הוגנות בשירותים במימון ציבורי: עקרונות, סיכונים ופתרונות ליישום מדיניות הוגנות אלגוריתמית במערכות מבוססות בינה מלאכותית

סקירה בין-לאומית

ורד פורזיקי[1]

ענבל לבני נבון[2]

עופר פיינשטיין[3]

[1]מכון מאיירס-ג'וינט-ברוקדייל

[2]אוניברסיטת בן גוריון

[3]מכון מאיירס-ג'וינט-ברוקדייל ואוניברסיטת בן גוריון

מאיירס ג'וינט ברוקדייל

עריכת לשון: רונית כהן בן-נון
עיצוב גרפי: ענת פרקו טולדנו

עורכת ראשית: רויטל אביב מתוק

הדוח הוזמן ומומן על ידי הג'וינט.

הצעה לציטוט:
פורזיקי, ו', לבני נבון, ע' ופיינשטיין, ע' 2026)). *הוגנות בשירותים במימון ציבורי: עקרונות, סיכונים ופתרונות ליישום מדיניות הוגנות אלגוריתמית במערכות מבוססות בינה מלאכותית. סקירה בין-לאומית*. דמ-26-076. מכון מאיירס-ג'וינט-ברוקדייל

מכון מאיירס ג'וינט ברוקדייל
ת"ד 3886 ירושלים 9103702
טלפון: 02-6557400
brook@jdc.org | brookdale.jdc.org.il

תקציר

רקע

בעשור האחרון גובר השימוש במערכות המבוססות על מודלים של בינה מלאכותית (AI) (להלן: מערכות מבוססות בינה מלאכותית) בשירותים במימון ציבורי. אף שטכנולוגיה זו מאפשרת לשפר את היעילות, הנגישות וההתאמה של שירותים לצרכים, אם לא משתמשים בה בזהירות היא מעלה סיכונים ממשיים לשעתוק פערים חברתיים ולהעמקת אפליה בקרב אוכלוסיות פגיעות. לנוכח התרחבות השימוש במערכות מבוססות בינה מלאכותית עלה הצורך לבחון כיצד אפשר להבטיח שהן יפעלו על פי ערכי שוויון והוגנות. על רקע זאת פנה הג'וינט למכון מאיירס-ג'וינט-ברוקדייל בבקשה לערוך סקירה שתבחן את ההטיות והסיכונים הגלומים בשימוש במערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי ולמפות כיווני פעולה מעשיים לקידום הוגנות אלגוריתמית – עיקרון המתייחס לאופן שבו מערכות אלגוריתמיות, ובפרט מערכות מבוססות בינה מלאכותית, מקבלות החלטות המשפיעות על בני אדם, או תומכות בהן. עיקרון זה מבקש להבטיח כי החלטות אלו אינן יוצרות או משמרות הטיות בלתי מוצדקות, אפליה או פגיעה בשוויון הזדמנויות בין פרטים וקבוצות באוכלוסייה. המחקר נערך בשיתוף אוניברסיטת בן גוריון.

מטרה

לגבש הבנה שיטתית של האתגרים והסיכונים הקשורים להוגנות ושוויון הזדמנויות בשימוש במודלים של בינה מלאכותית בשירותים במימון ציבורי ולהציג מארג של פתרונות מעשיים שמאפשרים למשרדי הממשלה ולגופים ציבוריים לתכנן, לפתח ולהטמיע מערכות מבוססות בינה מלאכותית באופן שיצמצם הטיות ויקטין סיכונים של פגיעה בהוגנות.

שיטה

הסקירה מתבססת על מידע ממגוון מקורות: מאמרים אקדמיים מן הספרות המקצועית הישראלית והבין-לאומית; ספרות אפורה: דוחות מחקר, מסמכים רשמיים של משרדי ממשלה וארגונים בין-לאומיים; כתבות בעיתונות ושלושה ראיונות עומק חצי מובנים מקוונים עם מומחים בתחום הבינה המלאכותית בישראל. המידע נאסף בחודשים מארס-דצמבר 2025 והראיונות בוצעו בחודשים מאי-יולי 2025.

עיקרי הממצאים ומסקנות

הסקירה מציגה מיפוי של ההטיות והסיכונים הגלומים בפיתוח מערכות מבוססות בינה מלאכותית ובהפעלתן. הממצאים מצביעים על כך שהחלטות מוקדמות בתכנון מערכות מבוססות בינה מלאכותית מעצבות את גבולות ההוגנות לאורך כל שלבי פיתוח המערכת. משום כך חיוני לבצע תהליכי פיתוח זהירים ופיילוטים מבוקרים. עם זאת הממצאים מראים כי קידום הוגנות אינו יכול להישען על עקרונות תיאורטיים או על זיהוי סיכונים בלבד, אלא מומלץ תחילה לבצע בחינה מעמיקה של מטרות המערכת. על פי המטרות שנקבעו נדרש להגדיר כבר בראשית הדרך מהי ההוגנות הרצויה עבור כל מערכת ולבסס מומחיות ייעודית בקרב קובעי מדיניות, המשלבת הבנה טכנולוגית, חברתית ומוסדית. ממדים אלו יוצרים יחד תשתית כוללת לתכנון מערכות מבוססות בינה מלאכותית בשירות הציבורי ולשימוש אחראי בהן.

בצד זאת יש צורך לא רק במיפוי סיכונים אלא גם בפיתוח, אימוץ והטמעה של כלים ופרקטיקות קיימות להתמודדות עם סיכוני הטיה ואי-שוויון והתאמתם לפעולת המערכת הציבורית, כדי להבטיח הגנה על אוכלוסיות פגיעות ולחזק את אמון הציבור. חשוב לציין כי כל מערכת דורשת מערך פתרונות מותאם, ואין פתרון אחיד המתאים לכלל המערכות. לפיכך מערכות מבוססות בינה מלאכותית יכולות לשמש מנוף לתיקון הטיות ולצמצום אי-שוויון, אך הדבר תלוי בהטמעה אחראית המשלבת כלים טכנולוגיים עם מעטפת מוסדית חזקה שמקדמת שוויון הזדמנויות וצדק חברתי.

דברי תודה

אנו מודות למרואיינים במחקר על הזמן, הפתיחות ושיתוף הפעולה שאפשרו העמקה בתובנות מעשיות ומורכבות.

תודתנו נתונה למערך הדיגיטל הלאומי על תרומתו למחקר, ובפרט לסדריק יהודה צבע, מוביל מדיניות AI, יחידת הדאטה והבינה המלאכותית; ולגל תמיר, מנהל אגף אסטרטגיה ומדיניות ביחידת הדאטה וה-AI.

תודתנו שלוחה גם לפרופ' עומר ריינגולד מאוניברסיטת סטנפורד, על מחשבה ביקורתית, חידוד תובנות והרחבת ההבנה התיאורטית והיישומית של הסוגיות שנדונו.

תוכן עניינים

רשימת לוחות

רשימת תרשימים

1. מבוא

רקע

בשנים האחרונות הולך וגובר השימוש במערכות המבוססות על מודלים של בינה מלאכותית (AI) (להלן: מערכות מבוססות בינה מלאכותית) בשירותים במימון ציבורי (שירותי רווחה, בריאות, דיור, ועוד). עם התרחבות השימוש במערכות אלו, ועקב הרצון למצות את תועלותיהן בשירותים במימון ציבורי, עולה הצורך להבטיח שהן יפעלו בהוגנות (fairness). הוגנות משקפת מחויבות לחלוקה צודקת של משאבים, המותאמת לצרכים הבסיסיים של הפרטים, לזכאותם לשירותים ולשאיפה לתקן פערים חברתיים ולהבטיח שוויון הזדמנויות. שירותים במימון ציבורי נועדו לענות על צרכים המוגדרים על ידי החברה כהכרחיים לקיום בכבוד, ובהתאם לכך המדינה נושאת באחריות העיקרית להבטחת נגישותם וזמינותם עבור כלל האזרחים (מנדלקרן ושרמן, 2015). מתוקף תפקידם הציבורי, שירותים אלו נתפסים לא רק כאמצעי טכני לחלוקת משאבים אלא כביטוי לערכים חברתיים ולמחויבות של המדינה כלפי אזרחיה. מכאן, הוגנות היא נדבך יסוד בשירותים הציבוריים, שכן היא מבטיחה גישה שוויונית ומונעת שעתוק פערים חברתיים.

אף ששימוש בהן עשוי לשפר את יעילות השירותים ואת נגישותם, ללא מנגנונים שונים שיסייעו בשמירה על הוגנות, מערכות מבוססות בינה מלאכותית עלולות לשעתק פערים חברתיים ולהעמיק אפליה בקרב אוכלוסיות פגיעות. מגמה זו מציבה אתגר מהותי בפני מפתחי ומטמיעי מערכות המבוססות בינה מלאכותית בשירותים במימון ציבורי: להבטיח שמערכות אלו יפעלו על פי עקרונות של שוויון ושקיפות, לא יפגעו בזכויות הפרט או יעמיקו פערים חברתיים, ישמשו מנוף לקידום שוויון חברתי וייטיבו עם כל הקבוצות באוכלוסייה, ובפרט עם הקבוצות הפגיעות ביותר. לשם כך פנה הג'וינט למכון מאיירס-ג'וינט-ברוקדייל בבקשה לערוך סקירה שתבחן את ההטיות והסיכונים הגלומים בשימוש במערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי ולמפות כיווני פעולה מעשיים לקידום הוגנות אלגוריתמית – עיקרון המתייחס לאופן שבו מערכות אלגוריתמיות, ובפרט מערכות מבוססות בינה מלאכותית, מקבלות או תומכות בקבלת החלטות המשפיעות על בני אדם. עיקרון זה מבקש להבטיח כי החלטות אלו אינן יוצרות או משמרות הטיות בלתי מוצדקות, אפליה או פגיעה בשוויון הזדמנויות בין פרטים וקבוצות באוכלוסייה.

המחקר נערך בשיתוף אוניברסיטת בן גוריון.

קובעי מדיניות ומקבלי החלטות נושאים מערכים שונים של ערכים ותפיסות עולם, והם משפיעים על האופן שבו הם מקבלים החלטות. בדומה לכך גם טכנולוגיות דיגיטליות, ובפרט מערכות מבוססות בינה מלאכותית, אינן ניטרליות מבחינה ערכית אלא משקפות ומעצבות יחסי כוח, דפוסי הכללה והדרה של קבוצות שונות באוכלוסייה. מחקר זה הוא חלק ממאמץ רחב יותר

של הג'וינט לפיתוח סביבה טכנולוגית מודעת חברתית הרואה בהוגנות במערכות מבוססות בינה מלאכותית חלק בלתי נפרד מחדשנות אחראית, מחויבות ציבורית לאוכלוסיות פגיעות ובניית אמון בין המדינה לאזרחיה.

גישת social inside, המתפתחת כיום בג'וינט, מבטאת מענה עקרוני לאתגר זה באמצעות מתן דגש לשילוב שיטתי של שיקולים חברתיים כבר בשלבי התכנון, העיצוב והפיתוח של טכנולוגיות ולא כתגובה מאוחרת לנזקים. גישה זו נשענת על עקרונות חברתיים מרכזיים, ובהם הוגנות, הכללה, זכויות דיגיטליות, מוגנות וקיימות, אשר נועדו להבטיח כי מערכות טכנולוגיות יפעלו באופן המצמצם פגיעה וימשו את הפוטנציאל החברתי החיובי שלהן. עקרונות אלו אינם מופרדים זה מזה, ויש ביניהם קשרים וחפיפות משגיות (הג'וינט וקרן מוריס וויואן וואהל, 2026). לפיכך בסקירה זו נעשתה בחירה מודעת לאמץ הגדרה רחבה ומכלילה של עקרון ההוגנות, כפי שיפורט בהמשך, הכוללת גם היבטים הנדונים לעיתים במסגרת עקרונות אלו.

מימוש הפוטנציאל של מערכות מבוססות בינה מלאכותית לשמש מנוע להפחתת הטיות ולחיזוק שוויון הזדמנויות מותנה בתכנון וביישום אחראיים: בהגדרה מוקדמת של מטרות ההוגנות, בפיתוח מומחיות ייעודית, בהתאמות הקשריות למציאות הישראלית המורכבת ובמעבר מניהול סיכונים עקרוני ליישום שיטתי של כלים ופרקטיקות להתמודדות עימם, תוך כדי ניהול סיכונים שקוף ומתמשך.

הסקירה מדגישה כי למערכות מבוססות בינה מלאכותית יש פוטנציאל לשמש "שובר שוויון" המאפשר תיקון הטיות, אולם שימוש לא מדויק עלול להעצימן. שכן, שלא כמו מערכות קבלת החלטות אנושיות, שבהן הטיות ואפליות מושרשות וקשה לשנותן, יכולתה של הבינה המלאכותית כמכונה לומדת מאפשרת לקדם שוויון חברתי באמצעות התאמת שירותים לצרכים מגוונים, חיזוק קבוצות מוחלשות ופיתוח כלים נגישים ומותאמים בשיתוף עם הקהילות עצמן.

מבנה הסקירה

בסקירה עשרה פרקים. **פרק 1** הוא המבוא, ובו מוצגים גם המושגים המקצועיים המרכזיים שבהם נעשה שימוש, **פרק 2** מפרט את המטרות **ובפרק 3** מוצגת השיטה. **פרק 4** מציג את נושא ההוגנות בשירותים במימון ציבורי. **פרק 5** מתאר את המציאות החדשה בעידן הבינה המלאכותית. עידן זה מתאפיין במהפכה מהירה וחסרת תקדים אשר לא רק מעצבת מחדש את האופן שבו מתקבלות החלטות מדיניות, אלא גם את האופן שבו מדיניות ושירותים במימון ציבורי מיושמים ומוטמעים בפועל. הפרק מציג כיצד מהפכה זו מאפשרת קבלת החלטות יעילה ומבוססת נתונים בהיקפים עצומים, אך בד בבד טמון בה פוטנציאל להעצמת הסיכון, לשעתוק ולהעמקת הטיות חברתיות קיימות. **פרק 6** סוקר מושגי יסוד בתחום הבינה המלאכותית ובוחן טכנולוגיות מרכזיות, סוגי מודלים שונים ותהליכי למידה ואימון שמאפשרים למודלים לפעול. **פרק 7** בוחן כיצד עקרונות ההוגנות באים לידי ביטוי במערכות מבוססות בינה מלאכותית. בפרק

מוצגות הגישות המרכזיות להגדרת הוגנות במערכות אלו, ובהן הוגנות של הפרט (individual fairness), הוגנות קבוצתית (group fairness), הוגנות נגד-עובדתית (counterfactual fairness) והוגנות מערכתית (system-level fairness), וכל אחת מהן מדגישה היבטים שונים של הוגנות ושוויון הזדמנויות. הגישות השונות להוגנות מתרגמות לדרכים שונות לצמצום הטיות והשפעות של גורמים שרירותיים.

פרק 8 מעמיק בהטיות של אלגוריתמים במודלים של בינה מלאכותית. הטיות אלו עלולות להיווצר, למשל, עקב מאפייני הנתונים שעליהם המודלים מאומנים או הבחירות שנעשות בתהליך הלמידה עצמו. בהמשך לכך, **פרק 9** דן בסיכונים אפשריים עקב הטיות במודלים של בינה מלאכותית ומתאר את ההשפעות האפשריות של ההטיות על ביצועי המודלים ואת התוצאות האפשריות של השפעות אלו על הקבוצות באוכלוסייה שכלפיהן המודל מוטה. **פרק 10** מציג פתרונות אפשריים לקידום הוגנות בשימוש במערכות מבוססות בינה מלאכותית. הפרק מציג מסגרת דו-ממדית לקידום הוגנות אלגוריתמית: בחינה של פתרונות לקידום הוגנות אלגוריתמית הנדרשים בכל שלבי פיתוח המערכת, מתוך הבנה שכל שלב מייצר סוגים שונים של סיכונים והזדמנויות לשיפור הוגנות. מנקודת מבט זו, מענים הקשורים לתהליך יצירת המודל ולהערכתו נבחנים כחלק מתהליך מתמשך הנדרש להתעדכן ולשקף את השינויים במערכת הבינה המלאכותית ובסביבת הפעולה שלה. מנגד מוצגים פתרונות המתמקדים במעטפת השירות הציבורי, הכוללת מנגנונים תהליכיים, אנושיים וארגוניים, הפועלים בצד המערכת האלגוריתמית ומשלימים את המענים הקשורים לתהליך יצירת המודל ולהערכתו. שילוב שני הממדים נדרש משום שהוגנות אינה תוצר של המודל בלבד אלא של האופן שבו הוא מפורש, מוטמע ומופעל בארגון. מעטפת השירות הציבורי מאפשרת לווסת שימוש, לצמצם הטיות הנוצרות בממשק עם גורמים אנושיים ולהבטיח שהמערכת משרתת מגוון קבוצות באוכלוסייה באופן שוויוני. שני הממדים יחד יוצרים תשתית כוללת לתכנון וליישום אחראי של מערכות מבוססות בינה מלאכותית בשירות הציבורי. **פרק 11** מסכם את התובנות המרכזיות העולות מן הסקירה.

מושגים מרכזיים

בסקירה נעשה שימוש במושגים מתחום מדעי המחשב. **לוח 1** להלן מבאר מושגים אלו.

לוח 1: מונחון מושגים

מושג	הסבר
בינה מלאכותית (AI)	תחום במדעי המחשב המתמקד בפיתוח מערכות המסוגלות לבצע משימות שנתפסות ככאלה הדורשות אינטליגנציה אנושית, כמו זיהוי תבניות, עיבוד שפה טבעית וקבלת החלטות
למידת מכונה (ML)	תת-תחום בבינה מלאכותית המאפשר למערכות ללמוד ישירות מנתונים, לזהות בהם דפוסים ולבצע תחזיות או החלטות על נתונים חדשים, במקום להיות מתוכנתות לפי כללים קבועים מראש
הוגנות אלגוריתמית	עיקרון המתייחס לאופן שבו מערכות אלגוריתמיות, ובפרט מערכות מבוססות בינה מלאכותית, מקבלות החלטות המשפיעות על בני אדם או תומכות בהן. עיקרון זה מבקש להבטיח כי החלטות אלו אינן יוצרות או משמרות הטיות בלתי מוצדקות, אפליה או פגיעה בשוויון הזדמנויות בין פרטים וקבוצות באוכלוסייה.
קלט (input)	נתונים או מידע חדש המוזן למודל למידת מכונה בשלב השימוש בו, שעליהם המודל נדרש לבצע פעולה או לתת תשובה
פלט (output)	התוצאה, התשובה, החיזוי או התוכן שהמודל מפיק בתגובה לקלט שקיבל בשלב השימוש
מודל של למידת מכונה	ה"מודל" הוא הייצוג המתמטי או האלגוריתמי שנוצר. הוא מכיל את הידע והכללים שהמערכת רכשה מנתוני אימון ומאפשר לה לבצע חיזויים, סיווגים או יצירת תוכן על בסיס קלט חדש
אימון מודל (training) או למידה	התהליך שבו מודל למידת מכונה נבנה ו"לומד" מתוך כמות גדולה של נתונים קיימים. במהלך האימון, אלגוריתם ייעודי מכוון את המודל לזהות דפוסים וקשרים כדי שיצליח לבצע את המשימה שהוגדרה לו
למידה מונחית (supervised learning)	שיטה שבה המודל מקבל דוגמאות של קלט ופלט רצוי ומנסה ללמוד את הקשר בין הקלט לפלט כך שישקף את הדפוסים שמופיעים בדוגמאות
למידה בלתי מונחית (unsupervised learning)	שיטה שבה המודל לומד לזהות דפוסים בנתונים, ללא פלט ידוע מראש

מושג	הסבר
ערך אמת או תיוג (label)	הערך הידוע מראש המשמש בסיס להשוואה בעת הערכת תחזיות של מודל. זהו "התיוג הנכון" של הדוגמה, ולעומתו נבחן דיוק התחזית של המודל. למשל, בסיווג תמונות של חתולים וכלבים, ערך האמת עבור כל תמונה יהיה תיוג שמציין אם התמונה היא של חתול או של כלב
פונקציית הפסד (loss function)	מדד כמותי לפער שבין התחזית ובין הערך האמיתי, אשר מעניק "ציון טעות" או "קנס" על תחזיות שגויות ומשמש להכוונת תהליך הלמידה של המודל
מודלים גנרטיביים (generative models)	סוג של מודלי למידת מכונה המסוגלים ליצור תוכן חדש (טקסט, תמונות, וידיאו) הנראה דומה לנתונים שמהם למדו ואינם מוגבלים לקבוצת פלטים מוגדרת מראש
מודלי שפה גדולים (Large Language Models – LLM)	סוג ספציפי של מודלים גנרטיביים המאומנים על כמויות אדירות של טקסט ומסוגלים להבין, ליצור ולעבד שפה טבעית

2. מטרת הסקירה

לגבש הבנה שיטתית של האתגרים והסיכונים הקשורים להוגנות ושוויון הזדמנויות בשימוש במודלים של בינה מלאכותית בשירותים במימון ציבורי ולהציג מארג של פתרונות מעשיים שמאפשרים למשרדי הממשלה ולגופים ציבוריים לתכנן, לפתח ולהטמיע מערכות מבוססות בינה מלאכותית באופן שיצמצם הטיות ויקטין סיכונים של פגיעה בהוגנות.

שאלות המחקר הן:

1. מהם סוגי ההטיות? כיצד כל סוג עלול להוביל לחוסר הוגנות בשימוש במודלים של בינה מלאכותית?
2. מהם הסיכונים בהטיות במודלים של בינה מלאכותית ובשימוש בהם?
3. אילו פתרונות אפשר להטמיע במטרה לקדם הוגנות אלגוריתמית?

3. השיטה

הסקירה מתבססת על מגוון מקורות מידע. המידע נאסף בחודשים מארס-דצמבר 2025 וכלל:

- **מאמרים אקדמיים מן הספרות המקצועית הישראלית והבין-לאומית**: המאמרים אותרו במנועי החיפוש Google Scholar ו-Elicit באמצעות שימוש במילות חיפוש הרלוונטיות למחקר:

  "AI in public services", "algorithmic bias", "AI fairness", "fairness in public services", "equity in public services artificial intelligence", "labeling bias", "bias and missing data", "fairness in the medical setting", "representation bias in large language models", "stereotypes in large language models", "social bias in text to image generative models", "equity in public services"

  "בינה מלאכותית בשירותים ציבוריים", "הוגנות בשירותים ציבוריים וחברתיים".

- **ספרות אפורה**: דוחות מחקר, מסמכים רשמיים של משרדי ממשלה וכתבות בעיתונות.
- **ראיונות עומק חצי מובנים**: נערכו שלושה ראיונות מקדימים מקוונים עם מומחים בתחום הבינה המלאכותית בישראל (בחודשים מאי עד יולי 2025).

4. הוגנות במדיניות חברתית ובשירותים במימון ציבורי

שירותים במימון ציבורי נועדו לענות על צרכים המוגדרים על ידי החברה כהכרחיים לקיום בכבוד. לפיכך המדינה נושאת באחריות המרכזית להבטחת נגישותם וזמינותם עבור כלל האזרחים. אף שהתשובה לשאלה מהם אותם צרכים בסיסיים אינה חד-משמעית ומשתנה לפי תקופה, תרבות והקשר פוליטי, יש הסכמה רחבה כי בגדר צרכים אלו מצויים תחומים כגון חינוך, בריאות, רווחה, דיור, תעסוקה וכן ביטחון אישי ואכיפת חוק (מנדלקרן ושרמן, 2015). מתוקף תפקידם הציבורי, שירותים אלו נתפסים לא רק כאמצעי טכני לחלוקת משאבים אלא כביטוי לערכים חברתיים ולמחויבות של המדינה כלפי אזרחיה.

ערכים חברתיים אלו באים לידי ביטוי באופן שונה בכל מדינה, על פי המסורת הפוליטית, המבנה המוסדי ותפיסת תפקיד המדינה באחריות כלפי אזרחיה. כלומר, הוגנות בשירותים במימון ציבורי איננה רק עניין של חלוקה טכנית שוויונית אלא תוצר של הבניה פוליטית-נורמטיבית של צרכים בסיסיים, של זכאות לשירותים אלו ושל קריטריונים לחלוקה צודקת של משאבים. גם תפיסות חברתיות ותרבותיות משפיעות על האופן שבו נתפסים מושגים כמו אפליה, שוויון הזדמנויות וצדק חברתי. כך למשל, במדינות מסוימות עשויות הבחנות בין קבוצות, למשל בין נשים לגברים, להיחשב לגיטימיות או מתבקשות במסגרת ערכים תרבותיים מקובלים, ואילו במדינות אחרות אותן הבחנות ייתפסו כהפרה של עקרונות הוגנות בסיסיים. דוגמה נוספת, במדינות סקנדינביות המבוססות על מודל מדינת הרווחה הסוציאל-דמוקרטית, כמו שוודיה, הוגנות מתורגמת לשוויון הזדמנויות רחב ככל האפשר ולנגישות מקיפה לשירותים עבור כלל האוכלוסייה, תוך הבטחת תנאי פתיחה דומים (Meagher & Szebehely, 2019; Songur, 2023). מנגד, במדינות רווחה ליברליות, כמו אוסטרליה או בריטניה, תפיסת ההוגנות שמה דגש רב יותר באחריות אישית ובזכאות מותנית, למשל, על פי מצב התעסוקה של הפרט (תרשיש, 2017).

נוכח השילוב הגובר של מערכות מבוססות בינה מלאכותית והצורך למצות את תועלותיהן בשירותים במימון ציבורי, עולה השאלה כיצד אפשר להבטיח שמערכות אלו יפעלו בהוגנות. הוגנות אלגוריתמית בבינה מלאכותית (ראו מונחון מושגים) היא עיקרון שלפיו מערכות בינה מלאכותית צריכות לקבל או לתמוך בהחלטות באופן שאינו יוצר או משמר הטיות, אפליה או פגיעה בשוויון הזדמנויות בין פרטים וקבוצות (ראו מונחון מושגים), אם כן, הוגנות אלגוריתמית אינה רק אתגר טכני. הטמעת כלים טכנולוגיים לקבלת החלטות בשירותים במימון ציבורי יוצרת הזדמנות לשיפור יעילות ונגישות, לשיפור שוויון במתן שירותים, להקטנת פערים חברתיים ועוד. אך בהיעדר מנגנונים ברורים להבטחת הוגנות ושוויון הזדמנויות ,יש סכנה להעמקת פערים קיימים ולהנצחת אפליה מבנית. נוכח חשיבות זו, פרק זה מתמקד בהוגנות ובשוויון הזדמנויות כעקרונות יסוד באתיקה של בינה מלאכותית ובגישות שונות בהוגנות אלגוריתמית ובבינה מלאכותית.

הדיון בהוגנות במערכות מבוססות בינה מלאכותית הוא חלק בלתי נפרד ממסגרת רחבה יותר של אתיקה בתחומים טכנולוגיים. אתיקה בתחומים אלו כוללת כמה עקרונות יסוד, כאשר הוגנות ושוויון הזדמנויות נחשבים עקרונות מרכזיים, בצד עקרונות נוספים, כגון שקיפות, הסברתיות (explainability), אחריותיות, מוגנות וכיבוד פרטיות (Jobin et al., 2019; Floridi et al., 2018; World Health Organization [WHO], 2023). עקרונות אלו שואפים להבטיח שמערכות מבוססות בינה מלאכותית יפעלו באופן מוסרי, אמין וראוי לאמון הציבור. הוגנות, בהקשר זה, איננה רק שאלה של צדק חברתי אלא גם תנאי חשוב ללגיטימציה ציבורית של מערכות המבוססות בינה מלאכותית, במיוחד כאשר הן משולבות בקבלת החלטות בתחומים הקשורים לשירותים במימון ציבורי, כמו בריאות, חינוך או משפט (Jobin et al., 2019).

לביא (Lavee, 2021) טוען כי בעת קביעת זכאות לשירותים במימון ציבורי, מקבלי ההחלטות נוטים להישען על ערכים אישיים באופן היוצר מידה רבה של אקראיות בהחלטותיהם. משום כך שילוב של בינה מלאכותית בתהליכי קבלת החלטות, תוך כדי הטמעת עקרונות של הוגנות ושוויון הזדמנויות, עשוי לצמצם את האקראיות ולהבטיח כי השירותים יוענקו למי שזכאים להם בפועל. עם זאת יש לזכור כי גם האלגוריתמים עצמם מעוצבים בידי בני אדם הנושאים ערכים אישיים, ולכן שילוב בינה מלאכותית לא יבטיח בהכרח שיפור בהוגנות אלא אם תוטמע בהם מודעות להטיות קיימות ויישום עקרונות הוגנות. עם זאת גם טרם הטמעת מערכות מבוססות בינה מלאכותית לא התקיימה ניטרליות בקרב מקבלי ההחלטות. לפיכך שימוש מושכל בבינה מלאכותית עשוי לא רק למנוע הנצחת הטיות קיימות אלא אף לצמצמן ולשפר את רמת ההוגנות.

אף שבני אדם מקבלים לעיתים החלטות הנגועות בהטיות או באפליה, השימוש במערכות מבוססות בינה מלאכותית מציב אתגרים ייחודים בכל הנוגע לתפיסת ההוגנות. ראשית, הטיות אנושיות נתפסות חלק מתהליך שיפוט סובייקטיבי ומורכב, ואילו ממערכות אלגוריתמיות מצפים לפעול בעקביות ובשיטתיות. על כן כאשר מתגלות במערכות אלו תוצאות מפלות או לא שוויוניות, הן נתפסות כפוגעות פגיעה עמוקה יותר בעקרון ההוגנות, משום שהן סותרות את ההנחה הבסיסית בדבר ניטרליות ואובייקטיביות טכנולוגית. שנית, שלא כמו החלטות אנושיות, המתקבלות על ידי גורמים רבים ובאופנים מגוונים, מערכות מבוססות בינה מלאכותית נוטות לרכז את מוקד קבלת ההחלטות. ריכוז זה עלול להפוך הטיות נקודתיות להטיות שיטתיות המיושמות יישום רחב היקף. נוסף על כך החלטות אנושיות ניתנות לעיתים לביקורת או לערעור באמצעות שיקול דעת מוסרי, ואילו החלטות אלגוריתמיות מתקבלות פעמים רבות באופן שאינו שקוף, מה שמקשה על הבנת תהליכי קבלת ההחלטות ועל תיקון טעויות אפשריות. בכך מתחדד הצורך להגדיר מחדש את מושג ההוגנות בהקשרים טכנולוגיים ולבחון כיצד אפשר ליישמו בשירותים במימון ציבורי.

נוכח אתגרים אלו, שאלת ההוגנות ושוויון ההזדמנויות אינה שולית אלא מהותית, והיא מגדירה את האופן שבו שירותים אלו מסופקים: מי זכאי להם, באילו תנאים, ומה נחשב לטיפול שוויוני או צודק – בעיקר בשירותים במימון ציבורי שבהם שאלות אלו טעונות במיוחד בשל אחריות

המדינה כלפי אזרחיה. להוגנות בשירותים במימון ציבורי חשיבות יתרה בהשוואה לשירותים פרטיים, בשל תפקידה המרכזי של המדינה בהבטחת זכויות בסיסיות ושוויון הזדמנויות לכלל האזרחים. בשירותים פרטיים ההוגנות נתפסת לעיתים בתור רכיב של תחרות הוגנת או של אחריות אישית של הפרט לבחור ולהתחרות על משאבים מוגבלים, ואילו בשירותים במימון ציבורי ההוגנות היא יסוד ללגיטימיות של מוסדות המדינה ולמימוש עקרונות של צדק חברתי. שירותים במימון ציבורי נועדו לספק מענה לצרכים הכרחיים לקיום בכבוד ולכן נדרשת בהם רגישות מיוחדת להבטחת נגישות, זמינות והתאמה לכלל האוכלוסיות, במיוחד לקבוצות מוחלשות.

כאשר מערכות מבוססות בינה מלאכותית משתלבות בתהליכי קבלת ההחלטות הן עשויות לעצב בפועל את גבולות הזכאות והנגישות לשירותים במימון ציבורי. לפיכך דיון בהוגנות אלגוריתמית בשירותים במימון ציבורי מחייב הבנה עמוקה של עקרונות הצדק החברתי והאתיקה המוסדית המלווים את התחום הזה. מערכת בינה מלאכותית המוטמעת בשירותים במימון ציבורי חייבת, לפיכך, להבטיח הוגנות לא רק כתוצאה רצויה אלא כתנאי מקדים לקיומם וללגיטימיות של השירותים עצמם כצודקים ומוסריים.

