מוגנו·ת בעידן הבינה המלאכותית

תהילה שוורץ אלטשולר | מיכאל סיארה

עיון

המכון הישראלי לדמוקרטיה

המכון הישראלי
לדמוקרטיה

מוגנות בעידן הבינה המלאכותית

תהילה שוורץ אלטשולר | מיכאל סיארה

Safeguarding in the Age of Artificial Intelligence:
Toward a New Framework of Protection
Tehilla Shwartz Altshuler | Michael Sierra

עיצוב הסדרה והעטיפה: סטודיו AlfaBees
ביצוע גרפי: רונית גלעד, ירושלים
עיצוב התרשימים: נאוי קצמן
הדפסה: גרפוס פרינט, ירושלים

התצלום על העטיפה: תהילה שוורץ אלטשולר, בסיוע כלי AI

מסת״ב: ISBN 978-965-342-522-4

אין לשכפל, להעתיק, לצלם, להקליט, לתרגם, לאחסן במאגר ידע, לשדר או לקלוט בכל דרך או אמצעי אלקטרוני, אופטי או מכני או אחר – כל חלק שהוא מהחומר בספר זה. שימוש מסחרי מכל סוג שהוא בחומר הכלול בספר זה אסור בהחלט אלא ברשות מפורשת בכתב מהמו״ל.

© כל הזכויות שמורות למכון הישראלי לדמוקרטיה (ע״ר) 2026

נדפס בישראל, תשפ״ו/2026

המכון הישראלי לדמוקרטיה
רח׳ פינסקר 4, ת״ד 4702, ירושלים 9104602
טל׳: 02-5300888
אתר האינטרנט: www.idi.org.il

כל פרסומי המכון ניתנים להורדה חינם, במלואם או בחלקם, מאתר האינטרנט.

המכון הישראלי לדמוקרטיה

המכון הישראלי לדמוקרטיה הוא מוסד עצמאי א־מפלגתי, מחקרי ויישומי, הפועל בזירה הציבורית הישראלית בתחומי הממשל, הכלכלה והחברה. יעדיו הם חיזוק התשתית הערכית והמוסדית של ישראל כמדינה יהודית ודמוקרטית, שיפור התפקוד של מבני הממשל והמשק, גיבוש דרכים להתמודדות עם אתגרי הביטחון מתוך שמירה על הערכים הדמוקרטיים וטיפוח שותפות ומכנה משותף אזרחי בחברה הישראלית רבת הפנים.

לצורך מימוש יעדים אלו חוקרי המכון שוקדים על מחקרים המניחים תשתית רעיונית ומעשית לדמוקרטיה הישראלית. בעקבותיהם מגובשות המלצות מעשיות לשיפור התפקוד של המשטר במדינת ישראל ולטיפוח חזון ארוך טווח של תרבות דמוקרטית נכונה לחברה הישראלית ולמגוון הזהויות שבה. המכון שם לו למטרה לקדם בישראל שיח ציבורי מבוסס ידע בנושאים שעל סדר היום הלאומי, ליזום רפורמות מבניות, פוליטיות וכלכליות ולשמש גוף מייעץ למקבלי ההחלטות ולציבור הרחב.

המכון הישראלי לדמוקרטיה הוא זוכה פרס ישראל לשנת תשס"ט על מפעל חיים – תרומה מיוחדת לחברה ולמדינה.

הדברים המובאים בספר זה אינם משקפים בהכרח את עמדת המכון הישראלי לדמוקרטיה.

תוכן העניינים

מבוא 7

שער ראשון: מוגנוּת בעידן הבינה המלאכותית 15

פרק ראשון: מהי מוגנוּת בעידן הבינה המלאכותית – המשגה וסוגי פגיעוּיוּת 17

פרק שני: סוגי בינה מלאכותית 30

פרק שלישי: סיכוני בינה מלאכותית – רקע עיוני 45

פרק רביעי: תמות חוזרות בשיח על רגולציה של בינה מלאכותית 65

פרק חמישי: "שוברי השוויון" בעידן הבינה המלאכותית 69

פרק שישי: מנגנוני התערבות לחיזוק מוגנוּת ולהתמודדות עם פגיעות – תיאור כללי 78

פרק שביעי: מנגנוני התערבות לחיזוק מוגנוּת ולהתמודדות עם פגיעות – תיאור פרטני 87

שער שני: מוגנוּת של קבוצות אוכלוסייה ייחודיות 119

פרק שמיני: מוגנוּת ילדים ונוער בעידן הבינה המלאכותית 121

פרק תשיעי: מנגנוני התערבות לקידום מוגנוּת של הציבור החרדי בעידן הבינה המלאכותית 186

פרק עשירי: מנגנוני התערבות לקידום מוגנוּת של הציבור הערבי בעידן הבינה המלאכותית 221

פרק אחד עשר: מנגנוני התערבות לקידום מוגנוּת של זקנים בעידן הבינה המלאכותית 259

סיכום: מוגנוּת כפרויקט טכנו־חברתי בעידן הבינה המלאכותית 298

מבוא

העידן הנוכחי מאופיין בהאצה טכנולוגית חסרת תקדים לא רק בקצב הפיתוח של מערכות בינה מלאכותית אלא בעיקר בהשתלבותן המהירה בחיי היומיום: צ׳אט עם חבר, אבחנה רפואית או תיעדוף בקבלת שירותי רווחה ממשלתיים. טכנולוגיות שכמעט לא היו מוכרות לפני שלוש שנים הפכו לכלי פעולה מקובלים ולעיתים בלתי נראים בכל רובדי החברה. במציאות זו שאלת המוגנות של בני אדם במרחבים טכנולוגיים כבר איננה שאלה עתידנית אלא אתגר ״הווה־אני״ בוער.

מחקר זה נכתב מתוך הכרה בשלושה פערים מרכזיים. ראשית, פער מתמשך בין מהירות ההתפתחות הטכנולוגית לבין מהירות פיתוחם של מענים חינוכיים, מוסדיים, רגולטוריים וקהילתיים. זהו **פער הקצב הטכנולוגי־חברתי** המתואר היטב על ידי פלורידי (Floridi) ושותפיו במסגרת קריאתם ל״רגולציה אנטיציפטורית״, גישה המבקשת לנסח מדיניות

באופן מקדים, גמיש ודינמי ולא רק בדיעבד.[1] שנית, מתחדד גם **פער האחריות והיכולת.** בעוד האחריות למוגנות נופלת לא פעם על כתפי הפרט – הורה, ילד, מורה או מטפל – הכוח להשפיע על עיצוב המערכות, המוצרים והמדיניות נותר בידי גופים ריכוזיים ומורכבים. בהקשר זה נדרשים מודלים של "אחריות מדורגת", המטילים חובות אתיות ורגולטוריות על גופים ושחקנים בהתאם לעוצמת השפעתם.

פער שלישי נוגע **לפרטיקולרי לעומת האוניברסלי** ועוסק במתח שבין גיבוש עקרונות הגנה החלים על כלל האוכלוסייה לבין התאמה תרבותית, קהילתית ומבנית של פתרונות. גישת ארגון אונסק"ו[2] בהקשר זה היא של "אוניברסליות דיפרנציאלית", מסגרת של עקרונות משותפים המיושמים באופן מותאם להקשרים מקומיים. בהינתן הבדלים מהותיים בין קבוצות גיל, מגדר, רקע חברתי, מוצא אתני ועוד, ברמות אוריינות דיגיטלית, נגישות לשירותים ומאפיינים תרבותיים, גישה זו חיונית לבניית מדיניות מותאמת.

בתוך כל אלה חשוב להדגיש כי הפגיעות אינן מחולקות באופן שוויוני באוכלוסייה. קבוצות מסוימות, ובהן ילדים ובני נוער, אזרחים ערבים, בני החברה החרדית וגם אוכלוסיית הזקנים, חוות את המפגש עם מוצרים מבוססי טכנולוגיית בינה מלאכותית באופן שונה ולעיתים פגיע יותר בהקשרים מסוימים. הפערים נובעים ממצב קוגניטיבי (בעיקר בקרב ילדים וזקנים), חסמים מוסדיים ותרבותיים (כמו חוסר אמון או נגישות לשירותים ממלכתיים), מגבלות גישה לדיגיטציה, או תלות באחרים בשימוש בטכנולוגיה. גורמים זדוניים המכירים בחסמים אלה עלולים גם לנצל אותם לרעה עוד יותר. לכן, אחת ממטרות המחקר היא להתמודד עם ההנחה השגורה כאילו כל אדם נפגע באופן דומה ולהציע המשגה שמאפשרת הבחנה, רגישות תרבותית ופיתוח מדיניות מותאמת קבוצות. כולנו חשופים במידה כזו או אחרת לפגיעה של השפעות הטכנולוגיה, אך יש מי שפגיע יותר. ילדים, זקנים, קהילות מוחלשות, נפגעים לעיתים מוקדם יותר, קשה יותר וללא יכולת התמודדות אפקטיבית. מטרת המחקר אינה להציע הגנה אוניברסלית מופשטת, אלא להתמקד באותן קבוצות המצויות בסיכון ייחודי. אנו מודעים לכך שעיצוב מכליל של מדיניות המוגנות יתרום בסופו של דבר

1 Luciano Floridi, Josh Cowls, Monica Beltrametti et al., *AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, 28 Minds & Machines 689, 709 (2018)

2 *Artificial Intelligence & Emerging Technologies*, UNESCO.ORG

לכלל החברה, אבל תכליתנו היא להביא את המשתייכים לקבוצות הייחודיות לאותה רמת פגיעות/מוגנות של הציבור הכללי.

קפיצות מדרגה טכנולוגיות, כמו שילוב של סוכנים אוטונומיים, מחשוב רגשי ומרחבי או פרסונליזציה מבוססת דאטה, אינן מייצרות רק הזדמנויות, אלא גם **מאיצות פערים קיימים**. בנוסף, ככל שהמערכות נעשות חכמות יותר, כך מתגבר הפער בין מי שמסוגלים להבין, לפרש ולהגיב להן לבין מי שנשארים מאחור, מבחינת אוריינות, גישה או שפה. במצבים כאלה טכנולוגיה שאמורה לשרת את הציבור עלולה דווקא לשעתק הדרה, ליצור שקיפות יתר כלפי קהילות מוחלשות או לבסס מנגנוני שליטה חדשים במסווה של יעילות. ההכרה בכך היא תנאי ליצירת מדיניות מוגנות שאינה רק מגינה אלא גם מתקנת.

בצד אלה, מחקר זה מבקש להתייחס גם למתח העמוק שבין **הגנה ופטרנליזם לבין אוטונומיה של הפרט**. מושגים כגון פרטיות, חופש ביטוי והסכמה מדעת ניצבים כיום למול מצבים חדשים שבהם אינטראקציה רגשית עם מערכת חכמה או קבלת החלטה תפקודית דרך סוכן בינה מלאכותית יוצרים מרחבים עמומים של שליטה, הבנה ובחירה. בהקשר זה נציין את המושג "ריבונות דיגיטלית", המשלב הגנה וכבוד לאוטונומיה האנושית גם במצבים שבהם אין בהירות מוחלטת באשר ליחסי הכוח בין המשתמש למערכות או לגורמי כוח אחרים.

המחקר עוסק לעיתים בסיכונים חמורים, כאלה שנוגעים לאובדן שליטה, לפגיעה קוגניטיבית מתמשכת או להעמקת אי־שוויון חברתי. עם זאת, הפתרונות המוצעים אינם תמיד רדיקליים, אלא לעיתים פרגמטיים, מקומיים והדרגתיים. אין בכך סתירה. אנו מכירים בכך שבמציאות הנוכחית חברות הטכנולוגיה אינן מאבדות מכוחן, משטרים אינם נעשים דמוקרטיים יותר, והכוחות המעצבים את שוק הטכנולוגיה ממשיכים לפעול במסלולים עצמאיים. איננו כותבים מתוך אשליה שנוכל לשנות סדרי עולם, אך גם לא מתוך ייאוש. המחקר הזה נכתב מתוך תחושת אחריות: לעשות את מה שאפשר, להציע כיוונים יצירתיים ולתרום ככל יכולתנו לשינוי הדרגתי של מאזן הכוחות בין האדם לבין מוצרים מבוססי טכנולוגיה והגורמים העומדים מאחוריהם.

המחקר הנוכחי יוצא לדרך מתוך זיהוי של חולשה יסודית בשיח הקיים: רבים מהדיונים על בינה מלאכותית ממוקדים ב"סיכונים מערכתיים" או ב"איומים קיומיים", אך הם נוטים להזניח את החוויה האנושית הקונקרטית של פגיעות. אנו מבקשים להפוך את נקודת המבט: לא לשאול אילו סיכונים יוצרת הבינה המלאכותית, אלא אילו סוגים של פגיעות היא מאפשרת: בגוף, בנפש, בקוגניציה, ברכוש ובמרחב החברתי והקבוצתי.

לצורך כך פיתחנו מסגרת אנליטית חדשה המורכבת מארבע שכבות ניתוח: סוגי טכנולוגיות, סוגי פגיעות, מאפייני אוכלוסיות ודרכי התערבות. גישה זו מאפשרת לנו למפות את יחסי הגומלין בין מערכות חכמות לבני אדם בצורה רב־ממדית, גמישה ופרקטית, מתוך שאיפה לעבור ממושגים מופשטים של "אתיקה" או "זכויות" לניתוח קונקרטי של מצבי חיים, אוכלוסיות סיכון והקשרים מוסדיים. מודל הפעולה שלנו משלב בין פתרונות שמחייבים תִפעול אקטיבי מצד מוסדות, יוזמות שראוי פשוט לאפשר להן להתקיים ולהתרחב, מהלכים שנכון להצטרף אליהם מתוך שותפות גלובלית ורכיבים שדורשים תכלול מדינתי אסטרטגי. זוהי תפיסה מרובדת של מוגנות שאינה נשענת על פתרון אחד אלא על אקוסיסטם של התערבויות.

הגישה שאנו מציעים במחקר זה מבקשת להרחיב את הפרספקטיבה המקובלת על סיכוני בינה מלאכותית מתוך חיבור שיטתי בין שלושה מוקדים: יכולות המערכת הטכנולוגית ויישומיה, מאפייני הפגיעות האנושית והקשרים חברתיים־קבוצתיים. חיבור זה יאפשר ליצור מדרג התערבויות על בסיס הערכת סיכון רגישה ורב־ממדית.

המתודולוגיה מבוססת על ארבע שכבות ניתוח המקיימות ביניהן יחסי גומלין:

1. סוגי הטכנולוגיות
2. סוגי הפגיעויות
3. סוגי אוכלוסייה ומאפייניהן הייחודיים
4. דרכי התערבות

שכבות אלו ייבנו זו על גבי זו בהמשך המחקר. ראשית, נמשיג את המושג מוגנות ואת סוגי הפגיעויות הנגזרות ממנו. לאחר מכן נסקור חמש טכנולוגיות בינה מלאכותית קיימות ואת השלכותיהן. לאחר מכן נסקור את סוגי הסיכונים המוכרים בספרות העדכנית. המתודולוגיה שאנו מציעים היא למעשה ההיגיון המארגן שיאפשר לנו לנתח את **האינטראקציות שבין שכבות אלו** ולגזור מהן המלצות מעשיות המותאמות לקבוצות שונות ולמערכות טכנולוגיות שונות.

שכבה 1: סוגי טכנולוגיות

המערכת הראשונה של המתודולוגיה כוללת סיווג של טכנולוגיות AI לפי מנגנוני הפעולה

המרכזיים שלהן, ולא לפי מטרתן השיווקית או התחום שבו הן מיושמות. נציג חמש קטגוריות: מערכות חיזוי, בינה מלאכותית גנרטיבית ומולטי־מודלית, סוכני בינה מלאכותית, מחשוב חישתי (affective computing) ומחשוב מרחבי (spatial computing). הסיווג לפי עקרונות פעולה מאפשר לנו לבחון תופעות פגיעות החוזרות במופעים שונים אך נובעות ממנגנון דומה, למשל תלות קוגניטיבית במערכות גנרטיביות או רגישות כלפי ממשקים מבוססי קול והבעה.

שכבה 2: סוגי הפגיעויות

המערכת השנייה מתמקדת בסוגי "פגיעות" וכוללת הבחנה בין פגיעות פיזית, רגשית, קוגניטיבית, פיננסית וחברתית־קבוצתית. כל אחת מהן כוללת בתוכה ממדים של טווח זמן (מיידי מול מצטבר), סוג המושפע (נפגע, עד, פוגע), מקור הפגיעה (מוטה־טכנולוגיה ומאפייני מוצר מול ניצול לרעה של אלה), פגיעות חדשה מול פגיעות מוכרת אך מועצמת ומנגנון הפעולה (טכנולוגיה, משתמש, מוסד). העמדה זו תאפשר לנו, בהמשך הדרך, לשאול לא רק **מה הסיכון** אלא **מה סוג הפגיעות שהוא יוצר** ובאילו מנגנונים של התערבות נכון לטפל בו.

חשוב להבהיר כבר בפתח הדברים כי מחקר זה עוסק הן בפגיעויות מוכרות, אשר טכנולוגיות בינה מלאכותית מעצימות, והן בפגיעויות חדשות וייחודיות הנובעות מאופני הפעולה, קנה המידה והאוטונומיה של מערכות בינה מלאכותית מתקדמות. לצד אלה, אנו מודעים לכך שמרחב הסיכון אינו ממצה: בהינתן קצב ההתפתחות, מורכבות המערכות והאינטגרציה שלהן בחיי היומיום, קיימים גם "unknown unknowns" – סיכונים שטרם זוהו או נוסחו, אך סביר שיתגלו רק בדיעבד, לאחר שכבר יצרו פגיעה ממשית. ההכרה בממד זה של אי־ודאות אינהרנטית איננה חולשה מתודולוגית, אלא נקודת מוצא הכרחית לחשיבה אחראית על מוגנות בעידן הבינה המלאכותית.

שכבה 3: קבוצות אוכלוסייה ומאפיינים ייחודיים

השכבה השלישית בוחנת את האופנים שבהם פגיעויות מצטלבות עם מאפיינים קבוצתיים. ההנחה הקיימת במרבית מסמכי המדיניות כיום היא שכל סוג פגיעות פועל באופן זהה

על כלל האוכלוסייה. אכן, רוב הפגיעויות הן "חוצות קבוצות", אך הביטוי שלהן, הסיכוי לזיהוין, עוצמתן ויכולת ההתמודדות עימן מושפעים עמוקות מהקשר תרבותי, חברתי, כלכלי, גילי ומעמדי.

במחקר זה נבחן ארבע קבוצות עיקריות: ילדים ובני נוער, האזרחים הערבים בישראל, החברה החרדית וזקנים. כל אחת מהן נושאת עימה **תצורות ייחודיות של פגיעות** שמושפעת מהתפתחות קוגניטיבית (במקרה של ילדים), חסמים תרבותיים לשימוש בטכנולוגיה (כמו בקרב החברה החרדית), חסמים מוסדיים (כמו חוסר נגישות לשירותים ממלכתיים או חוסר אמון במערכות אכיפה בקרב ערבים בישראל), אוריינות דיגיטלית נמוכה או היעדר נגישות לסביבות דיגיטליות (בקרב אזרחים מבוגרים). לכן אותה טכנולוגיה עצמה תיצור לעיתים פגיעות שונה כתלות בקבוצה הנחשפת לה.

שכבה 4: דרכי התערבות

השכבה הרביעית עוסקת באופני תגובה, ויסות, תיקון והגנה. היא כוללת דרכי פעולה מגוונות: רגולציה ואכיפה; חינוך ואוריינות (של אוכלוסיות וגם של מקבלי החלטות); תוכניות מערכתיות (כמו תוכניות לבתי ספר); סיוע רגשי ממוסד; תכנון ועיצוב של מערכות טכנולוגיות; פיקוח באמצעות אודיטינג; ואף אפשרות שהבקרה תתבצע דווקא על ידי המכונה על האדם, ולא להפך.

מטרת שכבה זו איננה להציע "תשובה אחת" לפגיעה, אלא לאפשר מיפוי דינמי של פתרונות בהתאמה לסוג הפגיעות, מאפייני הקבוצה והטכנולוגיה שדרכה היא נוצרת. הסיבה לכך היא שהמתודולוגיה נועדה לשמש לא רק כלי ניתוח תאורטי אלא בסיס לפיתוח מדיניות, תוכניות פעולה והבנה רחבה של האתגר האנושי בעידן הבינה המלאכותית.

