10

ביקורת על מערכות בינה מלאכותית
מדריך למבקרי המגזר הציבורי

מוסדות הביקורת העליונים של ברזיל, פינלנד, גרמניה, הולנד, נורווגיה ובריטניה

זהו תרגום של מסמך ההנחיות Auditing AI Systems שכתבו מוסדות הביקורת האמורים לעיל. הגרסה הראשונה של המסמך פורסמה בשנת 2020, והגרסה הנוכחית - ב-19 בדצמבר 2025.

תקציר

מסמך זה קובע הנחיות עבור ביקורת מערכות בינה מלאכותית (AI) המבוצעת על ידי מוסדות ביקורת עליונים (SAI). מטרתו לסייע למוסדות ביקורת עליונים ולמבקרים עצמם לבצע ביקורת של מערכות בינה מלאכותית המשמשות גופים במגזר הציבורי. הוא נועד למבקרים בעלי ידע מסוים בשיטות כמותיות, אך אינו מניח כי קיימת מומחיות בתחום של מערכות בינה מלאכותית או מודלים של למידת מכונה.

המסמך כולל קטלוג ביקורת, המורכב מקווים מנחים בנושאי ביקורת מוצעים מבוססי סיכונים, נוסף על בקרות מצופות. למסמך זה נלווה כלי עזר בפורמט Excel המסכם ומנחה את המבקרים בביצוע חלקים שונים של הביקורת.

1. תקציר מנהלים

בשנת 2017 חתמו מוסדות הביקורת העליונים של ברזיל, פינלנד, גרמניה, הולנד, נורווגיה ובריטניה על מזכר הבנות לשיתוף פעולה בנושא ניתוח נתונים. לאור ההכרה בטרנספורמציה הדיגיטלית המשנה את אופן פעולתם של גופי ציבור, הסכימו מוסדות הביקורת לשתף ידע ביניהם ולגבש גישות ביקורת חדשות. בשנת 2019 הם התחייבו להפיק במשותף מסמך הנחיות בנושא של בינה מלאכותית במגזר הציבורי. הגרסה הראשונה של מסמך זה פורסמה ב-2020 כסקירה טכנית, ועודכנה לראשונה ב-2023[1]. המסמך עבר עדכון משמעותי לגרסת 2025 הנוכחית עקב שינויים משמעותיים בשימוש בבינה מלאכותית ובסיכונים המזוהים עימה, לצד התפשטות השימוש בבינה מלאכותית יוצרת, חקיקתן של תקנות אסדרה (רגולציה) חדשות וזמינותן הנרחבת של מערכות בינה מלאכותית מסחריות.

בינה מלאכותית מציעה כיום בשימוש נרחב במגזר הציבורי, במטרה לשפר את השירותים ולהפחית עלויות. מערכות בינה מלאכותית יכולות לספק חיזויים (בינה מלאכותית לחיזוי), המלצות או החלטות, או לייצר תוכן כמו טקסט, תמונות או שמע (בינה מלאכותית יוצרת). מרבית מערכות הבינה המלאכותית מסתמכות על למידת מכונה (machine learning), שבמסגרתה מודלים לומדים דפוסים מתוך נתונים כדי להשיג מטרות ספציפיות.

טכנולוגיות חדשות מלוות על פי רוב בסיכונים חדשים. בזמן שהחקיקה הייעודית בנושא עודנה מתגבשת, קיים צורך במנגנוני בקרה ובביצוע ביקורות בתחום זה. קהילת הבינה המלאכותית מתמקדת יותר ויותר בכללי אתיקה ובהשפעתה החברתית של הבינה המלאכותית. מאז פרסום מסמך זה לראשונה, הונהג רישוי למבקרים המתמחים בבינה מלאכותית[2].

אף שרשויות הגנת המידע פיתחו קווים מנחים ייעודיים ומסוגלות ליטול על עצמן תפקיד פיקוחי להגנת המידע האישי, רבים מן הסיכונים המזוהים עם יישומי בינה מלאכותית חורגים אל מעבר לסוגיות של מידע אישי. לדוגמה, מערכות בינה מלאכותית אטומות (opaque) עלולות להפוך פרקטיקות בלתי הוגנות לאוטומטיות ולחזק אותן, ובעקבות זאת לערער את האמון במוסדות ציבור. מוסדות ביקורת עליונים צריכים להיות ערוכים לבצע ביקורת של יישומי בינה מלאכותית באמצעות ביקורות ציות וביקורות ביצועים כאחד. כמה ממוסדות הביקורת העליונים החתומים על מזכר ההבנות ערכו חקרי מקרה (case studies) או ביקורות פיילוט[3], במטרה לגבש מתודולוגיה גנרית לביקורת על יישומי בינה מלאכותית או לבחון את השימוש בבינה מלאכותית בשירות הציבורי.

המסמך הנוכחי מתווה את הסיכונים העיקריים הטמונים בשימוש בבינה מלאכותית בשירות הציבורי, ומציע שיטות לביצוע ביקורת על מערכות בינה מלאכותית. ההנחיות בו מבוססות על ניסיונם של מוסדות הביקורת העליונים שחיברו את המסמך (בברזיל, פינלנד, גרמניה, הולנד, נורווגיה ובריטניה), שכלל ביצוע ביקורות של למידת מכונה ופרויקטים בתחום התוכנה. זו איננה רשימת קריטריונים מחייבת, ויש להשתמש בה כמדריך לפיתוח או התאמה של מתודולוגיות הביקורת.

המסמך מתחלק לשני חלקים עיקריים. פרק 3 מציג קריטריונים אפשריים לביקורת, ובכללם תקנות לאומיות, תקנים בין-לאומיים וקווים מנחים הנמצאים בשימוש נרחב.

---

1 עדכון ראשון זה מתבסס על ניסיונם של המוסדות החתומים על מזכר ההבנות - ניסיון שנצבר דרך יישום המתודולוגיה שפותחה בביקורות שערכו אותם מוסדות.

2 ראו לדוגמה, הסמכת AAIA מטעם ISACA.

3 ראו לדוגמה את המסמכים הבאים: הבנת אלגוריתמים וכן ביקורת של 9 אלגוריתמים המשמשים את ממשלת הולנד, שפרסם בית המשפט לביקורת של הולנד; השימוש בבינה מלאכותית בממשלה המרכזית מטעם משרד הביקורת הלאומי של נורווגיה (גרסה מלאה בנורווגית בלבד); שיטות ניתוח נתונים ובינה מלאכותית בממשל הפדרלי הגרמני מאת בית הדין הפדרלי לביקורת חשבונות של גרמניה (בגרמנית בלבד); או שימוש בבינה מלאכותית בממשלה, שהפיק משרד הביקורת הלאומי הבריטי.

פרק 4 מציג קטלוג ביקורת[4] הבנוי סביב תחומי מפתח: ניהול וממשל של הפרויקט, נתונים, פיתוח המערכת, הערכה לפי פריסה, פריסה וניהול שינויים, ומערכות בינה מלאכותית בסביבה תפעולית. עבור כל אחד מן התחומים אנו כוללים סקירה כללית, מצביעים על סיכונים ספציפיים לבינה מלאכותית ומציעים בקרות. אנו כוללים גם כלי עזר שיכול לשמש מבקרים בהכנות לביקורת.

עם האתגרים העיקריים שזיהו מוסדות הביקורת העליונים נמנים אלה:

- מפתחים עשויים להתמקד בביצועים טכניים, ולהתעלם מציות, שקיפות והוגנות.
- תקשורת גרועה בין בעלי המוצר למפתחים עלולה להוביל למערכות לא מועילות או יקרות.
- ארגונים רבים חסרים את המיומנויות לפיתוח בינה מלאכותית בתוך הארגון ומסתמכים על ספקים חיצוניים, מצב שיש בו כדי להגביר סיכוני ציות.
- שוררת אי-ודאות לגבי השימוש בנתונים אישיים בבינה מלאכותית, עם אחריותיות לא ברורה ומבנים ארגוניים מוגבלים.

המבקרים זקוקים להכשרה ספציפית בתחומי המומחיות הנמנים להלן כדי לבצע הערכות בעלות משמעות ביישומי בינה מלאכותית ולספק המלצות הולמות:

- ביקורות בסיסיות (baseline audits): כדי לסקור תיעוד באופן מועיל, המבקרים נדרשים להיות בעלי הבנה טובה של העקרונות הבסיסיים של מערכות בינה מלאכותית ואלגוריתמים של למידת מכונה, ובעלי ידע עדכני על ההתפתחויות הטכניות המהירות בתחום זה.
- ביקורות מעמיקות (in-depth audits): כדי לבצע ביקורות הכוללות בדיקות ביצועים משמעותיות, מבקרים זקוקים למיומנויות מעשיות בשפות תכנות נפוצות וביישום מודלים, וכן לידע וליכולת להשתמש בכלי תוכנה מתאימים.
- תשתית טכנולוגיית מידע (IT infrastructure): מאחר שמערכות בינה מלאכותית משתמשות לעיתים קרובות בפתרונות מבוססי ענן כדי למלא אחר דרישות המחשוב הגבוהות, על המבקרים להיות בעלי הבנה בסיסית בשירותי ענן ובאופן שבו הם תומכים בפעולות בינה מלאכותית.

מסמך זה מציע את המסקנות וההמלצות הבאות עבור מוסדות ביקורת עליונים:

- על מוסדות ביקורת עליונים להיות ערוכים לבקר מערכות בינה מלאכותית כדי למלא אחר התפקידים המוקנים להם בחוק, ולהעריך אם השימוש במערכת המבוקרת מספק שירותים ציבוריים יעילים ומועילים המצייתים לחוק ולכללים.
- ביקורות בינה מלאכותית דורשות מן המבקרים לרכוש מיומנויות וידע מיוחדים. על מוסדות ביקורת עליונים להשקיע בכישורים של מבקריהם: מבקרים המבצעים ביקורות ביצועים, ציות או טכנולוגיית מידע צריכים להיות בעלי הבנה כללית טובה של סוגים שונים של מערכות בינה מלאכותית, מקרי השימוש שלהם, כללי אתיקה וסיכונים ספציפיים הנוגעים לבינה מלאכותית. לצורך בדיקות ביקורת טכניות ברמה שמעבר לניהול המערכת (ממשל), דרושים מבקרים בעלי מיומנויות בפיתוח בינה מלאכותית.
- קטלוג הביקורת וכלי העזר במסמך זה יושמו בהצלחה בכמה מקרי בוחן וניתן להתאימם לביקורות עתידיות.

המחברים מקווים כי מסמך הנחיות זה והכלים המלווים אותו יתמכו בקהילת הביקורת הבין-לאומית בהפקת ביקורות מועילות של בינה מלאכותית.

2. הקדמה

סוגי מערכות בינה מלאכותית ביישומי המגזר הציבורי

הבינה המלאכותית מתפתחת במהירות, והשימוש בה במגזר הציבורי הולך ומתרחב. גופים במגזר הציבורי מאמצים בינה מלאכותית כדי לשפר יעילות, לספק שירותים משופרים ולהפחית עלויות.

מערכות בינה מלאכותית הן טכנולוגיות מעשה ידי אדם שייעודן לבצע מטלות אשר באופן טיפוסי מצריכות אינטליגנציה (בינה). מערכות אלה לומדות מתוך נתונים ומתאימות את עצמן להשגת מטרות ספציפיות, כך שעל פי רוב הן משפרות את ביצועיהן בעצמן, ללא תכנות נוסף במפורש. בינה מלאכותית כוללת תחומים כמו למידת מכונה, עיבוד שפה טבעית, ראייה ממוחשבת, זיהוי קולי, תכנון ורובוטיקה. היא מאפשרת יכולות כמו תפיסה,

4 במושג קטלוג ביקורת כוונתנו למערכת קווים מנחים הכוללים הן את התוכן המוצע של נושאי ביקורת על בסיס סיכונים, הן את המתודולוגיה לביצוע בדיקות ביקורת הנוגעות לכך.

הנמקה, קבלת החלטות ופתרון בעיות. מרבית מערכות הבינה המלאכותית מסתמכות על למידת מכונה (machine learning), שבמסגרתה מודלים לומדים דפוסים מתוך נתונים כדי להשיג מטרות ספציפיות.

מערכות בינה מלאכותית לחיזוי (predictive AI) משתמשות בנתונים היסטוריים כדי לזהות דפוסים ולספק ניבויים לגבי מקרים חדשים. לדוגמה, הן מסוגלות להעריך את הסיכון להתפתחות מחלה על פי רשומות רפואיות

סיווג לפי טכנולוגיית בסיס, מורכבות ומקרי שימוש נפוצים

סוג מערכת בינה מלאכותית	טכנולוגיית בסיס	מורכבות	מקרי שימוש נפוצים במגזר הציבורי
מערכות מבוססות כללים	• מערכות מומחה • עצי החלטה • מנועי כללים עסקיים	**נמוכה**	• קביעת זכאות (כגון סינון לקבלת הטבות) • בדיקות ציות אוטומטיות (כגון אישור דוחות מס) • תמיכה מובנית בהחלטות
בינה מלאכותית לחיזוי (סיווג)	• רגרסיה לוגיסטית • יערות אקראיים • רשתות עצביות	**בינונית**	• גילוי מרמה (כגון סימון סיכון להונאת מס) • תעדוף מקרים (כגון מיון תביעות רווחה) • זיהוי חריגים (כגון סימון יעדים לביקורת מכס)
בינה מלאכותית לחיזוי (רגרסיה)	• רגרסיה לינארית • מודלים של סדרות עיתיות • צירוף מודלים	**בינונית**	• חיזוי ביקוש (כגון ביקוש למיטות אשפוז) • הקצאת משאבים (כגון אומדני התחייבויות פנסיוניות עתידיות) • אומדן תקציב (כגון תחזיות צריכת אנרגייה)
ראייה ממוחשבת	• רשתות עצביות קונבולוציוניות • מודלים של גילוי אובייקטים • פילוח תמונות	**גבוהה**	• עיבוד מסמכים (כגון התאמה לתמונת הדרכון) • ביקורת תשתיות (כגון איתור פגמים בכביש) • מעקב וניטור (כגון ניתוח תמונות לוויין)
מודלי שפה גדולים (LLM)	• ארכיטקטורת טרנספורמר • מודלי יסוד מאומנים מראש • כוונון עדין / RAG	**גבוהה מאוד**	• טיפול בפניות האזרח (כגון צ'טבוטים בשירותים לציבור) • סיכום מסמכים (כגון ניתוח מסמכי מדיניות) • סיוע בניסוח (כגון ניסוח תכתובות)
בינה מלאכותית יוצרת (מולטי מודאלית)	• מודלי דיפוזיה • טרנספורמרים מולטי מודאליים • רשתות יריבות יוצרות	**גבוהה מאוד**	• יצירת תוכן (כגון חומרי מידע לציבור) • סימולציה של הדרכה (כגון נתונים סינתטיים לאימון מודלים) • כלי נגישות (כגון תיאורי אודיו)

הערה: סיווג המורכבות משקף אתגרים ביכולת לבצע ביקורת במערכות אלה, לרבות היכולת לפרש את תוצאות המודלים, דרישות הנוגעות לנתונים והצורך במומחיות טכנית מיוחדת. מערכות בעלות רמת מורכבות גבוהה מצריכות על פי רוב בקרות ממשל ופיקוח קפדניות יותר.

לוח 2.1: סוגי מערכות בינה מלאכותית ביישומי המגזר הציבורי

או לסמן טעויות פוטנציאליות בדוחות מס. מערכות אלה משמשות באופן שכיח לסיווג (כמו מיון מקרים לקטגוריות) או רגרסיה (כמו ניבוי ערכים מספריים).

מערכות בינה מלאכותית יוצרת (generative AI) יודעות ליצור תוכן חדש (כמו טקסט, תמונות, וידאו או אודיו) על ידי למידה מתוך קובצי נתונים גדולים. הן מסוגלות לייצר תוצרים ריאליסטיים, על פי רוב עם הכוונה אנושית מעטה או ללא הכוונה כזו כלל.

מודלי יסוד (foundation models) או **מודלי בינה מלאכותית למטרות כלליות (general-purpose AI models)** מאומנים על קובצי נתונים גדולים ומגוונים מאוד. הם משתמשים בטכניקות למידה עמוקה (deep learning) וניתן להתאים אותם למגוון רחב של יישומים, מזיהוי תמונות ועד מחקר מדעי. **מודלי שפה גדולים (Large Language Models - LLMs)** הם אחד מהסוגים העיקריים של מודלי יסוד, המאומנים בעיקר על טקסטים. הם מסוגלים לבצע מטלות רבות כמו תרגום, סיכום ושיחה על ידי הפקת שפה הדומה לשפה אנושית.

השימוש בבינה מלאכותית בשירותים לציבור טומן בחובו הזדמנויות חדשות אך גם אתגרים, ובהם הצורך להגן על נתונים אישיים, הקפדה על כך שההחלטות תהיינה הוגנות וניתנות להסבר, והימנעות מהטיה ואפליה. מערכות בינה מלאכותית שאינן בנויות או מנוהלות כראוי עלולות להגביר את עומס העבודה, לגרום לעיכובים או לפגוע באמון הציבור.

כדי להתמודד עם סיכונים אלה, יש הכרח בבקרות פנימיות לצד ביקורות חיצוניות. גופים בין-לאומיים ולאומיים פיתחו עקרונות וקווים מנחים המסייעים לשימוש אחראי בבינה מלאכותית[5]. עם התפתחות חקיקה ייעודית לבינה מלאכותית, מוקמים גופי רגולציה ופיקוח חדשים. לדוגמה, חוק הבינה המלאכותית של האיחוד האירופי (להלן - EU AI Act) מסדיר את הקמתן של רשויות ייעודיות לאכיפת ציות ולתמיכה בביקורת. רשויות אלה כוללות גופים מוסמכים[6] ורשויות פיקוח שוק[7]. לגופים אלה מוקנות זכויות גישה נרחבות לתיעוד הטכני או לרכיבי הבסיס של מערכות בינה מלאכותית, לרבות הנתונים וקוד המקור. מסגרות של פיקוח על בינה מלאכותית יכולות להוביל לקביעת תקנים והסמכות מקצועיות, דבר שיקל על ביצוע ביקורות בתחום.

מוסדות ביקורת עליונים (SAI) ממלאים תפקיד מרכזי בדרישת אחריות מגופי ציבור. בין היתר הם יכולים לדרוש שהשימוש בבינה מלאכותית יהיה בטיחותי ואתי. אומנם ביקורת בינה מלאכותית עלולה להיות כרוכה באתגרים, אך היא הולכת ונעשית חיונית יותר ככל שהטכנולוגיות הללו נפוצות יותר במגזר הציבורי. הסמכה למבקרי בינה מלאכותית מתמחים מורשים מפותחת בימים אלה[8].

מסמך זה מציע הכוונה למוסדות ביקורת עליונים בכל הנוגע לביקורת על מערכות בינה מלאכותית המשמשות במגזר הציבורי. הוא מציג בהרחבה את הסיכונים המרכזיים לצד אמצעים להפחתתם, ומפרט אילו בקרות צריכים המבקרים לצפות למצוא בארגון המבוקר. ביקורות בינה מלאכותית יכולות לכלול אלמנטים של ביקורת ביצועים וכאלה של ביקורת ציות. זו איננה רשימת קריטריונים מחייבת, ויש להשתמש בה כמדריך לפיתוח או התאמה של מתודולוגיות הביקורת.

מערכות בינה מלאכותית כמעט לעולם אינן פועלות בנפרד; הן מוטבעות על פי רוב בתהליכי טכנולוגיית מידע (IT) רחבים יותר. על המבקרים להעריך תמיד את הסיכונים הקשורים לסביבת טכנולוגיית המידע הרחבה, והערכה זו עשויה להוביל לביקורת טכנולוגיית מידע מקיפה יותר. משימה זו עשויה להצריך מומחיות של אנשי מקצוע המתמחים בבינה מלאכותית או בטכנולוגיות מידע. מדריך זה מתמקד באופן ספציפי בביקורת של רכיב הבינה המלאכותית.

ההמלצות כאן שואבות ממחקרים שהתפרסמו ומניסיונם של מוסדות הביקורת העליונים שחיברו את המסמך, לרבות ביקורות של מערכות בינה מלאכותית ופרויקטים אחרים בתחום התוכנה. המסמך משקף את הניסיון בביקורות בינה מלאכותית באותם מוסדות ביקורת עליונים שחיברו את המסמך במועד הכתיבה[9], וייתכן שהוא יעודכן עם הצטברותו של ניסיון ביקורת נוסף וקבלת תוצאות של מחקרים חדשים.

פרק 3 מציג קריטריונים אפשריים לביקורת, ובכללם תקנות לאומיות, תקנים בין-לאומיים וקווים מנחים הנמצאים

---

5 אנו מציינים אותם לצד מקורות נוספים לקריטריוני ביקורת בפרק 3 של מסמך זה.

6 EU AI Act, נספח VII; סעיף 3 (22): גוף מוסמך (notified body) הוא ארגון שהוסמך רשמית מכוח דיני האיחוד האירופי לבדוק אם מוצרים או מערכות מסוימים עומדים בתקנים הנדרשים.

7 EU AI Act, סעיף 74; סעיף 3 (26): רשות פיקוח שוק (market surveillance authority) היא הגוף הלאומי האחראי על ניטור מוצרים בשוק ואכיפת הציות לכללי האיחוד האירופי.

8 לדוגמה, ההסמכה של ISACA לתחום הבינה המלאכותית.

9 עדכון אחרון: דצמבר 2025.

בשימוש נרחב. פרק 4 מציג קטלוג ביקורת[10] הבנוי סביב תחומי מפתח: ניהול וממשל של הפרויקט, נתונים, פיתוח המערכת, הערכה לפני פריסה - כולל שיקולים אתיים כמו הוגנות והסברתיות (explainability - נקרא גם הסבריות), פריסה וניהול שינויים, ומערכות בינה מלאכותית בסביבה תפעולית. עבור כל אחד מן התחומים אנו כוללים סקירה כללית, מצביעים על סיכונים ספציפיים לבינה מלאכותית ומציעים בקרות.

למסמך זה מצורפים נספחים המכילים מידע נוסף:

3. נספח 1, רכיבים של ביקורת טכנולוגיית מידע קלאסית בהקשר של בינה מלאכותית, מסכם את רכיבי הביקורת של טכנולוגיית מידע הרלוונטיים לבינה מלאכותית.
4. נספח 2, נתונים אישיים והתקנות הכלליות להגנת המידע (GDPR) בהקשר של בינה מלאכותית, כולל סיכונים הקשורים לנתונים אישיים ולהפרת ה-GDPR.
5. נספח 3, מדדי שוויון והוגנות במערכות סיווג, מספק סקירה של מדדי השוויון וההוגנות הרלוונטיים ביותר ליישומי סיווג.
6. נספח 4 מכיל מילון מונחים, בדגש על המונחים הטכניים המשמשים במסמך זה, לצד רשימת קיצורים וראשי תיבות המשמשים במדריך זה בנוסח המקורי שלו באנגלית.
7. נספח 5 מכיל סיכום של תפקידים בגוף המבוקר העשויים להיות רלוונטיים לביקורת מערכת בינה מלאכותית.

כלי עזר נפרד לביקורת עומד גם הוא לרשות הקוראים. כלי זה מציע מגוון של שאלות ביקורת הממופות לכל אחד משלבי מחזור החיים של בינה מלאכותית, ומסייע למבקרים לזהות ראיות ביקורת מתאימות.

הכנת האמצעים החזותיים במדריך זה וכן עיבוד המדריך וכלי העזר לנוסח הסופי נעשו בחלקם בסיוע כלי בינה מלאכותית, עם בדיקות יסודיות על ידי המחברים.