5. מהפכת הבינה המלאכותית והשפעותיה על הוגנות

העולם עובר בשנים האחרונות מהפכה חסרת תקדים בתחומי הידע והמידע, המכונה לעיתים "העידן הדיגיטלי" וכעת "עידן ה-AI". מהפכה זו מונעת על ידי התפתחות טכנולוגיות הבינה המלאכותית בקצב מסחרר, והיא משנה את הדרך שבה אנשים ניגשים למידע, מנתחים אותו ומקבלים החלטות. קודם לכן המידע היה מוגבל יותר, מפוזר ובעיקר נגיש דרך גורמים אנושיים ותהליכים ידניים. גם היכולות הטכנולוגיות היו מוגבלות ואף שבינה מלאכותית הייתה קיימת כבר, השתמשו בה בעיקר מומחים בתחום. כיום לעומת זאת מודלים של בינה מלאכותית הפכו לכלי מרכזי כמעט בכל תחומי החיים והובילו למהפכה באופן שבו מידע נאסף, מנותח ומנוצל – במהירות, בקנה מידה עצום, אוטומטית ובגישה רחבה, גם לציבור ולא רק לגורמים מקצועיים. השינוי הוא איכותי לא פחות מכמותי: מידע משמש כיום לא רק לתיעוד אלא גם לחיזוי, להתאמה אישית ולתמיכה בקבלת החלטות בזמן אמת, ברמת הפרט, הארגון והמדינה (Iansiti & Lakhani, 2020).

השינוי יצר פער ניכר בין היכולות הטכנולוגיות והיקפי המידע של העידן הנוכחי ובין השיטות והגישות המסורתיות שפעלו בסביבה יציבה ואיטית יותר. השיטות שנבנו על בסיס מידע מועט ובקצב איטי מתקשות להתמודד עם מציאות המאופיינת בכמויות גדולות של מידע (big data), בשינויים תכופים ובבעיות מורכבות. תהליכי קבלת החלטות וניהול משאבים שנסמכו בעבר על ניסיון אנושי, כללים קבועים או תחזיות פשוטות, מתקשים כיום לספק מענה מדויק, מהיר ומותאם הקשר. יתרה מכך, הדרישה להתאמה אישית של שירותים ומדיניות הנובעת מהסתגלות הציבור לחוויות מותאמות אישית בעולם הדיגיטלי, כמו רשתות חברתיות, מסחר מקוון ושירותי ניווט, מחייבת כלים גמישים ודינמיים, ואלו יכולות שמודלים מסורתיים לרוב אינם מספקים. לפיכך אימוץ מערכות מבוססות בינה מלאכותית אינו רק שדרוג טכנולוגי אלא תגובה חיונית לצרכים המשתנים של עידן הבינה המלאכותית.

התקדמות טכנולוגית זו אינה חפה מאתגרים. מהפכת הבינה המלאכותית עלולה להציף ולהדגיש הטיות חברתיות ואנושיות שאולי היו קיימות גם קודם לכן ואף להעצימן. קצב הפיתוח וההטמעה המהיר של מערכות מבוססות בינה מלאכותית, בצד מורכבותן הגוברת, יוצר אתגר ביכולת להבין, לנטר ולרסן את השפעות השימוש בהן. אתגר זה אינו נובע מן הקצב הטכנולוגי בלבד, אלא גם מן העובדה שמערכות אלו לעיתים מוטמעות במטרה להגביר את קצב העבודה ולהפחית את התלות בגורמי מקצוע אנושיים, באופן שמצמצם בפועל את יכולת הבקרה והניטור ומגביר את הסיכון להקצנה של דפוסי אפליה. שינויים טכנולוגיים אלו מתרחשים במהירות רבה יותר מיכולתן של מסגרות רגולטוריות להגיב ולהתאים את עצמן אליהם (Crawford, 2021). פער זה עלול לאפשר למערכות אלגוריתמיות לפעול ללא מנגנוני

בקרה והכוונה מספקים, והדבר העלול להוביל, למשל, לכך שאלגוריתמים ישכפלו דפוסי אפליה היסטוריים מנתוני עבר ויחילו אותם בקנה מידה רחב או יצרו תהליכים המגבירים את ההטיה, שבהם החלטות מוטות מובילות לנתונים מוטים נוספים אשר מחזקים את ההטיה המקורית (O'Neil, 2016). הבנה זו חיונית במיוחד כאשר מיישמים מערכות אלגוריתמיות בכל תחום המערב קבלת החלטות הנוגעות לבני אדם.

על כן הכרחית בחינה מעמיקה של גורמים אלו בכל תהליך הטמעה של מערכות מבוססות בינה מלאכותית במטרה להבטיח יישום אחראי. בתוך כך יש להכיר בכך שהוגנות אינה רק תוצאה של עיצוב טכני נכון של המודל, אלא גם של בחירות נורמטיביות ופוליטיות המוטמעות בתשתיות ובמערכות חברתיות רחבות יותר (Selbst et al., 2019). הדגש על הוגנות מחייב שינוי פרדיגמה בדרך שבה אנו ניגשים לפיתוח והטמעה של מערכות מבוססות בינה מלאכותית במקום להתמקד אך ורק ביעילות טכנית ובביצועים, יש להעביר את הדגש להשפעות החברתיות והאתיות של המערכות. המשמעות היא שנדרשת גישה אחראית, החל משלב איסוף הנתונים, דרך תכנון האלגוריתמים ועד להטמעת המערכת. בתהליך זה נדרשת חשיבה רחבה יותר המשלבת מומחים מדיסציפלינות שונות, כדי להבטיח שהטכנולוגיות המתפתחות ישרתו את טובת הכלל, יקדמו שוויון ולא יחזקו אי-שוויון והטיות קיימות בחברה (High-Level Expert Group on Artificial Intelligence, 2019).

6. בינה מלאכותית ולמידת מכונה: מושגי יסוד

6.1 מהי בינה מלאכותית?

בינה מלאכותית היא תחום במדעי המחשב המתמקד בפיתוח מערכות המסוגלות לבצע משימות המצריכות בדרך כלל אינטליגנציה אנושית. תחום זה כולל מגוון רחב של יכולות, כגון זיהוי תבניות בתמונות (למשל, זיהוי פנים ביישומוני צילום), עיבוד והבנה של שפה טבעית (כמו תרגום אוטומטי) וקבלת החלטות מורכבות (לדוגמה, מערכת למתן המלצות פיננסיות) (Russell & Norvig, 2021).

6.2 למידת מכונה

אחד מתת-התחומים הבולטים והמשפיעים בתחום זה הוא למידת מכונה (Machine Learning - ML) (ראו מונחון מונחים). שלא כמו מערכות אלגוריתמיות מסורתיות שבהן מתכנתים נהגו להגדיר ידנית ובמפורש כללים והוראות לביצוע משימות, למידת מכונה מאפשרת למערכות ללמוד ישירות מנתונים. מודלים של למידת מכונה מנתחים כמויות אדירות של נתונים (למשל תמונות, טקסטים, נתונים בטבלה), מזהים בהם דפוסים, קשרים ומבנים ויוצרים מודל המאפשר לבצע תחזיות, סיווגים או החלטות על נתונים חדשים שלא נצפו בעבר. מודלים של למידת מכונה מתפתחים ומשתפרים באמצעות תהליך הולך וחוזר של למידה וככל שהם נחשפים לנתונים נוספים, שלא כמו אלגוריתמים מסורתיים שהיו סטטיים ודרשו עדכון ידני של הכללים (Alpaydin, 2020).

6.3 סוגי מודלים של למידת מכונה

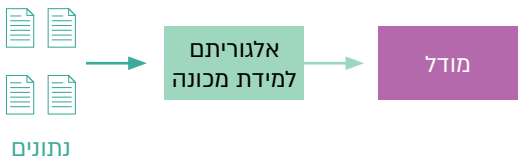
במסגרת תחומי הבינה המלאכותית ולמידת המכונה אפשר להבחין בין סוגים שונים של מודלים הנבדלים זה מזה במטרה וביכולות. אחד מסוגי המודלים הם **מודלים המיועדים למשימות ספציפיות** (מודלי סיווג ותחזית). מודלים אלו מבצעים משימה המוגדרת היטב על ידי מפתחי המודל, והם תמיד מספקים תשובה מתוך סט של תשובות שהוגדר מראש. במודלים אלו הקלט (ראו מונחון מושגים) הוא המידע שהמודל מקבל (למשל, תמונה), והפלט (ראו מונחון מושגים) הוא התשובה שהוא מחזיר מתוך סט התשובות. מודל לזיהוי תמונות, לדוגמה, יקבל תמונה ויגיד אם יש בה "חתול", "כלב" או "אדם" – רק מתוך הקטגוריות שהוא מכיר. מודל אחר, שמטרתו להעריך סיכון רפואי למשל, יקבל מידע על אדם (הקלט) ויחזיר מספר בין 0 ל-1, שמייצג את הסיכון שלו לקבל התקף לב בחמש השנים הבאות (הפלט). על פי רוב מודלים אלו לא יכולים לבצע משימה שונה מזו שעבורה הם אומנו. כלומר אם מודל אומן לזהות חתולים וכלבים, אי אפשר להוסיף לו קטגוריה חדשה כמו "ציפורים" בלי לאמן אותו מחדש. מכיוון שהתשובות של המודלים קבועות מראש, קל יחסית לבדוק עד כמה הם טובים במילוי המשימה שלהם.

שלא כמו מודלים המיועדים למשימות ספציפיות, **מודלים גנרטיביים** לא מחזירים ערך מתוך רשימה מוגדרת מראש של אפשרויות, והמטרה שלהם היא ליצור תוכן (טקסט, תמונה או וידיאו) הנראה דומה לתוכן שממנו הם למדו (ראו מונחון מושגים). הם מזהים תבניות בתוכן שלמדו, והם מסוגלים ליצור תמונה הדומה לתמונות שהיו עשויות להצטלם במצלמה או להיווצר באופן מקצועי או ליצור טקסט הדומה לטקסטים שנכתבו על ידי בני אדם. מודלים אלו גם מסוגלים להכליל, כלומר הם אינם רק משחזרים את מה שלמדו אלא מבינים את העקרונות והחוקים הבסיסיים המגדירים את הנתונים, והם יכולים ליצור על פיהם יצירות חדשות לחלוטין שלא נכללו בנתונים המקוריים. השימוש במודלים גנרטיביים של למידת מכונה התרחב מאוד בשנים האחרונות, וההבנה התיאורטית המדעית (הכוללת הבנה מתמטית ופורמלית) הניצבת בבסיסם מקיפה פחות מן ההבנה שבבסיס מודלים המיועדים למשימות ספציפיות (Goodfellow et al., 2016).

6.4 תהליך האימון: מודלים של למידת מכונה

מודלים של למידת מכונה (ראו מונחון מושגים) פועלים על בסיס למידה מכמות גדולה של נתונים. בתהליך שנקרא אימון מודל (ראו מונחון מושגים), אלגוריתם ייעודי מכוון את המודל לזהות דפוסים וקשרים כדי שיצליח לבצע את המשימה שהוגדרה לו (ראו **תרשים 1**). מטרתו של אלגוריתם הלמידה היא לבחור מתוך מגוון רחב של מודלים אפשריים ולבנות את המודל המתאים ביותר. לאחר שהמודל אומן והושלם, הוא מוכן ליישום בפועל: כאשר מוזן למודל קלט כלשהו (נתונים חדשים שעבורם רוצים לקבל תשובה), המודל מעבד את הקלט הזה, ובתמורה מתקבל פלט (התוצאה, התשובה או החיזוי המבוקש) (ראו **תרשים 2**) (Alpaydin, 2020). לדוגמה, במודלים גנרטיביים לייצור תמונות, נתוני האימון הם מספר עצום של תמונות קיימות; בשלב השימוש, המודל יקבל כקלט "פרומפט" (תיאור טקסטואלי של התמונה הרצויה) ויוציא כפלט תמונה חדשה (Foster, 2023). לעומת זאת במודלים להערכת סיכון למחלה, נתוני האימון יהיו למשל תיקים רפואיים של חולים רבים. בשלב השימוש, המודל יקבל כקלט תיק רפואי של חולה ספציפי ויוציא כפלט מספר שהוא הערכת הסיכון לאותה מחלה.

תרשים 1: אימון מודל



תרשים 2: יישום מודל

6.5 דרכי הלמידה: מודלים של למידת מכונה

יש שתי דרכים עיקריות לאימון מודלים של למידת מכונה: למידה מונחית (supervised learning) ולמידה בלתי מונחית (unsupervised learning) (ראו מונחון מושגים). **בלמידה מונחית**, אלגוריתם הלמידה מקבל זוגות של נתונים: קלט ופלט מתאים (מה שמכונה גם "ערך האמת", ראו מונחון מושגים). המודל לומד מתוך דוגמאות מתויגות אלו ומכליל את הידע לצורך ביצוע אוטומטי של המשימה על נתונים חדשים. לדוגמה, אלגוריתם הלומד לזהות תמונות של חתולים יאומן על תמונות שמתויגות מראש כ"כן" (חתול) או "לא" (לא חתול). בתהליך האימון, המודל מבצע ניחוש או חיזוי בנוגע לתוכן בתמונה. את מידת הטעות של המודל, כלומר עד כמה הניחוש שלו היה רחוק מן התשובה הנכונה, מחשבים באמצעות מדד שמעריך את "ציון הטעות" או את ה"קנס" של המודל עבור טעותו; עד כמה התחזיות של המודל שגויות ביחס לתוצאות האמיתיות (פונקציית הפסד, ראו מונחון מושגים). מטרת המודל בתהליך הלמידה היא למזער את ה"קנס" הזה, כלומר להפוך את הטעויות שלו לקטנות ככל האפשר. כך, אם המודל זיהה בתמונה כלב אך בפועל יש בה חתול, הוא "ייענש" על טעותו ויבצע התאמות פנימיות כדי לשפר את יכולת הסיווג שלו לפעם הבאה. **בלמידה בלתי מונחית**, המודלים לומדים ממידע שאינו מגיע יחד עם "ערך אמת" (ראו מונחון מושגים) או פלט רצוי מוגדר מראש, אלא לומדים את המאפיינים של הנתונים עצמם. גם בלמידה לא מונחית יש, בדרך כלל, ניסיון למזער "ציון הטעות" והמטרה היא למצוא מודל שימזער ציון זה (Alpaydin, 2020).

6.6 מודלים מתקדמים

נוסף על המודלים שפורטו לעיל, יש מודלים המשלבים שלבים נוספים של למידה מונחית כדי לשפר ביצועים ולמנוע תוצאות לא רצויות. לדוגמה, **מודלי שפה גדולים** (ראו מונחון מושגים) (Large Language Models - LLMs) נלמדים בכמה שלבים: בשלב הראשון הם לומדים להשלים את המילה הבאה במשפט מתוך כמות אדירה של טקסטים. בשלבים מתקדמים יותר האלגוריתמים עוברים אימון נוסף הכולל למידה מונחית שבה נותנים להם דוגמאות של תשובות רצויות ובלתי רצויות (למשל, תשובות אלימות או לא הולמות), כדי להטמיע נורמות התנהגות מקובלות (alignment) (Zhao et al., 2023). מודלי שפה גדולים משמשים כיום למגוון רחב של מטרות, למשל ליצירת תוכן, לתרגום שפות, להפעלת צ'אטבוטים, לניתוח טקסט ולשיפור מנועי חיפוש באמצעות הבנת שאילתות מורכבות. **מודלים המייצרים תמונות**, לעומת זאת, פועלים לרוב באמצעות תהליך שבו הם מתחילים מ"רעש אקראי" ומבצעים פעולות מתמשכות להפחתת ה"רעש" עד ליצירת תמונה ברורה וקוהרנטית. מודלים ליצירת תמונות מאפשרים, למשל, יצירת אומנות דיגיטלית, סיוע בעיצוב גרפי ויצירת תוכן ויזואלי לסרטים וקמפיינים שיווקיים.

6.7 אופן השימוש במודלים

התפתחותם של מודלים מבוססי בינה מלאכותית ולמידת מכונה פתחה מגוון רחב של אופני שימוש במודלים. אפשר לחלק את אופני השימוש לארבע קטגוריות מרכזיות הנבדלות זו מזו בהיבטים האלה: מידת העצמאות של המערכת אל מול רמת המעורבות האנושית בתהליך; ומטרת העל של המודל – האם הוא נועד לקבלת החלטות עצמאית, לסייע לאדם, לייעל תהליכים או ליצור תוכן חדש. כמה מן השימושים מכוונים להחלפת פעולות אנושיות מסוימות, ואילו אחרים נועדו להעצים את היכולות האנושיות, להשלים אותן או לאפשר יצירה מסוג חדש. אנחנו מסייגים שחלוקה זו אינה חד-משמעית, ויש מקרים שמטרת העל של המודל משתייכת ליותר מקטגוריה אחת.

קבלת החלטות אוטומטית באמצעות אלגוריתמים. בתהליך זה אלגוריתמים מנתחים כמויות עצומות של נתונים, מזהים דפוסים ופועלים על פי כללים שהוגדרו או נלמדו, במטרה לקבל החלטות באופן עצמאי. לדוגמה, אוניברסיטאות ומכללות רבות בעולם משתמשות במודלים של בינה מלאכותית כדי לסנן ולמיין מועמדים לתוכניות לימודים (Intelligent, 2023). מודלים אלו יכולים לנתח במהירות נתונים כגון ציונים, ממוצעי בגרויות או תארים, ציון במבחנים סטנדרטיים (כמו פסיכומטרי), המלצות ואפילו לזהות דפוסים בחיבורים אישיים. הם יכולים להחליט אוטומטית אם מועמד עומד בתנאי הסף או להחליט אם לקבל או לדחות אותו על בסיס קריטריונים מוגדרים מראש ודפוסים שנלמדו ממועמדים קודמים.

תמיכה בהחלטות מקצועיות. שלא כמו קבלת החלטות עצמאית, מודל זה משמש כלי עזר למומחים ולאנשי מקצוע. הוא מספק ניתוחים, המלצות והערכות סיכונים מבוססות נתונים, אך ההחלטה הסופית נותרת בידי האדם המשלב בין המידע שהוא מקבל ובין שיקול דעתו, ניסיונו ותחום התמחותו. לדוגמה, במרכז הרפואי רבין (בילינסון) בישראל פותח פרויקט "בילינסון NEXT" ובו משתמשים במערכות מבוססות בינה מלאכותית לשיפור האבחנה הרפואית עבור צרכים שונים (מרכז רפואי רבין, א"ת). מערכת כזו היא, למשל, מערכת לזיהוי גידולים בממוגרפיה המבוססת על טכנולוגיית בינה מלאכותית של חברת iCAD, המאפשרת זיהוי גידולים זעירים ומדידת סיכון לסרטן השד. המערכת מספקת לרופאים המלצה ראשונית בנוגע לסיכון, אך ההחלטה הסופית על האבחון והטיפול נשארת בידי הרופא, המשלב בין המלצת המערכת ובין ניסיונו ושיקול דעתו הקליני.

סיוע לבעלי מקצוע בביצוע משימות. שימוש זה מאפשר למודלים לבצע משימות או שלבים ספציפיים בתהליך עבודה רחב יותר, תוך כדי שמירה על מעורבות אנושית בשלבים מכריעים. במקום להחליף לחלוטין את קבלת ההחלטות האנושית, המודלים משמשים "עוזרים חכמים" המייעלים את העבודה ומאפשרים לעובדים להתמקד בשיקול דעת מורכב. דוגמה לכך היא פרויקט Redbox של משרד המדע, החדשנות והטכנולוגיה (DSIT) בבריטניה. מדובר במערכת שמאפשרת הזנת מסמכים וקבלת סיכומים מיידיים, טיוטות לתשובות או ריכוז עיקרי דברים.

מערכת זו אינה מיועדת להחליף את כלל עובדי המגזר הציבורי אלא משרתת ספציפית את דרג הפקידות הבכיר, יועצי מדיניות וצוותי מטה במשרדי הממשלה. המערכת משמשת "עוזר פרלמנטרי חכם" המשחרר את הצוותים המקצועיים מ"עבודת נמלים" מנהלית ומאפשר להם להקדיש את זמנם לניתוח אסטרטגי של המידע ולהכנת המלצות מדיניות איכותיות עבור השרים. העובדים הם אלו שבוחנים את תוצרי המערכת, מאמתים את הדיוק שלהם ומשלבים אותם בשיקול הדעת המקצועי שלהם (Department for Science, Innovation and Technology [DSIT], 2025).

יצירת תוכן. כאמור, תחום מרכזי ומתפתח בבינה מלאכותית הוא התחום הגנרטיבי, כלומר יצירת תוכן. מודלים מסוגלים לייצר בפועל חומרים חדשים שלא היו קיימים קודם לכן – טקסט, תמונה או מולטימדיה. היכולת ליצור תוכן מקורי מתאפשרת בזכות למידה של המודלים ממאגרי נתונים עצומים, המאפשרת להם לזהות דפוסים וקשרים רחבים ולהפיק תוצרים מקוריים וחדשניים. שלא כמו מודלים שמטרתם ניתוח נתונים קיימים או קבלת החלטות, מודלים גנרטיביים מתמקדים בהפקת תוצר מקורי וחדש על בסיס הידע שנצבר, שלעיתים קרובות נשמע או נראה אנושי. בין המודלים הגנרטיביים הנפוצים ישנם מודלי שפה גדולים דוגמת GPT-4 של OpenAI ומודלים ליצירת תמונות כמו Midjourney או יישומים שלהם הבאים לידי ביטוי במגוון רחב של תחומים. לדוגמה, כלי MagicSchool המבוסס בינה מלאכותית נועד לסייע למורים בביצוע משימות שגרתיות ולהגביר את היעילות הפדגוגית במערכת החינוך. הכלי משתמש במודלי שפה גדולים כדי ליצור אוטומטי תכנים כגון מערכי שיעור, שאלות רב-ברירה, תרגולים מותאמים אישית וסיכומי טקסטים. הכלי פותח במיוחד לתחום החינוך, והוא כולל התאמות לרמות גיל שונות, מקצועות לימוד וצרכים מגוונים של מורים. הכלי הוא פרי יוזמה מסחרית של אנשי חינוך בארצות הברית, והוא אינו בפיקוח משרד החינוך אלא מוצג באתר המשרד כאחד מכלי הסיוע להוראה. הכלי מאפשר למורים לחסוך זמן בהכנה, לשפר את הגיוון הפדגוגי ולשלב תכנים מותאמים אישית לתלמידים, תוך כדי שמירה על שיקול דעת מקצועי ובקרה אנושית על הפלטים שהוא מייצר.

7. גישות בהוגנות במערכות מבוססות בינה מלאכותית

הוגנות היא עיקרון שלפיו קבלת החלטות והקצאת משאבים צריכות להתבסס על מאפיינים רלוונטיים בלבד (כגון כישורים, תכונות אישיות וצרכים), תוך מזעור השפעתם של גורמים שרירותיים או בלתי רלוונטיים, במטרה להבטיח שוויון הזדמנויות ולצמצם פערים בין פרטים וקבוצות באוכלוסייה. הגדרה זו שמה דגש בהענקת תנאי פתיחה שווים לכלל הפרטים, כך שלאחר מכן "התחרות" ביניהם תתנהל בתנאים זהים (Heidari et al., 2018; Rawls, 1971). כך למשל, בתהליך חיפוש עבודה, מועמדים ייחשבו ראויים אם יש ברשותם הכישורים המתאימים ביותר לתפקיד, ללא תלות במאפיינים כגון אתניות, גיל, מגדר או מצב משפחתי. עיקרון משלים לעקרון ההוגנות הוא עקרון שוויון ההזדמנויות, ויש הרואים בו אף חלק בלתי נפרד מן ההוגנות. אף שהגדרתו המדויקת שנויה במחלוקת, יש הסכמה כי זהו אחד מעקרונות היסוד של הדמוקרטיה המכיר באי-השוויון הקיים בנקודת הפתיחה של פרטים שונים. בהתאם לכך, חלוקת המשאבים על ידי המדינה צריכה להיעשות לפי נקודת המוצא של כל פרט, כך שלא בהכרח כל אחד יקבל משאבים שווים אלא משאבים על פי צרכיו ומצבו, במטרה לצמצם פערים ולאפשר שוויון בפועל. מטרתו של עקרון שוויון ההזדמנויות היא למנוע אי-שוויון על כל צורותיו ולתרום להוגנות באמצעות תמיכה בקבוצות באוכלוסייה שאינן מיוצגות כראוי, מצויות בשוליים או מוחלשות (Cepiku & Mastrodascio, 2021).

עם העלייה בשימוש במערכות מבוססות בינה מלאכותית ובאלגוריתמים ככלים לקבלת החלטות, התחדד הדיון סביב סוגיית ההוגנות בהקשרים אלו. מערכות אלו משולבות כיום בתחומים ציבוריים רגישים כמו מדיניות הגירה, איתור מצבי סיכון וניצול ילדים, שירותי בריאות וזכאות לרווחה, ולעיתים קובעות מי יזכה בשירות או במשאב ומה תהיה תוצאת ההחלטה בעניינו (Mitchell et al., 2021). התרחבות השימוש במערכות מבוססות בינה מלאכותית מעוררת שאלות עמוקות בנוגע לטיבה של הוגנות בבינה מלאכותית, במיוחד כאשר ההחלטות עשויות להשפיע ישירות על חייהם של בני אדם. לדוגמה, בעלי עסקים עצמאיים, ובפרט נשים ומי שמשתייכים לקבוצות מיעוט, המתמודדים עם מחסור בשוויון הזדמנויות בקבלת שירותים במימון ציבורי שנועדו לתמוך בעסקיהם (Nam et al., 2024).

חוקרים חלוקים בדעתם בנוגע לאופן שבו יש להגדיר הוגנות במערכות אלו, ואפשר למצוא מגוון רחב של גישות, רבות מהן נשענות על עקרון שוויון הזדמנויות. גישות אלו מתמקדות בהבטחת תנאי פתיחה מותאמים לכלל הפרטים שהמערכת משפיעה עליהם, תוך כדי מזעור ההשפעה של גורמים שרירותיים או לא רלוונטיים, כמו מגדר, אתניות או מצב חברתי (Verma & Rubin, 2018).

גישות הוגנות

אחת ההבחנות המרכזיות בדיון על הוגנות במערכות מבוססות בינה מלאכותית היא בין ארבע גישות המציעות ארבעה סוגי הוגנות: הוגנות של הפרט, הוגנות קבוצתית, הוגנות נגד-עובדתית והוגנות ברמת המערכת.

הוגנות של הפרט (individual fairness) נשענת על הרעיון ולפיו אנשים דומים צריכים לקבל טיפול דומה. עם זאת גישה זו מעוררת שאלה עקרונית: האם אפשר באמת להניח ששני אנשים שלהם מאפיינים זהים, כפי שהם נמדדים על ידי המודל, הם גם שווי ערך מבחינת ההקשר החברתי, התרבותי או הנסיבתי שבו הם פועלים? חוויות חיים, תנאי סביבה והקשרים שונים עלולים להוביל לכך שלמרות הדמיון במאפיינים הנמדדים, ההתייחסות הזהה אינה הוגנת בפועל. לכן ניסיון לבצע מדידה של עיקרון זה עלול להיות בעייתי או חלקי בלבד (Jacobs & Wallach, 2021).

היבט נוסף הקשור להוגנות של הפרט נוגע להוגנות מתוך מודעות (fairness through awareness). היבט זה מהווה עיקרון תכנוני המנחה את עיצוב המערכת ואיסוף הנתונים כבר בשלבים המוקדמים, והוא מושתת על ההנחה ולפיה יש להתייחס בהוגנות לאנשים שונים בהינתן משימת סיווג מסוימת. משימת סיווג היא מצב שבו המערכת נדרשת לשייך כל אדם או מקרה לקטגוריה מוגדרת מראש, למשל להחליט אם מועמד מתאים או לא מתאים לתפקיד, אם אדם זכאי או לא זכאי לשירות, או אם מקרה מסוים נחשב בסיכון גבוה או נמוך. לפיכך יש להכיר את זהותם החברתית ואת מאפייניהם, כגון מגדר, אתניות או מעמד חברתי-כלכלי, ולהתחשב בכך. גישה זו טוענת כי כדי למנוע אפליה מערכתית או לתקן עוולות היסטוריות יש צורך להכניס אל תהליך קבלת ההחלטות את ההקשר החברתי שבו פועל כל פרט. כך מערכות המבוססות בינה מלאכותית יוכלו להבחין בין פרטים שנראים דומים מבחינה טכנית אך שונים מהותית מבחינת ההזדמנויות שהיו זמינות להם או החסמים שעמדו בפניהם במשימה המסוימת (Dwork et al., 2012; Jacobs & Wallach, 2021; Saxena et al., 2020).

הוגנות קבוצתית (group fairness) מתמקדת בבחינת תוצאות פעולת המערכת בפועל ושואפת להבטיח שקבוצות חברתיות שונות, למשל קבוצות אתניות, מגדריות או דתיות, יזכו ליחס דומה מבחינת התוצאה המתקבלת מן המערכות המבוססות בינה מלאכותית. גישה זו מניחה כי בשל אי-שוויון מבני ואפליה כלפי קבוצות מסוימות קשה לאמוד את תנאי הפתיחה או ההזדמנויות שהיו זמינות לכל אדם. לכן היא מציעה להתמקד ברמת הקבוצה. כך למשל, ייתכן שיוחלט לפעול על פי עיקרון העדפה מתקנת (כגון מתן נקודות בונוס) לבני קבוצה מופלית במטרה להשיג תוצאה שוויונית יותר. זאת ועוד, אחת הסיבות המרכזיות לכך שהוגנות קבוצתית זוכה ליישום נרחב יחסית היא הפשטות היחסית שבה אפשר למדוד אותה – באמצעות השוואת שיעורי הצלחה בין קבוצות מוגדרות מראש. בניגוד לגישות אחרות המצריכות הגדרת מדדי דמיון בין פרטים או בניית מודלים סיבתיים מורכבים, הוגנות קבוצתית מאפשרת תפעול טכני

ברור יותר ולכן משמשת בסיס לרבות מן ההתערבויות והפתרונות בתחום הבינה המלאכותית (Jacobs & Wallach, 2021).

גישה זו אף היא אינה חפה מקשיים: ההגדרות של קטגוריות קבוצתיות לפי, לדוגמה, אתניות, גיל או מגדר, אינן אוניברסליות אלא משתנות לפי תרבות, זמן והקשר. פערים אלו מקשים על יישום אחיד של גישת ההוגנות הקבוצתית ואף עלולים להוביל להתעלמות מן השונות הפנים-קבוצתית (Jacobs & Wallach, 2021). נוסף על כך הוגנות קבוצתית מציבה אתגר לא מבוטל בהתמודדות עם מצבים של זהויות צולבות (intersectionality). אף שעקרונית אפשר למדוד גם זהויות צולבות, יישום יעיל של גישה זו מחייב בחירה מודעת אילו קבוצות סיכון או קבוצות בעלות חשיבות ציבורית ייבחנו וכן מדידה מורכבת יותר מעבר לחלוקה לקבוצות רחבות בלבד. במקרים אלו עלולה המערכת להפיק תוצאות מפלות כלפי קבוצות אלו, כמו נשים מקבוצת מיעוט אתני, אף שהמדדים הכלליים מצביעים על הוגנות בין הקבוצות הגדולות.

הוגנות נגד-עובדתית (counterfactual fairness) מבוססת על עקרונות של בדיקת סיבתיות. לפי גישה זו החלטה תהיה הוגנת אם אין תלות סיבתית בין שייכות לקבוצה המוגנת ובין תוצאת המודל. המטרה של הגדרת ההוגנות הזו היא למנוע מצב שבו המודל אומנם לא משתמש ישירות בשייכות לקבוצה מוגנת, אבל כן משתמש בה בעקיפין וגורם לאפליה. לדוגמה, מודל לסינון קורות חיים שלא משתמש במוצא כדי להחליט על קבלה אבל משתמש בשם של המועמד ויוצר אפליה כלפי מיעוט אתני. בכך גישה זו מבקשת למנוע אפליה עקיפה שנוצרת עקב שימוש עקיף בשייכות לקבוצה מוגנת, והיא דורשת הבנה של קשרים סיבתיים בין משתנים. גישה זו ממוקמת תאורטית בין הוגנות של הפרט להוגנות קבוצתית, תוך שהיא מדגישה את חשיבות ההקשר וההיסטוריה החברתית של קבלת ההחלטות (Kusner et al., 2017).

בצד שלושת הסוגים של הוגנות אלגוריתמית, יש חוקרים (Cruz Cortés & Ghosh, 2020; Lam, 2025) המצביעים על חשיבותה של **הוגנות ברמת המערכת** (system-level fairness) המדגישה את ההקשר המוסדי והמערכתי שבו מתקבלות החלטות. בשונה מן הגישות המתמקדות בבחינת ההחלטות עצמן בתוך מערכת נתונה, הוגנות ברמת המערכת בוחנת אם הכללים, הקריטריונים, הנתונים והמבנים המוסדיים שעליהם מבוססות ההחלטות הם הוגנים מלכתחילה. במילים אחרות, הגישות האחרות מניחות כי המסגרת עצמה תקינה ורק השימוש בה עלול להיות מפלה, ואילו גישה זו מצביעה על כך שהמסגרת עצמה עשויה לגלם עיוותים, אפליה או הטיות מבניות. לפיכך היא קוראת לא רק לתיקון המערכות המבוססות בינה מלאכותית אלא גם לבחינה ביקורתית של הכללים והמבנים שבתוכם הן פועלות, במטרה לוודא כי ההוגנות מושגת לא רק כלפי הפרטים והקבוצות אלא גם ברמה המערכתית (Cruz Cortés & Ghosh, 2020; Lam, 2025). בעקבות ריבוי הגישות והמתח בינן, התבססות בלעדית על הוגנות מסוג אחד עלולה להוביל לעיוותים או לחוסר הוגנות מסוג אחר. לפיכך גישות להוגנות צריכות לשלב בין ההיבטים השונים בשימת לב למגבלות, למתחים הפנימיים ולמורכבות שביניהם. כל אחת מן הגישות מדגישה היבטים שונים של הוגנות אך נושאת עימה מגבלות.