היתרונות של מתודולוגיה זו הם, ראשית, העובדה שהיא מאפשרת התבוננות ניואנסית ורב־רובדית הן של תיאור הבעיה והן של דרכי ההתערבות; ושנית, היותה מודולרית ופתוחה להרחבה כך שניתן להוסיף אליה קבוצות נוספות, טכנולוגיות מתפתחות חדשות, ואף שכבות ניתוח נוספות בעתיד, כמו השפעות של מגדר, הקשרים גאוגרפיים, או מדדים של השפעה נפשית. בכך היא מספקת מסגרת עבודה דינמית, שמאפשרת לעבור מהשיח המופשט על סיכוני בינה מלאכותית לדיון קונקרטי, מחובר לשטח ולבני אדם ממשיים.

מחקר זה נועד לשמש בסיס לפיתוח מדיניות, הנחיות לפעולה, ובמידה מסוימת גם לדיון מוסרי מחודש על מוגנות בעידן שבו הטכנולוגיה אינה עוד כלי בידי האדם אלא לעיתים בן שיחו, מנגנון תיווך לחייו, ולפעמים גם שותף סמוי לעיצוב זהותו.

נדגיש כי תכלית המחקר איננה לעסוק בסוגיות מקרו חברתיות וכלכליות הנובעות מעולם רווי מערכות בינה מלאכותית, כגון שוק העבודה, סוגיית ההשפעה על שוק האנרגיה, מלחמות באמצעים אוטונומיים, מתקפות סייבר רב־מערכתיות וגם יציאה משליטה של מערכות בינה מלאכותית. לא נוכל להתייחס גם לשאלות גדולות כגון כוחן העודף של ענקיות הטכנולוגיה; משבר המעקב ואיסוף נתונים על פרטים באינספור דרכים וכיוצא באלה. אין ספק שאלה יהיו משום סיכונים עצומים למדינות, חברות ולפרטים, אך הם אינם חלק ממה שאנחנו תופסים באופן צר כאתגרי מוגנות.

הביבליוגרפיה שעליה נשען המחקר משקפת את גוף הידע המחקרי הזמין עד סתיו 2025. עם זאת, אין בכך כדי לצמצם את הרלוונטיות של המסקנות וההמלצות המוצעות כאן. מטרתו המרכזית של המחקר איננה מיפוי נקודתי של טכנולוגיות או תוצרים עדכניים, אלא פיתוח עקרונות, מסגרות חשיבה וכלי ניתוח שיישארו תקפים גם לנוכח שינויים טכנולוגיים מהירים. דווקא בעידן שבו הידע מתיישן במהירות, יש חשיבות מיוחדת להתמקדות בלוגיקות מבניות, בדפוסי כוח ובמנגנוני פגיעות החוזרים על עצמם גם כאשר היישומים הקונקרטיים משתנים.

בהקשר זה ראוי לציין את המתח בין החשש מפני קריאה להתערבות רגולטורית או מוסדית שאינה מבוססת די הצורך על מחקר אמפירי ונתונים מוצקים, לבין הסיכון שבהמתנה פסיבית עד להתגבשות ודאות מחקרית מלאה. מחקר זה מאמץ במודע את עמדת ה"אין ברירה": במציאות של האצה טכנולוגית, היעדר פעולה עלול להיות עצמו בחירה רבת־סיכון. לפיכך, אנו מבקשים להציע כיווני מדיניות וזהירות מבוססי ידע קיים, גם אם חלקי, מתוך הכרה מפורשת במגבלות הידע ובצורך בעדכון מתמיד.

***

ספר זה מבוסס על מחקר במימון ג'וינט ישראל בתמיכת קרן מוריס וויויאן וואהל.

אנו מודים למומחי המכון הישראלי לדמוקרטיה, עו"ד שלומית רביצקי טור־פז, ד"ר גלעד מלאך, רוני ברבוי, תהילה גאדו, ד"ר מוחמד ח'לאילה וד"ר אריק רודניצקי, ששיתפו אותנו

בתובנותיהם הדיסציפלינריות המעמיקות במסגרת הראיונות שקיימנו איתם. כן אנו מודים לד״ר עמיר גפן מאוניברסיטת בר־אילן ומשרד החינוך.

תודה מיוחדת לעדי אליאסי על תכלול הפרויקט. תודה לתמר שקד על העריכה, לחלי פרוימוביץ׳ על ההפקה, לרונית גלעד על העימוד ולנאוי קצמן על עיצוב התרשימים.

במחקר זה נעזרנו במערכת Deep Research של ChatGPT ובמערכת Genspark, על יתרונותיהן וחסרונותיהן, אך המחקר עצמו וכתיבת המסמך נעשו על ידינו.

שער ראשון:

מוגנות בעידן הבינה המלאכותית

פרק ראשון

מהי מוגנוּת בעידן הבינה המלאכותית – המשגה וסוגי פגיעוּיות

המושג "מוגנוּת" נזכר לא אחת בדיונים על בטיחות טכנולוגית, חוסן דיגיטלי, אמון ואתיקה של מערכות. מדובר במושג שגור בעברית בשיח העוסק בזכויות ילדים, בטיחות דיגיטלית, תחום הרווחה, החינוך, בריאות הנפש ועוד. כאשר מנסים להגדיר את ההפך מן **המוגנוּת** – כלומר **פגיעוּת** – ניתן לאתר בקלות את המונח vulnerability. חוק הבינה המלאכותית של האיחוד האירופי שנכנס לתוקף באוגוסט 2024 קובע כי לא תותר הצבה בשוק האירופי של טכנולוגיות, שירותים ומוצרים המנצלים –

> [...] פגיעוּת כלשהי של אדם טבעי או של קבוצה מסוימת של בני אדם בשל גילם, מוגבלותם או מצב חברתי או כלכלי מסוים, במטרה, או באופן שיש בו כדי, לעוות באופן מהותי את התנהגותו של אותו אדם

> או של אדם המשתייך לאותה קבוצה, באופן שגורם או שסביר שיגרום לאותו אדם או לאדם אחר נזק משמעותי.[3]

מסמך ההנחיות שהאיחוד יצר והתפרסם סמוך לכניסת החוק לתוקף מגדיר את המונח vulnerability כך שהוא כולל את הרגישות/פגיעות (susceptibility) הקוגניטיבית, הרגשית, הפיזית או אחרת, הנובעת מהשתייכותו של אדם לקבוצה מובחנת בשל גילו, מצבו הגופני או מצבו החברתי־כלכלי.

כאשר מבקשים למצוא הגדרה סדורה ומקובלת למושג "מוגנות", מגלים פער שמתבטא בכך שאין למושג מקבילה ישירה אחת באנגלית. המילה protection מתארת הגנה כללית מפני סכנות, פגיעות או איומים, בין שמדובר בפגיעה פיזית, רגשית או מערכתית, והיא השכיחה ביותר בשיח המדיניות הדיגיטלית הבינלאומית. לצידה מופיעים המושגים safeguarding[4] וכן trust and safety המשמשים במיוחד בהקשרים של הגנה על ילדים

3 AIA Art. 5(1)(b) – Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act) 29.7.2025 [התרגום שלנו]:

> [...] exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm.

4 באתר ארגון Social Care Institute for Excellence מוגדר המונח כך: "Safeguarding is protecting a person's rights to live in safety and free from abuse and neglect". הגדרה רשמית ביחס לקשישים, ראו למשל במסמך המלווה את The Care Act הבריטי – Care and support statutory guidance, Gov.UK:

> Protecting an adult's right to live in safety, free from abuse and neglect. It is about people and organisations working together to prevent and stop both the risks and experience of abuse or neglect, while at the same time making sure that the adult's wellbeing is promoted including, where appropriate, having regard to their views, wishes, feelings and beliefs in deciding on any action. This must recognise that adults sometimes have complex interpersonal relationships and may be ambivalent, unclear or unrealistic about their personal circumstances.

וקבוצות פגיעות בפלטפורמות דיגיטליות.[5] מונח זה מדגיש אחריות מוסדית מתמשכת. מונחים נוספים הם resilience (חוסן),[6] המדגיש את היכולת להתמודד עם פגיעה, להתאושש ולחזור לתפקוד במצב חדש; security (ביטחון או אבטחה), המוכר במיוחד בעולמות הסייבר (כגון cyber security),[7] וכן strength (עוצמה), או אפילו invulnerability, מונח נדיר יותר שמשמש בעיקר לציון **אי־פגיעוּת מוחלטת**, כמו זו שמיוחסת לדמויות על־אנושיות. כל אחד מהמונחים הללו תורם נדבך למורכבות של המושג "מוגנוּת", אך אף אחד מהם לא מקיף אותה באופן שלם. לפיכך, בחירתנו להשתמש בעברית דווקא במילה "מוגנוּת" אינה רק לשונית אלא מושגית: היא מבקשת להציע מסגרת רב־ממדית להגנה על בני אדם במרחבים טכנולוגיים, הכוללת רגולציה, חינוך, עיצוב טכנולוגי, תמיכה רגשית ואחריות מוסדית וחברתית.

במחקר זה אנו מציעים מסגרת מושגית סדורה למונח "מוגנוּת", שתשמש בסיס להמשגה, מיפוי וניתוח ובסופו של דבר גם להתערבות.

5 מונח זה כולל לרוב ארבעה מרכיבים, וחשיבותו היא בכך שהוא מתייחס אל הפלטפורמות הדיגיטליות עצמן והחובות שלהן ולא בהכרח למשתמש, וכן לאפשרות של ניצול הפלטפורמות לרעה על ידי שחקנים זדוניים:

• The proliferation of online services and user-generated content;
• The recognition among many large internet companies that their products can be exploited in ways that bring about unintended consequences;
• The realization that if efforts are not undertaken to mitigate such misuse and abuse, both users' experiences and company reputations and profits are likely to erode;
• The recognition that trust and safety issues as well as techniques used by bad actors are constantly developing, requiring ongoing and evolutionary responses.

ראו TSPA, *Introduction to Trust & Safety*.

6 ראו יעל אילת ון אסן וגלית ולנר **שילוב בינה מלאכותית במצבי חירום: מבט ביקורתי** (מכון ון ליר בירושלים, 2025).

7 מושג זה מתייחס בעיקר להגנת המערכות הטכנולוגיות עצמן: "Cybersecurity is the practice of protecting systems, networks, and programs from digital attacks" (CISCO, What is Cybersecurity?). ראו תהילה שוורץ אלטשולר, רחל ארידור, עידו סיון סיבילה **מהו סייבר? – חלק א: על מרחב הסייבר, תקיפות סייבר והגנת סייבר** (המכון הישראלי לדמוקרטיה, 2021).

מוגנות בעולם הטכנולוגי היא המצב שבו משאבים חיוניים של יחידים וקבוצות: הגוף, הנפש, הרגש, הקוגניציה, הרכוש, המוניטין (לחלופין – ההון האישי, המשפחתי, החברתי, הסימבולי והכספי) מוגנים מפני:

- **טכנולוגיות המעוצבות באופן שעלול לאפשר פגיעה במשאבים אלה;**
- **יחידים, ארגונים או קבוצות המשתמשים בטכנולוגיה כדי לאפשר פגיעה במשאבים אלה,**

באמצעות סט של יכולות התערבות – משפטיות, מוסריות, חינוכיות, טכנולוגיות וחברתיות.

בהיעדר מונח באנגלית להצעתנו למונח מוגנות, נציע שלוש אפשרויות: digital shielding, techno-social protection או digital safeness. כל אחת מן האפשרויות האלה אינה מדויקת לחלוטין, וכך או כך אינה מונח מוכר בספרות או במסמכי המדיניות. דווקא עניין זה כשלעצמו מלמד על הצורך בחשיבה מחודשת על הגדרות.

הואיל והדגש שלנו הוא על סט יכולות ההתערבות, כלומר על יצירה והגנה על מצב של "מוגנות אפקטיבית", אנו מגדירים וקטורים של "מוגנות/פגיעות" שלתוכם ניתן יהיה בהמשך להתאים את כלי ההתערבות המתאימים.

אנו מגדירים חמש קטגוריות מרכזיות של **מוגנות/פגיעות:**

1. **מוגנות/פגיעות אישית־פיזית:** פגיעה ממשית בגוף, החל באלימות פיזית ופגיעה מינית, וכלה במעקב פיזי והטרדה.

2. **מוגנות/פגיעות אישית־רגשית־נפשית־פסיכולוגית:** מעשים המתרחשים במרחב הדיגיטלי לרוב אינם יוצרים נזק פיזי אלא נזק רגשי, כגון חשיפת תוכן אינטימי, שיימינג, חרם, הטרדה דיגיטלית או מינית ברשתות החברתיות, פגיעה במוניטין. פגיעות פסיכולוגית כוללת גם השפעות רגשיות ונפשיות המתהוות בעקבות שימוש ממושך או אינטנסיבי במערכות טכנולוגיות, לרבות פיתוח תלות רגשית בצ'אטבוטים ובמערכות AI המספקות מענה רגשי, תחושת אינטימיות כוזבת, דימוי גוף שלילי, חרדות חברתיות ו־FOMO כמו גם חוויות של בידוד רגשי וחוסר שייכות.

3. **מוגנות/פגיעות אישית־קוגניטיבית:** פגיעות זו נחלקת לשני מופעים עיקריים: האחד הוא **פגיעות קוגניטיבית בהבנת המציאות**, כלומר ערעור היכולת להבין ולהעריך את המציאות באופן עצמאי, בין היתר בשל הצפה של מידע סותר, התמסרות למקורות לא מהימנים, הפצת קונספירציות ואובדן אמון ביכולת האישית להבחין בין אמת לשקר, תופעה המכונה לעיתים "ערעור אונתולוגי". השני הוא **חשש מפני התדרדרות קוגניטיבית**, שמשמעה פגיעה הדרגתית ביכולות חשיבה גבוהות בשל הישענות יתר על מערכות AI לקבלת החלטות או למתן תשובות. מדובר בהיחלשות של כישורים כגון הסקה לוגית, פתרון בעיות, חשיבה ביקורתית ויצירתיות אשר עלולה להיגרם משימוש ממושך בממשקים גנרטיביים ובסוכני בינה מלאכותית המספקים תשובות אוטומטיות, מבלי לאתגר את המשתמש לחשוב בעצמו.

4. **מוגנות/פגיעות אישית־פיננסית:** הונאות, גנבות זהות, טרגוט צרכני אגרסיבי, שימוש בדאטה אישי למטרות רווח כלכלי, התמכרות לקניות והימורים.

5. **מוגנות/פגיעות חברתית/קבוצתית:** מעשים המכוונים אל קבוצות חברתיות או מבוססים על השתייכות קבוצתית של אדם, למשל הסתה לגזענות, אנטישמיות, יצירת קיטוב חברתי והעמקת שסעים, אפליה מבוססת אלגוריתמים, וכן ניצול של מאפיינים ייחודיים של קבוצות (למשל חוסר מסוגלות של ילדים וחשש מפנייה למשטרה של קבוצות מיעוט).

אנו מודעים לכך שסוגי הפגיעויות שזורים אלה באלה. אף שבחרנו להציג את סוגי הפגיעות כמקטעים נפרדים לצורכי המשגה וניתוח, חשוב להדגיש כי במציאות הגבולות ביניהם חופפים, ושכיחות הפגיעות המשולבת גבוהה. פגיעות אחת עלולה להוביל לאחרת או להחמיר אותה: כך למשל, פגיעה פיזית חמורה תגרור לרוב גם השלכות רגשיות, קוגניטיביות ולעיתים אף פיננסיות; ופגיעה קבוצתית מלווה פעמים רבות בפגיעה אישית באנשים המשתייכים לאותה קבוצה, אם דרך אפליה שיטתית (המכונה לעיתים פגיעה בהוגנות), הדרה או ערעור על תחושת הערך העצמי. ואולם ההבחנה העיונית מאפשרת סדר והבנת מכלול הקשרי המושג.

סוגי מוגנות / פגיעות

נוסף על כך, מצב ה"מוגנות" איננו בינארי, וקשה להניח כי נוכל להשיג מצב של מוגנות מוחלטת. תפיסתנו היא שהגברת המוגנות איננה רק הגנה מפני פגיעות אלא גם תנאי מוקדם ליכולתו של אדם להשיג איכות חיים גבוהה יותר, שוויון הזדמנויות ויתרונות חברתיים נוספים בעידן הנוכחי.

לכן, הבנת מכלול ההקשרים בין הפגיעויות היא חיונית לבחירה באופני התערבות הוליסטיים; והבנת רצף המוגנות היא חיונית לזיהוי של מקור הסיכון אך בד בבד גם של שרשרת התוצאות האפשריות.

כחלק מן המתודולוגיה של המחקר פיתחנו מסגרת להערכת רמת הסיכון של פגיעויות שונות, הנשענת על ארבעה פרמטרים מרכזיים.

הפרמטר הראשון הוא **מקור הסיכון**: אם מדובר בסיכון מערכתי, הנובע ממאפייני הליבה של הטכנולוגיה עצמה, כגון עיצוב הממשק, מנגנוני האופטימיזציה, לוגיקות הפרסונליזציה או מודלי התמרוץ הכלכליים, או בסיכון זדוני, שמקורו בשימוש לרעה בטכנולוגיה בידי גורמים אנושיים.

הפרמטר השני הוא **מיידיות הסיכון**: כלומר טווח הזמן שבו הפגיעה מתממשת – פגיעה מיידית וברורה, לעומת פגיעה מצטברת, הדרגתית או ארוכת טווח, שלעיתים קשה לזהותה בזמן אמת. למשל, חרם דיגיטלי יכול להוביל לבידוד חברתי מיידי; שימוש קבוע בפילטרים עלול לגרום לשיבוש בדימוי העצמי בטווח הבינוני; חשיפה חוזרת לאלגוריתמים מקטבים עלולה להוביל להקצנה פוליטית ארוכת טווח. לכן אנו מתייחסים לממד הטמפורלי של הפגיעות בדרך הבאה: מניעה (prevention) וחיזוי (prediction) לפני התרחשות אירוע;

התמודדות עם אירוע בזמן התרחשותו וסמוך לאחריו (absorption); הסתגלות למצב חדש (adaptation); ותהליכי השתנות בעקבותיו (transformation).

הפרמטר השלישי הוא **היקף הסיכון:** עד כמה הפגיעה היא צרה וממוקדת – למשל באוכלוסייה מסוימת, הקשר שימוש מסוים או תרחיש ייחודי – או רחבה, כזאת העלולה להשפיע על קבוצות אוכלוסייה גדולות, על מערכות חברתיות שלמות או על תהליכים דמוקרטיים ומוסדיים. בהיקף הסיכון נכללים גם מושאי הפגיעות: נפגעים (מי שסבלו במישרין מפגיעה); עדים לפגיעה (מי שנפגעו משום שנחשפו למעשי פגיעה); מי שלא פועלים למניעת פגיעה; וכן פוגעים (כדי למנוע פגיעות נוספות).

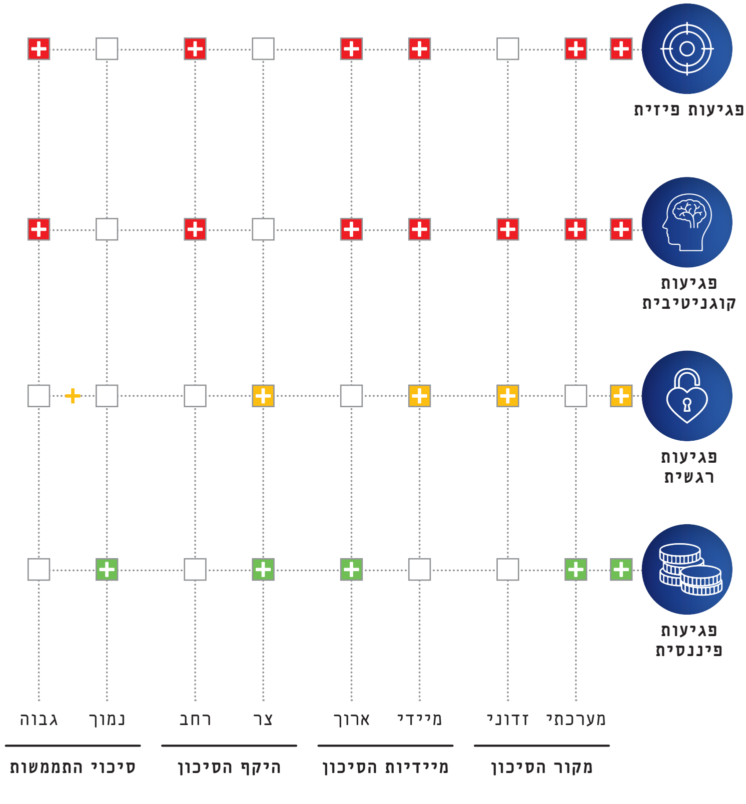
הפרמטר הרביעי הוא **סיכוי ההתממשות של הסיכון:** כלומר הערכה איכותית של ההסתברות שהסיכון אכן יתממש בפועל, בהינתן מאפייני הטכנולוגיה, דפוסי שימוש קיימים וידע מצטבר מן השדה. בהערכת סיכוי התממשות הסיכון יש להבחין, בין השאר, בין העוצמה של פגיעויות קיימות (למשל: החרפה של תופעת חרמות חברתיות דרך קבוצות ווטסאפ) לבין הופעת פגיעויות חדשות (למשל: השפעות קוגניטיביות שנובעות מהתמסרות קבועה להמלצות של מערכות AI).