3. מקורות של קריטריוני ביקורת

מאחר שבינה מלאכותית ולמידת מכונה מתפתחות ומשתנות במהירות, מסגרות החקיקה הקובעות את הכללים ליישומן עודן בתהליכי שינוי. מדינות שונות מגבשות תקנות אסדרה שמטרתן להנחות את הפיתוח של בינה מלאכותית ואת השימוש בה. פרק זה מציג סקירה כללית של הנוף האסדרתי במדינות שחיברו מסמך זה. הוא מציג גם תקנים וקווים מנחים בין-לאומיים שעשויים לשמש בידי המבקרים כדי להעריך אם משתמשים בבינה מלאכותית בצורה בטיחותית ומתוך שמירה על כללי האתיקה.

לא כל היבטי התכנון והפיתוח של מערכות בינה מלאכותית והשימוש בהן מאוסדרים באופן ברור. במקרים כאלה, המבקרים יכולים להיעזר בתקנות או בקווים מנחים קיימים מתחומים קרובים, כמו אבטחת מידע או הגנת מידע.

פרק זה מציג תחילה תקנות ויסודות משפטיים ייעודיים לתחום הבינה המלאכותית. בהמשך הוא מציין קווים מנחים אחרים, תקנים וגישות מומלצות (best practices), ומסביר בקצרה אילו דרישות מתחומים קרובים אחרים רלוונטיים בדרך כלל לבינה מלאכותית.

3.1 המסגרת המשפטית ליישומי בינה מלאכותית

מדינות שונות קובעות תקנות שמטרתן להנחות כיצד יש לפתח בינה מלאכותית וכיצד להשתמש בה. על המבקרים להבין את הנוף החקיקתי כדי להעריך אם ארגונים במגזר הציבורי מצייתים לחוק ומשתמשים בבינה מלאכותית באופן בטיחותי ואתי.

3.1.1 האיחוד האירופי: EU AI Act

EU AI Act[11] הוא דבר החקיקה המקיף הראשון להסדרת תחום הבינה המלאכותית באיחוד האירופי. מטרתו לוודא שמערכות בינה מלאכותית בשוק האיחוד האירופי הן בטוחות לשימוש ומכבדות זכויות יסוד. הגישה שנוקט החוק מתמקדת בבטיחות המוצר, בדומה לכללים הנוגעים לטכנולוגיות אחרות הכרוכות בסיכון גבוה.

3.1.1.1 תחולת ה-EU AI Act

החוק מגדיר בינה מלאכותית בצורה רחבה, אך אינו כולל מערכות מומחה פשוטות מבוססות כללים[12]. הוא מציין במפורש כי מערכות בינה מלאכותית מסיקות כיצד להפיק פלט מהקלט שהן מקבלות[13]. ארגונים יכולים למלא תפקידים שונים על פי החוק: הם יכולים להיות "ספקים" (providers) אם הם מפתחים או מזמינים מערכות בינה מלאכותית, או

10 במושג קטלוג ביקורת כוונתנו למערכת קווים מנחים הכוללים הן את התוכן המוצע של נושאי ביקורת על בסיס סיכונים, הן את המתודולוגיה לביצוע בדיקות ביקורת הנוגעות לכך.

11 הדברים נאמרים מפורשות ב-Recital 12 של ה-EU AI Act.

12 הדברים נאמרים מפורשות ב-Recital 12 של ה-EU AI Act.

13 הנציבות האירופית פרסמה קווים מנחים לגבי ההגדרה של מערכת בינה מלאכותית.

השוואה בין מסגרות אסדרה

גישות מרכזיות לממשל בינה מלאכותית לפי אזור

האיחוד האירופי - AI Act	ברזיל - PL 2338/2023	בריטניה - Pro-Innovation
גישה מרכזית סיווג מבוסס סיכונים המלווה בחובות תואמות **קטגוריות סיכון** • אסור (סיכון בלתי קביל) • סיכון גבוה (נדרש ציות לדרישות) • חובת שקיפות • סיכון מינימלי (אין דרישות) **דרישות עיקריות** • מערכות ניהול סיכונים • ממשל נתונים (נקרא גם משילות נתונים) • תיעוד טכני • פיקוח אנושי	**גישה מרכזית** מערכת שכבות סיכון המתואמות עם מסגרת האיחוד האירופי + הוראות חוק הגנת המידע על פי LGPD **קטגוריות סיכון** • סיכון מופרז • סיכון גבוה • סיכון בינוני • סיכון נמוך **דרישות עיקריות** • דרישות שקיפות • הסברתיות • פיקוח אנושי • הפחתת הטיות	**גישה מרכזית** מסגרת מבוססת עקרונות עם הנחיות ייעודיות למגזרים שונים **עקרונות מרכזיים** • בטיחות, אבטחה ועמידות • שקיפות והסברתיות • הוגנות • אחריותיות וממשל **יישום** • גופי האסדרה המגזריים מכניסים התאמות להנחיות • בקרות מידתיות • התמקדות בחדשנות • הערכות ספציפיות להקשר

הערה בנוגע לתחולת EU AI Act: ה-AI Act אינו חל על מערכות בינה מלאכותית המשמשות למטרות הגנתיות וצבאיות, למטרות ביטחון לאומי או למטרות מחקר ופיתוח.

סיכון גבוה יותר = נדרש תיעוד מקיף יותר וראיות לציות. על המבקרים להעריך אילו מסגרות אסדרה חלות, על פי תחום השיפוט וההקשר של פריסת המערכת.

לוח 3.1: השוואה בין מסגרות האסדרה המתוארות במדריך זה.

"מפעילים" (deployers) אם הם משתמשים במערכות בינה מלאכותית במסגרת פעילותם.

החוק חל על מרבית מערכות הבינה המלאכותית אשר מוצעות או נמצאות בשימוש באיחוד האירופי, אך יש חריגים מסוימים[14]. לדוגמה, הוא אינו מכסה מערכות המשמשות באופן בלעדי לצורכי ביטחון לאומי, הגנה או מחקר מדעי.

3.1.1.2 גישה מבוססת סיכונים

החוק מחיל מסגרת מבוססת סיכונים. הוא מסווג מערכות בינה מלאכותית לארבע קטגוריות[15]. איור 3.1 מציג את הקטגוריות השונות של מערכות בינה מלאכותית המוגדרות ב-EU AI Act.

- **בינה מלאכותית אסורה בשימוש:** שימושים מסוימים אסורים בחוק מלכתחילה, כמו דירוג חברתי, פרופיילינג לצורכי חיזוי, וזיהוי רגשות במקומות עבודה או בבתי ספר[16].
- **בינה מלאכותית בסיכון גבוה:** עם אלה נמנות, לדוגמה, מערכות בינה מלאכותית המשמשות

14 EU AI Act, סעיף 2.

15 הנציבות האירופית מספקת כלי לבדיקת ציות. מבקרים יכולים להשתמש בכלי זה כדי להעריך אילו הוראות של ה-AI Act חלות על מערכת בינה מלאכותית נתונה.

16 הנציבות האירופית פרסמה קווים מנחים לגבי בינה מלאכותית אסורה בשימוש.

במערכות אכיפת החוק, בתשתיות קריטיות (כולל מכשירים רפואיים) ובגישה לשירותים חיוניים. רבות מהוראות החוק מתמקדות במערכות אלה.

- **בינה מלאכותית עם חובת שקיפות:** מערכות שבאות במגע ישיר עם אנשים חייבות להיות מסומנות בבירור כבינה מלאכותית, ללא קשר לדרגת הסיכון.
- **מערכות בינה מלאכותית אחרות:** רוב מערכות הבינה המלאכותית האחרות אינן כפופות לכללים ספציפיים, אך עדיין חייבות לעמוד בחוקים הכלליים של האיחוד האירופי.

3.1.1.3 דרישות

במקרה של מערכות בינה מלאכותית בסיכון גבוה, אלה הדרישות מן הספקים[17]:

- הערכה ותיעוד של הסיכונים.
- ניהול רשומות (records) ויומני רישום (log) מפורטים של המערכת.
- פיקוח אנושי.
- עמידה בסטנדרטים של איכות נתונים, עמידות, אבטחת סייבר ודיוק.
- רישום מערכות בסיכון גבוה במסד נתונים של כלל האיחוד האירופי.

הנציבות האירופית, עם המדינות החברות, עתידה להקים מסד נתונים למערכות בסיכון גבוה. הוא יכלול תיאורים של יישום הבינה המלאכותית, של הנתונים ושל הפונקציונליות של המערכת, וכן מדריכים למשתמשים. תחול חובה לרשום במסד הנתונים גם מערכות שאינן מסווגות כמערכות בסיכון גבוה אך ורק בשל פטור.

עבור כל המערכות, בכל סיווגי הסיכון:

- אם מערכות בינה מלאכותית מקיימות מגע ישיר עם פרטים, חובה לסמן את המערכות ואת הפלטים שלהן כ"בינה מלאכותית" או "נוצר על ידי בינה מלאכותית"[18].
- יש להבטיח אוריינות בינה מלאכותית (AI literacy). ספקים ומפעילים חייבים לפעול כדי לוודא שהמשתמשים בעלי כשירות מספקת להשתמש במערכות בינה מלאכותית.

ה-EU AI Act מעודד קביעת קודי התנהלות מרצון המחילים את האסדרה של מערכות בסיכון גבוה על מערכות אחרות.

3.1.1.4 מודלים של בינה מלאכותית למטרות כלליות

החוק קובע כללים מיוחדים למודלים של בינה מלאכותית למטרות כלליות (general-purpose AI models).

מודלים עוצמתיים במיוחד מסוג זה מסווגים כמערכות הכרוכות בסיכון מערכתי (systemic risk). סיווג זה מבוסס על יכולת ההשפעה של המודל. מודל ייחשב ככרוך בסיכון מערכתי אם הוא אומן על ידי שימוש ביותר מ-1025 פעולות חישוב בנקודה צפה (FLOPs).

חשוב להבחין בין מודל בינה מלאכותית למטרות כלליות (כמו GPT-4) לבין מערכת שבנויה על גבי מודל כזה (כמו ChatGPT). האסדרה לגבי בינה מלאכותית למטרות כלליות חלה בעיקר על המודל עצמו[19]. מודלים של בינה מלאכותית למטרות כלליות אינם מסווגים כמערכות בינה מלאכותית החוסות תחת החוק, כך שהם אינם כפופים לכללים הנוגעים למערכות אסורות בשימוש או בסיכון גבוה. עם זאת, אם מודל למטרות כלליות משמש כחלק ממערכת בינה מלאכותית בסיכון גבוה, הספק והמשתמש של אותה מערכת חייבים לוודא שהמערכת עונה על כל הדרישות הרלוונטיות. ספקים של מודלי בינה מלאכותית למטרות כלליות מחויבים לספק מידע ותיעוד על המודל[20]. הדבר נועד לסייע למשתמשים להבין את יכולות המודל ומגבלותיו, ולסייע לארגונים לציית לחוק[21].

3.1.1.5 החלת ה-EU AI Act

ה-EU AI Act נכנס לתוקף באוגוסט 2024. הוראות החוק עתיד להיות מיושמות בהדרגה על פני כמה שנים. עבור מערכות בינה מלאכותית שכבר פועלות, הדרישות הרלוונטיות עשויות לחול באופן חלקי בלבד או במועד מאוחר יותר.

---

17 ראו פרטים נוספים בפרק 3 ל-EU AI Act. הנציבות האירופית עתידה לפרסם קווים מנחים לגבי מערכות בינה מלאכותית בסיכון גבוה (בנובמבר 2025).

18 הנציבות האירופית עתידה לפרסם קווים מנחים לגבי חובות שקיפות אלה (בדצמבר 2025).

19 לקבלת הקשר נוסף, ראו: S. Wachter, Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond", *Yale Journal of Law & Technology, 26, 3 (2024).*

20 הנציבות האירופית פרסמה כללי נוהל (Code of practice) לעבודה על מודלים של בינה מלאכותית למטרות כלליות.

21 לפי EU AI Act, סעיף 53 2, חובות אלה אינן חלות על ספקים של מודלי בינה מלאכותית למטרות כלליות המתפרסמים תחת רישיון קוד פתוח או רישיון חינם ואינם יוצרים סיכונים מערכתיים.

ה-AI Act הוחל על ידי מדינות EFTA של האזור הכלכלי האירופי (נורווגיה, ליכטנשטיין ואיסלנד) באיחור מסוים.

3.1.2 החקיקה וההנחיות בבריטניה

בריטניה מסתמכת על חוקים מגזריים ועל גופי אסדרה כדי לנהל בקרה על פיתוח בינה מלאכותית ושימוש בה במגזר הציבורי, וזאת במקום חקיקה רוחבית ייעודית לבינה מלאכותית. בשנת 2023 פרסמה הממשלה סקירה טכנית בשם "גישה מקדמת חדשנות לאסדרת בינה מלאכותית" (A pro-innovation approach to AI regulation)[22] המתווה עקרונות לגופי האסדרה. מסמך זה מתאר "חמישה עקרונות" לפיתוח ואסדרה של בינה מלאכותית. ואלה הם: בטיחות, אבטחה ועמידות (robustness); שקיפות נאותה והסברתיות (explainability); הוגנות (fairness); אחריותית (accountability) וממשל (governance); וכן חופש כניסה ויציאה (contestability) ופיצוי (redress). העקרונות מתארים מה חייבות מערכות בינה מלאכותית המשמשות במגזר הציבורי להשיג, וכן את הדרכים להשגת תוצאות אלה בפועל. מבקרים יכולים להשתמש בהנחיות הללו כדי להבין כיצד עשויות גישות מומלצות להיראות בשטח לאחר פריסת מערכת בינה מלאכותית.

רשויות קיימות מופקדות על יישום עקרונות אלה במגזרים שלהן, על פיתוח הנחיות אסדרה ועל הפיקוח על פריסת הטכנולוגיה. לדוגמה, נציבות המידע הבריטית (Information Commissioner's Office - ICO) עדכנה את חוק הנתונים (Data [Use and Access] Act) כך שיחול על נסיבות שבהן ארגון עשוי למכֵן קבלת החלטות (להפוך אותן לאוטומטיות) על בסיס עיבוד אוטומטי של מידע אישי[23].

ממשלת בריטניה מספקת עקרונות והנחיות למגזר הציבורי. לדוגמה, "המדריך לבינה מלאכותית במגזר הציבורי הבריטי" (AI Playbook for the UK Government) מספק הנחיות לשימוש בבינה מלאכותית באופן בטיחותי, מועיל ומאובטח על ידי עובדי המינהל הציבורי וארגונים ממשלתיים. הוא מתווה עשרה עקרונות לשימוש בבינה מלאכותית בממשלה ובמגזר הציבורי. מבקרים יכולים להשתמש בהנחיות אלה כדי לדעת מה הן הגישות המומלצות לרכש ופריסה של מערכות בינה מלאכותית[24]. כמו כן, התקן לתיעוד שקיפות אלגוריתמית (Algorithmic Transparency Recording Standard - ATRS) הוא כעת תקן מחייב עבור משרדי הממשלה[25]. ה-ATRS קובע שיטה אחידה עבור ארגוני המגזר הציבורי לפרסום מידע על אופן השימוש שלהם בכלים אלגוריתמיים ועל הסיבות לשימוש זה.

הממשלה הבריטית גם הרחיבה הנחיות קיימות בנושא וכללה בהן המלצות לגבי מערכות בינה מלאכותית. לדוגמה:

3. ***The AQuA Book*** קובע הנחיות לביצוע ניתוח מהימן התואם לצרכים העסקיים. הוא מתווה גורמים לקביעת רמת הבקרה והאימות הנדרשת, כולל הנחיות ספציפיות להבטחת איכות של מערכות בינה מלאכותית[26]. עם גורמים אלה נמנים קריטיות עסקית; חדשנות הגישה; רמת הדיוק הנדרשת בתוצרים; כמות המשאבים הזמינים לביצוע הבטחת האיכות; והשפעות על הציבור. המסמך כולל גם שיקולים בעיצוב מודלים של בינה מלאכותית, לרבות הערכת סיכונים, הערכת השפעות, ביקורת הטיות וביקורת ציות.
4. ה-***Magenta Book*** קובע הנחיות להערכת השפעתן של התערבויות בינה מלאכותית[27].
5. המרכז הלאומי לאבטחת סייבר (National Cyber Security Centre) קובע עקרונות לאבטחת מערכות למידת מכונה[28].
6. קיים גם מסמך הנחיות לרכש בינה מלאכותית ***Guidelines for AI Procurement***, המתאר גישות מומלצות לרכישת טכנולוגיות בינה מלאכותית בגופים ממשלתיים[29].

נוסף על כך קיימות הנחיות לאישור ולהבטחת איכות של בינה מלאכותית:

7. המסמך ***Introduction to AI Assurance*** של ממשלת בריטניה מביא סקירה כללית של מושגים בהבטחת

---

22 למידע נוסף ראו: A pro-innovation approach to AI regulation - GOV.UK.

23 למידע נוסף ראו: Data protection | ICO.

24 המדריך נגיש כאן: AI Playbook for the UK Government - GOV.UK.

25 למידע נוסף ראו: Algorithmic Transparency Recording Standard Hub - GOV.UK.

26 למידע נוסף ראו: Design - The AQuA Book 7.

27 למידע נוסף, ראו: Guidance on the Impact Evaluation of AI Interventions (HTML) - GOV.UK.

28 למידע נוסף, ראו: Principles for security of Machine learning ML - NCSC.GOV.UK.

29 למידע נוסף, ראו: Guidelines for AI procurement - GOV.UK.

איכות למערכות בינה מלאכותית. הוא אינו כולל פרטים טכניים להבטחת איכות של בינה מלאכותית, אך זהו מקור שימושי למידע רקע[30].

8. המסמך ***AI in Audit*** שפורסם מטעם המועצה לדיווח פיננסי (Financial Reporting Council) מציג חקר מקרים לבינה מלאכותית בביקורת לצד הנחיות לתיעוד כלי בינה מלאכותית[31].

9. הנחיות הממשלה הבריטית לבקרת הוצאות דיגיטל וטכנולוגיה כוללות הנחיות לגבי רכש בינה מלאכותית במסגרת ***Risk and Importance***[32].

3.1.3 חקיקה והנחיות בברזיל

ברזיל מפתחת מסגרת חוקית לאסדרת בינה מלאכותית. הצעת חוק מס' 2338 משנת 2023[33] שהוגשה לסנאט קובעת כללים לפיתוח בינה מלאכותית ושימוש בה ברחבי המדינה. הצעת החוק מכסה טווח רחב של נושאים, כולל פיתוח כלכלי, מדעים וטכנולוגיה, הדין האזרחי וזכויות הפרט.

הסנאט אישר את הצעת החוק והיא נבחנת כעת על ידי בית הנבחרים. מחוקקים הציעו תיקונים אחדים. אלה כללו דרישות לסימון תוכן שנוצר באמצעות בינה מלאכותית, קידום מחקר והכשרה, הגדלת הענישה על הכפשה באמצעות בינה מלאכותית והגנה על ילדים ומתבגרים. תיקונים נוספים מתייחסים לפיתוח בינה מלאכותית ושימוש בה, שינויים בחוק הגנת המידע הכללי (LGPD) של ברזיל, השתתפות הציבור באסדרה והחובות החוקיות החלות על מפתחי בינה מלאכותית. הצעת החוק גם מכסה נושאים של זכויות יוצרים, קניין רוחני, זכויות עובדים וחופש הביטוי. תחולת הצעת החוק היא רחבת היקף. היא מטפלת בנושאים של אחריות, בטיחות, אמון, כיבוד זכויות אדם, דמוקרטיה, הגנת הסביבה, פיתוח בר קיימה, סיווג סיכונים, ממשל תאגידי, הערכת השפעות, חבות אזרחית ועונשים.

מוסד הביקורת העליון של ברזיל (Federal Court of Accounts - TCU) פרסם גם הנחיות מעשיות לשימוש אתי בבינה מלאכותית. ביולי 2024 פרסם TCU מדריך לשימוש בבינה מלאכותית יוצרת. המדריך מיועד לכל העובדים, הקבלנים והמתמחים ב-TCU. מטרתו לתמוך בחדשנות ופרודוקטיביות ובה בעת להבטיח ביטחון נתונים, פרטיות ומהימנות[34].

מדריך ה-TCU קובע כללים ברורים:

- **פתרונות בינה מלאכותית מאושרים:** רק כלי בינה מלאכותית מסוימים מאושרים לשימוש עם מידע סודי. כלים שאינם מאושרים על ידי TCU, כמו ChatGPT ו-Gemini, מותרים לשימוש רק עם נתונים ציבוריים ובתנאים מחמירים.
- **אחריותיות המשתמשים:** עובדי הארגון ימשיכו לשאת באחריות מלאה לכל תוכן שמופק בעזרת בינה מלאכותית, גם אם הכלי הפיק טעויות.
- **פיקוח אנושי:** כל ההחלטות האוטומטיות מחייבות סקירה אנושית, במיוחד כשמדובר במידע אסטרטגי או כזה המיועד לציבור.
- **אימות תוכן (ולידציה):** תוכן שנוצר באמצעות בינה מלאכותית חייב להיבדק בהיבטים של דיוק, הוגנות והלימות. עובדי הארגון אינם רשאים להשתמש בבינה מלאכותית אם אין להם אפשרות לאמת את הפלט.
- **אבטחה וציות:** על השימוש בבינה מלאכותית להיות תואם למדיניות אבטחת המידע של TCU. אי-ציות עלול להוביל להליך משמעתי או הפרת חוזה.
- **קניין רוחני:** עובדי הארגון אינם רשאים להשתמש בתוכן שנוצר באמצעות בינה מלאכותית אם הוא עלול להפר זכויות יוצרים, ועליהם לסמן חומר שנוצר באמצעות בינה מלאכותית במקרה המתאים.
- **פיתוח פתרונות חדשים:** כלי בינה מלאכותית יוצרת (generative AI) חדשים חייבים באישור יחידת טכנולוגיית המידע (IT) וועדת ההנהלה. רק יחידת ניהול ה-IT רשאית לפתח יישומי בינה מלאכותית המיועדים לקהלים חיצוניים.

המדריך שם דגש גם על סיכוני ההטיה הטמונים בבינה מלאכותית. הוא מבחין בין הטיית מודל (מנתוני האימון) לבין הטיית אוטומציה (הנטייה האנושית לתת אמון בהצעות אוטומטיות). המדריך מדגיש כי כל שימוש בבינה מלאכותית חייב להתאים לקוד ההתנהלות של TCU ולמדיניות נגד אפליה.

---

30 למידע נוסף, ראו: Introduction to AI assurance. - GOV.UK

31 למידע נוסף, ראו: AI in Audit.

32 למידע נוסף, ראו: Risk and Importance Framework - GOV.UK.

33 PL 2338/2023.

34 למידע נוסף, ראו: Guia de uso de Inteligencia Artificial no Tribunal de Contas da Uniao (TCU).

3.2 תקנים, קווים מנחים וגישות מומלצות

ארגונים שונים כבר הגדירו תקנים וקווים מנחים עבור היבטים שונים בפיתוח בינה מלאכותית.

הארגון לשיתוף פעולה ופיתוח כלכלי (OECD) פרסם עקרונות לבינה מלאכותית (AI Principles) ב-2019 ועדכן אותם ב-2024[35]. עקרונות אלה מצביעים על חמישה ערכי ליבה: שקיפות, הסברתיות, עמידות, בטיחות ואבטחה. הארגון גם מספק המלצות, למשל השקעה במחקר ופיתוח בתחום הבינה המלאכותית.