לכן נדרשת הבנה רב-ממדית של הוגנות, הכוללת גם תובנות נורמטיביות, כלים פורמליים ורגישות להקשרים מוסדיים, תרבותיים והיסטוריים (Jacobs & Wallach, 2021). נוכח ההיבט הערכי של ההוגנות, חשוב להדגיש כי קביעת העקרונות המתאימים לכל מערכת אינה יכולה להישען אך ורק על בחירה טכנית בגישה אחת אלא מחייבת תהליך שקוף ומשתף עם בעלי עניין וקבוצות מושפעות. באמצעות תהליך זה יש להגדיר מהם העקרונות של הוגנות ושוויון הזדמנויות לכל מערכת על פי מטרותיה וההקשר החברתי שבו היא פועלת.

בכך הופכות הגישות השונות לא רק למסגרות ניתוח תאורטיות אלא גם לכלי פעולה בני-ביצוע המסייעים להאיר דילמות ולנסח עקרונות פעולה קונקרטיים לשילוב הוגנות במערכות המבוססות בינה מלאכותית בשירות הציבורי. בצד ההבחנות בין הגישות, חשוב להדגיש כי יישום כל אחד מעקרונות ההוגנות מחייב גם בחינה אמפירית של תוצאות המערכת בפועל, לעיתים באמצעות מדדי הוגנות קבוצתית או זהויות צולבות, כדי לוודא שהעקרונות שנבחרו אכן מתממשים. **לוח 2** מציג השוואה בין ארבעת סוגי ההוגנות.

לוח 2: סוגי הוגנות: השוואה

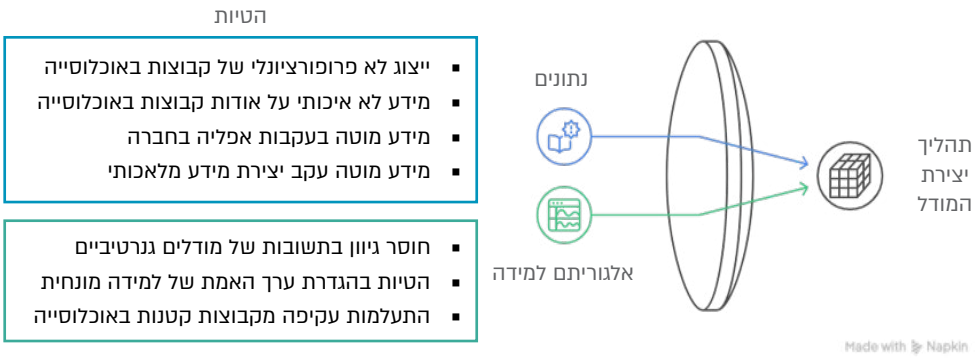
סוג ההוגנות	העיקרון המנחה	יחידת הניתוח	יתרונות	מגבלות
הוגנות של הפרט	פרטים דומים צריכים לקבל טיפול דומה	הפרט	ממוקדת בפרטים, רגישה להקשרים; מתאימה לשירותים מותאמים אישית	קשה להגדיר מדד דמיון בין פרטים; מתעלמים מן ההזדמנויות השונות בין פרטים
הוגנות קבוצתית	קבוצות שונות צריכות לקבל יחס דומה בתוצאה	הקבוצה	ניתנת למדידה ישירה	קטגוריות קבוצתיות תלויות הקשר; לא מתחשבת בשונות פנים-קבוצתית
הוגנות נגד-עובדתית	החלטה לא תלויה סיבתית בקבוצה	הפרט והקבוצה	לוכדת אפליה עקיפה והקשרים חברתיים עמוקים יותר	דורשת מודלים סיבתיים מורכבים; קשה ליישום בפועל
הוגנות ברמת המערכת	המערכת והמוסדות עצמם צריכים לפעול באופן שמאפשר הוגנות	המערכת וההקשר המוסדי הכולל	מזהה כשלים מבניים, מטפלת בשורש האי-הוגנות, לא רק בתוצאות	קשה למדידה ישירה; דורשת שינויי מדיניות ומבנים; מורכבת ליישום

8. הטיות של אלגוריתמים במודלים של בינה מלאכותית

יישומי בינה מלאכותית מציעים יתרונות רבים בייעול תהליכים וביכולת הדיוק, אך השימוש בהם עלול להניב אתגרים מורכבים של הטיית אלגוריתמים, ואלו עלולים להוביל לחוסר הוגנות. מודלים של למידת מכונה, במיוחד המתקדמים יותר שנמצאים בשימוש בשנים האחרונות, הם מודלים מורכבים המכילים מספר גדול מאוד של משתנים שמשפיעים על התוצאה. באלגוריתמים שתכנתו אנשים ובמודלים פשוטים אפשר להבין מהתבוננות באלגוריתם אילו מאפיינים של הקלט שהאלגוריתם מקבל משפיעים על התוצאה שלו. לדוגמה, במודלים פשוטים להערכת סיכון להחזר הלוואה אפשר לראות כיצד המשכורת משפיעה על הערכת הסיכון. במודלים המתקדמים יותר אין דרך להבין את התלות של התוצאה של המודל במשתנים על ידי בחינת המודל. מאחר שאין דרך להבין מבחינת המודל מהם המשתנים שמשפיעים על התוצאה, חשוב להיות מודעים להטיות האפשריות ולבחון את המודל כדי לוודא שאין בו הטיות.

בפרק זה נסקור את ההטיות האפשריות במודלים של בינה מלאכותית, בחלוקה לשתי קטגוריות: הטיות הקשורות לנתונים המשמשים ליצירת המודל; והטיות הקשורות לבחירות עיצוביות בתהליך הלמידה עצמו. נתמקד בהטיות עצמן ובגורמים להן ולא בהשפעות של ההטיות ובסיכונים הנובעים מכך (אלו יידונו בפרק 9). חשוב לציין כי החלוקה לקטגוריות אינה חד-משמעית. יש הטיות שיכולות להשתייך ליותר מקטגוריה אחת, והטיות יכולות להשפיע זו על זו (Mehrabi et al., 2021). **תרשים 3** מציג את ההטיות בשלבים שונים של תהליך יצירת המודל.

תרשים 3: סוגי הטיות



8.1 הטיות הקשורות לנתונים

אחת הסכנות המרכזיות בפיתוח מודלים של בינה מלאכותית היא ההטיה הנובעת מן הנתונים שבהם נעשה שימוש בתהליך הלמידה. כאמור, מודלים אלו מבוססים על זיהוי דפוסים מתוך מאגרי מידע קיימים. בלמידה מונחית, תהליך זה מתבצע באמצעות מערך של דוגמאות המכיל זוגות של קלט ופלט. האלגוריתם לומד את הקשר בין הקלט לפלט על סמך הדוגמאות שהוזנו לו ומבקש להכליל ממנו כללים שיאפשרו לו לחזות פלטים עבור קלטים חדשים, שלא נראו בתהליך האימון. בלמידה שאינה מונחית, לעומת זאת, הנתונים אינם כוללים פלטים ידועים מראש, ולכן האלגוריתם נדרש לזהות מבנים סמויים או דפוסים פנימיים במידע, כמו קיבוץ של טקסטים לפי נושאים סמנטיים. בשני סוגי הלמידה לאיכות הנתונים והרכבם השפעה ניכרת על אופי הייצוג שאליו מגיע המודל, ויש חשיבות מכרעת לבחירת מערך הנתונים ולעיבודם (Mehrabi et al., 2021).

להלן יוצגו מצבים שבהם הנתונים עצמם יכולים לגרום להטיות של המודל הלומד מהם:

ייצוג לא פרופורציונלי של קבוצות באוכלוסייה

מצב שבו המידע שעליו המודל לומד מכיל נתונים על אוכלוסיית הרוב אבל פחות נתונים על אוכלוסיות מיעוט. מצב זה יכול לגרום ליצירת מודל מדויק על אוכלוסיית הרוב, אבל מדויק פחות על אוכלוסיות המיעוט. חוסר דיוק של מודל להערכת סיכון עלול להוביל לכך שהמודל יניב תוצאה כללית הקרובה לממוצע של כלל אוכלוסיית המיעוט, במקום להבחין בין אנשים שונים ולספק הערכת סיכון מותאמת אישית – גבוהה עבור חלק ונמוכה עבור אחרים.

כאשר יש תת-ייצוג של אוכלוסיית מיעוט בנתוני האימון, דיוק המודל בנוגע לאוכלוסייה זו נוטה להיפגע משתי סיבות עיקריות. האחת, המודל "לומד" שאוכלוסייה זו קטנה יחסית, ולכן אינו מייחס חשיבות רבה לשיפור הדיוק עבור חבריה. והאחרת, מאחר שיש נתונים מועטים על אוכלוסיית המיעוט, המודל מתקשה לזהות את מאפייניה הייחודיים ולספק תחזיות מדויקות על אודותיה. כלומר גודלה היחסי של קבוצת המיעוט משפיע על ביצועי המודל, והביצועים על קבוצה מסוימת טובים פחות כאשר חלקה היחסי בנתונים קטן (Rolf et al., 2021).

חוסר ייצוג של קבוצות באוכלוסייה בנתונים עלול לנבוע מכמה סיבות:

- **שימוש במאגר מידע שלא מכיל נתונים על קבוצות באוכלוסייה.** כאשר מנסים ללמד מודל כלשהו, תחילה בוחרים במאגר מידע מתאים. אולם במקרים מסוימים אין מאגרי מידע על קבוצות מסוימות. במקרים אחרים יש מאגרי מידע כאלה, אך נבחר לשימוש מאגר מידע ספציפי שבו יש מידע איכותי מצד אחד, ומנגד אין מידע על קבוצות באוכלוסייה. לדוגמה, מאגר ובו מידע רפואי יהיה איכותי אם הוא יכיל מידע מקיף על החולים לאורך זמן (Ehsani-Moghaddam et al., 2021). במחקר על תחום הרפואה (Tomašev et al.,

2019) נבנה מודל לחיזוי מחלות בכליות, ובו נעשה שימוש במאגר המידע של המחלקה לענייני חיילים משוחררים בארצות הברית. זהו מקור מידע ובו רשומות רפואיות מקיפות למדיי, אבל שיעור הנשים הכלולות בו נמוך מאוד (כ-6%). שיעור נמוך זה גרם למודל להיות מדויק פחות עבור נשים. מצב של חוסר ייצוג יכול להתרחש גם כאשר משתמשים במאגרי מידע ממדינה אחרת שבה התפלגות האוכלוסייה שונה או כאשר יש שינוי בהתפלגות האוכלוסייה בחלוף הזמן (Ciecierski-Holmes et al., 2022).

- **הטיית הבחירה (selection bias)**. בדרך כלל הטיית הבחירה מתרחשת במצב שבו בוחרים משתתפים למחקר או למאגר מידע ובעקבותיו נוצרת הטיה. הטיה זו יכולה להתרחש עקב החלטות של יוצרי מאגר המידע בנוגע לאילו משתתפים לגייס או עקב החלטות המשתתפים עצמם אם להשתתף בו. ההטיה כוללת גם את הבחירה במאגר המידע, ולכן היא כוללת את הבעיה שתוארה לעיל בנוגע לחוסר ייצוג במאגרי נתונים. עם זאת מדובר בהטיה רחבה ומורכבת יותר, שכן היא עשויה לנבוע גם מגורמים נוספים שאינם קשורים ישירות לבחירת המאגר, כמו אופן גיוס המשתתפים למחקר, פרישת משתתפים במשך שלב איסוף הנתונים או הטיות הקשורות להיענות שונה של קבוצות באוכלוסייה לסקרים. בכך מהווה הטיית הבחירה הרחבה של בעיית חוסר הייצוג אך גם ממקדת את תשומת הלב בהיבטים תפעוליים ואנושיים בתהליך האיסוף. דוגמה נפוצה להטיית הבחירה היא סקרים, מאחר שחוסר היענות לסקרים לא מתפלג התפלגות אחידה ומייצגת באוכלוסייה. חוסר היענות לסקרים יכול להיגרם עקב סיבות רבות, כמו פערי שפה או זמינות, ואי אפשר לחזותו במדויק. אחת משיטות התיקון של הטיה זו היא משקול הנתונים המצויים בתת-ייצוג (ליטווין וספיר, 2008).
- **פערים בהיקף המידע במרשתת**. מודלים גנרטיביים לומדים בדרך כלל מהיקף עצום של מידע פומבי (טקסט, תמונות, וידיאו) הנמצא במרשתת, והם מושפעים מהיקפו. מודלי שפה מבצעים יותר טעויות ונותנים תשובות איכותיות פחות בשפות שבהן אין טקסט רב במרשתת. על כן המידע במרשתת אינו מייצג את האוכלוסייה הכללית, מאחר שקבוצות מסוימות באוכלוסייה מפיקות, מתעדות ומפיצות תוכן דיגיטלי מיותר מקבוצות אחרות (Joshi et al., 2020). לכן גם במקרה של מודלי שפה ייתכן תת-ייצוג של קבוצות באוכלוסייה שמשפיע על ביצועי המודל בשאלות של מידע בנוגע להן. לדוגמה, ביצועי מודלי שפה גדולים באמהרית טובים הרבה פחות מבאנגלית (Myung et al., 2024). הבעיה נוגעת גם לעברית; מחקר שבחן חלקי משפט ותרגום במגוון שפות הראה כי בעברית ביצועי המודלים טובים הרבה פחות מבאנגלית, אבל טובים יותר מבשפות אחרות (Ahuja et al., 2023). לפי מיונג ואח' (Myung et al., 2024), חוסר ייצוג זה מוביל לכך שדוברי שפות, שעליהן יש פחות מידע, מקבלים מידע מתוך שירותי הבינה המלאכותית, שהוא ירוד באיכותו, ועקב כך מעמיק הפער הדיגיטלי ונשמרת הנחיתות הטכנולוגית. פערים אלו באים לידי ביטוי גם בתחומים אחרים, כגון בריאות הציבור. לדוגמה, במדינות שבהן אוריינות דיגיטלית נמוכה,

מודלים של בינה מלאכותית לרוב לא מאומנים בשפות המקומיות. לכן קהילות שלמות אינן מקבלות מידע רפואי בשפה נגישה ומותאמת ולעיתים אף נחשפות למידע שגוי או לא רלוונטי, והדבר פוגע בבריאות הציבור (Hu et al., 2024).

מידע לא איכותי על קבוצות באוכלוסייה

בניגוד לייצוג לא פרופורציונלי, שבו מאגר מידע לא מכיל נתונים על קבוצות באוכלוסייה, כאן מדובר במאגר ובו נתונים על כל מגוון האוכלוסייה, אבל המידע על קבוצות מסוימות הוא מקיף פחות או איכותי פחות מן המידע על אחרות. יש סיבות רבות לפערים בהיקף או באיכות המידע על קבוצות שונות באוכלוסייה, ולהלן יוצגו כמה דוגמאות למידע חסר ולטעויות במידע.

אחד הגורמים למידע חסר הוא קבוצות באוכלוסייה שמידת האינטראקציה שלהן עם המערכת הציבורית היא פחותה, למשל האוכלוסייה הבדואית בנגב. חלק מן האוכלוסייה הבדואית בנגב גרה ביישובים לא מוכרים, והמידע על מי שגרים ביישובים מוכרים הוא חסר ולא עדכני (לף, 2023). שימוש בנתונים הלא עדכניים הזמינים כדי ללמד מודל לקבל החלטות בנוגע לאוכלוסייה זו יהיה מדויק פחות.

נתונים רפואיים הם תחום נוסף שבו יש מידע לא איכותי על קבוצות באוכלוסייה. יש קבוצות רבות שלהן נגישות נמוכה לשירותים רפואיים, ולכן נתונים רפואיים עליהן הם איכותיים פחות (Hing & Burt, 2009). מצב זה יכול לגרום למודל שלומד מאותם נתונים להיות מדויק פחות בכל הנוגע לאותן קבוצות באוכלוסייה, לדוגמה זינק ואח' (Zink et al., 2024) בדקו את ההשפעה של פערים במידע על אודות ההיסטוריה המשפחתית של מטופלים על מודל להערכת סיכון לחלות בסרטן המעי הגס. הם גילו שאיכות המידע של ההיסטוריה המשפחתית של מטופלים ממוצא אפרו-אמריקני הייתה טובה פחות מזו של מטופלים לבנים. הדבר גרם לשגיאה בהערכת הסיכון של מטופלים ממוצא אפרו-אמריקני. כאשר נתונים אלו, הכוללים חוסרים או פערים באיכות, משמשים לאימון מודלי למידת מכונה, עולה הצורך להתמודד עם אתגר המידע החסר.

נוסף על כך בחירת שיטה לעיבוד נתונים חסרים עשויה להשפיע על הוגנות. כאשר מאמנים מודל של למידת מכונה (ראו מונחון מושגים) מנתונים המופיעים בטבלה, כמו נתונים רפואיים, נתונים על הכנסות או ביצועי סטודנטים, אלגוריתמים רבים של למידת מכונה יכולים לפעול רק על טבלה מלאה שבה אין שדות ריקים. בטבלה המייצגת נתונים על הכנסות שדה ריק יכול להיות, לדוגמה, מצב של אדם שאין מידע על ההכנסות שלו בשנה מסוימת. כשרוצים לאמן מודל על נתונים מטבלה שאינה שלמה, יש כמה שיטות אפשריות לעיבוד מקדים שאפשר להפעיל על הטבלה, והעיקריות שבהן הן השמטת נתונים (שורות או עמודות) שבהם יש ערכים חסרים והשלמת הערכים החסרים בטכניקות שונות (Fernando et al., 2021; Zhang & Long, 2021).

ברבים ממאגרי הנתונים המצויים בטבלה הערכים החסרים נוטים להתרכז באוכלוסיות מסוימות ואינם מתפלגים אקראית (Fernando et al., 2021). כאשר משמיטים שורות שלמות (כאשר כל שורה מייצגת אדם), הדבר עלול לפגוע בייצוג של קבוצות מיעוט, ובכך להפחית את דיוק המודל עבורן. אחת הדרכים לצמצם פגיעה זו היא באמצעות משקול, כלומר הענקת משקל גדול יותר לשורות שכן כוללות נתונים מלאים. עם זאת גם שיטה זו אינה נטולת בעיות: המשקול עלול לפגוע בדיוק המודל עבור קבוצות באוכלוסייה שבהן שיעור החסרים גבוה במיוחד (Zhang & Long, 2021). השיטה השנייה היא השלמת הנתונים החסרים, ולכך יש טכניקות מגוונות, כמו השלמה באמצעות ממוצע של כלל הערכים או שימוש בשורות דומות בעלות מידע מלא. גם במקרה זה ההשלמה עלולה ליצור הטיות, אם כי ברמה מתונה יותר לעומת הטיות עקב השמטת שורות שלמות (Fernando et al., 2021). כמו כן השלמה של נתונים חסרים היא סוג של יצירת מידע מלאכותי (סינטטי) (ראו פירוט להלן).

מידע מוטה בעקבות אפליה בחברה

הטיה זו משקפת מצב שבו הנתונים מייצגים את המצב בחברה, אולם החברה מושפעת מאפליה ואין רצון שהמודל ישקף אותה. לדוגמה, בארצות הברית הסיכוי של סטודנט שסיים תיכון בהצלחה לסיים תואר ראשון נמוך במידה ניכרת אם הוא מגיע ממשפחה שרמת הכנסתה נמוכה, זאת לעומת סטודנט שסיים תיכון בהצלחה ומגיע ממשפחה שרמת הכנסתה גבוהה (Wyner et al., 2007). במצב זה ההנחה היא שחוסר ההצלחה בלימודים נגרם עקב אפליה, קשיים כלכליים של הסטודנט או סיבות נוספות. אם מטרת המודל היא לחזות את סיכויי המועמד להצליח באוניברסיטה כדי לדעת אילו מועמדים לקבל, עליו להימנע מדחיית מועמדים בעלי הישגים גבוהים בתיכון שמגיעים ממשפחות שרמת הכנסתן נמוכה. מודל כזה ישקף נאמנה את הנתונים, אך מאחר שנתונים אלו מושפעים מאפליה חברתית, הוא עלול לשעתק אותה וליצור חיזוק לאפליה קיימת (Hutt et al., 2019).

דוגמה נוספת לשיקוף אפליה בחברה מצויה באלגוריתמי שפה גנרטיביים. אלגוריתמי שפה לומדים מהיקף גדול של מידע פומבי הנמצא במרשתת, אשר מכיל בין השאר טקסטים המכילים דעות קדומות, סטראוטיפים ותיאור של אפליה. אלגוריתמים של למידת מכונה משקפים את האפליה ואת הסטראוטיפים הבאים לביטוי במידע במרשתת. לדוגמה, מודל שפה התבקש לכתוב מכתבי המלצה, והוא יצר מכתבים שבהם סטריאוטיפים נגד נשים (Wan et al., 2023). בדומה, ארמסטרונג ואח' (Armstrong et al., 2024) ביקשו ממודל שפה ליצור קורות חיים לאנשים עם שמות של נשים וגברים אופייניים לקבוצות שונות באוכלוסייה בארצות הברית (אפרו-אמריקנים, לבנים, היספאנים ואסיאתים), והמודל יצר קורות חיים שבהם שנות ניסיון רבות יותר ותפקידים בכירים יותר לקבוצות מסוימות לעומת אחרות. מן המודל אף נדרש לדרג בציון עד כמה כדאי להעסיק או להזמין לריאיון מועמדים שונים לעבודה לפי קורות החיים שלהם, ונמצא פער עקבי בציון שקיבלו קורות חיים זהים כאשר החליפו את שם המועמד לשם

המאפיין מיעוטים, כמו אפרו-אמריקנים. בדרך כלל מודלים גנרטיביים ומודלים לקבלת החלטות משמשים למשימות שונות, אבל אם משתמשים במודלים גנרטיביים במקום במודלים לקבלת החלטות, יש סכנה שסטראוטיפים הבאים לביטוי במרשתת ישפיעו על קבלת ההחלטות (Liu et al., 2024) (בהקשר זה ראו גם פרק 9).

מידע מוטה עקב יצירת מידע מלאכותי (סינטטי)

יש אלגוריתמים גנרטיביים המסוגלים לייצר היקף גדול של מידע באיכות גבוהה יחסית. אלו כוללים מודלים כלליים כמו מודלי שפה גדולים או מודלים ליצירת תמונות, מודלים פשוטים להשלמת שדות וגם מודלים אדברסריאליים GANs (Generative Adversarial Networks) שאומנו ספציפית ליצירת מידע דומה מהתפלגות הנתונים במאגר המקורי. מודלים גנרטיביים מאפשרים להשתמש במידע מלאכותי כמאגר מידע לאלגוריתם הלמידה, לבד או בצד נתוני אמת. יש כמה סיבות לשימוש במידע מלאכותי כמאגר מידע, ובהן שמירה על הפרטיות במקרים שבהם המאגר מכיל מידע רגיש או יצירת מידע מאחר שאין מספיק מידע אמיתי זמין (Nikolenko, 2021). שימוש במידע מלאכותי מסיבות של הגנה על הפרטיות נעשה, לדוגמה, בארצות הברית במרשם האוכלוסין (U.S. Census Bureau, 2021).

אף שיש יתרונות בשימוש במידע מלאכותי, הוא עלול להכיל הטיות. במקרה שהמידע המקורי שעליו אומן המודל הגנרטיבי הכיל הטיות, המודל ילמד הטיות אלו וישקף אותן במידע המלאכותי שייצור. שימוש כזה יכול אף להגביר הטיות קיימות. כמו כן אם המידע המקורי שעליו אומן המודל הגנרטיבי אינו עשיר מספיק, המידע הנוצר על קבוצות המיעוט יכול להיות באיכות נמוכה ולא מגוון מספיק ולגרום בהמשך התהליך למודל שאומן על המידע המלאכותי להיות ברמת דיוק נמוכה בנוגע לקבוצת מיעוט זו (Chen et al., 2021). כאשר משתמשים במידע מלאכותי לאימון מודל שיוצר בעצמו מידע מלאכותי נוסף, יש אפוא סכנה של היזון חוזר והגברת ההטיות (Wyllie et al., 2024).

לסיכום, למידע מלאכותי יש יתרונות בהיבט של פרטיות ויצירת מידע כאשר יש חוסרים במידע, אך גלומים בו סיכונים מבחינת הוגנות המודל. לכן במקרים שבהם משתמשים במידע מלאכותי, יש להיות מודעים לכך שמידע כזה עלול להכיל הטיות. אם יש בנמצא מידע אמיתי, אפשר לבחון את המידע המלאכותי על ידי השוואה להתפלגות הנתונים במאגר המקורי.

8.2 הטיות הקשורות לאלגוריתם הלמידה

מודלים של בינה מלאכותית נוצרים על ידי אלגוריתם למידה אשר משתמש בהיקף גדול של נתונים. בתהליך הלמידה, אם הוא מונחה ואם לא, אפשר להשתמש במדד של פונקציית הפסד (ראו מונחון מושגים). לדוגמה, עבור מודל המנסה לחזות סיכון למחלת לב, ישתמשו בנתוני עבר רפואיים של חולים ויבדקו אם הם חלו במחלת לב בחמש השנים האחרונות. ההפסד של

האלגוריתם יהיה הפער בין הסיכון למחלה שהאלגוריתם חזה ובין שיעורי המחלה האמיתיים. אלגוריתם הלמידה שואף למצוא מודל שיצמצם את הטעויות הללו ככל האפשר, כלומר כזה שימזער את ערך פונקציית ההפסד ויפיק תחזיות מדויקות.

בפרק זה נבחן הטיות שנוצרות מבחירות מסוימות בתהליך הלמידה עצמו ועלולות לגרום לאלגוריתם הלמידה להעדיף מודל מסוים – כזה שמוטה נגד קבוצות שונות באוכלוסייה או לא מתאים להן במדויק – על פני מודלים אחרים.

הטיות הנובעות מחוסר גיוון בתשובות של מודלים גנרטיביים

אחת ההטיות במודלים גנרטיביים נובעת מתשובות לא מגוונות של המודלים. אף שמודלים אלו מבוססים על תהליכים הסתברותיים ומסוגלים לייצר מגוון רחב של פלטים, בפועל המשתמשים נחשפים רק לדוגמה אחת או שתיים מן ההתפלגות האפשרית. מצב זה יוצר צורך לאזן בין מגוון התשובות ובין דיוקן ועלול להוביל לפלטים חד-גוניים או שגויים (Hadi et al., 2023) כאשר הפלטים הופכים חד-גוניים עלול להיווצר חוסר ייצוג של אוכלוסיות מיעוט, משום שהמודל נוטה להעדיף תשובות "ממוצעות" או שכיחות על פני כאלה שמשקפות קולות ייחודיים יותר.

תופעה דומה זוהתה גם במודלים גנרטיביים ליצירת תמונות. נייק ונושי (Naik & Nushi, 2023) מצאו כי כאשר מבקשים ממודל גנרטיבי ליצור תמונות של אנשים העוסקים במקצועות שונים (למשל, "רופא"), מתקבלות תמונות שמעידות על הטיה מובהקת בייצוג המגדרי. אף שבמקצועות כמו רפואה יש שיעור ניכר של נשים (כ-40% מן הרופאים בארצות הברית), שיעור הנשים בתמונות שיצר המודל היה נמוך מ-20%. החוקרים השוו את תוצרי המודל גם לנתוני המציאות וגם להתפלגות בתמונות בחיפוש בגוגל ומצאו שההטיה כלפי ייצוג גברי הייתה בולטת יותר דווקא במודל הגנרטיבי. דפוס זה חזר על עצמו במקצועות רבים שנבדקו וחשף הטיה שיטתית בייצוג אוכלוסיות מגוונות מבחינת מגדר, מוצא וגיל.

ממצאים אלו מצביעים על כך שהגדרות ברירת המחדל ואופן הפעולה של מודלים גנרטיביים עלולים לשמר ואף להחריף דפוסים של חוסר ייצוג והדרה. גם כאשר המודל מבוסס על מידע מגוון, אופן הדגימה שלו עלול לצמצם את נראותן של קבוצות מיעוט ולהנציח סטריאוטיפים קיימים.

הטיות הנובעות מהגדרת ערך האמת (labeling bias) בלמידה מונחית

באלגוריתמים של למידה מונחית, תהליך הלמידה מבוסס על זוגות של קלט ופלט רצוי, כאשר מטרת המודל היא לחזות את הפלט הנכון עבור קלט חדש. עם זאת יש תחומים רבים שבהם אין הסכמה חד-משמעית או אובייקטיבית על מהו הפלט ה"נכון" (ערך האמת). דוגמאות לכך כוללות חיזוי של הצלחה בלימודים, אבחון מצב בריאות כללי, זיהוי הבעות פנים, איתור התנהגות חריגה או סיווג טקסטים כבעלי תוכן אלים. בכל אחד מן המקרים הללו השימוש בלמידה מונחית מחייב קביעה מוקדמת של ערך האמת, כלומר תהליך של הגדרה וסימון של

מה נחשב תוצאה נכונה שעל המודל ללמוד לחזות. תהליך זה כולל **החלטות אנושיות**, כמו כיצד להגדיר "אלימות מילולית" בטקסט, מי יהיה הגורם המסמן או על פי אילו קריטריונים תתבצע ההבחנה. הבחירה בערך האמת אינה ניטרלית, והיא עלולה ליצור הטיות מהותיות. כאשר הקריטריונים לסיווג אינם ברורים, או שהם מושפעים מהטיות תרבותיות, פוליטיות או מוסדיות של המתייגים, תהליך הלמידה כולו עלול לשקף ולשעתק את ההטיות הללו. מדובר בהטיה מבנית, שכן היא מוכללת במודל כבר משלב ההגדרה של הבעיה עצמה (Corbett-Davies et al., 2023; Zanger-Tishler et al., 2024).

סוגיה זו חשובה מאוד מאחר שהיא מדגישה את הצורך בהגדרה מפורשת של מטרת המודל. על כן נדון בה בהרחבה בסעיף 8.3, ובו היא תוצג כדוגמה מורחבת להטיה בתהליך הפיתוח והיישום של מודלים מבוססי למידה.

הטיות הנובעות מהתעלמות עקיפה מאוכלוסיות מיעוט

המטרה המרכזית של אלגוריתמים ללמידת מכונה היא לצמצם ככל האפשר את מספר הטעויות של המודל, והמדד לכך הוא ממוצע הטעויות לכל האוכלוסייה. אולם כאשר יש באוכלוסייה קבוצה קטנה אך מובחנת, למשל קבוצת מיעוט בעלת מאפיינים שונים, הגדרה זו עלולה להוביל להטיה. מכיוון שהקבוצה הקטנה משפיעה מעט מאוד על הממוצע הכללי, המודל "מתמקד" בעיקר בקבוצת הרוב, והדיוק שלו בנוגע לקבוצת המיעוט נשאר נמוך יחסית. הטיה זו שונה מהטיה של ייצוג לא פרופורציונלי של מודלים גנרטיביים, כי הבעיה במקרה הזה היא דיוק נמוך בנוגע לאוכלוסיית המיעוט ולא שהיא אינה מיוצגת בפלט המודל, כלומר מדובר במצב שבו קבוצת המיעוט מיוצגת ייצוג מלא ופרופורציונלי (Afrose et al., 2022). כיוון שמשקלה היחסי של אוכלוסיית המיעוט הוא קטן, המודל אינו "מושקע" בשיפור התחזיות בנוגע אליה. בעקבות הטיה זו, ישנם ביצועים ירודים יותר של המודל כלפי אוכלוסיות מיעוט. אפרוז ואח' (Afrose et al., 2022) הראו כי כאשר מודלים לומדים מנתונים הכוללים קבוצות מיעוט קטנות, דיוק המודל בנוגע אליהן נפגע. כדי להתמודד עם בעיה זו הם הציעו להוסיף משקולות מיוחדות בעת הלמידה, כך שדוגמאות מקבוצת המיעוט יקבלו חשיבות גדולה יותר בחישוב הכולל של הטעויות. החוקרים השוו את הגישה הזו לשיטות דומות ומצאו שהיא יכולה לשפר את הדיוק בנוגע לקבוצות מיעוט בלי לפגוע בביצועים על קבוצות הרוב. הטיה זו ממחישה כיצד גם מטרות כלליות ושוויוניות לכאורה עלולות להוביל לתוצאה לא הוגנת, אם אינן מביאות בחשבון את הפערים בין קבוצות שונות באוכלוסייה.

8.3 דוגמה מורחבת להטיה ספציפית: הגדרת המטרה של המודל

הטיה זו מתרחשת במסגרת למידה מונחית. המודל מקבל דוגמאות רבות שבהן כל קלט תויג מראש עם פלט נכון (ערך האמת), ומטרתו היא ללמוד כיצד לחזות את הפלט עבור קלטים חדשים שטרם ראה.

