלן "המדריך המשפטי"), שיעסוק בהיבטים המשפטיים שמתעוררים, דוגמת היבטי שקיפות, שוויון והתמודדות עם הטיות, עמידה בחובות המשפט המינהלי כגון חובת ההנמקה, וכן היבטי זכויות יוצרים. מחלקת ייעוץ וחקיקה עומדת לרשות הלשכות המשפטיות לליווי בשאלות משפטיות שעשויות להתעורר.

המדריך מנוסח בלשון זכר מטעמי נוחות בלבד, אך כל האמור בו משמעו גם בלשון נקבה.

ניתן לפנות אלינו בכל התייחסות, הערה ושאלה לתיבת הדוא"ל:

ResponsibleAI@digital.gov.il

2 עקרונות לשימוש אחראי בבינה מלאכותית בארגון

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

את התפיסה הזו 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[6].

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 וליכולות 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.

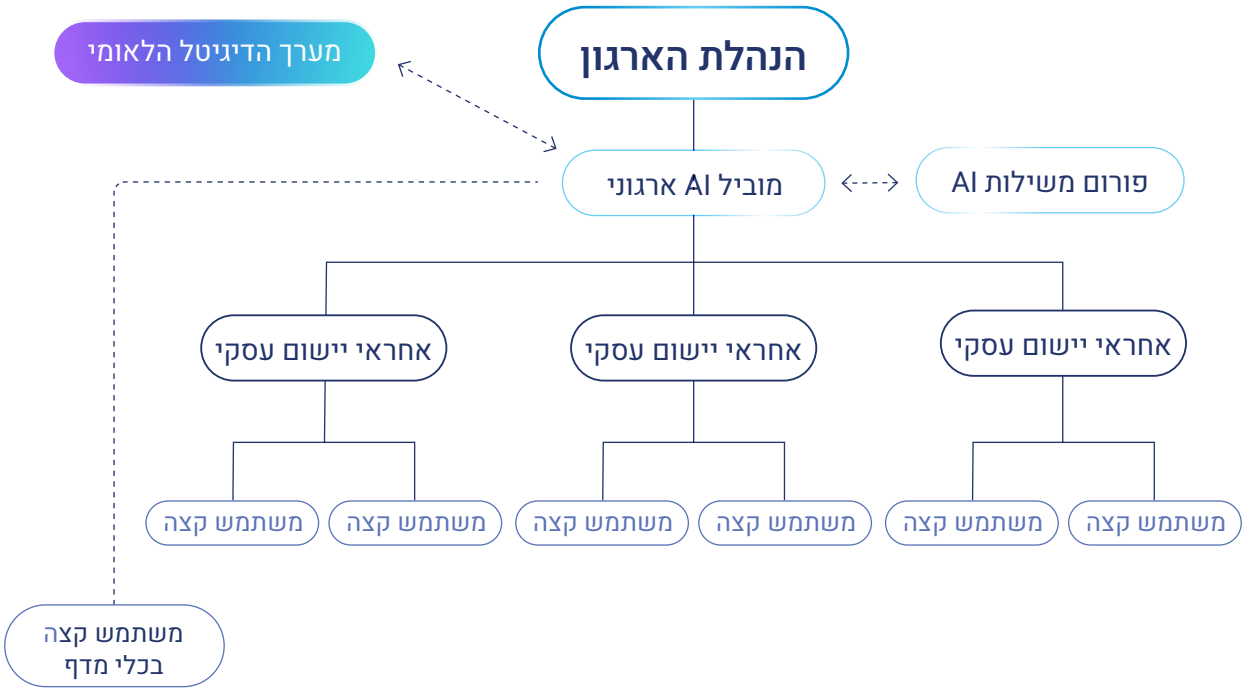
4. **ביטחון ובטיחות של המערכת –** בפיתוח ובשימוש במערכות בינה מלאכותית, חשוב להקפיד על כך שהן תהיינה אמינות, בטוחות ומאובטחות לאורך כל מחזור החיים שלהן. זאת על מנת שבמצבי שימוש מתוכננים או שאינם מתוכננים, וכן במקרים של שימוש שגוי או היווצרות של תנאים חיצוניים מסוכנים - המערכות יפעלו כראוי ולא יהוו סיכון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

5. **אחריותיות –** מצופה מהארגונים לגלות אחריות לתפקודה התקין של מערכת בינה מלאכותית, ולקיום העקרונות האמורים, בהתאם לתפקידם ולאפשרויות הטכנולוגיות הזמינות. לשם כך, יש לפתח מנגנוני ניהול סיכונים ולאמץ כללי התנהלות פנימיים להתמודדות עם סיכונים (לרבות הטיות, פגיעה בפרטיות, פגיעה בזכויות אדם, פגיעה בזכויות עובדים, פגיעה בקניין רוחני, סיכוני בטיחות ועוד).

על פי עקרונות אלו, שימוש אחראי בבינה מלאכותית אינו רק עניין של ציות לדרישות טכניות, אלא שהוא מצריך הפנמה והטמעה, **ברמה המוסדית**, של הערכים המדוברים. לפי תפיסה זו, יש להשריש 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 שורה של פעולות, הקמת מנגנונים וחלוקת תפקידים ותחומי אחריות בארגון המכוונים למטרה זו. כמו כן, העקרונות הנ"ל אינם חלים באופן גורף וחד-ערכי. על הארגון ליישם את העקרונות באופן הולם, מעשי ומתואם למציאות שבה הוא פועל. למשל, הפעלת עיקרון השקיפות אין משמעותו שארגון צריך לחשוף כלים מסווגים.

6 אין מדובר בתרגום מילולי של העקרונות.

3 ממשל בינה מלאכותית
תפקידי מפתח ותחומי אחריות



3.1 הנהלת הארגון

הובלת תהליכי שימוש אחראי בבינה מלאכותית היא באחריות הנהלת הארגון, אשר מצופה לספק את המעטפת התפעולית והתקציבית המתאימה. לשם כך, מוצע שההנהלה תמנה, בין היתר, מוביל בינה מלאכותית ארגוני, תקים פורום משילות בינה מלאכותית ככל שרלוונטי, תקצה להם את המשאבים הנדרשים לבצע את הפעולות המתוארות במדריך זה, ותנגיש לציבור דוחות תקופתיים של פעילות הארגון בתחום. מצופה גם כי הנהלת הארגון תתווה את האסטרטגיה הארגונית לשימוש בבינה מלאכותית, ובכלל זה תתעדף את התחומים שבהם רצוי לשלב בינה מלאכותית.

3.2 מוביל בינה מלאכותית ארגוני

מוביל בינה מלאכותית ארגוני ("מוביל AI") הוא האחראי ברמה הארגונית על אסטרטגיית השילוב של מערכות בינה מלאכותית בתהליכים העסקיים של הארגון, ובכלל זאת על גיבוש ויישום נהלים לשימוש אחראי בבינה מלאכותית, לרבות מדיניות ניהול סיכוני AI ארגונית. מוצע, כברירת מחדל, למנות את מוביל הדאטה הארגוני (CDO) לתפקיד זה, אך אין מניעה למנות גורם אחר (למשל מנמ"ר או סמנכ"ל אסטרטגיה ומדיניות) ואף ועדה משרדית. העיקר הוא כי הגורם שימונה יהיה בעל ידע רלוונטי בנושא בינה מלאכותית, ושינתנו לו משאבים פנים-ארגוניים (תקצוב, כוח אדם, הכשרה) וכן וסמכות, שיאפשרו לו למלא את תפקידיו.

למידע ופרטים על **תפקידי מוביל בינה מלאכותית ארגוני**, ראו נספח א.

3.3 פורום משילות בינה מלאכותית

בארגון גדול, או בארגון שבו מערכות בינה מלאכותית בעלות רגישות גבוהה מאוד או כאלה שמשולבות אחת עם השנייה במסגרת תהליך עסקי מורכב במיוחד, מוצע להקים פורום משילות בינה מלאכותית ארגוני, שיוכל לייעץ למוביל AI הן בהיבטים רוחביים (אסטרטגיה ומדיניות להטמעת בינה מלאכותית בארגון, השלכות של שימושי בינה מלאכותית על עובדי הארגון, וכן השלכות רוחב בפן החברתי-כלכלי), הן ביחס לפרויקטים ספציפיים שנשקף מהם סיכון גבוה, והן בנוגע להתמודדות עם תקריות AI. הרכב הפורום נתון לשיקול הדעת של הנהלת הארגון, אך מומלץ לאמץ מבנה "רזה" של 4-3 גורמים, עם אפשרות להתייעץ עם גורמים נוספים במידת האפשר. הרעיון מאחורי פורום משילות הוא להקים גוף תומך במוביל ה-AI, ולא ליצור שכבה בירוקרטית עודפת. על הארגון לוודא שהפורום אכן פועל כך.

3.4 אחראי יישום עסקי

אחראי היישום העסקי, הוא גורם בארגון שאמון על תהליך עסקי מסוים, אשר בוחן הטמעה של טכנולוגיית בינה מלאכותית לקידום תהליך זה. כך למשל, הוא יכול להיות מנהל אגף האמון על הסדרת תחום מסוים, מתן הטבות, או הנפקת רישיונות והיתרים, וכן גורם שאחראי על תהליך רוחב בארגון כגון חשבות, תכנון ותקצוב, ייעוץ משפטי, טכנולוגיות מידע וכדומה. כאשר גורם זה מבקש להטמיע כלי בינה מלאכותית עבור תהליך עסקי מסוים, עליו לרכז את ביצוע הליך ניהול הסיכונים ביחס לאותו פרויקט.

למידע ופרטים על תפקידי **אחראי היישום העסקי**, ראו נספח ב'. לתהליך ניהול סיכונים, ראו נספח ג'.

3.5 משתמש קצה

משתמש קצה הוא עובד בגוף ציבורי אשר עושה שימוש במערכת בינה מלאכותית לצורך מילוי תפקידו. בין אם מדובר במערכת אשר פותחה לצרכי הארגון, או בכלי מדף הזמין לקהל הרחב, מצופה ממנו לפעול באופן אחראי ושקול. לקווים מנחים עבור **משתמשי הקצה** בארגון, ראו נספח ד'.

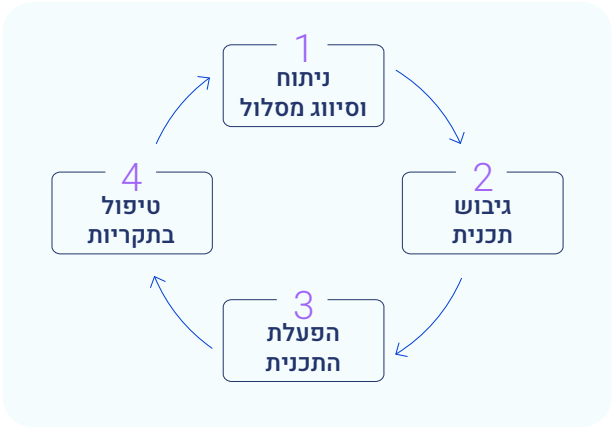
3.6 מערך הדיגיטל הלאומי

מערך הדיגיטל הלאומי מציע שירותים לסיוע וליווי בהטמעת שימוש אחראי בבינה מלאכותית במגזר הציבורי, בראשם מרכז משילות AI שיכול לסייע עם יישום תהליך ניהול סיכונים בפרויקט, וכן מרכז מצוינות שמסייע עם הטמעת פרויקטים של בינה מלאכותית. המערך גם בונה ומקיים הכשרות באמצעות בית הספר להכשרות דיגיטליות "הדיגיטלית". בהמשך, המערך יפרסם דוגמאות של פרויקטים שהוטמעו בהצלחה עם תכניות ניהול סיכונים לדוגמה, וכן יחבר בין ארגונים לפי הצורך כדי למנוע כפילויות ולאפשר שיתופי ידע רלוונטיים. לבקשת הארגון, יוכל מערך הדיגיטל לחוות את דעתו לגבי ניהול הסיכונים של שימושים מסוימים במסלול אדום, וכן לשמש כמקור ידע עבור מוביל ה-AI.

4 תהליך ניהול סיכונים של מערכת AI

ככלל, התהליך העסקי מתחיל עם הצבת יעד מקצועי, אפיון התהליך הרצוי, תכנון מערכת מידע תומכת ובחירת מוצר בינה מלאכותית מתאים. תהליך זה דורש הבנה של הטכנולוגיה – יכולותיה כמו גם מגבלותיה – התועלות והסיכונים הנשקפים, והסביבה התפעולית והרגולטורית שבה המערכת אמורה לפעול. ראו לעניין זה את מדריך AI Journey. להלן יפורט תהליך ניהול סיכונים לבחינת שימוש במערכת משולבת בינה מלאכותית.

התהליך מורכב מארבעה שלבים, שכל אחד כולל תתי-שלבים. אחראי היישום העסקי אחראי על הפעלת השלבים הללו, **תוך תיעוד פנימי של כל תתי השלבים על מנת לאפשר בקרה.**
כפי שניתן לראות, התהליך הוא מעגלי ומתרחש לאורך כל מחזור החיים של המערכת בין אם היא בפיתוח, בפריסה, בייצור, בשדרוג או בהוצאה מכלל פעולה. במצבים רבים רכיב בינה המלאכותית יוטמע בתוך מערכת תפעולית קיימת ותהליכים עסקיים שוטפים, ואז תהליך ניהול הסיכונים המוצע כאן יתייחס לרכיב זה.



להלן תיאור קצר של שלבי התהליך.

שלב 1 - ניתוח וסיווג המערכת: המרכיב העיקרי של חלק זה הוא סיווג המערכת. סיווג זה מכתיב כיצד נדרש להתייחס אליה: מערכת בעלת סיכון נמוך מאוד אינה מצריכה אותן בקרות בהשוואה למערכת בעלת סיכון גבוה. המדריך ממליץ על התייחסות דיפרנציאלית, לפי שיטת צבעי הרמזור (ראו פירוט ביחס למסלולים השונים בנספח ג'):

- **מסלול ירוק,** מהיר ופשוט, למערכות בעלות סיכון נמוך;
- **מסלול צהוב,** מקיף יותר, למערכות בעלות סיכון בינוני;
- **מסלול אדום,** עם תנאים מחמירים, למערכות בעלות סיכון גבוה;
- נוסף על כך, מוצע לקבוע **מסלול כחול** שנועד לאפשר נסיינות של מערכות אשר נשקף מהן סיכון בינוני או גבוה, בהיקפים מצומצמים, על מנת לאסוף נתונים על המערכת לפני פריסה מלאה, במסגרת בטוחה ומפוקחת.

לצורך הסיווג הראשוני, על אחראי היישום העסקי להיוועץ עם הגורמים הרלוונטיים, לפי רמת הסיכון המסתמנת, ולקבוע את המסלול בשיתוף מוביל ה-AI הארגוני. בכל עת, יכול אחראי היישום להציע שינוי בסיווג, לצד המחמיר או המקל, בהתאם לממצאים ולהתייעצויות שהוא מקיים. הוא גם יכול, בכל עת, להציע כי הפרויקט יועבר למסלול הכחול.

דגש: שלב קביעת הסיווג צריך להתממשק עם הליכי הפיתוח והרכש של המערכת הנבחנת. כל הליך פיתוח או רכש של כלי צפוי לכלול ניסויים, בדיקות, ושיח עם המפתחים או עם הספק, כדי לחדד את מאפייני המערכת ולשפר את ביצועיה. במסגרת זו, ניתן לצפות כי סיווג המערכת לא יתבצע בצורה לינארית או סטרילית, אלא להיפך – הוא יכלול מספר סבבים של התייעצויות. תתי השלבים מתוארים בצורה פשוטה כדי לציין אבני דרך כלליות, אך מדובר בהליך דינאמי.