המכון הלאומי לתקנים וטכנולוגיה של ארה"ב (National Institute of Standards and Technology - NIST) פרסם מסגרת לניהול סיכונים בתחום הבינה המלאכותית (AI Risk Management Framework) ב-2023[36]. מסגרת זו מסייעת לארגונים לזהות, להעריך ולנהל סיכונים המזוהים עם בינה מלאכותית. היא מכסה נושאים כמו ביסוס תרבות של מודעות לסיכונים, ניתוח סיכונים ופיתוח בקרות נאותות.

קבוצת מומחים לבינה מלאכותית באיחוד האירופי (EU High-Level Expert Group on Artificial Intelligence) פרסמה קווים מנחים אתיים לבינה מלאכותית מהימנה[37]. קווים מנחים אלה דורשים ממערכות בינה מלאכותית לעמוד בדרישות של חוקיות, אתיקה ועמידות. הם כוללים שבע דרישות לבינה מלאכותית מהימנה ורשימת תיוג שתכליתה לסייע לארגונים להעריך את מידת עמידתם בדרישות.

מכון פראונהופר לניתוח בינה ומערכות מידע (The Fraunhofer Institute for Intelligent Analysis and Information Systems - IAIS) תרגם דרישות אלה לקטלוג פרקטי של הערכות, המקל על ארגונים ליישם את הקווים המנחים בשטח[38].

בשנת 2024 אימצה המועצה האירופית את אמנת המסגרת לבינה מלאכותית וזכויות אדם, דמוקרטיה ושלטון החוק (Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law)[39].

האמנה דורשת מן המדינות החתומות עליה להגן על זכויות אדם, על תהליכים דמוקרטיים ועל שלטון החוק לאורך כלל מחזור החיים של מערכות בינה מלאכותית.

3.3 מקורות של קריטריוני ביקורת מתחומים משיקים

לא כל היבטי הבינה המלאכותית מוסדרים באסדרה ייעודית. במקרים של העדר אסדרה, המבקרים נדרשים לבחון תקנים ודרישות מקובלים מתחומים קרובים. אלה יכולים לספק יסודות חזקים להערכת שימוש בטוח ואחראי בבינה מלאכותית.

3.3.1 אבטחת מידע וממשל IT

מערכות בינה מלאכותית מעבדות לעיתים קרובות נתונים רגישים או תומכות בפעילות הקריטית של מוסד. לכן סטנדרטים של אבטחת מידע רלוונטיים במידה רבה. המבקרים יכולים לבדוק את הציות לתקני אבטחת מידע רלוונטיים - לאומיים או בין-לאומיים. לדוגמה, משרד אבטחת המידע הפדרלי הגרמני (BSI) פיתח גישה הנקראת "IT-Grundschutz". היא כוללת כמה פרסומים עיקריים[40]:

תקציר (קומפנדיום) של "IT-Grundschutz", המתאר איומים פוטנציאליים ודרישות אבטחה.

תקנים של ה-BSI, שבהם מומלצים תהליכים, שיטות ומדדים של אבטחת מידע.

בדומה לכך, מסגרות כמו COBIT של ISACA יכולים להנחות ממשל IT ולסייע להבטיח את קיומן של בקרות תקפות[41].

3.3.2 שימוש בשירותי ענן חיצוניים

מערכות בינה מלאכותית רבות, במיוחד אלה המשתמשות במודלי שפה גדולים (LLM), מסתמכות על שירותי ענן חיצוניים. שירותים אלה אומנם מציעים גמישות וסקלביליות (יכולת הרחבה), אבל הם גם יכולים להכניס סיכונים חדשים למערכת.

---

35 לאחר פרסום מסמך עקרונות הבינה המלאכותית (AI principles) המקורי, פורסם גם ארגז כלים (Toolkit) של מדעי הנתונים, שנועד לסייע ביישום העקרונות לתחום הבינה המלאכותית.

36 למידע נוסף, ראו: AI Risk Management Framework.

37 High-Level Expert Group on AI (2019): *Ethics Guidelines for Trustworthy Artificial Intelligence*.

38 למידע נוסף, ראו: AI assessment catalogue.

39 למידע נוסף, ראו: Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.

40 למידע נוסף, ראו: IT-Grundschutz.

41 ISACA, *COBIT 2019 Framework: Governance and management objective (2018)*.

פרסום ה-BSI "שימוש מאובטח בשירותי ענן" *(Secure Use of Cloud Services)*[42] מתאר כיצד אפשר להשתמש בשירותי ענן בבטחה. כדי לעיין בהערכות טכניות של שירותי ענן המבקרים יכולים להיעזר במסגרות שפרסם ה-BSI, כגון "קטלוג בקרות ציות במחשוב ענן" *(Cloud Computing Compliance Controls Catalogue - C5)*[43] ו"קטלוג קריטריונים לשירותי ענן לבינה מלאכותית" *(AI Cloud Services Criteria Catalogue - AIC4)*[44]. קיים גם המסמך "מטריצת בקרות ענן" *(Cloud Controls Matrix - CCM)* שפרסמה הברית לאבטחת ענן *(Cloud Security Alliance - CSA)*[45].

3.3.3 הגנת המידע

מערכות בינה מלאכותית משתמשות על פי רוב במאגרי נתונים גדולים העשויים לכלול נתונים אישיים. דיני הגנת המידע, כמו התקנה הכללית להגנת מידע (GDPR) של האיחוד האירופי, חלים הן על הפיתוח של בינה מלאכותית, הן על השימוש בה. כמה רשויות לאומיות להגנת המידע פרסמו קווים מנחים להגנת מידע ביישומי בינה מלאכותית; למשל ה-ICO בבריטניה[46], רשות הגנת המידע הנורווגית (Datatilsynet)[47], חוק הגנת המידע הכללי של ברזיל[48] וכן ועידת הגנת המידע הגרמנית (DSK)[49]. נספח 2 מסכם כמה מן האתגרים המרכזיים הקשורים לציות ל-GDPR עבור מערכות בינה מלאכותית.

3.3.4 שקיפות ואי-אפליה

ארגוני המגזר הציבורי מחויבים לנהוג בשקיפות לגבי תהליכי קבלת החלטות. זכותם של האזרחים להבין כיצד מתקבלות החלטות המשפיעות עליהם, לרבות החלטות שנעזרות במערכות בינה מלאכותית ואף מתקבלות באמצעותן. לדוגמה, סעיף 41(2c) למגילת זכויות היסוד של מדינות האיחוד האירופי (Charter of Fundamental Rights of the European Union) קובע כי חובתם של גופי ממשל לנמק את החלטותיהם. ברמה הלאומית, הזכות לשקיפות והסברתיות מעוגנת לא פעם בחוקי חופש המידע.

בדומה לחוקי הגנת המידע, חוקים נגד אפליה מגינים על זכויות יסוד, ועל כן חלים גם על יישומי בינה מלאכותית. עם זאת חוקים אלה מנוסחים פעמים רבות בדרך שמגינה על הפרט מפני אפליה במקרים פרטניים או במצבים ספציפיים. עשויה להידרש פרשנות רחבה כדי שיחולו על אפליה אלגוריתמית המעמידה קבוצות אנשים בעמדת נחיתות סטטיסטית.

3.3.5 אתיקה בבינה מלאכותית

דרישות אתיקה רבות הנוגעות לבינה מלאכותית מושרשות בזכויות יסוד כמו פרטיות, הוגנות וכבוד האדם. קווים מנחים בין-לאומיים יכולים לסייע למבקרים לגבש קריטריונים לבינה מלאכותית אתית. קווים מנחים אלה מבוססים על פי רוב על אמנות לזכויות אדם ועל חוקות של מדינות שונות. לדוגמה, קבוצת המומחים לבינה מלאכותית באיחוד האירופי ביססה חלקית את הקווים המנחים האתיים לבינה מלאכותית מהימנה על אמנות האיחוד האירופי, על מגילת זכויות היסוד של האיחוד האירופי ועל הדין הבין-לאומי לזכויות אדם[50]. גישה זו יכולה לשמש מודל לפיתוח קריטריוני ביקורת לבינה מלאכותית אתית, המעוגנים בזכויות אדם בסיסיות שניתן למצוא בחוקות רבות ובמסמכי התחייבות בין-לאומיים[51].

---

42 למידע נוסף, ראו: Secure Use of Cloud Services.

43 למידע נוסף, ראו: Cloud Computing Compliance Controls Catalogue (5).

44 למידע נוסף, ראו: AI Cloud Services Criteria Catalogue (AIC4).

45 למידע נוסף, ראו: Cloud Controls Matrix.

46 הנחיות מטעם משרד נציב המידע: UK GDPR guidance and resources וכן Guidance on AI and data protection.

47 למידע נוסף, ראו: Artificial intelligence and privacy.

48 למידע נוסף, ראו: LEI N° 13.709.

49 למידע נוסף, ראו: Guidance on Recommended Technical and Organisational Measures for the Development and Operation of AI Systems וכן Guidance on Artificial Intelligence and Data Protection (בגרמנית בלבד).

50 למידע נוסף, ראו: Ethics Guidelines for Trustworthy AI.

51 משרד הביקורת הלאומי של נורווגיה ביסס קריטריוני ביקורת לפרטיות ואוטונומיה של האדם על האמנה האירופית לזכויות אדם, בשילוב עם החוקה הנורווגית. הוא פרסם את הקריטריונים בדוח ביקורת על השימוש בבינה מלאכותית בממשלה המרכזית (בנורווגית בלבד; דוח מסכם ללא קריטריוני הביקורת זמין באנגלית).

3.3.6 קניין רוחני והשפעה סביבתית

השימוש בבינה מלאכותית, ובמיוחד השימוש במודלי שפה גדולים (LLM), מעורר שאלות לגבי קניין רוחני. אימון נתונים עשוי לכלול חומרים הכפופים לזכויות יוצרים. ה-EU AI Act והדירקטיבה לזכויות יוצרים של האיחוד האירופי (EU Copyright Directive)[52] הם דברי חקיקה המסדירים דיני זכויות יוצרים ביחס לבינה מלאכותית במדינות האיחוד האירופי. בריטניה דורשת מארגונים לציית לדין הבין-לאומי לזכויות יוצרים ושוקלת לעדכן את מדיניותה כדי לספק בהירות לגבי תחולתו על בינה מלאכותית[53].

מערכות בינה מלאכותית יכולות גם להשפיע משמעותית על הסביבה, בפרט אלה שמצריכות כוח מחשוב רב. יש תקנות הדורשות מארגונים לדווח על טביעת הרגל הסביבתית של מערכות הבינה המלאכותית שלהם. לדוגמה, ה-EU AI Act מעודד שימוש בבינה מלאכותית בדרך שממזערת פגיעה בסביבה[54], וקובע כי במקרה של מערכות בינה מלאכותית מסוימות, חלה דרישה לספק מידע על ההשפעה הסביבתית מבחינת צריכת האנרגייה. בבריטניה יש ארגונים במגזר הציבורי הנדרשים לתת גילוי נאות למידע מסוים על קיימות, וצפוי כי הנחיה זו תורחב כדי לכלול את ההשפעות של בינה מלאכותית. הערכת ההשפעה הסביבתית של בינה מלאכותית היא תחום חדש הראוי לתשומת לב בהקשר של ביקורת. אף שהשיטות למדידת השפעה זו עודן בתהליכי פיתוח, כלים מסוימים כבר הותוו על ידי ה-OECD[55].

4. קטלוג ביקורת

קטלוג הביקורת המתואר בפרק זה כולל קווים מנחים לביקורת מערכות בינה מלאכותית, הכוללים הן את התוכן המוצע של נושאי ביקורת על בסיס סיכונים, הן את הבקרות או האמצעים להפחתת הסיכונים שעל המבקרים לצפות לראות בגוף המבוקר.

מערכות בינה מלאכותית מוטמעות על פי רוב בתשתיות IT רחבות יותר. אף שיישומי בינה מלאכותית כרוכים באתגרים שונים מאלה המזוהים עם מערכות תוכנה או IT מסורתיות, עדיין ניתן לארגן את הביקורת על יישומי בינה מלאכותית ומודלי למידת מכונה בעזרת מסגרות עבודה המוכרות למבקרי IT - כמו התהליך התקני חוצה התעשיות לכריית מידע (Cross-Industry Standard Process for Data Mining), הידוע בשמו המקוצר CRISP-DM[56]. המסמך הנוכחי מתמקד בסוגיות הנוגעות במפורש ליישומי בינה מלאכותית, ובפרט בשימוש בטיחותי, אחראי ואתי בבינה מלאכותית. השיקולים העיקריים כוללים שקיפות והסברתיות של החלטות מערכות הבינה המלאכותית, שוויון והוגנות של החלטות אלה ואבטחת המערכות. על היבטים אלה להיות משולבים בתהליך הפיתוח לכל אורכו, בהתאם לעקרונות של הוגנות מובנית בתכנון ("by design"), שקיפות מובנית בתכנון ופרטיות מובנית בתכנון. לפיכך, אין ייעודו של קטלוג ביקורת זה לשמש כרצף לינארי של צעדים יחידים, אלא שהוא מייצג לולאה רציפה של שלבים המקיימים אינטראקציה ביניהם.

4.0 עומק הביקורת ומבנה הביקורת

ביקורת בינה מלאכותית יכולה להתבצע ברמות שונות, המחייבות דרגות שונות של מומחיות טכנית מצד המבקרים ודרגות משתנות של גישה לרכיבים הטכניים של המערכת:

1. ברמה הבסיסית (baseline), הביקורת מתמקדת בסקירת התיעוד כדי לוודא שיש התייחסות לכל רכיבי הממשל העיקריים, שהסיכונים מזוהים ושהותוו אסטרטגיות מתאימות להפחתת הסיכונים. גישה זו יכולה להעריך את מידת השליטה של הארגון בגורמי הסיכון, אך היא לא קובעת אם מערכת הבינה המלאכותית מפירה בפועל עקרונות כמו יחס שוויוני.
2. כאשר המבקרים מקבלים גישה להרצת המודל ולבדיקתו באמצעות נתוני ייצור אמיתיים או נתונים סינתטיים, הם יכולים להעריך באופן ישיר את ההתנהגות ואת התוצרים של המערכת. זה מאפשר להם לבקר את השפעת המערכת בפועל. לדוגמה, המבקרים יכולים לבדוק מערכות חיזוי מבוססות בינה מלאכותית על ידי שינוי משתנה קלט יחיד והשוואת התוצאות, או לחפש פגיעויות במערכת בינה מלאכותית יוצרת על ידי שימוש בפרומפטים שנוסחו במיוחד לצורך זה.
3. בביקורת מקיפה של מערכת בינה מלאכותית, בדיקות טכניות נוספות יכולות להקנות תובנות רבות ערך לגבי

---

52 למידע נוסף, ראו: Directive 2019/790.

53 למידע נוסף, ראו את הנחיות ממשלת בריטניה: Copyright and Artificial Intelligence.

54 EU AI Act, Article 95.

55 למידע נוסף, ראו: OECD, Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications.

56 ראו בנספח 1 תיאור של מבנה ביקורת בינה מלאכותית בהתאם למסגרת CRISP-DM.

סקירה כללית של קטלוג הביקורת
נושאים עיקריים ושאלות ביקורת לדוגמה לפי תחום

4.1 ניהול וממשל של הפרויקט

נושאים עיקריים
- הצעת פרויקטים של בינה מלאכותית
- דרישות חוקיות ואתיות החלות בתחום השיפוט הספציפי
- ניהול מחזור החיים של בינה מלאכותית
- הערכת סיכונים של בינה מלאכותית

באילו תהליכים עסקיים ישמש היישום?
מה הן הראיות לניתוח עלות-תועלת?

4.2 נתונים

נושאים עיקריים
- כמות הנתונים ואיכותם
- יצירה מבוססת-אחזור (RAG) ומודלי שפה גדולים (LLM)
- סוגיות של פרטיות בהקשר של מערכות בינה מלאכותית

מה הם מקורות הנתונים?
כיצד תעריכו את איכות הנתונים?

4.3 פיתוח המודל

נושאים עיקריים
- תהליך הפיתוח ומאפייני הביצוע
- הבטחת איכות, מהימנות ועמידות בפיתוח
- ציפיות לגבי תיעוד ובקרת גרסאות

כיצד בחרתם את המאפיינים ששימשו ליצירת המודל?
מה הם מאגרי הנתונים המשמשים אתכם לאימות (ולידציה)?

4.4 הערכה לפני פריסה

נושאים עיקריים
- בדיקות ביצועים וקבלה
- שקיפות והסברתיות
- הוגנות ומזעור פגיעה
- אבטחה
- השפעה סביבתית

כיצד אתם מבטיחים את הוגנות המודל?
כיצד אתם מגינים על היישום מפני שימוש לרעה?

4.5 פריסה וניהול שינויים

נושאים עיקריים
א. מוכנות עסקית וניהול שינויים
ב. קריטריונים לעלייה לאוויר ושערים

מה הם החלקים המרכיבים את מערכת למידת המכונה?
כיצד היישום מוטבע בארכיטקטורת המערכת הסובבת?

4.6 מערכות בינה מלאכותית בסביבה תפעולית

נושאים עיקריים
1. ניטור ומעקב ביצועים
2. תגובה לתקריות ועדכוני המודל
3. ציות שוטף ונתיבי ביקורת

כיצד אתם מנטרים את ביצועי המודל לאורך זמן?
מה יניע אתכם לסקור את המודל או לבצע אימון מחדש?

כל תחום ביקורת כולל סיכונים ובקרות מצופות. השאלות לדוגמה נועדו להמחשה, ויש להתאימן לכל מערכת בינה מלאכותית העומדת לבחינה.

לוח 4.1: סקירה כללית של קטלוג הביקורת. השאלות לדוגמה נועדו להדגמת כלי העזר לביקורת.

מרכיבי המערכת:

- ביקורת מדוקדקת של הנתונים וסקירת הקוד יכולות להקנות ביטחון רב יותר לגבי מידת הדיוק של התיעוד, בפרט בנוגע למפרטים של מודל למידת מכונה.
- עבור מערכות חיזוי מבוססות בינה מלאכותית ייתכן כי יידרש שחזור של האימון, הבדיקות, הדירוג ומדדי הביצועים של המודל (או חלקים ממנו) כדי להבין ולאמת את פרטי המודל, ביצועיו והדירותו, וכן את ההשלכות שלו לגבי הוגנות. הדבר דורש תשתית שמסוגלת לתמוך בתהליך אימות מסוג זה. יש להביא בחשבון את העלויות הקשורות בכך ואת התועלות הפוטנציאליות של גישה זו כאשר שוקלים לנקוט אותה.
- פיתוח חלופות מתאימות למודל יכול להועיל בהבלטת ליקויים ודרכים למנוע אותם. כמו בסעיף (3ב), נדרשת בחינה קפדנית בטרם מחליטים לפתח מודלים חלופיים, מפני שמהלך זה עלול להוביל לתפקידים החורגים מתפקידיו של מוסד הביקורת העליון הקבועים בחוק. כמו כן יש לשקול את העלות והתועלת של גישה זו. במקרה של מערכות בינה מלאכותית מורכבות ו/או מערכות שנרכשו מצד שלישי, יש להניח כי גישה כזו לא תהיה ישימה עבור המבקרים.

מבקרים המעוניינים לבצע את כל רמות הסקירה צריכים להיות בעלי ידע מעמיק על הטכנולוגיה של המערכת כדי לתת המלצות שיפור בעלות משמעות. כאשר הארגון המבוקר נעדר מומחיות פנימית, יש לזהות כל ליקוי ולהציע המלצות מתאימות. משימה זו עשויה להצריך מומחיות של אנשי מקצוע המתמחים בבינה מלאכותית או בטכנולוגיות מידע. עומק הביקורת המתאים ייקבע על פי הסיכונים הקשורים למערכת הבינה המלאכותית, מטרת הביקורת ושאלות הביקורת הכלליות[57].

גורם מגדיר נוסף בעת תכנון ביקורת בינה מלאכותית הוא בשלותו של הגוף המבוקר ושאיפותיו ביחס לבינה מלאכותית, כמו גם השאלה האם המערכת פותחה על ידי הגוף המבוקר או על ידי גוף חיצוני שממנו נרכשה. רמת התיעוד הזמין ואופיו ישתנו משמעותית, קרוב לוודאי, בהתאם למבנה הארגוני של הגוף המבוקר, השקעתו בתחום הבינה המלאכותית והיקף הפרויקט שלו. במקרה של מערכות הנרכשות מגורם חיצוני, כאשר קוד המקור ומפרט המודל המפורט אינם ידועים לגוף המבוקר (או שאין לו אפשרות לאמת אותם), הגוף המבוקר עדיין אחראי לספק תיעוד מקיף המוכיח את עמידתו בסטנדרטים נאותים של תהליכי מינהל ציבורי.

על המבקרים להניח כי הנתונים, או מקטעי נתונים רלוונטיים, זמינים לצד התיעוד. עם זאת ייתכן שהקוד והמודל עצמו אינם בפורמט נגיש למבקרים[58], במיוחד במקרה של מערכות בינה מלאכותית שנרכשו מגוף חיצוני. במקרים כאלה, שיתוף פעולה הדוק עם הארגון המבוקר הוא הכרחי לצורך בדיקות ו/או הדגמות באתר הארגון, שיבוצעו בפיקוחם של מבקרים בעלי הבנה נאותה בנושא.

4.1 ניהול וממשל של הפרויקט

ניהול וממשל מועילים של הפרויקט הם יסוד חיוני להצלחת הפיתוח והפריסה של מערכות בינה מלאכותית במגזר הציבורי.

ביקורת של ניהול וממשל בפרויקט בינה מלאכותית מתמקדת בפיקוח הארגוני, ולא בביצועים הטכניים של המערכת. היא מבוצעת ברמת הממשל ומתבססת בראש ובראשונה על סקירת מסמכים וראיונות, ואינה כרוכה בבדיקות ישירות של מערכת הבינה המלאכותית.

השלב הראשון בתהליך הוא בחינת הגודל והמבנה של הארגון הנבדק, שכן גורמים אלה משפיעים על הסדרי הממשל ועל היקף התיעוד הזמין. בארגונים גדולים ייתכן שכבר גובשו אסטרטגיות ועקרונות לתחום הבינה המלאכותית, ואילו בארגונים קטנים ייתכן שהפרויקט מנוהל באופן פחות רשמי. כאשר היקף התיעוד מוגבל, ראיונות יכולים להקנות תובנות חשובות לגבי פרקטיקות ניהול הפרויקט. לכל הפחות, על המבקרים לצפות לראות ראיות למרכיבי הממשל העיקריים, כפי שמתואר בתתי-הסעיפים הרלוונטיים שכותרתם "בקרות מצופות".

4.1.1 הצעת פרויקטים בתחום הבינה המלאכותית

הבנת מקורותיו ומטרותיו של פרויקט בינה מלאכותית היא צעד ראשון קריטי. מרבית הארגונים דורשים שהצעה לפרויקט תאושר לפני תחילת העבודה, אם כי רמת

57 לדוגמה, במסמך "שימוש בבינה מלאכותית בממשלה המרכזית" (The Use of Artificial Intelligence in the Central Government), שאלות הביקורת הכלליות נגעו לשימוש אחראי בבינה מלאכותית בממשלה המרכזית, ואילו הפרטים הטכניים של המערכות שהיו בשימוש בעת ההיא לא היו חלק מתכולת הביקורת. כך שהביקורת של ארבע מערכות בינה מלאכותית שהוצגו כחקרי מקרה במסמך בוצעה בפורמט של ביקורת בסיסית (baseline) שבדקה היבטים של ממשל בינה מלאכותית.