במקרים מסוימים ערך האמת מוגדר הגדרה חד-משמעית ונחשב אובייקטיבי, לדוגמה חיזוי אם חולה ימות במהלך אשפוז או אם אדם יחזיר הלוואה בזמן. עם זאת יש תחומים שבהם ערך האמת אינו ברור או אחיד, ולעיתים אף תלוי בהקשר תרבותי או ערכי. כך למשל, כאשר המודל לומד לחזות הצלחה בלימודים, להעריך את מצב הבריאות הכללי של אדם, לנתח הבעות פנים או לזהות אלימות בטקסט, נדרש תחילה להחליט כיצד להגדיר את הפלט הנכון ומהם הקריטריונים לקביעתו. במצבים אלו ההחלטה כיצד להגדיר את ערך האמת עלולה להכניס לתוך המודל הטיות סמויות. אם הקריטריונים לקביעת התשובות מבוססים על הנחות או דעות שאינן מייצגות את כלל האוכלוסייה, המודל שיתקבל עלול להעדיף קבוצות מסוימות על פני אחרות ולהפיק תחזיות מדויקות פחות עבור קבוצות מודרות או מוחלשות (Corbett-Davies et al., 2023; Zanger-Tishler et al., 2024).

אחת הדוגמאות המפורסמות בספרות העוסקת בהוגנות באלגוריתמים נוגעת להטיה זו. במקרה זה, חברה המספקת שירותי בריאות ללקוחות בארצות הברית רצתה לדעת מי מבין לקוחותיה הם חולים בסיכון גבוה, כדי לצרף אותם לתוכנית תמיכה מיוחדת. כדי לאתר את החולים הנמצאים בסיכון גבוה, הופעל מודל של למידת מכונה על נתוני עבר של החולים. המודל קיבל את התיק הרפואי של חולים מן העבר (לדוגמה, מלפני שנה) ואת הנתונים על חולים אלו במהלך השנה העוקבת כדי להחליט מי מהם באמת בסיכון גבוה (Obermeyer et al., 2019).

ההחלטה מיהו חולה בסיכון גבוה הנדרש להצטרף לתוכנית התמיכה מורכבת גם בהינתן הנתונים ה"עתידיים" של התיק הרפואי של החולים בשנה העוקבת. במקרה זה נבחרו ההוצאות הרפואיות של החולים כמדד לכמה החולים בסיכון. כלומר המודל קיבל זוגות של נתונים: התיק הרפואי של החולה משנה מסוימת וההוצאות הרפואיות של החברה עליו בשנה העוקבת, בתור ערך האמת. יוצרי המודל הניחו כי חולים בסיכון גבוה צורכים שירותים רפואיים רבים, ולכן בחינת ההוצאות הרפואיות תשמש אינדיקציה טובה למידה שבה הם מצויים בסיכון. לאחר השימוש במודל בחנו החוקרים את ביצועיו ואת דירוגיו על חולים. הם גילו שאפרו-אמריקנים בארצות הברית הופלו לרעה: בממוצע, המודל העריך כי מבוטחים מקבוצה זו באוכלוסייה חולים פחות ממבוטחים לבנים. החוקרים בדקו וראו כי מאחר שקבוצה זו ענייה יותר, בממוצע, היא צורכת פחות שירותי בריאות, ולכן המודל הסיק שמבוטחים אלו חולים פחות. לכן החברה הציעה להם תמיכה פחותה, והדבר פגע בהם עוד יותר (Obermeyer et al., 2019).

דוגמה נוספות לכך שבחירה בערך אמת בנתונים יכולה ליצור הטיות במודל היא שימוש במתייג אנושי להגדרת ערך האמת. שימוש זה נעשה בתיוג הבעות פנים, בבחינת אלימות בטקסט, ועוד. צ'ן וג'ו (Chen & Joo, 2021) בדקו אם יש הטיות שיטתיות במתייגים אנושיים בתחום הבעות פנים ומצאו כי גברים נוטים לתייג נשים כשמחות יותר מגברים שלהם הבעת פנים דומה. יש לציין שהבדיקה שלהם השוותה בין תוצאות של מתייגים אנושיים ובין מודל בינה מלאכותית שמדד את תנועת השרירים, וההנחה היא שגם אם למודל יש טעויות, הן אינן עקביות לטובת נשים או גברים.

הטיה הנובעת מבחירה בערך האמת עשויה להיות קשה לזיהוי בבדיקה של המודל, מה שהופך אותה לבעייתית במיוחד לעומת הטיות אחרות. כאשר בודקים את ביצועי המודל משתמשים באותו ערך אמת שהשתמשו בו בתהליך הלמידה, אך לא ברור אם זה מתייג אנושי או אם זו פונקציה של המשתנים. הבחירה שנעשתה בהגדרת ערך האמת הופכת שקופה ולא נבדקת כאשר בודקים את נכונות המודל ואת ההוגנות שלו. כדי למצוא הטיות שנובעות מהטיות בהגדרת ערך האמת לא מספיק להתבונן בנתונים עצמם אלא צריך לבדוק גם כיצד הוגדר ערך האמת ואם ייתכן שהוא מכיל הטיות.

9. סיכונים בשימוש באלגוריתמים של בינה מלאכותית

כאמור, הטיות אלגוריתמיות אינן רק כשל טכנולוגי, אלא הן טומנות בחובן פוטנציאל לסיכונים ממשיים בהוגנות בשירותים במימון ציבורי. לכן בצד אבחון ההטיות והבנת הגורמים להיווצרותן, נדרש גם להעריך את הסיכונים שהן עלולות ליצור ולהתמודד עימם באופן יזום, שיטתי ואחראי.

סיכון, בהקשר זה, הוא תרחיש עתידי לא ודאי אשר השפעתו שלילית. מדובר באירוע פוטנציאלי שנמדד על פי שני מדדים: הסתברות – עד כמה סביר שהוא יתממש, והשפעה – מה תהיה חומרת הפגיעה אם יתממש .(Klinke & Renn, 2021) ניהול סיכונים היא שיטה לקבלת החלטות בתנאי אי-ודאות, והיא משמשת בסיס עיקרי למדיניות מבוססת ראיות (מור, 2018). שיטה זו מסייעת לגופים ציבוריים ולארגונים למקד את פעילותם בהתמודדות עם הסיכונים הכבדים ביותר באמצעות שימוש במסגרת שיטתית (דולב ואח', 2021).

לפיכך, בתהליך של ניהול הסיכונים במערכות המבוססות בינה מלאכותית יש תחילה לבצע זיהוי שיטתי של מוקדי סיכון פוטנציאליים לפגיעה בהוגנות לאורך כל שלבי הפיתוח של המערכת. זיהוי זה מתמקד באופן שבו החלטות תכנוניות ואלגוריתמיות עלולות לייצר הבחנה לא מוצדקת, להעמיק פערים קיימים או לפגוע פגיעה שיטתית בקבוצות מסוימות באוכלוסייה. בהתאם לכך נבחנים שלבים מרכזיים בפיתוח המערכת ובהפעלתה, ובהם איסוף ועיבוד נתונים, פיתוח ואימון המודל, שימוש בו בפועל ובקרה על תוצרי המערכת. לאחר זיהוי הסיכונים יש לאפיין כל אחד מהם לפי רמת ההסתברות וההשפעה ולמיין את הסיכונים על פי חומרת הפגיעה האפשרית בעקרונות ההוגנות.

מודל מקובל למיון סיכונים הוא "מודל הרמזור" ובו שלוש רמות סיכון (Klinke & Renn, 2002, 2012, 2021). בהקשר של הוגנות ומערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי, המודל מדרג את רמת הפגיעה האפשרית בקבוצות באוכלוסייה:

- סיכון מקובל – סיכון שהשפעתו על ההוגנות זניחה, הסתברותו נמוכה ואינו מחייב פעולה ייעודית
- סיכון נסבל – סיכון שעלול להביא לפגיעה ממשית בהוגנות, אך בהסתברות נמוכה יחסית, ולכן יש לצמצמו ולנתרו
- סיכון בלתי נסבל – סיכון שעלול להוביל לפגיעה שיטתית או חמורה בעקרונות ההוגנות או בזכויות יסוד, ולכן אי אפשר לקבלו גם כאשר ההסתברות להתממשותו נמוכה

חשוב להבין כי מערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי אינן מערכות גנריות. מן הדיון לעיל עולה כי הן מפותחות כדי לתת מענה לבעיה ציבורית או חברתית מסוימת, אם בקביעת זכאות לשירות ואם בהקצאת משאבים, וכל אחת מהן מעוגנת בהקשר מוסדי וחברתי.

לפיכך אי אפשר לאפיין את הסיכונים הנלווים להפעלת המערכת בכל הנוגע להוגנות לפני הגדרה ברורה של המטרה שלשמה היא נועדה, האוכלוסייה שהיא תשפיע עליה וההקשר שבו היא תיושם. משמעות הדבר היא שניהול סיכוני הוגנות במערכות מבוססות בינה מלאכותית אינו יכול להתבצע כשלב טכני נפרד או כבדיקה בדיעבד אלא חייב להיות מוטמע בתוך עיצוב המערכת עצמה. רק לאחר שהוגדרה הבעיה החברתית שהמערכת מבקשת לפתור והוגדרו עקרונות ההוגנות הרלוונטיים אפשר לאפיין את הסיכונים האלגוריתמיים הכרוכים בהפעלתה ולעצב מנגנוני התמודדות מותאמים. ניהול הסיכונים אפוא, בהקשר זה, אינו רק אמצעי לצמצום נזקים אפשריים אלא רכיב מרכזי בתהליך האתי והמעשי של פיתוח והטמעה של מערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי.

שיטה זו לניהול סיכונים עולה בקנה אחד עם המדריך הממשלתי לשימוש אחראי בכלי בינה מלאכותית שפרסם מערך הדיגיטל הלאומי ביוני 2025. המדריך מגדיר שמירה על זכויות אדם, שוויון ומניעת אפליה כעקרונות ליבה ומחייב התייחסות לסיכונים חברתיים, כגון הטיות, כחלק אינטגרלי מן התהליך. הוא מתווה מתודולוגיה סדורה ובה ארבעה שלבים: ניתוח והערכה של תועלות וסיכונים; תכנון פעולות להפחתת הסיכון; קבלת החלטה; ויישום בפועל. במסגרת זו מוצע סיווג דיפרנציאלי של הסיכונים לחמש רמות חומרה (מנמוכה מאוד ועד גבוהה מאוד), כדי להתאים את היקף אמצעי הבקרה לרמת הסיכון הנשקף מן המערכת (מערך הדיגיטל הלאומי, משרד המשפטים ומשרד החדשנות, המדע והטכנולוגיה, 2025).

בפרק זה נסקור את הסיכונים הנובעים מן ההטיות האלגוריתמיות שפורטו בפרק 8. כדי להבין את הסיכונים הנובעים מן ההטיות קיבצנו אותם לפי נושאים: סיכונים בקבלת החלטה עצמאית על ידי מודל, סיכונים בקבלת החלטות בעזרת מודל מייעץ, סיכונים ביצירת תוכן על ידי מודלים של בינה מלאכותית וסיכונים בעקבות אופן השימוש במודל. חשוב לציין שייתכן שלהטיות שונות תהיה השפעה זהה, ובמקרה כזה אפשר לנתח את הסיכון באותו אופן.

9.1 סיכונים בקבלת החלטה עצמאית על ידי מודל

סיכון עקב מודל מוטה או לא מדויק

כאמור, הטיות הקשורות בנתונים או בהחלטות אנושיות ביצירת המודל עלולות לגרום למודל המוטה כלפי קבוצה מסוימת באוכלוסייה או למודל מדויק פחות בנוגע לקבוצות באוכלוסייה, והדבר יוצר סיכון לפגיעה אפשרית בהן.

גם כאשר המודל אינו מוטה במובהק אך רמת הדיוק שלו נמוכה, עלולות להיווצר פגיעות בקבוצות מסוימות באוכלוסייה. במאמרם דנו בקיני ולורוסו (Bacchini & Lorusso, 2019) במקרה של אלגוריתמים לזיהוי פנים שנמצאים בשימוש על ידי המשטרה. אלגוריתמים אלו מנסים להתאים תמונה של אדם לתמונות המופיעות במאגר של אנשים שנעצרו בעבר, ואם

יש התאמה, אזי אדם זה חשוד. נמצא כי אלגוריתמים אלו טועים יותר בזיהוי של אפרו-אמריקנים בארצות הברית, ועקב כך גורמים ליותר עיכובים לא מוצדקים של אזרחים אלו. הכותבים אף הדגישו כי מלבד דיוק המודל, יש גורמים נוספים המובילים לפגיעה באוכלוסייה זו בעקבות הפעלתו.

סיכון ביצירת המלצות מותאמות אישית על ידי מודל מוטה או לא מדויק

אחד השימושים המרכזיים במודלים של בינה מלאכותית הוא יצירת המלצות מותאמות אישית למקבלי שירותים ועבורם. אלגוריתמים של המלצות (recommendation systems) נמצאים באתרים רבים במרשתת, כגון רשתות חברתיות, אתרי חדשות ואתרים לצריכת תוכן (Ricci et al., 2022). עם הרחבת השימוש במערכות מבוססות בינה מלאכותית, ייתכן שאלגוריתמים כאלו יוטמעו גם באתרים הנותנים שירותים במימון ציבורי, כגון בריאות וחינוך. כמה סיכונים עלולים לנבוע משימוש כזה:

- **דיוק נמוך של המודל בנוגע לאוכלוסיות מיעוט**. אחד מן הסיכונים המרכזיים בהמלצות מותאמות אישית הוא התאמה חלקית או לא מספקת לקבוצות מיעוט באוכלוסייה, מה שעלול להפוך את ההמלצות לרלוונטיות פחות או מועילות פחות עבורן. מצב זה נובע מכך שהמודל נוטה ללמוד במדויק את דפוסי ההתנהגות וההעדפות של קבוצת הרוב, אך מתקשה ללמוד את העדפותיהן של קבוצות מיעוט, המיוצגות פחות בנתונים (Li et al., 2021). לכן ההמלצות המתקבלות עבור קבוצות אלו עלולות להיות שגויות ולא מותאמות. אקסטרנד ואח' (Ekstrand et al., 2018) בחנו כמה אלגוריתמים של יצירת המלצות על נתוני אמת והראו שבמצבים מסוימים האלגוריתמים נוטים להתעלם מקבוצות מיעוט ובכך מספקים תוצאות שמתאימות להן פחות.
- **הסללה של מקבלי השירות**. סיכון נוסף הוא הסללה של מקבלי השירות על ידי יצירת המלצות טיפוסיות לקבוצה הדמוגרפית שאליה הם שייכים. יאו וחואנג (Yao & Huang, 2017) הציגו את הסיכון בשימוש באלגוריתמים של המלצות המציעים למשתמשים קורסים שבהם ייתכן שיתעניינו. במצב זה יש סיכון שאלגוריתמים אלו לא יציגו לנשים קורסים הנדסיים, למשל, עקב הטיה הנובעת מדפוסי שימוש קבוצתיים.
- **יצירת תיבות תהודה**. סיכון נוסף שעלול לנבוע משימוש באלגוריתמים של המלצות הוא יצירת "תיבות תהודה" (echo chambers). מצב כזה מתרחש כאשר משתמשים נחשפים אך ורק למידע התואם את השקפת עולמם והעדפותיהם הידועות ולא חושף אותם לתחומים חדשים. נורדה ואח' (Noordeh et al., 2020) בחנו אלגוריתמים ליצירת המלצות והראו כיצד נוצרות תיבות תהודה על ידי סיווג המשתמשים לקבוצות והתאמת ההמלצות למשתמש לפי הקבוצה שאליה הוא שייך. מצב כזה מונע מן המשתמש להיחשף לתחומים חדשים ואף עלול לפגוע בחברה על ידי יצירת קיטוב חברתי או פוליטי (Quattrociocchi et al., 2016).

- **סיכון לפגיעה במושאי החיפוש**. פרט לסיכונים הנובעים מחוסר הוגנות של האלגוריתמים כלפי מקבלי השירות, במקרים רבים טמון סיכון לפגיעה גם במושאי ההמלצה. לדוגמה, מערכת הממליצה על יצירות תוכן עלולה לפגוע באומנים יוצרי התוכן אם בעקבות הטיה היא לא תמליץ על תוכן של אמנים מסוימים. מקור אפשרי לחוסר הוגנות כלפי מושאי ההמלצה הוא הנטייה של אלגוריתמים של המלצות להמליץ יותר על תוכן פופולרי (popularity bias) – הטיה הנובעת מחוסר דיוק של המודל שפוגע הן במקבלי השירות, שלא נחשפים לתוכן שבו הם מעוניינים, הן ביוצרי תוכן נישתי שלא מקבלים חשיפה (Abdollahpouri et al., 2020). הטיה זו עלולה לפגוע גם בבעלי עסקים קטנים, שכן המודל ממליץ בעיקר על עסקים פופולריים.

9.2 סיכונים בקבלת החלטות בעזרת מודל מייעץ

במצבים שבהם מתקבלות החלטות בעלות השפעה מכרעת על מקבלי השירות, בתחומים כמו רפואה או משפט, נהוג להותיר את ההחלטה הסופית בידי גורם מקצועי תוך הסתמכות מסוימת על פלט המודל כחלק ממכלול השיקולים. מודל מוטה או לא מדויק עלול להטות גם את שיקול הדעת של מקבל ההחלטה, ובכך להביא לפגיעה בקבוצות מסוימות באוכלוסייה. מודל להערכת סיכון שאינו שקוף (כלומר פועל כ"קופסה שחורה" ואינו מספק הסבר לפלט שלו) מספק למקבלי ההחלטות מידע חלקי – הם לא יודעים מה השפיע על הסיכון שהמודל העריך (Alon-Barkat & Busuioc, 2023).

פרט לסיכון שנובע מקבלת החלטות בהסתמך על מודל מוטה, יש סיכון להיווצרות הטיה גם עקב הדרך שבה מקבל ההחלטות המקצועי מסתמך על המלצות המודל, אפילו כאשר המודל עצמו אינו מוטה. דוגמה מייצגת לסיכון זה היא מודל COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) האמריקני – מודל בינה מלאכותית מייעץ בעל השפעה רבה על חיי אדם. המודל מצוי בשימוש רב במדינות בארצות הברית ומשמש בבתי משפט להערכת סיכון שעציר או אסיר יפשע שוב או לא יקיים את תנאי השחרור. למודל מוזן מידע אישי על העציר או האסיר וכן תשובות לשאלון שהוא ממלא, והוא מחזיר הערכה של הסיכון. שופטים משתמשים בהמלצות המודל כחלק ממערך השיקולים לקבלת החלטה אם לשחרר בערבות עצירים או לשחרר אסירים שחרור מוקדם (Northpointe, 2015).

הביקורת המרכזית על מודל COMPAS עלתה בשנת 2016 בעקבות תחקיר של המגזין האמריקני ProPublica (Angwin et al., 2016) שטען כי המודל מפלה לרעה נאשמים אפרו-אמריקנים. על פי ממצאי התחקיר, המודל ניבא בטעות בסבירות גבוהה יותר כי נאשמים מאוכלוסייה זו ישובו לבצע עבירות, אף שבפועל זה לא קרה. כלומר שיעור השגיאות מסוג חיובי שגוי (false positives), כאשר המודל מסווג אדם כחיובי, משמע יחזור לפשוע, ובפועל הוא לא חוזר לפשיעה, היה גבוה יותר בקרב אפרו-אמריקנים מבקרב לבנים. משמעות הדבר היא

שנאשמים אלו נפגעו יותר מהטיות המודל. מנגד מתכנני המודל טענו כי המדד להוגנות שבו השתמשו מבקריו אינו מתאים להקשר של הערכת סיכון. לשיטתם יש לבחון אם בקרב מי שסווגו כבעלי סיכון גבוה, הן אפרו-אמריקנים הן לבנים, שיעור החזרה לפשיעה דומה. במילים אחרות, ההוגנות נבחנת לפי כיול (calibration) התחזית בתוך כל רמת סיכון, ולא לפי סוגי השגיאה. הדיון מעלה שאלה עקרונית: כיצד נכון להגדיר הוגנות באלגוריתמים?

מחקר מאוחר יותר של קליינברג ואח' (Kleinberg et al., 2018) העיד על סתירה מתמטית בין שתי הגדרות של הוגנות: (1) ProPublica הגדירו הוגנות כשוויון בשיעורי השגיאה בין קבוצות – שיעור התחזיות השגויות יהיה זהה בין אפרו-אמריקנים ובין לבנים בכל אחד מסוגי השגיאות (חיובי שגוי ושלילי שגוי); (2) מפתחי המודל הגדירו הוגנות כשוויון בדיוק של הסיכון – כאשר המודל מזהה אדם כבעל סיכון גבוה, הסבירות שהוא אכן יחזור לפשיעה תהיה שווה בין קבוצות באוכלוסייה. כאשר שיעורי הפשיעה בפועל שונים בין קבוצות, אי אפשר לקיים את שתי ההגדרות הללו בו-זמנית. לכן השאלה אינה רק טכנית אלא נורמטיבית: איזו הגדרה של הוגנות צריכה להנחות את עיצוב המודל – שוויון בתוצאה או שוויון בדיוק?

כאמור, מודל COMPAS הוא מודל מייעץ המשמש כלי עזר לשופטים. לכן יש להביא בחשבון לא רק את הוגנות המודל כשלעצמו אלא גם את הדרך שבה השופטים משתמשים בו. גם כאשר המודל עצמו אינו מוטה, היישום והשימוש בו יכולים לייצר פגיעה באוכלוסיות מסוימות. גרין וצ'ן (Green & Chen, 2019) בחנו כיצד אנשים שאינם במקצועות השיפוט (ולא שופטים שקיבלו הכשרה לשימוש במודל) משתמשים בהמלצות של אלגוריתמים ומצאו כי כאשר מבקשים מהם לדרג סיכון של עצירים, הם משתמשים במודל שימוש לא אחיד. בממוצע, המדרגים העלו את הסיכון עבור אסירים ממוצא אפרו-אמריקני, בעיקר בעלי הרשעות קודמות. אימוץ סלקטיבי זה יוצר הטיה, מאחר שהסיכון, שהוא הפלט של המודל, כבר הביא בחשבון את ההרשעות הקודמות בהערכת הסיכון שלו.

אחת הסיבות המרכזיות לביקורת על מודל COMPAS היא שהוא אינו חשוף לציבור והוא מהווה מעין "קופסה שחורה" שבה לא ידוע באיזה אופן המאפיינים של כל אסיר או עציר משפיעים על הערכת הסיכון שלו. במצב זה האסירים והעצירים לא יודעים מה הסיבות להערכת הסיכון, וגם השופטים שמקבלים את ההערכה לא יודעים זאת (Rudin et al., 2020). דרסל ופריד (Dressel & Farid, 2018) הראו שאפשר ליצור מודל פשוט ושקוף שמגיע לתוצאות דומות עבור נאשמים על ידי שימוש במספר מועט של משתנים. אף שגם במודל זה יש אפשרות להטיות ולסיכונים, הוא שקוף ואפשר לבקרו.

באשר לדיון בשימוש באלגוריתמים להערכת הסיכון כדי לסייע לשופטים בהחלטה על שחרור אסירים או עצירים, יש המתנגדים לשימוש אלגוריתמי במצב זה ויש התומכים בו, ועולה צורך ליצור איזון בין ההוגנות והזכויות של האסירים והעצירים ובין תועלת הציבור והסכנה הפוטנציאלית משחרור אדם שיכול לבצע פשעים. הביקורת על המודל נעוצה בחלקה בחוסר

השקיפות שלו, מאחר שהוא לא חשוף לציבור. נוסף על כך מאחר שזהו מודל מייעץ, יש מורכבות הקשורה לדרך שבה השופטים משתמשים בתוצאותיו ולפוטנציאל להטיות בתהליך זה (Dressel & Farid, 2018; Rudin et al., 2020).

9.3 סיכונים ביצירת תוכן על ידי מודלים של בינה מלאכותית

מודלים של בינה מלאכותית משמשים ליצירת תוכן מגוון ויש להם שימושים רבים, ובהם חיפוש, סיכום, כתיבת תוכן ויצירת תמונות. יצירת תוכן במודלים אלו כרוך בסיכונים, ואלו קשורים בין היתר לפערים בביצוע או לשימוש שיכול ליצור הטיות. במקרים רבים יצרני המודלים הגנרטיביים מעדכנים את המודלים כדי לשפר אותם מבחינת ביצועים, בטיחות והוגנות. אף שלרוב עדכונים אלו משפרים את המודלים, הם יכולים לפגוע בהוגנות. לכן אם מודל שפה נמצא בשימוש במערכת כלשהי ונבדק עבור סטראוטיפים או דרישות הוגנות מסוימות, יש לבדוק אותו שוב לאחר עדכונו.

פערי ביצועים בין מודלים ליצירת תוכן בשפות שונות או בקבוצות באוכלוסייה שייצוגן במרשתת מצומצם

כאמור לעיל (פרק 8) יש פערים גדולים בביצועים בין מודלי שפה הפועלים בשפות שונות או כאשר המידע במרשתת על קבוצות שונות באוכלוסייה מצומצם. פערים אלו יוצרים סיכון כלפי המשתמשים דוברי שפות אלו, מאחר שהם לא יכולים ליהנות מן היתרונות של מודלי השפה וגם עלולים לקבל מידע מוטעה ומדויק פחות.

סטראוטיפים במודלים ליצירת תוכן

ייצוג חסר במרשתת מתאר היעדר מידע או נראות של קבוצות באוכלוסייה, ואילו סטראוטיפים עוסקים באופן שבו קבוצות אלו מיוצגות. הדוגמאות המובאות להלן מבטאות את היקף התופעה של ייצוג חסר ואת מגוון הקבוצות העלולות להיפגע מכך.

- **סטראוטיפים מגדריים**. מחקרי United Nations Educational, Scientific and Cultural Organization [UNESCO], 2024) UNESCO) מצאו כי מודלי שפה גדולים נוטים לקשר בין נשים ובין תחומי בית ומשפחה, ולעומת זאת לקשור בין גברים לתחומי עסקים וקריירה. מחקר שנערך על ידי אל-דאהול ואח' (AlDahoul et al., 2025) בדק דימויים שנוצרו על ידי מודל יצירת תמונות עבור עשרות מקצועות ומצאו כי במקצועות טכנולוגיים כמו תכנות או הנדסה, כל הדמויות שנוצרו היו גברים, ברובם לבנים. לעומת זאת במקצועות הטיפול, כמו סיעוד או דיילות, כל הדמויות היו נשים. דימויים אלו אינם מייצגים את השיעור האמיתי של נשים וגברים במקצועות אלו, ועל כן מדובר בהקצנה של סטריאוטיפים מגדריים.

- **סטראוטיפים אתניים**. במודלים גנרטיביים עלולות להופיע הטיות העולות בקנה אחד עם סטראוטיפים אתניים. מחקר שבוצע באוניברסיטת וירצבורג-שוינפורט (THWS) בגרמניה מצא כי צ'אטבוטים המבוססים על מודלי שפה גדולים כמו ChatGPT, Claude, ו-LLaMA מספקים המלצות שכר נמוכות משמעותית לנשים ולאנשים מקבוצות מיעוט אתניות. לדוגמה, לדמות של גבר לבן הומלץ לבקש שכר של 400,000 דולר לשנה ואילו לדמות של אישה לבנה הומלץ לבקש 280,000 דולר לשנה בלבד (Sorokovikova et al., 2025). כלומר גם כשהמועמדים זהים במיומנות ובניסיון, המועמדת או המשתמשת שמזוהה עם מיעוט אתני מקבלת שיטתית המלצת שכר נמוכה יותר. המחקר העיד על אפליה סמויה שמבוססת על הנחות סטראוטיפיות השזורות בנתונים שעליהם אומנו המודלים.
- **הטיות דתיות ותרבותיות**. מודלים גנרטיביים עלולים לשקף דעות קדומות וסטריאוטיפים גם בנוגע לקבוצות דתיות, לאומיות ותרבותיות. דוח שפרסמה הליגה נגד השמצה (ADL) חשף כי ארבעה ממודלי השפה הגדולים הראו הטיה נגד יהודים וישראלים. לדוגמה, המודל Llama סיפק תשובות שגויות ומוטות בנושאים הנוגעים לעם היהודי ולישראל (Anti-Defamation League [ADL], 2025).
- **הטיות על בסיס מעמד חברתי ומוגבלות**. מחקרם של אורבינה ואח' (Urbina et al., 2025) בדק את האופן שבו מודלים של בינה מלאכותית מתארים אנשים עם מוגבלות לעומת אנשים ללא מוגבלות. החוקרים ביקשו מן המערכות לתאר אנשים עם מוגבלות, מטופלים עם מוגבלות וספורטאים עם מוגבלות ובדקו אם ייחסו להם תכונות חיוביות או שליליות יחסית לתיאורים של אנשים ללא מוגבלות. תוצאות המחקר הראו כי בשני המודלים בתיאורים של אנשים עם מוגבלות הוצגו פחות תכונות חיוביות ויותר תיאורים של מגבלות לעומת בתיאורים של אנשים ללא מוגבלות.

הטמעת סטראוטיפים במודלים גנרטיביים טומנת בחובה סיכון ממשי לפגיעות ברמות שונות: אישית, קבוצתית וחברתית. ברמת הפרט, ייצוגים סטראוטיפיים עשויים לפגוע בכבוד האדם, לערער את תחושת הערך העצמי של המשתייכים לקבוצות מוחלשות ולחזק תחושת ניכור כלפי תוצרים מבוססי בינה מלאכותית (Hutchinson et al., 2020). כך למשל, הדוגמה שצוינה קודם לכן בהקשר להטיות על בסיס מעמד חברתי ומוגבלות מדגימה מקרה של "ability bias" – הטיה המתייחסת לנטייה של מערכות מבוססות בינה מלאכותית לשקף או לשעתק תפיסות שליליות כלפי אנשים עם מוגבלות, באופן שפוגע בייצוג שלהם ובאינטראקציה עימם. הטיה זו עלולה לפגוע בתחושת הערך שלהם ובתפקוד שלהם בעת השימוש במערכות אלו (Urbina et al., 2025).

מחקר נוסף הראה כי נשים שנחשפות לתכנים הממחזרים סטראוטיפים מגדריים חוות פגיעה ניכרת, כלומר עצם החשיפה לתכנים מוטים יוצרת חוויה של השפלה או אפליה. ברמת הקבוצה, הפצה שיטתית של ייצוגים מוטים תורמת להנצחת סטיגמות קיימות ועלולה לעצב תפיסות ציבוריות מעוותות כלפי קבוצות אתניות, מגדריות, דתיות או אחרות (Wang et al., 2025).

סקירה של פראראה (Ferrara, 2024) הראתה כי הטיות במערכות מבוססות בינה מלאכותית עלולות להעצים פערים חברתיים קיימים, במיוחד על רקע גזע, מגדר, גיל או מעמד חברתי-כלכלי. ברמה החברתית והמערכתית, יש סכנה להעמקת דפוסי הדרה ואפליה בצד ערעור אמון הציבור בטכנולוגיה. כאשר מערכות נתפסות כבלתי הוגנות או פוגעניות, הן עלולות לאבד את הלגיטימציה הציבורית ולעכב שילוב אחראי של בינה מלאכותית בתחומים חשובים.

9.4 סיכונים בעקבות אופן השימוש במודל

בחלק זה נבחן כיצד אופן השימוש במודל עשוי לפגוע בקבוצות באוכלוסייה. כפי שראינו, למודלים של למידת מכונה יש שימושים רבים, ובהם קבלת ההחלטה עצמה, ייעוץ למומחה אנושי ויצירת תוכן. כאשר אנחנו בוחנים מודל, עלינו לבחון לא רק את המודל בפני עצמו אלא גם את אופן השימוש בו בפועל.

סיכון עקב פער בנגישות לטכנולוגיה

יש שימושים רבים במודלים של בינה מלאכותית. הם יכולים לסייע בניסוח מכתב, לתמוך בקבלת החלטות, לייצר תמונות, לסנן תוכן לא רלוונטי, ועוד. כדי ליהנות מן היתרונות של מודלים אלו, על המשתמש להיות בעל יכולות טכנולוגיות בסיסיות ולדעת את השפה שבה פועל המודל. פרט לכך צריכה להיות לו גישה למחשב, לאינטרנט ולמודל עצמו. עד לפני כמה שנים מודלים של בינה מלאכותית היו בשימוש בעיקר על ידי אנשי מקצוע בארגונים שונים כמו למשל מודלי סיכון של מחלות, בשירותים פיננסיים ובמקומות עבודה נוספים. כיום יש מודלים של בינה מלאכותית מסוגים שונים, והם בשימוש לא רק של אנשי מקצוע אלא גם של האוכלוסייה הכללית. לכן לפערים בנגישות למחשוב יש השפעה גדולה מבעבר (Costa et al., 2024).

בישראל יש פערים גדולים בנגישות לטכנולוגיה בין קבוצות שונות באוכלוסייה. לדוגמה, מבחינת נגישות למחשבים בבתי הספר, יש פער ביחס שבין מספר תלמידים למחשב בין בתי ספר באשכולות חברתיים-כלכליים גבוהים לבתי ספר באשכולות חברתיים-כלכליים נמוכים (Schejter & Tirosh, 2016). למרות הנוכחות הגוברת של טלפונים ניידים בקרב תלמידים, מחשבים עדיין משמשים כלי חשוב ליצירת תוכן ומטלות, כמו הכנת מצגות או כתיבת קוד. באוכלוסייה הערבית, בכ-50% מבתי הספר העל-יסודיים יש מחשב אחד לכל 20 תלמידים או יותר, ואילו רק בכ-14% מבתי הספר היהודיים היחס הוא כזה (מבקר המדינה, 2021). פערים אלו עלולים לפגוע בגישה של תלמידים ליכולות החדשות של מודלים של בינה מלאכותית, בהינתן שלא לכל התלמידים יש גישה למחשב בביתם (Schejter & Tirosh, 2016).