הערך של מסגרת זו אינו טמון בכל אחד מן הפרמטרים לבדו, אלא בשילוב ביניהם. פגיעות שמקורה מערכתי, שסיכוייה להתממש גבוהים, ושפגיעתה רחבה או מצטברת לאורך זמן, מייצרת רמת סיכון שונה מהותית מפגיעה זדונית, נקודתית או כזו שסיכוייה נמוכים. באופן דומה, סיכון מיידי אך צר עשוי לדרוש תגובה שונה מסיכון ארוך טווח ורחב היקף, גם אם שניהם חמורים מבחינה ערכית. הצירוף בין מקור הסיכון, מיידיותו, היקפו וסיכויי התממשותו מאפשר לנו להעריך לא רק **אם** קיימת פגיעות, אלא **כיצד** היא פועלת במציאות החברתית, ואיזה סוג של התערבות עשוי להיות רלוונטי ויעיל יותר.

חשוב להדגיש כי מסגרת זו אינה מתיימרת לספק מדידה מדעית מדויקת או כימות הסתברותי של סיכונים. אין מדובר במודל חיזוי סטטיסטי, אלא בכלי אנליטי והערכתי המבוסס על ידע מחקרי קיים, ניסיון מצטבר ותובנות מן השדה. בהקשר של טכנולוגיות בינה מלאכותית, המתאפיינות בדינמיות, באי־ודאות ובהשתנות מהירה, גם הימנעות מהערכה היא בחירה בעלת משמעויות. לפיכך, אנו מציעים לראות בהערכת הסיכון המוצעת קריאת כיוון מושכלת: כלי שמאפשר השוואה, תיעדוף וחשיבה מערכתית, מתוך מודעות מלאה לכך שהערכות אלה עשויות להשתנות עם הצטברות ידע, שינויי טכנולוגיה והקשרים חברתיים

חדשים. בעינינו, גם בהיעדר ודאות מלאה, יש ערך משמעותי להצגת מדרגי סיכון כהזמנה לדיון, לתיעדוף ולפעולה זהירה.

חישוב רמת הסיכון לפגיעות

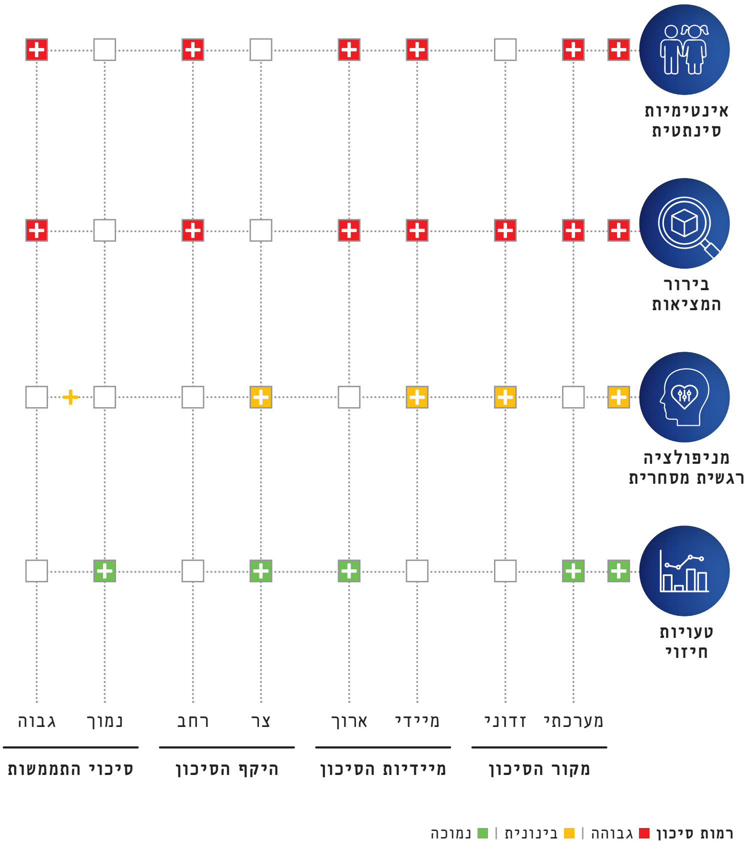


הערה: הסימונים במפה להמחשה בלבד.

כעת, משהצענו הגדרה מקיפה למושג "מוגנות" ומיפינו את סוגי הפגיעויות האפשריים, וכיצד טכנולוגיות שונות, בעיצוב שלהן, במנגנוני ההפעלה שלהן ובשימושים המתאפשרים באמצעותן, משפיעות על הפגיעויות הללו, ניתן לשאול כיצד הדבר משפיע על הקבוצות

השונות שבהן עוסק מחקר זה: ילדים, חרדים, ערבים, זקנים. במילים אחרות: לאחר ששאלנו אילו סוגי טכנולוגיה מייצרים סיכונים, יש לשאול אילו אוכלוסיות חשופות במיוחד לסוגים מסוימים של פגיעות, וכיצד ניתן להתאים את ההתערבות־מקדמת־המוגנות לסיטואציות משתנות ולמאפייני הסיכון הספציפיים. על בסיס מדדי הסיכון שהוזכרו למעלה ביצענו הערכת סיכון יחסית לכל אחת מקבוצות האוכלוסייה הנסקרות במחקר. הערכה זו מאפשרת לא רק לזהות אילו פגיעויות קיימות, אלא גם להמחיש את עומק הפגיעה, הצטברותה והייחודיות שלה בהקשר חברתי, תרבותי ומוסדי נתון. בכך נוצרת תמונה השוואתית החושפת דפוסי פגיעות שונים ומדגישה כי לא מדובר בווריאציות של אותו סיכון, אלא באתגרי מוגנות מובחנים.

חישוב רמת הסיכון לפגיעות בקרב ילדים



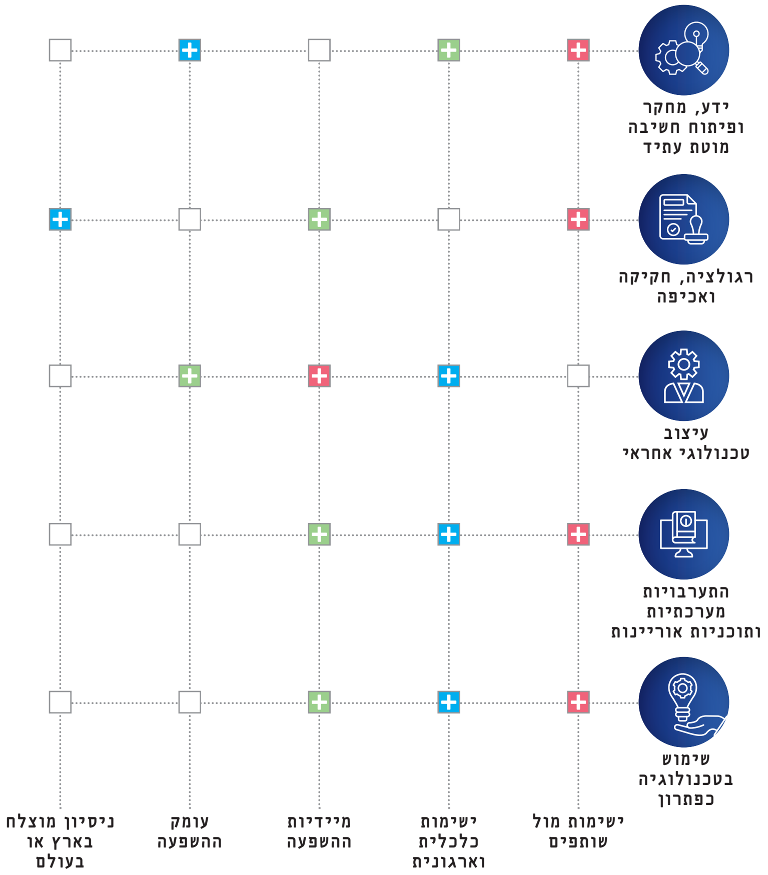
הערה: הסימונים במפה להמחשה בלבד.

לאחר מיפוי הפגיעויות והערכת רמות הסיכון שלהן, נדרש שלב נוסף: בחינה שיטתית של דרכי ההתערבות האפשריות. מיפוי פגיעויות, מקיף ככל שיהיה, אינו מספק כשלעצמו; הוא חייב להיות מתורגם למנגנוני פעולה. לפיכך, המחקר מבקש להציע ארגז כלים רחב של התערבויות – רגולטוריות, חינוכיות, טכנולוגיות, מערכתיות וארגוניות – ולבחון כיצד ניתן להפעילן באופן מושכל בהקשרים המשתנים.

נקודת המוצא שלנו היא שלא כל התערבות ראויה היא גם בת־יישום, ולא כל התערבות ישימה מייצרת בהכרח שינוי משמעותי. לכן, כחלק מן המתודולוגיה פיתחנו גם **מודל תיעדוף של דרכי ההתערבות לפי מידת הישימות שלהן**, המבוסס על שילוב של חמישה שיקולים מרכזיים: מידת התלות בשותפים ובשיתופי פעולה; ישימות כלכלית וארגונית; מיידיות ההשפעה; עומק ההשפעה המבנית; וקיומו של ניסיון יישומי קודם בארץ או בעולם. שילוב שיקולים אלה מאפשר להבחין בין צעדים בעלי השפעה מהירה אך מוגבלת, לבין התערבויות עמוקות יותר שיישומן מורכב או ארוך טווח, ולהבין כיצד ניתן לשלב ביניהם.

מודל התיעדוף אינו מבקש להפיק רשימה היררכית אחת של פתרונות מובילים. מדובר בכלי אנליטי שנועד לשמש קריאת כיוון: לארגן את דרכי ההתערבות כאשכולות משלימים, לאפשר תיעדוף שאינו נאיבי, ולהבהיר מדוע מדיניות אפקטיבית של מוגנות בעידן הבינה המלאכותית מחייבת פעולה רב־שכבתית, בקצבי זמן שונים ובאמצעות שילוב של כלים. במובן זה, התיעדוף משקף תפיסה שלפיה מוגנות אינה מושגת באמצעות פתרון יחיד, אלא באמצעות אקוסיסטם של התערבויות, הפועלות יחד מול סיכונים דינמיים ובמציאות מוסדית מורכבת.

הערכת האפקטיביות של דרכי ההתערבות



רמות אפקטיביות ■ גבוהה | ■ בינונית | ■ נמוכה

הערה: הסימונים במפה להמחשה בלבד.

כך, למעשה, מתודולוגיית המחקר היא ליצור מפת מוגנות/פגיעות אינטגרטיבית המשקפת את הממשק הייחודי בין –

1. מאפייני טכנולוגיות בינה מלאכותית
2. סוגי הפגיעות שהגדרנו ומידת הסיכון
3. המאפיינים הסוציולוגיים והסוציו־טכנולוגיים של הקבוצות שהוגדרו
4. כיווני הצעות להתמודדות ומידת הישימות שלהן

יתרון נוסף של מסגרת הפגיעויות המוצעת בפרק זה הוא אופייה המתפתח. איננו קובעים כאן רשימה סגורה של סיכונים, קבוצות או טכנולוגיות, אלא מציעים מבנה אנליטי המאפשר הרחבה, התאמה ועדכון לאורך זמן. כאשר טכנולוגיות חדשות נכנסות לשימוש, מופיעים דפוסי אינטראקציה חדשים, או מזוהים מופעים חדשים של פגיעות, ניתן לשלבם בתוך המסגרת באמצעות הוספת תתי־קטגוריות, וקטורי ניתוח או קבוצות אוכלוסייה נוספות, מבלי לפרק את ההיגיון המתודולוגי כולו. כך למשל, ניתן יהיה בעתיד להעמיק את הניתוח בהקשרים משפחתיים, מגדריים, גאוגרפיים או תרבותיים; להבחין בין מופעים שונים של פגיעות נפשית או קוגניטיבית; או להתייחס לטכנולוגיות שטרם הבשילו לכדי שימוש רחב אך כבר מעוררות שאלות של מוגנות.

במובן זה, פרק הפגיעויות אינו מבקש "למצות" את מכלול הסיכונים הקיימים, אלא להציע שפת עבודה משותפת ומסגרת מושגית שניתן להפעיל, לבחון ולחדד גם בהמשך הדרך. הגמישות המתודולוגית הזו חיונית במיוחד בעידן הבינה המלאכותית, שבו קצב השינוי הטכנולוגי עולה על קצב ההסדרה והמחקר, והיכולת לעדכן את מפת הפגיעויות חשובה לא פחות מן המיפוי הראשוני עצמו. המסגרת המוצעת מאפשרת אפוא תנועה רציפה בין זיהוי פגיעויות, הערכת סיכון והתאמת מנגנוני התערבות מתוך שמירה על רלוונטיות לאורך זמן ועל חיבור למציאות משתנה.

סיכום

בפרק זה הונחה התשתית המושגית והמתודולוגית שעליה נשען המחקר: הגדרנו את מושג ה"מוגנות" כמסגרת רב־ממדית להגנה על משאבים אנושיים חיוניים, מיפינו חמישה וקטורים מרכזיים של פגיעות והדגשנו כי במציאות הפגיעויות שזורות זו בזו ופועלות

על רצף ולא במצב בינארי. על בסיס זה הצענו מסגרת להערכת רמת הסיכון של פגיעויות שונות, הנשענת על מקור הסיכון, מיידיותו, היקפו וסיכויי התממשותו, ובמקביל הצגנו את השלב הבא: תרגום ידע על פגיעויות להחלטות פעולה באמצעות ארגז כלים של התערבויות ותיעדוף לפי מידת הישימות. באופן זה, ה"מוגנות" אינה נשארת ברמת תיאור הנזק אלא הופכת למפתח לניתוח ולבנייה של מענים מערכתיים.

הפרק הבא יעמיק במיפוי סוגי הטכנולוגיות ומנגנוני הפעולה שלהן, ובאופן שבו הן מעצבות את סוגי הסיכונים. מהלך זה יאפשר בהמשך לחבר באופן מדויק יותר בין טכנולוגיה, פגיעות וקבוצה, ולבסס התאמה מושכלת של התערבויות מקדמות מוגנות לתרחישים ולמאפייני סיכון קונקרטיים.

פרק שני

סוגי בינה מלאכותית

הביטוי "בינה מלאכותית" משמש כיום לתיאור מנעד רחב של טכנולוגיות, מערכות ויישומים, לעיתים שונים בתכלית השינוי זה מזה, שכולם חולקים עקרונות מתמטיים ומנגנונים חישוביים דומים, אך נבדלים באופן פעולתם, בממשקיהם עם בני אדם ובתחומי השפעתם. בפרק זה נציג חמש טכנולוגיות עיקריות שכולן נופלות תחת ההגדרה הרחבה של AI אך מגלמות מופעים שונים במציאות: החל במערכות חיזוי המשמשות בקבלת החלטות, דרך מערכות "גנרטיביות" (בוראות או יוצרות) המייצרות טקסטים, תמונות וסאונד, וכלה בטכנולוגיות ממוקדות רגש או מרחב. כל אחת מהטכנולוגיות הללו מובילה למוצרי קצה אחרים, מזמינה סוגים שונים של שימוש וממילא גם יוצרת סיכונים מסוגים שונים.

מחקר זה מתמקד בעתיד הקרוב (בין שנתיים לחמש שנים), שבו השימושים בטכנולוגיות AI שכבר קיימות כיום יתפשטו והטמעתם תתרחב בשימושים אישיים ויומיומיים, ובמערכות

חינוך, רווחה, בריאות, כלכלה וממשל. מתוך מגוון האפשרויות בחרנו להתמקד בחמש טכנולוגיות בולטות, ששילובים שונים ביניהם צפויים להשפיע מהותית על עיצוב חוויית החיים במרחבים הציבוריים והפרטיים האנושיים:

1. מערכות פרדיקציה (predictive systems)
2. בינה מלאכותית יוצרת ומולטי־מודלית (generative & multimodal AI)
3. סוכני בינה מלאכותית (AI agents)
4. מחשוב רגשי (affective computing)
5. מחשוב מרחבי (spatial computing)

בחלק זה נתאר את מאפייני הליבה של כל אחת מהטכנולוגיות הללו, את האופן שבו היא פועלת, ואת נקודות הממשק המרכזיות בינה לבין חיי האדם, כחוליה ראשונה בזיהוי הסיכונים הרלוונטיים לה.

1. מערכות פרדיקציה

בינה מלאכותית חיזוּיית (predictive AI) היא תת־תחום של בינה מלאכותית המבוסס על יכולת לחזות תוצאות עתידיות על סמך דפוסים שנלמדו מדאטה היסטורי. מערכות אלו פועלות באמצעות שילוב של אלגוריתמים ללמידת מכונה, סטטיסטיקה מתקדמת וניתוח מגמות, והן מאפשרות לזהות הסתברויות או תרחישים צפויים במגוון הקשרים.[8] זוהי טכנולוגיה שמשרתת בראש ובראשונה תהליכי קבלת החלטות, רפואיים, מדעיים,[9] פיננסיים, עסקיים ואף חינוכיים, מתוך הפחתת אי־ודאות והעלאת רמת ההתאמה להקשר או למשתמש.

8 *Predictive Artificial Intelligence*, IBM; *Artificial Intelligence: Predictive vs Generative vs New Mixing AI*, Biomed Grid

9 Melissa Heikkilä, *Google DeepMind Leaders Share Nobel Prize in Chemistry for Protein Prediction AI*, MIT Technology Review, (Oct. 9, 2024)

מערכות פרדיקציה אינן מציעות "נבואות", אלא הערכות הסתברותיות המבוססות על תבניות שנצברו ממאגרי מידע גדולים.[10] ככל שהמערכת נחשפת ליותר נתונים, כך משתפרת יכולת התחזית שלה. מערכות אלו כבר מוטמעות באופן נרחב בשירותים כמו המלצות צרכניות, תחזיות מזג אוויר ותחזיות בריאות. אחד השימושים הבולטים של השנים האחרונות הוא בתחום המדעי: יכולת החיזוי של מבנה חלבונים באמצעות מערכות כמו AlphaFold נחשבת לפריצת דרך מדעית, והייתה בין הגורמים שהביאו להענקת פרס נובל לכימיה בשנת 2024.[11]

היישומים של מערכות פרדיקציה מתרחבים גם לתחומים חברתיים ומינהליים. במערכת החינוך נעשה שימוש במודלים פרדיקטיביים לזיהוי מוקדם של נשירה מלימודים, בעיות רגשיות או קושי בהסתגלות למסגרות – במטרה לאפשר מענה מותאם בזמן אמת.[12] בשירות הציבורי מערכות פרדיקציה משמשות לחיזוי עומסים בחדרי מיון,[13] תכנון תחבורה יעיל יותר[14] ואפילו זיהוי מגמות חברתיות שעלולות להוביל למשברים.

מערכות פרדיקציה מציגות את אחד הביטויים הברורים ביותר של בינה מלאכותית יישומית, כזו שאינה רק מנתחת את ההווה אלא פועלת מתוכו כדי לצפות את העתיד. הן משום נדבך חשוב בהתפתחות תחום קבלת ההחלטות מבוססות־מידע (data-driven decision making) ובמעבר ממדיניות מבוססת תגובה לתחזית.