58 מודלי למידת מכונה רבים מפותחים בשפת פייתון או R, אבל יש גם אפשרויות אחרות לכתיבת התוכנה. לא סביר אפוא לדרוש שהמבקרים יכירו כל שפת קוד ומסגרת תוכנה שקיימות. העלויות הקשורות לרכישת תוכנה נוספת ו/או כוח מחשוב נוסף עלולות ליצור בעיות נוספות.

הפורמליות עשויה להשתנות. הצעה נאותה תכלול פרטים אלה:

- המטרות והיעדים של הפרויקט.
- ההיקף המיועד והתוצרים.
- מדדי ביצועים וקריטריונים להצלחה.
- הערכה של דרישות מוקדמות, כמו זמינות נתונים ומומחיות רלוונטית.
- ניתוח עלות-תועלת.
- זיהוי בעלי עניין מרכזיים.
- הערכה של טכנולוגיות מתאימות.
- הסדרים לניהול מערכת הבינה המלאכותית לכל אורך מחזור החיים שלה.

הערכת דרישות מוקדמות כמו נתונים זמינים ומומחיות נדרשת, וגיבוש הבנה ברורה של הפוטנציאל והמגבלות של טכנולוגיית הבינה המלאכותית, דורשים על פי רוב שיתוף פעולה בין מומחים טכניים למומחים מקצועיים לתחום העיסוק (domain experts) בשלבים המוקדמים ביותר.

אפילו בפרויקטים גמישים, שהתוכניות לגביהם עשויות להשתנות, חשוב להגדיר מטרות, דרישות ומדדי ביצוע בשלבים הראשונים.

כאשר קיימים כללי מדיניות פנימיים שמתייחסים לבינה מלאכותית, תוכניות הפרויקט צריכות להפנות אליהם כדי להבטיח תיאום עם עקרונות הממשל הארגוניים - גם עבור מערכות שנרכשו או פותחו מחוץ לארגון.

עבור פרויקטים מורכבים או בעלי עלות גבוהה, הוכחת היתכנות (proof of concept) יכולה לסייע לבדוק את הישימות ולהקטין את הסיכון לתוצאות בלתי מספקות. בתהליך זה יש להעריך הן את הביצועים הטכניים, הן את היכולת לעמוד בדרישות פונקציונליות. פרויקטים של בינה מלאכותית מעל היקף מסוים עשויים להוביל לדרישות רגולטוריות נוספות בנוגע למחקרי ישימות ותיעוד.

מערכת בינה מלאכותית צריכה להראות יתרונות ברורים ומדידים בהשוואה לגישות קונוונציונליות כדי להצדיק את מורכבותה, עלותה והסיכונים הכרוכים בה. אף שמודלי למידה עמוקה נפוצים בשימוש, הם גוררים לעיתים קרובות עלויות שאינן נראות לעין מייד. לדוגמה, אימון או כוונון עדין של מודלי למידה עמוקה מצריכים שימוש בחומרה מתמחה, כגון יחידות עיבוד גרפיקה (GPU), שניתן לפרוס בארגון באופן מקומי או לשכור מפלטפורמות ענן. פריסה מקומית בארגון מצריכה השקעה משמעותית בחומרה, באנרגייה ובקירור, ונלווה לה הסיכון הנוסף של התיישנות עם הזמן. פריסה מבוססת ענן מתומחרת בדרך כלל במודל של תשלום לפי שימוש (pay-as-you-go), כך שהגבייה היא לפי עומס העבודה הנדרש לאימון והרצה של המודל. העלויות עלולות לעלות במהירות בהתאם למשך הזמן והאינטנסיביות של תהליכים אלה.

למרות אתגרים אלה, מערכות בינה מלאכותית עשויות לספק יתרונות מהותיים. ביניהם אפשר לציין שיפורי יעילות ופרודוקטיביות באמצעות מיכון מטלות, הפחתת טעויות אנוש והגברת החדשנות. מודלי למידת מכונה, לדוגמה, מסוגלים לזהות דפוסים וקשרים שאינם ניכרים לעין האנושית בנקל, כך שהם מאיצים תהליכים כמו גילוי תרופות או ניתוח נתונים בקנה מידה גדול. בינה מלאכותית יכולה גם לחזק אבטחת סייבר ולתרום לניצול מיטבי של מערכות אנרגייה, כולל תכנון תחנות כוח ואמינות אספקת החשמל. במגזר הציבורי, בינה מלאכותית מציעה הזדמנויות לניהול יעיל ושירותים מותאמים אישית בעלות נמוכה יותר.

על המבקרים לוודא כי בוצע ניתוח עלות-תועלת יסודי לפני הטמעת הפרויקט, וזאת בפרט במקרה של מודלי יסוד ו-LLM, שבהם תמחור המבוסס על צריכה עלול לגרום להוצאות שוטפות בהיקף משמעותי. ניתוח עלות-תועלת יכלול:

- **הגדרת היעדים** שהגוף המבוקר חותר להשיג באמצעות מערכת הבינה המלאכותית, ומדדי ביצוע עיקריים (KPI) הקשורים לאותם יעדים.
- **הערכה השוואתית** כנגד מערכות קיימות וחלופות ישימות, המביאה בחשבון ביצועים, עלויות והשפעה עסקית. הערכה זו צריכה לכלול ניתוח של רמת הדיוק של המודל ועלויות הקשורות לטעויות, הוצאות פיתוח ותחזוקה וחיסכון פוטנציאלי. כדי לקבוע שהמודל מפיק תוצאות מהימנות, יש ליישם שיטות בדיקה מעמיקות כמו השוואות נתונים היסטוריות, מחקרי פיילוט או בדיקות A/B.
- **תרגום ביצועים טכניים לערך רלוונטי בהיבט העסקי** (פיננסי, תפעולי, אסטרטגי).
- **זיהוי עלויות נסתרות ולאורך מחזור החיים**, כולל השפעה סביבתית.
- **ניתוח עלויות טוקנים/API** עבור מודלי יסוד, והשפעתן על כדאיות בטווח הארוך.
- **מדידת שיפורי פרודוקטיביות** וסימנים ל"בינה מלאכותית לשם עצמה".
- **תחזיות של עלות התחזוקה לאורך מחזור החיים**, כולל שיקולים של סוף החיים.
- **הערכת עלויות משפטיות, עלויות ציות ועלויות הזדמנות.**

4.1.1.1 סיכונים שיש להביא בחשבון

- אימוץ פתרונות בינה מלאכותית ללא ערך מוסף ברור או התאמה מובהקת.
- הערכה לא נאותה של גישות חלופיות, כמו שיטות אנליטיות או מבוססות כללים.
- הבנה טכנית מוגבלת בקרב אנשי ההנהלה, או מודעות בלתי מספקת למציאות התפעולית בקרב הצוותים הטכניים, שעלולות לגרום לתקשורת לא טובה ולציפיות בלתי מציאותיות.
- היעדר אימות של ישימות המערכת באמצעות הוכחת היתכנות.
- היעדר הגדרה של מטרות או מדדי ביצוע.
- חוסר התאמה לעקרונות הפנימיים של ממשל בינה מלאכותית.
- עלויות מחשוב מופרזות, כמו חומרה, אנרגייה או חיובי ענן, עקב שימוש במודלים מורכבים שלא לצורך.
- עלויות טוקן או API גבוהות עקב שימוש ב-LLM היכן שמודלים פשוטים יותר יספיקו.

4.1.1.2 בקרות מצופות

- מעורבות מוקדמת ושוטפת של בעלי עניין. לשם כך אפשר להיעזר במנגנונים כמו סדנאות בשיתוף גורמים שונים, אסיפות אזרחים או התייעצויות עם הציבור.
- הצהרות מוגדרות כראוי של בעיות והנחות מוסכמות.
- פיתוח משותף של מטרות פרויקט הבינה המלאכותית, ותוכנית פרויקט הכוללת את הבעלים של הפרויקט ובעלי עניין רלוונטיים, כדי להבטיח ישימות והבנה משותפת.
- יכולת לקשר באופן ברור בין המטרות לדרישות הטכניות והפונקציונליות (traceability), אשר מתועדות כראוי ומתורגמות לדרישות מודל המאפשרות מדידה וניטור.
- הגדרת אינדיקטורים למדידת הערך המוסף של מערכת הבינה המלאכותית מראש, המבטיחים יכולת להעריך את היתרונות לכל אורך מחזור החיים.
- תיעוד של ניתוח עלות-תועלת, ניתוח סיכונים והמלצות לגבי הטכנולוגיה.
- הוכחת היתכנות במקרה של פרויקטים גדולים.
- התייחסות להשפעה הסביבתית לפני פריסה.
- מיפוי פרויקט הבינה המלאכותית אל מול אסטרטגיית הבינה המלאכותית של המוסד ו/או הנחיותיו בנושא, במקרה הרלוונטי.

4.1.2 דרישות חוקיות ואתיות החלות בתחום השיפוט הספציפי

פרויקטים של בינה מלאכותית כרוכים לרוב בטווח רחב יותר של שיקולים משפטיים ואתיים לעומת מערכות IT מסורתיות, בפרט כאשר מדובר במערכות מורכבות, מערכות אוטונומיות או כאלה המשפיעות באופן ישיר על בני אדם. הרלוונטיות של תקנות ספציפיות תלויה בגורמים כמו שימוש בנתונים אישיים, רמת האוטומציה, המשתמשים המיועדים ומטרות המערכת. לדוגמה, מערכות בינה מלאכותית שייעודן רק לתמוך בתהליכים פנימיים, ואינן משפיעות על אזרחים ישירות, יצריכו בדיקת ביקורת מעמיקה פחות בסוגיות של אתיקה בהשוואה למערכות קבלת החלטות אוטומטית, שיש להן השפעה ישירה או עקיפה על אזרחים.

אומנם תקנות אסדרה וחוקים רבים אינם מתייחסים לטכנולוגיה ספציפית, אך תרגום עקרונות כמו מניעת אפליה ופרטיות לדרישות טכניות עלול להיות מאתגר. הסטנדרטים לפירוש מושגים משפטיים - כמו אפליה בהקשר של אלגוריתמים - ולתרגום שלהם לדרישות טכניות, נמצאים עדיין בתהליך של התפתחות[59]. על המבקרים לבדוק ולאשר שצוות הפרויקט כולל אנשים בעלי מומחיות משפטית מספקת, ושהקווים המנחים הפנימיים בארגון מתייחסים לאתיקה ולסיכונים המשפטיים. הערכת החוקים והתקנות הרלוונטיים תכלול הן תפעול רגיל, הן תופעות לוואי פוטנציאליות, לרבות כאלה הנובעות מפגמים או תקלות במודל.

אתגרים אלה מצטמצמים במידה מסוימת בזכות גיבושן ויישומן של תקנות אסדרה ייעודיות לבינה מלאכותית בתחומי שיפוט רבים (ראו חלק 3). ה-EU AI Act לדוגמה, בהיותו אסדרה של בטיחות המוצר בהגדרתו, מכיל יותר פרטים טכניים מחקיקה שמתמקדת בזכויות יסוד. כך הוא מקשר את דרישות החוק למפרטים טכניים בהתבסס על מאפייני מערכת הבינה המלאכותית. אף שהדבר יכול לסייע

59 סקירה מקיפה של ההשלכות הפרקטיות של עקרונות בינה מלאכותית, כולל המקרה של בינה מלאכותית יוצרת ותיאור פערי מחקר, ניתן למצוא כאן: A. Oesterling et al., *Operationalizing the Blueprint for an AI Bill of Rights: Recommendations for Practitioners, Researchers, and Policy Makers (2024).*

למפתחים להטמיע עקרונות אתיים בפועל, הוא גם דורש הערכה קפדנית של תחולת האסדרה. ייתכן שהמפתחים ינסו להימנע מחובות החוק על ידי ניצול מצבים של חוסר בהירות או פטורים בתחולת האסדרה - כמו כללי סיווג, רמות סיכון או קטגוריות של מפעילים (משתמשי המערכת).

כאשר ארגונים משתמשים בשירותים או מודלים חיצוניים, עליהם להעריך את ההשלכות במישור המשפטי והאתי, לרבות ציות לדינים מקומיים ואותנטיות של נתוני האימון. ייתכן מצב שבו בפיתוחם של מודלים חיצוניים הופרו תקנות מקומיות, או שהפיתוח עורר חששות אתיים כמו הפרות פוטנציאליות של זכויות קניין רוחני במהלך האימון של LLM. על הארגונים לשקול גם את הסיכונים לטווח ארוך של הסתמכותם על ספקים זרים או על מודלים שפותחו מחוץ לארגון. כמו כן תקנות האסדרה עשויות להגביל את מיקומם הפיזי של שירותים מתארחים, ומגבלות מסוג זה יכולות להשתנות על פני זמן - כפי שהודגם בשינויי החקיקה להעברות נתונים בין האיחוד האירופי לארה"ב בעקבות פסקי הדין בעניין (Schrems I, Schrems II)[60].

4.1.2.1 סיכונים שיש להביא בחשבון

- ידע בלתי מספיק או חוסר תשומת לב להוראות של חוקים וכללי אתיקה.
- הפרת דיני הגנת המידע או חוקים נגד אפליה.
- הערכה מאוחרת או חסרה של מסגרות אסדרה, שעלולה להוביל לעיכובים בפרויקט, לעלויות מוגדלות או לעצירת פרויקטים.
- עקיפת דרישות האסדרה באמצעות טקטיקות כמו המעטה בחשיבות פונקציונליות מערכת הבינה המלאכותית או הגבלתה באופן מלאכותי, תיוג שגוי של מערכת הבינה המלאכותית או סוג הפרויקט, או טכניקות דומות שמטרתן התחמקות מציות.
- חוסר מודעות לסיכונים משפטיים או אתיים הקשורים לפיתוח מודלים חיצוניים או למיקומם הפיזי של שירותים חיצוניים.
- אי-ודאות לגבי הישימות ארוכת הטווח של שימוש בשירותי ספקים זרים (בחו"ל).

4.1.2.2 בקרות מצופות

- צוותי פרויקט רב-תחומיים בעלי מומחיות משפטית, מומחיות טכנית ומומחיות בתחום העיסוק (domain experts).
- קווים מנחים פנימיים בדבר שימוש אתי בבינה מלאכותית ובדבר סיכונים משפטיים.
- הערכה מקיפה של החוקים והתקנות הרלוונטיים.
- ניתוח השלכות משפטיות ואתיות של שימוש בשירותים או מודלים חיצוניים.

4.1.3 ניהול מחזור החיים של בינה מלאכותית

ניהול מקיף של מחזור החיים יבטיח שפרויקטים של בינה מלאכותית יהיו מתוכננים כראוי וייתמכו משלב ההקמה ועד סיום חיי המערכת, ובכלל זה:

- הגדרה של המטרה, הגבולות והמגבלות של יישום הבינה המלאכותית.
- תכנון לקראת שלבי העיצוב, הפיתוח, אבטחת האיכות, הפריסה, התחזוקה, העדכונים והוצאת המערכת משירות פעיל בסיום חייה.
- הטמעת עקרונות אתיים - כמו הגנת מידע, שקיפות ואי-אפליה - בכל שלבי מחזור החיים ("Ethical AI by Design").
- עבור מערכות שנרכשות או מתקבלות דרך מיקור חוץ, יש להבטיח שניהול מחזור החיים יכסה את הרכש, ניהול החוזה ותיאום עם עקרונות ממשל מקומיים.

4.1.3.1 סיכונים שיש להביא בחשבון

- חוסר בהירות של התפקידים ותחומי האחריות בארגון המבוקר עבור כל שלבי הפרויקט, המוביל לעיכובים או לפגיעה בפונקציונליות.
- הגנה בלתי מספקת על עקרונות האתיקה, אם אלה אינם מוטבעים בתכנון עצמו לאורך מחזור החיים הכולל של המערכת.
- חוסר יכולת לתמוך במערכת הבינה המלאכותית לאורך כלל מחזור החיים שלה, כולל תחזוקה, עדכונים וניהול תקריות.
- שקיפות ושליטה מוגבלות כאשר הפיתוח, התשתית או רכיבים של המערכת מתנהלים במיקור חוץ, מצב המוביל לבעיות הבאות:

1. מידע שאין די בו כדי לקבל החלטות מושכלות או לבקר את המערכת;
2. תפקידים ותחומי אחריות לא ברורים, המגדילים את הסיכונים ופוגעים ביכולת להגיב לתקריות או לנתח כשלים;

60 Reuters: *EU seals new US data transfer pact, but challenge likely*.

3. בעיות של המשכיות ביצועים בסביבות המשתנות במהירות, כאשר המודלים הבסיסיים או הפלטפורמות מתעדכנים לעיתים קרובות ללא שליטה של הגוף המבוקר.

- רכישת מערכות או שימוש במערכות שפותחו מחוץ לארגון ללא תמיכה לטווח הארוך, בשילוב עם היעדר מומחיות פנימית בארגון, העלולים להוביל לגריעת ביצועים ולהגברת סיכוני האבטחה.

4.1.3.2 בקרות מצופות

- תוכנית מחזור חיים מתועדת, לרבות תפקידים[61], תחומי אחריות, צירי זמן ומשאבים.
- עקרונות אתיקה המובנים מראש בתכנון (by design).
- במקרה של מערכת נרכשת או כזו שחלק מהפיתוח שלה בוצע במיקור חוץ, יש לנקוט אמצעים כדי לוודא שהגוף המבוקר שומר על שליטה מספקת על השימוש בבינה מלאכותית, לרבות:

1. תיעוד מקיף המגדיר בבירור את התפקידים ותחומי האחריות - הן של הגוף המבוקר, הן של הספק החיצוני;
2. במקרים רלוונטיים, הסכמי רמת שירות (SLA) המבטיחים תמיכה שוטפת ומספקים בקרה נאותה על תשתיות חיצוניות או שירותים חיצוניים, כדי להבטיח ניהול תקריות וניתוח כשלים יעיל.

- תוכנית הבטחת איכות מוסכמת המתואמת עם מטרות הפרויקט, והדרישות הטכניות והפונקציונליות הקשורות בכך.

4.1.4 הערכת סיכונים

מערכות בינה מלאכותית מייצרות סיכונים אשר לא בהכרח מקבלים מענה במסגרות ניהול הסיכונים המקובלות. ככל שהטכנולוגיה הולכת ומתפתחת, כך מתפתחת גם ההבנה שלנו לגבי הסיכונים הטמונים בה. רבים מסיכונים אלה נובעים מהדרך שבה מערכות בינה מלאכותית לומדות מנתונים, וכן ממגבלות השקיפות והשליטה על הדפוסים שהן מזהות. לדוגמה, בינה מלאכותית עלולה לחזק הטיות קיימות שלא במתכוון, לחשוף מידע סודי המוטמע בנתוני האימון, או להיחשף למניפולציות באמצעות טכניקות הטעיה. סיכונים אלה גוברים כאשר ארגונים משתמשים במודלי יסוד שפותחו מחוץ לארגון, ושלא תמיד ברור מה היו נתוני האימון שלהם ומה הם אמצעי ההגנה.

כאשר מבצעים ביקורת כדי לבדוק את הניהול של בינה מלאכותית בגוף המבוקר, יש לבדוק אם מסגרת ניהול הסיכונים של הארגון עודכנה כדי לשקף חששות הנוגעים לבינה מלאכותית באופן ספציפי. ברמת הפרויקט, אופיים וחומרתם של סיכוני הבינה המלאכותית ישתנו בהתאם ליישום. עם התחומים העיקריים הראויים לתשומת לב המבקרים נמנים תחומי האבטחה, ההוגנות, השקיפות וההסברתיות - כל אחד מהם נבחן ביתר פירוט בתתי-הסעיפים הרלוונטיים של סעיף 4.4.

אחת השאלות המרכזיות שעל המבקרים לבדוק היא האם הארגון ערך ניתוח סיכונים המתייחס מפורשות לסיכוניה הייחודיים של בינה מלאכותית. ניתוח זה צריך להתייחס לא רק לסיכוני הפרויקט, אלא גם להשפעותיו הפוטנציאליות על פרטים ועל החברה.

4.1.4.1 סיכונים שיש להביא בחשבון

- הערכות סיכונים סטנדרטיות עלולות להתעלם מסיכונים ייחודיים של בינה מלאכותית.
- סיכונים תפעוליים, טכניים וחברתיים הנובעים מבינה מלאכותית לא תמיד מזוהים באופן מלא, לא מוסברים, או שלא ננקטים אמצעים להפחתתם.

4.1.3.2 בקרות מצופות

- במקרים הרלוונטיים, הערכת שיקולי אתיקה נוסף על סיכוני פרויקט מוכרים וסיכוני בטיחות.
- הכללת סיכונים הקשורים להסתמכות על תשתיות חיצוניות או מודלים חיצוניים.
- כאשר משתמשים במודלים שפותחו מחוץ לארגון או בקוד פתוח, הבאה בחשבון של סיכוני בטיחות הנובעים מגורמים זדוניים, כגון החדרת דלתות אחוריות (backdoors) או בעיות בהתאמת המודל למטרותיו.
- התייחסות למסגרות מבוססות לניהול סיכוני בינה

61 ראו בנספח 5 רשימה של תפקידים אפשריים.

מלאכותית[62] וקווים מנחים להערכות הנדרשות על פי חוק[63], והסתייעות בהם כדי לזהות סיכוני בינה מלאכותית ייעודיים ולטפל בהם, גם כאשר הדבר אינו מתחייב מהוראות החוק.

- פיתוח מודיעין אסטרטגי כדי לצפות מראש שיבושים ולנהל סיכונים מתפתחים, באמצעות כלים כמו סריקת אופק וניתוח תרחישים[64].
- בחינה והשוואה של הסיכונים הכרוכים בגישות חלופיות בשלב של הצעת הפרויקט ובחירת המודל (ראו סעיפים 4.1.1 ו-4.3).
- מיפוי שיטתי של נזקים פוטנציאליים כאשר מדובר בתחומי פעולה קריטיים (כמו שירותי בריאות או אכיפת החוק), ושימוש בסולמות מדידה מתוקננים להערכת חומרת ההשפעות, לרבות אלה הקשורות למידע מוטעה, אבטחה, אפליה והפרעה חברתית-כלכלית.
- סקירת תיעוד לצורך הערכת סיכונים:

1. סקירת התיעוד הטכני, כולל שפות קוד, גרסאות של פלטפורמות, תוצרי המודל והיסטוריה של הפרומפטים למודל והגדרות יצירת התוכן עבור בינה מלאכותית יוצרת.
2. בדיקת הציות לדיני הפרטיות, להנחיות בנושא אתיקה ולמדיניות הפנימית לבינה מלאכותית.
3. הערכת בקרות איכות, לרבות תיעוד תהליכי MLOps/LLMOps, פעילויות צוות אדום (המחפש נקודות תורפה) וניטור לוחות מדדים (דשבורד) (ראו סעיף 4.3).
4. וידוא קיומו של תיעוד מפורש לכל הנחות היסוד שעליהן התבססו הערכות המערכת.

4.2 נתונים

האפקטיביות של מערכות בינה מלאכותית תלויה בנתונים שעליהן הן בנויות. אף שממדי איכות נתונים מקובלים - כמו שלמות, תקפות וייחודיות - עדיין חשובים, הבינה המלאכותית מייצרת אתגרים נוספים. עם אלה נמנים האיכות, הכמות והייצוגיות של נתוני האימון, וכן הסיכונים הקשורים להטיה, התאמת יתר (overfitting - ראו להלן) והצורך בניטור ותחזוקה שוטפים. פרק זה מתאר את הסיכונים העיקריים הקשורים בנתונים ואת הבקרות הרלוונטיות שיש לבחון במסגרת ביקורת מערכת בינה מלאכותית.