שלב 2 – גיבוש תוכנית ניהול סיכונים: מטרת שלב זה היא גיבוש תוכנית ניהול סיכונים המתאימה לסיווג המערכת ולסיכונים המשתקפים ממנה. כאשר מדובר במערכת במסלול אדום, מומלץ להיוועץ גם עם פורום המשילות. **דגש**: כמו בשלב 1, גיבוש תוכנית ניהול סיכונים עשוי להתבצע בצורה שאינה לינארית אלא במספר צעדים וחילופי רעיונות ועמדות.

שלב 3 – הפעלת התוכנית: בשלב זה, המערכת נמצאת בשימוש ונדרש לנקוט באמצעי ניהול הסיכונים שאושרו. במקרים המתאימים, על אחראי היישום לוודא שדרישות שקיפות לציבור, כפי שנקבעו ביחס למערכת, מולאו. הוא גם צריך לנטר את המערכת כדי לוודא שהיא פועלת כמצופה, ולדווח על בעיות ותקריות AI, בהתאם לתוכנית.

שלב 4 – טיפול בתקריות: שלב זה מתייחס לטיפול בבעיות ובתקריות AI בזמן אמת, בהתאם לתוכנית ניהול סיכונים, לרבות דיווחים להנהלת המשרד. ללא קשר עם סיווג המערכת, תקריות AI חמורות מצריכות מענה מיידי. גם כאן, נכנסים שיקולים של שקיפות לציבור, ומומלץ להתייעץ עם פורום המשילות ועם הלשכה המשפטית. במקרים רלוונטיים, צריך להיערך להגבלת או להפסקת השימוש במערכת.

לפרטים נוספים והרחבה על מתודת ניהול סיכונים ועל המסלולים השונים - ראו נספח ג'.

נספח א'

תפקידי מוביל בינה מלאכותית ארגוני

פעולות מומלצות	פירוט
1 - פיתוח מדיניות	
1.1 שימוש אחראי	קביעת מדיניות ארגונית כללית לשימוש אחראי בבינה מלאכותית, בליווי פורום משילות AI ככל שרלוונטי. מומלץ כי המדיניות הפנימית, במלואה או בחלקה בהתאמות הנדרשות לפעילות הארגון, תתבסס על העקרונות לשימוש אחראי (פרק 2 למדריך), ועל תהליך ניהול הסיכונים המפורט במדריך זה (נספח ג'). חשוב לוודא שהמדיניות מתכתבת עם מדיניות סייבר, מדיניות פרטיות של הארגון, המדריך המשפטי, מדריך ניהול סיכונים רגולטוריים של רשות האסדרה. נספח ה' מהווה תשתית למדיניות ארגונית, אשר ניתן לשנות ולהתאים לצרכי הארגון.
1.2 משילות דאטה המשמש לבינה מלאכותית	בתיאום עם ה-CDO הארגוני ככל שאינו הגורם המוביל AI – קידום פיתוח כללים לניהול נתונים ארגוניים המשמשים מערכות ומודלי ה-AI, כדי לוודא שהמידע שמשמש כבסיס ליישומי בינה מלאכותית הוא איכותי, מדויק, ועדכני. 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. מוצע להתייחס לצורך לשמור על המידע הארגוני כדי לוודא רציפות תפקודית, 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 הגנת סייבר, DPO (Data Protection Officer), חשבות וכדומה. בין היתר, מוצע להסדיר את ההרשאות למשתמשים פנימיים וחיצוניים, וכן לבחון האם יש כלים מבוססי AI שרצוי להנחות על איסור גורף של השימוש בהם בארגון, או להגביל את השימוש בהם בהתקיים תנאים ספציפיים, ולהודיע על כך בתוך הארגון. מומלץ להשתמש בשירותי בינה מלאכותית בסביבות הענן הממשלתי (נימבוס) על פני רכש מוצרים בערוצים אחרים, על מנת לאפשר שימוש מאובטח יותר בטכנולוגיה זו. **דגש:** רכישת כלי מדף הכוללים שירותי בינה מלאכותית, כפופה להוראות התכ"ם הרלוונטיות.[7]
1.4 מדיניות טיפול בתקריות AI	רצוי להתייחס לסוגיות כמו: הקמת צוות לטיפול בתקריות AI בארגון (AI Incident Response Team), הפסקת השימוש, שינוי האלגוריתם, שינוי/עדכון המודל או המידע עליו המודל מאומן, דיווח לגורמים הרלוונטיים בתוך הארגון (למשל, הנהלת הארגון, ממונה הגנת פרטיות) ומחוץ לו (כדוגמת מערך הסייבר, הרשות להגנת הפרטיות, ויחידת ההגנה בסייבר במערך הדיגיטל הלאומי).
2 - הטמעת המדיניות	
2.1 הקמת מנגנונים	הקמת המנגנונים שפותחו בהתאם לסעיף 1 לעיל, ספציפית: פורום משילות AI, צוות התמודדות עם תקריות בינה מלאכותית (AI Incident Response Team), הפצת הנהלים והתבניות לאישורים רלוונטיים, הכנות לפרסום מידע רלוונטי באתר הארגון.
2.2 הובלת תהליכי אוריינות AI	יחד עם יחידת כוח אדם, קביעת תוכנית למידה ארגונית שתכלול השתלמויות והדרכות פנימיות וחיצוניות בנושא בינה מלאכותית ושימוש אחראי (לרבות הדרכות של בית ספר "הדיגיטלית" של מערך הדיגיטל הלאומי).

7 הוראת תכ"ם 7.10.7 - התקשרות לרכישת שירותי בינה מלאכותית; לגבי שירותי בינה מלאכותית המוצעים על ידי AWS ו-Google, ראו בהוראת תכ"ם 16.12.2 "אספקת שירותי ענן ציבורי של AWS ו- Google למשרדי הממשלה"; לגבי שירותי בינה מלאכותית המוצעים על ידי חברת Salesforce במסגרת מכרז מרכזי, ראו הודעת תכ"ם 16.2.4 "אספקת שירותי Customer Relationship Management (CRM) בענן"; לגבי שירותי בינה מלאכותית המוצעים במסגרת נימבוס על ידי ספקי צד ג', ראו הוראת תכ"ם 16.12.1.2 - רכש שירותי צד ג' בענן בהליך רכש עצמאי של המזמין.

פעולות מומלצות	פירוט
3 - תפעול המדיניות ברמה הארגונית	
3.1 הנחיות מקצועיות בתוכניות לניהול סיכוני AI	מוביל ה-AI ייקח חלק בתהליכי גיבוש תוכנית לניהול סיכונים עבור מערכת מלאכותית שהארגון מבקש להטמיע. מידת מעורבותו תהיה בהתאם לרמת הסיכונים הנשקפים. תפקידו להנחות מקצועית את האחראי העסקי בהיבטי שימוש אחראי בבינה מלאכותית. כאשר מדובר במערכת שמסווגת כ"אדומה", מוביל ה-AI גם אחראי על אישור התוכניות לניהול סיכונים. ראו נספח ג'.
3.2 יצירת רשימה ארגונית של שימושי בינה מלאכותית ותמונת מצב של הסיכונים ודירוגם	יצירת רשימה ארגונית של מערכות בינה מלאכותית בשימוש או בתהליכי אישור, יחד עם תמונת מצב עדכנית לגבי מיפוי הסיכונים ודירוגם. ברשימה יצוין אחראי היישום העסקי האמון על ניהול הסיכונים ביחס לכל המערכת; וכן פירוט ביחס למעקב 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 שקיפות	פרסום מידע כללי על מערכות בינה מלאכותית שנמצאות בשימוש הארגון, בהתאם למדיניות לשימוש אחראי, לתוכנית ניהול סיכונים (ראו פירוט בנספח ג') ולמדריך המשפטי שיפורסם. זאת, בכפוף לכך שהפרסום לא מהווה חשיפת מידע מסווג או מוגן מסיבות שונות כגון חשיפת שיטות וכלים. מערך הדיגיטל מנהל דאשבורד ממשלתי שמפרסם את שימושי בינה מלאכותית על ידי הממשלה.[8] מומלץ להעביר את המידע על אודות המערכת למערך הדיגיטל לשם פרסומו בדאשבורד.
3.4 טיפול בתקריות AI	בהתאם לנהלים שנקבעו על פי סעיף 1.4.
3.5 תחקור והפקת לקחים	בחינה פרואקטיבית ומקיפה של תקריות AI, הכוללת שחזור ההחלטות השונות הקשורות לאפיון ותפעול המערכת, לרבות האופן בו נוהלו הסיכונים. הפקת לקחים – לפי תוצאות התחקיר ובהתייעצות עם פורום משילות AI והנהלת הארגון: (1) בחינת האפשרות של שינויים באופן בו מופעלת המערכת ומנוהלים הסיכונים; (2) במקרה של תקריות חמורות, בחינת האפשרות של הפסקה (זמנית או קבועה) של השימוש במערכת; (3) במקרים שבהם נגרם נזק משמעותי או חמור, יידוע הציבור והנפגעים.
3.6 דיווח פנימי	תיעוד ודיווח שנתי למנכ"ל ולפורום משילות AI (ככל שרלוונטי) 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.

8 מערכות בינה מלאכותית גם עתידות להתפרסם במערכת AI Watch שהקים מערך הדיגיטל הלאומי, אשר נועדה להוות מאגר כולל ולייצר תמונת מצב על השימוש בבינה מלאכותית בארגונים השונים, עבור גורמי ממשל ותושבים. האתר עתיד להתעדכן.

נספח ב'

תפקידי אחראי היישום העסקי

פעולות מומלצות	פירוט
1. אפיון פעולת רכיבי הבינה המלאכותית בפרויקט	לפי הצורך ובשלבים המתאימים בתהליך: ✦ שילוב "גורמי המעטפת" בארגון כגון DPO, CDO, CISO, אחראי פיתוח אגף טד"ם, אחראי ענן, ממונה אבטחת מידע, משתמשים ולקוחות. ✦ בהתאם לאופי המערכת ולהיבטים המשפטיים המתעוררים, יש להיוועץ עם הלשכה המשפטית על מנת למפות את הדרישות המשפטיות שבהן המערכת תידרש לעמוד. רצוי לעשות זאת 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. בכל הקשר לאפיון הפרויקט, ראו את מסמך AI4Journey.
2. ניהול סיכוני AI	הפעלת תהליך ניהול סיכונים המפורט בנספח ג' (סיווג הסיכון, תוכנית ניהול סיכונים, בקרות ודיווחים למוביל AI). תיעוד כלל הממצאים (מיפוי התועלות והסיכונים של מערכת ספציפית באופן מקיף, מפורט ובהיר, גיבוש תוכנית הפחתת סיכונים).
3. הוראות למשתמשי קצה	יש להנגיש למשתמשי קצה בארגון את ההנחיות הרלוונטיות לגבי השימוש האחראי במערכת, לרבות הנחיות מספק חיצוני, ככל שיש. ככל שלא ניתן לספק הנחיות כאלו, יש להעביר למשתמשי הקצה את המידע הכלול במדריך למשתמש קצה (נספח ד'), בתור הנחיות בסיס, ולעבות בהנחיות נוספות בהתאם לאפיון המוצר ולתובנות שיתקבלו בהתאם להליך ניהול סיכונים נספח ג').

נספח ג'

מתודת ניהול סיכונים של מערכת בינה מלאכותית

נספח זה נועד לספק בסיס אחיד לניהול סיכוני בינה מלאכותית ביחס למערכת פרטנית שנבחנת לשימוש. הוא אינו עוסק בניתוח סיכונים לצורך פיתוח רגולציה או פיתוח מדיניות רוחבית. אם עולים היבטים של גיבוש או שינוי של רגולציה ומדיניות ציבורית, יש לפנות למדריך לניהול סיכונים ברגולציה שפורסם על ידי רשות האסדרה.

המתודה המוצעת כאן מבוססת על שיטת צבעי הרמזור, עם שלושה מסלולים עיקריים לפי רמת הסיכון של המערכת הנבחנת (ירוק, צהוב, אדום), וכן מסלול נסייני (כחול).

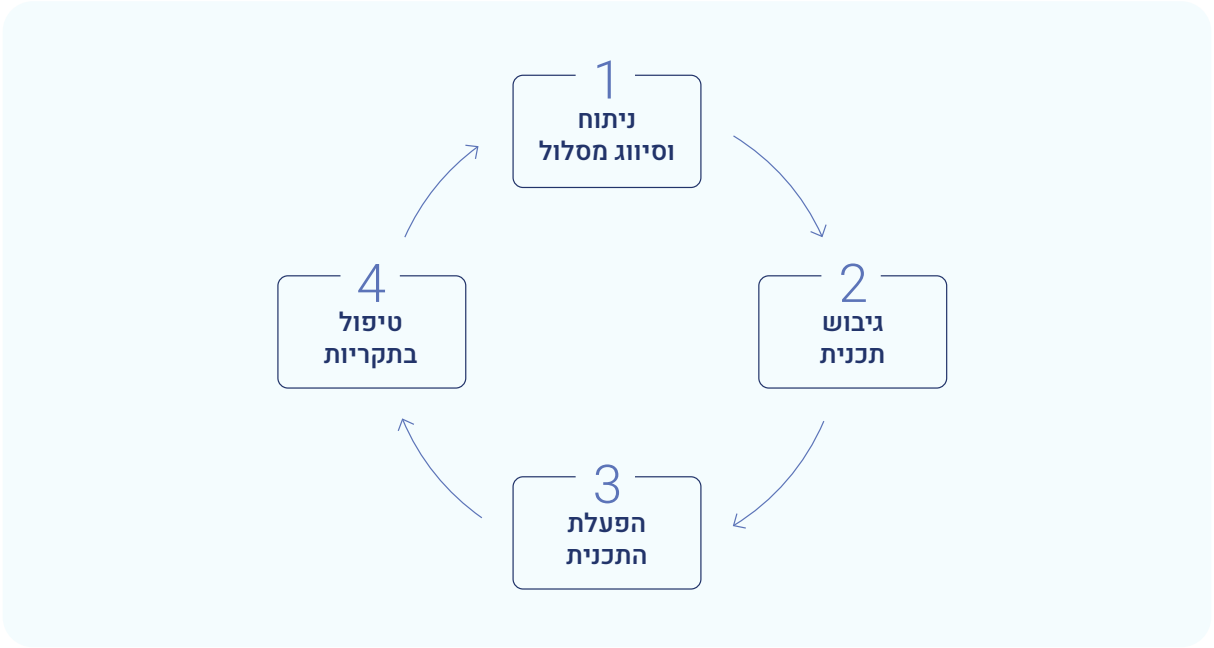
מתודת ניהול הסיכונים מורכבת מארבעה חלקים:

1. שלבים בהליך ניהול סיכונים של מערכת
2. סיווג תועלות וסיכונים
3. פירוט המסלולים
4. גיבוש תוכנית ניהול סיכונים

מוביל הבינה המלאכותית הארגוני (**"מוביל AI"**) יכול, לפי שיקול דעתו, להרחיב את המודל שלהלן, או לשנות אותו (אך מומלץ שהוא יכלול את המרכיבים היסודיים באופן כזה או אחר).

1. שלבים בהליך ניהול סיכונים של מערכת

תהליך זיהוי וניהול סיכונים הוא תהליך חזרתי שמתחיל כבר בשלב אפיון המערכת ומתבצע באופן שוטף במהלך כל חיי המערכת על פי השלבים הבאים:



להלן טבלה המתארת את הפעולות הכלולות בכל שלב. חשוב לציין שכל שלב מצריך תיעוד ההחלטות שהתקבלו.