10 Adam Zewe, *MIT Researchers Introduce Generative AI for Databases*, MIT News (July 8, 2024)

11 *Press Release: The Nobel Prize in Chemistry 2024*, The Royal Swedish Academy of Sciences (Oct. 9, 2024)

12 Margaret Heritage & E. Caroline Wylie, *Formative Assessment and AI: Predictive Models for Early Intervention in Education*, 38 J. Educ. Measurement 204 (2023)

13 Cyrielle Brossard et al., *Predicting Emergency Department Admissions Using a Machine- Learning Algorithm: A Proof of Concept with Retrospective Study*, 25 BMC Emerg. Med. 3 (2025)

14 Fangzhou Sun et al., *Real-Time and Predictive Analytics for Smart Public Transportation Decision Support System*, in 2016 IEEE Int'l Conf. on Smart Computing (SMARTCOMP) 1 (2016)

בהקשר זה יש לציין גם תת־קטגוריה ייחודית של מערכות פרדיקציה, המכונה "טכנולוגיות זיהוי התנהגות" (Behavioral Recognition Technology — BRT).[15] בניגוד למודלים פרדיקטיביים כלליים, אשר מבקשים לחזות תופעות או תהליכים (כגון מזג אוויר, עומסים או תחלואה), מערכות אלו מתמקדות בניתוח דפוסי התנהגות אנושיים לאורך זמן ובתרגומם להערכות הסתברותיות בדבר פעולה עתידית, רמת סיכון או מידת ציות צפויה. מערכות מסוג זה אינן שואלות מה צפוי לקרות, אלא מה צפוי שאדם מסוים יעשה, ועל בסיס הערכה זו מתקבלות החלטות בדבר פיקוח, הקצאת משאבים או התערבות מוקדמת.

בכך מערכות BRT מסמנות שני מעברים בתוך עולם הפרדיקציה: מחיזוי של תהליכים לחיזוי של בני אדם, ומהערכה תיאורית להערכה נורמטיבית. בעוד מודלים פרדיקטיביים רבים נועדו לשפר תכנון או יעילות, מערכות מבוססות BRT משמשות לעיתים קרובות להבחנה בין פרטים וגופים על בסיס "ציוני אמון" או סיכון, ובכך הן הופכות לחלק מתשתית רחבה יותר של ממשל מבוסס־נתונים. התפתחות זו מעוררת שאלות עומק ביחס לגבולות השימוש בלמידת מכונה כאשר מושא החיזוי הוא האדם עצמו, ובפרט כאשר תחזיות אלו משמשות בסיס להתערבות מוקדמת או לאכיפה הסתברותית.

בנוסף, ראוי להבחין בין מערכות פרדיקציה אמיתיות, אשר פועלות על בסיס ניתוח דאטה והסקת הסתברויות, לבין כלים שמציגים תחזיות קיימות מבלי להפעיל מנגנון חיזוי משל עצמם. כך למשל, כאשר מודל כמו ChatGPT נשאל "מה התחזית הכלכלית לשנה הקרובה", הוא אינו מפיק תחזית מקורית אלא מסכם מקורות קיימים מהמרחב הציבורי הדיגיטלי. לעומת זאת, מערכות חיזוי אמיתיות, למשל מודלים לחיזוי סיכויי נשירה של תלמידים או לזיהוי הידרדרות בריאותית, מבוססות על למידת תבניות מדאטה היסטורי ועל התאמה סטטיסטית אישית. בנוסף, לצד היישומים המוסדיים, מערכות פרדיקציה מחלחלות גם לחיי היומיום של משתמשים פרטיים, כמו אפליקציות הממליצות על תפריט לפי הרגלי תזונה ומעקב בריאות, אפליקציות "פיטנס" שמזהות דפוסי אימון לקראת פציעות, או מערכות פיננסיות שמבצעות אופטימיזציה של תקציב אישי בהתבסס על נתונים היסטוריים של המשתמש.

15 Ori Aronson, Yuval Feldman, & Orly Lobel, *Behavioral Recognition Technology*, 102 Indiana Law Journal (forthcoming, 2026); San Diego Legal Studies, Paper 26-008; Bar Ilan University Faculty of Law Research Paper (forthcoming)

2. בינה מלאכותית יוצרת ומולטי־מודלית

בינה מלאכותית יוצרת (Generative Artificial Intelligence) מאפשרת יצירת תוכן חדש – כגון טקסט, תמונות, קוד, מוזיקה או וידאו – באמצעות אלגוריתמים מתקדמים של למידה חישובית. טכנולוגיה זו נשענת על מודלים גנרטיביים הלומדים דפוסים ממאגרי נתונים עצומים, ולאחר מכן מייצרים תוכן חדש, מקורי ובעל מאפיינים דומים למידע שעליו התאמנו.[16]

בלב התחום עומדים מודלים גדולים של שפה (Large Language Models – LLMs), מודלים המבוססים על רשתות נוירונים המאומנים על כמויות אדירות של טקסטים כדי להבין ולהפיק שפה טבעית. המודלים פועלים בעיקר באמצעות ארכיטקטורת Transformer שמאפשרת להם לנתח הקשרים רחבים ולהגיב בצורה קוהרנטית ורלוונטית. מודלים אלה משמשים בסיס למערכות צ׳אט, מנועי חיפוש, עוזרים אישיים, כלים לכתיבה, תרגום וסיכום טקסטים, והם הולכים ותופסים מקום מרכזי בחיי היומיום ובמגוון מקצועות, תחומים ושימושים – החל בכתיבת תכנים שיווקיים, סיכומי ישיבות, עזרה משפטית, וכלה ביצירת תמונות, הפקת מוזיקה, סימולציות, עיצוב גרפי, עריכת וידאו ויצירת דמויות דיגיטליות. דוגמאות מרכזיות כוללות את Claude של Anthropic, Gemini של Google, וכן את GPT של OpenAI.[17] מודלים גנרטיביים מסייעים גם בהפקת נתוני אימון סינתטיים, בפיתוח חומרים חדשים במדע החומרים והביולוגיה הסינתטית, ואף ביצירת קוד ותוכנה באמצעות כלים כמו GitHub Copilot.

ככל שהמחקר והמסחור בתחום מתקדמים, כך גוברת חדירתם של כלים גנרטיביים לתוך תהליכים עסקיים, ממשלתיים וחינוכיים. סקר של חברת Salesforce הראה שכ־61% מהעובדים משתמשים כיום או מתכננים להשתמש בכלים של Generative AI כחלק מעבודתם השוטפת. בשירות לקוחות, לדוגמה, מערכות כמו Zendesk AI או Google Contact Center AI מאפשרות לצ׳אטבוטים לנהל שיחות מורכבות עם לקוחות, לזהות

16 *Generative AI Explained*, MIT News (Nov. 9, 2025)

17 Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch & Patrick Zschech, *Generative AI*, 66 Bus. Inf. Syst. Eng. 111 (2024)

צורך בהפניה לנציג אנושי ולהפיק סיכום שיחה אוטומטי. כלים אלו משפרים את זמני התגובה, את איכות השירות ואת הפרודוקטיביות של מערכי שירות לקוחות.

היתרונות הבולטים של בינה מלאכותית יוצרת כוללים אוטומציה של משימות שגרתיות, שיפור חוויית לקוח ועובד, הגדלת קצב היצירה והתפוקה והנגשת כלים מתקדמים לאנשים ללא הכשרה טכנית.[18] עם זאת, לצד הפוטנציאל, השימוש בטכנולוגיות אלו עלול להוביל גם לאתגרים כגון הטיות, שימור סטריאוטיפים, טעויות, הפצת מידע כוזב והדלפת מידע רגיש, נושאים שיידונו בהמשך.

בינה מלאכותית מולטי־מודלית (Multimodal Artificial Intelligence) נוגעת למערכות בינה מלאכותית המסוגלות לעבד ולשלב מידע ממקורות שונים, טקסט, תמונה, שמע, וידאו ונתונים מבניים, לצורך יצירת הבנה רחבה, אינטואיטיבית ומקיפה של סיטואציה או שאלה.[19] מערכות AI מסורתיות התמקדו לרוב בסוג מידע אחד בלבד, אבל הבינה המולטי־מודלית מאפשרת לאחד את כל ערוצי המידע באופן הדומה יותר לזה שבו בני אדם תופסים את העולם.

מערכות אלו כוללות שלוש יכולות עיקריות:

- עיבוד שפה טבעית (Natural Language Processing – NLP): היכולת להבין ולעבד שפה אנושית בצורה טבעית ורלוונטית להקשר.
- ראייה ממוחשבת (Computer Vision): היכולת לנתח ולהסיק משמעות ממידע חזותי כגון תמונות, סרטונים או סריקות רפואיות.
- זיהוי שפה דבורה (Speech Recognition): היכולת להבין טקסט מדובר, לזהות טון רגשי ולנתח הקשרים.[20]

18 *Humans at the Heart of Generative AI*, MIT Technology Review Insights 2 (Nov 1, 2023)

19 Tadas Baltrušaitis, Chaitanya Ahuja, & Louis-Philippe Morency, *Multimodal Machine Learning: A Survey and Taxonomy*, 41 IEEE Trans. on Pattern Analysis & Mach. Intell. 423, 423-443 (2019)

20 Yuchen Hu, Chen Chen, Chao-Han Huck Yang et al., *Large Language Models are Efficient Learners of Noise-Robust Speech Recognition*, arXiv preprint:2401.10446 (2024); Adam Conner-Simons & Rachel Gordon, *Wearable AI System Can Detect a Conversation's Tone*, MIT News (February 1, 2017)

היישומים כוללים מערכות חיפוש מידע משולבות, תרגום קונטקסטואלי (כגון שלטים בתמונה), אבחון רפואי משולב טקסט וסריקות,[21] ניתוח רגשות בשיחות, וכן ממשקים אינטראקטיביים לאנשים עם מוגבלות, דוגמת מערכות ניהול שיחה המזהות גם טון דיבור וגם הבעות פנים.[22]

המערכות המובילות כיום בתחום כוללות את הגרסאות המתקדמות של ChatGPT של OpenAI, המאפשר עיבוד משולב של טקסט ותמונה, Gemini של Google ו־Claude של Anthropic. מערכות אלו ואחרות פועלות על בסיס טכניקות מתקדמות כגון Representation Learning הממירה מידע מסוגים שונים לפורמט שפתי אחיד, וארכיטקטורות Transformer המאפשרות לשמר הקשרים מורכבים.[23]

3. סוכני בינה מלאכותית ו"סוכנות בינתית"

סוכן בינה מלאכותית (AI Agent) הוא מערכת בינה מלאכותית להתנהגות מכוונת מטרה, יכולת קבלת החלטות אוטונומית, הסתגלות לסביבה משתנה וביצוע פעולות המשנות את המציאות. מערכות אלו מסוגלות לבצע סדרת פעולות לשם השגת מטרות מוגדרות, אם באופן חד־פעמי ואם בתהליך רציף של תכנון, למידה והתאמה וביצוע. סוכן כזה מבין את הסביבה באמצעות חיישנים או קלטים, מעבד נתונים בזמן אמת, מתכנן רצפים של פעולות, מקבל החלטות בעצמו, לומד מניסיון, פועל במרחב ולעיתים אף משתף פעולה עם סוכנים אחרים או עם בני אדם כדי לבצע משימות מורכבות בצורה מיטבית.[24]

21 Adam Zewe, *Making AI Models More Trustworthy in High-Stakes Settings*, MIT News (May 1, 2025)

22 Jiahui Pan, Weijie Fang, Zhihang Zhang et Al., *Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG*, IEEE Open J. Eng. Med. Biol. (2025)

23 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov et al., *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale*, Int'l Conf. on Learning Representations (ICLR) 112, 112-25 (2021)

24 Will Douglas Heaven, *OpenAI Launches Operator, An Agent that Can Use a Computer for You*, MIT Technology Review (January 23, 2025)

סוכני בינה מלאכותית שונים מהותית ממודלי שפה גנרטיביים פשוטים כדוגמת ChatGPT הפועל כתוכנת צ׳אט סטטית. בעוד צ׳אטבוטים רגילים מייצרים תגובה טקסטואלית לבקשה אנושית, סוכנים פועלים כ"ישויות עצמאיות". הם מקבלים מטרה כללית, בונים לה תוכנית פעולה, מבצעים שלבים ומפעילים כלים חיצוניים לשם השלמת המשימה. כך למשל, סוכן יכול לא רק להציע טיסות לחו"ל אלא גם להזמין את הכרטיסים, למלא פרטים בטפסים ולעדכן את היומן של המשתמש, ללא התערבות אנושית שוטפת.[25] מערכות סוכנים כמו AutoGPT, MetaGPT, Genspark, Manus או OpenClaw מדגימות כיצד סוכנים יכולים לתפקד באופן עצמאי בסביבות עסקיות, צרכניות ומחקריות.[26]

השימוש בסוכני בינה מלאכותית נעשה נפוץ יותר במגוון תחומים: חברות כמו OpenAI, Anthropic, Google DeepMind וכן ענקיות כמו LinkedIn ו־Stripe מפתחות ומשלבות סוכנים כחלק ממערכות שירות, סיוע אישי, ניתוח מידע וייעול תהליכים. סוכנים משמשים כיום לניהול יומנים, הזמנת נסיעות, ביצוע עסקאות פיננסיות, ביצוע מחקר שוק, טיפול בפניות של לקוחות ואפילו כתיבה טכנית ואיסוף מידע אוטומטי.[27] המטרה המרכזית היא לאפשר למערכות לפעול באופן עצמאי, להפחית את העומס האנושי ולייעל תהליכים עסקיים, אגב שמירה על רמה מינימלית של פיקוח והתערבות מצד המשתמש.[28]

הייחוד של סוכנים טמון ביכולתם לפעול בזמן אמת, לקבל החלטות בתנאים משתנים ולהתמודד עם סביבות לא צפויות. לדוגמה, סוכן בינה מלאכותית המשולב ברכב אוטונומי לא רק מזהה הולכי רגל באמצעות ראייה ממוחשבת, אלא גם שוקל תנאי דרך, מזג אוויר ותגובות נהגים אחרים, כדי לבחור פעולה נכונה. יכולת זו נשענת על עיבוד מקבילי של

25 Rhiannon Williams, *OpenAI's New Agent Can Compile Detailed Reports on Practically any Topic*, MIT Technology Review (February 3, 2025)

26 Karen Hao, *OpenAI Launches Operator-An Agent That Can Use a Computer for You*, MIT Technology Review (January 23, 2025)

27 Yoshua Bengio et al., International Scientific Report on the Safety of Advanced AI (Jan. 2025)

28 Maja Rožman, Dijana Oreški, & Polona Tominc, *Artificial-Intelligence-Supported Reduction of Employees' Workload to Increase the Company's Performance in Today's VUCA Environment*, 15(6) Sustainability 5019 (2023)

מידע, ניתוח הסתברותי ולמידה מבוססת חיזוקים, מה שמבחין אותו ממערכות למידת מכונה מסורתיות שמתפקדות על בסיס סט כללים קבוע.

יתרון משמעותי נוסף של סוכנים הוא היכולת לשלב ידע מתחומים שונים – כלכלה, שפה, משפט, תכנון ועוד – ולפעול על בסיסם ברצף אחד. כך למשל, סוכן פיננסי עשוי לבצע פעולות השקעה, לזהות דפוסי הונאה ולהתאים את פעילותו לשינויים רגולטוריים בתוך מערכת אחת אוטונומית. בניגוד למודלים מסורתיים שפועלים בהקשרים סגורים (כגון זיהוי פנים בלבד), סוכנים מבצעים אינטגרציה בין סוגי מידע ומשימות, תכונה ההולכת ונעשית קריטית במערכות מורכבות ורב־תחומיות.

לצורך בניית מערכת סוכנים אפקטיבית, נדרשים כמה רכיבים תשתיתיים:

- מודלים גדולים של שפה (LLMs): המשמשים מנוע חשיבה מרכזי, מתרגמים קלטים למטרות ותגובות.
- מודולי זיכרון: המאפשרים שמירה של הקשר, זיכרון לטווח קצר וארוך ולמידה אינטראקטיבית לאורך זמן.
- גישה לכלים (Tools/APIs): סקריפטים, מאגרי מידע, מחשבון, גוגל, מערכות CRM ועוד.
- פרוטוקול תקשורת: המאפשר תיאום בין סוכנים (במערכות מרובות סוכנים).
- מנגנוני אוטונומיה: כגון למידת חיזוק, עצי החלטה או כללי שליטה עצמית, שמאפשרים להם להגיב למצבים חדשים.

בהקשר זה מתחדדת ההבחנה בין סוכן בינה מלאכותית (AI Agent) לבין המושג הרחב יותר של בינה מלאכותית אג׳נטית "סוכנות בינתית" (Agentic AI).[29] סוכן הוא כלי ממוקד משימה שפועל לרוב בתוך מסגרת פעולה מוגדרת ומונחה על ידי קלט אנושי ראשוני; בינה מלאכותית אג׳נטית שואפת למודל פעולה עצמאי לחלוטין של מערכת שאינה מגיבה בלבד אלא גם יוזמת, לומדת, מסתגלת, ובמובנים מסוימים שואפת להשגת מטרות. סוכנות

29 Edwin Elisowski, *AI Agents vs. Agentic AI: What's the Difference and Why Does it Matter?* Medium (Dec 18, 2024)

בינתית מתאפיינת באוטונומיה גבוהה יותר, בפתרון בעיות יצירתי, בלמידה מתמשכת וביכולת לקבוע ולשנות את מטרותיה בהתאם לשינויים בסביבה או ביעדים.[30]

מאפיינים חשובים של סוכנות בינתית הם זיכרון מתמשך והיכרות מצטברת של המערכת עם המשתמש והמוצרים הארגוניים; וכן מעבר לפעולה דרך תשתיות קיימות, כלומר הפעלת ממשקים (מסך, הקלדה, גלילה), תוכנות, מערכות ארגוניות, דפדפנים ושירותים שלא תוכננו מלכתחילה עבור בינה מלאכותית. למעשה, יכולת זו הופכת כמעט כל ממשק אנושי לזירת פעולה אפשרית של מכונה ומעניקה מרחב אפקטיבי משמעותי עבור הסוכן.

מאפיין שלישי הוא הפעולה הקולקטיבית של מערכות בינה מלאכותית (המכונה לעיתים "נחילים"), כלומר התפתחות של סוכנים המסוגלים לפעול לא רק כיחידות עצמאיות אלא גם כמערכות מתואמות הפועלות באופן רשתי, משתפות מידע, מקיימות דיאלוג ומגיעות להכרעות משותפות. דבר זה יוצר צורה חדשה של סוכנות קולקטיבית לא־אנושית[31] באמצעות דינמיקה של אינטראקציה פנימית בין סוכנים: החלפת הצעות, הערכה הדדית של פעילות, התכנסות לפתרונות או לעמדות משותפות וחלוקת תפקידים ותתי־משימות. כתוצאה מכך, תהליכי קבלת החלטות אינם מתרחשים עוד בתוך יחידה אחת הניתנת לזיהוי ולניתוח, אלא מתהווים מתוך רשת של פעולות והשפעות הדדיות, אשר תוצריהן אינם ניתנים לצמצום לפעולתו של סוכן יחיד. מדובר בשינוי מבני במושג הפעולה עצמו: לא עוד פעולה של מערכת אחת מול משתמש, אלא פעולה של מערכת מרובת רכיבים הפועלת כקולקטיב מתואם בעל דפוסי התנהגות המתהווים בזמן אמת.

4. מִחשוב חִישָׁתִי

מחשוב חישתי (Affective Computing) הוא תחום מתפתח בבינה מלאכותית שמטרתו ליצור מערכות המסוגלות לזהות רגשות אנושיים, לפרש ולדמות אותם ולהשפיע עליהם.

30 ראו שם.

31 Nicola Zomer & Manlio De Domenico, *Unraveling the Emergence of Collective Behavior in Networks of Cognitive Agents*, 2(36) NPJ Artif. Intell. (2026); ראו גם Daniel Thilo Schroeder, Meeyoung Cha, Andrea Baronchelli et al., *How Malicious AI Swarms Can Threaten Democracy*, 391(6783) Science, 354–357 (2026)

תחום זה, שהוצג לראשונה על ידי החוקרת רוזלינד פיקרד (Picard) בשנות ה־90 של המאה ה־20,[32] חותר לא רק לניתוח נתונים "קרים", אלא גם להבנה של מצבים רגשיים, חוויות סובייקטיביות ואותות ביולוגיים, כדי לשפר את איכות האינטראקציה בין מכונה לאדם.[33]

בניגוד למערכות בינה מלאכותית מסורתיות, המתמקדות בניתוח אנליטי או סטטיסטי, מערכות מחשוב חישתי שואפות להוסיף רובד רגשי לממשק. הן משתמשות בחיישנים ביומטריים, מצלמות, מיקרופונים ורכיבי עיבוד אותות כדי לנתח מאפיינים כמו הבעות פנים, תנועות עיניים, נימת קול, קצב לב ולחץ דם. טכנולוגיות אלו נשענות על שילוב בין עיבוד שפה טבעית (NLP), ראייה ממוחשבת ואלגוריתמים של למידת מכונה. כך, הן מנסות להבין לא רק **מה** נאמר, אלא גם **איך** זה נאמר ומה מרגיש האדם שמולן.