4.2.1 כמות הנתונים ואיכותם

מהימנותה של כל מערכת בינה מלאכותית תלויה באיכות נתוני האימון שלה ובהיותם מספיקים (דיות נתונים). נתונים בלתי הולמים או מוטים עלולים להוביל למודל שמייצר תוצאות בלתי הוגנות ולא מדויקות. לדוגמה, אם נתוני האימון אינם משקפים את אוכלוסיית העולם האמיתי, המודל עלול להעדיף באופן שיטתי קבוצות מסוימות על פני אחרות.

בעיות שכיחות אחרות באיכות הנתונים במערכות בינה מלאכותית הן תוצר לא מדויק או לא שלם - טעויות, השמטות או מידע שעבר זמנו עלולים לערער את ביצועי המודל. חוסר עקביות עלול להיווצר כאשר מצרפים נתונים ממקורות שונים, דבר העלול להוביל לאי-התאמות של הפורמט או המשמעות.

במקרה של מערכות בינה מלאכותית שפותחו על ידי גוף חיצוני או נרכשו ממנו, ייתכן שלארגון תהיה גישה מוגבלת לנתוני האימון, והדבר יקשה להעריך את התאמתם להקשר שבו פועל הארגון.

נתונים בלתי מספיקים מייצגים אתגר משמעותי בפיתוח מערכות של בינה מלאכותית. כמות הנתונים הנדרשת גדלה יחד עם מורכבות הבעיה שהמערכת מיועדת לטפל בה. כאשר הנתונים הזמינים מוגבלים מדי יחסית למורכבות המודל, קיים סיכון של התאמת יתר (overfitting). זהו מצב שבו המודל לומד את נתוני האימון - כולל הרעש שבהם - באופן מדויק מדי, ואינו מצליח להפיק מהם הכללות לגבי מצבים חדשים. במצב ההפוך, אם המודל פשוט מדי למשימה, הוא עלול להימצא בהתאמת חסר (underfitting). אז הוא יחמיץ דפוסים חשובים או לא יצליח לתפוס את המבנה שמאחורי הנתונים.

בלמידה מונחית (supervised learning) דרושים נתונים מספיקים כדי לבצע אימון, אימות ובדיקות. נתונים בלתי מספיקים בכל אחד משלבים אלה, או שימוש באותם נתונים לשלבי האימון והבדיקות גם יחד, יובילו למדדי ביצוע

62 לדוגמה: National Institute of Standards and Technology. (2023), Artificial Intelligence Risk Management Framework (AI RMF) או the IBM AI Risk Atlas: F. Bagehorn et al., AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources (2025).

63 לדוגמה, קיימים קווים מנחים המתארים מתי וכיצד לבצע הערכת השפעות על הגנת המידע (כנדרש על פי GDPR באיחוד האירופי), ברמה בין-לאומית וברמה לאומית. אנו מצפים כי יתפרסמו קווים מנחים דומים לגבי הערכת השפעות על זכויות יסוד, כנדרש על פי ה-EU AI Act במקרים ספציפיים.

64 OECD (2024), "Framework for Anticipatory Governance of Emerging Technologies", OECD Science, Technology and Industry Policy Papers, No. 165, OECD Publishing, Paris.

מנופחים. טכניקות כמו הגדלת מספר הדגימות מקטגוריות פחות שכיחות (upsampling) או הקטנת מספר הדגימות מקטגוריות נפוצות יותר (downsampling) יכולות לסייע לשמור על איזון וייצוגיות בנתונים. סיכון נוסף מכונה דליפת יעד (target leakage): מצב שבו נתוני האימון כוללים בשוגג מידע שלא יהיה זמין בזמן החיזוי בפועל. הדבר עלול להוביל לביצועים שנראים מרשימים במיוחד בשלב הפיתוח, אך לתוצאות חלשות בתנאי אמת. דליפת יעד נובעת ממשתנים הכוללים בטעות מידע המרמז על התוצאה הצפויה.

מודלים של בינה מלאכותית פגיעים גם לסיכונים כמו הרעלת נתונים, מצב שבו גורמים זדוניים מחדירים נתונים פגומים למאגר נתוני האימון כדי לערער את תקינות המודל ואמינותו.

גם תשתית הנתונים חשובה. מבני סילו של נתונים (data silos) - מצב של פיצול מידע בין מערכות שונות - יכולים לפגום ביכולת לבצע שילוב וניתוח של נתונים. ככל שנפח הנתונים גדל, נדרשת תשתית ניתנת להרחבה לצורך תמיכה בפיתוח ובתפעול השוטף. ההתייחסות לשיקולי אבטחה ופרטיות חיונית, במיוחד כשמטפלים במידע רגיש או אישי.

4.2.1.1 סיכונים שיש להביא בחשבון

- שימוש בנתונים בלתי אמינים, מוטים או בלתי מייצגים.
- הכנסת הטיה במהלך עיבוד הנתונים.
- הפרדה בלתי מספקת בין נתוני האימון לנתוני הבדיקה.
- יכולת הכללה גרועה של נתונים חדשים.
- הרעלת נתונים או מניפולציה לצורך הטעיה.
- דליפת יעד.
- היעדר שקיפות לגבי ייצוגיות נתוני האימון במודלים המפותחים מחוץ לארגון.

4.2.1.2 בקרות מצופות

- הערכת האותנטיות של מאגרי הנתונים ששימשו לאימון וכוונן עדין שלהם, בפרט כשמשתמשים בנתונים מהאינטרנט או נתונים סינתטיים, כדי להבטיח ציות לדרישות קניין רוחני והגנת המידע.
- וידוא שפיצולי הנתונים לצורכי אימון, תיקוף ובדיקה שומרים על ייצוגיות ומטופלים כראוי.
- השוואה של איכות הנתונים ואיכות המודל בכל מערכות הנתונים - לצורכי אימון, בדיקה ותיקוף.
- הגדרת המאפיינים העיקריים של האוכלוסייה ובדיקה אם הם מיוצגים כראוי בנתוני האימון, הבדיקה והתיקוף.
- סקירה ועדכון סדורים של נתוני האימון השמורים כדי למנוע התיישנות שלהם.
- עבור מערכות שפותחו על ידי גוף חיצוני או נרכשו ממנו - השגת מידע מספיק לגבי נתוני האימון מהגורם המפתח, אימות ביצועי המערכת עם נתונים מקומיים ובדיקה לגילוי הטיות פוטנציאליות.

4.2.2 מודלי שפה גדולים (LLM) ויצירה מבוססת אחזור (RAG)

יצירה מבוססת אחזור (retrieval-augmented generation - RAG) היא גישה המאפשרת למודלי שפה גדולים לשאוב ממקורות מידע חיצוניים בזמן שהם יוצרים תשובות. הזנת מסמכים או נתונים רלוונטיים לתוך ההקשר של המודל מאפשרת ל-RAG להתמודד עם אתגרים ידועים בבינה מלאכותית יוצרת, כמו הזיות, מידע מיושן והיעדר ציטוטי מקורות. עם זאת טכניקה זו מכניסה סיכונים חדשים הנוגעים לאיכות הנתונים המאוחזרים ואמינותם, וכן בקשר לבנייה ותחזוקה של בסיס הידע של המערכת[65].

מידע על הנתונים ששימשו לאימון מודלי יסוד שפיתחו ארגונים חיצוניים בדרך כלל אינו זמין. נתונים הנאספים אוטומטית ממקורות פתוחים (לדוגמה, מהאינטרנט) עלולים להיות פגיעים להרעלת נתונים. אם נתונים מסוג זה משמשים לאימון, לכוונון עדין או במערכות בינה מלאכותית שעלו לאוויר (לדוגמה, סוכני בינה מלאכותית המשתמשים בחיפוש באינטרנט), על המבקרים לצפות לראות ראיות לכך שהסיכון להרעלת נתונים הובא בחשבון ונוהל. סיכון זה רלוונטי גם ל-RAG, היכן שבסיס הידע נבנה על ידי שימוש בנתונים הנאספים באופן אוטומטי.

לשם בדיקה והערכה של פלטי LLM - כולל אלה המופקים באמצעות RAG - נדרשות לעיתים קרובות כמויות גדולות של נתוני דוגמאות המשקפים את מגוון שאילתות המשתמש האפשריות ואת התשובות הצפויות. בפועל, איסוף ידני של מאגרי נתונים כאלה כמעט אף פעם אינו ישים, בפרט ביישומים מורכבים או מתמחים. גישה חלופית היא לייצר נתונים סינתטיים - הנוצרים באופן מלאכותי במקום להיאסף מאינטראקציות בעולם האמיתי - ולהשתמש בהם כדי ליצור מערכי בדיקה מגוונים וסקלביליים לצורכי אימון, כוונון עדין או אימות של המודל. למרות יתרונותיהם, נתונים סינתטיים מצריכים אימות קפדני כדי לוודא שהם

65 פרטים טכניים נוספים על ארכיטקטורת RAG והערכת תוצרי RAG מובאים בסעיפים 4.3 ו-4.4.

לא מכניסים הטיות חדשות או פוגמים בביצועי המודל. איכותם של מאגרי נתונים סינתטיים תלויה במידה שבה הם משקפים את מקרי השימוש המיועדים ואת גיוון התרחישים בעולם האמיתי.

4.2.2.1 סיכונים שיש להביא בחשבון

- נתונים מורעלים הנכללים במודלים שפותחו מחוץ לארגון, בבסיס הידע של מערכת RAG או במועד השאילתה.
- הערכה בלתי אמינה של איכות תוצרי המודל.
- הגברת הטיות או בעיות ביצוע אחרות כאשר משתמשים בנתוני בדיקה סינתטיים שנוצרו על ידי LLM.

4.2.2.2 בקרות מצופות

- שימוש במודלים חיצוניים ממקורות מהימנים, עם תיאור של נתוני האימון היכן שזה אפשרי.
- אימות איכות הנתונים בנתוני קלט שנאספו אוטומטית.
- אימות נתונים סינתטיים כדי למנוע חיזוק הטיות קיימות או הכנסת הטיות חדשות, או פגיעה בביצועי המודל.
- אימות אנושי של נתונים סינתטיים שנוצרו על ידי LLM, אם הדבר ישים.

4.2.3 סוגיות של פרטיות בהקשר של מערכות בינה מלאכותית

השימוש בנתונים אישיים בבינה מלאכותית כפוף לדרישות קפדניות המעוגנות בחוק, לרבות הגבלות על המטרה, מזעור נתונים, מידתיות ושקיפות. בסעיף 3.3.3 אנו דנים בדרישות החלות בתחומי שיפוט ספציפיים.

האחריות לציית לחוק מוטלת על הארגון המפתח את מערכת הבינה המלאכותית או על הארגון שמטמיע אותה (המשתמש). המבקרים נדרשים לבחון תיעוד כגון הערכת השפעות על הגנת המידע (DPIA), כדי לוודא שאכן מקיימים את חובות הגנת המידע. יש להקדיש תשומת לב לכל סוגי הנתונים ששימשו במהלך הפיתוח, לא רק לתוצרי המודל הסופי. לדוגמה, מאפיינים (features) שנבדקו אך בסופו של דבר לא הוכנסו לשימוש עדיין יכולים לכלול עיבוד של נתונים אישיים.

במקום שמשתמשים בו בנתונים אישיים, יש להעריך את חשיבותם לביצועי המודל. כאשר מבקרים וארגונים קובעים את מידת תרומתם של הנתונים האישיים לרמת הדיוק או לאפקטיביות של מערכת הבינה המלאכותית, הם יכולים להעריך אם השימוש באותם נתונים הכרחי ומידתי ביחס למטרה המיועדת. במקרה שמתעורר ספק או חשד לאי-ציות להוראות החוק, יש להפנות את המקרה לרשויות הגנת המידע המתאימות.

4.2.3.1 סיכונים שיש להביא בחשבון

- הפרות של תקנות הגנת המידע (ראו נספח 2).

4.2.3.2 בקרות מצופות

- הערכת הסיכונים הנוגעים לנתונים אישיים, המסייעת לקבוע אם נדרש DPIA.
- DPIA, במקרה הרלוונטי.

4.3 פיתוח המודל

פיתוח המודל הוא שלב מכריע במחזור החיים של מערכת בינה מלאכותית, שבו העיצובים המושגיים מתורגמים למודל תפעולי. עבור המבקרים, שלב זה מספק את התמונה הברורה ביותר של אופן הטמעתן של דרישות טכניות, ארגוניות ואתיות במערכת.

סעיף זה מסביר את התהליך של בניית מודל והבטחת איכותו בשלב הפיתוח, וכולל את הנושאים של בחירת מודל והתאמתו (כולל שימוש במודלי יסוד ו-LLM), הנדסת מאפיינים, תכנון פרומפטים ופרמטרים, הבטחת איכות פנימית ושיטות להבטחת אמינות, עמידות והדירות. אנו מתייחסים לבדיקות קבלה (acceptance testing) מבוססות תוצרים ולהחלטות אור ירוק/אדום בסעיף 4.5, וסעיף 4.6 עוסק בבדיקות תפעוליות וסביבות חיות.

4.3.1 תהליך הפיתוח ומאפייני הביצוע

פיתוח של מודל בינה מלאכותית כרוך בבנייה, אימון ושכלול של המודל לצורך ביצוע משימות ספציפיות. תהליך זה כולל על פי רוב הכנת נתונים (ראו סעיף 4.2), בחירה או התאמה של אלגוריתמים, הערכת ביצועים ואופטימיזציה של הפרמטרים כדי לוודא שהמודל יספק ביצועים מדויקים כאשר יוזנו אליו נתונים חדשים. כיום הנטייה בארגונים היא להתאים מודלים שאומנו מראש או מודלי יסוד, כמו LLM, במקום לפתח מודלים מאפס. על המבקרים להעריך אפוא את מקור המודל, הרישוי, שיטות ההתאמה (כמו כוונון עדין, הנדסת פרומפטים או RAG) וכל קשרי התלות ההדדית עם ספקים חיצוניים.

בחירת המודל הנכון היא קריטית, מכיוון שהיא משפיעה ישירות על הביצועים, הסקלביליות והיכולת להכליל למצבים חדשים. הגורמים העיקריים המשפיעים על החלטה זו - ובכלל זה, האם לפתח מודל ייעודי לארגון או לבצע התאמות במודל שאומן מראש - הם אלה:

- זמינות נתונים: כמות הנתונים, איכותם והרלוונטיות שלהם.
- מומחיות טכנית: מיומנויות הארגון בלמידת מכונה ובתחום העיסוק (domain experts).
- משאבי מחשוב: היכולת לאמן ולהריץ מודלים.

- אילוצי זמן: לו"ז הפרויקט ודרישות המסירה.
- צורכי הסברתיות: דרישות לגבי שקיפות והיכולת לפרש את תוצאות המודל (interpretability).
- תקציב: משאבים כספיים לצורכי הפיתוח והתפעול השוטף.

לדוגמה, ארגונים עשויים להשוות בין הביצועים של אלה: (1) שיטות פשוטות יותר של למידת מכונה; (2) למידה עמוקה שתעבור תהליך כוונון עדין או מודלי שפה קטנים; (3) LLM הנגיש דרך API. בחלק מן המקרים, שיטות פשוטות יותר של למידת מכונה או מודלים קטנים יותר שעוברים כוונון עדין עשויים לספק ביצועים טובים יותר ממודלים מורכבים או יקרים המיועדים למטרות כלליות. על המבקרים לצפות מהארגון לתעד את הנימוקים לבחירת המודל, ולוודא שבחירתו נאותה מן הבחינה הטכנית, וכי היא מוצדקת ושקופה.

הנדסת מאפיינים - התהליך של הפיכת נתונים גולמיים למשתנים בעלי משמעות עבור המודל - חשובה במיוחד. השיטות ששימשו ליצירה, לחילוץ ולבחירה של מאפיינים צריכות להיות מתועדות כהלכה. על המבקרים לשים לב לאופן שבו מאפיינים ספציפיים משפיעים על ביצועי המודל, והאם יש מאפיינים העלולים להכניס מורכבות מיותרת או סיכוני הטיה.

במקרה של בינה מלאכותית יוצרת ומודלי יסוד, התהליכים של עיצוב פרומפטים וניהול גרסאות הם קריטיים, מכיוון שההוראות או ההקשר שנותנים למערכת יכולים להשפיע על התוצרים באופן משמעותי. נוסף על הנדסת פרומפטים, הגדרת הפרמטרים לדגימה ממלאת תפקיד מרכזי בעיצוב התנהגות המודל ובעקביות תשובותיו[66]. תיעוד הגדרות אלה חיוני להדירות, לשקיפות ולאחריותיות.

הערכת גישות של למידת מכונה מונחית מתבצעת באופן טיפוסי על ידי מדידת ביצועים על נתוני הבדיקה. במקרה של מערכות בינה מלאכותית יוצרת ומודלים למטרות כלליות, יש צורך להעריך מדדים אחרים. אחד האתגרים העיקריים בפיתוח ביקורת עבור מערכות אלה הוא הסיכון של הזיות (hallucinations). הזיות מתרחשות כאשר המודל מפיק תוצרים שנשמעים סבירים אך הם שגויים עובדתית, מטעים או חסרי פשר. סיכון זה גבוה במיוחד ב-LLM, שיכול לייצר תשובות נחרצות שאינן מעוגנות בנתונים אמינים. חיוני לאמת את התוצרים כנגד מקורות מהימנים כדי לגלות טעויות ולמנוע מידע שגוי - הדבר יסייע לשמר תחושת ביטחון במערכת. עם המדדים האפשריים להערכת מערכות בינה מלאכותית מודרניות נמנים אלה:

- שיעור ההזיות ורמת הדיוק העובדתי.
- עמידות לפרומפטים או קלטים שונים.
- מדדים של בטיחות ורעילות.
- התאמה לתחומי ידע ספציפיים.
- יעילות הטיפול בהקשר ובנתונים.

אין מדד יחיד שיכול לשקף את איכותו הכוללת של מודל, וההערכה דורשת על פי רוב לאזן בין תחומי עדיפות מתחרים, כמו דיוק לעומת הוגנות. על המבקרים לצפות מהארגון להגדיר ולנמק את מדדי הביצוע שבחר.

הדירות (reproducibility) - היכולת לשחזר תוצאות כשמשתמשים באותם נתונים, פרמטרים וסביבה - היא עדיין אבן יסוד בפיתוח מערכת בינה מלאכותית אמינה. מבקרים מתחילים בדרך כלל בסקירת תיעוד, כדי לבסס קו התחלתי לביקורת. אם התיעוד אינו מספק, אינו ברור או מכיל סתירות, ייתכן צורך בשחזור המודל ו/או הניבויים שלו. כאשר שחזור מלא אינו אפשרי (כגון במודלים גדולים או קנייניים), חשוב לבחון תיעוד מעמיק של פרומפטים, הקשרים לאחזור והגדרות תצורה כדי לתמוך בביקורת.

4.3.1.1 סיכונים שיש להביא בחשבון

- טעויות בהכנת הנתונים או בהנדסת המאפיינים היוצרות הטיות או פוגמות באמינות המודל (ראו גם סעיף 4.2, נתונים).
- הערכת יתר של ביצועי המודל, בפרט אם הבדיקות אינן נאותות או אינן מייצגות.
- חוסר יציבות והעדר עקביות של התוצרים, כולל רגישות לשינויי פרומפטים או לשינויים בהקשר במערכות יוצרות.
- חוסר יכולת לאמת טענות לגבי יכולות של מודל יסוד, במקרים רבים, כאשר לרשות המבקרים עומדים משאבים מוגבלים ואין להם גישה למשקלות המודל.

4.3.1.2 בקרות מצופות

תיעוד מועיל ובקרת גרסאות הם מרכיבים חיוניים לשקיפות, לעקיבות וליכולת לבצע את הביקורת. על המבקרים לצפות למצוא את אלה:

- רשומות ברורות לגבי בחירת המודל, התאמתו לצורכי הארגון והיסטוריית גרסאות.
- הערכות ביצועים המשתמשות במדדי השוואה כלליים ובבדיקות ייעודיות ליישום.

66 פרמטרים כמו "טמפרטורה" או "top-k" שולטים בגיוון ובאקראיות של תשובות המודל. לדוגמה, טמפרטורה גבוהה יותר מגבירה את השונות, ואילו ערכים נמוכים יפיקו תוצאות עקביות יותר.


AI

- ראיות להדירות, כולל נתונים, פרמטרים ופרטים סביבתיים.
- הערכות סיכונים משולבות שבחנו השפעות תפעוליות וחברתיות.
- במקרה של בינה מלאכותית יוצרת ומודלי יסוד, על התיעוד לכלול עיצוב פרומפטים וניהול גרסאות, הגדרות של פרמטרים, ופרטים על APIs חיצוניים או מודלים מאומנים מראש שהארגון משתמש בהם.
- בחירת המודל: תיעוד ברור של הקריטריונים והבדיקות ששימשו להשוואת המודל, או נימוקים לכך שלא בוצעה השוואה.
- בקרת גרסאות: ניהול גרסאות קפדני עבור הנתונים, תוצרי המודל, פרומפטים, תוצרי ניבוי והקוד שמחבר בין תהליכי המערכת (pipeline code). יש לכלול פרטים על הנדסת פרומפטים, ניהול גרסאות והגדרות דגימה.
- רשומות מקיפות: מידע על ארכיטקטורת המודל, נתוני האימון, מגבלות ידועות, מדדי ביצועים, תהליכי פיתוח והחלטות פריסה.
- בינה מלאכותית יוצרת ומודלי יסוד: תיעוד השיטות המשמשות לגילוי הזיות ולצמצום התופעה, ובדיקת המידתיות של שיטות אלה אל מול הסיכונים המזוהים עם השימוש המיועד במערכת.
- הנדסת מאפיינים: תיאור האופן שבו משתנים נוצרים ונבחרים, בדגש על מאפיינים העשויים לפעול כתחליפים עקיפים לתכונות רגישות, וזאת כדי להבטיח עמידה בסטנדרטים של הגנת המידע ולמנוע הטיות.

4.3.2 הבטחת איכות, מהימנות ועמידות בפיתוח

הבטחת איכות (quality assurance) בפיתוח מתייחסת לאיכות הקוד, איכות המודל וגישות מומלצות להבטחת מהימנות ועמידות לפני פריסת המערכת.

- **מהימנות (reliability)** מתייחסת ליכולת של המודל לספק ביצועים עקביים ונכונים בתנאים הצפויים, בפרט בעת טיפול בנתונים הדומים לאלה ששימשו לאימון המודל.
- **עמידות (robustness)** היא היכולת של המודל לשמור על רמת ביצועיו כאשר הוא עומד בפני קלט בלתי צפוי - כמו נתונים בהתפלגות שונה מהנתונים שעליהם הוא אומן; נתונים "מחוץ להתפלגות" (out of distribution - OOD); רעש; או שינויים בסביבה.

שיקולים אלה רלוונטיים במיוחד למודלי יסוד ולמודלים יוצרים, שבהם ההבחנה בין נתוני האימון לנתונים התפעוליים לעיתים מטושטשת. מודלים מודרניים לראייה ממוחשבת ומודלי שפה רבים נוצרים על ידי כוונון עדין של מודל יסוד שאומן במקור על מאגר נתונים נפרד, גדול הרבה יותר.

התוצאה היא שהנתונים המשמשים לכוונון עדין הם רק חלק מן המכלול שמעצב את המודל הסופי, מה שמקשה לקבוע במדויק אילו נתונים הם "מחוץ להתפלגות". ראיות בשטח מעלות כי מודלים שהותאמו מתוך מודלי יסוד נוטים לבצע הכללות טובות יותר לנתונים חדשים, לעומת אלה שאומנו מאפס באמצעות למידה מונחית. עם זאת הסיבות להכללה משופרת זו עדיין נחקרות.