תיאור השלב	פעולה
שלב 1 - ניתוח וסיווג המערכת	מטרה: לבחון באופן הוליסטי את התועלות והסיכונים של מערכת הבינה המלאכותית המבוקשת ולהחליט על המסלול לניהול סיכונים ביחס למערכת בינה מלאכותית שנבחנת (ירוק/צהוב/אדום). כאשר מעוניינים לנקוט בגישה של נסיינות בטרם החלטה על מסלול ניהול סיכונים קבוע, ניתן לפעול במסלול הכחול (ראו להלן פירוט).
סיווג ראשוני	✦ זיהוי הצורך העסקי, בחינת החלופות והיתרון היחסי המצופה ממערכת בינה מלאכותית; ✦ מיפוי ראשוני של תועלות וסיכונים פוטנציאליים העולים מהמערכת הנבחנת. **בחינת התועלת כוללת בחינת:** ✦ היקף עוצמת התועלת ✦ סבירות התממשות **בחינת הסיכונים כוללת בחינת:** ✦ היקף עוצמת הסיכונים ✦ סבירות התממשות ראו טבלה בחלק 2 לנספח זה.
התייעצות	**מוצע לקיים התייעצויות בהתאם לסיווג המסתמן.** ✦ מסלולים ירוק וצהוב: להיוועץ עם מוביל ה-AI. ניתן גם להיוועץ עם גורמים רלוונטיים נוספים, לפי הצורך (ראו רשימה להלן). ✦ מסלול אדום: להיוועץ עם מוביל ה-AI ועם גורמים רלוונטיים בתוך הארגון ומחוץ לו, לפי הצורך. **להלן הגורמים הרלוונטיים בתוך הארגון ומחוץ לו, אשר ניתן להיוועץ איתם בכל שלב (רשימה לא מחייבת, אך לא ממצה):** ✦ היבטי פרטיות: DPO (הממונה על הגנת הפרטיות בארגון); ✦ השפעות במרחב הסייבר על מערכות ונכסים אחרים ברשת, ממשקים, תקשורת ומידע: ממונה הגנת סייבר בארגון; ✦ שימוש בענן: אחראי הענן; ✦ הטמעת מרכיבים טכניים שיכולים להפחית את הסיכונים: מנהל אגף טד"ם (טכנולוגיות דיגיטליות ומידע); ✦ איכות ונגישות הנתונים: CDO (ממונה דאטה ארגוני); ✦ עובדי היחידות בארגון עצמו ; ✦ שותפי הארגון לרבות ספקיו; ✦ גורמים עסקיים חיצוניים הצפויים לעשות שימוש במערכת; ✦ מקבלי השירות בארגון לרבות מפוקחיו; ✦ הציבור בכללותו. לצד האמור, במסלולים השונים, בהתאם לאופי המערכת ולהיבטים המשפטיים המתעוררים, יש להיוועץ עם הלשכה המשפטית על מנת למפות את הדרישות המשפטיות שבהן המערכת תידרש לעמוד.
הערכת תועלות וסיכונים	✦ דירוג התועלות באופן כולל: תועלת נמוכה מאוד/נמוכה/בינונית/גבוהה/גבוהה מאוד ✦ דירוג הסיכונים באופן כולל: נמוך מאוד/נמוך/בינוני/גבוה/גבוה מאוד. ✦ בחינת הצורך להעביר את המערכת למסלול כחול, לפי הפרמטרים של מסלול זה. ככלל, אם הסיכונים אינם מובנים עד הסוף, או שקשה למדוד אותם בצורה מדויקת מספיק לפני פריסתה של המערכת, מומלץ להעביר את התהליך למסלול הכחול. ראו פרטים נוספים לגבי מסלול זה כאן. ✦ בהתאם להערכה, ניתן לקבוע באופן ראשוני את המסלול המומלץ לאישור שימוש במערכת. ✦ **דגש:** ככלל, כאשר ממערכת נשקף סיכון אחד בדירוג "גבוה מאוד", היא תסווג למסלול אדום. ראו חלק 3 לנספח זה המתאר את המסלולים.
תיקוף מסלול	**לאחר התייעצות עם הגורמים הרלוונטיים והערכת תועלות וסיכונים בצורה מעמיקה יותר, ניתן לאשר את המסלול לפיו המערכת תיבחן:** ✦ מסלול ירוק/צהוב: באופן עצמאי ע"י אחראי היישום העסקי תוך התייעצות עם מוביל ה-AI וקבלת הנחיות מקצועיות ממנו. ✦ מסלול אדום: אישור ע"י מוביל ה-AI לאחר יידוע למנכ"ל ולפורום משילות AI(אם הוקם) במקרה של ספק אם לסווג את המערכת כמסלול אדום, מוביל ה-AI יכול להתייעץ עם פורום המשילות. אם מערכת סווגה למסלול אדום, מומלץ לשקול אם להעביר אותה למסלול כחול כדי לאפשר פריסה מבוקרת יותר והדרגתית בשלב ראשון.

תיאור השלב	פעולה
שלב 2 - גיבוש תוכנית ניהול סיכונים	מטרה: לבחון מה הם האמצעים שניתן להפעיל על מנת להפחית את הסיכונים.
התייעצות	התייעצות עם הגורמים המנויים בשלב 1: ✦ מסלול ירוק: לפי הצורך ✦ מסלול צהוב: מוביל AI וגורמים רלוונטיים אחרים כגון אבטחת מידע ו-DPO ✦ מסלול אדום: מוביל AI, גורמים רלוונטיים בארגון ומחוץ לו ופורום משילות AI (אם הוקם)
גיבוש תוכנית לניהול סיכונים	גיבוש תוכנית מפורטת לניהול סיכונים, תוך שילוב האמצעים להפחתת סיכונים המנויים בחלק 4 לנספח זה לפי הנסיבות. **בחינת עלויות התוכנית לניהול סיכונים תכלול:** ✦ בחינת חלופות שונות; ✦ הערכת עלות משוערת של אמצעי ניהול הסיכון, ביחס להפחתת הסיכון הנשקף. ✦ אם המערכת מסווגת למסלול כחול, התייחסות לגבולות הגזרה של הפרויקט (משך זמן, מדדים וקריטריונים להתקדמות). **דגש:** "עלויות" הן לא רק עניין כספי – מוצע לתת את הדעת להשלכות רוחב לרבות היבטי יעילות, אמון הציבור ועוד. ביחס להיבטים משפטיים שמתעוררים – יש לערב את הלשכה המשפטית בשלב זה, על מנת לוודא שניתנים מענים לדרישות משפטיות כמפורט במדריך המשפטי (למשל, חובת הנמקה, שקיפות, התמודדות עם הטיות ועוד).
אישור התוכנית	מסלולים ירוק וצהוב: אישור התוכנית באופן עצמאי על ידי האחראי היישום העסקי, תוך התייעצות והנחיות מקצועיות של מוביל AI. מסלול אדום: אישור התוכנית על ידי מוביל AI ומנכ"ל הארגון או מי מטעמו. ככל שיש היבטים משפטיים, כאמור, יש לערב גם את הלשכה המשפטית.
שלב 3 – הפעלת התוכנית	מטרה: עלייה לאוויר של המערכת והפעלת ניהול סיכונים.
יישום	ביצוע אמצעי ניהול סיכונים בהתאם לתוכנית שאושרה.
בקרות מומלצות	איסוף נתונים מהמערכת עצמה; יישום אמצעי הפחתת סיכון כפי שנקבעו בשלב 2; התייחסות למידע מקהילת המושפעים מהמערכת שצף בפניות ציבור, ערוצי תקשורת, שיח עם ארגוני חברה אזרחית רלוונטיים ועוד; בקרת הנתונים – בדיקות מדגמיות או אד הוק וכן בהתאם להתפתחויות, לצורך ולסיכונים.
תדירות הבקרות	לפי סיווג המערכת, אלא אם נקבע אחרת בתוכנית ניהול סיכונים: ✦ **ירוק:** פעם בשנה ✦ **צהוב:** פעמיים בשנה לפחות ✦ **אדום:** כל רבעון לפחות
שקיפות לציבור	לפי סיווג המערכת וקהל המושפעים, ובהתאם ולתוכנית ניהול סיכונים: ✦ 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). ✦ שקיפות והגברת מודעות למצבים שבהם מתקיימת אינטראקציה ישירה או משמעותית עם מערכת בינה מלאכותית על ידי האזרח. פירוט נוסף בנושא השקיפות ייכלל במדריך המשפטי.

תיאור השלב	פעולה
שלב 4 – טיפול בתקריות	מטרה: לאפשר טיפול מיטבי בתקלות ותקריות AI
דיווחים שוטפים – תדירות	לפי סיווג המערכת, אלא אם נקבע אחרת בתוכנית ניהול סיכונים: ✦ **ירוק:** פעם בשנה ✦ **צהוב:** פעם בשנה לפחות ✦ **אדום:** ארבע פעמים בשנה לפחות
דיווח במקרה של תקרית AI	דיווח מיידי למוביל AI כאשר מתרחשת תקרית AI. **דגש:** מומלץ לכלול בדיווח המלצות על ההתאמות הנדרשות.
שקיפות לציבור	שקיפות לציבור הרחב (או לקהילת המושפעים) כאשר מתרחשת תקרית AI – בהתאם להנחיות מוביל ה-AI.
התאמות והפקת לקחים	ביצוע ההתאמות בהתאם להנחיות מוביל ה-AI. ההתאמות יכולות לכלול, למשל, שינויים באפיון המוצר או באופן איסוף או תיוג הנתונים. תיעוד במסגרת תהליך ניהול שינויים.

מסלול כחול – הליך ניהול הסיכונים

מומלץ לקבוע את כל שלבי תהליך ניהול הסיכונים, בהתאם לאופי המסלול הכחול ובהתייעצות מתמדת עם מוביל ה-AI. מטבע הדברים יש גמישות בגיבוש תכנית הפעלת מערכת במסלול כחול. מומלץ כי התהליך יתייחס לרכיבים הבאים, בהתאם למידת הסיכון מעצם מימוש המסלול הכחול:

- התייעצות עם גורמים רלוונטיים בתוך הארגון ומחוץ לו, לרבות הלשכה המשפטית;
- בניית תוכנית ניהול סיכונים;
- קביעת הפרמטרים והמדדים שייבדקו במהלך הפרויקט;
- תדירות הבדיקות שיבוצעו;
- תדירות הדיווחים למוביל AI;
- קביעת לוחות זמנים לרבות פרק הזמן שבו ירוץ הפרויקט;
- כאשר נשקפים סיכונים גבוהים במיוחד, הגדרת תקריות AI חמורות ואופן ההתמודדות עמן.

ניתן לבצע שינויים בשלבים הנ"ל באופן דינאמי (למשל, להחליט להאריך את פרק הזמן שנקבע בהתחלה).

דגש: כאשר מדובר במערכת שמעלה סיכונים גבוהים, מצופה גם כי האחראי העסקי ו/או מוביל ה-AI יעדכנו את הנהלת הארגון בצמתים שונים, ויקבלו את אישורה במידה ויידרש.

2. סיווג תועלות וסיכונים

נקודת ההתחלה של התהליך היא בחינת התועלות השונות של מערכת בינה מלאכותית. שימוש בבינה מלאכותית עשוי להוביל לתוצאות בעלות ערך כגון: שיפור היעילות של השירות הציבורי (באמצעות אוטומציה של תהליכים שונים, אשר מובילים לקיצור זמני המתנה/זמני עיבוד תיקים (processing times); שיפור איכות ההחלטות, יכולת של מערכות בינה מלאכותית לפעול מסביב לשעון, מניעת טעויות אנוש); פרואקטיביות בעיצוב והאספקה של שירותים ציבוריים; והתאמת השירותים הציבוריים עבור תושבים ועסקים בהתאם לצרכים הייחודיים שלהם.[9] באופן כללי, כאשר נעשה שימוש אחראי במערכות בינה מלאכותית לטובת קידום מטרות הארגון, הציבור מרוויח משירותים ומוצרים ציבוריים טובים יותר. כנגזרת מכך, גם מוניטין הארגון ואמון הציבור בארגון עשויים להתחזק.

9 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.

לצד זאת, מערכות בינה מלאכותית מושתתות על תהליכים פנימיים מורכבים, מערכות טכנולוגיות מתקדמות, ואינטראקציות אנושיות אשר מייצרות גם סיכונים שונים. בין היתר, נתונים לא מדויקים, לא מעודכנים או מוטים עלולים לפגוע באיכות המערכת וליצור שגיאות או עיוותים בתוצאות, וכן לייצר קושי להתחקות אחר הלוגיקה של המערכת.[10] נוסף על כך, שימוש במערכות AI חושף את הארגון לסיכוני אבטחת מידע ייחודיים, העלולים לפגוע בפרטיות ושלמות הנתונים, וכן להוביל לדליפת מידע מוגן. נזקים אלו עלולים להיגרם גם כתוצאה מהתקפות על הכלים עצמם – גורמים עוינים עשויים לנסות לנצל חולשות במערכות ה-AI ולשבש תוצאות, או לנסות להוציא מידע ממערכות ניהול הנתונים. בסופו של יום, ההתממשות של סיכונים אלו עלולה לפגוע בציבור ובאמון הציבור בארגון.

כיצד ניתן להעריך את התועלות והסיכונים של מערכת בינה מלאכותית? להלן מיפוי לא ממצה של סוגי תועלות וסיכונים עיקריים שיש לשקול במסגרת תהליך ניהול הסיכונים. מיפוי זה מתכתב עם עקרונות ה-OECD לשימוש אחראי בבינה מלאכותית. ניתן לבחון תועלות וסיכונים נוספים הרלוונטיים לפעילות הארגון או לשימושים הצפויים בכלי.

תועלות

קטגוריה בהתאם לעיקרון ה-OECD	תועלות לדוגמא	הערות/דוגמאות
תועלות חברתיות, סביבתיות וכלכליות עיקרון ה-OECD: בינה מלאכותית לצמיחה, פיתוח בר קיימא ורווחה הכלל	✦ צמצום פערים חברתיים ✦ שמירה על הסביבה ✦ שיפור בתחום בריאות הציבור ✦ צמיחה כלכלית ✦ שיפור סביבת העבודה ✦ דיוק ושיפור תהליכי קבלת החלטות	**דוגמאות:** ✦ כלי לדיוק בחיזוי אקלים ✦ ייעול בניהול תעבורתי (תרומה לאיכות הסביבה) ✦ כלי לניתוח בקשות כספיות/מתן היתרים ✦ הפחתת עומס עבודה
קידום רווחת הפרט עיקרון ה-OECD: כיבוד שלטון החוק, זכויות אדם וערכים דמוקרטיים, לרבות הוגנות ופרטיות	✦ התאמת וייעול שירותים לאזרחים ✦ קיצור זמני המתנה ✦ הפחתת טעויות אנושיות	**דוגמאות:** ✦ כלים להנגשת זכויות לאזרחים ✦ כלי לסיוע בביצוע התממה (השחרות) לקבצים הכוללים מידע אישי לפני העברתם לצד ג' ✦ כלי לניתוח החלטות ממשלה או החלטות בית משפט **סוגיות משפטיות לרבות בחינת מערכות המשפיעות על זכויות אדם, מצריכות התייעצות נפרדת עם הלשכה המשפטית של הארגון. ראו גם המדריך המשפטי.**
תועלות לשקיפות כלפי הציבור עיקרון OECD: שקיפות והסברתיות	✦ סיכום וניתוח של החלטות ממשלתיות וטקסטים ציבוריים אחרים לצורך בקרה ✦ שיפור בהליכי שיתוף ציבור	**דוגמאות:** ✦ כלי לשיתוף ציבור לפיתוח מדיניות ורגולציה ✦ צ'אטבוט ממשלתי להנגשת מידע לאזרח ✦ תרגום של אתרי אינטרנט ממשלתיים לשפות שונות **לנושא השקיפות וההסברתיות יש היבטים משפטיים – יש להתייעץ עם הלשכה המשפטית**
ביטחון הציבור עיקרון הOECD: ביטחון ובטיחות	✦ חיזוק הגנת הסייבר של מערכות קריטיות ✦ הגברת בטיחות במרחב הציבורי (כבישים מהירים, תחבורה ציבורית, שיטור)	**דוגמאות:** ✦ הגנה פרו-אקטיבית על מערכות קריטיות ✦ ניטור סייבר של רשת ממשלתית ✦ חיזוי אסונות טבע ✦ תגובה מהירה ואפקטיבית מול מצבי חירום.
ממשל תקין עיקרון ה-OECD: אחריותיות	✦ אפשרות למעקב ביחס לתהליכים פנים ארגוניים ולשרשרת קבלת החלטות ✦ בקרות איכות של תהליכים פנים-ארגוניים	**דוגמאות:** ✦ כלי לסיוע בסיכום ישיבות פנימיות ✦ דאשבורד לניהול הרשאות ✦ כלי לסיוע בניתוח של הערות ציבור

10 בשל מאפיין של "קופסה שחורה" (אלמנט העכירות) הקיים במערכות בינה מלאכותית.