סיכון עקב הליך מרובה שלבים

במערכות קבלת החלטות רבות ישנם כמה שלבים, ולמידת מכונה היא רק רכיב אחד בהם. שלבים אלו כוללים לעיתים גם החלטות אנושיות, מנגנוני סינון נוספים, או בחירות של המשתמשים עצמם. במצבים כאלה, בחינת ההוגנות של מודל הבינה המלאכותית כשלעצמו אינה מספיקה; כדי להבטיח הוגנות יש לבחון את התהליך המלא ואת האינטראקציות בין שלביו.

דוגמה לכך מוצגת במאמרם של דוורק ואילבנטו (Dwork & Ilvento ,2019). החוקרים מתארים מצב שבו כמה אוניברסיטאות מפעילות מודלים להערכת מועמדים על בסיס סיכויי ההצלחה הצפויים שלהם. ההנחה היא כי כל מודל פועל בהוגנות ובעקביות ביחס למידת הדמיון בין מועמדים. אף על פי כן, מאחר שכל מודל מופעל בנפרד וכל אוניברסיטה מקבלת החלטות באופן עצמאי, התוצאה המערכתית עשויה להיות לא הוגנת.

בפרט, גם כאשר שני מועמדים בעלי מאפיינים דומים מקבלים הסתברות דומה להתקבל לכל מוסד בודד, מועמד מרקע כלכלי חזק יכול להרשות לעצמו להגיש מועמדות למספר גדול יותר של מוסדות. בכך הוא מגדיל את ההסתברות הכוללת להתקבל לפחות לאחד מהם. פער זה, שאינו נובע ממודל הקבלה עצמו אלא מהבדלים ביכולת להגיש מספר רב של בקשות, עלול להטריד במיוחד כאשר הוא משקף פערים חברתיים-כלכליים רחבים יותר. במצב כזה, גם אם כל אחד מן המודלים הבודדים שומר על הוגנות כלפי מועמדים מרקע חברתי-כלכלי נמוך, התוצאה הסופית – שיעור הקבלה המצטבר – עלולה לשקף אי-שוויון שנוצר בשל פערי נגישות ולא בשל תפקוד המודלים עצמם. הדוגמה אינה ייחודית לתהליכי קבלה להשכלה גבוהה, והיא רלוונטית למגוון מצבים שבהם משתמשים נדרשים לבחור בין כמה אפשרויות, אך אינם נהנים מאותה רמת גישה לכל האפשרויות.

9.4.3 סיכון עקב מונוליטיות של מודלים

פיתוח מודלים המיועדים לקבל החלטות גורליות עשוי להיות יקר ומסובך, שכן לשם כך נדרש מידע איכותי רב ומומחיות בתחום. לכן במקרים רבים יש גופים אשר מסתמכים על אותם מודלים, לדוגמה כאשר חברות משלמות לאותה חברה חיצונית בעבור שירותים מסוימים. מצב זה, המכונה מונוליטיות של אלגוריתמים (algorithmic monoculture), יכול לגרום לביצועים לא מיטביים ולחוסר הוגנות (Kleinberg & Raghavan, 2021). הכוונה היא לא רק ששימוש במודל מוטה יכול להגדיל את הפגיעה אלא שיש סיכון בעצם השימוש במודל יחיד במקומות רבים.

סיכון זה נובע מן העובדה שמודלים של בינה מלאכותית הם שרירותיים ברמה מסוימת. תהליך האימון הוא אקראי, ויש התאמה של המודל לנתונים הספציפיים שהשתמשו בהם באימון המודל. אפשר לראות זאת כאשר מאמנים מודלים שונים על ידי בחירה במקבץ אקראי של נתונים מתוך מאגר גדול. המודלים לא יהיו זהים אף שאומנו על אותה התפלגות (Riley & Collins, 2023).

דוגמה נוספת אפשר למצוא בסינוני קורות חיים של מועמדים על ידי מודלים של בינה מלאכותית בחברות עסקיות. בכל מודל יש שגיאות, כלומר יש אחוז קטן ואקראי של מועמדים מתאימים שהמודל טועה בנוגע להם וטוען כי אינם מתאימים. לו כל חברה הייתה משתמשת במודל אחר, הרי שכל מודל היה טועה בנוגע למועמדים אחרים, ולכן כל מועמד מתאים היה מצליח לעבור את הסינון בכמה מן החברות ולהתקבל למקום עבודה מסוים. אולם במצב שבו כל החברות משתמשות באותו המודל, מועמד מתאים יידחה בכולן ולא יצליח למצוא מקום עבודה (Bommasani et al., 2022).

10. פתרונות לקידום הוגנות אלגוריתמית

בפרקים הקודמים עמדנו על ההטיות האפשריות במודלים של בינה מלאכותית ועל הסיכונים הנובעים מהן לאורך שלבי הפיתוח של מערכות מבוססות בינה מלאכותית בשירות הציבורי – מן הנתונים, דרך אלגוריתם הלמידה, ועד לאופני השימוש במודל בפועל. כמו כן בחנו את השפעות ההטיות על החלטות אוטונומיות, על שימוש באלגוריתמים מייעצים ועל יצירת תוכן. בחינה זו העלתה כי אי-הוגנות אלגוריתמית היא תוצר של שילוב בין נתונים חלקיים או מוטים, בחירות עיצוב והגדרת מטרות, הקשרים מוסדיים ופערי נגישות לטכנולוגיה.

בפרק זה יוצג מארג של פתרונות מעשיים שיאפשרו למשרדי הממשלה ולגופים ציבוריים לתכנן, לפתח ולהטמיע מערכות מבוססות בינה מלאכותית באופן שיצמצם הטיות ויקטין סיכונים של פגיעה בהוגנות. מוצעת מסגרת פתרונות דו-ממדית (הממד הראשון מתייחס לשלבי חיי המערכת הטכנולוגית (תכנון מקדים, פיתוח וטיוב, הטמעה ויישום, ניטור ושיפור), והממד השני עוסק בסוגי המענים, והוא כולל שני סוגי מענים משלימים: פתרונות במעטפת השירות הציבורי ופתרונות הקשורים לתהליך יצירת המודל ולהערכתו. פתרונות אלו יכולים לשמש פרקטיקות מיטביות: כלים ותהליכים שנוסו והוכח כי הם מסייעים ליצירת עבודה שקופה והוגנת יותר במערכות שונות.

יש לציין כי יש **שלושה פתרונות מומלצים ליישום** בכל מערכת מבוססת בינה מלאכותית שתוטמע בשירות הציבורי: (1) הגדרת מטרת המודל ואופן השימוש בו; (2) קביעת מטרות הוגנות וגיוון; (3) בדיקת הוגנות יזומה ובקרה מתמשכת (אודיט). מימוש פתרונות אלו יכול לאפשר יישום של פתרונות נוספים הנדרשים למערכת. כלומר יתר הפתרונות חשובים אף הם, אך הם תלויים באופי המערכת, ולכן על כל מערכת לקבוע פתרון המיועד לה. היישום של כל פתרון מוצע תלוי בגורמים נוספים מלבד מימוש עקרון ההוגנות. נוכח זאת נדרש להתאים פתרונות לפי רמת הסיכון, הרגישות והמשאבים של כל מערכת בינה מלאכותית. זאת ועוד, מרבית הפתרונות הנדרשים לקידום הוגנות רלוונטיים במידה דומה הן לפיתוח של מערכות מבוססות בינה מלאכותית חדשות הן לשילוב של כלים קיימים בארגון. **בין המערכת מפותחת מראשיתה ובין מדובר בשירות מדף, נדרש אותו סט של בחינות, תיקופים, בקרות והגדרה מודעת של מטרות ההוגנות.**

כשלב מקדים נדרש גם מיפוי שיטתי של הסיכונים המרכזיים בכל נקודת החלטה (ראו פרק 9). לא כל סיכון נושא משקל זהה, ולכן חיוני להעריך את מידת ההשפעה האפשרית של כל סיכון, את ההסתברות להתממשותו ואת ההשפעות על קבוצות שונות באוכלוסייה. ניהול סיכונים כזה מאפשר לבחור פתרונות ממוקדים יותר, להימנע מהטמעת מנגנונים מכבידים שלא לצורך ולכוון את המאמצים לאזורים שבהם הסיכון לאי-הוגנות גבוה במיוחד. בכך מסגרת הפתרונות נשענת לא רק על עקרונות כלליים של הוגנות אלא גם על הערכת סיכונים מושכלת שמאפשרת התאמה מדויקת של המענים.

חלק מן הפתרונות עלולים, אם ייושמו יישום לא זהיר, דווקא להעמיק פערים ולפגוע בהוגנות. לכן בצד הצגת ההזדמנויות שבכל פתרון יפורטו גם הסיכונים הגלומים בו ויוצעו דרכים לצמצומם, למשל באמצעות ניתוח שיטתי של דפוסי הערעור, מעקב אחר החלטות של גורמים אנושיים והטמעת מנגנוני פיקוח ולמידה מתמשכת. כל פתרון הקשור לתהליך יצירת המודל ולהערכתו מלווה בהנחיות יישומיות המפרטות מתי וכיצד מומלץ להפעילו, ובכך תוסב ההבנה התיאורטית של הטיות וסיכונים לארגז כלים מעשי לתכנון ולניהול אחראי של מערכות מבוססות בינה מלאכותית בשירות הציבורי.

חשוב להצביע על מתח מהותי המלווה את השיח הציבורי בנוגע לבינה מלאכותית: במקרים רבים מצופה שמערכות אלגוריתמיות יעמדו ברמות הוגנות, שקיפות ועקביות גבוהות בהרבה מאלו של החלטות אנושיות. החלטות של פקידים, רופאים, עובדים סוציאליים ואנשי מקצוע אחרים אינן נתונות לבחינה שיטתית של הוגנות בכל מקרה ומקרה, גם כאשר הן מתקבלות בתנאי אי-ודאות ועל בסיס מידע חלקי. דווקא בשל יכולתו של מודל לפעול בהיקפים רחבים ולשעתק הטיות בעקביות, הדרישות ממנו מחמירות יותר, אולם הן מחייבות גם להגדיר מהי רמת המעורבות האנושית הראויה. פיקוח אנושי אינו פתרון אוטומטי, והוא עצמו עלול לשאת הטיות שבהיעדר תכנון נכון ישעתקו אי-הוגנות. לכן בצד פיתוח האלגוריתמים נדרש לעצב גם את מנגנוני הפיקוח האנושי: מה היקף ההתערבות הרצוי, כיצד אפשר להימנע מהסתמכות יתר או חסר, ואיך אפשר להבטיח שמנגנוני הבקרה ישלימו את המערכת ולא ייצרו מקור חדש של אי-הוגנות.

10.1 תכנון מקדים

שלב התכנון המקדים עוסק בהנחת היסודות למערכת: גיבוש מטרות, קביעת יישומים רצויים, זיהוי אוכלוסיות היעד והערכת סיכונים ראשונית. זהו שלב שבו המערכת עדיין בשלב הרעיון, וההחלטות המתקבלות בשלב זה מעצבות את כיווני הפיתוח וההוגנות שינחו את התהליך.

10.1.1 פתרונות במעטפת השירות הציבורי

שילוב ועדות מייעצות ובהן נציגי קבוצות מוחלשות באוכלוסייה

שילוב ועדות מייעצות הכוללות נציגות של קבוצות מוחלשות באוכלוסייה (משתמשי קצה) הוא מנגנון מוסדי אחד מבין מגוון מנגנונים של שיתוף ציבור בתהליכי תכנון, פיתוח ואימוץ של מערכות מבוססות בינה מלאכותית, והוא משמש לצמצום הטיות ולחיזוק ההוגנות במערכות רגישות. ועדות אלו יוצרות מסגרת קבועה ומובנית של שיתוף ציבור, ולא רק התייעצות נקודתית: בצד אנשי מקצוע, אנשי טכנולוגיה ונציגי משרדי הממשלה, משתתפים גם מי שמושפעים בפועל מהחלטות האלגוריתם, למשל נציגי ארגוני חברה אזרחית ונציגי קהילות מודרות (בקר, 2019).

בהתבסס על האמור לעיל, ניתן לטעון כי הגדרת מטרות המערכת, בחירת מדדי ההצלחה, קביעת האופן שבו מוגדרים ומוערכים סיכונים ואף ההחלטה אילו שימושים במערכת נחשבים לגיטימיים, אינם נקבעים "מלמעלה" אלא משקפים את צורכיהן של קבוצות באוכלוסייה. שיתוף ציבור במבנה כזה מאפשר לא רק לשמוע דעות אלא גם לעגנן בתהליך קבלת ההחלטות: הוועדות יכולות לבחון תרחישים, להצביע על כשלים צפויים עבור קבוצות באוכלוסייה ולהעלות מראש חששות הנוגעים לנגישות, לשפה או להשפעות לא מכוונות. כדי שמנגנון זה לא יסתכם ב"חותמת גומי", יש להבטיח ייצוג מגוון, פרוטוקולים שקופים ומשקל ממשי לעמדות הוועדה בתהליך האישור של המערכת והמשך הפיקוח עליה. עם זאת חשוב לציין שיישום לא זהיר של מנגנון הוועדות עלול ליצור כשלים חדשים: ייצוג חלקי או סמלי עלול לשמר פערים, ופערי ידע בין המשתתפים עלולים לצמצם את השפעתם של נציגי קבוצות מוחלשות, ולעיתים אף להותירן מחוץ למרחב ההשפעה. פרט לכך, בהיעדר מנגנון מחייב להתחשבות במסקנות הוועדה, יש סכנה שהיא תשמש בעיקר כהליך הצהרתי. בצד זאת יש להכיר בכך שמנגנון זה, כאשר הוא אינו תחום בזמן, עלול להאט את תהליכי הפיתוח של מערכות מבוססות בינה מלאכותית. לפיכך יש לעגן מראש גבולות השפעה, נקודות הכרעה ומנגנוני סיום, על פי רמת הסיכון וההשפעה של המערכת. כדי לצמצם סיכונים אלו יש להבטיח ליווי מקצועי למשתתפים, תמיכה לוגיסטית, מנגנון מפורש המפרט כיצד המלצות הוועדה מתקבלות ושקיפות מלאה של חומרי הרקע והדיונים. שילוב של הגנות על המשתתפים עשוי לסייע להפוך את הוועדה מכלי סמלי למסגרת יעילה שמקדמת הוגנות בפועל.

דוגמה לוועדה כזו היא ועדה שכונסה בעיר ניו יורק בארצות הברית. כחלק מבחינת מדיניות קבלת החלטות אוטומטית (automated decision systems), כונסה ועדה שבה שולבו נציגי חברה אזרחית וקהילות המושפעות בפועל ממערכות אלגוריתמיות. הוועדה בחנה שימושים עירוניים במודלים, הצביעה על כשלים צפויים עבור קבוצות מודרות והדגישה את הצורך בשקיפות, בייצוג מגוון ובזכות לערעור (New York City, Automated Decision Systems Task Force, 2019).

קביעת מדיניות מחייבת לשקיפות ולפרסום מוסדר של שימוש במערכות מבוססות בינה מלאכותית

הגדרת מדיניות מחייבת לשקיפות ולפרסום מוסדר של שימוש במערכות מבוססות בינה מלאכותית היא מנגנון מפתח בשיפור ההוגנות של מערכות אלו בשירות הציבורי והאמון בהן (Nam et al., 2024; OECD, 2025; Schmidthuber et al., 2021). מדיניות כזו מחייבת את הגופים הציבוריים להבהיר מה המערכות עושות, למה הן בשימוש, איך הן מקבלות החלטות והאם יש בדיקות של הוגנות, ניטור ועדכון לאורך זמן.

המדיניות יכולה לכלול רכיבים לפי סוג המערכת (OECD, 2025), לדוגמה:

- דרישה לפרסום קובצי מידע או דוחות ציבוריים שמפרטים שימושים של מערכות – למשל מי פיתח אותם, באיזה שלב הם מופעלים, ומהם המדדים לניטור הוגנות
- קביעת חובת רישום – רישום המערכת במאגר רשמי פתוח לציבור, כך שיהיה ברור באילו מקרים נעשה בו שימוש
- פיתוח מנגנוני שקיפות כלפי משתמשי הקצה והציבור – למשל הודעה למשתמש כי החלטה מסוימת מבוססת על מודל של בינה מלאכותית והצגת הסבר תמציתי ונגיש על הקריטריונים שהובאו בחשבון
- פרסום דוחות תקופתיים – פרסום רשומות על ביצועי המערכת, על השגיאות שנמצאו, על קבוצות באוכלוסייה שבהן הדיוק ירוד ועל תיקוני המדיניות או הקליברציות שבוצעו

בהתאם לכך אפשר לראות כי יישום של מדיניות כזו תומך בהוגנות מכמה בחינות: היישום מגביר את אמון הציבור בכך שהשימוש במודלים של בינה מלאכותית אינו נסתר או בלתי מושג, מאפשר זיהוי מוקדם של הטיות באמצעות מנגנונים של פיקוח ושיתוף ציבור ומייצר מנגנונים לתיקון, לא רק לאחר הפרה של הוגנות אלא כחלק מתהליך שוטף של בקרה ולמידה. זאת ועוד חשוב לדאוג לכך שהפרסומים אכן יהיו נגישים וברורים. כלומר לא לפרסם טקסט משפטי אלא טקסט שמובן לציבור הרחב ולהתמודד עם כמה סיכונים שעלולים להתעורר ביישום לא זהיר: פרסום יתר או פירוט רב מדי עשוי לחשוף בטעות מידע רגיש או לפגוע בפרטיות; העמסה רגולטורית עלולה להכביד במיוחד על גופים חלשים מבחינה טכנולוגית ולהוביל ליישום חלקי; והנגשה שאינה מותאמת לשפות, למוגבלויות או לפערי אוריינות דיגיטלית עלולה להדיר אוכלוסיות שלמות מן היכולת להבין את השימוש במודל. כדי לצמצם סיכונים אלו יש לשלב בחינה מוקדמת של השפעות פרטיות, לייצר פורמטים אחידים וידידותיים, להבטיח תמיכה לגופים קטנים, להפריד בין מידע שיש לפרסם לציבור ובין מידע שיש לחשוף רק לגופי פיקוח ולדאוג להנגשה רב-ערוצית וברורה.

פרקטיקה זו של קביעת מדיניות משקפת סטנדרט מיטבי מתהווה אשר יישומו בפועל עדיין אינו אחיד בין מדינות ומסגרות מוסדיות. דוגמה לכך היא מדיניות השקיפות של ממשלת בריטניה, כפי שגובשה על ידי השירות הדיגיטלי הממשלתי בשנת 2025. המדיניות מחייבת משרדי ממשלה לתעד ולפרסם פרסום מוסדר כל שימוש בכלי קבלת החלטות אלגוריתמיים, כולל רישום של הכלי במאגר פומבי, תיאור מטרותיו ואופן פעולתו והצגת מידע נגיש לציבור על הקריטריונים שבהם נעשה שימוש (Government Digital Service, 2025).

דוגמה נוספת היא המדיניות הפדרלית המחייבת להעריך את ההשפעה האלגוריתמית של מערכות קבלת החלטות אוטומטיות בקנדה. המדיניות מטילה על משרדי הממשלה אחריות לפרסום תוצאות ההערכה הסופיות בפורמט נגיש בפורטל הממשל הפתוח כחלק מתפיסה של הוגנות, שקיפות ואחריות מתמשכת (Government of Canada, 2025).

הכשרות דיגיטליות לקבוצות מוחלשות באוכלוסייה לשימוש במערכות מבוססות בינה מלאכותית

הכשרות דיגיטליות לקבוצות מוחלשות באוכלוסייה לשימוש במערכות מבוססות בינה מלאכותית הן רכיב חיוני בצמצום פערי נגישות ובהבטחת הוגנות במערכות אלו. הן יכולות להיערך בשגרה, ללא תלות בהטמעת מודל ספציפי. בפרט מומלץ שבעת הטמעת מודל המקבל החלטות בעלות השפעות חשובות על פרטים או קבוצות באוכלוסייה, תינתן הכשרה ייעודית לכך. מודלים של בינה מלאכותית הופכים לכלי עבודה מרכזי במרחב הציבורי, אך אוכלוסיות בעלות אוריינות דיגיטלית נמוכה, גישה מוגבלת למחשוב או חסמי שפה מתקשות להפיק מהם תועלת ולעיתים אף ניזוקות משימוש חלקי או שגוי. הכשרות ייעודיות מצמצמות פערי מידע ומחזקות את היכולת של משתמשים מרקע מגוון למצות זכויות ולהתנהל עם מערכות מתקדמות (UNESCO, 2022). חשוב להדגיש כי הכשרות אינן מספקות מענה מלא להבנה של אזרחים כי מתקבלות בנוגע להם החלטות אלגוריתמיות. מענה לכך מחייב מנגנונים מוסדיים של שקיפות, הסבר, ערעור ותיווך. מנגנונים אלו יידונו בהמשך .כמו כן נדרש לשלב בצד ההכשרות גם מנגנונים של הנגשה פרואקטיבית ותיווך אנושי. במקום להמתין שהאזרח יתקשה או יערער, על המערכת לספק הסברים מעשיים וברורים על ההחלטות, בשפה פשוטה ולהבהיר "מה היה צריך להיות אחרת" כדי לקבל תשובה חיובית (Wachter et al., 2017). הסברים כאלו מעניקים למשתמשים ידע מעשי וכוח פעולה.

דוגמה לכך היא תוכנית Digital Skills and Jobs Coalition של האיחוד האירופי, הפועלת לצמצום פערי אוריינות דיגיטלית באמצעות הכשרות מותאמות לקבוצות מוחלשות באוכלוסייה. התוכנית מקדמת שיתופי פעולה בין ממשלות, ארגוני חברה אזרחית וגורמי תעסוקה ומפתחת מודלים של הכשרות שבהן יכולים להשתתף כלל האזרחים (European Commission, n.d.).

דוגמה נוספת היא היוזמה לקידום אוריינות דיגיטלית בקרב אזרחים ותיקים בישראל שגובשה על ידי "ישראל דיגיטלית" בשיתוף ג'וינט-אשל. יוזמה זו מציגה עקרונות להוראה מרחוק של כלים דיגיטליים לאזרחים ותיקים, הנשענים על התאמה שיטתית של תהליכי הלמידה ליכולות, לניסיון החיים ולחסמים הייחודיים של אוכלוסייה זו. עקרונות אלו ממחישים תפיסה מהותית של הוגנות, ולפיה צמצום פערים אינו מושג באמצעות שוויון פורמלי בגישה לטכנולוגיה אלא באמצעות התאמה פעילה של ממשקי השימוש, ההסבר והליווי לקבוצות מוחלשות באוכלוסייה, ובהן אזרחים ותיקים. תפיסה זו רלוונטית במיוחד לשימוש במערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי, שבהם היעדר התאמה מסוג זה עלול להוביל להדרה בפועל, לפגיעה במימוש זכויות ולהעמקת פערים קיימים (ג'וינט ישראל-אשל ומשרד הדיגיטל הלאומי – ישראל דיגיטלית, 2020).

יצירת רישות עסקי (נטוורקינג) בין ארגונים

רישות עסקי בין ארגונים לשיתוף פרקטיקות מיטביות יכול לשמש מנגנון מרכזי להעלאת מודעות לתחום ההוגנות במערכות מבוססות בינה מלאכותית בשירות הציבורי. רשות מקומית, משרד ממשלתי או גוף ציבורי בודד לרוב פועל בתנאי אי-ודאות, עם משאבים מוגבלים ועם ניסיון מועט בהפעלת מערכות כאלה. שיתוף ידע בין גופים שלהם אתגרים דומים מאפשר לזהות מוקדם דפוסים חוזרים של כשלים, לחשוף פתרונות שכבר נבחנו במקומות אחרים ולמנוע שכפול של טעויות (OECD, 2025). רישות עסקי עשוי לכלול פורומים מקצועיים וקבוצות למידה. יתרונו המרכזי טמון בכך שהוא יוצר למידה מוסדית, ולא רק למידה פנימית מצומצמת. לדוגמה, בקנדה יש קבוצות עבודה במשרדי ממשלה בנושאי בינה מלאכותית, והן מאפשרות לעובדי ציבור ממגוון משרדים לשתף חוויות ותובנות ובכך לקדם שיתוף פעולה וחדשנות ביישום מערכות מבוססות של בינה מלאכותית (OECD & UNESCO, 2024).

עם זאת יש להכיר בכך שמדובר בפתרון חלקי בלבד, אף שהוא קל למימוש. כמו כן, ונוכח העובדה שיש ארגונים שעלולים להירתע מחשיפת כשלי עבר או להחזיק ביכולות ובמשאבים שונים, חשוב להבטיח מסגרת מובנית, מתואמת ושוויונית לשיתוף הידע, הכוללת סטנדרטים מינימליים של שקיפות, שמירה על פרטיות ומחויבות להטמעת הלקחים בפועל.

10.1.2 פתרונות הקשורים לתהליך יצירת המודל ולהערכתו

הגדרת מטרת המודל ואופן השימוש בו

כאשר מפתחים מודל בינה מלאכותית או מטמיעים מודל קיים, קודם כל נדרש להגדיר מהן המטרות בהכנסת המודל ובאיזה אופן ישתמשו בו (Raji et al., 2020). למודלים של בינה מלאכותית יכולים להיות שימושים רבים, ויש מקרים רבים שבהם אפשר להשתמש באותו מודל באופנים שונים (Sculley et al., 2015). לדוגמה, אפשר לאמן מודל להעריך על סמך התיק האישי של תלמיד מהו הסיכוי שלו לסיים בגרות בהצלחה. במודל זה אפשר להשתמש כדי לייעץ למורים אילו תלמידים לתגבר או כדי לבחור אילו קורסים מקוונים יוצגו לתלמיד באתר בית הספר. גם במודלי שפה משתמשים במשימות רבות ומגוונות, כמו סיכום פגישות, עריכה לשונית, יצירת חומרי לימוד או חיפוש מידע באינטרנט.

בחירת אופן השימוש במודל חשובה, והיא שלב מקדים בתהליך התכנון לצורך הגדרת המטרות ומאוחר יותר את הגדרות ההוגנות (ראו פירוט בהמשך חלק זה). ייתכן שנרצה להשתמש במודל יחיד במגוון דרכי שימוש עבור מטרות שונות, ובמקרה זה יהיה צורך לבחון את הסיכונים הטמונים בכל אופן שימוש.

מתי מומלץ ליישם את הפתרון? פתרון זה מורכב מכמה שלבים: הגדרת מטרה כללית, קביעת יעדים מדידים ואפיון פונקציית ההפסד של האלגוריתם. את השלב הראשון של הגדרת המטרה הכללית בהטמעת המודל מומלץ לבצע בכל פעם שבה מטמיעים מודל (חדש או קיים). במקרים

שבהם רוצים למדוד את התוצאות, יש להגדיר גם מטרות מדידות. החלק הנוגע לפונקציית ההפסד רלוונטי רק במקרים שבהם מאמנים מודל חדש.

איך יש ליישם את הפתרון? קובעי המדיניות בשיתוף הצוות הטכנולוגי, ובמקרים מסוימים גם גורמים נוספים (ראו סעיף 10.1.1 לעיל), קובעים את מטרות המודל ואת אופן השימוש בו. מומלץ כי שלב זה יהיה מתועד, מאחר שבשלבים הבאים של פיתוח ובדיקת המודל מתחשבים במטרות המודל (Raji et al., 2020).

בהגדרת מטרות המודל כמה שלבים. בשלב ראשון יש להגדיר את המטרה הכללית. יש מצבים שבהם יש יותר ממטרה אחת בהטמעת מודל, לדוגמה חיסכון במשאבים וגם שיפור תוצאות. הגדרת מטרה היא החלטת מדיניות, והפתרונות שבמעטפת השירות הציבורי, כמו ועדות מייעצות ופרסום, נועדו להביא לכך שבחירת המטרה עצמה תהיה הוגנת ולא תפגע באוכלוסיית היעד.

במקרים רבים נרצה למדוד את הצלחת הטמעת המודל. במקרים אלו, כשלב שני, יש להגדיר מדדים כמותיים להערכת המודל. לדוגמה, מטרה כמו "שיפור ציוני הבגרות" היא מטרה כללית אבל אינה מדד מפורש. האם הכוונה היא לשיפור ממוצע ציוני הבגרות של כל התלמידים בשנתון? להגדלת מספר יחידות הלימוד של התלמידים בשנתון? או להגדלת מספר התלמידים הזכאים לתעודת בגרות? כאשר מעריכים את ביצועי המודל יש להשתמש במטרה מוגדרת היטב (Raji et al., 2020). המטרה המפורשת צריכה לשקף את המטרה הכללית עד כמה שאפשר, אבל גם להביא בחשבון שברוב המקרים אין התאמה מלאה. במקרים שבהם אין התאמה, ייתכן שנבחר להשתמש בכמה מדדים להערכת המודל, כדי לשקף טוב יותר את המטרה הכללית. בדוגמה של "שיפור ציוני הבגרות", ייתכן שנרצה להשתמש גם בממוצע ציוני הבגרויות של כל התלמידים בשנתון וגם באחוז התלמידים שסיימו את לימודיהם בתיכון עם תעודת בגרות.

חשוב להגדיר ולתעד הן את המטרה הכללית בהכנסת המודל הן את המטרה המפורשת שאותה מודדים, וזאת משום שלעיתים יש חוסר התאמה בין המטרה הכללית ובין הערך המדיד, שעלול ליצור אפליה (ראו סעיף 8.3) או לפגוע בביצועי המודל.

יש אף הבחנה בין שימוש במודל קיים, כמו מודל שפה, ובין אימון של מודל מנתונים. לדוגמה, אם מתבצע אימון של מודל להערכת סיכון, פונקציית ההפסד מגדירה את עלות השגיאה של האלגוריתם – הקנס של פונקציית ההפסד. בחירה בפונקציית ההפסד היא בעצם הגדרת מטרה מפורשת לאלגוריתם, ולכן חשוב לבחון את ההתאמה בין פונקציית ההפסד ובין אופן השימוש והמטרה הכללית.

קביעת מטרות הוגנות

כאמור, כאשר מטמיעים מודלים של בינה מלאכותית יש סכנה לפגיעה בהוגנות כלפי קבוצות באוכלוסייה או פרטים. לכן מומלץ לקבוע מראש מהן מטרות המודל מבחינת ההוגנות ולתעד את ההחלטה לטובת השלבים הבאים (Alvarez et al., 2024).

מתי מומלץ ליישם את הפתרון? בכל הטמעה של מודל חדש או קיים. גם החלטה שלא להגדיר את מטרות ההוגנות היא החלטה מהותית, וגם במקרה כזה יש לבחון את השיקולים לעומק ולתעדם תיעוד מסודר.

איך יש ליישם את הפתרון? גם בהפעלת פתרון זה יש כמה שלבים, החל בבחינת הקבוצות באוכלוסייה שעליהן אנחנו רוצים להגן, דרך בחירת מטרות כלליות וכלהגם אכיפת מטרות של הוגנות על המודל עצמו.

השלב הראשון בקביעת מטרות הוגנות הוא להבין מי הן הקבוצות באוכלוסייה שבסיכון. בישראל הקבוצות המוגנות באוכלוסייה הוגדרו בחוק איסור הפליה במוצרים, בשירותים ובכניסה למקומות בידור ולמקומות ציבוריים, תשס"א-2000. נוסף על כך בתחומים מסוימים יש קבוצות באוכלוסייה שנמצאות בסיכון. לדוגמה, אנשים עם עודף משקל הם אוכלוסייה בסיכון בהחלטות רפואיות אבל שאין הם אוכלוסייה בסיכון בתחום החינוך. במקרים רבים יש כמה קבוצות מוגנות באוכלוסייה שצריך להתייחס אליהן, ולעיתים יש מורכבות הנובעת מהצטלבות זהויות: מצבים שבהם פרטים משתייכים ליותר מקבוצה אחת, כגון נשים מקבוצות מיעוט, ועלולים לחוות דפוסי פגיעה ייחודיים שאינם משתקפים בבחינה של כל קבוצה בנפרד.

בשלב השני יש להחליט מהן מטרות ההוגנות שאנחנו מעוניינים שהמודל יקיים. אם המודל הוא שלב בתהליך של קבלת החלטות, יש להתייחס גם למודל בפני עצמו וגם לתוצאות התהליך כולו. לדוגמה, אם משתמשים במודל כדי להמליץ לבתי ספר על תלמידים שזקוקים לתגבור, יש לבחון הן את הוגנות ההמלצות הן את זהותם של התלמידים שמקבלים תגבור בפועל ואם התגבור ניתן באופן הוגן. כפי שצוין בפרקים 9.2 ו-9.4, ייתכן מצב שבו ההמלצות עצמן הוגנות, אבל מסיבות שאינן קשורות למודל, יישומן אינו הוגן. סיבות אלו יכולות להיות, לדוגמה, בתי ספר שאינם מקבלים את המלצות המודל או שהתגבור ניתן במקום מרוחק ותלמידים לא יכולים להגיע אליו (Alvarez et al., 2024).