טכניקות כמו RAG יכולות לשפר מהימנות על ידי הוספת מידע ממקורות חיצוניים לתוצרי המודל, דבר המסייע להתמודד עם בעיות כמו הזיות או ידע לא עדכני. פיתוחים בארכיטקטורת מודלים, כמו חלונות הקשר מורחבים המאפשרים למודל לשקול כמות רבה יותר של מידע סביבתי כשהוא יוצר תשובות, תומכים בתוצרים מהימנים יותר, מדויקים יותר ושהולמים יותר את ההקשר.

על המבקרים לצפות לראות ראיות לכך שהמהימנות והעמידות הובאו בחשבון מהשלב הראשון, בליווי בקרות נאותות לניהול הסיכונים.

4.3.2.1 סיכונים שיש להביא בחשבון

- איכות קוד נמוכה או היעדר בדיקות, המובילים לתוצרים בלתי יציבים או בלתי אמינים.
- פגיעויות הנובעות מקשרי תלות חיצוניים, כמו עדכוני ספקים או שינויים במודל של צד שלישי.
- סיכוני מידע שגוי, כולל הזיות או הנצחת הטיות.
- נתונים שעברו שינויים מינימליים המובילים להתנהגות מודל מוטעית או בלתי אמינה.

4.3.3.2 בקרות מצופות

- עקביות התוצרים: בדיקת עקביות התוצרים על ידי הרצת אותו פרומפט כמה פעמים, במיוחד במודלים יוצרים. הערכת הדירות באמצעות תתי-קבצים של נתוני אימון או נתונים עצמאיים (שיכולים להיות סינתטיים). אם עולה חשד להתאמת יתר או למאפיינים מיותרים, יש לאמן מחדש את המודל עם פחות מאפיינים ולהשוות את התוצאות.
- רגישות לפרטים אישיים: במודלים המשתמשים בנתונים אישיים, יש לבצע אימון מחדש עם כמות מצומצמת של נתונים אישיים או ללא נתונים אישיים כלל, כדי להעריך את שקלול התמורות בין ביצועים לפרטיות.

- מעקב פרומפטים והקשר: תיעוד וריאציות בפרומפטים, בהקשר וכן בהקשרי RAG כדי לתמוך בשחזור ממוקד, כאשר שכפול מלא של המודל אינו אפשרי.
- ניתוח מתאמים חיצוניים: אם הקוד או המודל אינם נגישים, יש להשתמש בנתונים הזמינים כדי לנתח מתאמים בין מאפיינים וניבויים של המודל, ומכך להסיק על התנהגותו. שיטות אפשריות כוללות LIME או SHAP.
- הדגמת וריאנטים: יש לבקש מהגוף המבוקר להדגים אימון מחדש או זרימות עבודה פשוטות, כדי להבין את התנהגות המודל בתנאים שונים.
- יציבות של תפעול למידת מכונה (machine learning operations - MLOps), כולל בקרת גרסאות, בדיקות יחידה, סקירות קוד, בדיקות אינטגרציה ובדיקות מקצה לקצה.

4.4 הערכת המודל לפני פריסה

לפני פריסה של מערכת בינה מלאכותית, יש לבצע הערכה כדי לוודא שהמערכת עומדת בסטנדרטים שנקבעו בשלב הצעת הפרויקט ובמהלך פיתוח המודל. שלב זה קריטי כדי לוודא שהמערכת מתאימה למטרתה, פועלת כמצופה ועונה על דרישות החוק וכללי האתיקה. ההערכה צריכה להיות מידתית לסיכונים ולהקשר השימוש, ויש לבצעה באופן חוזר במקרה של שינוי במערכת או בסביבתה.

שימוש בבינה מלאכותית ככלי עזר בתהליכי ההערכה ילך וייעשה נפוץ יותר ככל שמערכות בינה מלאכותית יהיו גדולות יותר ומורכבות יותר, כך שבדיקות והערכות ידניות ייעשו מאתגרות יותר. אחת הדוגמאות היא "LLM כשופט" (LLM-as-a-judge): שימוש ב-LLM להערכה ודירוג של תוצרי מערכות בינה מלאכותית.

על המבקרים לוודא כי כל שימוש נוסף בבינה מלאכותית הוערך באופן מספק על ידי הבעלים של הפרויקט.

פרק זה עוסק בנושאים הבאים:

- בדיקות ביצועים וקבלה
- שקיפות והסברתיות
- הוגנות ומזעור פגיעה
- אבטחה
- השפעה סביבתית

תהליך ההערכה מיועד לוודא שמערכת הבינה המלאכותית עומדת בסטנדרטים שנקבעו לגבי ממשל (סעיף 4.1) ופיתוח המודל (סעיף 4.3), וכי היא מתאימה בפריסה לסביבת הייצור. סעיף 4.5 עוסק בעלייה לאוויר של המערכת ותוכנית שינויים לאחר שתהליך ההערכה הושלם בהצלחה. סעיף 4.6 עוסק בניטור שוטף ואימון מחדש לאחר פריסה. אם הניטור לאחר פריסה חושף בעיות של הוגנות, אבטחה או רמת דיוק, יש לחזור שנית על בדיקות ההערכה הרלוונטיות המתוארות כאן.

4.4.1 בדיקות ביצועים וקבלה

מערכות בינה מלאכותית צריכות להיבדק בתנאי מציאות טרם פריסתן, לרבות תיקוף טכני של התוצרים ובדיקות משתמשים של זרימות עבודה המשקפות את השימוש הצפוי במערכת בפועל. הבדיקות צריכות להתבסס על מדדים עסקיים וקריטריוני קבלה (acceptance) שהוגדרו בתחילת הפרויקט.

נקודות תורפה פוטנציאליות ופעולות מתקנות יתועדו. תוצאות בדיקות אלה, לצד ראיות לעמידה בקריטריוני הקבלה, יהוו את הבסיס להחלטה אם לתת אור ירוק לפריסה (ראו סעיף 4.5). האמור חל הן על פיתוח מערכות בארגון, הן על מערכות הנרכשות מבחוץ.

4.4.1.1 סיכונים שיש להביא בחשבון

- מערכת הבינה המלאכותית אינה עונה על הדרישות העסקיות.
- הביצועים פחות טובים כאשר משתמשים בנתונים שטרם נראו.

4.4.1.2 בקרות מצופות

- קריטריוני הקבלה ברורים ומוסכמים מראש.
- המערכת נבדקת עם נתונים מייצגים.
- משתמשי קצה מעורבים בבדיקות הקבלה.
- תוצאות הבדיקות וההחלטות מתועדות.

4.4.2 שקיפות והסברתיות

שקיפות והסברתיות חיוניות לבניית אמון במערכות בינה מלאכותית ולהבטחת אחריותיות. שקיפות משמעה מתן מידע ברור על השימוש במערכת - איך, מתי ומדוע משתמשים בה. הסברתיות היא היכולת לתאר כיצד המערכת מגיעה לתוצרים או להחלטות.

הערכת הסברתיות עלולה להיות מאתגרת, במיוחד במקרה של מערכות מורכבות או מצד שלישי. מודלים רבים של בינה מלאכותית פועלים כ"קופסאות שחורות", כך שלא קל להבין את ההיגיון הפנימי שלהן. מערכות של בינה מלאכותית יוצרת מתבססות לעיתים קרובות על הסתברות, כך שהן מקשות על שחזור פלטים ספציפיים או על מעקב שיראה כיצד נוצרה תשובה ספציפית. סעיף 4.3 מסביר כיצד ניתן לנתח מתאמים בין מאפיינים וניבויים כדי להסיק מהם על התנהגות המודל.

רמת ההסברתיות הנדרשת תלויה במטרת המערכת ובקהל היעד של ההסבר (כמו רגולטור או אדם מהציבור). ההסברתיות אמורה להיות מידתית לרמת הסיכון שמייצגת המערכת. חשוב למצוא את האיזון בין המשאבים הנדרשים

לבין רמת ההסבר ההכרחית.

כאשר מערכות נרכשות מגורם חיצוני, על הארגון לבקש מידע מהספק על קלטים במעלה הזרם - כמו ארכיטקטורת המודל, נתוני אימון, תהליכי כוונון עדין, מגבלות ידועות ומדדי ביצועים. כדי לסכם מידע זה, משתמשים על פי רוב בכרטיסי מודל (model cards) וגיליונות נתונים. מודלי יסוד המבוססים על קוד פתוח עשויים לספק גישה לקוד, למשקלות המודל ולמדריכים לשימוש אחראי.

4.4.2.1 סיכונים שיש להביא בחשבון

- חוסר יכולת להסביר או לנמק החלטות, המגבילה את יכולתו של הארגון לשפוט את תחזיות המודל ואת אמינותו.
- אי-ציות לתקנות הדורשות הליכים שקופים וביסוס החלטות (דרישות שכיחות במינהל הציבורי).
- מידע בלתי מספק עבור משתמשי הקצה, כך שאינם יכולים לערער על התוצאות.
- הבנה מוגבלת של הנתונים, המודל והניבויים שעלולה להוביל לאי-גילוי של הטיה ואפליה.
- "Fairwashing" - מניפולציה בטכניקות הסבר בדיעבד כדי להצדיק תוצאות בלתי הוגנות.

4.4.2.2 בקרות מצופות

- ניהול תיעוד על מטרת המערכת, המבנה שלה, תוצריה הצפויים ומגבלותיה. התיעוד יכול לכלול כרטיסי נתונים (data cards - מקורות והטיות של קובצי נתונים), כרטיסי מערכת (system cards - ארכיטקטורה, נתוני אימון, מגבלות) וכרטיסי ביקורת (audit cards - דרישות רגולציה וכימות סיכונים).
- הצגת הודעה בפני משתמשים כאשר הם באינטראקציה עם בינה מלאכותית.
- מתן מנגנונים בידי המשתמשים לערער על החלטות, ווידוא קיומם של הסדרים הנותנים למשתמשים אפשרות לקבל פיצוי אם נפגעו.
- הצדקת השימוש במודלים בעלי "קופסה שחורה" במקרה הרלוונטי.
- הבהרת אחריות משפטית, לרבות דרישות למתן גילוי נאות לגבי תוכן סינתטי.
- שימוש בפתרונות כמו סימני מים (גילויים הממוקמים על גבי תוכן שנוצר באמצעות בינה מלאכותית) וזיהוי מקור ואותנטיות (איתור מקור התוכן וההיסטוריה שלו), כדי לסייע בזיהוי חומר שנוצר באמצעות בינה מלאכותית.

4.4.3 הוגנות ומזעור פגיעה

תהליכי הפיתוח והבדיקה של מערכת בינה מלאכותית צריכים להבטיח שהמערכת תגרום נזק מזערי ותבטיח הוגנות. הוגנות (fairness) פירושה יחס צודק לפרטים ולקבוצות והימנעות מהשפעות שליליות בשל מאפיינים כמו מגדר, מוצא או מיקום. פלט לא הוגן של מערכת עלול לגרום לאפליה, להפסד כספי או לנחיתות חברתית בדרך אחרת.

ישנן דרכים רבות להגדיר הוגנות ולמדוד אותה. הגישה המתאימה תהיה תלויה במטרת המערכת ובהקשר. על המבקרים לצפות מארגונים מבוקרים לתעד ולנמק את בחירתם במדדי הוגנות, ולהראות כיצד מדדים אלה תואמים את דרישות החוק והמדיניות. יש להתייחס הן להוגנות כלפי הפרט, הן להוגנות כלפי קבוצות.

מערכות בינה מלאכותית כמעט לעולם אינן פועלות במנותק ממערכות אחרות. כאשר מתקיימת אינטראקציה בין כמה מודלים, חשוב להעריך כיצד האינטראקציה עלולה לייצר סיכונים או להגביר סיכונים. התמקדות בהערכות של מודל יחיד עלולה לגרום להערכת חסר של הסיכונים הקיימים במערכת הכוללת.

עבור מערכות נרכשות או מודלי יסוד, יש לספק תיעוד המראה כיצד הטיה זוהתה וטופלה. התיעוד יכלול פרטים על מקורות נתוני האימון, שיטות להפחתת הטיות ומדדי הוגנות. התיעוד ובדיקות ההוגנות שמבצעים מפתחים מצד שלישי צריכים להיות רלוונטיים למקרה השימוש של הארגון ולנתונים שלו.

בדיקות הוגנות יכולות לכלול בדיקות של תרחישים חלופיים (counterfactual testing) ושימוש במאגרי נתונים המשמשים כעוגן להשוואה (benchmark datasets). אלה מסייעים לזהות אם המערכת מתייחסת למקרים דומים באופן עקבי, ללא קשר למאפיינים אישיים מוגנים בחוק (כגון מגדר או דת). בדיקת תרחישים חלופיים יכולה לשמש להערכת הוגנות - משנים משתנים רגישים בנתוני הקלט (כמו מגדר או מיקום) ורואים אם התוצרים משתנים באופן לא הולם. קובצי נתונים להשוואה יכולים לשמש להערכת ביצועים והוגנות של מערכת בינה מלאכותית עבור קבוצות שונות.

הוגנות בסיווג שמבצעת הבינה המלאכותית ניתן להעריך על ידי השוואת מדדי ביצוע שונים עבור קבוצות שונות, כדי לזהות הטיה או אפליה פוטנציאליות. על המבקרים להיות מודעים לכך שאף מודל לא יכול לעמוד בכל קריטריוני ההוגנות בבת אחת, במיוחד כאשר שכיחות הקבוצות שונה. לפרטים נוספים על מדדי הוגנות למודלים של סיווג, ראו נספח 2.

הטיה עלולה להיווצר מנתוני אימון בלתי מייצגים או בעלי איכות ירודה, או מהמבנה של המערכת עצמה. מאחר שביצועי המודל משתפרים ככל שכמות הנתונים גדלה (ראו

סעיף 4.2), התוצאות משקפות לא פעם העדפה לקבוצות רוב, בשעה שמיעוטים נותרים בעמדת נחיתות. מודלים מסוג בינה מלאכותית יוצרת פגיעים לכך במיוחד, מפני שהם מאומנים בדרך כלל על קובצי נתונים גדולים ובלתי מסוננים מהאינטרנט שעשויים לשקף הטיות חברתיות. התוכן שנוצר יכול לשקף או לחזק סטריאוטיפים מזיקים.

כאשר המבקרים רוצים להעריך אם קיימת הטיה בנתוני האימון, עליהם לבחון את המקור והאיכות של מאגרי הנתונים, במיוחד אם הם נלקחו מהאינטרנט או נוצרו באופן סינתטי. על הארגון להראות כי הוא טיפל בנושאים של זכויות קניין רוחני, הגנת המידע וסטנדרטים של אתיקה בעת איסוף הנתונים והשימוש בהם. תיעוד ברור צריך להיות זמין למבקרים.

הפיקוח האנושי חשוב, בפרט ביישומים הכרוכים בסיכון גבוה או ברגישות מיוחדת. על המבקרים לוודא כי בודקים אנושיים מעורבים באימות הפלט (אימות מסוג "אדם בתוך התהליך" - human in the loop), במיוחד כאשר השפעתן של החלטות אוטומטיות עשויה להיות משמעותית. על הארגון להראות כי הוא מקיים ניטור שוטף לגילוי בעיות של הוגנות וטיפול בהן מרגע שהן מתעוררות לאחר עליית המערכת לאוויר (ראו מידע נוסף בסעיף 4.6).

4.4.3.1 סיכונים שיש להביא בחשבון

- מערכת הבינה המלאכותית מחזקת או מרחיבה פערים קיימים.
- הטיה או אפליה אינן מתגלות בשל בדיקות הוגנות בלתי מתאימות או בלתי מספקות.
- אינטראקציה בין מערכות או מודלים שונים יוצרת סיכונים חדשים שאינם מתגלים על ידי הערכת כל רכיב בנפרד.
- מערכות שנרכשו מבחוץ או שייכות לצד שלישי נעדרות שקיפות לגבי בדיקות הוגנות או הפחתת הסיכון להטיות.
- הסתמכות על מדד הוגנות יחיד או בסיסי השוואה גנריים עלולה להקנות תחושת ביטחון כוזבת במהימנות התוצרים.

4.4.3.2 בקרות מצופות

הוגנות היא בין השיקולים החשובים ביותר לאורך שלבי התכנון, הפיתוח והפריסה של מערכות בינה מלאכותית. על המבקרים לצפות לראות ראיות לבדיקות הוגנות עמידות ותואמות הקשר, תיעוד ברור וניטור שוטף. כאשר אין אפשרות להבטיח הוגנות באופן מוחלט, חשוב להכיר בסיכונים בשקיפות ולנהל אותם.

- תיעוד של מקורות נתוני האימון, שיטות להפחתת הטיות ומדדי הוגנות. הכללת הרציונל לפי המקרה.
- שימוש בבדיקת תרחישים חלופיים (שינוי משתנים רגישים כדי לבדוק אם מתקבלים שינויי תוצר בלתי הולמים) ובקובצי נתונים להשוואה כדי להעריך הוגנות.
- הבניית פיקוח אנושי לתוך תהליך האימות, במיוחד ביישומים הכרוכים בסיכון גבוה או ברגישות מיוחדת.
- ניטור שוטף המבטיח גילוי בעיות של הוגנות בתפעול לאחר העלייה לאוויר וטיפול בבעיות אלה.
- במקרה של בינה מלאכותית יוצרת, בדיקת פרומפטים ואת תשובות המערכת כדי לוודא שאינה לוקה בדעות קדומות או פוגעניות.
- זיהוי סיכוני אפליה צפויים, מעקב אחריהם וטיפול בהם.
- במודלים של צד שלישי או מודלי יסוד, הארגון המבוקר יוודא כי בדיקות ההוגנות והתיעוד המסופקים על ידי הגוף המפתח רלוונטיים למקרה השימוש המיועד.

4.4.4 אבטחה

אבטחה מתייחסת להגנה על נתונים, נכסים ופונקציות של המערכת מפני גישה בלתי מורשית, ניצול לרעה או נזק. ככל שמערכות בינה מלאכותית מרחיבות את יכולותיהן ונעשות אוטונומיות יותר, כך עולים סיכוני הבטיחות והאבטחה. הדבר נכון במיוחד במקרה של מערכות שמסוגלות לפעול באופן עצמאי או לעבד מידע רגיש.

מערכות בינה מלאכותית מתמודדות עם אותם אתגרי אבטחה שקיימים במערכות IT אחרות. אולם כמויות הנתונים הגדולות וכוח המחשוב האדיר המעורבים בפיתוח בינה מלאכותית משמעם שהאבטחה של תשתית מבוזרת ומבוססת ענן חשובה במיוחד.

4.4.4.1 סיכונים שיש להביא בחשבון

- מערכות בינה מלאכותית יכולות לקבל גישה לנתונים אישיים או סודיים, לחשוף אותם או להשתמש בהם באופן שגוי. קיים גם סיכון של ביצוע שינויים בלתי מורשים, גרימת נזק או אובדן נתונים.
- אי-ציות לדיני הגנת המידע, כמו GDPR, עלול להוביל להשלכות משפטיות ולפגיעה במוניטין. סיכון זה גדל אם הנתונים מעובדים או מאוחסנים במדינות שקיימות בהן תקנות אסדרה שונות.
- גורמים זדוניים עלולים לבצע פעולות מניפולטיביות בקלט הנתונים כדי להטעות את המודל, ולהוביל לפלטים שגויים או מזיקים. דוגמאות לכך כוללות הרעלת נתונים ומתקפת הזרקת פרומפטים.

- תוקפים עלולים להצליח לחדור למודלים או לקוד הבסיס, ולבצע בהם מניפולציות או שימוש לרעה.
- מערכות בינה מלאכותית בעלות רמת אוטונומיות גבוהה, במיוחד סוכני AI, עשויות לקבל החלטות או לבצע פעולות ללא פיקוח אנושי מתאים.

4.4.4.2 בקרות מצופות

נושא האבטחה נשקל בכל שלב במחזור החיים של הבינה המלאכותית. הערכות סיכונים וסקירות קבועות מתבצעות כדי לוודא שהבקרות עדיין מועילות ככל שהטכנולוגיה והאיומים מתפתחים. דיון נוסף על ניטור אבטחה שוטף מובא בסעיף 4.6.

מגוון המערכות והיישומים המבוססים על בינה מלאכותית מקשה לחזות מראש את כל הסיכונים הפוטנציאליים. ארגונים נדרשים לבצע הערכת סיכונים יסודית ולהיות מודעים לסיכוני האבטחה הספציפיים העשויים להתעורר בשלבים שונים של מחזור החיים של המערכת - מהפיתוח ועד הפריסה. האבטחה צריכה להיות שיקול כאשר מחליטים אם לפתח מערכת בינה מלאכותית בארגון או לרכוש מצד שלישי. במקרה של סוגיות אבטחה ידועות, יש ליישם אסטרטגיות להפחתת הסיכונים ולבדוק את יעילותן. כל מערכת צריכה לעמוד לסקירות קבועות כחלק מאסטרטגיית הפחתת הסיכונים, בפרט במקרה של התפתחויות חדשות.

בקרות אפשריות:

- מניעת גישה בלתי מורשית לנתונים, למודלים ולקוד באמצעות בקרות גישה, ניטור רשתות וקריפטוגרפיה (כולל הצפנה).
- הפרדת סביבת פיתוח הבינה המלאכותית מרשתות ה-IT הראשיות, כדי להקטין את הסיכון של פגיעה נרחבת.
- סקירות קבועות של המערכות לגילוי פרצות וסימני תקיפה. בדיקת בעיות בפגיעות המערכת ועדכון אמצעי ההגנה לפי הצורך (צוותים אדומים).
- הכשרת אנשי צוות הארגון לזהות איומי אבטחה ולהבין את אחריותם.
- יישום אמצעי הגנת מידע כדי להבטיח ציות לחוקים ולתקנות הרלוונטיים. לדוגמה, בדיקות שמסננות נתונים זדוניים טרם כניסתם למערכת, ובדיקות של נתונים יוצאים טרם יציאתם מהמערכת (בדיקות אלה עונות על צורכי הבטחת איכות והגנת מידע). במקרה של בינה מלאכותית יוצרת, שימוש במסנני תוכן למניעת חשיפה של מידע רגיש.
- בדיקת עמידתם של ספקי שירותי ענן וספקי מודלים חיצוניים בתקני האבטחה, ובדיקת הסכמי רמת השירות כדי לוודא שהם מכסים ניהול תקריות ותמיכה.
- קביעת נהלים לגילוי תקריות אבטחה, דיווח עליהן ותגובה, כולל חזרה למצב קודם (rollback) או הוצאה משירות במקרה הנדרש.

4.4.5 השפעה סביבתית

השפעתן הסביבתית של מערכות בינה מלאכותית, במיוחד LLM, יכולה להיות משמעותית. האימון והפריסה של מערכות אלה מצריכים על פי רוב כוח מחשוב מהותי, המוביל לצריכת אנרגייה גבוהה ולפליטות פחמן מוגברות. ככל שקהלים רחבים יותר מאמצים בינה מלאכותית, כך גוברת החשיבות של הבנת טביעת הרגל הסביבתית של מערכות מסוג זה וניהולה.

כאשר מערכת נרכשת מצד שלישי, על המבקרים להתייחס לביצועים הסביבתיים של הספק - בין היתר שימוש באנרגייה מתחדשת, תשתית חסכונית באנרגייה ופרקטיקות עסקיות המבוססות על תפיסת הקיימות. רבים מהספקים של שירותי ענן כיום מציעים כלים להערכת טביעת הרגל הפחמנית של פעילויות מחשוב ענן.

4.4.5.1 סיכונים שיש להביא בחשבון

- עלות סביבתית שעולה על יתרונותיה של מערכת הבינה המלאכותית.
- אי-ציות לתקנים של דיווח סביבתי או לחקיקה.