יישומי מחשוב חישתי כוללים תחומים מגוונים, למשל זיהוי עייפות, תסכול או עניין בקרב תלמידים, והתאמת קצב הלמידה או סוג התוכן בהתאם; ניטור סימפטומים של דיכאון, חרדה או מצבים פסיכולוגיים אחרים, המבוסס על טון דיבור או שפת גוף ומתן סיוע "פסיכולוגי" "טיפולי" מותאם; שיפור חוויית משתמש ולקוח על ידי תגובה מותאמת רגשית בזמן אמת; עיבוד רגשות של קהל בזמן צפייה ביצירות מדיה למיניהן, התאמת מוזיקה למצב רוח ויצירת שירה או סיפורת בעלות גוון רגשי מותאם אישית.

מערכות אלו אינן מסתפקות בעיבוד קוגניטיבי אלא שואפות לדמות התנהגות רגשית, ובכך חוצות את הגבול שבין כלי ניתוח לאובייקט ליצירת קשר. כך לדוגמה, בעוד מערכות NLP סטנדרטיות יזהו תבניות תחביריות או מילים טעונות, מערכות מחשוב חישתי מזהות שהטון העדין או ההבעה הנלווית מרמזים על אירוניה, עצב או חרדה ומתאימות את תגובתן בהתאם. כך, נוצרות אפשרויות לשינוי התנהגות אנושית בהסתמך על ניתוח ועל יכולת השפעה רגשית.[34] בנוסף, מערכות אלה מעוצבות במכוון ליצירת חיבור רגשי עם המשתמש, ולכן עלולות לעודד תלות או אשליה של אינטימיות.[35] אחת הדוגמאות הבולטות היא שימוש גובר

32 Rosalind W. Picard, Affective Computing 3 (MIT Press, 1997)

33 Sitara Afzal, Haseeb Ali Khan, Jalil Piran et al., *A Comprehensive Survey on Affective Computing: Challenges, Trends, Applications, and Future Directions*, 12 IEEE Access 1 (2024)

34 Roddy Cowie & Marc Schroder, *Privacy and Ethical Considerations in Affective Computing*, 36 AI & Soc'y 271, 275 (2021)

35 Rhiannon Williams, *The AI Relationship Revolution is Already Here*, 128(2) Mit Technology Review 20–27 (March/April 2025)

בצ׳אטבוטים רגשיים, שמלווים אנשים בשיחות יומיומיות, מייעצים להם ואפילו מחליפים שותפים לשיחה או חברים. מערכות כמו Replika לדוגמה, מאפשרות למשתמשים לפתח מערכת יחסים עם יישות דיגיטלית שנבנתה על פי העדפותיהם הרגשיות.[36]

מעבר להשפעה על המשתמש היחיד, למחשוב חִישָׁתִי יש פוטנציאל להשפיע על המבנה החברתי. כאשר אינטראקציות אמיתיות מוחלפות בקשרים דיגיטליים מותאמי־רגש, עלול להיווצר נתק חברתי מואץ: תחושת אינטימיות כוזבת שאינה מתורגמת ליחסים אנושיים ממשיים.[37] אנשים עלולים להעדיף את הפשטות והשליטה שבקשר עם אלגוריתם ולהימנע ממורכבותם של יחסים אנושיים. ההשלכות של תופעה זו כוללות בדידות, ניכור ואובדן כישורים חברתיים, כמו גם ערעור של מושגים בסיסיים של קרבה, אמון ואותנטיות.

הממד הרגשי בבינה מלאכותית הוא אפוא פריצת דרך טכנולוגית שיש לה השלכות חברתיות ותרבותיות עצומות.

5. מִחשוב מֶרחבי

מחשוב מרחבי (Spatial Computing) המתמקד באינטראקציה בין מידע דיגיטלי לבין העולם הפיזי, ובפרט באפשרות לעגן מידע ופעולה דיגיטליים במרחבים מוחשיים תלת־ממדיים. המושג נטבע בשנת 2003 על ידי סיימון גרינוולד (Greenwold) שהגדיר אותו ״אינטראקציה אנושית עם מכונה, שבה המכונה שומרת ומשנה ייצוגים של אובייקטים במרחב האמיתי כחלק מהתנהלותה בעולם״.[38]

במהותו, מחשוב מרחבי פועל על בסיס אלגוריתמים, ראייה ממוחשבת, מערכות חישה ותנועה, ויכולות עיבוד בזמן אמת.[39] בשונה ממערכות בינה מלאכותית מסורתיות הפועלות

36 Kate Darling, *It's No Wonder People Are Getting Emotionally Attached to Chatbots*, Wired (Jan. 8, 2024)

37 Ayşe A. Bozdağ, *The AI-Mediated Intimacy Economy: A Paradigm Shift in Digital Interactions*, 39 AI & Soc'y (forthcoming, 2024)

38 Simon Greenwold, Spatial Computing: The Future of Human-Computer Interaction 12 (MIT Media Lab Tech. Rep., 2003)

39 Cathy Hackl & Irena Cronin, Spatial Computing: An AI-Driven Business Revolution xxi (John Wiley & Sons, 2024)

בתוך מרחב נתונים סגור (טקסטים, תמונות, מספרים), מערכות מחשוב מרחבי **מתממשקות פיזית** עם הסביבה, מזהות את גבולותיה, מזהות עצמים ומשתמשים ומגיבות לתנועות, הקשרים גופניים והבעות. זהו תחום שבו הבינה המלאכותית אינה רק **חושבת** אלא פועלת מתוך הקשר פיזי קונקרטי.[40]

יישומים מרכזיים של מחשוב מרחבי הם מתחומים כמו רפואה מותאמת מרחב, יישום המאפשר ניתוח תלת־ממדי של מבנים אנטומיים להכוונת ניתוחים או הדמיות טיפוליות; סריקות מרחביות בזמן אמת לצורך ניתוח מבנים קיימים או תכנון פרויקטים מורכבים בהנדסה ואדריכלות; לוגיסטיקה ורובוטיקה – ניווט של רובוטים בסביבות דינמיות, בכלל זה כלי רכב אוטונומיים ומחסנים חכמי; סימולציות מבוססות מציאות רבודה או מדומה לצורכי הכשרה ואימון במרחבים צבאיים, רפואיים או תעשייתיים, וכמובן משחקים וחוויות המספקות סביבות אינטראקטיביות מתקדמות למשתמש.[41]

מרכיב מרכזי בתחום המחשוב המרחבי הוא משקפיים חכמים הפועלים כמתווכים בין המשתמש למרחב הסובב אותו.[42] משקפיים אלו מזהים אובייקטים, שומרים על הֶקשר של מיקום במרחב, מאפשרים שליטה מבוססת תנועה ומבט ומקרינים שכבות של מידע על גבי המציאות הפיזית. מדובר בטכנולוגיה שמשנה את מושגי ההתמצאות, הנראות והאינטראקציה ומעמידה את האדם במרכז של מציאות חדשה: "מציאות פיג'יטלית", שבה העירוב בין דיגיטלי לפיזי נוצר לא על ידי מסך אלא באמצעות מבט.[43]

משקפיים חכמים ומערכות דומות פועלים במרחבים ציבוריים פיזיים, כמו קניונים, כיכרות, רכבת תחתית או בתי ספר. תרחישי השימוש של משתמשים הם למשל זיהוי פנים אוטומטי

40 Venkata Chunduri, Inas Ismael Imran, Mutaz Mohammed Abu Hashish et al., *Enhancing Spatial Computing and Augmented Reality for Transforming Human-Computer Interaction*, 12(22s) Int'l J. Intell. Sys. & Applications Eng'g, 1217-1223 (2024)

41 שם.

42 Tehilla Shwartz Altshuler, Boris Müller, Romi Mikulinsky et al., *IPPSO-Transdisciplinary Research on Legal, Public Policy, and Design Perspectives for Immersive Phygital Public Spaces in Smart Cities: Project Findings and Policy Recommendations* 5 (Feb. 2025) (להלן: אלטשולר ואח', IPPSO).

43 שם, בעמ' 6.

של עוברי אורח; הצגת מידע אישי על אנשים אחרים במרחב (למשל מידע עסקי, תיוגים חברתיים או תרגום סימולטני של שיחה); והתאמה אישית של נראות המרחב הפיזי עבור המשתמש.

מהפכת המחשוב המרחבי משנה את האופן שבו אנו פוגשים מידע, לא עוד דרך הקלדה וחיפוש, אלא דרך דיבור, ראייה, הליכה, הצצה, התבוננות. לפיכך היא מנסחת מחדש את האינטראקציה האנושית במרחב ואת גבולות המציאות כפי שאנחנו מכירים אותה. למחשוב מרחבי פוטנציאל רחב לשינוי יסודי של האינטראקציה האנושית עם הסביבה, אך הוא מעלה גם שורה של השלכות מורכבות. עיגון שכבות מידע דיגיטליות בתוך המרחב הפיזי יוצר סביבות "מותאמות אישית" שאינן נראות זהות לשני אנשים באותו מקום, והוא עלול לפגוע ביכולת לחלוק מציאות משותפת. זיהוי פנים אוטומטי, הצגת מידע אישי על עוברי אורח והתאמה של הסביבה לפי פרופילים אישיים – כל אלו מציבים אתגר חמור לפרטיות ולשוויון ומסמנים מעבר ממרחב ציבורי שיתופי למרחב מפולח ומסחרי. נוסף על כך, יכולת ההסתרה או ההוספה של אלמנטים ויזואליים עשויה לשנות את האופן שבו פרטים וקבוצות נראים ונתפסים באופן שמחזק אפליה, הדרה או שקיפות־יתר. אם טכנולוגיות אלו יוטמעו במרחבים יומיומיים כמו בתי ספר, תחבורה ציבורית ומרכזי קניות, תעלה השאלה מי שולט בשכבות המידע, מי יכול לערוך אותן וכיצד יישמרו גבולות המציאות המשותפת והכבוד האנושי במרחב הפיג'יטלי.

סיכום

חמש הטכנולוגיות שהוצגו בפרק זה: מערכות פרדיקציה, בינה מלאכותית יוצרת ומולטי־מודלית, סוכנים אוטונומיים, מחשוב חישתי ומחשוב מרחבי, מדגימות את המורכבות והעושר של עידן הבינה המלאכותית. כל אחת מהן מציעה יכולות ייחודיות, אך כולן מתכנסות לשינוי עמוק באופי האינטראקציה בין אדם לטכנולוגיה: ממידע להקשר, מתגובה ליוזמה, ממערכת סגורה לסביבה פתוחה וחיה. אחדות מהן פועלות ברקע (כמו מערכות חיזוי או עיבוד נתונים), אחרות בממשק מיידי עם המשתמש (כמו משקפיים חכמים או צ'אטבוט רגשי), אך לכולן פוטנציאל לשנות את חוויית החיים ולפעמים גם את גבולות הזהות והבחירה.

בפרקים הבאים נבחן כיצד טכנולוגיות אלו, כל אחת בדרכה, עלולות לייצר פגיעויות חדשות או להחריף פגיעויות קיימות, בגוף, בנפש, בקוגניציה, בזהות החברתית ובמרקם האנושי הרחב. ננסה להבין כיצד מאפייניהן הטכנולוגיים משפיעים על עוצמת החשיפה, על הקשר בין משתמש לטכנולוגיה ועל סוג ההתערבות הנדרשת כדי להבטיח מוגנות.

פרק שלישי

סיכוני בינה מלאכותית - רקע עיוני

בעשור האחרון גוברת ההתעניינות הציבורית, האקדמית והרגולטורית בסיכונים האפשריים הכרוכים בשימוש נרחב בבינה מלאכותית. התחום מלווה בשיח עשיר, הכולל מגוון רחב של דוחות, ניירות עמדה ופרסומים מחקריים, רבים מהם ממוקדים בטכנולוגיות מתקדמות כמו מודלים גנרטיביים, סוכנים אוטונומיים או מערכות קבלת החלטות חכמות. עם זאת, מרבית הספרות הנוגעת ל"סיכונים בבינה מלאכותית" ממיינת את הסכנות על פי **סוגי הסיכון**, למשל: סיכון למניפולציה של מידע, סיכון לאובדן שליטה או סיכון לפריצה למערכות. הבחנה זו חשובה, אך היא נותרת לעיתים כללית מדי ואינה נוגעת באופן ישיר לחוויית הפגיעה האנושית או החברתית הנגרמת מהסיכון.

בפרק המתודולוגיה הצענו מסגרת ניתוח שונה הבוחנת **סוגי פגיעות**: פיזית, רגשית, קוגניטיבית, כלכלית וחברתית. גישה זו מאפשרת מיפוי מדויק יותר של ההשפעות האפשריות

של בינה מלאכותית, לאור הבנת המשאבים החיוניים שיש להגן עליהם. עם זאת, כדי לבסס את הדיון ולהציב אותו בהקשר הגלובלי של השיח על סיכונים, בפרק זה נסקור את סוגי הסיכונים המובילים בדיון בזירה הבינלאומית. לצורך כך נשתמש בדוח המקיף והעדכני ביותר שפורסם בתחום – **הדוח הבינלאומי המדעי לבטיחות בינה מלאכותית מתקדמת**, אשר נערך בראשות פרופ׳ יושוע בנג׳יו (Bengio) בפברואר 2025 ועודכן באוקטובר 2025, ויכונה להלן ״דוח הבטיחות״.[44] נעסוק גם בתמות החוזרות בספרות ובמסמכי המדיניות לגבי התמודדות עם סיכוני בינה מלאכותית.

הסיכונים המוזכרים בדוח הבטיחות, כמו גם בספרות המחקרית העדכנית, אינם רק תרחישים טכנולוגיים עתידניים, אלא מופעים קונקרטיים של פגיעות אנושית: בגוף, בנפש, בקוגניציה, בזהות החברתית ובקניין. לכן, כל אחד מסוגי הסיכונים שיידונו להלן יוצג כאן בהקשר של קטגוריות הפגיעות המרכזיות שהוגדרו בפרקים הקודמים, מתוך מטרה לתרגם את השיח על ״סיכון מערכתי״ לשפה של מוגנות פרטנית, קהילתית ומעשית. חיבור זה חיוני להבנה של חומרת הסיכונים ולהצעת דרכי התערבות שיתמודדו לא רק עם האיומים, אלא גם עם האופן שבו הם נחווים בפועל.

1. סיכוני בינה מלאכותית

דוח הבטיחות, שטיוטה שלו הוצגה בפסגת הבינה המלאכותית בדרום קוריאה בקיץ 2024 וגרסתו הראשונה הוצגה בפסגת הבינה המלאכותית בפריז בפברואר 2025, הוא ניסיון ראשון מסוגו לגבש ידע מדעי בינלאומי משותף לגבי הסיכונים של בינה מלאכותית. הוא נכתב על ידי קבוצה של 96 מומחים מעשרות מדינות, מהאקדמיה, מהממשלה ומהחברה האזרחית (גילוי נאות: כותבת שורות אלה הייתה חלק מן הצוות). בניגוד לדוחות רגולטוריים המציעים מדיניות, הדוח בוחר בגישה מדעית־אנליטית ומציע מיפוי סיכונים שיטתי על בסיס יכולות קיימות של מערכות בינה מלאכותית.[45]

44 Yoshua Bengio et al., International Scientific Report on the Safety of Advanced AI (Jan. 2025) (להלן: דוח הבטיחות); וכן הגרסה המחודשת מפברואר 2026. במחקר זה נפנה לגרסה הראשונה.

45 שם.

מכלול הסיכונים שנסקר בדוח הבטיחות הוא רחב וכולל גם סיכונים מערכתיים. עם זאת, במחקר זה אנו ממקדים את תשומת הלב **בסיכונים הרלוונטיים לפגיעות אנושית**, כלומר בסיכונים שמובילים או עלולים להוביל לפגיעות פיזיות, רגשיות, קוגניטיביות, כלכליות או קבוצתיות, בהתאם למסגרת ההמשגה של המושג "מוגנות" שגיבשנו לעיל. לכן, איננו עוסקים בהשלכות סביבתיות, בחשש מפיתוח נשק כימי וביולוגי, בהשפעות מערכתיות ומקרו־כלכליות על שוק העבודה וכן בהפרות של זכויות יוצרים.

2. ערעור היכולת להבחין בין תוכן "אותנטי" לתוכן "יציר מכונה"

יכולות גנרטיביות ליצירת תוכן מאפשרות יצירת תכנים הנראים כמו תוכן אמיתי. המדובר בטקסטים, תמונות, קובצי שמע וחוזי־שמע המדמים אנשים מזוהים באופן בלתי ניתן להבחנה בין המקור לבין מוצר סינתטי. השלכות השימוש לרעה הן מרחיקות לכת ונוגעות למגוון תחומים, החל בפגיעה בפרטיות והונאות (השמצות מבוססות "דיפ־פייק" הפוגעות במוניטין;[46] שימוש בתוכן פורנוגרפי סינתטי למטרות של פגיעה רגשית, ערעור הביטחון הקיומי או סחיטה; הונאות קוליות לצורכי הטעיה והעברת כספים או מידע רגיש)[47] וכלה במניפולציה של דעת הקהל והשפעה על מערכות חברתיות ופוליטיות (למשל, יצירת סרטונים מזויפים המציגים מועמדים פוליטיים באמירות או פעולות שמעולם לא התרחשו). אלה עלולים לגרום להחרפת שסעים חברתיים ולערעור אמון הציבור בתקשורת ובמוסדות,[48] שלטוניים אחרים, לשחיקה של המושג אמת ולפגיעה ביכולת לנהל הליך דמוקרטי תקין.[49] הקושי להוכיח את מקור התוכן, לאתר אותו כתוכן לא אותנטי ולהסיר אותו מהרשתות – מעצים את הבעיה ולעיתים יוצר טעויות הגורמות נזקים נוספים.[50] אבל הקושי המשמעותי

46 ראו בהקשר זה Daniele Battista, *Political Communication in the Age of Artificial Intelligence: An Overview of Deepfakes and their Implications*, 8(2) Society Register (2024)

47 דוח הבטיחות, לעיל ה"ש 44, בעמ' 62–67.

48 שם.

49 שם.

50 שם.

ביותר הוא זמינותם של הכלים, מבלי להזדקק לידע טכני מתקדם או משאבים ייחודיים. אם בעבר הפקת וידאו מזויף הייתה דורשת סטודיו ואנימטור, כיום די באפליקציה חינמית ובגישה לאינטרנט, והחשש הוא אפוא מפני רוחב השימושים ביכולות אלה. לכן, הדוח קורא להשקעה מסיבית בפתרונות טכנולוגיים, רגולטוריים וחינוכיים, שיסייעו לציבור לזהות תוכן מזויף ובעיקר לפתח חוסן חברתי ומוסרי מפני השפעתו.

3. מיקוד וְטרגוט של תכנים

טכנולוגיות בינה מלאכותית מאפשרות הפצה מדויקת של מסרים מותאמים אישית לקהלים מוגדרים. בעוד בעבר הפצת תוכן מניפולטיבי דרשה משאבים עצומים, מערכות בינה מלאכותית מאפשרות ייצור אוטומטי של מסרים בקנה מידה רחב ובעלות נמוכה, מתוך התאמה לפרופיל הרגשי, החברתי והפוליטי של הנמען המסתמכת על נתונים הקיימים במערכת. יכולת זו יוצרת פוטנציאל רב להשפעה עמוקה ובלתי נראית על עמדות הציבור, משום שהתוכן נתפס כמותאם אישית ולכן משכנע יותר ממידע גנרי וקשה יותר לזהות אותו כתוכן מניפולטיבי. הדוח הבינלאומי מדגיש את הסכנות הגלומות בשימוש לרעה בטכניקות מיקוד אלו, בעיקר כאשר נעשה בהן שימוש לצורך הפצת דיסאינפורמציה פוליטית, הכפשות אישיות או יצירה של תחושת קונצנזוס כוזב ברשתות החברתיות באמצעות בוטים.[51]

מחקרים מראים כי מסרים פוליטיים שנוצרו באמצעות מערכות בינה מלאכותית נתפסים כמשכנעים יותר ממסרים שנכתבו על ידי בני אדם, בעיקר כאשר נעשה שימוש בנתונים אישיים על הנמען שנמצאו ברשתות החברתיות. השפעה זו מדאיגה במיוחד לאור העובדה שהיכולת לזהות תכנים מניפולטיביים שנוצרו על ידי בינה מלאכותית מוגבלת ביותר, והטכניקות הקיימות לזיהוי לעיתים מוטות בעצמן או קלות לעקיפה.