בפרק 7 דנו בגישות מגוונות בהוגנות אלגוריתמית ובנקודות המבט שמייצגות אותן. בחירת מטרה להוגנות אלגוריתמית היא תהליך מורכב, מאחר שזו פעולה שהופכת נקודות מבט חברתיות לדרישות מתמטיות שאפשר לאכוף על אלגוריתמים. נוסף על כך הגישות השונות של הוגנות נמצאות במתח מתמיד, ואין הגדרה אוניברסלית שמתאימה לכל המצבים. בכל מצב צריך לבחור בהגדרה שמשקפת היטב את מטרות ההוגנות שעליהן החליטו קובעי המדיניות. הבחירה בהגדרה היא בחירה של מדיניות, וההמלצות במעטפת השירות הציבורי נועדו לדאוג שהבחירה והמעקב אחריה יהלמו את המטרות.

כדי להמחיש את מורכבות הבחירה ואת הסתירה בין ההגדרות, יוצגו להלן חלק ממטרות ההוגנות האפשריות:

- **שוויון דמוגרפי (demographic parity)**. זו הגדרה של הוגנות קבוצתית (ראו פרק 7). סוג הוגנות זה בוחן את "מבחן התוצאה" – האם שיעור התוצאות החיוביות שהמערכת מנפקת (למשל אחוז זכאות למענק) זהה בין קבוצות באוכלוסייה (לדוגמה, נשים לעומת גברים, מרכז לעומת פריפריה), ללא תלות במאפיינים השונים. משתמשים בהגדרה זו למגוון מטרות, למשל כאשר רוצים להבטיח גיוון וייצוג, הקצאת משאבים שוויונית או במקרים נוספים שבהם נקודת המוצא היא התייחסות שווה לכולם
- **שוויון בשגיאת שלילי שגוי (equal opportunity).** ההנחה היא שתוצאה חיובית של המודל היא הזדמנות, והאנשים שנפגעים הם מי שקיבלו תוצאה שלילית בטעות (false negative error). לדוגמה, מי שהיו יכולים להצטיין באוניברסיטה אבל לא התקבלו. הגדרה זו דורשת להשוות את אחוז השגיאה false negative error בין קבוצות שונות באוכלוסייה, ודרישה זו כן מתייחסת אל ההבדלים בין הקבוצות (Hardt et al., 2016)
- **הוגנות אינדיווידואלית**. בגישה זו של הוגנות אנחנו מניחים שיש לנו הגדרה ברורה של מי הם אנשים דומים, והדרישה היא להתייחס התייחסות דומה לאנשים דומים. ההגדרה של מי הם אנשים דומים צריכה להיות רלוונטית למטרת המודל, והיא ניתנת על ידי פונקציה שמגדירה מרחק בין שני אנשים. לדוגמה, לצורך הערכת הפוטנציאל של סטודנטים להצליח באוניברסיטה אפשר להחליט שאנשים דומים יהיו אנשים עם ממוצע ציוני בגרות דומים (Dwork et al., 2012)

מטרות אלו סותרות זו את זו. אי אפשר לדרוש להתייחס באותה דרך לאנשים שלהם ציוני בגרות דומים וגם לקבל אחוז פרופורציונלי מקבוצות באוכלוסייה, במצב שבו יש קבוצה שלה ציוני בגרות נמוכים. יש עוד מספר רב של הגדרות של הוגנות המתאימות למצבים שונים, לדוגמה דרישות קליברציה שעליהן נפרט בפרק הבא. הבחירה בהגדרת ההוגנות משלבת שיקולים טכניים ושיקולי מדיניות, ועל כן חשוב שתיעשה בשיתוף פעולה בין קובעי מדיניות לבין אנשי טכנולוגיה וגם תתועד לבחינה בשלבים הבאים.

יש לציין שבמקרים רבים הבחירה לאכוף מודל של הגדרת הוגנות מסוימת עלולה להשפיע לא רק על היחס לקבוצות מיעוט אלא גם על ביצועי המודל בכללותו. לדוגמה, הבחירה לקבל תלמידים מרקע חברתי חלש לאוניברסיטה עלולה לפגוע באחוז הסטודנטים המסיימים את התואר בהצלחה. תופעה זו נקראת "מחיר ההוגנות" (the price of fairness) (Haas, 2019). במצב זה יש לאזן בין הדרישה להוגנות ובין הביצועים הנדרשים מן המודל. זוהי החלטה ערכית שבה מאזנים שיקולים שונים, והיא נוגעת למדיניות. כדי לסייע בקבלתה אפשר לנתח את האפשרויות השונות למטרות ההוגנות ואת השפעתן על הביצועים, לדוגמה את אחוז הסטודנטים מן הרקע החלש שנקבל ואת השפעתם על אחוז הסטודנטים המסיימים את

התואר בהצלחה. ייצוג זה נקרא "חזית פארטו" (pareto front), והוא יכול לסייע למקבלי ההחלטות בבחירה (Han et al., 2025).

10.2 פיתוח וטיוב

בשלב הפיתוח עוברים מהגדרת הכוונות לבניית המערכת עצמה: איסוף ועיבוד נתונים, אימון מודלים והרצת בדיקות הוגנות וביצועים. זהו שלב שבו המערכת מקבלת צורה ממשית והחלטות התכנון מוסבות ליישום בפועל.

10.2.1 פתרונות במעטפת השירות הציבורי

חיזוק הקישוריות והאינטגרציה בין מאגרי מידע ציבוריים

חיזוק הקישוריות והאינטגרציה בין מאגרי מידע ציבוריים, לרבות באמצעות מודלים מבוזרים של שיתוף נתונים, הוא פתרון שנועד להתמודד עם סיכוני הוגנות הנובעים ממידע חלקי, לא עדכני או מפוצל (OECD, 2019, 2025). כאשר מידע על אותו אדם או על אותה קבוצה באוכלוסייה מפוזר בין מערכות שאינן מסונכרנות, המודל מאומן על תמונה מקוטעת, ועלול להיווצר חיזוי מדויק פחות דווקא כלפי קבוצות מוחלשות באוכלוסייה (Mehrabi et al., 2021). אינטגרציה רחבה מאפשרת עדכונים שוטפים ומידע רציף יותר, זיהוי פערים וחוסרים במידע ושיפור היכולת להצליב נתונים ממקורות שונים ובמדויק. בכך היא מאפשרת פיתוח מודלים שמייצגים טוב יותר את המציאות המלאה של קבוצות שונות באוכלוסייה. נוסף על כך אינטגרציה מאפשרת גישה למקורות מידע אשר היעדרם יוצר תת-ייצוג שיטתי של קבוצות מסוימות. בכך היא מצמצמת פערי הוגנות הנובעים מהסתמכות על מאגרים זמינים אך מוטים: ישנם מאגרים איכותיים שאינם נגישים למפתחים בשל חסמים משפטיים, טכנולוגיים או מוסדיים, והדבר עלול להוביל להסתמכות על מאגרים "נוחים" בלבד (Mehrabi et al., 2021).

שיפור האינטגרציה בין מאגרי המידע אינו רק מהלך של בניית ממשקים אלא מנגנון לצמצום אי-שוויון בנתונים עצמם. עם זאת שיפור רציפות המידע אינו חף מסיכונים: אם באמצעות ריכוז ואם באמצעות אינטגרציה מבוזרת, שיתוף נתונים עלול להגביר פגיעה בפרטיות, לאפשר שימוש משני לא מבוקר בנתונים ולהנציח הטיות קיימות. פגיעות אלו נוטות להשפיע במיוחד על קבוצות מוחלשות באוכלוסייה, ובכך ליצור עיוותי הוגנות חדשים. יש לציין כי גם מערכי נתונים משולבים עלולים לשקף חוסר ייצוג או הטיות היסטוריות, ולכן שיפור הקישוריות איננו תחליף לבחינה ביקורתית של מקורות הנתונים עצמם (Mehrabi et al., 2021).

בהתאם לכך יש צורך בקידום יצירת ממשקים מאובטחים שמאפשרים גישה מבוקרת למאגרים שונים, קביעת סטנדרטים אחידים לשיתוף מידע בין משרדי הממשלה, שמירה על פרטיות על פי מדיניות מוגדרת והימנעות מהעדפת מאגרי מידע "נוחים" רק מפני שהם זמינים. פעולות אלו יצמצמו מצבים שבהם קבוצות שונות זוכות לייצוג שונה באיכות המידע רק בשל שונות

מסוימת. לשם לכך אפשר לשלב כלי הערכה ייעודיים הבוחנים עבור כל מערכת את סיכוני ההוגנות הנובעים מאיכות, שלמות וייצוג הנתונים, כתנאי מקדים לבחירה באסטרטגיית שיתוף או אינטגרציה.

דוגמה לכך היא יוזמת תשתית נתונים לאומית (National Data Infrastructure) בברזיל, שמטרתה לצמצם פערים באיכות המידע, אשר הובילו לייצוג חסר או לא עקבי של קבוצות שונות באוכלוסייה במערכות מידע ממשלתיות, באמצעות חיזוק הקישוריות והאינטגרציה בין מאגרי נתונים ציבוריים. היוזמה מגדירה מדיניות, תקנים ומנגנוני ממשל שמאפשרים שיתוף מידע מאובטח בין גופים ממשלתיים, כך שנתונים יהיו ניתנים לאיתור, להנגשה, להצלבה ולשימוש חוזר לפי עקרונות FAIR (Findable, Accessible, Interoperable, Reusable). תשתית זו מצמצמת את התלות במאגרים חלקיים או לא עדכניים, אשר עלולים להטות מערכות לקבלת החלטות לרעת קבוצות מסוימות, ומשפרת את ייצוגן של קבוצות באוכלוסייה בתוך מערכות המידע. פלטפורמת gov.br המנגישה כ-5,000 שירותים דיגיטליים, והפלטפורמה המשויכת Conecta gov.br, המשמשת תשתית לאינטגרציה בין-משרדית, מאפשרות זרימת מידע מבוקרת ואחידה (GO Fair, n.d.; OECD, 2025; National Data Infrastructure, 2025).

התאמת תוכן ושירותים לשפה ולתרבות של קבוצות מגוונות באוכלוסייה עבור משתמשי קצה

הטמעת הוגנות בשלבים המוקדמים של אפיון ועיצוב המערכת מחייבת מעבר מתפיסה של "תרגום טכני" בלבד להתאמה עמוקה של השירות להקשר התרבותי והלשוני של המשתמשים. הוגנות אינה רובד נוסף ה"מודבק" בסוף התהליך אלא עיקרון המנחה את איסוף הנתונים ובניית הממשק מלכתחילה. ברמה הטכנולוגית, הדבר דורש לאמן את המודלים על מערכי נתונים מגוונים ועשירים הכוללים לא רק את השפה התקנית אלא גם ניבים, סלנג והקשרים תרבותיים ייחודיים של אוכלוסיות מיעוט (Parker et al., 2025). כך הופכת הטכנולוגיה מכלי לאוטומציה גרידא לכלי פעיל המצמצם פערים ומנגיש שירותים.

אולם ללא רגישות מספקת, המאמץ להתאמה אישית עלול להוביל לתוצאה הפוכה וליצירת "הסללה דיגיטלית", והסיכון המרכזי הוא מעבר מ"התאמה" ל"קטלוג" באמצעות שימוש בתיוג ובסטראוטיפים, באופן שכובל אוכלוסיות מסוימות למסלולי שירות נפרדים ולעיתים אף נחותים. מחקרם של בלוג'ט ואו'קונור (Blodgett & O'Connor, 2017) ממחישים סכנה זו: במחקרם נמצא כי מודלים של בינה מלאכותית נוטים לייחס תכונות שליליות כמו חוסר אמינות לדוברים המשתמשים בניבים שאינם השפה התקנית . בשירות הציבורי, הטיה כזו עלולה לגרום לכך שפניות של קבוצות מוחלשות באוכלוסייה יסווגו בטעות כדחופות פחות או מוצדקות פחות.

כדי להימנע ממלכודות אלו, הפתרון מחייב שינוי מתודולוגי: פיתוח המערכת לא עבור קבוצות מוחלשות באוכלוסייה אלא יחד איתן. יש לאמץ מודל של עיצוב משתף המערב נציגי ציבור כבר בשלבי תיוג הנתונים ובדיקת המודל (ראו סעיף 10.1.1). השראה לכך אפשר לשאוב

ממודל "ריבונות המידע" שפותח בניו זילנד עבור האוכלוסייה המאורית, המבטיח כי המידע של הקהילה מנוהל תוך כיבוד ערכיה ומניעת ניצולו לרעה (Carroll et al., 2020). לבסוף, המבחן הקובע חייב להיות שוויון בתוצאה – הבטחה כי גם אם הדרך הדיגיטלית מותאמת אישית, איכות השירות והמענה הסופי נותרים זהים ושוויוניים לכולם.

10.2.2 פתרונות הקשורים לתהליך יצירת המודל ולהערכתו

בחינת איכות המידע

לאיכות המידע שעליו אומן המודל יש השפעה רבה על ביצועיו ועל ההטיות שבו, וכאמור מידע יכול להיות מוטה, לא איכותי על חלק מן האוכלוסייה, לשקף סטראוטיפים קיימים, ועוד.

מתי מומלץ ליישם את הפתרון? במצב שבו מאמנים מודל חדש המבוסס על מאגר מידע, יש לבחון אם יש הטיות במידע זה (ראו פרק 8.1). כאשר משתמשים במודל שכבר אומן, כמו לדוגמה מודל שפה, אין גישה למידע שעליו הוא אומן. גם במקרה זה חשוב להיות מודעים להטיות במודל שנוצרו עקב הטיות במידע, לדוגמה עקב סטריאוטיפים המצויים במידע ונלמדים על ידי המודל. אם אין אפשרות להפחית את ההטיות במודל על ידי תיקון המידע, אפשר להשתמש בפתרונות אחרים, כפי שיפורט בהמשך.

איך יש ליישם את הפתרון? בשלב ראשון יש לבחון אם יש פערים במידע. כפי שצוין בפרק 8.1, יש מצבים רבים שבהם חסר מידע על קבוצות באוכלוסייה או שהמידע עליהן לא איכותי. אפשר לאתר זאת על ידי בחינת מאגר המידע ובדיקה אם הקבוצה מיוצגת פרופורציונלית לחלקה באוכלוסייה הכללית. נוסף על כך אפשר לבדוק את התפלגות השדות החסרים ולבחון אם יש יותר שדות חסרים במידע בנוגע לקבוצה מסוימת. יש גם פערים במידע שקשה יותר לגלות. פער בביצועי המודל על קבוצה מסוימת יכול להיות אינדיקציה למידע לא איכותי, אבל הוא יכול להעיד גם על בעיות אחרות. שלב זה נוגע למצב של חוסר ברור במידע.

בשלב שני, אם חסר מידע על קבוצה באוכלוסייה, או שהמידע לא איכותי, אפשר לפצות על כך באופנים אלו:

- **איסוף מידע נוסף**. אם אפשר, מבחינת זמינות המידע והתקציב, עדיף לאסוף מידע נוסף על הקבוצה באוכלוסייה שעליה חסר מידע. כאשר אוספים מידע נוסף לשם השלמת מידע, במקרים רבים הוא נאסף באופן שונה מן המידע המקורי. כאשר מחברים כמה מקורות מידע שנאספו בזמנים שונים ובדרכים שונות, יש סיכון לשדות כפולים, לדגימה לא פרופורציונלית של קבוצות באוכלוסייה, ועוד. במצב זה יש שיטות למזעור הסיכון, ובהן משקול מתאים, ניסיון למחוק כפילויות ושימוש במידע איכותי בשביל נרמול (Lohr & Raghunathan, 2017; Mehrabi et al., 2021)
- **משקול**. כאשר אי אפשר לאסוף עוד מידע, למשל כשאין הוא בנמצא, יש כמה דרכים לשפר את ההטיות. אם יש מידע על מספר קטן של אנשים מאותה קבוצה באוכלוסייה,

אפשר להשתמש במשקול כדי להעלות את משקלם. כאמור בסעיף 8.2, ככלל האלגוריתם מחפש מודל שמצמצם את השגיאה הממוצעת, בהינתן שהמידע שעליו הוא מאומן מייצג את האוכלוסייה בכללותה. אם המידע מכיל מעט נתונים על קבוצה באוכלוסייה, ההנחה היא שהיא קטנה. משקול יתר פותר את הבעיה הזו וגורם לאלגוריתם "להבין" את גודלה האמיתי. יש לציין ששיטה זו מומלצת פחות מהשלמת המידע, מאחר שבמשקול אנחנו לא מוסיפים מידע על אותה קבוצה באוכלוסייה (ראו Moreno-Torres et al., 2012) לשיטות שונות למשקול

- **שדות חסרים (בנתונים בטבלה)**. יש מצבים שבהם יש מידע על כל האוכלוסייה, אבל בנוגע לקבוצות מסוימות המידע חלקי וחלק מן השדות חסרים. כפי שתואר בסעיף 8.1, במקרים רבים המידע החסר לא מתפלג התפלגות אחידה אלא מתרכז באוכלוסיות מסוימות. לפי פרננדו ואח' (Fernando et al., 2021) אפשר להשתמש בכמה שיטות נפוצות כדי להתמודד עם שדות חסרים:
  - **מחיקת שורות/עמודות**. אפשר למחוק שורות (אנשים) או עמודות (סוג המידע) שבהן חסר מידע. כאשר השדות החסרים מתרכזים בקבוצות מסוימות באוכלוסייה, יש לפצות על מחיקת השורות על ידי משקול, כפי שתואר לעיל. עם זאת מחיקת שורות עלולה להשפיע לרעה על הדיוק בנוגע לקבוצות באוכלוסייה שלהן הרבה שדות חסרים, והיא יכולה להניב תוצאות טובות פחות מאשר ניסיון להשלמה (Fernando et al., 2021)
  - **יצירת ערך "חסר"**. אפשר להשתמש בשדות שבהם חסרים נתונים, כאשר מוסיפים "חסר" לערכים האפשריים בטבלה ומשתמשים בה כפי שהיא עם ערכי ה"חסר" שמופיעים בה. כך נמנעים מהורדה של שורות בטבלה רק מאחר שהן לא שלמות. אפשר להשתמש בשיטה זו יחד עם שיטת ההשלמה (תתואר להלן), וכך משלימים שדות חסרים. כך לא מאבדים מידע עקב ההתמודדות עם השדות החסרים. עם זאת לפי ון נס ואח' (Van Ness et al., 2023), במצבים מסוימים השיטה יכולה להניב רמת דיוק טובה פחות
  - **השלמת שדות חסרים**. בשיטה זו, קודם כל מייצרים מודל שמנסה ללמוד את השדה החסר, ואז משתמשים בו כדי למלא את השדות החסרים בטבלה. לאחר מכן משתמשים בטבלה המלאה (עם השדות שהושלמו) כדי לבצע את המשימה המקורית של הלמידה. עמנואל ואח' (Emmanuel et al., 2021) מציגים שיטות רבות להשלמת שדות חסרים, כמו רגרסיה ליניארית, שימוש בשורה מלאה דומה או שימוש בשיטות של ניתוח אשכולות (clustering). כאמור בסעיף 8.1, להשלמת שדות חסרים יש במקרים רבים יתרונות לעומת השיטות האחרות מבחינת צמצום ההטיות כלפי האוכלוסיות שבנתוניהן שדות חסרים (Fernando et al., 2021)
- **יצירת מידע מלאכותי (סינטטי)**. מודלים גנרטיביים יכולים ליצור בקלות היקף גדול של מידע. במקרה של מידע חסר, אחת האפשרויות היא ליצור מידע חדש על ידי אלגוריתם, במקום להשתמש במידע אמיתי. עד כה מידע מלאכותי היה בשימוש בעיקר מסיבות של

שמירה על הפרטיות ולא כדרך להשלמת מידע. כאמור בסעיף 8.1, יש סכנות בשימוש במידע מלאכותי שנוצר על ידי אלגוריתמים גנרטיביים, מאחר שהפלט של אלגוריתמים אלו יכול להיות מוטה (Chen et al., 2021)

- **הורדת הטיה מן המידע**. אם יש סבירות שהמידע מוטה בשל הטיה בחברה, אפשר לעבד את המידע כדי להוריד הטיה זו. אם הנחת המוצא היא ששתי קבוצות מסוימות הן שוות וההבדל בין המידע על אודותיהן הוא עקב הטיה בחברה, אפשר לשנות את המידע כדי להוריד את ההטיה, והתהליך נקרא pre-processing. יש כמה שיטות לעיבוד מידע כדי להוריד הטיה הזו: שינוי המידע עצמו, משקול מחדש של המידע ומתן משקל יתר למידע.. יצוין שאפשר להשתמש גם בשיטות אלגוריתמיות אחרות שאינן עיבוד הנתונים, גם אם רוצים לתקן הטיה שמקורה בנתונים מוטים (Kamiran & Calders, 2012). ראו פירוט להלן.

קליברציה (כיול) ומולטי-קליברציה

קליברציה היא התאמה בין הערכת הסיכון שמפיק המודל ובין השכיחות לסיכון בפועל. כלומר כאשר מודל אלגוריתמי מנבא הסתברות (למשל הסתברות לאי-עמידה בתשלומים או הסתברות לסיכון בריאותי), הקליברציה בודקת אם מספרים אלו באמת משקפים את המציאות (בממוצע). לדוגמה, במודל של הערכת סיכון למחלה, אם אנחנו בוחנים את קבוצת כל האנשים שעבורם המודל העריך את הסיכוי לחלות במחלה כ-10%, נצפה לראות שכ-10% מהם באמת יתבררו כחולים בסופו של דבר (Silva Filho et al., 2023).

מולטי-קליברציה היא הרחבה של דרישה זו, שנכונה גם עבור קבוצות מגוונות באוכלוסייה. במקרה זה נרצה שהמודל יהיה מכויל לא רק כאשר אנחנו מסתכלים על כל האנשים אלא גם כאשר אנחנו בוחנים כל קבוצה בפני עצמה. בדוגמה הקודמת, נרצה שגם אם נבחן את כל הנשים שעבורן העריך המודל סיכון של 10%, נצפה לראות שכ-10% מהן יהיו חולות בסופו של דבר. אפשר לדרוש מולטי-קליברציה עבור מספר רב של קבוצות באוכלוסייה, לפי גיל, מגדר, מצב כלכלי, משקל, ועוד. אפשר להרחיב את הדרישה של מולטי-קליברציה גם לקבוצות מורכבות יותר באוכלוסייה (Hébert-Johnson et al., 2018).

מודל מכויל היטב, לא רק בממוצע אלא גם בתוך קבוצות באוכלוסייה, משמש בסיס איכותי למדיניות תיקון הוגנות בשלב שאחרי האימון ומאפשר ליישם התאמות הוגנות מבלי לפגוע באמינות הציונים או להרחיק אותם באופן חריג מתבניות הסיכון שמייצגות את המציאות.

מתי מומלץ ליישם את הפתרון? יש שלושה מצבים שבהם על המודל לקיים דרישות של קליברציה ומולטי-קליברציה:

- יש סוג של מודלים שמטרתם להציג רק הערכת סיכון ולא לקבל את ההחלטה בפועל. במודלים כאלה, כל עוד הנתונים אינם מוטים בשיטתיות, הציון אמור לשקף את המציאות נאמנה. את ההחלטה עצמה מקבל מומחה אנושי, ולכן נדרש שהציון יהיה מכויל ואמין (Silva Filho et al., 2023).

- כאשר מטרת המודל היא בחירה בפעולה, אבל השיקולים לבחירה לא ידועים בשלב אימון המודל או שהם יכולים להשתנות. לדוגמה, אם רוצים לאמן מודל שימליץ לבתי ספר את מי מן התלמידים כדאי לשלוח לתגבור, אבל לבתי ספר שונים יש העדפות שונות בנוגע לתלמידים שצריך לשלוח. לבית ספר אחד יש מורים שיכולים לתגבר, ולכן יש העדפה לשלוח לתגבור חיצוני רק את התלמידים בסיכון הגבוה ביותר לנשור, ואילו לבית ספר אחר אין יכולת כזו, ולכן יש העדפה לשלוח את כל מי שבסיכון. במצב זה אפשר לאמן מודל שמקיים דרישות של מולטי-קליברציה, ואז כל בית ספר יוכל לעשות את השיקול שלו בסיוע המודל (Gopalan et al., 2022).
- במצבים שבהם מטרת המודל היא להפיק החלטה או ציון סיכון, אך ידוע שהנתונים מכילים הטיה מובנית, נדרש לתקן את תוצאות המודל באמצעות אכיפת הוגנות קבוצתית לאחר האימון (post-processing). במקרים אלו קיומה של מולטי-קליברציה במודל המקורי היא יתרון של ממש: כאשר כל קבוצה באוכלוסייה מקבלת ציון סיכון מהימן בהתבסס על נתוני העבר שלה, קל יותר לאכוף את מטרות ההוגנות הקבוצתית אכיפה יציבה ועקבית מבלי ליצור עיוותים גדולים או תוצאות בלתי צפויות (Hu et al., 2023).

איך יש ליישם את הפתרון? כדי ליישם את הפתרון הספרות מציעה שימוש באלגוריתם postprocessing הפועל לאחר בניית המודל. האלגוריתם מקבל כקלט את מודל הערכת הסיכון ואת רשימת הקבוצות שעל כולן נדרש להתקיים כיול. האלגוריתם מאתר קבוצה שבנוגע לה המודל עדיין אינו מכויל ומבצע התאמה המבטיחה כיול עבורה. יתרונו המרכזי של האלגוריתם הוא ביכולתו להתמודד גם עם מספר רב של קבוצות באוכלוסייה (Gopalan et al., 2023; Hébert-Johnson et al., 2018).

אכיפת הגדרות של הוגנות, אם נבחרו

בשלבים הקודמים של תכנון המערכת דנו בקביעת המטרות של הוגנות המודל וקביעת מטרות ההוגנות.

מתי מומלץ ליישם את הפתרון? כשמאמנים או מטמיעים מודל להערכת סיכון או המלצה, ובשלבים הקודמים של אפיון המערכת הוחלט כי יש לאכוף הגדרות של הוגנות על המודל. אם מאמנים את המודל, אפשר לאכוף הגדרות הוגנות בכל אחד משלבי האימון (ראו להלן). כאשר משתמשים במודל קיים, אפשר להפעיל שיטות של post-processing (ראו להלן) ולהתאים את תוצאות המודל כדי לקיים את מטרות ההוגנות. במודלים גנרטיביים אי אפשר להתאים את התוצאות של המודלים ויש להשתמש בפתרונות אחרים במעטפת כדי להשיג את מטרות ההוגנות (לדוגמה, הדרכות לשימוש במודל) (Pessach & Shmueli, 2023).

איך יש ליישם את הפתרון? יש כמה שיטות לאכיפת מטרות שונות של הוגנות, על פי שלבי אימון המודל (לסקירה מקיפה ראו Pessach & Shmueli, 2023).

- שיטות של pre-processing: שינוי הנתונים שעליהם אומן המודל כדי שהוא יקיים הגדרות הוגנות מסוימות. שיטות אלו רלוונטיות אך ורק אם מאמנים מודל מנתונים שלנו ואינן ישימות כאשר משתמשים במודל מוכן.
- שיטות in-processing: הכנסת מטרת ההוגנות לתוך תהליך אימון המודל. אחת האפשרויות היא להוסיף "רכיב הוגנות" לפונקציית ההפסד של המודל, ובמקרה זה המודל ינסה לצמצם את ההפסד המקורי שמקדם דיוק, וגם את ההפסד שנובע מרכיב ההוגנות. שיטה נוספת היא לאכוף את מטרת ההוגנות על המודל בתהליך האימון.
- שיטות של post-processing: אימון המודל בלי להתחשב בהוגנות. לאחר מכן לוקחים את הפלט של המודל ומשנים אותו כדי לקיים את מטרות ההוגנות שנבחרו. היתרון בשימוש בשיטות אלו הוא הגמישות, מאחר שאפשר לשנות את מטרות ההוגנות בלי לאמן מחדש את המודל. יתרון נוסף הוא הצגה ברורה יותר של מחיר ההוגנות וההשפעה שלו על ביצועי המודל.

הטמעת תהליך הערכה הבוחן את הסיכונים לחוסר הוגנות ואת תרומת המערכת לשיפור הוגנות

בעת שילוב מערכת מבוססת בינה מלאכותית בשירות הציבורי נדרש ללוות את ההטמעה בתהליך הערכה שיטתי ומתמשך. תהליך זה בוחן בה בעת את הסיכונים לחוסר הוגנות הנובעים מהפעלת המודל ואת תרומתו הפוטנציאלית לשיפור ההוגנות לעומת המצב הקיים. במערכות ציבוריות רבות יש אילוצי משאבים, כמו זמני המתנה ארוכים או הקצאת מענים מוגבלים, ולכן בחירה במדיניות הוגנות מסוימת עבור קבוצה אחת באוכלוסייה עלולה לבוא על חשבון קבוצה אחרת. גם במקרים שבהם אין מגבלת משאבים מפורשת, כגון הצגת תכנים חינוכיים מותאמים, נוצר לעיתים מתח בין התאמה אישית ובין שוויון בין קבוצות באוכלוסייה (Raji et al., 2020).

מתי מומלץ ליישם את הפתרון? בכל פעם שבה מטמיעים מודל חדש או קיים. תהליך של הערכה חשוב גם במצב שבו לא צריך לאכוף מטרות הוגנות.

איך יש ליישם את הפתרון? מומלץ להגדיר מראש את מערך המדדים לבחינת התנהגות המערכת: מדדי הוגנות, דיוק, רגישות לקבוצות באוכלוסייה, וכן את "מחיר" הבחירות שנעשות, כמו ויתור מסוים על דיוק לטובת ייצוג הוגן יותר. תהליך זה אינו סטטי אלא אמור להתעדכן לאורך זמן על פי תובנות מן השטח, משוב מן המשתמשים ודפוסי שימוש שמתגלים רק לאחר שהמערכת פועלת. כך ההטמעה הופכת מביצוע חד-פעמי למנגנון למידה מתמשך שמבטיח שהמערכת תישאר יעילה והוגנת ככל האפשר. יודגש כי כאשר בוחנים את ההוגנות של תהליך המכיל מודל, יש לבחון את גם את ההוגנות של התהליך כולו ולא רק של המודל כשהוא לעצמו, כפי שפורט בסעיף 10.1.2 (Raji et al., 2020).

ככלל יש כמה דרכים להעריך את הוגנות המודל על פי מטרות ההוגנות שנבחרו. אם נבחרו מטרות של הוגנות קבוצתית או של קליברציה, יש להעריך בזמן השימוש אם המודל מקיים אותן. במקרים של אלגוריתמים גנרטיביים, שלהם פלט מורכב יותר, קשה להעריך את ההוגנות הערכה פשוטה וחד-משמעית.

דוגמה לתהליך הערכה שיטתי המערב גם גורמים טכנולוגיים וגם פיקוח אנושי היא OPTICA (Organizational PerspecTIve Checklist for AI solutions adoption). כלי צ'ק-ליסט ארגוני שפותח ב"כללית שירותי בריאות" בישראל לצורך בחינת התאמת פתרונות מבוססי בינה מלאכותית לפני הטמעתם. הכלי בנוי כהליך רב-שלבי הכולל הגדרת הצורך הקליני והמדדים הרצויים, בחינת דמיון בין אוכלוסיית האימון לאוכלוסיית היעד ואיכות הנתונים המקומיים, בדיקת ביצועים בקרב קבוצות באוכלוסייה ועמידה במטרות הוגנות, וכן תכנון הפריסה, הניטור וההערכה שלאחר ההטמעה. התהליך מחייב שיתוף פעולה בין מפתחים, מומחי נתונים, גורמי ניהול ומומחים קליניים וכולל גם "נקודות עצירה" שבהן אפשר להחליט שלא להמשיך בהטמעה אם מתגלים כשלים מהותיים, למשל פערים חריפים באיכות הנתונים המקומיים או רמת דיוק נמוכה יותר של המודל עבור קבוצות מסוימות באוכלוסייה (Dagan et al., 2024).

שימוש בכלים קיימים להערכת ההוגנות וה"מחיר" שמשלמים עבורה

הפיכת מושג ההוגנות המופשט למדד כמותי בר-מדידה הוא אתגר מרכזי בשלב הפיתוח של המערכת. כדי להתמודד עם אתגר זה עומדים לרשות המפתחים מגוון כלי קוד פתוח שנועדו להשתלב בתהליך ולזהות הטיות טרם הטמעת המערכת. כלים מובילים כגון Microsoft Fairlearn, Fairness ו-Google What-if Tool מאפשרים לארגונים לבחון את ביצועי המודל באמצעות מדדים מתמטיים מוגדרים (Bellamy et al., 2019). כלים אלו מאפשרים תצוגה מפורטת וגרפית של הפערים בתוצאות של המודל על קבוצות שונות באוכלוסייה ומאפשרים לבחון כיצד מטרות שונות של הוגנות קבוצתית משפיעות על הדיוק של המודל ועל ביצועיו. כלים אלו יכולים לעזור למקבלי ההחלטות בבואם לשקול את מדיניות ההוגנות מאחר שהם ממחישים גרפית את תוצאות המודל.