4.4.5.2 בקרות מצופות

על המבקרים לחפש תיעוד המוכיח ציות לתקני דיווח סביבתי או לחקיקה אחרת. מאחר שהמתודולוגיות למדידת השפעה סביבתית עדיין אינן מפותחות במידה מספקת, ביקורת חוקרת עלולה לעורר אתגרים. ככל שהתקנים והכלים למדידת השפעה סביבתית ימשיכו להתפתח, על המבקרים להתעדכן בגישות המומלצות ובדרישות החדשות בתחום זה.

- תיעוד של הציות לתקני דיווח סביבתי ולחקיקה הרלוונטית.
- בחינת הביצועים הסביבתיים של ספקים מצד שלישי.
- שימוש בכלים זמינים להערכת טביעת הרגל הפחמנית של מערכות בינה מלאכותית והמידה שבה הן מנצלות משאבים.
- הערכה ותיעוד של צריכת אנרגייה, צריכת מים ופליטות גזי חממה, היכן שניתן לעשות זאת.

4.5 פריסה וניהול שינויים

פריסת מערכת בינה מלאכותית היא שלב קריטי שמעביר את הטכנולוגיה משלב הפיתוח אל העלייה לאוויר. שלב זה מצריך תכנון קפדני כדי להבטיח שהמערכת תשתלב בצורה חלקה בתהליכים העסקיים ותספק את התועלת המצופה ממנה. ניהול שינויים אפקטיבי מסייע לארגונים להסתגל לדרכי עבודה חדשות ולצמצם את הסיכון להפרעות. פריסה מוצלחת צריכה גם לאפשר עיצוב מחדש של תהליכים ותמיכה במימוש היתרונות המצופים.

פרק זה עוסק בנושאים הבאים:

- מוכנות עסקית וניהול שינויים.
- קריטריונים לעלייה לאוויר ושערי איכות.

פריסה וניהול שינויים קשורים בקשר הדוק לשלבים אחרים במחזור החיים של בינה מלאכותית. הנושאים של הצדקה עסקית, מטרות הפרויקט ומעורבות בעלי עניין הנדרשת לפריסה מפורטים בסעיף 4.1. סעיפים 4.3 ו-4.4 עוסקים בתכנון טכני, בדיקות משתמשים וקריטריוני קבלה שעליהם מבוססת הפריסה. סעיף 4.6 מתייחס לניטור שוטף, אימון מחדש ומעקב ציות לאחר הפריסה.

4.5.1 מוכנות עסקית וניהול שינויים

פריסת מערכות בינה מלאכותית דורשת לעיתים קרובות שינויים משמעותיים בתהליכי העבודה ובמבנה הארגוני. כדי לממש את כל יתרונותיה של הבינה המלאכותית, ארגונים נדרשים לעיתים לעצב מחדש תהליכים קיימים ולוודא שזרימות העבודה החדשות תואמות את יכולות המערכת החדשה. תהליך זה יכול לכלול מיפוי של התהליכים הנוכחיים וזיהוי הזדמנויות לאוטומציה או שיפור. יש להגדיר בבירור את התפקידים האנושיים ואת אלה של הבינה המלאכותית, ולטפל בצעדים מיותרים או צווארי בקבוק באופן מיטבי.

כל בעלי העניין צריכים לקבל מידע על המטרה והפונקציה של המערכת ולהבין אותן. עם בעלי העניין נמנים כל משתמשי הקצה, צוותי IT וההנהלה. הדרכת העובדים היא היבט חיוני במוכנות הארגון, ובמיוחד הדרכה לגבי אי-הוודאות במערכות הסתברותיות, "הזיות" של LLM ואמצעי הבטחת איכות נאותים. אסטרטגיות תקשורת אפקטיביות חשובות לטיפוח בעלות ואמון בקרב בעלי העניין. מנגנונים כמו התייעצויות עם הציבור וסדנאות משותפות יכולים לשמש ככלים לעידוד המעורבות של בעלי עניין.

מסמכי מדיניות בנושא בינה מלאכותית, כמו אסטרטגיית בינה מלאכותית פנימית או קווים מנחים לשימוש בטיחותי ואתי, יכולים לסייע להכין עובדים בכל חלקי הארגון להסתגל לפתרונות מבוססי בינה מלאכותית. מסמכים אלה אמורים להבהיר את התפקידים ותחומי האחריות של כל אחד מהעובדים ולוודא שמשתמשי הקצה מודעים למגבלות המערכת. צריכות להיות הנחיות המבהירות את יחסי הגומלין בין מערכות בינה מלאכותית לעובדים אנושיים, כולל סמכות ואחריותיות במקרה של מחלוקת.

יש חשיבות קריטית להגדרת תהליכי המעבר מהפיתוח לייצור, וכן לתכנון המשכיות והסדרי מסירה למועד שבו אנשי מפתח יעזבו את הפרויקט. על הארגון לשקול להקים פורום של בעלי המומחיות הטכנית הנדרשת - נפרד מצוות הפיתוח - כדי להעריך באופן ביקורתי את השימוש בתוצרי המודל: הן יחידות של בקרה פנימית, הן מנגנונים לטיפול בתלונות של משתמשים או בסוגיות הקשורות לנתונים.

כדי להבטיח שיתרונות המערכת יתממשו, על הארגון לקבוע יעדים ברורים (כמו רווחי התייעלות, חיסכון בעלויות, איכות שירות משופרת או שיעורי טעות מופחתים) ולפתח אינדיקטורים והסדרי ניטור שיעקבו אחר מידת התממשות היתרונות המצופים לאחר הפריסה. האחריות למימוש היתרונות צריכה להיות מוטלת על גורם ספציפי, ויש להפיק לקחים וללמוד מהם כבסיס לפרויקטים עתידיים.

4.5.1.1 סיכונים שיש להביא בחשבון

- ריכוז הידע הטכני על מודל מסוים אצל מספר קטן של עובדים או יועצים חיצוניים, דבר המוביל לחוסר תקשורת וחוסר יעילות בהטמעת המערכת.
- משתמשים במערכת החדשה שאינם מוכנים לשנות תהליכי עבודה מוכרים או מרגישים עודף ביטחון או סקפטיות לגבי היכולות של מערכת הבינה המלאכותית.
- תהליכי עבודה נוקשים או חסרי גמישות המעכבים את אימוץ המערכת או יוצרים חסמים לכך.
- היעדר שקיפות והסברתיות עבור משתמשי הקצה.
- מעבר לא תקין מהפיתוח לייצור.
- משתמשי קצה שאינם מודעים במידה מספקת למגבלות מערכת הבינה המלאכותית, דבר המוביל לחוסר שקיפות או חוסר הוגנות בהחלטות.
- הדרכה או הנחיות בלתי מספקות למשתמשי הקצה, המובילות לשימוש שגוי או חוסר הבנה.
- היעדר תהליכים מוגדרים למעבר, תכנון המשכיות ומסירה בסיום תפקיד.
- תקשורת לא מסודרת ומעורבות לקויה של בעלי עניין, המובילות להתנגדות או לחוסר אמון.
- היעדר תכנון מחדש של התהליכים, או היעדר הגדרת יעדים למימוש יתרונות המערכת. התוצאה היא החמצת הזדמנויות לשיפור.

4.5.1.2 בקרות מצופות

- מתן הדרכות למשתמשים כדי להבטיח שיהיו להם המיומנויות וההבנה הנדרשות כדי להשתמש במערכת בצורה מועילה, כולל כישורי הערכה ביקורתית.
- גיבוש מדיניות או הנחיות ברורות בנוגע לאינטראקציה בין האדם לבינה המלאכותית, המבהירות נושאים של סמכות ואחריות.
- הגדרת תהליכים עבור המעבר משלב הפיתוח לייצור, תכנון המשכיות ומסירת המידע הלאה בסיום תפקיד.
- הקמת פורומים לבקרה פנימית בלתי תלויה וטיפול בתלונות חיצוניות.
- שימוש באסטרטגיות תקשורת כדי לעורר מעורבות בקרב בעלי עניין ולבנות אמון.
- מיפוי ועיצוב מחדש של תהליכים עסקיים כדי לשלב את הבינה המלאכותית בצורה מועילה.
- קביעת יעדים למימוש יתרונות המערכת ומעקב התקדמות.
- הקצאת אחריות למימוש יתרונות ושיפורים מתמידים.

4.5.2 קריטריונים לעלייה לאוויר ושערים

עלייה לאוויר תלויה בכך שהמערכת עברה את כל ההערכות טרום פריסה, כולל עמידה בקריטריוני קבלה בהיבטים הטכניים, האתיים והמשפטיים (ראו סעיפים 4.3 ו-4.4). תוצאותיהן של הערכות אלה, לרבות בדיקות קבלה על ידי המשתמשים וראיות לעמידה במדדי KPI עסקיים, חייבות להיות מתועדות ומאושרות בחתימת בעלי עניין רלוונטיים.

הקריטריונים לעלייה לאוויר צריכים לכלול אישור לכך שהמערכת מצייתת לחוקים ולכללים הנדרשים, חבילות תיעוד (כגון כרטיסי מודל, כרטיסי מערכת וכרטיס ביקורת) ופרסום מידע רלוונטי לגבי שקיפות ואחריותיות (ראו סעיף 4.4). חייבים להיות תהליכי ניהול שינויים הקובעים כיצד ינוהלו עדכונים עתידיים, הגירת נתונים וחזרה למצב קודם - כולל בקרת גרסאות, בדיקות תאימות ומנגנונים של חזרה למצב קודם המבטיחים המשכיות ומהימנות.

כאשר מערכות בינה מלאכותית נרכשות מבחוץ או מסתמכות על מודלים חיצוניים, הסכמי רמת השירות צריכים להבטיח תמיכה שוטפת ושליטה על תשתיות או שירותים חיצוניים במידה מספקת לניהול תקריות וניתוח כשלים (ראו סעיף 4.6).

4.5.2.1 סיכונים שיש להביא בחשבון

- מערכת שאינה בשלה או לא עמדה ברף הביצועים הנדרש, ולא עמדה בקריטריונים להעלאתה לאוויר - ובכל זאת עוברת לסביבת הייצור.
- היעדר אפשרות לשחזר התנהגות של גרסאות ייצור קודמות לאחר שגרסה חדשה עולה לאוויר.
- תמיכה בלתי מספקת או שליטה בלתי מספקת במערכות שנרכשו או פותחו מחוץ לארגון.

4.5.2.2 בקרות מצופות

- וידוא קיומם של הקריטריונים לעלייה לאוויר ושערים (נקודות בקרה), כולל אישורי ציות וחבילות תיעוד.
- חתימה על הסכמי רמת שירות עבור מערכות שנרכשו או פותחו מחוץ לארגון, המבטיחים תמיכה ושליטה באופן רציף.
- ניהול בקרת גרסאות, בדיקות תאימות ומנגנוני חזרה למצב קודם לניהול שינויים (ראו סעיף 4.6).

4.6 מערכות בינה מלאכותית בסביבה תפעולית

מרגע שמערכת בינה מלאכותית נפרסת בארגון, היא פועלת בתנאי העולם האמיתי, אשר עשויים להיות שונים מאלה ששררו במהלך שלבי הפיתוח והבדיקה. היות שכך, מערכות בסביבה התפעולית חשופות לכמה סיכונים חדשים.

נתונים "חיים" לעיתים שונים ממאגרי נתונים היסטוריים או כאלה שנאספו ממקומות שונים, ששימשו לאימון ולבדיקה. נתונים חיים יכולים לכלול קלט בלתי צפוי או "מלוכלך". מרגע שבינה מלאכותית הופכת לחלק מהתהליכים התפעוליים של הארגון, משתמש בה צוות שלא בהכרח היה מעורב בפיתוחה. הדבר מגביר את הסיכון לשימוש שגוי או שימוש למטרות החורגות מהכוונה המקורית. סביבות תפעוליות גם כרוכות בדרישות גדולות יותר מבחינת היציבות הטכנית, המהירות והזמינות.

עם הזמן, שינויים בהתנהגות המשתמשים, גורמים דמוגרפיים או גורמים חיצוניים יכולים לשנות את התפלגות נתוני הקלט. מצב זה ידוע כ"סחף נתונים" (data drift). גם סחף של המודל יכול להתרחש - מצב שבו ביצועי המודל הולכים ונגרעים לאורך זמן. שינויים אלה עלולים לפגוע בביצועי המערכת או להכניס בה הטיות חדשות. כמו כן שינויים בנורמות המקובלות, במדיניות או בחקיקה עשויים להשפיע על דרישות הציות.

אפקטיביות הניטור והניהול חיונית כדי להבטיח שהמערכת תמשיך לספק ערך, תמשיך לציית לתקנות ותפעל בצורה בטוחה ואתית. ארגונים נדרשים לבחון תצפיות בקביעות ולדרוש מהבעלים של המערכת בארגון להיות מודעים

לביצועיה. אם מערכת הבינה המלאכותית אינה עומדת עוד בדרישות הביצוע או הציות, או שהצורך העסקי השתנה, יש לנקוט פעולה מתקנת (כגון אימון מחדש, התאמות טכניות או הוצאת המערכת משירות). אימון מחדש צריך להתבצע בהתאם לאותם סטנדרטים של תיעוד, בדיקה וקבלה שחלו בעת פיתוח המודל הראשוני (ראו סעיף 4.3). שינויים של מאפיינים, פרומפטים או ארכיטקטורת המודל חייבים להיות מסומנים במספרי גרסאות ומתועדים כראוי.

מערכות בינה מלאכותית רבות מסתמכות על מודלים, APIs או שירותי ענן מצד שלישי. ספקים עשויים לעדכן מודלים או להוציא אותם מהשוק, לשנות APIs או לשנות את תנאי השירות. שינויים אלה עלולים להשפיע על הביצועים או להכניס סיכונים חדשים. אם עדכון על ידי הספק או שינוי המודל משפיעים על התנהגות המערכת באופן מהותי, על הארגון להפעיל שוב את קריטריוני הקבלה, ולהשיב את המערכת לסביבה התפעולית רק לאחר שהיא עוברת את המבחנים הרלוונטיים.

4.6.1 סיכונים שיש להביא בחשבון

- סחף ביצועים (performance drift): המערכת עשויה להיתקל בנתונים חיים שהם שונים מנתוני האימון או הבדיקה שסוננו בקפידה, מה שיוביל לירידה ברמת הדיוק או האמינות של המערכת. ייתכן שהבדלים אלה יהיו נוכחים מההתחלה, וייתכן שיופיעו בהדרגה בצורת סחף נתונים.
- שימוש שגוי או לא בהתאם לכוונה: מערכות בינה מלאכותית יכולות לשמש בדרכים שאינן צפויות מראש במהלך הפיתוח. משתמשים חדשים עשויים ליישם את מערכת הבינה המלאכותית בהקשרים שונים או לטעות בפרשנות התוצרים שלה. זה עלול להוביל להחלטות לא נאותות ולשימוש בניבויים של המערכת מחוץ להקשר שנועדה לו.
- ירידת ביצועים עקב שימוש שגוי: מערכת הבינה המלאכותית עשויה לספק ביצועים כמצופה, אך להוסיף להיכשל בהשגת ה-KPIs אם ההליכים סביבה אינם מיושמים כראוי. לדוגמה, ייתכן שהיא לא תספק את התועלת המצופה ממנה אם העובדים לא משתמשים בניבויים שלה בצורה מועילה.
- היעדר בעלות: ללא אחריות ברורה על הניהול השוטף, כולל בעלות עסקית והעברת אחריות כשאדם עובר תפקיד, ייתכן שהמערכת לא תנוטר כראוי או לא תעודכן במועד הנדרש.
- פערי ציות: חוקים ותקנות עשויים להשתנות עם הזמן, ומערכות יכולות להפוך לכאלה שאינן עומדות בחוק אם לא מקפידים על סקירות יזומות.
- שינויים מטעם הספק ושינויי גרסה של המודל: ספקים עשויים לעדכן מודלים או להוציא אותם מהשוק, לשנות APIs או לשנות את תנאי השירות. אירועים אלה יכולים לפגום בביצועים או לעורר סיכונים חדשים.

4.6.2 בקרות מצופות

כדי לנהל סיכונים אלה, על הארגון לבצע את הפעולות האלה:

- מעקב רציף אחר רמת הדיוק, האמינות, העמידות ומדדי הביצוע העיקריים שנקבעו בשלב תכנון הפרויקט.
- מעקב אחר איכות נתוני הקלט וייצוגיות.
- קביעת בדיקות אוטומטיות או מתוזמנות שתכליתן לבדוק אם מתרחש סחף נתונים או סחף מודל, והגדרת רמה קבילה של סיפי ביצוע. אם מערכת הבינה המלאכותית פותחה על ידי צד שלישי ואין בידי הארגון פרטים על הנתונים, עליו לקבל מהצד השלישי דוחות קבועים על איכות הנתונים.
- גיבוש מדיניות לשימור או עדכון של מודלים במקרה שהתגלה סחף. תיעוד אירועי אימון מחדש, כולל הנתונים ששימשו לאימון, השינויים שנעשו בהם ותוצאות ההערכה. וידוא כי שינויים במאפיינים, בפרומפטים או בארכיטקטורת המודל מסומנים במספרי גרסאות ומתועדים כראוי.
- מעקב אחר הטיות והוגנות בפלטים של המערכת, לצד מעקב אחר משוב ותלונות משתמשים.
- קביעת נהלים לגילוי, רישום וחקירה של תקריות. הודעה לבעלי הענין לפי הנדרש ונקיטת פעולה מתקנת, כולל החזרת המערכת למצב קודם או הוצאתה משירות לפי הצורך.
- סקירות קבועות של ציות המערכת לחוקים ולתקנות עדכניים, במיוחד עבור מערכות שמעבדות נתונים אישיים, מקבלות החלטות המשפיעות על פרטים או מופעלות במגזרים מפוקחים.
- מתן הנחיות ברורות והדרכה למשתמשים, כדי לוודא שהם מבינים מהו השימוש המיועד של המערכת ומהן מגבלותיה. יש לזהות שימושים שגויים שניתן לצפותם מראש באופן סביר, לתעד אותם ולנהל אחריהם מעקב.
- הבטחת בעלות ברורה על הניהול השוטף, ניטור ועדכונים.
- הגדרת טריגרים להוצאה משירות, נהלים להעברת נתונים לארכיון או מחיקתם, ותחומי אחריות לניהול מעבר חלק ליום שבו המערכת תוצא משימוש.

- מעקב אחר גרסאות הספק/מודל שבשימוש. בדיקות שמטרתן לגלות נסיגה בביצועים או בעיות תאימות לאחר עדכוני הספק. ניהול מנגנונים המאפשרים להחזיר את המערכת למצב קודם או פתרונות חלופיים, ותיעוד כל השינויים והשפעתם.

5. סיכום ומסקנות

גופים במגזר הציבורי מפתחים ומטמיעים מערכות בינה מלאכותית בשירותים לציבור כדי לשפר את איכות השירות ולהפחית עלויות. ואולם טכנולוגיה חדשה זו מלווה באתגרים חדשים. מוסדות ביקורת עליונים נדרשים להצטייד בכל הדרוש כדי לבצע ביקורת מועילה במערכות בינה מלאכותית. כמה מדינות כבר החלו להפעיל תוכניות פיילוט או שכבר ביצעו את הביקורות הראשונות במערכות מסוג זה במגזר הציבורי. מסמך זה נועד לסייע לקהילת הביקורת הבין-לאומית על ידי התוויית הנחיות ושיטות מומלצות לביצוע ביקורת במערכות בינה מלאכותית.

קטלוג הביקורת בנוי סביב שישה תחומי מפתח בביקורת, המתמקדים בניהול וממשל של הפרויקט, נתונים, פיתוח המערכת, הערכה לפני פריסה, פריסה וניהול שינויים, וניהול שוטף של מערכות בינה מלאכותית בסביבה תפעולית. הקטלוג מנחה את המבקרים לאורך תהליך טיפוסי של פיתוח בינה מלאכותית. טכניקות הביקורת המתוארות כאן יושמו בביקורות שביצעו המוסדות החתומים על מזכר ההבנות. מטרת הביקורות הייתה להעריך את התועלת והמהימנות של יישומי למידת מכונה, וכן את היעילות והאפקטיביות של ההטמעה והתפעול של יישומים אלה.

בינה מלאכותית נמצאת בשימוש במגוון רחב של מגזרים, ובכל אחד מהם סיכונים שונים החלים על יישומים שונים. תחום הבינה המלאכותית עודנו מתפתח, חקיקה חדשה ממשיכה להתגבש ומוקמים גופי אסדרה ופיקוח ייעודיים. קטלוג ביקורת זה משקף את התקן לביקורת בינה מלאכותית בקרב מוסדות הביקורת העליונים שחיברו את המסמך, במועד הכתיבה[67]. ייתכן שהמסמך יעודכן ככל שיצטבר ניסיון ביקורת נוסף ויתקבלו תוצאות של מחקרים חדשים.

למסמך זה מצורף כלי עזר לביקורת בינה מלאכותית המאפשר למבקרים לבחור שאלות מתוך רשימה מוכנה וליצור שאלון מותאם ההולם את הביקורת המסוימת שהם מבצעים. כלי העזר לביקורת מספק שאלות ביקורת מומלצות, ומציין מה הן ראיות הביקורת הנדרשות.

נספחים

נספח 1: רכיבים של ביקורת טכנולוגיית מידע קלאסית בהקשר של בינה מלאכותית

מערכות בינה מלאכותית חולקות מאפיינים רבים עם תוכנה מסורתית, כך שתקני ביקורת מבוססים כמו התהליך התקני חוצה התעשיות לכריית מידע (Cross-Industry Standard Process for Data Mining) יכולים להנחות את סקירתן[68]. התקן משקף תהליך פיתוח סטנדרטי של מערכות בינה מלאכותית, גם אם הגורם המפתח לא משתמש בו באופן מפורש. מבקרי IT יכולים אפוא לבצע סקירה ברמה גבוהה ללא מומחיות בבינה מלאכותית. התקן מחולק לשישה שלבים:

1. הבנה עסקית
2. הבנת הנתונים
3. הכנת הנתונים
4. בניית המודל ופיתוח
5. הערכת המודל לפני פריסה
6. פריסה ותהליכי ניהול השינויים המלווים אותה

הביקורת בחלק מן השלבים, כמו בשלב ההבנה העסקית, מתבצעת באותו האופן כמו ביקורת של פרויקטים רגילים לפיתוח תוכנה. שלבים אחרים, כמו הכנת הנתונים ובניית המודל, עשויים לחייב מומחיות בנושא. הגישה המומלצת היא להרכיב צוותי ביקורת המכילים תערובת של מיומנויות. צוות מאוזן יכול לכלול מבקרים מומחים בתחום, מבקרי IT ומדעני נתונים. אם חסר בצוות ניסיון בתחום מדעי הנתונים, על אנשי הצוות לבקש חוות דעת מעמיתיהם שהם בעלי מומחיות זו. באותו אופן, צוותים שהיכרותם עם ביקורות IT טובה פחות, צריכים לערב מבקרי IT מנוסים. גישה זו תסייע להבטיח שכל היבטי המערכת ייסקרו כראוי.

בית הדין הפדרלי לביקורת חשבונות של גרמניה (Bundesrechnungshof) נקט גישה זו בביקורת הראשונה שביצע למערכת בינה מלאכותית. צוות הביקורת כלל מבקר מומחה בעל ידע על הארגון המבוקר ושני מבקרים טכניים: אחד בעל רקע ב-IT, והאחר - במדעי הטבע. מבקרים אלה הובאו מיחידות אחרות כדי להשתתף בצוות ייעודי לביקורת בינה מלאכותית.