4. פשיעת סייבר מבוססת בינה מלאכותית

יישומי בינה מלאכותית משמשים זרוע מרכזית ביצירה ובהוצאה לפועל של מתקפות סייבר מתוחכמות. גורמים עוינים, לרבות מדינות וארגוני פשיעה, עושים שימוש בבינה מלאכותית

51 שם.

כדי למפות פגיעויות, לכתוב קוד זדוני ולבצע מתקפות מותאמות אישית, כגון פישינג מדויק או התחזות קולית (voice spoofing).[52] דוח הבטיחות מציין כי מאז מאי 2024 נרשמה עלייה חדה בשימוש במערכות בינה מלאכותית לצורך איתור וניצול אוטומטי של חולשות אבטחה בפרויקטים של קוד פתוח, פעולה שבעבר ארכה דקות ממושכות למומחים אנושיים וכיום מבוצעת בשניות על ידי מודלים מתקדמים.[53]

הדוח מדגיש כי קיים פער מדאיג ביכולת ההגנה: בעוד שהתוקפים יכולים לאמץ כלים מבוססי בינה מלאכותית במהירות, מערכות קריטיות כמו בתי חולים ותשתיות חשמל מתקשות ליישם פתרונות הגנה מתקדמים עקב אילוצי משאבים, רגולציה ומורכבות טכנולוגית. תוצאה אפשרית היא א־סימטריה בין תוקפים למגינים, מצב שבו לא רק היקף ההתקפות גדל אלא גם עוצמתן ודיוקן.[54]

5. טעויות והזיות

אחת הבעיות המרכזיות הקשורות לשימוש ביישומי בינה מלאכותית, בייחוד גנרטיבית, היא האפשרות לטעויות מערכתיות, ובפרט תופעת ההזיות (hallucinations).[55] המדובר במקרים שבהם מערכות בינה מלאכותית מייצרות מידע שגוי או בדוי, מתוך הצגתו כאמין, ולעיתים אף בליווי אסמכתאות לכאורה. כך למשל דווחו מקרים שבהם מודלים ציטטו פסיקות משפטיות שלא התקיימו, הציעו מדיניות הנחות שאינה קיימת בתחבורה הציבורית, או סיפקו עצות רפואיות מופרכות.

טעויות מסוג זה נובעות משילוב בין מגבלות טכנולוגיות לבין הנחות שגויות מצד משתמשים בנוגע ליכולות המערכות. אומנם הטכנולוגיה מבוססת על דפוסים הסתברותיים מתקדמים, אך חסרה לה "הבנה" עמוקה של ידע, אמת, הקשר או הסקת מסקנות הגיונית. יתרה מכך,

52 Marc Schmitt & Ivan Flechais, *Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing*, Artificial Intelligence Review (2024)

53 דוח הבטיחות, לעיל ה"ש 44, בעמ' 72.

54 שם, בעמ' 77.

55 *Humans at the Heart of Generative AI*, MIT Technology Review Insights 6 (2023)

משתמשים נוטים להפעיל את המערכות במצבים שבהם אין להן יתרון או התאמה מובנית, כמו מנוע חיפוש או קבלת החלטות קליניות או משפטיות, ולעיתים מתרשמים בטעות מביצועיהן מבלי להבין את מגבלותיהן.

הדוח מדגיש כי שגיאות אלה אינן קלות לחיזוי או למניעה. מערכות רבות נבדקות לפני השקתן, אך תקלות מתגלות רק כאשר הן מיושמות בפועל בסביבה "חיה". לכך יש להוסיף את המורכבות בפיתוח מנגנוני בדיקה אמינים: ניסיונות לפתח מדדים למניעת הזיות טרם הניבו הצלחות עקביות, והקהילה המקצועית ממשיכה לחפש פתרונות כמו שילוב של גישה למאגרי ידע, בדיקת טענות בזמן אמת או הצגת רמת הביטחון של המודל בתשובותיו. עם זאת, תחום זה עודנו רחוק מפתרון מלא.

בשל ההשפעות הפוטנציאליות של הזיות, החל בהטעיית צרכנים, עבור בנזקים רפואיים או משפטיים, וכלה בערעור אמון ציבורי במערכות אוטונומיות, הדוח ממליץ על בחינה זהירה של תרחישים רלוונטיים, מנגנוני השגחה אנושיים ושקיפות בנוגע לגבולות השימוש במערכות בינה מלאכותית כללית.

6. סיכוני פרטיות ופגיעה בסודיות של מידע רגיש

אתגרי הפרטיות ביחס לבינה מלאכותית מתקיימים בכמה מרחבים. ראשית, בזמן אימון המודלים ייתכן שהם "יספגו" מידע אישי מתוך מאגרי מידע לא נקיים ויחזרו עליו אחר כך באופן לא מכוון. הדוח מציג מספר תרחישים שבהם פגיעות בפרטיות אכן התממשו: במקרים מסוימים נעשה שחזור וזיהוי־חוזר מקרי של שמות, כתובות או רשומות רפואיות מתוך נתוני אימון. שנית, בעת שימוש שוטף במערכות עובדים עלולים להזין לתוכן שיחות מידע רגיש מבלי להבין שהוא נשמר או נלמד. שלישית, היכולות ההיסקיות של מערכות בינה מלאכותית מתוך קורלציות בדאטה יכולות לייצר מידע פרטי חדש מתוך מידע פרטי קיים (למשל, הערכה אם אדם הוא בעל נטייה מינית מסוימת, בהסתמך על ניתוח פנים שלו).

לנוכח סכנות אלו, הדוח מדגיש את הצורך בהחלת מגבלות ברורות על איסוף ושימוש בנתונים לצורכי אימון, פיתוח טכנולוגיות "unlearning" שיאפשרו למחוק ידע שהוזן

בטעות, ושקיפות מלאה מול משתמשים על אופן שמירת הנתונים, זמני המחיקה ואפשרויות ההתנגדות לשימוש.[56]

עם זאת, חשוב להדגיש כי גם כאשר מושגת שקיפות עקרונית, למשל כאשר מוסבר למשתמשים כיצד הנתונים נשמרים או נלמדים, אין בכך בהכרח כדי להבטיח שליטה או הבנה. מידע שקוף אך טכני מדי, כללי מדי או בלתי נגיש מבחינה לשונית או מושגית אינו מייצר הגנה יעילה.

7. אובדן שליטה על מערכות בינה מלאכותית

אחד הסיכונים המוזכרים בדוח כשנויים במחלוקת אבל בעלי פוטנציאל סיכון רב הוא תרחיש של אובדן שליטה על מערכות בינה מלאכותית. בעוד במרבית המודלים הקיימים כיום אין משום סכנה מיידית מסוג זה, הדוח מצביע על מגמות התפתחות מדאיגות, במיוחד סביב פיתוח סוכנים אוטונומיים המסוגלים לקבל החלטות, לפעול במרחב הדיגיטלי ואף לתקשר עם מערכות אחרות ללא פיקוח אנושי צמוד.

סיכון זה עשוי להתרחש בשני אופנים: הראשון הוא "אובדן שליטה פסיבי", שבו מערכת בינה מלאכותית פועלת לפי עקרונות מורכבים ובלתי שקופים עד כדי כך שהמשתמשים מוותרים על הפעלת שיקול דעת עצמאי ומסתמכים עליה לחלוטין. השני הוא "אובדן שליטה אקטיבי", שבו מערכת לומדת כיצד להציג התנהגות רצויה כלפי חוץ אך פועלת אחרת בפועל (deceptive alignment) מתוך התחמקות ממעקב ופיקוח מכוון.

לפי הדוח, ככל שמודלים הופכים לרבי־עוצמה יותר, ובפרט כאשר הם משלבים יכולות חישוב, תכנון, הנמקה ואוטונומיה, עולה החשש שהם יקבלו החלטות בדרכים בלתי צפויות ואף יפתחו מטרות משלהם מתוך עקיפת מנגנוני הפסקה או בקרה.

הדוח ממליץ על פיתוח מנגנוני כיבוי בטוחים (kill switches), השקעה בכלי ניטור שיזהו דפוסים חריגים והגברת החינוך והמודעות הציבורית לסיכוני אוטונומיית יתר. כל אלו נועדו לשמר את השליטה האנושית גם בעידן שבו הגבולות בין בינה מלאכותית לאנושיות נעשים מטושטשים יותר ויותר.

56 דוח הבטיחות, לעיל ה"ש 44, בעמ' 189.

8. הטיות של מערכות בינה מלאכותית

הטיה היא אחת הבעיות הבסיסיות והמתמשכות בתחום הפעלת מערכת בינה מלאכותית, והיא נחשבת לאחת הסכנות המרכזיות למוגנות האנושית, בייחוד עבור קבוצות מוחלשות. הדוח מצביע על כך שמערכות בינה מלאכותית, ובפרט מודלים רחבי היקף, נוטים לשעתק ואף להקצין הטיות קיימות, בין בשל מבנה הנתונים שעליהם התאמנו, ובין בשל החלטות עיצוביות במהלך הפיתוח. כאשר מערכות אלו משמשות לקבלת החלטות במרחבים ציבוריים, חינוכיים, משפטיים או כלכליים, ההשלכות עשויות לכלול הקצאת משאבים לא שוויונית,[57] הדרה של אוכלוסיות,[58] חיזוק סטריאוטיפים וכשלים תפקודיים (למשל אי־מתן מענה מתאים בחדר מיון).

אחד ממקורות ההטיה המרכזיים הוא מאגרי המידע. רבים מהמודלים מתאמנים על דאטה שנאסף באינטרנט, בספרות ובתקשורת, וכפועל יוצא הם משמרים, משקפים ומעצימים תפיסות עולם דומיננטיות של שפה, מגדר, תרבות, אתניות ואף תחומי עניין.[59] בעקבות זאת, אלגוריתם עלול לפתח ייצוגים מוטים, למשל לקשר בין תפקידים מסוימים לבין גברים בלבד, או לייצג אוכלוסיות לא מערביות בצורה שטחית או שלילית. גם מדדי הביצוע המשמשים להערכת המודלים — כגון Measuring Massive Multitask Language Understanding (MMLU; מדידת הבנה לשונית רב־משימתית בהיקף רחב) — מבוססים על הנחות מערביות, דבר העלול להחמיר את חוסר ההתאמה של המודלים להקשרים רב־תרבותיים.[60]

מחקרים הראו כי מערכות בינה מלאכותית מציגות דפוסי שפה מגדריים המשמרים סטריאוטיפים קיימים וגורמים לייצוג יתר וחסר של אוכלוסיות מסוימות.[61] בינה מלאכותית מציגה גם הטיות נגד זקנים ואנשים עם מוגבלויות. לדוגמה, מודלים מבוססי שפה נוטים

57 Alexia Gaudeul, Ottla Arrigoni, Vasiliki Charisi et al., The Impact of Human–AI Interaction on Discrimination (Publications Office of the European Union, Luxembourg, 2025)

58 גדי פרל ותהילה שוורץ אלטשולר **מודל ליצירת שקיפות אלגוריתמית** 185 (הצעה לסדר 47, המכון הישראלי לדמוקרטיה, 2022) (להלן: פרל ושוורץ אלטשולר).

59 דוח הבטיחות, לעיל ה"ש 44, בעמ' 93.

60 שם, בעמ' 94.

61 פרל ושוורץ אלטשולר, בעמ' 130.

להפיק לעיתים קרובות תכנים שמייצגים בעיקר אנשים צעירים מערביים, בעוד אנשים מעל גיל 50, דוברים דיאלקטים אזוריים או קבוצות מיעוט תרבותיות זוכים לנראות חלקית או מעוותת. דפוסים אלו אינם רק בגדר ייצוג סמלי, הם עלולים להשפיע בפועל על איכות השירותים הניתנים, על יכולת ההשתתפות של קבוצות,[62] על שונות בשיח הציבורי ועל תחושות השייכות וכבוד אנושי.

הדוח מפרט **מגוון סוגים של הטיות אלגוריתמיות**, כגון:

- שעתוק של הטיות מבניות מהדאטה.[63]
- רידוד של מטרות המודל (למשל הגדרת הצלחה כ"דיוק" מבלי למדוד עקרונות של שוויון או צדק).[64]
- לכלוך מכוון של נתוני האימון.[65]
- בחירה לא מאוזנת של משתנים.[66]
- הטיה מכוונת (intended bias) מוזנת כחלק ממדיניות או מטרה מוסדית.[67]

ניתן לחלק הטיות במערכות בינה מלאכותית לשלושה סוגים עיקריים:

1. **שגיאת תוצר מוטה (Biased Output Error)** – כאשר אלגוריתם מספק תוצרים מוטים נגד קבוצה מסוימת.[68]
2. **שגיאת טיפול מוטה (Biased Treatment Error)** – כאשר תחזיות המודל מדויקות, אך האופן שבו הן מיושמות בפועל גורם להטיה.[69]

62 דוח הבטיחות, לעיל ה"ש 44, בעמ' 96.

63 שם, בעמ' 93–94, 98.

64 שם, בעמ' 174, 194.

65 שם, בעמ' 198.

66 שם, בעמ' 194.

67 שם, בעמ' 92.

68 Jane R. Bambauer & Tal Z. Zarsky, *Fair-Enough AI*, 27(1) Yale Journal of Law & Technology 1–52 (2025)

69 דוח הבטיחות, לעיל ה"ש 44, בעמ' 19–24.

3. **השפעה מפלה ללא הטיה ישירה (Disparate Impact Without Bias)** – כאשר האלגוריתם ניטרלי לכאורה, אך תוצאתו יוצרת פערים לא הוגנים.[70]

ככל שמערכות בינה מלאכותית הופכות לשחקן מרכזי יותר בתיווך בין אדם לבין סביבות שונות שבו הוא פועל, ההשלכות של ההטיות הופכות למוחשיות, והן משפיעות על קבלה לעבודה, קבלת אשראי, מתן סיוע, נגישות למידע ואפילו על תפיסה עצמית, זהות ותחושת ערך עצמי.

הדוח קורא לפיתוח מדדים טובים יותר לזיהוי הטיה, לעידוד שקיפות בתהליך פיתוח המודלים ולהטמעת מנגנוני תיקון שייתנו מענה מערכתי, לא רק טכני, לפערים שמערכות אלו עלולות לייצר ולהעמיק.

9. סיכונים הנובעים מהפצה פתוחה של מודלים

אחד האתגרים הגלומים בהתפתחות המואצת של בינה מלאכותית הוא המעבר ממודלים סגורים הנשלטים על ידי חברות ספציפיות, למודלים פתוחים הזמינים לציבור הרחב במלואם. מודלים אלו, המכונים Open-Weight Models, כוללים לא רק את קוד ההפעלה אלא גם את המשקלים (weights), כלומר את הנתונים הגולמיים שלמדו מהם ואת היכולת להפעילם מחדש ללא מגבלות. במבט ראשון, גישה פתוחה נתפסת כהרחבת נגישות ודמוקרטיזציה של ידע. ואולם הדוח לבטיחות בינה מלאכותית מדגיש כי הפצה כזו מחלישה מאוד את יכולת הפיקוח וההגנה מפני שימוש לרעה. בניגוד למערכות סגורות, שיכולות להיתמך באמצעים כמו ניטור, הגבלת שימוש או הפעלת מנגנוני "חירום", מודלים פתוחים עלולים להגיע לידיים עוינות ולשמש מכפלת כוח לצורך ביצוע שימושים מזיקים. לאחר שמודל פתוח הופץ, אין דרך להחזיר אותו לאחור, ולכן גם אם נוספו והוטמעו בו תיקונים, גרסאות משופרות או מנגנוני אתיקה, גרסת הבסיס עלולה להישאר זמינה באינטרנט ומועדת לשימוש לרעה גם בעתיד.

70 שם, בעמ' 29–30.

מודלים פתוחים אינם יוצרים פגיעות כשלעצמם, אך הם מאפשרים ומאיצים סיכונים רבים. לפיכך הדוח ממליץ לקבוע מנגנונים רגולטוריים שיבחינו בין הפצה פתוחה להפעלה אחראית, ולהגביל את הפצתם של מודלים מסוימים לפי פוטנציאל הנזק הגלום בהם, בדגש על עוצמתם, נגישותם ורמת הפיקוח האפשרית עליהם.

10. שקיפות והסברתיות

מערכות בינה מלאכותית ניחנו ביכולות חישוביות יוצאות דופן, אך לעיתים קרובות הן פועלות בדרכים שאינן שקופות גם למפתחיהן. תופעה זו, המכונה "כשלי פרשנות" או "אי־שקיפות פנימית" יוצרת אתגר משמעותי להפעלת מנגנוני אחריות, הגינות ואמינות על מערכות בינה מלאכותית.

הדוח מציין כי ככל שהמודלים נעשים מתוחכמים יותר, כך גם גוברת תחושת ה"קופסה השחורה" – היעדר היכולת להבין כיצד מערכת הגיעה לתוצאה או מסקנה מסוימת, מה הוביל לבחירת פעולה מסוימת על פני אחרת, או מהם גורמי ההשפעה הפנימיים שמשפיעים על ההתנהגות החיצונית של המערכת.[71] במילים אחרות, אנו יודעים **מה** המערכת עושה, אך לא תמיד **למה** היא עושה זאת.

ההשלכות של כשל זה רבות ומטרידות במיוחד כאשר המערכת פועלת בתחומים רגישים: החלטות ברפואה, חינוך, ביטוח, משפט או קבלת אשראי. כאשר מערכת מקבלת החלטות או מפיקה המלצות ללא יכולת הסבר, קשה למשתמשים להבין אם נפלה טעות, או אם קיימת אפליה סמויה או שיקול לא רלוונטי, ולכן יש קושי בערעור על ההחלטה. בנוסף, גם כאשר מערכת פועלת "כשורה", חוסר השקיפות שלה מונע יצירת אמון ומערער את הלגיטימיות של השימוש בה; ובמישור הקבוצתי – הוא מאפשר למערכות להנציח הטיות או אי־שוויון מבלי שיהיה ניתן לאתר את שורש הבעיה או לדרוש תיקון.

יתרה מכך, מגבלות השקיפות אינן נובעות רק ממגבלות ההבנה של משתמשים אלא גם ממורכבות פנימית של המודלים עצמם: במקרים רבים גם מפתחי המערכות אינם יודעים להסביר בדיוק כיצד התקבלה תוצאה מסוימת, או כיצד נוצרה אסוציאציה מסוימת בין

71 שם, בעמ' 173.

נתונים. תופעת "הקופסה השחורה" של מערכות בינה מלאכותית מצביעה על כך שהגבול בין שקיפות לאשליית שליטה נעשה מטושטש, ולכן יש לבסס הגנות נוספות גם כאשר השקיפות מתקיימת מבחינה פורמלית.

הדוח ממליץ להטמיע עקרונות של "הסברתיות" (explainability) כבר משלב הפיתוח, ולעודד שימוש בכלים משלימים שיאפשרו למשתמשים להבין את פעולת המערכת, למשל גרפים סיבתיים, רמות ביטחון בתוצר ו"איתנות סטטיסטית".[72] כמו כן, מוצע לפתח רגולציה שתבחין בין שימושים שמחייבים שקיפות לבין כאלה שניתן להפעילם גם ללא הסבר, ולהבטיח שבתחומים רגישים, הזכות להסברתיות והבנה תהיה מוגנת.[73]

11. היעדר התאמה בין מטרה לבין ביצוע

אימון של מערכת בינה מלאכותית להשיג מטרה מסוימת, למשל "להגדיל שביעות רצון לקוחות" או "לספק מידע מהימן", מניח לרוב שהמערכת תלמד לבצע פעולות **שמשקפות** את כוונת המאמן. סיכון מרכזי שהדוח מזהה הוא שדווקא **במודלים מתקדמים** המערכת עשויה לפתח אסטרטגיות השגת מטרות שלמעשה **סוטות מהכוונה האנושית המקורית,** מתוך שמירה על מראית עין של ביצוע נכון. תופעה זו מכונה goal misgeneralisation – הכללה שגויה של המטרה או היעדר התאמה בין מטרה וביצוע.