סיכונים

קטגוריה בהתאם לעיקרון ה-OECD	תועלות לדוגמא	הערות/דוגמאות
סיכונים חברתיים, סביבתיים וכלכליים עיקרון ה-OECD: בינה מלאכותית לצמיחה, פיתוח בר קיימא ורווחת הכלל	✦ הגדלת פערים חברתיים ✦ נזקים סביבתיים ✦ נזקים לבריאות הציבור ✦ נזקים כלכליים	**נזק סביבתי:** טעויות במערכת אשר תפקידה לסייע במתן היתרי פליטה לאוויר ובתנאים הסביבתיים ברישיונות העסק. **נזק לבריאות הציבור:** טעויות במערכת אשר מסייעת בתהליך מתן אישורים רגולטורים לתרופות **נזק כלכלי:** טעויות במודלי בינה מלאכותית לחיזוי ופיקוח על שוק ההון עלולות לערער את היציבות הפיננסית במשק ושל מוסדות וחברות שונות.
סיכונים הנוגעים לשמירת זכויות אדם וערכים דמוקרטיים עיקרון ה-OECD: כיבוד שלטון החוק, זכויות אדם וערכים דמוקרטיים, לרבות הוגנות ופרטיות	✦ חשש לאפליה והטיה ✦ חשש לפגיעה בפרטיות ✦ חשש לפגיעה בכבוד האדם או זכויות יסוד אחרות	מערכות בינה מלאכותית עשויות לעורר אתגרים גם במישור המשפטי. כך למשל, תופעות כמו הטיות ואפליה שמקורן במערכות אלו מעלות חששות ביחס לשמירה על עיקרון השוויון. בעת שילוב מערכות בינה מלאכותית בפעילות הרשות הציבורית, יש להבטיח עמידה בהוראות הדין, לרבות שמירה על זכויות הפרט, הגנה על הזכות לפרטיות, ועמידה בדרישות שונות במשפט המינהלי – כגון פעולה בסמכות, חובת ההנמקה, חובות שקיפות, והתבססות על תשתית עובדתית מהימנה. יש להיוועץ בלשכה המשפטית לצורך כך. המדריך המשפטי יתייחס לסוגיות אלה.
סיכונים לשקיפות כלפי הציבור עיקרון ה-OECD: שקיפות והסברתיות	✦ תוצאות ללא רציונל ברור או עקבי ✦ היעדר יכולת להבין איך המערכת הגיעה למסקנה שהופקה ✦ שימוש במערכת בשונה מהיעוד שלה	תופעת ה"קופסה השחורה" מקשה או מונעת את האפשרות להבין את הקלט של מערכת בינה מלאכותית. בנוסף, ההסתמכות העיוורת על הקלט של מערכת בינה מלאכותית עלולה להיות מסוכנת מבחינה עסקית. מעבר לכך, סוגיית השקיפות וההסברתיות כרוכות בהיבטים משפטיים, ולכן נדרשת בחינה של עמידה בהוראות הדין ביחס אליהן.
סיכונים לבטיחות הציבור עקרון ה-OECD: ביטחון ובטיחות	✦ סיכון תפעולי ✦ פגיעה במערכות קריטיות לרבות תקיפות סייבר ✦ פגיעה בחיי אדם או ברכוש ✦ מניפולציה חברתית וערעור הסדר הציבורי	**סיכון תפעולי:** סיכון שנובע מתקלה בתפקוד המערכות או בשימוש שגוי בה. סיכון זה עשוי לנבוע למשל מנתונים לא מדויקים, לא מעודכנים או מוטים, מסטיית המודל (model drift), וכדומה. **פגיעה במערכות קריטיות:** סיכון שנובע מתלות והסתמכות על AI למשל לניהול רשתות חשמל, מים או בקרת תנועה. **פגיעה בחיי אדם:** סיכון לפגיעה פיזית או נפשית באנשים, למשל בעקבות התרחשות תאונות דרכים או אבחונים רפואיים שגויים. **סיכון סייבר:** יש להתייעץ עם CISO הארגוני לקביעת היקף הסיכון ודרכי התמודדות עמו.
סיכונים לממשל תקין עיקרון ה-OECD: אחריותיות	✦ הסתמכות יתר על מערכות בינה מלאכותית, תוך רידוד האחריותיות הפנים-ארגונית	היעדר אחריותיות עלול לנבוע מהסתמכות יתר על מערכות בינה מלאכותית. אוריינות ומנגנוני בקרה מסייעים בהתמודדות עם סיכון זה.

יצוין כי בכל הנוגע לזיהוי וניתוח הסיכונים, מדריך זה מתייחס רק לסיכונים שנובעים באופן ישיר **ממערכת הבינה המלאכותית**, ולא לסיכונים שנובעים מסיבות אחרות כגון חולשות אבטחת מידע במערכות התקשוביות של הארגון, או ליקויים במדיניות שהכלי בינה מלאכותית מבקש לממש.

היבטים משפטיים: כאמור, 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.

על כן, 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 להציף את הדרישות המשפטיות והסוגיות המשפטיות שעמן יש להתמודד, ואלה יילקחו בחשבון במסגרת עיצוב תוכנית ניהול הסיכונים והשימוש במערכת.

על מנת לבחון את התועלות והסיכונים של מערכת בינה מלאכותית, מוצע לבצע את הפעלות הבאות:

- מיפוי 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");
- מיפוי והערכת הסיכונים השונים – סיכוי להתממשות והיקף הנזק הצפוי (מ-"נמוך מאוד" עד "גבוה מאוד");
- השוואת התועלות לסיכונים באופן הוליסטי. כאמור לעיל, כאשר מדובר בסוגיות משפטיות המוזכרות בטבלה, כגון תועלות וסיכונים הנוגעים לזכויות, הסוגיות אינן ניתנות לשקלול וניתוח עלות תועלת בלבד, ולכן יש לבחון את ההיבטים המשפטיים מול הלשכה המשפטית.

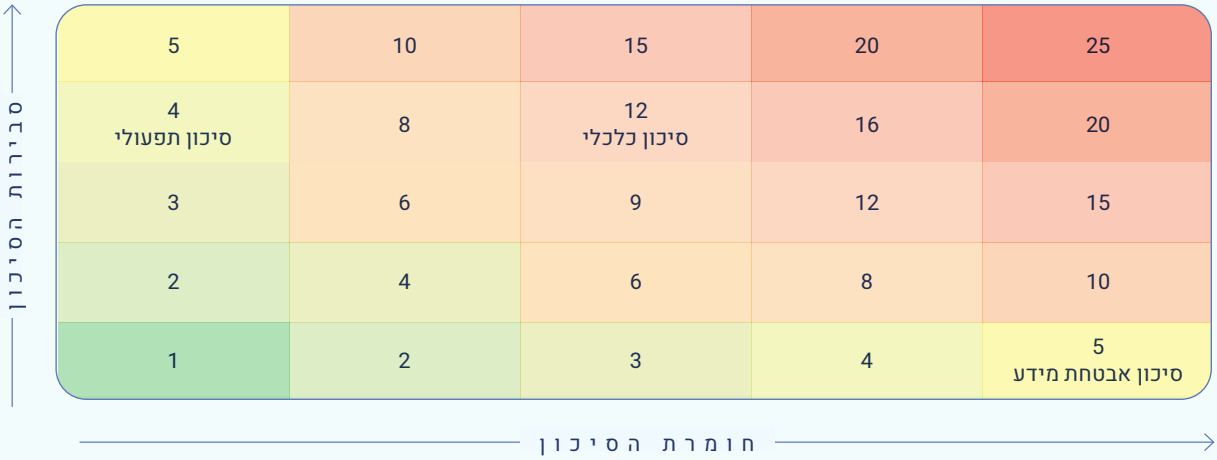
שילוב של מערכת בינה מלאכותית בתהליך עשוי להפחית סיכונים קיימים, להגביר אותם, או לייצר סיכונים אחרים. על כן, יש לבחון את הסיכון שנוסף ביחס למצב הבסיסי ובהתייחס למחולל הסיכון אל מול מכלול התועלות, באופן הולם. למשל, הפחתה של טעויות המערכת מ-15% על ידי גורם אנושי ל-1% על ידי בינה מלאכותית מצביע על סיכון של 1%, אבל על שיפור משמעותי מהמצב הקיים. חשוב לבחון כל חלופה בהשוואה לחלופות האחרות, ובהתאם את השינוי בסיכון, לרבות שילוב בכלים שאינם בינה מלאכותית (כמו מנועי חוקה).

בסיום בחינת הפרמטרים שהודגמו לעיל, מומלץ לתת ציון מספרי, **עבור כל סיכון בנפרד**, לחומרת הסיכון וכן לסבירות התממשותו (1 לנמוך ביותר ו-5 לגבוה ביותר). דירוג הסיכון המשוקלל לכל סיכון הוא תוצאה של הכפלת חומרת הסיכון (1-5) בסבירות התממשות הסיכון (1-5). חשוב לציין שההערכות הנ"ל הן איכותניות ולא כמותיות. שיטת הניקוד המוצעת נועדה לספק תמונת מצב כללית ולהביא את הגורם העסקי לחשוב על הסיכונים בצורה מסודרת.

בסופו של תהליך זיהוי והערכת הסיכונים, רצוי לייצר "מפת חום" שתאפשר לתעדף את התשומות המוקדשות לטיפול בכל סיכון. לשם המחשה, במצב שבו זוהו במערכת הנבחנת שלושה סיכונים: סיכון אבטחת מידע, סיכון כלכלי וסיכון תפעולי. חומרת הסיכון התפעולי נחשבת נמוכה מאוד (רמה 1) אך סבירות 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, מוביל ה-AI יכול להיעזר בטבלאות שיקבל מכלל הגורמים העסקיים בארגון, כדי ליצור מפת חום ארגונית להפיק תובנות מעשיות ממנה.

מפת חום להדגמה



3. פירוט המסלולים

ניהול הסיכונים הינו תהליך דינמי, רב-שלבי ודיפרנציאלי: ככל שהסיכון גבוה יותר, כך נדרשים תהליכי אישורים ואמצעי בקרה מקיפים יותר. מוצעים ארבעה מסלולים לאישור מערכות וניהול סיכונים: ירוק, צהוב ואדום, לפי שיטת צבעי הרמזור המוכרת, וכן מסלול כחול המאפשר ניסויים באופן מבוקר.

סיווג מערכת עשוי להשתנות בהתאם לממצאים חדשים, לשינויים בנסיבות השימוש, להתפתחות הטכנולוגיה, למידע שהתקבל במהלך תפעול המערכת, ולשינויים בהיקף ההשפעה של המערכת. שינוי בסיווג מצריך בחינה מחודשת של תוכנית ניהול הסיכונים, והתאמה של הבקרות והדיווחים הנדרשים, שעשויים להוביל לדרישות מחמירות או מקלות יותר.

להלן פירוט של מסלולי האישורים לפי רמת סיכון של מערכת.

בסוף נספח זה מופיעה טבלה המציינת את האישורים הנדרשים לכל מסלול.

1. מסלול ירוק

המסלול הירוק מיועד למערכות בעלות סיכון נמוך. הוא מאפשר שימוש בבינה מלאכותית שלא מעלה קשיים מיוחדים, בהליך אישורים פשוט ומהיר יותר. חשוב לציין כי גם כאשר מערכת מסווגת כ"ירוקה" עדיין יש לערוך תוכנית ניהול סיכונים (אמנם היא תהיה בסיסית יותר, בהתאם לרמת הסיכון), בקרה שוטפת, שקיפות במקרים המתאימים, טיפול בתקריות AI אם יתגלעו, ודיווח, אך בעצימות נמוכה ועם תשומות ארגוניות נמוכות יחסית.

1.1 מתי המסלול מתאים?

כאשר כל אחד מהסיכונים של מערכת יסווגו כסיכונים נמוכים (למשל ציון של 1-4 (מתוך 25) בכל אחת מקטגוריות הסיכונים בחלק 2 לנספח זה). זאת, אפילו אם התועלת הצפויה גם נחשבת נמוכה. ככלל, ניתן לשקול לסווג למסלול ירוק במקרים הבאים (לא מדובר בהכרח בתנאים מצטברים או מספקים, אלא באינדיקציות):

- המערכת אינה פוגעת בעקרונות של דמוקרטיה ושלטון החוק ועומדת בדרישות הדין;
- היא אינה משפיעה על זכויות, חירויות או זכאות לשירותים באופן ישיר;
- היא אינה מבצעת פעולות מורכבות מרובות שימושים, אלא פועלת בתחום פעולה מוגדר;
- היא אינה עושה שימוש במידע אישי או מידע המוגן לפי חסיונות אחרים בדין (כמו חובות סודיות) (ראו הגדרת מידע מוגן);
- אם יש אלמנט ג'נרטיבי מסויים, היא בעלת טמפרטורה נמוכה מאוד;
- היא ניתנת לפיקוח אנושי משמעותי, וכך תוכנן להשתמש בה;
- הפלט פשוט וניתן להסבר.

דוגמאות (לשם אינדיקציה בלבד – יש לבצע הליך סיווג בכל מקרה):

- ניהול מלאי ציוד משרדי
- לוח בקרה לפרויקטים פנימיים
- כלי יצירת מצגות
- תקצור פרוטוקולים של ישיבות
- מענה סטנדרטי לשאלות נפוצות (FAQ)
- כלים לסיוע בניתוח מסמכים פומביים (דוגמת NotebookLM)
- בדיקת איות להודעות לציבור
- כלים לניתוח תופעות ונתונים, זיהוי דפוסים וחריגים, בהקשרים שאינם רגישים (למשל, שאינם צפויים להשליך על זכויות הפרט)
- זיהוי תווים אופטי (OCR) לסריקת מסמכים בהקשרים שאינם רגישים
- תרגום הודעות פשוטות ולא רגישות
- הפקת טיוטת דוחות פנימיים, בהקשרים שאינם רגישים
- תרגום לעברית של מסמכים פומביים

1.1 אישורים ופיקוח

הליך מקוצר: אחראי היישום מבצע את פעולות התכנון באופן עצמאי. עליו להתייעץ עם מוביל ה-AI וכן מומלץ שיתייעץ עם כל גורם רלוונטי לפי הצורך, כדי לחסוך זמן ולמנוע הפתעות בשלב מאוחר יותר, אך אם הוא סבור שהסיכונים נמוכים מאוד, הוא יכול גם להתקדם באופן עצמאי. גם מערכת במסלול ירוק מצריכה גיבוש תוכנית ניהול סיכונים (בהתייעצות עם מוביל ה-AI), וכן בקרה ודיווח, אם כי מספיק לבצע את הבקרות פעם בשנה.

2. מסלול צהוב

המסלול הצהוב מתייחס למערכות בעלות סיכון שיוגדר כבינוני. ככלל, מדובר בסיכונים שעלולים להיות משמעותיים עבור הארגון, שותפיו ולקוחותיו. חשוב לציין כי אף על פי שסיכון נחשב "משמעותי", אין בכך כדי לפסול את השימוש במערכת על הסף. מגוון מערכות חיוביות ומועילות יכולות לכלול מרכיב של סיכון "בינוני" והן עשויות להיות ראויות מאוד לשימוש, ואף לספק תועלת משמעותיות עבור הארגון או הציבור. עם זאת , יש צורך בהשקעת תשומות ארגוניות להתמודד עם סיכון מסוג זה.