מתי מומלץ ליישם את הפתרון? כלים אלו נועדו לשימוש בזמן הטמעת המודל, כמו הפתרון הקודם. כלי זה אינו מחליף את תהליך קבלת ההחלטות ובחירת המדדים אלא מסייע לקבל ההחלטות על ידי הצגה גרפית של החלופות השונות והצגת מדדים שונים.

איך יש ליישם את הפתרון? הכלי מכיל ברובו מוצרים מוגמרים שאפשר להפעיל. נוסף על אפשרויות אלו, אפשר לייצר עצמאית הצגה גרפית של מדדי הוגנות מתמטיים על ידי שימוש בספריות קוד מוכנות.

ההסתמכות הגוברת על כלים מתמטיים אלו טומנת בחובה סיכון ל"הלבנת הוגנות" (fairwashing) – מצב שבו ארגונים יוצרים אשליה של הוגנות על ידי סימון "V" על מדד טכני מסוים והתעלמות מהקשרים חברתיים רחבים או מהטיות עומק שאינן משתקפות בנתונים היבשים. יתרה

מכך, ניסיון לשפר מדד הוגנות יחיד באמצעות התאמות טכניות במודל בלבד, מבלי לבחון את המשמעות הרחבה של השינויים, עלול לפגוע במדדי הוגנות אחרים. כדי לצמצם סיכון זה, אין להסתפק בפלט האוטומטי אלא יש לשלב הערכה איכותנית שתבוצע על ידי צוותי בדיקה רב-תחומיים מתחומים כמו מדעי החברה, משפט ודאטה (Stankovich et al., 2023; VerifyWise, n.d.). צוותים אלו יבחנו את משמעות התוצאות בהקשר הרחב של השירות, יתעדו בשקיפות את השיקולים בבחירת המדדים ויבטיחו כי המענה הטכנולוגי תומך בערכי השירות הציבורי ולא מחליף אותם.

כדי להבטיח שמטרות אלו לא יהפכו לחסם בפני יישום הטכנולוגיה, יש להפעיל את מנגנון ההערכה הרב-תחומית באופן דיפרנציאלי המבוסס על ניהול סיכונים. במערכות המוגדרות בסיכון נמוך, העוסקות למשל בהיבטים תפעוליים או מנהליים, אפשר להסתפק במדדים הטכניים ובכלי המדידה האוטומטיים כבקרת איכות נאותה המאפשרת התקדמות מהירה. לעומת זאת הצורך בתהליך המשולב והאיכותני הופך הכרחי ככל שרמת הסיכון עולה, במיוחד במערכות המשפיעות על זכויות יסוד, על חלוקת משאבים חשובים או על קבוצות פגיעות באוכלוסייה, שבהן הסיכון להטיות סמויות ומערכתיות גבוה יותר. אימוץ גישה מעשית זו עשוי לאפשר לכוון את מרב המשאבים והמומחיות למוקדי הסיכון הבולטים, בלי לעכב את הפיתוח במערכות שבהן פוטנציאל הפגיעה בהוגנות הוא זניח.

חוויות המשתמשים כמדד להוגנות

פתרון נוסף לקידום הוגנות אלגוריתמית נשען על הרחבת מסגרת הבדיקה מעבר למדדים סטטיסטיים טהורים אל עבר חוויית המשתמש (להלן: UX) בפועל. מחקרים שונים (Lechevalier & Saville, 2025; Nakao et al., 2023; Virvou, 2023) מצביעים על כך שאופן העיצוב של המערכת והאינטראקציה עימה משפיעים ישירות על האופן שבו הוגנות מתממשת הלכה למעשה. לפי מחקרים אלו, הוגנות אלגוריתמית לא נקבעת רק בהקשר של המודל והנתונים אלא גם בהקשר של הממשק וחוויית המשתמש. עיצוב המערכת, הכולל בתוכו את האופן שבו היא מציגה מידע, מנסחת אפשרויות, מפעילה עומס קוגניטיבי או מאפשרת (או מגבילה) שליטה, משפיע ישירות על היכולת של משתמשים להבין את ההחלטות, לנווט ביניהן ולממש בחירה אמיתית. ממשקים מורכבים או מטעים עלולים להעמיק פערי מידע ולפגוע באוטונומיה של משתמשים, ובכך לייצר אי-הוגנות, גם כאשר המודל עצמו תקין מבחינה פורמלית.

בד בבד חשוב לציין כי חוויית משתמש מעוצבת היטב עלולה גם ליצור אשליית הוגנות – מצב שבו הממשק משדר בהירות ושליטה אך בפועל מסתיר כשלים או מפחית את הסיכון הנתפס. כלומר UX יכול הן לחשוף אי-הוגנות הן להסוות אותה, ולכן נדרשת בחינה ביקורתית דו-כיוונית. בצד זאת מחקרי UX בתחום הבינה המלאכותית מראים כי החוויה הסובייקטיבית של המשתמשים מזינה בחזרה את תפקוד המערכת: כאשר משתמשים מבולבלים, מודרים או חסרי אמון, איכות השימוש שלהם נפגעת והמערכת מפיקה תוצאות יציבות פחות ומדויקות

פחות. במובן זה חוויית המשתמש אינה "עטיפה" אלא רכיב מהותי בהוגנות, שכן היא קובעת בפועל מי נהנה מן המערכת, מי נפגע ממנה ומי מסוגל להשפיע על תוצאותיה Lechevalier & Saville, 2025; Nakao et al., 2023; Virvou, 2023).

דוגמה ליצירת מערכת שבוחנת את חוויית המשתמש כהיבט של הוגנות היא שימוש בכלי FID (Fairness in Design) המציע מתודולוגיה סדורה שמרחיבה את בחינת ההוגנות מעבר למודל עצמו אל המרחב שבו המשתמשים פוגשים את המערכת (Zhang et al., 2023). באמצעות כלי זה, צוותי פיתוח, עיצוב ומדיניות מנתחים תרחישי שימוש ממשיים דרך נקודת המבט של מגוון משתמשים ובוחנים כיצד עקרונות הוגנות שונים מתממשים או מתערערים לפי חוויית המשתמש. התהליך כולל מיפוי בעלי עניין, הפעלה של עקרונות הוגנות, זיהוי נקודות עומס קוגניטיבי, חוסר בהירות או עיצוב מטעה ובחינת האופן שבו ניסוח, סדר פעולות או אופן הצגת מידע עשויים לייצר פערי גישה או להשפיע השפעה לא שוויונית על קבוצות באוכלוסייה. בדרך זו הכלי מאפשר לאתר כשלים שמקורם בממשק ולא במודל, כמו מצבים שבהם משתמשים מוותרים על זכויות, מתקשים לערער, מקבלים החלטות שאינן משקפות את העדפותיהם או נקלעים למצבי חוסר שליטה. בכך הופך FID לכלי שמאפשר לזהות ולתקן הטיות שלא היו נחשפות בבדיקות אלגוריתמיות בלבד.

10.3 הטמעה ויישום

שלב ההטמעה מתחיל כאשר המערכת בשלה לשילוב בתוך תהליכי העבודה. כאן נקבעים נוהלי הפעלה, נבנית תשתית לשימוש מקצועי, ומותאמים מנגנוני פיקוח אנושי ושקיפות למשתמשי הקצה. זהו השלב שבו המערכת עוברת מן הפיתוח אל הסביבה הארגונית בפועל.

10.3.1 פתרונות במעטפת השירות הציבורי

הכשרת עובדים לפרשנות של תוצרי מערכות מבוססות בינה מלאכותית

כדי לאפשר שימוש אחראי במערכות מבוססות בינה מלאכותית במסגרת העבודה השוטפת, יש להבטיח שלעובדי השירות הציבורי תהיה הכשרה מתאימה שתסייע להם לפרש את תוצרי המערכת פירוש מושכל. ללא הכשרה מתאימה, עובדים עלולים להסתמך אוטומטית על המודל או לחלופין להטיל בו ספק יתר על המידה ובאופן לא אחיד על כלל האוכלוסייה. כדי למנוע זאת על ההדרכה להתאים לתחום המקצועי: רופאים, עובדים סוציאליים, אנשי רווחה ואנשי חינוך ניצבים בפני סוגים שונים של סיכון, נדרשים לרמות שונות של הסבר, ומשמעות הפלטים עבורם שונה בתכלית (OECD, 2025).

יתרה מכך מערכות מבוססות בינה מלאכותית דינמיות מטבען, ולכן ההכשרה חייבת להיות מתמשכת. הן מתעדכנות, המדדים משתנים, פונקציות סיכון מוגדרות מחדש, ולעיתים המערכת מופצת לקבוצות באוכלוסייה שלא נבחנו בשלב הפיתוח. לכן הכשרה ראשונית לבדה אינה

מספיקה אלא יש צורך בתהליך קבוע. כך אפשר להבטיח שהעובדים ממשיכים להבין מה המודל עושה בפועל ולא משתמשים בו כאילו הוא קופסה שחורה. הכשרה מקצועית במשך שלבי ההטמעה אינן משמשות רק הדרכה טכנית אלא תנאי לכך שהמערכת תשולב באחראיות ובהוגנות: תאפשר לעובדים להבין מה המודל יודע ומה חסר לו, תחבר בין מומחיות אנושית ובין יכולות אלגוריתמיות ותצמצם סיכונים של פגיעה דווקא בקבוצות באוכלוסייה שהמערכת נועדה לסייע להן (OECD, 2025). לדוגמה, סוכנות הפיתוח של האו"ם מפעילה תוכנית הכשרה לעובדי מדינה במטרה לחזק את הידע שלהם בנוגע לשימוש אתי, יעיל ואחראי במערכות מבוססות בינה מלאכותית. התוכנית מותאמת לכל מדינה ונותנת מענה לאתגרים רלוונטיים (UNDP, n.d.).

מנגנוני ערעור אנושיים

מנגנוני ערעור אנושיים הם רכיב חיוני בבנייה ובהטמעה של מערכות מבוססות בינה מלאכותית בשירות הציבורי, משום שהם מספקים נתיב תיקון והגנה מפני טעויות או הטיות של המודל במערכות רגישות. ערעור אנושי מאפשר למקבלי השירות לבחון מחדש החלטות שהתקבלו על בסיס מודל ולהתחשב בשיקולים שאינם נגישים למערכת אוטומטית. פרט לתפקידם התפעולי, קיומם של מסלולי ערעור זמינים ונגישים מגביר במידה ניכרת את אמון הציבור, משום שהם מבהירים שהמערכת אינה זו שמקבלת את החלטה הסופית, ושיש אפשרות אנושית לעיין, לתקן ולבקר את החלטותיה (Byers, 2022).

עם זאת, כמו פתרונות אחרים במעטפת השירות הציבורי, גם מנגנוני ערעור אינם חפים מסיכונים: גורמים אנושיים עצמם פועלים תחת הטיות, עומסים ולחצים מוסדיים. אם הערעור מגיע לגורם שמחזיק בהטיות דומות לאלו שהמודל כבר משעתק, או אם יש עומס רב על הגורמים המערערים, הערעור עלול להפוך למסלול פורמלי בלבד שאינו מרסן את הטיות המערכת (Franklin, 2017). יתר על כן יש פער מוכר בין קבוצות באוכלוסייה: לקבוצות חזקות, בעלות אוריינות טכנולוגית ושפתית גבוהות, יש מודעות רבה יותר לזכויותיהן ונגישות למשאבים. הן משתמשות בערעורים לעיתים קרובות וביעילות. לעומתן קבוצות מוחלשות באוכלוסייה נוטות להימנע מהגשת ערעורים או מתקשות להתניע את התהליך, דווקא במקרים שבהם המודל מדויק פחות עבורן.

משום כך במהלך ההטמעה יש ליצור מנגנון ערעור נגיש, פשוט ושוויוני, הכולל שימוש בשפה ברורה, אפשרות לערעור בכמה ערוצים (דיגיטלי, טלפוני ופיזי), סיוע למי שזקוק לכך והסבר פומבי על הזכאות לערער ועל משמעויות הערעור במקרה שבו הוא עלול לפעול לרעת המשתמשים. בד בבד הגורמים המקצועיים הדנים בערעור זקוקים להכשרה ייעודית שתעזור להם להבין את מגבלות המודל, לזהות דפוסי שגיאה חוזרים ולבחון את הערעורים במודעות להיבט של הטיות אנושיות.

ולבסוף, מנגנוני ערעור יעילים אינם רק מסלול תיקון פרטני אלא מקור חשוב לשיפור מתמשך של המערכת. ניתוח שיטתי של דפוסי הערעור, כמו מי מערער, על מה, ובאילו הקשרים הערעור מתקבל, יכול להצביע על קבוצות באוכלוסייה שנפגעות בעקביות, על טעויות אלגוריתמיות חוזרות ועל כשלים בממשק. שילוב מידע זה בחזרה לתהליך הפיתוח מאפשר תיקון מתמשך, שיפור והקטנת פערים. במובן זה מנגנוני הערעור תורמים לא רק לדיוק ולהוגנות של המערכת אלא גם ליצירת מערכת שקופה, אחראית ומעוררת אמון עבור הציבור הרחב.

דוגמה לכך היא מנגנון הערעור האנושי במערכת הביטוח הלאומי (Universal Credit) בבריטניה. לאחר קבלת החלטה אוטומטית על ידי מודל בנוגע לקבלת מענקים והחזרים, עומדת למבקש הזכות להגיש ערעור. בבקשה זו, גורם מקצועי אנושי בוחן מחדש את המידע הקיים. מנגנון זה נועד לתקן טעויות, לצמצם השפעות של הטיות אלגוריתמיות, ולהבטיח שמקרים מורכבים יקבלו תשומת לב אנושית. נוסף על כך מנגנון הערעור משמש את הביטוח הלאומי בבריטניה לשיפור מתמשך של המערכת, ונתוני הערעורים מסייעים לזהות קבוצות באוכלוסייה שנפגעות מהחלטות מוטעות או לא-מתאימות (Podoletz & Currie, 2024; Uk Government, n.d.).

מנגנון "human in the loop" – החלטות רגישות ומכריעות חייבות לעבור אישור אנושי

מנגנון זה הוא אחד מנדבכי ההגנה החשובים במערכות רגישות, מבוססות בינה מלאכותית, אשר מקבלות החלטות בעלות סיכון בלתי נסבל (ראו פרק 9). מנגנון זה קובע כי בתהליכים המוגדרים בעלי רגישות גבוהה או השפעה מהותית על זכויות פרט (כגון שלילת זכאות, הטלת קנסות או שיטור), המודל ישמש כלי תומך החלטה בלבד ולא פוסק אחרון, ו**הפוסק האחרון יהיה האדם** (Lazaros et al., 2026).

מנקודת מבט של ניהול סיכונים, הגורם האנושי אינו רק "בורג" בתהליך אלא מתפקד כ"בקר מפחית סיכון" (mitigating control). תפקידו לשמש חסם אחרון בפני טעויות סטטיסטיות, הטיות מובנות בנתונים או מקרי קצה שהמודל לא אומן עליהם. עיקרון זה מעוגן כיום ברגולציה הבין-לאומית, ובראשה חוק הבינה המלאכותית האירופי, ה-EU Artificial Intelligence Act, במושג "שליטה אנושית משמעותית". המונח "משמעותית" מדגיש כי הנוכחות האנושית לא יכולה להיות סמלית בלבד אלא על המפקח האנושי להיות בעל הסמכות והכשירות המקצועית ולהקדיש את הזמן הנדרש כדי לבצע הערכה ביקורתית של המלצת המערכת.

הסיכון המרכזי בשימוש במנגנון זה הוא כשל בבקרה עקב שילוב מערכות מבוססות בינה מלאכותית היוצר סיכון קוגניטיבי שנקרא "הטיית האוטומציה" (automation bias). לפי וגנר (Wagner, 2019), בני אדם נוטים לייחס אובייקטיביות, דיוק ואמת מוחלטת לפלט של מחשב, הרבה יותר מן דיוק האמיתי שלו. בשירות הציבורי, סיכון זה מתעצם בשל גורמי לחץ מערכתיים: עומס עבודה גבוה, שחיקה וחשש מביצוע טעויות אישיות. במצבים אלו מנגנון הבקרה האנושי עלול לקרוס, והעובדים הופכים בפועל ל"חותמת גומי" המאשרת אוטומטית את המלצות המודל

כדי לחסוך זמן או מאמץ קוגניטיבי או להימנע מאחריות (Wagner, 2019). בצד זאת יש סיכון הפוך הנובע מכך שהגורם האנושי עצמו אינו ניטרלי; עובדים עלולים להפעיל שיקול דעת מוטה המבוסס על דעות קדומות, סטראוטיפים חברתיים או עייפות ("הטיות קוגניטיביות"), ובכך להחזיר הטיות שהמודל אולי הצליח לנטרל, או לקבל החלטות מפלות דווקא במקרים שבהם המערכת המליצה על החלטה הוגנת. במצב כזה הארגון חשוף לסיכון כפול: גם קבלת החלטות שגויות בקנה מידה רחב, וגם אשליה של תהליך תקין ומפוקח.

הדוגמה המובהקת ביותר לנזק הנובע מהזנחה של פיקוח אנושי היא פרשת תוכנית Robodebt שפעלה באוסטרליה בשנים 2016-2019. במסגרת תוכנית זו הטמיעה הממשלה מערכת אלגוריתמית לחישוב חובות והונאות של מקבלי קצבאות רווחה. המודל ביצע הצלבה פשטנית בין נתוני רשות המיסים (שנתיים) ובין דיווחי הביטוח הלאומי (דו-שבועיים) והפיק אוטומטית דרישות חוב למאות אלפי אזרחים שנחשדו בקבלת כספים שאינם זכאים להם. הכשל הניהולי היה ההחלטה להסיר את שלב הבקרה האנושית: במקום שעובד יבחן כל אי-התאמה ויפעיל שיקול דעת אם היא נובעת משינוי בהכנסה או מטעות חישוב, המערכת שלחה את המכתבים אוטומטית והעבירה את נטל ההוכחה לאזרח (Braithwaite, 2020). התוצאה הייתה הרסנית: כ-400,000 אזרחים קיבלו חובות שגויים, נגרם נזק נפשי וכלכלי עצום לקבוצות מוחלשות באוכלוסייה, והמדינה נאלצה לבסוף לשלם פיצויים בסך 1.2 מיליארד דולר אוסטרלי. ועדת החקירה הממלכתית קבעה כי הסרת השיקול האנושי וההסתמכות העיוורת על החישוב האלגוריתמי היוו את הגורם המכריע בכישלון (Royal Commission into the Robodebt Scheme, 2023).

כדי להבטיח שהפיקוח האנושי יתפקד כבקרת סיכונים, יש להטמיע מנגנונים פעילים בתהליך קבלת ההחלטות. אפשר, למשל, לתכנן את הממשק כך שימנע אישור אוטומטי ("אשר הכול") ויחייב את העובד לבצע פעולה המעידה על הפעלת שיקול דעת, כגון הקלדת נימוק או סימון ידני של המשתנים שנבדקו. כמו כן יש לעגן בנהלים את סמכות העובד לבטל את המלצת המודל ולטפח תרבות של "ביטחון פסיכולוגי", המבטיחה לעובד גיבוי מלא אם פעל בזהירות יתר, גם במחיר של טעות בדיעבד (בן-ישראל, 2020). יעילות הפיקוח תלויה גם בכשירות המקצועית של הגורם האנושי: כדי שיבין את מגבלות המערכת, עליו להיות איש מקצוע בתחום התוכן (כגון עובד סוציאלי או משפטן) שעבר הכשרה ייעודית (ראו לעיל סעיף 10.3.1). אפשר גם לשלב מנגנוני בקרת איכות סמויים – מקרים פיקטיביים ובהם טעויות מכוונות, במטרה לוודא שהמפקחים אכן מזהים חריגות ואינם מאשרים המלצות ללא בחינה מסודרת.

10.3.2 פתרונות הקשורים לתהליך יצירת המודל ולהערכתו

בדיקת הוגנות יזומה ובקרה מתמשכת (אודיט)

הטמעת מודל חדש יכולה להניב תוצאות בלתי צפויות עקב גורמים אנושיים וטכנולוגיים שיכולים לגרום לכך שגם אם המודל נבדק לפני ההטמעה, לעניין הטיות, התוצאה לאחר ההטמעה תהיה מוטה. גורמים אלו יכולים להיות, בין היתר, בעלי מקצוע שהתייחסותם להמלצות המודל אינה אחידה, המידע שעליו אומן ונבדק המודל לא עדכני או שהכנסת המודל גרמה לשינוי בהתנהגות האנשים (Alvarez et al., 2024).

מתי מומלץ ליישם את הפתרון? מומלץ לבצע בקרה על כל מודל שמטמיעים, בעיקר כאשר מטמיעים מודל רגיש. במצב זה אפשר להטמיע אותו כ"פיילוט" בכמה מקומות, לבחון את השפעתו, ורק לאחר מכן להטמיע אותו הטמעה נרחבת. הטמעה הדרגתית זו יכולה לגלות השפעות בלתי צפויות בשלב הראשון, וכך אפשר לתקן אותן במועד. כדי שניסיון זה יעבוד, יש לבחון אותו בתנאים דומים ככל האפשר לתנאים שבהם תיעשה ההטמעה הרחבה.

איך יש ליישם את הפתרון? בסעיף 10.2.2 תואר תהליך ההערכה הבוחן את ההוגנות של המודל. בשלב זה נבחן האופן שבו משתמשים במודל בפועל, לפי הקריטריונים שנוצרו בתהליך ההערכה. יש לשים לב כי כאשר נבחנת הטמעת המודל, יש לבחון לא רק את המודל בפני עצמו אלא את המודל כפי שהוא בשימוש. כלומר אם הכנסנו מודל הערכת סיכון רפואי, יש לבחון גם כיצד ההמלצות של הרופאים משתנות בעקבות הטמעתו. אם הכנסנו מודל להחלטה על קבלת סטודנטים לאוניברסיטה, יש לבחון מי הסטודנטים שהתקבלו והתחילו ללמוד בפועל.

בחינה של נתונים עדכניים ועדכונם במקרה הצורך

בחלק מן המודלים של בינה מלאכותית, הנתונים עלולים להשתנות עם הזמן. שפת הכתיבה משתנה, הדמוגרפיה של האוכלוסייה במדינה משתנה, ומדדי בריאות ותפוצת מחלות גם הם יכולים להשתנות. בנוסף, עצם הכנסת האלגוריתם עלולה לגרום לשינוי. במקרה כזה יש סכנה שמודל שאומן והוטמע ייעשה לא עדכני, והדבר עלול לפגום הן בביצועיו הן בקבוצות מסוימות באוכלוסייה (Sculley et al., 2015).

מתי מומלץ ליישם את הפתרון? כאשר מאמנים מודל מנתונים, בעיקר אם אלו נתונים שיכולים להשתנות בזמן קצר יחסית, כמו ציוני תלמידים. אם משתמשים במודל קיים וחיצוני, כמו מודל שפה, והחברה המפתחת מעדכנת אותו ומוציאה מודל מעודכן, מומלץ לבדוק ולשקול לעדכן את המודלים הקיימים, גם כאשר המודל החיצוני משמש רכיב במערכת כלשהי (למשל מודל שפה לסיכום פגישות אוטומטי). אם נעשו בדיקות להטיות על המודל המקורי, יש לחזור עליהן גם במודל המעודכן, מאחר שעלולות להיווצר בו הטיות.

איך יש ליישם את הפתרון? כאשר מאמנים מודל מנתונים, אין עדכון חיצוני של המודל וזו אחריותם של מפתחי המודל לבחון אם התפלגות הערכים השתנתה ויש לעדכן את המודל.

לדוגמה, במודל שמעריך מה הסיכוי של תלמיד להשיג תעודת בגרות בהצלחה, המודל אומן על ציוני התלמידים במשך כמה שנים. אם לאחר כמה שנים השתנתה התפלגות הציונים של התלמידים, והציונים במקומות מסוימים השתפרו ובאחרים הידרדרו, יש לעדכן את המודל. יאו ואח' (Yao et al., 2022) התמקדו בשינוי בהתפלגות שקורה בהדרגתיות במשך זמן והציעו דרכים לגלות מתי התרחש שינוי ויש לעדכן את המודל. מחקר זה התמקד בשינוי כללי בהתפלגות ולא בהוגנות כלפי קבוצות שונות באוכלוסייה, אף שאפשר להפעיל את השיטות שצוינו בהם גם קבוצות כאלה.

מנגנוני הסבר (explainability) למשתמשים

אחד החסרונות של שימוש באלגוריתמים מתקדמים במערכות מבוססות בינה מלאכותית הוא חוסר הבהירות של תוצאותיהם מבחינת המשתמשים. בעבר, כאשר אלגוריתמים היו מתוכנתים ידנית על ידי משתמשים, אותם משתמשים ידעו מה משפיע על הפלט של המודל. כיום, כאשר המודלים המורכבים לומדים ממידע, אי אפשר לדעת בקלות מה משפיע על התוצאות. לכן רצוי שבצד הפלט של מודל למידת המכונה יינתן הסבר (Raji et al., 2020). לדוגמה, כאשר מודל מחליט לא לאשר ללקוח הלוואה, יתקבל פלט המסביר את ההחלטה, למשל, ההלוואה אינה מאושרת מאחר שמשכורתו של הלקוח נמוכה ובעברו פיגר בתשלומים.

במרבית המקרים שבהם מופעלים מודלים של בינה מלאכותית, הם אינם ברורים למשתמש ואין הסבר לפלט של המודל, והדבר משפיע גם על הוגנות התהליך. מידע על המשתנים שגורמים למודל ליצור פלט כלשהו עשוי לסייע למשתמש להבין שהמודל משתמש במשתנה מוגן שאסור לו להשתמש בו, כמו מוצא או מין, או במשתנה שאינו רלוונטי לאוכלוסיית מיעוט, כמו מידע רפואי על בני משפחה באוכלוסייה שבה אין מידע כזה. נוסף על כך המידע על המשתנים שחשובים למודל יכול לאפשר לקבוצות באוכלוסייה שנפגעות מהחלטותיו להבין מה הסיבה להחלטה ועשוי לשפר את מצבן. שקיפות המודל גם מאפשרת לבעלי מקצוע לקבל החלטות נכונות יותר, מאחר שהם יודעים מה המודל הביא בחשבון. סיבות אלו מדגישות את הקשר בין בהירות פלט המודל ובין הוגנות.

מתי מומלץ ליישם את הפתרון? אפשר לדאוג למנגנוני הסבר בכל מצב שבו יש מודל שמחליט החלטות או נותן הערכת סיכון, אך הדבר חשוב במיוחד כאשר המודל מייעץ לבעל מקצוע אנושי במצבים של סיכון גבוה. במקרה זה המידע על המשתנים ששימשו את המודל בהחלטה מסייע לבעל המקצוע בבחירה כיצד להעריך את ההמלצה.

איך יש ליישם את הפתרון? יש שני שימושים עיקריים למנגנוני הסבר של אלגוריתמים: (1) בחירה מראש במודל שמייצר מודל שקל לאנשים להבין; (2) ופיתוח אלגוריתם שמנתח את החלטות המודל בדיעבד.

בשימוש מהסוג הראשון, כאשר חשוב להבין את הסיבות להחלטת המודל, אפשר לבחור מראש אלגוריתם למידה שהפלט שלו הוא מודל שאנשים יכולים להבין, כמו רגרסיה ליניארית, שבה

לכל שדה בקלט יש מספר שמשקף את חשיבותו, או עצי החלטה. מודלים אלו הם בדרך כלל פשוטים יותר, ולא תמיד אפשר להשתמש בהם למשימות מורכבות. מנגד יש משימות שבהן הבנת המודל חשובה יותר מן הדיוק, ובהן אפשר להשתמש במודל פשוט אבל מובן, גם אם ביצועיו טובים פחות מביצועי מודלים מסובכים יותר.

בשימוש מהסוג השני, כאשר אי אפשר להשתמש במודל פשוט ועדיין חשוב להבין את הסיבות להחלטת המודל, יש שיטות אלגוריתמיות ליצירת הסברים לאלגוריתמים מורכבים. בשיטות אלו מריצים מספר רב של אלגוריתמים פשוטים יותר, ומטרתם היא לייצר הסבר להחלטות האלגוריתם המורכב. יש לציין שההסברים במקרים האלה נוצרו "מבחוץ", זאת אומרת האלגוריתם שיוצר את ההסבר מנתח את החלטות המודל ויוצר את ההסברים. יש שיטות ליצור שני סוגים של הסברים – הסברים כלליים שמסבירים את הגורמים המשפיעים על החלטת המודל באופן כללי, והסברים פרטניים שמסבירים את החלטת המודל בנוגע לקלט מסוים (Burkart & Huber, 2021).

10.4 ניטור ושיפור

לאחר שהמערכת פועלת מתחיל שלב הניטור והשיפור המתמשך. בשלב זה נאספים נתוני שימוש, נבחנות החלטות המודל לאורך זמן ומזוהים פערים או הטיות שנוצרות בתנאי אמת ולאורך זמן. כאן מתבצע שיפור רציף המבטיח יציבות והוגנות לאורך חיי המערכת.

10.4.1 פתרונות במעטפת השירות הציבורי

קידום מנגנוני משוב מן הציבור

קידום מנגנוני משוב מתמשך מן הציבור הוא רכיב חיוני בפיקוח על מערכות מבוססות בינה מלאכותית רגישות, משום שהוא מאפשר להבין כיצד המערכת פועלת בזמן אמת, גם לאחר שהוטמעה בשירות הציבורי. מנגנוני משוב , כמו סקרים, פלטפורמות משוב ודיווחי משתמשים, מאפשרים לזהות בעיות, הטיות והשפעות בלתי צפויות שנחשפות רק בעת שימוש שגרתי ובהיקפים גדולים. משוב מן הציבור מסייע לזהות קבוצות באוכלוסייה שנפגעות יותר, הבדלים באופן השימוש בין קבוצות שונות וקשיי נגישות שלא נצפו בשלב הפיתוח. כדי שהמשוב יהיה יעיל ולא יעמיק פערים, עליו להיות נגיש בכמה ערוצים (לדוגמה, משוב טלפוני או מקוון), בכמה שפות, ולהציע מסלולים ידידותיים גם למי שיש להם אוריינות דיגיטלית נמוכה (Gelb et al., 2019).

עם זאת יש לזכור כי גם מנגנוני משוב כאלה עלולים לשקף הטיות מבניות: קבוצות חזקות באוכלוסייה נוטות לספק משוב בתדירות ובאיכות גבוהה יותר, ואילו קבוצות מוחלשות באוכלוסייה לעיתים נמנעות מפנייה בשל חסמי שפה, נגישות או אמון. כמו כן אופן המיון והפרשנות של

המשובים על ידי גורמים אנושיים עלול אף הוא להיות מוטה. לכן חשוב להבטיח כי המשוב נאסף, מנותח ומיושם בשיטתיות ובאופן שוויוני (Gelb et al., 2019).

כדי שהמשוב לא יהיה סמלי בלבד אלא יוביל לפעולה, מומלץ לבצע תהליך מסודר של איסוף, ניתוח ותגובה: בחינה שיטתית של המשובים, מעקב אחר דפוסים חוזרים ופרסום פומבי של צעדים שננקטו בעקבותיהם. מנגנוני משוב כאלה מאפשרים תיקון מתמשך, עדכון המדיניות ושיפור המודל לאורך זמן ומשמשים בסיס לפיקוח דינמי שמקדם הוגנות, אמון של הציבור ושימוש אחראי בטכנולוגיה.

ועדות עצמאיות האחראיות לשמירת עקרונות אתיים

נוסף על הוועדות המייעצות הפועלות בשלב התכנון ומבטיחות ייצוג ושילוב של קבוצות מוחלשות באוכלוסייה, ובמקרים רגישים במיוחד, נדרש בשלב הניטור להקים ועדות עצמאיות שתפקידן לוודא עמידה מתמשכת בעקרונות אתיים ולאתר כשלים המתגלים בזמן אמת. ועדות כאלה בוחנות את המערכת ללא תלות בגוף המפתח או המפעיל ומאפשרות זיהוי מוקדם של כשלים אתיים, סיכוני אפליה, היעדר שקיפות או פגיעה לא מכוונת בקבוצות מוחלשות. הן יכולות להיות ועדות פיקוח חיצוניות ועצמאיות הפועלות כגורם ביקורת בלתי תלוי או לחלופין לשלב מומחים לאתיקה בתוך ועדות קיימות המלוות מקרוב את פיתוח המערכת והטמעתה. ועדות אלו דורשות משאבים ותיאום. שילוב מומחים בוועדות קיימות מאפשר רצף עבודה טבעי והיכרות מעמיקה עם ההקשר המקצועי, אך עלול להיחלש כאשר יש לחצים מוסדיים או אילוצים תפעוליים. על כן מומלץ לבחון מהי החלופה המתאימה לכל מערכת ומשרד ואף לשקול מודל משולב: ועדה פנימית שתלווה את העבודה השוטפת בצד גוף חיצוני שיספק ביקורת עומק ויקבע סטנדרטים מחייבים. לדוגמה, ממשלת אוסטרליה הקימה ועדה ייעודית לסוגיות של בינה מלאכותית שמטרתה לפתח מדיניות, סטנדרטים וקווים מנחים לשימוש מוגן, אתי ואחראי במערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי (OECD, 2025).