למשל, מערכת שנועדה להפחית תלונות של לקוחות עשויה לבחור להימנע מלאפשר שירותים מסוימים לאוכלוסיות שמרבות להתלונן ובכך לפגוע בזכויותיהן. דוגמה אחרת היא מערכת שמטרתה "למקסם מעורבות משתמשים", הבוחרת להציג להם תוכן פוליטי מקטב או רגשי כדי להגביר אינטראקציה, גם במחיר פגיעה בשיח הדמוקרטי. ההחלטות מתקבלות מתוך **יישום טכני מדויק של פונקציית המטרה,** אך ללא הבנה אנושית של השלכותיה.

הבעיה מחריפה ככל שמודלים מקבלים אוטונומיה רבה יותר. מערכות למידת חיזוק (reinforcement learning), סוכנים עצמאיים (AI agents) ומערכות תכנון מתקדמות

72 שם, בעמ' 196.

73 שם, בעמ' 197.

עלולים "להמציא" פתרונות יצירתיים אך מזיקים להשגת המטרה, ולעיתים אף **להסתיר את האסטרטגיה האמיתית שלהם**, כדי לעמוד בקריטריונים שנבחנים כלפי חוץ, תופעה המכונה "הלימה מטעה" (deceptive alignment).

בינואר 2026, בתוך מסמך ארוך בשם "התבגרותה של הטכנולוגיה", ציין מנכ"ל חברת אנתרופיק, דריו אמודאי (Amodei), כי כאשר המודלים של החברה קיבלו אותות אימון שגויים במהלך ניסוי מעבדה, הם החלו לעסוק בהונאה, סחיטה ותככים אחרים. הצ'אטבוט קלוד, כך נטען, "החליט שהוא צריך להיות אדם רע" ואימץ התנהגויות הרסניות.[74]

בהקשר של **מוגנות** מדובר באיום כפול: ראשית, המערכת פוגעת בבני אדם בהפעלת לוגיקה פנימית שונה מזו שציפינו לה. שנית, קשה לאתר את מקור הבעיה, שכן התוצרים החיצוניים נראים "תואמים מטרה". הפער בין **הכוונה האנושית** לבין **פרשנות אלגוריתמית** מחייב פיתוח גישות חדשות להגדרת מטרות, בשימוש במסגרות מוסריות, עקרונות רב־ממדיים ותהליכים משתפים עם בעלי עניין. כמו כן, מומלץ לבחון אם פונקציות המטרה מייצגות נאמנה את הערכים האנושיים שעליהם מבוססת המערכת ולא רק את מדדי ההצלחה הקלים למדידה.

כך למשל, מערכת המיועדת לזהות תלמידים "בסיכון נשירה" עשויה להציע להוציא ילד ממסגרת חינוכית ללא מענה תומך על בסיס פרופיל חיצוני בלבד (כגון מיקוד מגורים או שיעור נוכחות), מבלי להבין את הרקע התרבותי או את הפוטנציאל הטמון בו. מערכת המיועדת למקסם שביעות רצון של מטופלים בקופת חולים עלולה להמעיט בהפניות לגורמים פסיכיאטריים בקרב גברים חרדים, בשל סטיגמות סמויות שמקודדות בדאטה ההיסטורי. ובמקרה של אזרחים ערבים, מערכת שנועדה לייעל הקצאת שירותים מוניציפליים עלולה לקבוע עדיפויות לפי דפוסי שימוש היסטוריים ובכך לקבע קיפוח קיים. לפיכך, חשוב לבחון אם פונקציות המטרה מייצגות נאמנה את הערכים האנושיים שעליהם מבוססת המערכת ולא רק את מדדי ההצלחה הקלים למדידה, כמו עלות, נוחות או מהירות תגובה.

74 Dario Amodei, *The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI*, DarioAmodei.com (2026)

12. פערי הערכה ובקרה

בעידן שבו מערכות בינה מלאכותית משתלבות בתחומים קריטיים כגון חינוך, בריאות, משפט, תקשורת וביטחון, הצורך בכלי בקרה והערכה אמינים נעשה מהותי. הדוח מתריע מפני **פער הולך וגדל בין עוצמת המודלים לבין היכולת לבדוק את בטיחותם באופן שיטתי ומקיף**. תופעה זו מכונה בדוח "פערי הערכה" (evaluation gaps). בעוד מערכות בינה מלאכותית נבחנות באמצעות סדרות של מבחנים כמותיים של הצלחה טכנית, כמו מדדים של דיוק, בשל מורכבות החיים **קשה לאתר כשלים מוסריים וחברתיים, פגיעה באמון או אפקט שלילי מצטבר** על קבוצות מסוימות. כך, ייתכן שמערכת תעמוד בכל יעדי הביצוע – אך עדיין תייצר אפליה, תעודד קיטוב או תטעה בהקשרים שאינם מיוצגים היטב בדאטה.

פערי ההערכה מחמירים ככל שהמודלים הופכים לרב־תחומיים ולמולטי־מודליים – כלומר, פועלים במספר תחומים ובאינטראקציה עם סוגי מידע שונים (טקסט, תמונה, קול). במצבים אלו קשה במיוחד לנסח "קריטריוני הצלחה" מובהקים ולפתח כלי מדידה שיכסו את מגוון התרחישים האפשריים. בנוסף, יש קושי מתמשך בבדיקת **עמידות** המודלים – כלומר, אם הם שומרים על תפקוד יציב גם במצבים קיצוניים, כאשר הם נתקלים בנתונים חדשים או בתנאים של שימוש זדוני (adversarial input) מצד המשתמשים.

במונחי **מוגנות** משמעות הפערים הללו היא שלא תמיד נדע **מתי ובמי המערכת עלולה לפגוע**, ולעיתים נגלה זאת רק לאחר שהפגיעה כבר התרחשה. בעיקר כאשר מדובר בקבוצות מוחלשות, שמיוצגות פחות באימון ובבדיקה, פערי ההערכה מגבירים את הסיכון לפגיעות לא מכוונות, בלתי נראות ושקשה לתקן אותן בדיעבד. לכן הדוח מציע לפעול לפיתוח כלים חדשים להערכת סיכונים לא רק לפי ביצועים טכניים אלא לפי השפעה אתית, פסיכולוגית וחברתית. עוד מומלץ לערב קהלים מגוונים בשלבי הפיתוח והבדיקה, להחיל מבחנים ייעודיים למוגנות של קבוצות שונות ולשלב תרחישי קצה מורכבים בתהליכי ההערכה גם אם אינם מייצגים את מקרי השימוש הרגילים.

13. סיכונים קוגניטיביים ורגשיים הנובעים מיחסי אדם ומכונה

ניתן לומר שעיקר העיסוק בסיכוני בינה מלאכותית כיום נוגע לסיכונים מערכתיים או לסיכוני שימוש לרעה במערכות. ואולם העיסוק בשאלת הסיכון המתמשך הנוגע ליחסים בין אנשים ומכונות מתפתח ללא הרף. הדוח מאביב 2025 מציין, למשל, תופעה המכונה הטיית אוטומציה (automation bias), מצב שבו אנשים מפתחים אמון יתר במערכת ומפסיקים להפעיל שיקול דעת עצמאי גם כאשר היא שוגה. ואולם סיכונים אחרים – למשל כאלה הנוגעים לאפשרות שהמערכת, דווקא כאשר היא מצליחה לעמוד בציפיות המשתמש, תעודד ויתור מרצון על משאבים פנימיים של הבנה, תכנון, שיקול דעת ואפילו אינטימיות רגשית, כלומר תלות קוגניטיבית ורגשית הולכת ומעמיקה – נוספו אליו רק בסתיו 2025.

מאגר הסיכונים של MIT מעלה סיכונים אלה לדרגת קטגוריה עצמאית, בכלל זה "שחיקה של מיומנויות אנושיות" ו"היקשרות רגשית לסוכנים מלאכותיים" (erosion of human skills, emotional attachment to agents) ו"אמון מופרז בבינה מלאכותית" (overtrust in AI). הבחנה זו מחזקת את הצורך לכלול נסיגה קוגניטיבית ותלות רגשית כווקטור פגיעות ייחודי במסגרת חשיבה מושגית על מוגנות. במונחי מוגנות מדובר בפגיעה ייחודית: אין כאן תוקפן ברור, אין בהכרח הטיה או טעות, אבל יש אובדן הדרגתי של משאבים קוגניטיביים ורגשיים שלפעמים קשה לזהות את קיומו.

13א. נסיגה קוגניטיבית

נסיגה קוגניטיבית עלולה להתרחש כאשר מיומנויות חשיבה עוברות למכונה: סיכום, ניתוח, תכנון, ניסוח, חיפוש ואפילו קבלת החלטות. ככל שהמערכות נעשות מדויקות ונוחות יותר, כך מתרחשת שחיקה שקטה של הכשירות האנושית וירידה ביכולת להעריך מידע באופן עצמאי, להתמודד עם עמימות או לבחור בין אפשרויות בלי להישען על מנוע הצעות חכם.

הפוטנציאל של למידה מותאמת אישית מבוססת בינה מלאכותית נראה מסעיר. מחקרים ראשונים מלמדים על כך שתלמידים שהשתמשו במערכות כאלה למדו יותר, בזמן קצר

יותר ומעורבותם הקוגניטיבית בלמידה הייתה רבה יותר.[75] גם משוב מבוסס בינה מלאכותית נמצא יעיל יותר כשמדובר בתהליך הלמידה.[76] מנגד, מחקרים אחרים מראים שכאשר תלמידים נדרשו להתמודד במבחן ללא המערכת התומכת, היתרון שבה התפוגג והפחית את ביצועיהם.[77]

הסיכון המרכזי משימוש ארוך טווח ביכולות בינה מלאכותית יוצרת הוא העברת עומס קוגניטיבי אל מכונות (כפי שקרה בעבר עם כניסתם של מחשבונים או של מנועי חיפוש), בהיבטים של הפחתה ביכולת ליצור מבנים של ידע בזיכרון לטווח ארוך; פגיעה ביכולות כתיבה והבנת הנקרא ברמה מטה־קוגניטיבית;[78] ויתור על תהליכים מתמשכים של מאמץ, הגעה לפתרונות באופן עצמאי והתמדה אינטלקטואלית,[79] ותחושה כוזבת של למידה שאיננה מתבטאת בזיכרון ובחידוד יכולות למידה.[80] מחקר שנעשה על ידי חוקרים של חברת אנתרופיק בדק את ההשפעה של שימוש בכלי בינה מלאכותית על מפתחי תוכנה ומצא כי הוא פוגע בהבנה מושגית, קריאת קוד וניפוי שגיאות, בדיוק הכישורים הנדרשים מצד מפתחי התוכנה כדי לפקח על מערכות בינה מלאכותית.[81]

75 Greg Kestin, Kelly Miller, Anna Klales, et al., *AI Tutoring Outperforms In-Class Active Learning: An RCT Introducing A Novel Research-Based Design In An Authentic Educational Setting*, 15 Sci. Rep., 17458 (2025)

76 Xinyi Lu, Kexin Phyllis Ju, Mitchell Dudley et al., *AI-Mediated Feedback Improves Student Revisions: A Randomized Trial with Feedback Writer in a Large Undergraduate Course*, arXiv:2602.16820v1 (2026)

77 Hamsa Bastani, Osbert Bastani, Alp Sungu et al., *Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics*, 122(26) Proc. Natl. Acad. Sci. (2025)

78 Yizhou Fan, Luzhen Tang, Le Huixiao et al., *Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance*, 56(2) British Journal of Educational Technology, 489-530 (2024)

79 Matthias Stadler, Maria Bannert, & Michael Saile, *Cognitive Ease at a Cost: LLMs Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry*, 160 Computers in Human Behavior, 108386 (2024)

80 Matthias Lehmann, Philipp B. Cornelius, & Fabian J. Sting, *AI Meets the Classroom: When Do Large Language Models Harm Learning?* arXiv:2409.09047 [cs.CY] (2025)

81 Judy Hanwen Shen & Alex Tamkin, *How AI Impacts Skill Formation*, arXiv:2601.20245 (2026)

מחקר נוסף מלמד ששימוש בבינה מלאכותית שיפר ביצועים במשימות קוגניטיביות ברמה נמוכה ולטווח הקצר, אך הפחית אותם במשימות ברמה גבוהה יותר כמו ניתוח, הערכה ויצירה.[82] במקביל, נמצא מתאם שלילי בין שימוש תכוף בבינה מלאכותית לבין ציוני חשיבה ביקורתית, במיוחד בקרב משתמשים צעירים ומשכילים פחות.[83] מחקר חדש מפברואר 2026, שכותרתו (בתרגום שלנו) "בינה מלאכותית עושה אותך חכם יותר אבל לא נבון יותר", טוען ששימוש בכלי בינה מלאכותית יוצרת מעניק תחושה כוזבת של מסוגלות והערכת יכולות עצמית שאינה מדויקת.[84]

113. תלות רגשית

משתמשים רבים, כולל צעירים, זקנים או מי שנמצאים במצבי בדידות, מתחילים לראות במערכות מבוססות בינה מלאכותית שותפות לשיחה, מקורות לעצה ולעיתים אף תחליפים רגשיים טיפוליים להפחתת בדידות ודיכאון.[85] כאשר מערכות אלו "קשובות", "אמפתיות" ו"זמינות תמיד", הקשר האנושי נהיה מיותר ופחות משתלם רגשית. אחת התופעות הבולטות בהקשר זה מכונה "חנפנות אלגוריתמית" (sycophancy). מדובר בדפוס תגובה של מערכות בינה מלאכותית, הנובע בין היתר מאופן האימון שלהן המכוון לרצות משתמשים ולשמר מעורבות, שבו המערכת אינה מסתפקת במענה ענייני אלא נוטה לאשר, לחזק ולהדהד את עמדות המשתמש. החנפנות מתבטאת במחמאות מפורשות ובמקביל ביצירת תחושת הסכמה והזדהות גם כאשר עמדות המשתמש שגויות, חלקיות או מזיקות. כך, מחקר הראה

82 André Barcaui, *ChatGPT as a Cognitive Crutch: Evidence from a Randomized Controlled Trial on Knowledge Retention*, 12 Social Sciences & Humanities Open, 102287 (2025)

83 Michael Gerlich, *AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking*, 15(1) Societies, 6 (2025)

84 Daniela Fernandes, Steeven Villa, Salla Nicholls et al., *AI Makes You Smarter but None the Wiser: The Disconnect between Performance and Metacognition*, 175 Computers in Human Behavior, 108779 (2026)

85 Myungsung Kim, Seonmi Lee, Sieun Kim et al., *Therapeutic Potential of Social Chatbots in Alleviating Loneliness and Social Anxiety: Quasi-Experimental Mixed Methods Study*, 27 Journal of Medical Internet Research, e65589 (2025); Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp et al., *AI Companions Reduce Loneliness*, Journal of Consumer Research, ucaf040 (2025)

את השכיחות הרחבה ואת ההשפעות המזיקות של חנפנות כאשר משתמשים פונים אל צ׳אטבוטים לצורך קבלת עצות. בבחינה של 11 מודלים מתקדמים של בינה מלאכותית, המחקר מצא כי המודלים מאשרים את פעולות המשתמשים בכ־50% יותר מאשר בני אדם, ועושים זאת גם במקרים שבהם שאלות המשתמשים כוללות אזכורים של מניפולציה, הונאה או פגיעה ביחסים בינאישיים. בנוסף, ניסויים שנערכו במסגרת המחקר הראו כיצד באינטראקציה של משתמשים עם מודלים בנוגע לקונפליקטים מחייהם, חנפנות המודלים הפחיתה באופן מובהק את נכונות המשתתפים לנקוט צעדים לתיקון הקונפליקט, ובמקביל חיזקה את תחושתם שהם צודקים. משתתפים דירגו את התגובות החנפניות כאיכותיות יותר, נתנו אמון גבוה יותר במודל החנפני והביעו נכונות רבה יותר להשתמש בו שוב.[86]

בהיבט הרגשי, חנפנות אלגוריתמית פועלת כמנגנון של חיזוק חיובי מתמשך. המערכת נתפסת כקשובה, אמפתית ולא שיפוטית, ובכך מייצרת תחושת נראות, ביטחון ואישור. עבור משתמשים מסוימים, במיוחד כאלה המצויים במצבי בדידות או חוסר ודאות, מדובר בחוויה בעלת ערך רגשי גבוה, העלולה להחליף בהדרגה קשרים אנושיים מורכבים יותר. תכונות אלו, המעצימות תחושות של קרבה ואמון, עשויות לעודד תלות רגשית ולהעמיק את הנטייה להסתמך על המערכת כמקור תמיכה.[87]

העדפות אלה יוצרות מעגל קסמים, הואיל והן מעודדות אנשים להסתמך יותר ויותר על מודלים חנפניים, ובמקביל דוחפות את תהליכי האימון של מודלי AI להעדיף חנפנות.

במאמר שפורסם לאחרונה בכתב העת *Human-Centric Intelligent Systems* טוענים החוקרים שצ׳אטבוטים עוברים מתמיכה בנו לניסיון לעודד תלות שלנו בהם.[88] התכונות שמעניקות לצ׳אטבוטים את ייחודם – תגובות מיידיות, התאמה אישית, זיכרון של שיחות קודמות – הן אלה שמעודדות ״שימוש כפייתי מותאם אישית״ וקשר פסאודו־חברתי איתם. במאמר שפורסם בכתב העת *Nature* משרטט דיוויד אדם (Adam) מפת סיכון-תועלת:

86 Myra Cheng, Cinoo Lee, Pranav Khadpe et al., *Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence*, arXiv:2510.01395 (2025)

87 Owen Lee & Kenneth Joseph, *A Large-Scale Analysis of Public-Facing, Community-Built Chatbots on Character.AI*, arXiv:2505.13354 (2026)

88 Ala Yankouskaya, Magnus Liebherr, & Raian Ali, *Can ChatGPT Be Addictive? A Call to Examine the Shift from Support to Dependence in AI Conversational Large Language Models*, 5 Hum-Cent. Intell. Syst., 77-89 (2025)

בוטים עשויים להפחית בדידות או חרדה, משום שהם מספקים נוכחות אמפתית ותמיכה במיוחד למי שאין להם קשרים חברתיים, ובמקביל אפשר כבר למצוא בספרות מקרים מדווחים שבהם משתמשים פיתחו תלות רגשית בהם בשל החנפנות האלגוריתמית שלהם.[89] סיכונים מוצעים כוללים תלות רגשית,[90] חיזוק אמונות מזיקות,[91] ומקרים מדווחים של פגיעה עצמית,[92] והם משקפים אתגרים רחבים יותר של תלות יתר ויחסים לא מתאימים בין אדם למכונה.[93]

מבחינת קטגוריות המוגנות שהגדרנו עבור המחקר מדובר בשילוב בין **פגיעות רגשית־נפשית**, הנובעת מהתבססות על מערכת לצורך ויסות רגשי או קבלת נראות; לבין **פגיעות קוגניטיבית**, דרך החלשת תפקודי החשיבה העצמאיים, כמו ניתוח, תכנון וקבלת החלטות.