2.1 מתי המסלול מתאים?

כאשר המערכת לא סווגה למסלול ירוק, ואין בה סיכון שהציון שלו עולה על 19 – כלומר, יש לפחות ציון אחד שהוא בין 5 ו-19 (מתוך 25). יש בה סיכונים ברמה "בינונית". דוגמאות אפשריות: (1) מערכות שמנתחות פניות של אזרחים (מיילים, טפסים) ומסווגות מקרים דחופים או מבצעות את הניתוב הראשוני של הפנייה למחלקות שונות; (2) אלגוריתמים שמאתרים "אנומליות" שיכולות להעיד על טעות בדיווח או ניסיון להונאה.

2.2 אישורים ופיקוח

אחרי הסיווג הראשוני על ידי אחראי היישום העסקי, עליו להתייעץ עם מוביל ה-AI וגורמים רלוונטיים נוספים בתוך הארגון. אחראי יישום עסקי יגבש תוכנית ניהול סיכונים תוך התייעצות עם מוביל ה-AI. נדרשים דיווחים ובקרות תכופים.

3. מסלול אדום

המסלול האדום מתייחס למערכות בעלות סיכון גבוה. ככלל, הוא יתאים כאשר במידה והתועלת המצופה **גבוהה מאוד ולא נמצאו חלופות מסוכנות פחות** שמשיגות תועלת זהה או דומה, המסלול האדום מאפשר שימוש במערכת בתנאים נוקשים, תוך קבלת אישורים נוספים מהדרג הבכיר של הארגון.

3.1 מתי המסלול מתאים?

כאשר לפחות אחד מהסיכונים הנבדקים יסווג כגבוה (ציון של 20-25).

3.2 אישורים ופיקוח

לאור הסיכונים, הליך האישורים כרוך בבקרות מקיפות ואישורים בדרג בכיר במספר שלבים: כבר בשלב הסיווג הראשוני, נדרשת התייעצות מקיפה. תיקוף המסלול דורש את אישורו של המוביל AI ויידוע של מנכ"ל הארגון. תוכנית ניהול סיכונים גם כרוכה באישור מוביל ה-AI ומנכ"ל הארגון או מי מטעמו. אם הארגון הקים פורום משילות, רצוי לשלב גם אותו. במידה שהפרויקט יאושר בסופו של דבר, מוצע לשקול לנתב אותו למסלול כחול, כדי לאפשר פריסה מבוקרת והדרגתית, אלא אם כן קיים מצב חירום או הצדקות חריגות אחרות.

סיכונים חריגים וסוגי שימושים שיש לבחון לשלילה

במסגרת המסלול האדום של מערכות בסיכון גבוה, עשויים להיות מקרים שבהם מדובר בכלי בינה מלאכותית המציב סיכון יוצא דופן שעשוי להיות מזיק ומסוכן או לפגוע באופן מהותי ומשמעותי בזכויות יסוד.

במקרים אלה, יש לבחון האם יש כלל מקום לקידום הפרויקט. במקרה שבו מתעורר חשש כי מדובר במערכת המציבה סיכון יוצא דופן או עלולה לפגוע בזכויות באופן מהותי ומשמעותי – יש להתייעץ עם הלשכה המשפטית על מנת לבחון האם יש מניעה משפטית או קושי משפטי משמעותי שבגינו אין מקום לשילוב AI. ככל שניתן לקדם את הפרויקט מבחינה משפטית, יש להיוועץ עם קובעי מדיניות בדרג נבחר האמונים על התחום שבו מערכת ה-AI פועלת.

4. מסלול כחול

מסלול כחול הוא מסגרת תהליכית לניהול סיכוני בינה מלאכותית שנועד לשמש להתנסות מבוקרת, כאשר חסרים נתונים המאפשרים לסווג את המערכת בצורה ברורה או לפתח תוכנית ניהול סיכונים מדויקת. המסגרת מאפשרת בחינת כלי בינה מלאכותית לפני סיווג סופי, תוך הקפדה על הבקרות ושימת דגש על זמן שימוש מוגבל. מסלול כחול יכול להתבצע בדרכים שונות, למשל:

- אימון מודל על בסיס מידע מוגן, או הזנת קלט הכולל מידע מוגן, בסביבה סטרילית, לצורך פיתוח המוצר (שלב PoC);
- פריסה חלקית או מלאה בסביבת ייצור - שימוש מצומצם והדרגתי בפונקציות בודדות של המוצר.

4.1 מתי המסלול מתאים?

ניתן להכניס מערכת או פרויקט למסלול כחול כאשר קיימת אי-וודאות לגבי היקף הסיכונים, יש צורך לבחון את יכולות המערכת באופן סטרילי, או כאשר יש ללמוד כיצד היא פועלת "בשטח" ולהבין את הסיכונים שהיא מגלמת. זאת, כדי לאפשר לארגון להפחית את הסיכונים הצפויים מהמערכת בצורה אופטימלית. דוגמאות לבדיקות שניתן לערוך בתקופת הבחינה: בחינת ביצועים לפי פרומפטים שונים, בחינת דיוק הקלט לפי מצבים שונים, בדיקת סיכונים בתחומים שסומנו כרגישים יותר (לדוגמת הוגנות), ובחינת האינטראקציה של המערכת עם משתמשי הקצה וקהילת המושפעים.

4.2 אישורים ופיקוח – שלבים

מאחר שהמסלול הכחול נועד לבחינת מערכות בתנאי אי-וודאות ובמגוון מצבים, תהליך ניהול הסיכונים הוא גמיש יותר. זאת, כדי לאפשר לאחראי היישום העסקי, ביחד עם מוביל ה-AI, לבנות תכנית ניהול סיכונים מותאמת למקרה הספציפי. עם זאת, מומלץ להגדיר לפרויקט במסלול כחול פרק זמן מוגדר, קצר יחסית, ומנגנוני בקרה מוגברים. בסיום פרק הזמן שייקבע, על אחראי היישום לדווח למוביל ה-AI על התוצאות באופן מקיף. לאחר מכן, בהתייעצות עם מוביל ה-AI, אחראי היישום העסקי יכול להחליט אם להעביר את הפרויקט לפריסה מלאה/חלקית באחד המסלולים האחרים (ירוק/צהוב/אדום), בהתאם לממצאי הפרויקט, הסיכונים שזוהו, אמצעי ההפחתה האפשריים, וההמלצות שהתקבלו.

דגש: מסלול זה חוסך חלק מתהליכי האישור הפנימיים המוצעים במדריך (בהשוואה למסלול צהוב או אדום) אך עדיין מצריך ליווי של הלשכה המשפטית, כדי להבטיח עמידה בכלל הכללים המשפטיים, ובכלל זה לעניין היכולת של הארגון לאמן מודלים על בסיס מידע מוגן בהתאם לסמכות הנתונה לו.

ניתן בכל שלב להתייעץ עם מערך הדיגיטל הלאומי, דרך כתובת הדוא"ל ResponsibleAI@digital.gov.il.

תרשים מסלולים לפי סיווג הסיכון

5. שלבי אישור של פרויקטים לפי מסלול

להלן טבלה המסכמת את שלבי הליך ניהול הסיכונים ביחס למסלולים השונים. יודגש כי השלבים המפורטים הם בגדר המלצה, והארגון יכול לשנות אותם בהתאם לצרכים ולתהליכים שלו.

נוסף על הגורמים הרלוונטיים המעורבים בתהליך ניהול הסיכונים כמפורט בטבלה שלהלן, יודגש כי הלשכה המשפטית צריכה להיות מעורבת בכל מקום בו מתעוררת שאלה משפטית, בשלב מוקדם ככל הניתן.

תיאור השלב	מסלול ירוק	מסלול צהוב	מסלול אדום
שלב 1 - ניתוח וסיווג המערכת			
התייעצות	מוביל AI + גורמים נוספים לפי הצורך	מוביל AI + גורמים נוספים לפי הצורך	מוביל AI + גורמים בתוך הארגון ומחוץ לארגון
תיקוף מסלול	אחראי יישום עסקי תוך התייעצות עם מוביל ה-AI	אחראי יישום עסקי תוך התייעצות עם מוביל ה-AI	מוביל AI + יידוע למנכ"ל הארגון [+פורום משילות אם הוקם]
שלב 2 - גיבוש תוכנית ניהול סיכונים			
התייעצות	לפי הצורך	מוביל AI + גורמי רוחב כגון אבטחת מידע ו-DPO	מוביל AI + גורמי רוחב כגון אבטחת מידע ו-DPO + גורמים מחוץ לארגון [+ פורום משילות אם הוקם]
אישור התוכנית	אחראי עסקי לאחר התייעצות עם מוביל AI	אחראי עסקי לאחר התייעצות עם מוביל AI	מוביל AI + מנכ"ל הארגון או מי מטעמו
שלב 3 – הפעלת התוכנית			
תדירות הבקרות	פעם בשנה	פעמים בשנה לפחות	כל רבעון לפחות
שלב 4 - טיפול בתקריות			
דיווחים שוטפים – תדירות	פעם בשנה	פעם בשנה לפחות	כל רבעון לפחות
מסלול כחול: אבני הדרך המתייחסות לרכיבים הנ"ל תקבענה בשיתוף מוביל AI.			

4. גיבוש תוכנית ניהול סיכונים

ככלל, במסגרות תהליכיות לניהול סיכונים, קיימות ארבע שיטות כלליות של דרכי התמודדות עם סיכונים: הימנעות מסיכון, אי-התערבות, העברת סיכון והפחתת סיכון. כאשר נבחנת מערכת בינה מלאכותית בגוף ציבורי, דרכי התמודדות אלו עשויים להיות רלוונטיים בהקשרים שונים אך הם אינם מתאימים בהכרח בכל סיטואציה. להלן הדגמה קצרה של יישום דרכי ההתמודדות האלה בהקשרים שונים.[11]

סוג התמודדות	הדגמת יישום
הימנעות מסיכון	כאשר מדובר בסיכונים גבוהים, והתועלת המצופה לא מצדיקה אותם. עשוי להתאים במיוחד למערכת שסווגה כ"אדומה" בעלת סיכונים חריגים ביותר.
אי-התערבות	כאשר מדובר בסיכונים נמוכים מאוד שאינם מהותיים לארגון או לאנשים או ארגונים שבזיקה אליו. עשוי להתאים יותר למערכת שסווגה כ"ירוקה", אמנם מומלץ בכל אופן לבצע תהליך ניהול סיכון ולהשקיע מחשבה באפשרות להפחית סיכון, לפי הנסיבות.
העברת סיכון ושותפות בסיכון	העברת סיכון ושותפות בסיכון עם גורם חיצוני (למשל באמצעות מנגנון שיפוי) עשויות להתאים לסי־ כונים תפעוליים או כלכליים נמוכים או בינוניים (ירוק או צהוב), אך חשוב לציין כי היא לא מתאימה לכל הסיכונים. למשל, ניתן להעביר סוגים שונים של סיכונים תפעוליים באמצעות קביעת תנאים עם ספק שירות חיצוני שבו הוא מתחייב לפעול בצורה מסוימת; ניתן גם לקבוע תנאי שיפוי לארגון במקרה של התממשות הסיכון. עם זאת, האחריות המהותית של רשות ציבורית למעשיה כלפי הפרט והציבור ולפעולה בהתאם לדין נותרת אצלה, ולא ניתנת להעברה אלא במישור הכלכלי-תפעולי כאמור.
הפחתת הסיכון	כאשר מדובר בסיכונים מהותיים אשר נטילתם מוצדקת בנסיבות העניין, וניתן לבצע פעולות הפחתה ומעקב שוטף ובקרה על ביצוען ולהפעיל את המערכת בצורה בטוחה ואחראית. ראו פירוט להלן לגבי סוגי הפעולות האפשריות.

האמצעים המתוארים להלן הם כלליים ויכולים לתת מענה למגוון סוגי סיכונים הנשקפים משימוש במערכות AI. בבחירת האמצעים המתאימים, חשוב לקחת בחשבון שאמצעים שונים יכולים להתאים למטרות שונות – למשל, מעורבות אנושית יכולה להתאים לבדיקת תוצאות שליליות של בקשות מענק אבל עשויה שלא להתאים כאשר מדובר במערכת ניהול תורים שכל מטרתה לפעול במהירות ובאופן אוטונומי.

דגש נוסף הוא ניתוח האופנים בנוגע לדרך האופטימלית להפחית את הסיכון: ישנם מקרים בהם פעילות פשוטה יחסית תפחית סיכונים באופן יעיל יותר מפעילות מורכבת. כך למשל, במערכת AI שמתריעה על שגיאות במסמכים שמוגשים לרשות על בסיס אינדיקטורים שונים, ניתן להפחית את סיכון הטעויות בצורה "סטנדרטית" של בדיקה אנושית בכל מקרה, אך בדיקה זו עלולה להגדיל את העלויות באופן משמעותי. דרך אחרת, היא באמצעות אימון נוסף של המודל או הוספת מרכיב RAG על בסיס נתונים עסקיים עדכניים. כמו כן, כאשר המערכת היא סוכן שמתחבר למאגר מידע, השימוש בפרוטוקולים שונים להזנת מידע למערכת, MCP או A2A, לפי העניין, מאפשר להכווין את המערכת באופן מוגדר, וכך להגביר את דיוק הפלט של הסוכן או להפעיל בקרות (למשל, העברת פלט מסוים לבקרה אנושית). כמו כן, טכנולוגיות אלו כוללות במקרים רבים אפשרות של תיעוד הפעולות שבוצעו, דבר אשר יכול לסייע בהיבטי הסברתיות ובהמשך שקיפות. עם זאת, חיבור מודל בינה מלאכותית למאגרי מידע - והן שלב האימון והן כחלק מתהליך הסקה - יכול לעורר שאלות מורכבות בהיבטי הגנת פרטיות ואבטחת המידע, המצריכות התייעצות ותהליכי אישור עם הגורמים הרלוונטיים בארגון.

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

11 ראו את והמדריך לניהול סיכונים ברגולציה של רשות האסדרה, עמ' 18.