כלי העזר לביקורת הכלול במסמך זה נבנה בהתאם למחזור החיים של הבינה המלאכותית, ומציג שאלות ביקורת המתאימות למבקרים מומחים, מבקרי IT ומדעני נתונים.

---

67 עדכון אחרון: דצמבר 2025.

68 A. Clark (2018): *The Machine Learning Audit-CRISP-DM Framework*.

נספח 2: נתונים אישיים והתקנות הכלליות להגנת המידע (GDPR) בהקשר של בינה מלאכותית

מערכות בינה מלאכותית מסתמכות לעיתים קרובות על מאגרי נתונים גדולים שעשויים להכיל נתונים אישיים. תקנות הגנת המידע, כמו התקנה הכללית להגנת מידע (GDPR) באיחוד האירופי, חלות הן על הפיתוח של מערכות אלה, הן על השימוש בהן. להלן שיקולים רלוונטיים למבקרים בתחום זה:

- מגבלות הנובעות מהמטרה: יש לאסוף נתונים אישיים אך ורק למטרה ברורה ומסוימת. כל עיבוד נוסף חייב להתאים למטרה המקורית, למעט החרגות בהיקף מוגבל. כאשר בינה מלאכותית מאומנת על נתונים היסטוריים (שייתכן כי נאספו לפני תחילת הפרויקט), המטרה המקורית חייבת להצדיק עיבוד נוסף מסוג זה.
- מזעור נתונים: השימוש בנתונים אישיים יוגבל רק לרמה ההכרחית למימוש המטרה שלשמה הם נאספו. האמור חל הן על המודל הסופי, הן על נתונים כלשהם ששימשו במהלך הבדיקה או האימון. על המבקרים לבדוק אילו נתונים היו בשימוש, מהו תפקידם בביצועי המערכת וכיצד הם מטופלים לאורך תהליך הפיתוח.
- מידתיות: חשוב שהכמות והסוג של הנתונים יהיו מידתיים ביחס למטרה ושלא יהיו פולשניים מדי. לדוגמה, זיהוי פנים כדי לעקוב אחר נוכחות תלמידים בבית הספר הוא מן הסתם שימוש מופרז.
- שקיפות: היכולת להסביר תהליכים והחלטות היא דרישה כללית בשירותים לציבור, במיוחד אם נתונים אישיים מעורבים בכך. עיבוד של נתונים אישיים צריך להתבצע בשקיפות.
- הארגון שמפתח את מערכת הבינה המלאכותית ומשתמש בה אחראי לפעול בהתאם לכללי הגנת המידע ולציית ל-GDPR. אם הבינה המלאכותית מעמידה את זכויות הפרט בסיכון גבוה, נדרש לבצע הערכה של ההשפעות על הגנת המידע (DPIA). ה-DPIA צריך להציג ראיות לכך שנשקלו חלופות הכרוכות בסיכונים נמוכים יותר. המבקרים יכולים לבדוק ציות בדרך כלל על ידי עיון בתיעוד.

נספח 3: מדדי שוויון והוגנות במערכות סיווג

הערכת מודלי סיווג מתבצעת בדרך כלל בעזרת שורה של מדדים סטנדרטיים הנגזרים ממטריצת בלבול (confusion matrix) - טבלה המשווה את הניבויים של מודל הסיווג כנגד התוצרים בפועל, כדי למדוד את רמת הדיוק ומדדי ביצוע אחרים הקשורים בכך. חלק ממדדים אלה עשויים לחול על מודלים של בינה מלאכותית יוצרת, אולם סעיף זה מתייחס ספציפית לבעיות בסיווג.

עם המדדים האלה נמנים:

- **שכיחות:** שיעור המקרים החיוביים בפועל בכל קבוצה.
- **שכיחות מנובאת:** שיעור הניבויים החיוביים בכל קבוצה.
- **שיעור תוצאות חיוביות נכונות (true positive rate - TPR):** שיעור המקרים החיוביים בפועל שזוהו נכונה על ידי המודל.
- **שיעור תוצאות חיוביות שגויות (false positive rate - FPR):** שיעור המקרים השליליים שסווגו באופן שגוי כחיוביים.
- **דיוק (ערך ניבוי חיובי):** שיעור הניבויים החיוביים שזוהו נכונה.
- **ערך ניבוי שלילי** - (negative predictive value (NPV: שיעור הניבויים השליליים שזוהו נכונה.

הוגנות ניתנת להערכה בכמה דרכים:

- **יחס מפלה או אי-שוויון הזדמנויות:** הדרישה היא כי שיעורי התוצאות החיוביות הנכונות יהיו דומים עבור קבוצות שונות, בהתמקד על סיכויים שווים לקבל תוצאות חיוביות נכונות.
- **השפעה בלתי שוויונית או צדק חלוקתי:** זו הערכה של מידת הדיוק, והמטרה היא להשיג רמת דיוק דומה של ניבויים חיוביים עבור קבוצות שונות.
- **שוויון סטטיסטי (שוויון דמוגרפי):** הדרישה היא כי שיעור הניבויים החיוביים יהיה זהה עבור כל הקבוצות, ללא קשר להבדלים הקיימים בתוצאות בפועל.

מדדים נוספים של הוגנות כלפי כל הקבוצות:

- **שקילות סיכויים (equalized odds):** הדרישה היא שה-TPR וה-FPR יהיו שווים עבור כל הקבוצות, או שהתראות שווא וזיהויים נכונים יתרחשו בשיעורים דומים עבור כל הקבוצות.
- **דַיות (sufficiency) (שקילות בשיעור הניבוי):** הדרישה היא לרמת דיוק שווה וערך ניבוי שלילי שווה עבור כל הקבוצות. לדוגמה, כאשר המודל מנבא שאדם יעבור שוב עבירה, ההסתברות שהמודל צודק זהה עבור כל הקבוצות.

חשוב לציין שבפועל, אף מודל בלתי מושלם לא יכול לעמוד בכל קריטריוני ההוגנות בבת אחת, במיוחד כאשר שיעורי התוצאות החיוביות שונים בבסיסם בין הקבוצות. המבקרים נדרשים להביא בחשבון מגבלות אלה ולהתמקד

בהשפעה הפרקטית של הבדלים כלשהם.

מושגי הוגנות נוספים:

- **הוגנות באמצעות חוסר מודעות:** השמטה של תכונות רגישות לא תבטיח הוגנות, מכיוון שמשתנים אחרים עשויים להימצא במתאם עם תכונות אלה.
- **הוגנות אישית:** אנשים דומים אמורים לקבל ניבויים דומים, אם כי פרמטר זה קשה בדרך כלל למדידה.
- **הוגנות של תרחישים חלופיים (counterfactual fairness):** הניבוי של המודל לא משתנה אם משנים תכונה רגישה, כאשר כל יתר התנאים שווים.

סקירה טובה של מושגי הוגנות אלה ואחרים מופיעה בספרות[69]. חשוב לציין כי רשימת המדדים הנכללת כאן אינה ממצה, והיא מובאת כנקודת מוצא עבור המבקרים.

נספח 4: מילון מונחים

מילון המונחים מוסיף פרטים על המינוח המשמש לאורך מסמך זה. הודגשו במיוחד המונחים הטכניים הנוגעים לאלגוריתמים ובינה מלאכותית.

בינה מלאכותית (Artificial Intelligence - AI)

מערכות בינה מלאכותית הן טכנולוגיות מעשה ידי אדם שייעודן לבצע מטלות אשר באופן טיפוסי מצריכות אינטליגנציה (בינה). מערכות אלה לומדות מתוך נתונים ומתאימות את עצמן להשגת מטרות ספציפיות, כך שעל פי רוב הן משפרות את ביצועיהן בעצמן, ללא תכנות נוסף במפורש. בינה מלאכותית כוללת תחומים כמו למידת מכונה, עיבוד שפה טבעית, ראייה ממוחשבת, זיהוי קולי, תכנון ורובוטיקה. היא מאפשרת יכולות כמו תפיסה, הנמקה, קבלת החלטות ופתרון בעיות. מרבית מערכות הבינה המלאכותית מסתמכות על למידת מכונה, שבמסגרתה מודלים לומדים דפוסים מתוך נתונים כדי להשיג מטרות ספציפיות.

בינה מלאכותית יוצרת (Generative AI)

מערכות בינה מלאכותית יוצרת יודעות ליצור תוכן חדש (כמו טקסט, תמונות, וידאו או אודיו) על ידי למידה מתוך קובצי נתונים גדולים. הן מסוגלות לייצר תוצרים ריאליסטיים, על פי רוב עם הכוונה אנושית מעטה או ללא הכוונה כזו כלל.

בינה מלאכותית לחיזוי (Predictive AI)

מערכות בינה מלאכותית לחיזוי משתמשות בנתונים היסטוריים כדי לזהות דפוסים ולספק ניבויים לגבי מקרים חדשים. הן מסוגלות להעריך את הסיכון להתפתחות מחלה על פי רשומות רפואיות או לסמן טעויות פוטנציאליות בדוחות מס. מערכות אלה משמשות באופן שכיח לצורכי סיווג (כמו מיון מקרים לקטגוריות) או רגרסיה (כמו ניבוי ערך מספרי).

הדירות (Reproducibility)

הדירות היא היכולת לשחזר תוצאות כשמשתמשים באותם נתונים, אותם פרמטרים ואותה סביבה.

הוגנות (מובנית בתכנון) (Fairness (by design))

הוגנות פירושה יחס שוויוני לפרטים ולקבוצות והימנעות מהשפעות שליליות בשל מאפיינים כמו מגדר, מוצא או מיקום. פלט לא הוגן של מערכת עלול לגרום לאפליה, להפסד כספי או לנחיתות חברתית בדרך אחרת. הוגנות מובנית בתכנון היא תפיסה המטמיעה הוגנות מהרגע הראשון, במהלך שלב התכנון והפיתוח של מערכת בינה מלאכותית, במקום תיקון בדיעבד.

הזיה (Hallucination)

הזיה של בינה מלאכותית היא מצב שבו מערכת מפיקה פלט שמכיל מידע שגוי או שאינו מבוסס על עובדות.

היפר-פרמטר (Hyperparameter)

היפר-פרמטרים משמשים לשליטה בארכיטקטורת המודל ובתהליך הלמידה בלמידת מכונה. פרמטרים אלה נשלטים על ידי כותבי הקוד, בשונה מערכי הפרמטרים של המודל, הנוצרים באמצעות אימון.

הנדסת מאפיינים (Feature Engineering)

הנדסת מאפיינים היא התהליך של הפיכת נתונים גולמיים למשתנים בעלי משמעות עבור המודל.

הסברתיות (מובנית בתכנון) (Explainability (by design))

הסברתיות היא היכולת לתאר כיצד המערכת מגיעה לתוצרים או להחלטות. הסברתיות מובנית בתכנון היא מתודולוגיה הדורשת להבטיח הסברתיות מהרגע הראשון, במהלך התכנון והפיתוח של מערכת בינה מלאכותית.

---

69 ראו לדוגמה: S. Verma & J. Rubin (2018): *Fairness Definitions Explained*

הרעלת נתונים (Data Poisoning)

הרעלת נתונים היא הוספת דוגמאות שנעשו בהן מניפולציות לנתוני האימון, כדי לפגום בביצועי מערכת הבינה המלאכותית או לשנות את התנהגותה בדרך ספציפית.

התאמת יתר (Overfitting)

התאמת יתר היא מצב שבו מערכת בינה מלאכותית מתאימה את עצמה מדי לנתוני האימון, ובעקבות זאת היא לא מספקת ביצועים טובים על נתונים חדשים.

יצירה מבוססת אחזור (Retrieval Augmented Generation - RAG)

יצירה מבוססת אחזור היא גישה המאפשרת למודלי שפה גדולים לשאוב ממקורות מידע מהימנים כשהם יוצרים תשובה. הזנת מסמכים או נתונים רלוונטיים לתוך ההקשר של המודל מאפשרת ל-RAG להתמודד עם אתגרים ידועים בבינה מלאכותית יוצרת, כמו הזיות, מידע מיושן והיעדר ציטוטי מקורות.

כוונון עדין (Fine-tuning)

כוונון עדין הוא תהליך שבו מכניסים התאמות למודל שפה גדול (LLM) באמצעות דוגמאות ספציפיות, כדי שתשובותיו יהיו מדויקות יותר עבור נושאים מסוימים.

למידה עמוקה (Deep Learning)

בלמידה עמוקה, מערכות בינה מלאכותית משלבות כמה שכבות עיבוד כדי לפתור משימות מורכבות.

מהימנות (Reliability)

מהימנות מתייחסת ליכולת של המודל לספק ביצועים עקביים ונכונים בתנאים הצפויים, בפרט בעת טיפול בנתונים הדומים לאלה ששימשו לאימון המודל.

מודלי יסוד (Foundation Models) או מודלי בינה מלאכותית למטרות כלליות (General-purpose AI Models)

מודלי יסוד או מודלי בינה מלאכותית למטרות כלליות מאומנים על מאגרי נתונים גדולים ומגוונים מאוד. הם משתמשים בטכניקות למידה עמוקה, וניתן להתאים אותם למגוון רחב של יישומים, מזיהוי תמונות ועד מחקר מדעי.

מודלי שפה גדולים (Large Language Models - LLMs)

מודלי שפה גדולים הם סוג עיקרי של מודלי יסוד, המאומנים בעיקר על טקסטים. הם מסוגלים לבצע מטלות רבות, כמו תרגום, סיכום ושיחה, על ידי הפקת שפה הדומה לשפה אנושית.

נתוני אימון (training), אימות (validation) ובדיקה (test)

בתחילה משתמשים בנתוני אימון כדי לאמן מודל למידת מכונה לביצועים מיטביים. לאחר מכן משתמשים בנתוני אימות כדי להעריך את התאמת המודל תוך כדי כוונון ההיפר-פרמטרים. נתוני הבדיקה הם האחרונים, ומטרתם לתת הערכה סופית של ביצועי המודל.

עמידות (Robustness)

עמידות היא היכולת של המודל לשמור על רמת ביצועיו כאשר הוא עומד בפני קלט בלתי צפוי - כמו נתונים בהתפלגות שונה מהנתונים שעליהם הוא אומן; נתונים "מחוץ להתפלגות" (out of distribution - OOD); רעש; או שינויים בסביבה.

פרומפט (Prompt)

פרומפט (או הנחיה) הוא קלט משתמש שניתן למערכת בינה מלאכותית יוצרת. מונח זה משמש בעיקר בהקשר של LLM.

פרטיות (מובנית בתכנון) ((Privacy (by design)

פרטיות כוללת הגנה על כל הנתונים האישיים שמערכת הבינה המלאכותית אוספת, משתמשת בהם, משתפת או מאחסנת. פרטיות מובנית בתכנון היא תפיסה המטמיעה פרטיות מהרגע הראשון, במהלך שלב התכנון והפיתוח של מערכת בינה מלאכותית, במקום תיקון בדיעבד.

שקיפות (מובנית בתכנון) ((Transparency (by design)

שקיפות משמעה מתן מידע ברור על השימוש במערכת - איך, מתי ומדוע משתמשים בה. שקיפות מובנית בתכנון היא מתודולוגיה הדורשת להבטיח שקיפות במהלך כל שלב בתהליך התכנון והפיתוח של מערכת בינה מלאכותית.

תפעול למידת מכונה (Machine Learning Operations - MLOps)

MLOps הוא מכלול של פרקטיקות המסייעות לנהל, לבדוק ולעדכן מודלים של למידת מכונה.

Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM הוא תהליך תקני לתכנון ופיתוח של מערכות למידת מכונה ובינה מלאכותית. הוא מורכב משישה שלבים:

- הבנה עסקית
- הבנת הנתונים
- הכנת הנתונים
- בניית מודל
- הערכה
- פריסה

קיצורים באנגלית

AI Artificial Intelligence

API Application Programming Interface

BSI German Federal Office for Information Security

CRISP-DM Cross-Industry Standard Process for Data Mining[70]

DPIA Data Protection Impact Assessment

FLOP Floating Point Operation

GDPR General Data Protection Regulation[71]

GPU Graphics Processing Unit

ICO Information Commissioner's Office

LGPD Brazil's General Data Protection Law

LLM Large Language Model

ML Machine Learning

NIST U.S. National Institute of Standards and Technology

OECD Organisation for Economic Co-operation and Development

RAG Retrieval Augmented Generation

SAI Supreme Audit Institution

TCU Brazil's Federal Court of Accounts

נספח 5: תפקידים

ברשימה הבאה ננסה לסכם את אחריותם של בעלי תפקידים בפרויקט בינה מלאכותית שעשויים להיות רלוונטיים לביקורת. המטרה היא לסייע למבקרים לזהות אנשי קשר בארגון המבוקר. תפקידים בתחום הבינה המלאכותית משתנים ככל שהטכנולוגיות וגישות העבודה מבשילות, כך שאין זו רשימה ממצה, וקרוב לוודאי שתשתנה עם הזמן כדי לשקף כלים, מתודולוגיות ושיקולי אתיקה חדשים.

ארגונים מבוקרים קטנים יחסית עשויים לשלב כמה מהתפקידים הנמנים למטה באדם יחיד או בצוות יחיד.

חלק מהתפקידים יכולים גם להתבצע על ידי יועצים חיצוניים; אולם במקרה שהביקורת מתקיימת לאחר שתפקיד היועץ הסתיים, צריך להיות כוח אדם פנימי שרכש את הידע לגבי התפקיד הרלוונטי. לפיכך, באחריות המבוקר להבטיח תיעוד מספק של כל עבודה המתבצעת על ידי היועצים החיצוניים.

ארכיטקט/מהנדס בינה מלאכותית

אדם זה מגדיר את הארכיטקטורה הכוללת של מערכת בינה מלאכותית ואת אסטרטגיית השילוב שלה. הוא מסייע להגדיר כיצד מודלים, צנרות נתונים, APIs ויישומים ישתלבו יחד. הוא בוחר מסגרות, פלטפורמות ושירותי ענן לפריסת המערכת, ודואג להיבטים של סקלביליות ואבטחה.

בעלי הפרויקט/בעלי המוצר

הצוות או היחידה בארגון המבוקר האחראים למשימה שמעתה תיתמך במערכת בינה מלאכותית או תבוצע אוטומטית בעזרתה. הם מחליטים אילו מדדי ביצוע נדרשים ממערכת הבינה המלאכותית ומה הם תנאי הקבלה של התוצרים בסיום פרויקט הפיתוח של המערכת.

חשב

האדם שעורך ביקורת של פרויקטים ובודק עמידה בעקרונות ממשל.

אנליסט/מדען נתונים

האדם שעובד עם הנתונים שיוזנו לתוך מערכת הבינה המלאכותית ומנתח אותם. הוא אחראי להבין את הנתונים, ועליו להיות מעורב באופן מעמיק בתהליך הפיתוח. הוא מסייע לבעלים של המוצר, על ידי תרגום דרישותיו למפרטים ודרישות שיועברו למפתחים.

מהנדס נתונים

האדם האחראי על ההיבטים הטכניים של הנתונים הגולמיים (מחסני נתונים, איכות נתונים, בקרת גישה) וכן על הבנת הנתונים הגולמיים ומקורותיהם. הוא גם אחראי על אספקת הנתונים וניהול הנתונים.

מומחה לתחום העיסוק (Subject matter expert)

מונח גנרי המתייחס לאדם בעל ידע מומחה בתחום (domain) ספציפי.

---

70 A. Clark (2018): The Machine Learning Audit-CRISP-DM Framework.

71 הפרלמנט האירופי ומועצת האיחוד האירופי (2016): תקנה (האיחוד האירופי) 2016/679 של הפרלמנט האירופי של המועצה מיום 27 באפריל 2016 בעניין הגנה על אישיות טבעית ביחס לעיבוד נתונים אישיים ובעניין תנועה חופשית של נתונים כאמור, וביטול דירקטיבה EC/95/46 (התקנה הכללית להגנת מידע).

ממונה על אבטחת IT

זהו האדם הבכיר ביותר האחראי על אבטחת טכנולוגיית המידע בארגון המבוקר. עליו להכיר את כל היבטי האבטחה של הפיתוח והתפעול של מערכת הבינה המלאכותית.

ממונה על הגנת מידע ופרטיות

זהו האדם הבכיר ביותר האחראי על הגנת מידע בארגון המבוקר. עליו להיות מודע לכל חשש שעולה לגבי שימוש בנתונים אישיים על ידי מערכת הבינה המלאכותית. תפקידו לוודא שהתוכנה מצייתת להוראות דיני פרטיות המידע והתקנות בתחום זה, כדוגמת ה-GDPR באיחוד האירופי.

ממונה על התקציב

אדם זה אחראי לתקציב של הארגון המבוקר, ועל כן אחראי על כל הוצאה הקשורה לפרויקטים של פיתוח מערכות בינה מלאכותית, רכש שלהן או ייעוץ לגביהן. בידיו הסמכות לקבוע אם הפיתוח והתפעול של תוכנה כזו מייצגים ניצול משתלם של תקציב הארגון המבוקר, ועליו להחזיק בכל המידע התקציבי על הפיתוח והתפעול של המערכת.

מנהל הפרויקט

האדם האחראי לכל נושאי הניהול/ממשל של הפרויקט. עליו להיות מסוגל לספק את כל המסמכים הנדרשים הקשורים לניהול הפרויקט.

מפַתח

האדם או האנשים שמייצרים את מערכת הבינה המלאכותית בהתאם למפרטים ולדרישות שהוסכמו עם בעלי המוצר (ומאמנים את המודל, עבור מודלים שנדרש בהם שלב אימון). המפתחים אחראים להפיכת הנתונים הגולמיים למשתנים הסופיים שבהם ישתמש המודל ("הנדסת מאפיינים"), והם עובדים בשיתוף עם מדעני הנתונים ומהנדסי הנתונים, מנהל הפרויקט ובעלי המוצר.

משתמש/אחראי עיבוד

האדם או היחידה שאמורים להשתמש בתוצאות מערכת הבינה המלאכותית לצורך מילוי תפקידם בארגון. הם גורם מכריע בקביעת הצלחתם של פרויקטים, מכיוון שעליהם להבין את ההצעות או התוצאות ממערכת הבינה המלאכותית וליישם אותן במשימות (שגרתיות).

סמנכ"ל מערכות מידע (CIO)

ה-CIO של הארגון המבוקר אחראי לכל מערכות טכנולוגיית המידע של הארגון, ועל כן עליו להכיר את כל מערכות הבינה המלאכותית שכבר פועלות ואת כל הפרויקטים העוסקים בפיתוח תוכנה מסוג זה.

קו חם לתהליכי עיבוד/מוקד תמיכה למשתמשים

הצוות המופקד על מתן תמיכה למשתמשים או לאנשים העוסקים בעיבוד. עליו לדעת להשיב על כל השאלות המתעוררות במהלך התפעול השגרתי של התוכנה.

כלי עזר לביקורת

כלי העזר לביקורת הוא כלי אקסל המכיל שאלות ביקורת מפורטות וראיות ביקורת שעל המבקרים לחפש בעת ביצוע ביקורת בינה מלאכותית. השאלות מסודרות בהתאם לשלבים השונים במחזור החיים של המערכת.

המבקרים יכולים לאסוף שאלות רלוונטיות לביקורת הספציפית וליצור שאלון מותאם לצורכיהם. בדרך זו, הכלי יכול לסייע להכין ולהנחות את המבקרים לאורך שלבי הביקורת של מערכות בינה מלאכותית. הכלי נועד לסייע כעמדת מוצא, ונועד להקיף את ההיבטים החשובים ביותר המתוארים במדריך זה.

הורידו כאן את כלי העזר לביקורת