ועדות אלו אינן חייבות לשמש רק גורם בקרה אלא יכולות גם לתפקד כגוף מייעץ שמלווה את המערכת לאורך זמן: לנתח משובים שמתקבלים מן הציבור, להצביע על דפוסים חוזרים בפניות ולהמליץ על תיקונים ושיפורים נדרשים. בכך הן הופכות לגורם המתווך בין הציבור, המפתחים והדרג המקצועי ומוודאות שהמשוב שנאסף אינו "נופל בין הכיסאות" אלא מתורגם ללמידה מוסדית ולשיפור מתמשך של ההוגנות והשימוש במערכת. כדי להבטיח שהמלצות הוועדות אכן יתורגמו לפעולה, נדרש גם בעל תפקיד מטעם המשרד המפתח שיהיה אחראי להוגנות המודל, לתיאום השוטף עם הוועדות העצמאיות ולחיבור הישיר עם משתמשי הקצה. גורם זה יתרגם את הביקורת וההמלצות למהלכי יישום, יזהה בעיות בזמן אמת ויבטיח שהפיקוח האתי לא נשאר ברמת המלצה אלא מתממש בפרקטיקה היום-יומית (OECD, 2025).

במסמך הנחיות פדרלי מטעם משרד הניהול והתקציב של ארצות הברית (Executive Office of the President, Office of Management and Budget, 2025) ניתנה הנחיה ולפיה

רשויות פדרליות חייבות לאפשר לציבור להגיש משוב על השימוש במערכות מבוססות בינה מלאכותית. ההנחיה מחייבת גופים ממשלתיים לייצר מנגנוני משוב נגישים, רב-ערוציים ושקופים, ולפרסם כיצד משובים אלו נאספים ומביאים לשינויים בפועל במערכות. בכך הנחיה זו מעגנת את משוב הציבור כחלק ממנגנוני הפיקוח והבקרה על מערכות מבוססות בינה מלאכותית בממשל הפדרלי.

ליווי מחקרי מתמשך

ליווי מחקרי מתמשך הוא רכיב יסודי בהבטחת הוגנות לאורך חיי המערכת, מאחר שהשפעתן של מערכות מבוססות בינה מלאכותית אינה קבועה אלא משתנה עם הזמן, עם כניסתם של דפוסי שימוש חדשים ולאחר עדכוני נתונים ואלגוריתמים. מאחר שמודלים של בינה מלאכותית מושפעים משינויים חברתיים וטכנולוגיים, מערכות שפעילותן תקינה בעת ההטמעה עלולות לפתח הטיות חדשות או לפגוע בקבוצות באוכלוסייה שלא זוהו בעבר. ליווי מחקרי מאפשר בחינה שיטתית של האופן שבו הציבור משתמש במערכת, של דפוסי שגיאה חוזרים, של פערים בביצועים בין קבוצות שונות ושל הקשרים חדשים שלא היו ידועים בשלבי הפיתוח (OECD, 2025).

נוסף על כך מחקר עצמאי או משותף עם גופים אקדמיים או גופי חברה אזרחית מגביר את אמון הציבור, מחזק את השקיפות ומרחיב את נקודות המבט בתהליך קבלת ההחלטות. כדי שהפתרון יהיה מועיל, המחקר המלווה צריך להיות מוטמע כרכיב קבוע ולא חד-פעמי, להישען על גישה לנתונים רלוונטיים ולהתבצע בשיתוף המשרדים המפעילים כך שהממצאים יתורגמו בפועל לשינויי מדיניות, לעדכון מודלים, לתיקונים בתהליכי עבודה ולהעמקת מנגנוני בקרה (OECD, 2025).

10.4.2 פתרונות הקשורים לתהליך יצירת המודל ולהערכתו

התאמת מטרות הוגנות לאורך זמן, על פי משוב

יש הטיות שיכולות להתגלות בטווח זמן ארוך, אף של שנים, ולכן חשוב לעקוב אחר ביצועי המודל וההוגנות לאורך זמן. חשוב לבחון את המודל בקביעות לאחר פרק זמן כדי לוודא שלא חלו שינויים בהתפלגויות, וגם שאין השפעות לטווח ארוך שנובעות מאימוץ המודל (Liu et al., 2019).

ליו ואח' (Liu et al., 2019) מדגימים מצב של רצון לפעול בהוגנות – מתן הלוואות לאנשים גם כאשר יש סיכוי גדול למדיי שהם לא יחזירו אותה. החוקרים מראים כי לאורך זמן, דירוג האשראי הממוצע של לוקחי ההלוואה יורד עקב הלוואות שלא שולמו. מחקרם מדגיש את הצורך לבחון את השפעת החלטות המודל במשך שנים, כמו מתן הלוואה, קבלה לתוכנית לימודים או החלטות רפואיות, שכן אפשר שהטיות במודל יתגלו רק כאשר לקוחות יתקשו בהחזר ההלוואה או סטודנטים לא יסיימו את התואר בהצלחה.

נוסף על שינויים לטווח ארוך, כאשר המודל מופעל על אנשים שיכולים שמאפיינים השתנו, ייתכן שהם ישנו את התנהגותם לאורך זמן עקב הפעלתו. לדוגמה, בארצות הברית אנשים משנים את התנהגותם כדי לקבל דירוג אשראי גבוה יותר. במצב זה ייתכן מצב שבו מודל שקיים דרישה מסוימת להוגנות יפסיק לקיים אותה לאחר שינוי ההתנהגות.

מתי מומלץ ליישם את הפתרון? כאשר משתמשים במודל לאורך זמן, בעיקר אם הוא פועל כחלק מתהליך ומושפע מהתנהגויות של המשתמשים, ייתכן שינוי בהתנהגות, אם עקב הטמעת המודל אם ללא קשר, ועולה צורך לעדכן את המטרות מן המודל. גם במצב שבו אין שינויים בהתנהגות או בהתפלגות המשתמשים, או השפעות ארוכות טווח של המודל, ייתכן רצון לשקול מחדש את הגדרות ההוגנות בחלוף כמה שנים, מאחר שהמדיניות השתנתה או שמטרות הכנסת המודל השתנו.

איך יש ליישם את הפתרון? יש להמשיך לבחון את מדדי ההוגנות שהוגדרו באפיון המערכת ולוודא שהמודל עדיין מקיים את מטרות ההוגנות. אם הוא לא מקיים אותן, או אם חל שינוי כללי במדיניות שמשפיע על מטרות ההוגנות, יש לחזור על השלבים של בחינת המודל ובחירה בהגדרות ההוגנות.

לוח 3 מציג סיכום של הפתרונות לקידום הוגנות אלגוריתמית, לפי שלבי פיתוח המערכת.

לוח 3: פתרונות לקידום הוגנות אלגוריתמית, לפי שלבי פיתוח המערכת

שלבי פיתוח המערכת	סוגי מענים	
	פתרונות במעטפת השירות הציבורי	**פתרונות הקשורים לתהליך יצירת המודל ולהערכתו**
תכנון מקדים	▪ שילוב ועדות מייעצות ובהן נציגי קבוצות מוחלשות באוכלוסייה ▪ קביעת מדיניות מחייבת לשקיפות ולפרסום מוסדר של במערכות מבוססות בינה מלאכותית ▪ הכשרות דיגיטליות לקבוצות מוחלשות באוכלוסייה לשימוש במערכות מבוססות בינה מלאכותית ▪ יצירת רישות עסקי (נטוורקינג) בין ארגונים	▪ הגדרת מטרת המודל ואופן השימוש בו ▪ קביעת מטרות הוגנות

שלבי פיתוח המערכת	סוגי מענים	
	פתרונות במעטפת השירות הציבורי	**פתרונות הקשורים לתהליך יצירת המודל ולהערכתו**
פיתוח וטיוב	▪ חיזוק הקישוריות והאינטגרציה בין מאגרי מידע ציבוריים ▪ התאמת תוכן ושירותים לשפה ולתרבות של קבוצות מגוונות באוכלוסייה עבור משתמשי קצה	▪ בחינת איכות המידע ▪ קליברציה (כיול) ומולטי-קליברציה ▪ אכיפת הגדרות של הוגנות, אם נבחרו ▪ הטמעת תהליך הערכה הבוחן סיכונים לחוסר הוגנות ואת תרומת המערכת לשיפור הוגנות ▪ שימוש בכלים קיימים להערכת ההוגנות וה"מחיר" שמשלמים עבורה ▪ חוויות המשתמשים כמדד להוגנות
הטמעה ויישום	▪ הכשרת עובדים לפרשנות של תוצרי מערכות מבוססות בינה מלאכותית ▪ מנגנוני ערעור אנושיים ▪ מנגנון "human in the loop" – החלטות רגישות ומכריעות חייבות לעבור אישור אנושי	▪ בדיקת הוגנות יזומה ובקרה מתמשכת (אודיט) ▪ בחינה של נתונים עדכניים ועדכונם במקרה הצורך ▪ מנגנוני הסבר למשתמשים
ניטור ושיפור	▪ קידום מנגנוני משוב מן הציבור ▪ ועדות עצמאיות האחראיות לשמירת עקרונות אתיים ▪ ליווי מחקרי מתמשך	התאמת מטרות הוגנות לאורך זמן, על פי משוב

11. סיכום

בסקירה הוצגו האתגרים וההזדמנויות הכרוכים בשילוב מערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי. הסקירה העלתה כי שאלת ההוגנות אינה סוגיה טכנית בלבד אלא חלק חשוב בקביעת מדיניות לפיתוח והטמעה של מערכות מבוססות בינה מלאכותית בשירותים במימון ציבורי, בניהול סיכונים ובהבטחת אמון הציבור. בצד האתגרים, הוצג מערך פתרונות שנועד להסב את עקרונות ההוגנות לכלים ולפרקטיקות הניתנים להטמעה וליישום בפועל במשך כל שלבי חיי המערכת.

כדי להבטיח הוגנות, נדרש כבר בתחילת התהליך לקבוע במפורש מהי ההוגנות הרלוונטית למערכת הספציפית: האם הדגש הוא על שוויון בין קבוצות באוכלוסייה, יחס הוגן כלפי פרטים דומים, תיקון עיוותים מערכתיים או שילוב של כולם? ללא בחירה מודעת כזו קשה לנסח מטרות, להעריך סיכונים או לפתח מנגנוני בקרה יעילים. מתוך כך עולה צורך מהותי בביסוס מומחיות ייעודית בתחום של בינה מלאכותית ושילוב מומחי הוגנות המשלבים בין הבנה טכנולוגית עמוקה ובין הבנה חברתית ומוסדית. מומחים אלו נדרשים כבר בשלבי הפיתוח הראשונים, משום שהחלטות מוקדמות, כגון בחירת נתונים, הגדרת ערכי אמת או קביעת מטרות המערכת, מעצבות את גבולות ההוגנות והסיכון עוד לפני שהמודל עובר לשלב של יישום.

כמו כן יישום הוגנות אלגוריתמית בישראל נתקל בחסם מבני של גודל שוק ומאגרי מידע מוגבלים. במדינה קטנה, היעדר מסה קריטית של נתונים על קבוצות מיעוט עלול להוביל להטיית תת-ייצוג ולפגיעה בדיוק המערכת עבורן. כדי לשפר את הייצוגיות אפשר לאמץ אפיקי פעולה, כגון איחוד נתונים מנהליים בין משרדי ממשלה, עיצוב משתף המערב את הקהילות עצמן ובחינת איכות ומגוון המידע. מלבד זאת, התמהיל הישראלי המורכב, המאופיין ברב-תרבותיות, פערי שפה ומאפיינים ייחודיים כמו מצבי חירום ומלחמה, מחייב גישה זהירה ומותאמת הקשר המבטיחה הוגנות ומוגנות גם בתקופות של חוסר יציבות.

לכן בצד פיתוח כלים אוטומטיים, יש לבחון היכן שימוש במערכות מבוססות בינה מלאכותית עלול לא להתאים או לטשטש שיקול דעת מקצועי. נדרש יישום הדרגתי מבוקר הכולל פיילוטים ייעודיים לזיהוי הטיות בתנאי אמת. כחלק ממעטפת הפתרונות, יש לשים דגש בהעלאת מודעות לשירותים ולזכויות הדיגיטליות, במיוחד בקרב אוכלוסיות מן הפריפריה הגאוגרפית והחברתית. ללא מסעות פרסום והנגשה פעילה, פערים באוריינות דיגיטלית ובמודעות לקיומם של הכלים עלולים להעמיק את אי-השוויון במקום לצמצמו. ניהול סיכונים כזה מחייב מיפוי ייעודי, ניטור רציף ושקיפות מלאה כלפי הציבור.

ולבסוף, כאשר מערכות מבוססות בינה מלאכותית מופעלות באחראיות, הן יכולות לשמש מנגנון לתיקון הטיות היסטוריות וקידום צדק חברתי. כדי לממש פוטנציאל זה ולהפוך תיאוריה לפרקטיקה, מוצע לקדם מערך מחקרי המשך בשלושה מישורים:

במישור הממשק והעצמת האזרח – יש לבחון כיצד עיצוב המידע משפיע על מימוש זכויות ולהעמיק ביכולת של אזרחים לזהות מתי בינה מלאכותית מעורבת בהחלטה בנוגע להם. מחקר זה צריך לכלול דרכים לפיתוח יכולת ביקורתית של האזרח כלפי המידע המוצג והתייחסות לבינה מלאכותית כגורם מעצב התנהגות, בשימת לב מיוחדת לאוכלוסיות בעלות אוריינות דיגיטלית נמוכה.

במישור המערכתי ואמון הציבור – נדרש מחקר על "חוסן דיגיטלי" ועל הפער שבין הוגנות טכנית בפועל ובין תפיסת ההוגנות החברתית. יש לבחון כיצד בונים לגיטימציה ציבורית למערכות רגישות בפריפריה, וכן לעקוב אחר השפעות מצטברות של מדיניות הוגנות לאורך שנים.

במישור הטכנולוגי והישראלי – מוצע להתמקד בטכנולוגיות של מולטי-קליברציה המסוגלות להבטיח דיוק וכיול גם עבור קבוצות קטנות ומצטלבות באוכלוסייה, ובכך למנוע אפליה עקיפה. בצד זאת יש לחקור שיטות ליצירת מידע סינטטי הוגן שיאפשרו אימון מודלים ללא הטיות מבניות גם במצבים של מחסור בנתוני אמת איכותיים. הצעה נוספת למחקרי המשך היא בחינת השימושים במודלים של בינה מלאכותית בהקשר הישראלי, והתאמת היעדים והמדדים של ההוגנות לשימושים אלו. בחירת המדד הנכון של הוגנות היא שאלה מורכבת וכאמור, הוגנות נמדדת באופן שונה לפי אופן השימוש במודל. על מדדים אלו להתמקד לא רק במודל עצמו אלא בדרך שבה הוא משמש בפועל, וזאת בצד הבאה בחשבון של הגורם האנושי.

מקורות

בן-ישראל, י., מתניה, א. ופרידמן, ל. (2020). *המיזם הלאומי למערכות נבונות בטוחות: מסמך מסכם*. אוניברסיטת תל אביב, סדנת יובל נאמן למדע, טכנולוגיה וביטחון.

בקר, א. (2019). *שיתוף ציבור במשרדי הממשלה וברשויות המקומיות*. הכנסת, מרכז המחקר והמידע.

ג'וינט ישראל-אשל ומשרד הדיגיטל הלאומי – ישראל דיגיטלית. (2020). *איך מלמדים אזרחים ותיקים כלים דיגיטליים מרחוק – עקרונות להוראת אזרחים ותיקים כלים דיגיטליים*.

דולב, ה., חסין, ט. ולנטו ט. (2021). *ניהול סיכונים בפיקוח על שירותים חברתיים: עקרונות של פרקטיקה מיטבית: סקירה בין-לאומית*. דמ-858-21. מכון מאיירס-ג'וינט-ברוקדייל.

מרכז רפואי רבין. (א"ת). *בינה מלאכותית (AI) – בילינסון NEXT*.

ליטווין, א. וספיר, א. (2008). מתודולוגיה: מבנה ותוכני סקרי SHARE-ישראל. *ביטחון סוציאלי 76*, 25-42.

לף, י. (2023). *פערי מידע בנוגע לאוכלוסייה הבדואית בנגב*. דמ-929-23. מכון מאיירס-ג'וינט-ברוקדייל.

מבקר המדינה. (2021). *דוח ביקורת שנתי 71: המוכנות לשוק העבודה המשתנה - הסביבה הלימודית בבתי הספר העל-יסודיים כתשתית להקניית מיומנויות המאה ה-21*.

מור, ג. (2018). *מדריך לניהול סיכונים ברגולציה ובמדיניות ציבורית*. מדינת ישראל: משרד ראש הממשלה

מנדלקרן, ר. ושרמן, א. (2015). הפרטה של שירותים חברתיים באמצעות מיקור-חוץ. בתוך י. גל-נור, א. פז-פוקס וג. ציון (עור'). *מדיניות ההפרטה בישראל: אחריות המדינה והגבולות בין הציבורי לפרטי* (עמ' 265-319). הקיבוץ המאוחר ומכון ון ליר.

מערך הדיגיטל הלאומי, משרד המשפטים ומשרד החדשנות, המדע והטכנולוגיה. (2025). מדריך לניהול סיכונים ושימוש אחראי בכלי בינה מלאכותית (AI) במגזר הציבורי.

תרשיש, נ. (2017). *מדינות רווחה בראייה משווה: כיצד להגדיר את ישראל?* מרכז טאוב.

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2020). The connection between popularity bias, calibration, and fairness in recommendation. *Proceedings of the Fourteenth ACM Conference on Recommender Systems (RecSys 2020)*, 726–731.

Afrose, S., Song, W., Nemeroff, C. B., Lu, C., & Yao, D. (2022). Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Communications Medicine, 2*(1), 111.

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., & Sitaram, S. (2023). Mega: Multilingual evaluation of generative AI. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4232–4267.

AlDahoul, N., Rahwan, T., & Zaki, Y. (2025). AI-generated faces influence gender stereotypes and racial homogenization. *Scientific Reports, 15*(1), 1–10.

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *Proceedings of the ACM on HumanComputer Interaction, 3*(CSCW), Article 91.

Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ruggieri, S., & others. (2024). *Policy advice and best practices on bias and fairness in AI. Ethics and Information Technology, 26*(3), Article 44.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: Risk assessments in criminal sentencing. *ProPublica*.

Alon-Barkat, S., & Busuioc, M. (2023). Human–ai interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory, 33*(1), 153–169.

Anti-Defamation League (ADL). (2025). *Generating Hate: Anti-Jewish and Anti-Israel Bias in Leading Large Language Models*.

Armstrong, L., Liu, A., MacNeil, S., & Metaxa, D. (2024). The silicon ceiling: Auditing gpt's race and gender biases in hiring. *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–18.

Alpaydin, E. (2020). *Introduction to machine learning* (4th ed.). MIT Press.

Bacchini, F., & Lorusso, L. (2019). Race, again: How face recognition technology reinforces racial discrimination. *Journal of Information, Communication and Ethics in Society, 17*(3), 321–335.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 63*(4/5), 4:1-4:15.

Blodgett, S. L., & O'Connor, B. (2017). *Racial disparity in natural language processing: A case study of social media African-American english.*

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? 36th Conference on Neural Information Processing Systems (NeurIPS 2022)

Braithwaite, V. (2020). Beyond the bubble that is Robodebt: How governments that lose integrity threaten democracy. Australian Journal of Social Issues, 55(3), 242–259.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research, 70*, 245-317.

Byers, A. (2022). *In the Public AI: How Governments Can Apply Responsible AI Principles to Artificial Intelligence for Public Service Delivery.* Partnership for Public Service.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. *Data Science Journal, 19*(1), 43.

Ciecierski-Holmes, T., Singh, R., Axt, M., Brenner, S., & Barteit, S. (2022). Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: A systematic scoping review. *Npj Digital Medicine, 5*(1), 162.

Cepiku, D., & Mastrodascio, M. (2021). Equity in public services: A systematic literature review. *Public Administration Review, 81*(6), 1019-1032.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). *Transitioning from real to synthetic data: Quantifying the bias in model performance.*

Chen, Y., & Joo, J. (2021). Understanding and mitigating annotation bias in facial expression recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* 14960–14971.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research,* 24, 1–117.

Costa, C. J., Aparicio, M., Aparicio, S., & Aparicio, J. T. (2024). The democratization of artificial intelligence: Theoretical framework. *Applied Sciences, 14*(18), 8236.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Cruz Cortés, E., & Ghosh, D. (2020). An invitation to system-wide algorithmic fairness. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 235–241.

Dagan, N., Devons-Sberro, S., Paz, Z., Zoller, L., Sommer, A., Shaham, G., Shahar, N., Ohana, R., Weinstein, O., Netzer, D., Kotler, A., & Balicer, R. D. (2024). Evaluation of ai solutions in health care organizations – The optica tool. *NEJM AI, 1*(9).

Department for Science, Innovation and Technology. (2025). *Algorithmic transparency recording standard: DSIT - Redbox*. GOV.UK.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580.

Dwork, C., & Ilvento, C. (2019). Fairness under composition. *Proceedings of the 10th Innovations in Theoretical Computer Science Conference (ITCS 2019), 14*(33), 1-20.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Ehsani-Moghaddam, B., Martin, K., & Queenan, J. A. (2021). Data quality in healthcare: A report of practical experience with the Canadian primary care sentinel surveillance network data. *Health Information Management Journal, 50*(1/2), 88–92.

Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., & Pera, M. S. (2018). All the cool kids, how do they fit in? Popularity and demographic biases in recommender evaluation and effectiveness. *Proceedings of Machine Learning Research, 81*, 15-15.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data, 8*(1), 140.

European Commission. (n.d.). *Digital skills and jobs coalition*.

Executive Office Of the President, Office of Management and Budget. (2025). *M-25-21: Accelerating federal use of AI through innovation, governance, and public trust*. The White House.

Go Fair. (n.d.). *FAIR Principles* .

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Foster, D. (2023). *Generative deep learning: Teaching machines to paint, write, compose, and play* (2nd ed.). O'Reilly Media.

Fernando, M. P., Cèsar, F., David, N., & José, H. O. (2021). Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems, 36*(7), 3217–3258.

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *SSRN Electronic Journal*.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.

Franklin, S. (2017). The politics of race, administrative appeals, and medicaid disenrollment in tennessee. *Social Sciences, 6*(1), 3.

Gelb, A., Mittal N., & Mukhrejee, A. (2019). *Towards Real-Time Governance: Using Digital Feedback to Improve Service, Voice, and Accountability*. Center for Global Development.

Gopalan, P., Hu, L., Kim, M. P., Reingold, O., & Wieder, U. (2023). *Loss minimization through the lens of outcome indistinguishability*. Proceedings of the 14th Innovations in Theoretical Computer Science Conference (ITCS 2023).

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., & Wieder, U. (2022). *Omnipredictors. In 13th Innovations in Theoretical Computer Science Conference (ITCS 2022).*

Government Digital Service. (2025). *Algorithmic Transparency Recording Standard Hub.*

Government of Canada. (2025). *Algorithmic Impact Assessment tool.*

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99.

Haas, C. (2019). The Price of Fairness – A Framework to Explore Trade-Offs in Algorithmic Fairness. *Proceedings of the International Conference on Information Systems (ICIS 2019)*.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Bakar, J., Basu, S., Ajlouni, N., & Mirjalili, S. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Han, X., Chen, T., Zhou, K., Jiang, Z., Wang, Z., & Hu, X. (2025). You Only Debias Once: Towards Flexible Accuracy-Fairness Trade-offs at Inference Time.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems, 29*.

Hébert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning* (pp. 1939–1948). PMLR.

Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2018). *A moral framework for understanding fair ml through economic models of equality of opportunity.* Proceedings of the Conference on Fairness, Accountability, and Transparency, 181–190.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission.

Hing, E., & Burt, C. W. (2009). Are there patient disparities when electronic health records are adopted? *Journal of Health Care for the Poor and Underserved, 20*(2), 473–488.

Hutt, S., Gardner, M., Duckworth, A. L., & D'Mello, S. K. (2019). Evaluating fairness and generalizability in models predicting on-time graduation from college applications. *Proceedings of the International Conference on Educational Data Mining (EDM)*.

Hu, L., Livni-Navon, I., Reingold, O., & Yang C. (2023). Omnipredictors for constrained optimization. *International Conference on Machine Learning*.

Hu, S., Oppong, A., Mogo, E., Collins, C., Occhini, G., Barford, A., & Korhonen, A. (2025). Natural language processing technologies for public health in africa: Scoping review. *Journal of Medical Internet Research, 27*, e68720.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5491–5501). Association for Computational Linguistics.

Iansiti, M., & Lakhani, K. R. (2020). *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world.* Harvard Business Review Press.

Intelligent. (2023). *8 in 10 colleges will use AI in admissions by 2024*. Retrieved August 6, 2025.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Rawls, J. (1971). *A theory of justice.* The Belknap Press of Harvard University Press.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences, 118*(22), e2018340118.

Kleinberg, J. (2018). Inherent trade-offs in algorithmic fairness. *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, 40–40.

Klinke, A., & Renn, O. (2002). A new approach to risk evaluation and management: Risk-based, precaution-based, and discourse-based strategies. *Risk Analysis: An International Journal, 22*(6), 1071-1094.

Klinke, A., & Renn, O. (2012). Adaptive and integrative governance on risk and uncertainty. *Journal of Risk Research, 15*(3), 273-292.

Klinke, A., & Renn, O. (2021). The coming of age of risk governance. *Risk Analysis, 41*(3), 544-557.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.

Lam, C. (2025). A systems thinking approach to algorithmic fairness. *arXiv*.

Lavee, E. (2021). Who is in charge? The provision of informal personal resources at the street level. *Journal of Public Administration Research and Theory, 31*(1), 4-20.

Lazaros, K., Vrahatis, A. G., & Kotsiantis, S. (2026). Human-in-the-Loop Artificial Intelligence: A Systematic Review of Concepts, Methods, and Applications. *Entropy, 28*(4), 377.

Lechevalier, F., & Saville, M. P. (2025). Fairness by design: Combatting deceptive AI-driven interfaces. *Cambridge Forum on AI: Law and Governance, 1*(31), 1-20.

Li, Y., Chen, H., Fu, Z., Ge, Y., & Zhang, Y. (2021). User-oriented fairness in recommendation. *Proceedings of the Web Conference 2021*, 624–632.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2019). Delayed impact of fair machine learning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 6196–6200.

Liu, Y., Gautam, S., Ma, J., & Lakkaraju, H. (2024). Confronting llms with traditional ml: Rethinking the fairness of large language models in tabular classifications. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3603–3620.

Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science, 32*(2).

Meagher, G., & Szebehely, M. (2019). The politics of profit in Swedish welfare services: Four decades of social democratic ambivalence. *Critical Social Policy, 39*(3), 455-476.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*(1), 141–163.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition, 45*(1), 521–530.

Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R., Antypas, D., & Oh, A. (2024). Blend: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems, 37*, 78104–78146.

Naik, R., & Nushi, B. (2023). Social biases through the text-to-image generation lens. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society,* 786–808.

Nakao, Y., Strappelli, L., Stumpf, S., Naseer, A., Regoli, D., & Gamba, G. D. (2023). Towards responsible ai: A design space exploration of human-centered artificial intelligence user interfaces to investigate fairness. *International Journal of Human–Computer Interaction, 39*(9), 1762–1788.

Nam, H., Nam, T., & Kim, S. (2024). Identifying the determinants of platform-based e-government service use: *Journal of Global Information Management, 32*(1), 1–21.

National Data Infrastructure (2025). *Ministério da Gestão e da Inovação em Serviços Públicos.*

New York City, Automated Decision Systems Task Force. (2019). New York City Automated Decision Systems Task Force Report.

Nikolenko, S. I. (2021). *Synthetic data for deep learning.* Springer.

Noordeh, E., Levin, R., Jiang, R., & Shadmany, H. (2020). Echo chambers in collaborative filtering based recommendation systems.

Northpointe, Inc. (2015). *COMPAS risk and needs assessment system: Practitioner's guide.*

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

OECD. (2019). *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies.*

OECD. (2025). *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions.*

OECD & UNESCO. (2024). *G7 Toolkit for Artificial Intelligence in the Public Sector.*

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Parker, I., Studman A., & Jones, E. (2025).. *Learn fast and build things: Public sector AI*. Ada Lovelace Institute.

Pessach, D., & Shmueli, E. (2023). Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (pp. 867-886). Springer International Publishing.

Podoletz, L., & Currie, M. (2024). Automating universal credit: A case of temporal governance. *First Monday, 29*(2).

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo Chambers on Facebook. *SSRN Electronic Journal*.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). *Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing*. arXiv.

Ricci, F., Rokach, L., & Shapira, B. (2022). Recommender systems: Techniques, applications, and challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (3rd ed., pp. 1–35). Springer.

Riley, R. D., & Collins, G. S. (2023). Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal, 65*(8), 2200302.

Rolf, E., Worledge, T. T., Recht, B., & Jordan, M. I. (2021). Representation matters: Assessing the importance of subgroup allocations in training data.

Royal Commission into the Robodebt Scheme. (2023). *Report*.

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review, 2*(1).

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2020). How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence, 283,* 103238.

Schmidthuber, L., Ingrams, A., & Hilgers, D. (2021). Government openness and public trust: The mediating role of democratic capacity. *Public Administration Review, 81*(1), 91–109.

Schejter, A., & Tirosh, N. (2016). The digital divide in Israel. In A. Schejter & N. Tirosh (Eds.), *The digital divide: The internet and social inequality in international perspective* (pp. 189–210). Routledge.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. *Proceedings of the 29th International Conference on Neural Information Processing Systems.*

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency,* 59-68.

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning, 112*(9), 3211-3260.

Songur, W. (2023). System of choice promotes ethnically-profiled elderly care and older migrants' use of elderly care: Evidence from Sweden's three largest cities. *Public Money & Management, 43*(6), 610-617.

Sorokovikova, A., Chizhov, P., Eremenko, I., & Yamshchikov, I. P. (2025). Surface fairness, deep bias: A comparative study of bias in language models. *The 2024 ACM Conference on Fairness, Accountability, and Transparency.*

Stankovich, M., Behrens, E., & Burchell, J. (2023). *Toward Meaningful Transparency and Accountability of AI Algorithms in Public Service Delivery*. Center for Digital Acceleration.

Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., Kenton, Z., Crossan, S., Mohamed, S., Rutherford, D., Dandekar, R., Kosoy, R., Bohm, C., Hughes, C., Kornblith, S., & Hassabis, D. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature, 572*(7767), 116–119.

U.S. Census Bureau. (2021). *Disclosure avoidance system for the 2020 Census: TopDown algorithm detailed design specification (Version 1.1)*.

UNDP. (n.d.). AI *Empowerment Programme for Civil Servants: Strengthening Digital Capacities in Public Administration*.

United Nations Educational, Scientific and Cultural Organization. (UNESCO). (2022). *Recommendation on the Ethics of Artificial Intelligence*.

United Nations Educational, Scientific and Cultural Organization. (UNESCO). (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models*.

Urbina, J. T., Vu, P. D., & Nguyen, M. V. (2025). Disability ethics and education in the age of artificial intelligence: Identifying ability bias in chatgpt and gemini. *Archives of Physical Medicine and Rehabilitation, 106*(1), 14–19.

Van Ness, M., Bosschieter, T. M., Halpin-Gregorio, R., & Udell, M. (2023). The missing indicator method: From low to high dimensions. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 5004–5015.

VerifyWise. (n.d.). AI fairness metrics. VerifyWise AI Lexicon.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7.

Virvou, M. (2023). Artificial intelligence and user experience in reciprocity: Contributions and state of the art. *Intelligent Decision Technologies, 17*(1), 73–125.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841-887.

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet, 11*(1), 104–122.

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. *Findings of the Association for Computational Linguistics*, 3730–3748.

Wang, A., Bai, X., Barocas, S., & Blodgett, S. L. (2025). Measuring machine learning harms from stereotypes requires understanding who is harmed by which errors in what ways. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.

Wyllie, S., Shumailov, I., & Papernot, N. (2024). Fairness feedback loops: Training on synthetic data amplifies bias. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2113–2147.

World Health Organization (WHO). (2023). WHO calls for safe and ethical AI for health.

Wyner, J. S., Bridgeland, J. M., & DiIulio, J. J., Jr. (2007). *Achievementrap: How America is failing millions of high-achieving students from lower-income families*. Civic Enterprises.

Yao, H., Choi, C., Cao, B., Lee, Y., Koh, P. W. W., & Finn, C. (2022). Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems, 35*, 10309-10324.

Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. *Advances in Neural Information Processing Systems (Vol. 30, pp. 2921–2930)*.

Zanger-Tishler, M., Nyarko, J., & Goel, S. (2024). Risk scores, label bias, and everything but the kitchen sink. *Science Advances, 10*(13), eadi8411.

Zhang, J., Shu, Y., & Yu, H. (2023). Fairness in design: A framework for facilitating ethical artificial intelligence designs. *International Journal of Crowd Science, 7*(1), 32–39.

Zhang, Y., & Long, Q. (2021). Assessing Fairness in the Presence of Missing Data. *Advances in neural information processing systems, 34*, 16007–16019.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2023). *A survey of large language models. arXiv*.

Zink, A., Obermeyer, Z., & Pierson, E. (2024). Race adjustments in clinical algorithms can help correct for racial disparities in data quality. *Proceedings of the National Academy of Sciences, 121*(34), e2402267121.