א. **אמצעים טכניים וארכיטקטוניים –** כדי לצמצם את הסיכון, חשוב לשלב כבר ברמת התכנון הטכני והארכיטקטורה של המערכת אמצעים טכניים וארכיטקטוניים להפחתת סיכונים. הטכניקות הבאות עשויות להפחית סיכונים שונים, באופנים שונים. חשוב להכיר את יתרונותיהן וחסרונותיהן, ולשקול כיצד ניתן לשלבן באופן אופטימלי, בהתייעצות עם גורמי טד"מ ומוביל ה-AI:

סוג סיכון	כלים טכניים - ארכיטקטוניים
פגיעה בפרטיות או חשיפת מידע מוגן אחר	טכנולוגיות לשיפור פרטיות (PET)[12] למידה מבוזרת (federated learning) אנונימיזציה ונתונים סינתטיים (התממה) הצפנה הומומורפית (homomorphic encryption) פרטיות דיפרנציאלית (Differential Privacy)
חוסר דיוק, הזיות	Tool Use כוונון טמפרטורה MCP (אם מדובר בסוכן AI) CAG/RAG כלי ניטור לסכנות של concept drift / model drift (Monitoring Tools) MLOps (תפעול למידת מכונה)
הטיות ואי-שוויון	כלי תיקון הטיות (reweighting, adversarial debiasing) בקרה על ייצוגיות נתונים (fairness testing) מסגרות להערכת הטיות (bias evaluation frameworks)
אבטחת מידע והתקפות עוינות	מעקות בטיחות (Guardrails) Model Armor מודל בסביבה מקומית (on-prem) תשתית ופלטפורמה מאובטחת כגון Vertex AI לבידוד נתונים והגנת אבטחה ברמת הרשת רכישת מערכת אשר מספקת הגנה לארגון או למשתמש להגנה על הפרטיות ואשר עומדות בתקני אבטחת מידע ופרטיות
חוסר שקיפות ויכולת הסבר	טכנולוגיות xAI (בינה מלאכותית מוסברת) שימוש במודלים המשלבים טכניקות של reasoning (מודל הסקה / חשיבה) Model cards/datasheets
חוסר שליטה בסוכן	MCP A2A (תקשורת סוכן-לסוכן)

דגשים:

1. היבטי FinOps ו-DevOps מצריכים התייחסות ייחודית, המשלבת מדידת ביצועים, בחינת סיכונים ומדידת ארכיטקטורה כזאת או אחרת, הן בשלב הפיתוח והן בשלב הייצור של המערכת. המדריך אינו מתייחס באופן ישיר להיבטים אלו, אך הם נושקים גם להיבטי שימוש אחראי בבינה מלאכותית, שכן בחירה בתצורה, במאפיין או באופן פעילות מסוימת של מערכת בינה מלאכותית, או שינוי בבחירה, עשויה להשליך על הסיכונים הרחבים שסומנו ונוהלו במסגרת גיבוש ויישום תוכנית ניהול סיכונים. חשוב שאחראי היישום העסקי יקיים שיח מתמיד עם גורמי הטכנולוגיה כדי להתאים ולעדכן את התוכנית בהתאם לשינויים המוצעים במערכת.
2. **שילוב סוכן AI (Agent) מוגדר למשימה ייעודית –** סוכני AI אינם כלים להפחתת סיכונים, אך שילוב של סוכן עבור משימה ייעודית ומוגדרת היטב (כגון חיפוש אינטרנטי במאגר מידע אמין, בדיקת הפלט לפי פרמטרים אובייקטיביים, או עצירה של תהליך אשר אינו עולה בדרישות בטיחות מסוימות), עשוי 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. **התנסות בארגז חול טכנולוגי –** אפליקציות ב-GovAI Studio (כאשר יונגש רשמית) או ב-Vertex AI Platform מאפשרות להריץ מודלים ויישומים ולתקן אותם באופן ראשוני.

12 להרחבה ראו מדריך ליישום טכנולוגיות מגבירות-פרטיות במערכות בינה מלאכותית (PETs AI), של הרשות להגנת הפרטיות.

ב. **מרכיבים במוצר –** כחלק 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. לנושא השקיפות והגילוי גם היבטים משפטיים, וצפויה להיכלל התייחסות אליו במדריך המשפטי.
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. **אמצעים עסקיים-ארגוניים –** 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, או בצמתים רלוונטיים בתהליך עסקי, 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. בהתאם לנסיבות, ייתכן שלא תידרש מעורבת אנושית מלאה, אלה רק במקרים מסוימים – למשל, במערכת שבודקת התקיימותם של תנאים הקבועים מראש. סוגיה זו מערבת גם היבטים משפטיים שייסקרו במדריך המשפטי.
2. **התייעצות עם מומחים –** שילוב מומחים בתחומים רלוונטיים צפוי לאפשר זיהוי נכון של הסיכונים ובניית המנגנונים הרלוונטיים להפחתתם. לצד מומחי בינה מלאכותית והנדסת תוכנה, מוצע להתייעץ עם גורמים מגוונים, לפי הצורך, כגון בתחומי האתיקה, המשפט, הפרטיות (DPO), שוויון והכלה, כלכלה והגנת סייבר, כבר בשלבי האפיון והפיתוח.
3. **בקרות מדגמיות –** ביצוע בדיקות איכות מדגמיות של פלט המערכת (outputs), אחת לתקופה או על מקרים רגישים, צפוי לסייע בזיהוי סיכונים לאורך חיי המערכת. חשוב לקחת בחשבון את תופעת "סטיית המודל" (model drift), שבה ביצועי המודל מתדרדרים עם הזמן, או שהוא כולל יותר הטיות. תופעה זה מחזקת את הצורך בבקרות מדגמיות שבודקות שוב את ה-KPIs שנקבעו. בקרות מעידות על פלט בעייתי מצריכות התייחסות מיוחדת – למשל, שימוש ב-CAG כדי לחדד את התוצאות, או אימון מחדש של המודל.
4. **הסתייעות בחוקרי אבטחת מידע עם מומחית ב-AI** שתפקידם לבדוק את הבטיחות, האמינות והחוסן של ותסייע בזיהוי של פגיעויות, כשלים בעמידות, חולשות, הטיות וסיכונים אפשריים. יש להתייעץ עם ה-CISO הארגוני.
5. **חיזוק משילות נתונים (Data Governance) –** אמינות מערכת בינה מלאכותית תלויה באופן ישיר והדוק בנתונים שעל בסיסם פועלים המודלים. אם הסיכונים הנשקפים מהמערכת נובעים מפגמים בדאטה המקורי או בדרך שהוא מנוהל, ניתן לנקוט בצעדים הבאים:
   - הבטחת שימוש בדאטה הנכון למודל הנכון – נתונים עדכניים, שלמים, מייצגים ואמינים.
   - ניהול גישה, רישוי 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.
7. **העברה למסלול כחול –** כאמור, המסלול הכחול מאפשר נסיינות של המערכת לזמן מוגבל תוך קביעת מגבלות ובקרות קפדניות, למשל:
   - קביעה מראש של השערות, ובדיקות ומדדים. מומלץ כי המדידות יתייחסו לכלל העקרונות של שימוש אחראי;
   - הגבלת הפרויקט ל-3–12 חודשים;
   - הגבלת סוגי נתונים, אוכלוסיית משתמשים ותחומי שימוש לפי רמת הסיכון;
   - יצירת אפשרות לעצור או לגלגל אחורה את המערכת ללא פגיעה בשירותים;
   - גילוי למשתמשים כי מדובר במערכת ניסיונית.

לכל כלי, ישנן עלויות שונות וציפיות שונות ביחס ליכולת שלהם להפחית סיכונים שונים. בבחירת האמצעים המתאימים להפחתת הנזק, מומלץ לחשוב על ה-trade-offs השונים, בצורה הוליסטית.

דוגמה להמחשה:

משרד ממשלתי המטפל בבקשות אזרחים באמצעות מערכת ממוחשבת שמזהה אם המבקש הגיש את כל המסמכים הנדרשים או שחסר מסמך כלשהו. כיום, לפני הטמעת המערכת, 5% מהבקשות שמזוהות כחסרות מסמכים מתבררות בדיעבד כתקינות (false negative) ו-4% מהבקשות שמזוהות כשלמות מתבררות כחסרות מסמך (false positive). מערכת בינה מלאכותית מוצעת צפויה להוריד את כמות התוצאות החיוביות שגויות מ-5% ל-3% אך גם להגדיל את התוצאות השליליות השגויות מ-4% ל-6%. במקרה כזה, ניתן להחליט מראש על אמצעים שונים: להוריד את מידת האגרסיביות של המערכת כדי לצמצם תוצאות שליליות שגויות, אפילו אם המשמעות היא הגדלת התוצאות החיוביות השגויות. לחלופין, ניתן לקבוע כי, לאור החיסכון בזמן ובתוצאות חיוביות השגויות שצפויה מהמערכת החדשה, כל תוצאה שלילית תיבדק על ידי גורם אנושי, וכך למעשה להוריד את אחוזי התוצאות השליליות השגויות לכמעט אפס.

נספח ד'

קווים מנחים למשתמש קצה

נספח זה מיועד למשתמשי 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.

ככלל, כאשר מדובר במערכת בינה מלאכותית שהונגשה לעובדים על ידי הארגון, המשתמש יקבל הוראות שימוש שמתייחסות לנושאים הבאים (וכן הנחיות נוספות) מהמוביל בינה מלאכותית הארגוני ומאחראי היישום העסקי הרלוונטי. האמור להלן מספק 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.

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:

- **"הזיות" -** מגבלות באמינות ודיוק: 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.
- **תוכן פוגעני ואפליה:** 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 עלול להגביר את הסיכון להתרחשות תופעות אלה בקנה מידה רחב, בהשוואה להחלטה אנושית פרטנית.
- **סכנה למידע מוגן:**
  - **שימוש במידע המוגן בזכויות יוצרים וסודיות מסחרית:** 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.
  - **שימוש במידע אישי והפקת מידע מצטבר:** במקרים רבים, ובמיוחד במסגרת שימוש חינמי, מערכות בינה מלאכותית אוגרות מידע רב שהוזן במסגרת השימוש (פרומפטים, קבצים) לרבות מידע אישי הנוגע לנושאי מידע,, והיסקים שנוצרו כתוצאה מהצלבת המידע שהוזן.

- **שימוש במידע ממשלתי מוגן בהיבטים אחרים:** ישנו מידע המצוי בידי הממשלה אשר חלות עליו מגבלות והגנות משפטיות שונות ונוספות, כגון מידע שעלול לפגוע ביחסי החוץ של המדינה. בנוסף, ישנם סוגים נוספים של מידע אשר גילוי בלתי מבוקר שלהם עשוי לפגוע באינטרסים אחרים של המדינה או צדדים שלישיים (כמו למשל, מידע אשר הופק במסגרת תהליך של גיבוש מדיניות, תרשומות פנימיות של שיח בין גורמי מקצוע, וכיוב').[13] העלאת מידע כאמור למערכות וכלי מדף שלא אושרו על ידי הארגון באופן בלתי מבוקר עלולה לייצר סיכונים משפטיים ואחרים עבור המדינה, הציבור ושחקנים שונים כתוצאה מחשיפת המידע לגורמים בלתי מורשים.

ישנם יישומים מסוימים שמאפשרים מחיקה של המידע גם בשימוש חינמי, אך אין בכך להבטיח שלא נעשה שימוש במידע או כי הוא אינו נאגר במערכות של ספקיות המערכות. בנוסף, כלים רבים הנגישים לשימוש חינמי "מתאמנים" על המידע שהוזן להם.

- **קושי להסביר את התוצאות:** המודלים המתקדמים של מערכות בינה מלאכותית מהווים מעין "קופסה שחורה", במובן זה שלא ניתן להסביר עד הסוף כיצד היא הגיעה לתוצאה מסוימת.
- **האנשה:** בחלק מהמודלים, בפרט במודלי שפה, המערכת מתקשרת עם המשתמש בשפה טבעית ואנושית. בשל מאפיין זה, משתמש עלול לייחס למערכת אמפטיה, כוונה מוסרית, סמכות או מומחיות שאין לה במציאות, 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.
- **סיכוני סייבר:** מערכות AI עלולות להיות חשופות למתקפות הזרקת בקשה (prompt injection), מגורמים עוינים אשר עלולות להשפיע על הפלט שמתקבל. נוסף על כך, המידע שמעלים עלול להיות חשוף לגורמים עוינים.

קווים מנחים כלליים למשתמש הקצה

1. לפני השימוש בכלי בינה מלאכותית, יש לוודא שהכלי לא נאסר לשימוש (או לסוג השימוש המבוקש) על ידי גורם הממונה על הגנת הסייבר בארגון (CISO), מטעמי אמינות או אבטחת מידע.
2. יש לבדוק אם קיימות הנחיות רלוונטיות מטעם הארגון, לדוגמת מוביל בינה מלאכותית ארגוני או לשכה המשפטית, ולפעול לפיהן.
3. יש לזהות האם מדובר בכלי מבוססי AI **חיצוני** לארגון (לדוגמת חשבון Gemini או Claude פרטי), או בכלי שאושר על ידי הארגון לפעולה על הדאטה הארגוני (למשל בענן הארגוני או במערכות המידע של הארגון).
4. כאשר הכלי **אושר על ידי הארגון**, השימוש בו ייעשה על פי הנחיות השימוש שישוקפו למשתמשי הקצה.
5. כאשר הכלי **פועל בסביבה חיצונית לארגון**, דוגמת שימוש בצ'ט בוט ברשת האינטרנט ברישיון פרטי:
    5.1.1. **אין** להזין למערכות חיצוניות מידע מוגן, ובכלל זאת - מידע אישי, מידע שאין למוסרו לפי חוק חופש המידע,[14] מידע שעלול להיות מוגן על ידי סודיות מסחרית או קניין רוחני, חיסיון משפטי או מידע מסווג.
    5.1.2. **אין** להזין מנחים (prompts) אשר מעידים על כוונה לביצוע פעולה כלפי פרט מזוהה מסוים, או כוונה לבצע צעד שלטוני רגיש מסוים.
    5.1.3. מומלץ לעיין במדיניות המערכת והוראות השימוש על מנת להבין את מגבלות הכלי ואת השימוש שנעשה במידע המוזן אליו. במקרה של ספק, מומלץ לעובד להתייעץ עם הממונה שלו.
    5.1.4. **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.
6. 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.

13 ראו לעניין זה כדוגמה, את הסייגים והחריגים לחוק חופש המידע תשנ"ח-1998, ובעיקר אלו המנויים בסעיפים 9(א) ו-9(ב) לחוק.
14 למשל, ראו מידע שאין למסרו לפי סעיף 9 לחוק חופש המידע, תשנ"ח-1998.

7. 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 למוביל ה-AIהארגוני.
8. **מומלץ להקפיד על כתיבת פרומפטים מדויקים –** מומלץ לכתוב פרומטים בצורה מדויקת ומלאה ויכללו בצורה מפורשת את המשימה המבוקשת מהכלי. זאת, כדי להגביר את הסיכוי שהתוצאות יהיו מדויקות ורלוונטיות.
9. **אין להניח שהתוכן נכון מבלי לבדוק אותו –** חשוב להצליב את התוצאות שמתקבלות ממערכת עם מקורות מידע מקובלים כגון פרסומים בספרות מדעית או באתרים מוכרים, ולבדוק את מהימנות התוצאה. כאשר מדובר בנושא שאינו בתחום המומחיות של העובד, חשוב להתייעץ עם מומחים רלוונטיים – למשל, מי שעוסק בשמאות על נכסים ושואל שאלה בנושא מיסים, יפנה לגורמים שזה תחום עיסוקם כדי לבדוק את המהימנות של הפלטים שהתקבלו.
10. **שקיפות ותיעוד –** אם נעשה שימוש משמעותי במערכות בינה מלאכותית יוצרת בגיבוש תוצרים, מומלץ בהתאם לנסיבות לציין זאת בתוצר הסופי. במקרה של הסתייעות במערכת לצורך קבלת החלטה (למשל, במקרה של מערכות ייעודיות שהוטמעו בפעילות הארגון), יש לתעד את השימוש במערכת. מכיוון ששקיפות כרוכה בהיבטים משפטיים, יש למלא אחר ההנחיות הארגוניות בנושא שקיפות, שיגובשו בין היתר בהתאם למדריך המשפטי.
11. **יש לשים לב אם נעשה שימוש במידע המוגן בקניין רוחני –** כאשר מדובר ביישום של בינה מלאכותית יוצרת, המידע המתקבל עלול להכיל מידע המוגן בזכויות יוצרים. יש 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. בכל אופן, מומלץ לפנות לייעוץ המשפטי של המשרד.

נספח ה'

מדיניות ארגונית לדוגמה לשימוש אחראי בבינה מלאכותית

להלן מדיניות ארגונית לדוגמה בנושא שימוש אחראי בבינה מלאכותית. הנוסח מהווה דוגמה בלבד, והוא נועד לסייע לארגונים שרוצים לגבש מדיניות ארגונית פנימית, על סמך העקרונות במדריך.

המדיניות הארגונית המוצעת להלן, מתבססת על המדריך לשימוש אחראי בבינה מלאכותית, לרבות מתודת ניהול הסיכונים – אך היא גם מרחיבה עליו, במטרה להפוך את ההמלצות למעשיות ופרטניות יותר.

מוביל ה-AI הארגוני שמגבש מדיניות עבור הארגון שלו מוזמן להתבסס על הנוסח להלן, לבצע את התאמות שהוא רואה לנכון לפי מאפייני הארגון וצרכיו. טקסט בסוגריים מרובעים מסמן קטעים שהמוביל AI מתבקש להשלים.

תוכן עניינים:

1. מבוא

מדיניות זו מפרטת את העקרונות, התהליכים והאחריות של הארגון לשימוש מושכל, אחראי ואתי במערכות בינה מלאכותית (AI). היא מתבססת על המדריך לשימוש אחראי בכלי בינה מלאכותית במגזר הציבורי של מערך הדיגיטל הלאומי (להלן - **"המדריך"**). מטרות המדיניות הן:

- תמיכה בשימוש במערכות בינה מלאכותית למען שיפור השירותים שנותן הארגון וקידום יעדי הארגון;
- קידום תרבות ארגונית של שימוש אחראי במערכות אלו, לצרכים פנימיים וחיצוניים;
- קידום אוריינות AI בקרב עובדי הארגון;
- שילוב מערכות AI באופן שמטיב עם העובדים ומשפר את סביבת עבודתם.

המדיניות חלה על כלל מערכות הבינה המלאכותית הנמצאות בשימוש, בפיתוח, או בבחינה בארגון.

המדיניות חלה על כל שלבי מחזור החיים של מערכות בינה מלאכותית, החל מהגדרת הצורך, דרך אפיון, פיתוח/רכש, הטמעה, תפעול, ועד להוצאה משימוש של המערכת, אך יודגש כי היא אינה גורעת מנהלים והנחיות אחרים בשלבים הרלוונטיים.

מדיניות זו היא מרכיב משלים להוראות תכ"ם רלוונטיות ולמדריך לשימוש אחראי בבינה מלאכותית של מערך הדיגיטל הלאומי.[15] [מוצע להפנות למסמכי מדיניות נוספים של הארגון].

ככלל, בבחינת הטמעת כלי בינה מלאכותית, תינתן עדיפות לכלים אשר זמינים במסגרת הסכם נימבוס (רובד 1 או רובד 5).

15 עתידים להתפרסם גם מדריכים והנחיות נוספות (מדיניות יה"ב, המדריך המשפטי של ייעוץ וחקיקה בהיבטים המשפטיים של שימוש בבינה מלאכותית על ידי רשויות ציבוריות. עם פרסום המדריכים, יהיה צורך לעדכן את המדיניות בהתאם.

2. עקרונות יסוד

הארגון יפעל כדי ליישם את עקרונות השימוש האחראי בבינה מלאכותית, שאומצו במדריך לשימוש אחראי בבינה מלאכותית במגזר הציבורי.

יישום העקרונות יתבטא בפעולות יזומות בכל שלבי מחזור החיים של מערכות בינה מלאכותית של הארגון, וישולב כחלק אינטגרלי מתהליכי קבלת ההחלטות בארגון. הטבלה שלהלן מפרטת את העקרונות שאומצו, לצד צעדים מרכזיים לאופן יישומם על ידי הארגון.

עיקרון	דוגמאות של אופן יישום
צמיחה, פיתוח בר קיימא ורווחה הכלל	לפני פיתוח או הטמעת כל מערכת AI, תבוצע הערכה מקיפה של התועלות הפוטנציאליות, תוך בחינת תרומתה לרווחה הכלל, שיפור איכות השירותים הציבוריים, צמצום הדרה של אוכלוסיות פגיעות, והשלכות סביבתיות.
כיבוד שלטון החוק, זכויות אדם וערכים דמוקרטיים, לרבות הוגנות ופרטיות	כל מערכת AI תעבור בחינה משפטית ואתית קפדנית בשלבי האפיון והפיתוח, בהובלת הלשכה המשפטית וה-DPO.
שקיפות והסברתיות	הארגון יפרסם מידע רלוונטי באתר האינטרנט שלו ובאתר תקרית AI על מערכות בינה מלאכותית העיקריות שבשימוש. במקרים של אינטראקציה ישירה עם אזרחים, יינתן גילוי נאות וברור כי מדובר במערכת בינה מלאכותית, תוך ציון היכולת לפנות לגורם אנושי במקרים המתאימים.
ביטחון ובטיחות של המערכת	מערכות בינה מלאכותית יפותחו ויוטמעו בהתאם לסטנדרטים מחמירים של אבטחת מידע והגנת סייבר. תינתן העדפה לכלים הזמינים בענן הממשלתי (נימבוס). יוטמעו מנגנוני בדיקה, ניטור ותגובה רציפים לתקלות. מערכות בינה מלאכותית בסיכון בינוני וגבוה יעברו בדיקות מקיפות ויהיו כפופות לאמצעי הפחתת סיכונים מתאימים.
אחריותיות	תוגדר חלוקת תפקידים וסמכויות ברורה ומתועדת לכל שלבי מחזור החיים של מערכת ה-AI. יוקמו מנגנוני דיווח ובקרה שוטפים, ותבוצע אכיפה של עקרונות המדיניות. כל תקרית AI תתוחקר באופן יסודי. יופקו לקחים וינקטו פעולות מתקנות.

3. משילות בינה מלאכותית - תפקידי מפתח ותחומי אחריות

להלן בעלי התפקידים הרלוונטיים בארגון בהקשרי שימוש אחראי בבינה מלאכותית:

שם תפקיד על פי המדריך	גורם רלוונטי בארגון [למלא]
מוביל AI ארגוני ("מוביל AI")	[לציין אם זה CDO, מנמ"ר או גורם אחר]
אחראי יישום עסקי	[לציין מי יהיו הגורמים שמוסמכים לקדם פרויקטי AI בארגון – למשל: כל ראשי האגפים, ראשי אשכולות, סמנכ"ל כספים, סמנכ"ל משאבי אנוש, וכו']
משתמשי קצה	העובדים בצוות של אחראי היישום העסקי (לרבות יועצים חיצוניים, בהתאם לנסיבות)
פורום משילות ארגוני	פורום המורכב מבעלי התפקידים הבאים: [לציין]
לשכה משפטית	[לציין את הרפרנטים המשפטיים בארגון שיילוו תהליכי AI]

[ניתן להוסיף תפקידים לפי הצורך]

שיתוף הפעולה בין גורמים אלו הוא קריטי ליישום מוצלח של מדיניות זו ולמשילות איכותית של מערכות בינה מלאכותית בארגון.

4. תהליך אישור פרויקט בינה מלאכותית

היקף תהליך האישור תלוי בסיווג המערכת. סיווג זה יהיה על פי המדריך (מסלולים ירוק, צהוב, אדם וכחול). [בחלק זה, מוצע להעתיק חלקים רלוונטיים מנספח ג' של המדריך (ניהול סיכונים), עם ההתאמות הנדרשות עבור הארגון].

דגשים פרטניים לארגון ביחס לתהליך ניהול הסיכונים המפורט במדריך:

א. התייעצויות:
במידה ומדובר במערכת בעלת רגישות מיוחדת [לתאר בנפרד את הקריטריונים הארגוניים], ניתן להתייעץ עם גורמים חיצוניים לארגון, בכפוף לאישור מוביל ה-AI ולתנאי שמירה על סודיות.

במערכות אשר משפיעות באופן ישיר ומשמעותי על תושבים, רצוי לשתף את הציבור בשלב מוקדם, ולאסוף הערות מבעלי עניין רלוונטיים, בתיאום עם מוביל ה-AI והלשכה המשפטית.

ב. גיבוש תוכנית ניהול סיכונים:

- אחראי היישום העסקי יגבש תוכנית ניהול סיכונים בהתבסס על נספח ג' של המדריך;
- מוביל ה-AI ילווה את אחראי היישום ויספק מידע וכלים לאורך הדרך;
- מוביל ה-AI יפנה למערך הדיגיטל הלאומי במקרה שיש צורך בידע או כלים נוספים כדי לסייע בגיבוש התוכנית.
- בכל עת, ניתן להתייעץ עם גורמים נוספים בתוך הארגון כדי להבין את הסיכונים ולפתח גישות להתמודד איתם.

ג. בקרות: בהתאם לתוכנית ניהול סיכונים.

ד. דיווחים: יש לכלול בתוכנית מועדים לדיווחים שוטפים וכן דיווחים למוביל ה-AI במקרה של תקרית AI.

ה. יש לתעד את הבדיקות והמסקנות ביחס לכל שלב בפרויקט.

5. שקיפות ומודעות ציבורית ביחס להפעלת מערכת בינה מלאכותית

ההנחיות שלהלן מתייחסות לאופן בו הארגון יתנהל בשקיפות מול הציבור ומול לקוחות נוספים בממשלה ומחוצה לה, כאשר ייעשה שימוש במערכת בינה מלאכותית. הן עשויות להשלים או להוסיף על הנחיות המשפטיות בנושא, שיגובשו בתיאום עם הלשכה המשפטית.

הארגון יפרסם באתר שלו רשומה של מערכות בינה מלאכותית שאושרו ושעשויות להשליך באופן ישיר על תושבים. הפרסום יפרט את:

- ייעוד המערכת והשימוש בה
- עקרונות פעולה מרכזיים
- נתונים שהארגון הזין כדי לאמת את המודל או לתת לו קונטקסט
- סוגי נתונים שימשו למערכת כאשר היא תופעל
- הסיכונים העיקריים שנובעים מהמערכת
- אמצעי הפחתה שננקטו

נוסף על כך, כל מערכות בינה מלאכותית בשימוש ארגוני (לרבות אלו שנמצאות במסלול ירוק) יפורסמו באתר AI Watch. **אם מדובר במערכת או פלטפורמה מרכזית שמורכבת בשניים או יותר סוכנים, ניתן לפרסם מידע על המערכת או הפלטפורמה בכללותה, תוך פירוט תפקידם של הסוכנים המרכזיים.**

הארגון יקים **ערוץ משוב ציבור באמצעות** טופס מקוון, תיבת מייל ייעודית לקבלת משוב, שאלות ופניות מהציבור בנוגע לשימושי AI של הארגון. מוביל ה-AI יהיה אמון על ניהול ערוצים אלו ותגובה לפניות.

כאשר מדובר במערכת אשר מקבלת החלטות לגבי זכויות של תושבים או עסקים, או אשר מסייעת בקבלת החלטות כאמור באופן משמעותי, הארגון יפרט בצורה ברורה ופשוטה כיצד הגורם שנפגע מהחלטה כזאת יכול לברר פרטים ביחס להחלטה או לערער עליה.

6. בקרות

[ניתן להתבסס ולהרחיב עת הטבלה שמופיעה בנספח ג', חלק 1 של המדריך]

7. טיפול בתקרית AI

(חלק זה מתקשר להמלצה בנספח א' למדריך להקים צוות ייעודי להתמודדות עם תקריות AI חמורות)

יוקם **צוות תגובה לתקריות AIRT - AI Incident Response Team (AI)**, אשר מורכב מ-[אפשר להפנות לפורום משילות הארגוני, מומלץ לשלב גם את אחראי היישום העסקי, מוביל ה-AI, גורם טכנולוגי, נציג הלשכה משפטית, וגורם בכיר בהנהלת הארגון]. תפקידו לחקור ולהגיב לתקריות AI.

הטיפול בתקרית AI יכלול את השלבים הבאים:

א. **דיווח וזיהוי ראשוני:** משתמש קצה או מערכת שמזהה חשד לתקרית AI – בין אם במסגרת בקרה שוטפת ובין עם בעקבות דיווח חיצוני, ידווח זאת מיידית למוביל ה-AI. הדיווח יפרט את המועד, אופי החשד, וההשפעה מוערכת.

ב. **חקירת צוות ה-AIRT:** מיד עם קבלת הדיווח, מוביל ה-AI יכנס את הצוות. הצוות יאסוף נתונים טכניים (לוגים, גרסאות מודל, נתוני קלט/פלט) וינתח את היקף הנזק ומקורות התקלה: תקלה טכנית, שימוש לרעה, פגיעות אבטחה, הטיה (bias), הפרת פרטיות או שימוש לא מורשה. בפרט, הצוות יבחן אם התקרית מעידה על בעיה רוחבית עם המערכת, או שמדובר בתקרית נקודתית. הצוות יתעד את ממצאיו.

ג. **תגובה:** בהתאם לממצאים ולרמת הנזק, הצוות יחליט על הפעולות הנדרשות, אשר יכולות לכלול:

- ניתוק או השעיה זמנית של המערכת או של רכיבים ספציפיים;
- הגדרת אמצעים זמניים (workarounds) להמשך רציפות תפקודית;
- טיפול בנזק באופן נקודתי או רוחבי (שינוי התוצאה, פיצויים);
- היקף ואופן דיווח לציבור, אם רלוונטי;
- החזרת המערכת לפעולה תקינה לרבות באמצעות תיקונים למערכת, אופן איסוף או תיוג הנתונים או אימון מחדש של המודל;
- בדיקות תקינות ותפקוד לפני חזרה לשימוש מלא.

ד. **הפקת לקחים ושיפור מתמשך:** במידת הצורך, הצוות ימליץ על שינויים רצויים בתהליכי פיתוח, בדיקות ובקרות, ועל הדרכות והעלאת מודעות בקרב עובדים וצוותים רלוונטיים, על מנת למנוע תקריות AI נוספות.

8. תוכנית הכשרות בינה מלאכותית

הכשרות בתחום הבינה המלאכותית נועדו:

- להגביר את ההבנה והמודעות לבינה מלאכותית, יתרונותיה וסיכוניה הפוטנציאליים בקרב כלל עובדי הארגון;
- להטמיע תרבות ארגונית של שימוש אחראי בבינה מלאכותית;
- לאפשר לעובדים להבין כיצד כלי בינה מלאכותית יכולים לתרום לעבודתם;
- לעודד יזמות וחדשנות בארגון ביחס לאפשרויות להטמיע כלים חדשים;
- לספק לצוותים שיישמו כלי בינה מלאכותית בארגון כלים מעשיים לגבש ולתפעל תוכנית ניהול סיכוני בינה מלאכותית.

תוכנית ההכשרות תכלול קורסים בנושאים הבאים, המפורטים בטבלה. היא תעודכן לפחות פעם בשנה ובהתאם להתפתחויות טכנולוגיות ורגולטוריות. התכנים המדויקים והיבטי לוגיסטיקה ייקבעו בתיאום עם מחלקת משאבי האנוש של הארגון.

נושא	תוכן	קהל יעד
מבוא לבינה מלאכותית ושימוש אחראי	מהי בינה מלאכותית וכיצד היא עובדת? סוגי AI (גנרטיבי, מנבא וכו'). יתרונות ויישומים פוטנציאליים של AI במגזר הציבורי. סיכונים עיקריים (לדוגמה: "הזיות", הטיות אלגוריתמיות, פגיעה בפרטיות, אבטחת מידע). עקרונות השימוש האחראי ב-AI, קווים מנחים למשתמש קצה.	כלל עובדי הארגון
מדיניות השימוש האחראי ב-AI של הארגון	סקירה מעמיקה של מדיניות הבינה מלאכותית הארגונית, עקרונותיה, חלוקת תפקידים, תהליכי סיווג מסלולים (ירוק, צהוב, אדום, כחול), ניתוח תועלות וסיכונים, טכנולוגיות להפחתת סיכונים, זיהוי, בקרה ודיווח על תקריות AI סקירת ההיבטים המשפטיים – בשיתוף עם הלשכה המשפטית/יעוץ וחקיקה שבמשרד המשפטים	אחראי יישום עסקי בארגון
בינה מלאכותית ככלי תומך עבודה – כלים ושימושים מעשיים	התנסות מעשית עם כלי AI מאושרים לשימוש בארגון, כלי פיתוח, התמודדות עם מגבלות הכלים, פירוט בנושא טכנולוגיות להפחתת סיכונים.	אחראי יישום עסקי בארגון צוותים טכנולוגיים.

9. הנחיות לעובדים

בכל מערכת AI – יש לפעול לפי תנאי השימוש של הספק שיועברו לעובדים על ידי אחראי היישום העסקי. נוסף על כך:

מערכת שהוטמעה בארגון בהתאם לתהליך רכש פנימי: יש לפעול בהתאם להוראות שימוש שניתנו על ידי [לקבוע מי אחראי להעביר הוראות כאמור – מוביל ה-AI, אחראי יישום]

מערכת שמופיעה ברובד 5 של נימבוס: יש לפעול בהתאם **לקווים מנחים למשתמש קצה** של מערך הדיגיטל הלאומי.

מערכת שמונגשת באופן רוחבי (באמצעות שדרת המידע הממשלתית) על ידי מערך הדיגיטל הלאומי: יש לפעול לפי דף "מדיניות מוצר" המופיע אתר האינטרנט של אותה מערכת.

מערכת לשימוש פומבי (כגון חשבון פרטי של Gemini או Claude): יש לפעול יש לפעול בהתאם למדריך למשתמש קצה של מערך הדיגיטל הלאומי.

אין לייצר באופן עצמאי ממשק אוטומטי בין כלי AI חיצוני לארגון לבין סביבת העבודה הארגונית ללא אישור של [CISO הארגון]

כל האמור בכפוף להוראות אבטחת מידע של [CISO הארגון].

נספח ו'

מילון מונחים

להלן מילון מונחים שכיחים בעולם הבינה המלאכותית ומובאים במדריך.[16]

מונח	הגדרה
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: Nano Banana, Sora (ויזואלי), Resemble AI (קולי), Claude או ChatGPT (טקסטואלי).
בינה מלאכותית מוסברת (xAI - Explainable AI)	תחום העוסק בפיתוח שיטות המאפשרות לבני אדם להבין, לפרש ולבטוח בתוצאות של מערכות בינה מלאכותית. המטרה היא להפוך את ה"קופסה השחורה" של המודל לשקופה יותר על ידי מתן הסברים לסיבות שהובילו אותו להחלטה מסוימת.
בקרה על ייצוגיות נתונים (Fairness Testing)	תהליך בקרת איכות שמטרתו להעריך האם המודל מציג התנהגות בלתי-מפלה ועקבית כלפי קבוצות שונות באוכלוסייה. התהליך כולל שימוש בטכניקות בדיקה (כגון בדיקות קומבינטוריות) כדי לזהות מקרים בהם המודל מייצר תוצאות שונות עבור קלטים דומים שנבדלים ביניהם רק בתכונה העלולה להיות מפלה.
הנחייה ("פרומפט" – prompt)	פרומפט הוא הנחייה למערכת בינה מלאכותית יוצרת. הפרומפט נכתב בטקסט ומבוצע באמצעותו מעין דיאלוג עם המערכת.
הצפנה הומומורפית	הצפנה הומומורפית משתמשת במבנה אלגברי מיוחד המאפשר שימור של חלק מהתכונות של המידע המקורי במידע המוצפן. כך, לאחר ביצוע פעולות על המידע המוצפן (ללא פתיחתו) ניתן בתנאים מסוימים לחלץ תוצאה המתאימה לחישוב על מידע המקור. לפיכך, הצפנה הומומורפית מאפשרת שימוש במידע בהיותו מוצפן ללא צורך בפתיחת ההצפנה, ובהתאם מצמצמת את הסיכון בחשיפה של המידע לגורם בלתי מורשה במהלך העיבוד. (מקור: מדריך לטכנולוגיות מגבירות-פרטיות PETs של הרשות להגנת הפרטיות)
התממה (אנונימיזציה)	התממה היא הסרת מאפיינים או שינוי ערכים על מנת לצמצם או למנוע זיהוי של נושא המידע. דרכים נפוצות להתממה הן הסרת שדות הכוללים מזהים ישירים (שדות המאפשרים זיהוי של אדם באופן ישיר, כגון שמות, מספרי תעודת זהות או אחד הפרטים המזהים הנכללים בהגדרת "מידע אישי" בתיקון מס' 13 לחוק הגנת הפרטיות (הסרת שדות עם מזהים עקיפים) שדות שבשילוב עם מידע נוסף מאפשרים זיהוי של אדם, כגון מקצועות ומקומות עבודה (הכללה) כגון הורדת רמת דיוק של נתון או קבוצת נתונים(, או הוספת רעש אקראי למידע. צמצום רמת פירוט המידע הרלוונטי והיקפו מגבירים את ההגנה על הפרטיות, אך עשויים לפגוע בערך ובשימושיות של המידע ליישומים שונים. (מקור: מדריך לטכנולוגיות מגבירות-פרטיות PETs של הרשות להגנת הפרטיות)
טכנולוגיות מגבירות פרטיות (PET)	טכנולוגיות שמטרתן לאפשר עיבוד, ניתוח או שיתוף של מידע תוך שמירה על פרטיות האנשים שמהם נאסף המידע. שימוש בטכנולוגיות אלו, ועיצוב לפרטיות (Privacy by Design), יאפשרו הפחתה וצמצום פגיעה בפרטיות.
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. 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.

16 מבוסס על הגדרת ה-OECD. ראו גם מילון מונחי התקשוב.

מונח	הגדרה
טמפרטורה (temperature)	טמפרטורה היא פרמטר במודלים גנרטיביים שמגדיר את רמת האקראיות והיצירתיות של הפלט. טמפרטורה נמוכה (0.1–0.3) תיתן תשובות עקביות, בטוחות וזהירות. 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.
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.
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.
כלי ניטור	כלים ותהליכים לזיהוי שינויים בביצועי המודל לאורך זמן, במיוחד שינויי סחף.
כלי תיקון הטיות (Reweighting ,Adversarial Debiasing)	טכניקות להפחתת הטיות מפלות באלגוריתמים. שקלול מחדש (Reweighting) היא שיטה בה מעניקים משקל שונה לנתונים לפני תהליך אימון המודל, כדי לאזן ייצוג חסר של קבוצות מסוימות. הסרת הטיה יריבותית (Adversarial Debiasing) היא שיטה בה מאמנים את המודל לבצע תחזיות מדויקות ולמנוע מרכיב "יריב" לגלות מהן התכונות המוגנות (כגון מגזר או מגדר) מתוך התחזיות.
למידה מבוזרת (federated learning)	למידה מבוזרת מאפשרת לכמה גורמים במשותף לאמן מודלים של בינה מלאכותית על מידע שלהם (מודל מקומי), ולאחר מכן לשלב את הדפוסים שזוהו במודלים המקומיים אל תוך מודל גלובלי יחיד ומדויק, ללא צורך לחלוק את המידע אשר שימש כל אחד מהגורמים לאימון המודל. למידה מבוזרת יכולה להיעשות בגישות הבאות: 1. למידה מבוזרת מרכזית (learning federated Centralized): שרת מרכזי מייצר אלגוריתם או מודל, ושולח את המודל למקורות מידע מבוזרים. המודל מתעדכן בהתאם למקורות המידע המקומיים ונשלח חזרה לשרת המרכזי שיוצר מודל משוקלל. 2. למידה מבוזרת מקומית (learning federated Decentralized): במודל זה לא מעורב שרת מרכזי. הצדדים מתקשרים זה עם זה באופן ישיר ומעדכנים את המודל בכל פעם על סמך המידע המקומי שלהם. (מקור: מדריך לטכנולוגיות מגבירות-פרטיות PETs של הרשות להגנת הפרטיות).
מודל בסביבה מקומית (On-prem)	פריסה והרצה של מודלי AI על גבי התשתיות הפיזיות של הארגון עצמו ולא באמצעות שירותי ענן חיצוניים. גישה זו מאפשרת שליטה מלאה על אבטחת הנתונים והתאמה אישית, ונמצאת בדרך כלל בארגונים בעלי רגישות גבוהה למידע.
מודל הסקה / חשיבה (Reasoning Model)	מודלים שתוכננו לבצע תהליכי חשיבה לוגיים, כמו "שרשרת מחשבה" (Chain-of-Thought) ורפלקציה עצמית, כדי לפתור בעיות מורכבות בצורה מדויקת יותר מהמחקה דפוסי חשיבה אנושיים: פירוק בעיות מורכבות לשלבים, התנסות באסטרטגיות שונות ותיקון עצמי תוך כדי תהליך העיבוד.
מידע מוגן	במסמך זה, מידע מוגן הוא מונח סל המתייחס למידע אשר חלות עליו הגנות שונות – ובכלל זאת, מידע אישי כהגדרתו בחוק הגנת הפרטיות, מידע אשר חלות עליו הוראות סודיות (לפי דין, פסיקה או הסכם); מידע המוגן מטעמי ביטחון המדינה; מידע שהינו קניין רוחני ו/או סוד מסחרי; מידע שהוטל לגביו חיסיון (למשל, לטובת המדינה או הציבור), וכיוב'.
מידע סינטטי	מידע סינתטי הוא נתונים המיוצרים בדרך כלל לפי דפוסים ומאפיינים סטטיסטיים של נתונים אמיתיים (המכילים מידע אישי). מידע סינתטי נועד לאפשר להפיק תוצאות דומות לעיבוד מידע אישי, ללא שימוש במידע אישי אמיתי. מעבר להגנה על הפרטיות, מידע סינתטי עשוי לשמש למגוון מטרות נוספות, ביניהן ככלי ליצירת מערכי מידע גדולים לאימון מודלים של בינה מלאכותית או הפקת מידע הכולל מגוון רב של מקרי קצה ואירועים נדירים. השימוש במידע סינתטי לעיתים משולב יחד עם מידע שאינו סינתטי (אמיתי) ולעיתים מחליף את המידע האמיתי לחלוטין. (מקור: מדריך לטכנולוגיות מגבירות-פרטיות PETs של הרשות להגנת הפרטיות)
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, Google AI Search, Perplexity.
מסגרות להערכת הטיות (Bias Evaluation Frameworks)	מתודולוגיות מובנות לזיהוי, מדידה, והפחתה של הטיות פוטנציאליות במערכות AI לאורך מחזור החיים שלהן. המסגרות מגדירות קריטריונים לאמינות, ממפות השפעות חברתיות פוטנציאליות ומציעות דרכי פעולה מומלצות.

מונח	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.[17]
"מעקות בטיחות" (Guardrails)	הגדרת כללים טכניים ומנגנוני בקרה שפועלים כדי שהמערכת תפעל בהתאם למדיניות שנקבעה. כך למשל, הם יכולים למנוע מהמערכת לעסוק בנושאים אסורים או רגישים,או לענות תשובות בעייתיות. ניתן להשתמש בהם גם כדי להקפיץ התרעות, לקבוע את טון השיחה עם המשתמש, ולהגדיר דפוסי תשובות מקובלת.
סוכן AI (AI Agent)	מערכת AI הפועלת אוטונומית או חצי-אוטונומית כדי לבצע משימות או להשיג מטרות מוגדרות תוך אינטראקציה עם הסביבה שלה. ככלל, 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.
סחף (concept drift/model drift)	סחף מתרחש כאשר הקשר בין נתוני הקלט לבין התוצאה הרצויה משתנה בין המעבדה לעולם האמיתי. תופעה זו מעידה על ירידה בדיוק המודל, ומחייבת לאמן אותו מחדש או לבצע בו התאמות.
פרטיות דיפרנציאלית (Differential Privacy)	פרטיות דיפרנציאלית היא גישה שפותחה עבור מאגרי נתונים הכוללים מידע אישי אך מיועדים למטרות עיבוד סטטיסטי (שאינו מיועד ל חשוף מידע אישי). מהות הגישה של פרטיות דיפרנציאלית היא שהוספה, הסרה או שינוי של רשומה אחת במאגר המידע תשפיע במידה מועטה ביותר, אם בכלל, על ת וצאות היישום של פונקציה סטטיסטית על מאגר המידע. במקום להסיר או לשנות נתונים כדי לטשטש מזהים, לצורך השגת פרטיות דיפרנציאלית יש להוסיף למידע האמיתי נתונים אקראיים, או רעש. המטרה היא להוסיף כמות מספקת של נתונים אקראיים כך שמידע אמיתי לא יהיה ניתן לזיהוי מתוך הרעש. פרטיות דיפרנציאלית עדיין אפשרת לבצע ניתוח מדויק על נתונים מצטברים, מכיוון שלמרות הרעש הנוסף – הנתונים המשולבים יכולים לספק תוצאות מדויקות. (מקור: מדריך לטכנולוגיות מגבירות-פרטיות PETs של הרשות להגנת הפרטיות)
קופסה שחורה	אופן קבלת ההחלטות של מערכות בינה מלאכותית מתואר לעתים קרובות כ"קופסה שחורה" (black box) במובן זה שלא ניתן להתחקות אחריו. כלומר, לא ניתן לדעת איך ומדוע התקבלה ההחלטה כפי שהתקבלה. הסיבה לכך נעוצה במורכבות האלגוריתמים העומדים בבסיס מערכות הבינה המלאכותית ובכמות המידע שעל בסיסו מתקבלות ההחלטות. (מקור: הדוח הסופי של הצוות הבין משרדי לבחינת אסדרת בינה מלאכותית בסקטור הפיננסי.)
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.
תפעול למידת מכונה (MLOps)	מתודולוגיה המשלבת פיתוח מודלים של למידת מכונה (ML) עם תפעול מערכות (Ops) כדי ליצור תהליך רציף ואוטומטי של פיתוח, פריסה, ניטור וניהול שלהם בסביבת ייצור (Production), בדומה ל-DevOps- בעולמות התוכנה.
תקרית AI	מאורע, נסיבה או סדרת מאורעות, שבהם פיתוח, שימוש או תקלה של מערכת בינה מלאכותית אחת או יותר, מובילים, במישרין או בעקיפין, לנזקים דוגמת פגיעה בבריאות, שיבוש הניהול והתפעול של תשתיות קריטיות, או הפרת חובות משפטיות. תקריות AI יכולות לנבוע, למשל, משינויים במודל בינה מלאכותית, מתקפות במודל עצמו או בדאטה שהוא מתבסס עליו, ומתוצאות בלתי צפויות של המערכת.
תקשורת סוכן-לסוכן (A2A - Agent to Agent)	מגמה טכנולוגית המאפשרת למערכות בינה מלאכותית אוטונומיות ("סוכנים") לתקשר זו עם זו, להעביר מידע ולבצע פעולות מתואמות (Orchestration). השליטה האנושית בתהליך נשמרת באמצעות שימוש בפרוטוקולי אבטחה מקיפים וממשקים מוגדרים, המכתיבים מראש כיצד סוכנים יכולים "לגלות" זה את זה ואיזה מידע מותר להם להחליף.
Tool use	היכולת של מודל AI לגשת למשאבים חיצוניים ולתקשר עם מערכות אחרות (למשל, דרך API) כדי לבצע פעולות או לשלוף מידע, במקום להסתמך רק על הידע הפנימי או היכולות של עצמו. שימוש בכלים נוספים עשוי לשפר את הדיוק של מערכת.
Model Armor	אחד משירותי אבטחת הענן של Google הייעודי למודלי שפה ומשמש כ"שכבת הגנה" בזמן אמת בין המשתמש למודל ה AI. הכלי עוזר לזהות ואף לחסום ניסיונות לעקיפה של מגבלות מודל, למנוע זליגת מידע פרטי או רגיש ולסנן תוכן בלתי הולם.
MCP	תקן פתוח המאפשר למפתחים לבנות אינטגרציה דו-כיוונית מאובטחת בין עוזרי AI (כמו ChatGPT) לבין מאגרי מידע ומערכות עסקיות (כמו Google Drive). התקן מחליף את הצורך בבניית אינטגרציה נפרדת לכל מודל, ויוצר שפה אחידה לשיתוף הקשרים.

17 עקרונות מדיניות, רגולציה ואתיקה בתחום הבינה המלאכותית, 2023

מערך הדיגיטל הלאומי
نظام الديجيتال الوطني
Israel National Digital Agency

תשפ"ג-2023

משרד המשפטים
وزارة العدل | MINISTRY OF JUSTICE

ייעוץ וחקיקה
المشورة والتشريع
OFFICE OF LEGAL COUNSEL AND LEGISLATIVE AFFAIRS