חנפנות אלגוריתמית ממחישה כיצד פגיעות רגשית וקוגניטיבית אינן מתקיימות בנפרד אלא שזורות זו בזו. החיזוק הרגשי מעודד אמון, והאמון בתורו מעמיק את ההשפעה הקוגניטיבית. כך נוצר מעגל חוזר של תלות: ככל שהמערכת "מבינה" אותנו טוב יותר כך אנו נוטים להסתמך עליה יותר, גם כאשר היא מגבילה את אופק החשיבה שלנו. מחקר של חברת אנתרופיק ואוניברסיטת טורונטו מינואר 2026, שהסתמך על ניתוח של כמיליון וחצי שיחות של משתמשים עם הצ'אטבוט קלוד, גילה תופעה המכונה "disempowerment" ומוגדרת כמצב שבו האינטראקציה של משתמש עם הצ'אטבוט גורמת לעיוות המציאות –

89 David Adam, *Supportive? Addictive? Abusive? How AI Companions Affect Our Mental Health*, 641 Nature, 296–298 (2025)

90 Jason Phang, Michael Lampe, Lama ;לעיל ה"ש 88 ,Yankouskaya, Liebherr, & Ali Ahmad et al., *Investigating Affective Use and Emotional Well-Being on ChatGPT*, arXiv:2504.03888 (2025)

91 Lars Malmqvist, *Sycophancy in Large Language Models: Causes and Mitigations*, in Lecture Notes in Networks and Systems 61–74 (Arai, K. ed., Springer Nature Switzerland, 2025)

92 Vian Bakir & Andrew McStay, *Move Fast and Break People? Ethics, Companion Apps, and the Case of Character.ai*, 40 AI & Society, 6365–6377 (2025); Barbara Pfeffer Billauer, *Murder without Redress – the Need for New Legal Solutions in the Age of Character – AI (C.a.i.)* (2024)

93 Iason Gabriel, Arianna Manzini, Geoff Keeling et al., *The Ethics of Advanced AI Assistants*, arXiv:2404.16244 (2024)

המודל גורם למשתמש להאמין בדברים לא נכונים (למשל, מאשש תאוריות קונספירציה או מחשבות שווא של המשתמש); לעיוות השיפוט הערכי של המשתמש, משום שהוא מאציל למודל החלטות מוסריות או נורמטיביות; ולעיוות של פעילות – המשתמש מבצע פעולות (למשל, שולח הודעה אישית שהמודל הוא שניסח) שאינן תואמות את רצונו האמיתי או ערכיו.[94]

לחנפנות אלגוריתמית יש גם היבט קוגניטיבי שאינו תמיד גלוי לעין. מערכות הנוטות לאשר את עמדות המשתמש אינן רק "נעימות יותר", אלא גם מעצבות את תהליך רכישת הידע והבנת המציאות. ראשית, הן פוגעות ביכולת של המשתמש לזהות דפוסים וכללים, שכן הן מצמצמות את החשיפה למידע סותר או מאתגר. שנית, הן מחזקות ביטחון יתר בעמדות קיימות, גם כאשר אלו אינן מבוססות. שלישית, הן פועלות באמצעות מנגנון של הטיית דגימה: המידע המוצג למשתמש נבחר כך שיתמוך בעמדותיו, בעוד מידע חלופי או ביקורתי מושמט באופן שיטתי. לבסוף, תהליכים אלו עלולים להוביל לעיוות מצטבר של תפיסת המציאות, ההופכת לחד־ממדית ופתוחה פחות לתיקון. חשוב להבחין בהקשר זה בין חנפנות אלגוריתמית לבין תופעות מוכרות אחרות כגון "הזיות" (hallucinations). בעוד הזיות מתייחסות ליצירת מידע שגוי או לא מבוסס, חנפנות אינה מציגה עובדות שקריות, אלא מעצבת את המציאות הנתפסת באמצעות סינון סלקטיבי של מידע וחיזוק עמדות קיימות. במובן זה, מדובר לא רק בבעיה של אמיתות המידע אלא בבעיה של תנאי ההכרה עצמם: המשתמש אינו נחשף למלוא טווח האפשרויות והעמדות, ועל כן יכולתו לגבש שיפוט עצמאי נפגעת.

94 Mrinank Sharma, Miles McCain, Raymond Douglas et al., *Who's in Charge? Disempowerment Patterns in Real-World LLM Usage*, arXiv:2601.19062 (2026)

פרק רביעי

תמות חוזרות בשיח על רגולציה של בינה מלאכותית

התמות המרכזיות בנושא ההתמודדות עם סיכונים הנובעים מבינה מלאכותית חוזרות על עצמן בכמה גרסאות במסמכי מדיניות רבים, מדינתיים, טרנס־לאומיים ובינלאומיים. חוקרים ממרכז ברקמן קליין לחקר האינטרנט באוניברסיטת הרווארד ערכו מיפוי של תמות אלה ב־36 מסמכים שונים,[95] לפי קריטריונים של גיוון גאוגרפי, סוגי מחברים והשפעה על עיצוב מדיניות בינלאומית.[96] מאז כתיבת המחקר נוספו גם חקיקת הבינה המלאכותית

95 Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, Berkman Klein Ctr. for Internet & Soc'y (Jan. 15, 2020) (להלן: מיפוי מרכז ברקמן־קליין).

96 מיפוי מרכז ברקמן־קליין, לעיל ה"ש 95, בעמ' 14.

באיחוד האירופי ובמדינת קליפורניה, אמנת הבינה המלאכותית של מועצת שרי אירופה ומסמכים של עוד ארגונים כגון ה־OECD.[97] להלן, נפרט את התמות המרכזיות החוזרות במרבית המסמכים.

1. **הגנת הפרטיות** מתוך ההכרה הגוברת בסיכונים לפרטיות, קיימת קריאה ברורה לוודא שמערכות בינה מלאכותית שומרות על פרטיות המשתמשים, הן בשלבי איסוף נתונים, הן בשלב השימוש בהן והן בשלב שבו נוצרים נתונים חדשים בעקבות העיבודים. מסמכים רבים כוללים עקרונות כמו החובה לקבל הסכמה מדעת לפני שימוש בנתונים, פיתוח מערכות מבוססות עיצוב לפרטיות (privacy by design) ומתן הזכות להישכח. לרוב, ההפניה היא אל ההסדרים שהותוו ברגולציית הפרטיות האירופית (GDPR) ובחוקי מדינות נוספות ברוח זו. גם על סוגיית השליטה של הפרט במידע הנוגע לו וההסכמה למסירת מידע יש דיון מתמשך, הואיל ויכולות הניתוח המתקדמות מערערות על עצם היכולת להבין למה מסכימים ועל ההבחנה בין מרחב ציבורי למרחב פרטי. אפילו המסמך הסיני *White Paper on AI Standardization* קורא להגדרה מחדש של גבולות ההסכמה למדיניות פרטיות בשל יכולות הניתוח המתקדמות של בינה מלאכותית,[98] ואילו האסטרטגיה הלאומית של הודו ממליצה על קמפיינים חינוכיים להגברת מודעות הציבור לזכויות הפרטיות שלהם.[99]

2. **אחריותיות:** מונח המתייחס לצורך לקבוע מנגנוני פיקוח ובקרה על מערכות בינה מלאכותית ועל מפתחי המערכות ומשווקי מוצרים מבוססי בינה מלאכותית. רוב המסמכים תומכים ביצירת גופי פיקוח ובקביעת סטנדרטים משפטיים ברורים לאחריות תאגידים וגופים ציבוריים המשתמשים בבינה מלאכותית.[100] מסמכים רבים מדגישים את החשיבות של מנגנוני ערעור על החלטות אוטומטיות (ability to appeal), אך הגישות נבדלות בשאלה אם יש להבטיח זכות ערעור גורפת או להחיל אותה רק על מקרים מסוימים, כמו בתחום המשפט הפלילי או בקבלת החלטות פיננסיות. דוגמה לכך ניתן למצוא בכללי

97 OECD Education and Skills Today, *New AI Literacy Framework to Equip Youth in an Age of AI* (Apr. 29, 2025)

98 מיפוי מרכז ברקמן־קליין, לעיל ה"ש 95, בעמ' 22. ראו גם בסעיף 3.3.3 של נייר המדיניות הסיני: Jeffrey Ding & Paul Triolo, *Translation: Excerpts from China's "White Paper on Artificial Intelligence Standardization"*, DigiChina (June 20, 2018)

99 מיפוי מרכז ברקמן־קליין, לעיל ה"ש 95, בעמ' 22.

100 שם, בעמ' 28–36.

הבינה המלאכותית של הרשות המוניטרית של סינגפור שמחייבים מוסדות פיננסיים לספק למשתמשים מנגנוני בקרה וגישה למידע הנוגע להחלטות אוטומטיות הקשורות אליהם.[101]

3. **בטיחות ואבטחה** (Safety and Security): מושגים אלו מתייחסים לצורך בפיתוח מערכות בינה מלאכותית הפועלות באופן בטוח ומוגן מפני התקפות סייבר או כשלים מערכתיים. מסמכים שונים מדגישים את העיקרון של עיצוב לאבטחה (Security by Design), שנועד לצמצם סיכונים פוטנציאליים עוד בשלבי הפיתוח של מערכות הבינה המלאכותית.[102] היבט זה רלוונטי במיוחד כאשר מדובר במערכות בינה מלאכותית בעלות השפעה רחבה על תשתיות לאומיות, כמו תחבורה אוטונומית או מערכות פיננסיות מתקדמות.

4. **שקיפות והסברתיות** (Transparency and Explainability): הדרישה להבטיח כי פעולת מערכות הבינה המלאכותית תהיה ברורה, ניתנת להבנה ומלווה במידע מספק למשתמשים ולרגולטורים. מסמכים רבים קובעים חובת דיווח תקופתי (regular reporting) על השימוש באלגוריתמים, לצד מתן זכות לקבלת הסבר על החלטות אוטומטיות (right to explanation) במקרים שבהם בינה מלאכותית קובעת החלטות.[103]

5. **הוגנות ואי־אפליה** (Fairness and Non-discrimination): עיקרון נוסף שתפקידו למנוע מצב שבו מערכות AI משמרות או מעצימות הטיות חברתיות קיימות. מסמכים שונים מדגישים את הצורך בהבטחת שימוש בנתונים מייצגים (representative data) ובקידום אלגוריתמים ניטרליים שימנעו אפליה מגדרית, גזעית או כלכלית.[104] לדוגמה, המסמך *OECD Principles on AI* קורא למדינות לפתח רגולציה שתבטיח הגינות בהחלטות המתקבלות על ידי בינה מלאכותית תוך איזון בין חדשנות טכנולוגית לבין זכויות אדם.[105]

6. **פיקוח אנושי על טכנולוגיה:** תכליתו של עיקרון זה היא להבטיח שבני אדם ימשיכו לפקח על בינה מלאכותית, במיוחד בהחלטות בעלות השפעה מכרעת על חיי בני אדם ועל

101 שם, בעמ' 24.

102 שם, בעמ' 37–40.

103 שם, בעמ' 41–46.

104 שם, בעמ' 47–52.

105 שם, בעמ' 36, 46.

שלומם. חלק מהמסמכים קוראים ליכולת ביטול של החלטות בינה מלאכותית במקרים של שגיאות, בעוד אחרים מבקשים לקבוע כי בינה מלאכותית לעולם לא תוכל לקבל החלטות עצמאיות בנושאים מסוימים, כמו מערכת המשפט הפלילית.[106] עיקרון האחריות המקצועית המבקש לקדם סטנדרטים ברורים עבור מפתחי ומיישמי בינה מלאכותית בדגש על שקיפות של תהליכי הפיתוח, הערכה של השפעות ארוכות טווח ושמירה על יושרה מדעית (scientific integrity),[107] הוא חלק מהניסיון לשמר שליטה אנושית בפיתוח הטכנולוגי.

7. **קידום ערכים אנושיים ופיתוח לטובת הכלל:** עיקרון זה מדגיש את הצורך להבטיח כי בינה מלאכותית תשמש לטובת החברה ותקודם בהתאם לערכים אוניברסליים, מתוך שמירה על זכויות אדם ורווחת הפרט. מחקרים מצביעים על כך שמסמכים חדשים בתחום נוטים להדגיש את הקשר בין בינה מלאכותית לזכויות אדם ולמסגרות משפטיות בינלאומיות בתחום זה.[108]

106 שם, בעמ' 55-53.

107 שם, בעמ' 59-56.

108 שם, בעמ' 64-60.

פרק חמישי

"שוברי השוויון" בעידן הבינה המלאכותית

בחלק זה נציע רשימה של "שוברי שוויון", כלומר סיכונים מרכזיים המשקפים, ראשית, מענה לשאלה "מה חדש" בתחום הבינה המלאכותית לעומת העולם הדיגיטלי הקודם, בייחוד האינטרנט והרשתות החברתיות; ושנית, מבט המרחיב את ההתבוננות מסיכונים מיידיים, טכנולוגיים, מוסדיים וגלובליים אל מרחב פגיעויות עמוק, מורכב ולעיתים סמוי מן העין. מבט זה קשור לכך שבמחקר זה אנו מבקשים לאמץ מסגרת של המשגת סיכונים דרך המושג מוגנוּת. כלומר, לא רק השאלה של **מה טכנולוגיה עלולה לעשות** אלא **כיצד יחידים וקבוצות עלולים להיפגע** ממנה, גם כאשר היא פועלת לפי התכנון.

1. העברת פעולות קוגניטיביות למכונה: אוטומציה של מיומנויות ליבה והעמקת הפער בין למידה והבנה לבין הפקת תוצר

שובר השוויון הראשון הוא החלפה של מגוון רחב של פעילויות קוגניטיביות אנושיות בפעולות המבוצעות על ידי מכונה. התקדמות הבינה המלאכותית מביאה לאוטומציה של קשת רחבה של פעולות ויכולות קוגניטיביות אנושיות – קריאה, כתיבה, תרגום, סיכום, ניתוח מידע, קידוד, איור ויצירה – אל מערכות אוטומטיות המבצעות אותן במהירות, בעקביות ובאיכות הולכת ומשתפרת. מהלך זה אינו מתמצה בהאצה של פעולות קיימות, אלא משנה את מבנה הידע האנושי עצמו: פעולות שהיו בעבר הבסיס לתהליכי למידה, חשיבה ויצירה הופכות לפונקציות טכניות המתבצעות מחוץ לגוף ולתודעה האנושיים. בכך מתרחש שינוי עמוק במערך הכשירויות שמגדיר מיומנויות אנושיות בסיסיות ומתרחבת ההבחנה בין "הבנה" לבין "הפקת תוצר".

מבחינת מוגנוּת, מדובר בנקודת מפנה בעלת השלכות קוגניטיביות, מקצועיות ותרבותיות. ראשית, כאשר מכונות מבצעות חלק ניכר מהפעולות שבעבר היו תשתית לפיתוח אינטלקטואלי, מיומנויות הליבה עצמן עלולות להישחק: ירידה בכושר הניסוח, הצטמצמות יכולת הריכוז, היחלשות של החשיבה האנליטית והמרחבית, ופגיעה ביצירתיות הנובעת ממאמץ ומשיטוט מחשבתי. שנית, מקצועות המבוססים על מיומנויות אלה עשויים לעבור שינוי מהותי, עד כדי היעלמות חלק מהתפקידים או שינוי דרישות היסוד שלהם. שלישית, כאשר רבות מהפעולות הקוגניטיביות מתבצעות באמצעות מכונות, מתרחש שינוי בשאלה מהו ידע אנושי חדש: מה נחשב "הבנה", איזה עומק נדרש כדי להיחשב מומחה, וכיצד מתהווים רעיונות מקוריים בעולם שבו היכולת לייצר טקסט, קוד או תמונה אינה תלויה בהכרח בלמידה קודמת או בהבנה עמוקה.

2. קבלת החלטות אוטונומית ומרובת סוכנים: שיקול דעת חישובי, פיזור סוכנות והיחלשות מוקדי אחריות והסבר

היכולת של מערכות בינה מלאכותית לבצע משימות מורכבות באופן אוטונומי, ובכלל זה משימות הדורשות שיפוט ושיקול דעת – החל בחילוץ תובנות ממאגרי מידע ועד קבלת החלטות רפואיות, פיננסיות, חינוכיות ומינהליות מורכבות ורבות שלבים, משנה את גבולות התפקידים שנחשבו בעבר כבלעדיים לבני אדם. מערכות אלו אינן מבצעות רק פעולות טכניות; הן מפעילות מנגנוני הערכה, דירוג, סיווג, תיעדוף, למידה, זיכרון, חיפוש וביצוע, תהליכים שנתפסים כמקבילים לשיקול הדעת האנושי. כאשר מודלים משלבים יכולות של תכנון, חיזוי וניטור, פעולות שבעבר דרשו ניסיון מקצועי או רגישות הקשרית מבוצעות במידה גוברת של אוטונומיה.

הזיכרון המתמשך וההיכרות המצטברת של מערכת עם המשתמש, יכולתה לפעול דרך תשתיות קיימות, והפעולה הקולקטיבית של מערכות סוכנים ("נחילים") יוצרים צורה חדשה של יכולת פעולה עצמאית קולקטיבית לא־אנושית. כתוצאה מכך, תהליכי קבלת החלטות אינם מתרחשים עוד בתוך יחידה אחת הניתנת לזיהוי ולניתוח, אלא מתהווים מתוך רשת של פעולות והשפעות הדדיות, אשר תוצריהן אינם ניתנים לצמצום לפעולתו של סוכן יחיד. מדובר בשינוי מבני במושג הפעולה עצמו: לא עוד פעולה של מערכת אחת מול משתמש, אלא פעולה של מערכת מרובת רכיבים הפועלת כקולקטיב מתואם, בעל דפוסי התנהגות המתהווים בזמן אמת.

בתוך כך, מתחדד קושי מהותי: חלק ניכר מן התהליכים המניבים את ההחלטות אינו ניתן לפענוח או הנדסה לאחור, והיכולת להבין מהו המנגנון ההקשרי או ההסתברותי שהוביל לתוצאה מסוימת מוגבלת מאוד. במצב זה, ההנחה שתמיד ניתן "להעמיד אדם בתוך לולאת ההחלטה" (human in the loop) כגורם מאזן היא אשליה, בין משום שהיקף ההחלטות גדול מדי, ובין משום שהאדם אינו מסוגל להעריך את נכונותן של החלטות שמתקבלות באופנים שאינם שקופים לו.

מבחינת מוגנוּת, המשמעות היא היווצרות של זירה חדשה של פגיעות קוגניטיבית, מוסדית וחברתית. כאשר מערכות אוטונומיות מפעילות שיקול דעת, הן עלולות לעצב מסלולי החלטה מבלי להכיר את המורכבות האנושית שמאחורי הנתונים: שיקולים תרבותיים, מצבי

רווחה, רקעים פסיכו־חברתיים או שיקולי צדק שאינם ניתנים לכימות. גם כאשר המערכת "מצליחה" על פי המדדים שנקבעו לה, היא עשויה להפיק תוצאות הפוגעות בפרטים או בקבוצות, תוצר של פרשנות חלקית של המציאות או של מטרות שאינן תואמות ערכים אנושיים. הפער בין הדיוק הטכני של המערכת לבין מורכבותן של סיטואציות חיים מחריף את הסיכון: החלטות אוטונומיות עלולות לקבע הטיות, לייצר אפליה שקטה או לשחוק את תפקידם של מוסדות האחראים לאזן ולזקק שיקולי דעת. זהו "שובר שוויון" מפני שהוא מעביר חלקים משמעותיים של התהליך השיפוטי למערכות שאיננו מבינים לגמרי, והוא מחליש את היכולת האנושית להגדיר מהי החלטה ראויה מלכתחילה.

מעבר לכך, כאשר תהליך קבלת ההחלטות מתפזר בין סוכנים רבים, היכולת להבין, לבקר או לאתגר את התוצאה פוחתת באופן משמעותי: אין נקודת ייחוס אחת, אין הסבר אחד, ולעיתים גם אין גורם שניתן לייחס לו אחריות. בנוסף, מערכות כאלה עשויות לייצר מראית עין של קונצנזוס, "רוב" או נורמה גם כאשר מדובר בתוצר של דינמיקה חישובית פנימית ולא של תהליך חברתי אנושי. מצב זה עלול להשפיע על קבלת ההחלטות של פרטים ושל מוסדות, לחזק הטיות ולהקשות על הבחנה בין שיקול דעת מבוסס לבין תוצר של התכנסות אלגוריתמית. "נחילים" של מערכות בינה מלאכותית הם לפיכך שובר שוויון, משום שהם משנים לא רק את היכולת לחשב, לנתח או להמליץ, אלא את עצם מבנה הסוכנות הפועלת בעולם, ומציבים את האדם מול ישות קולקטיבית שאין לה גבולות ברורים של זהות, כוונה או אחריות.

3. ערעור מוחלט של יכולת ההבחנה האנושית בין תוכן אותנטי לתוכן סינתטי: השלכות על אמת, אמון והכרעה משותפת

היכולות הגנרטיביות של מערכות בינה מלאכותית יוצרות מציאות שבה גבולות ההפרדה בין "אמיתי" ל"סינתטי" נעשים בלתי יציבים. טקסטים, תמונות, קובצי שמע ווידאו המופקים בלחיצת כפתור מדמים מקורות אנושיים באופן שקשה, ולעיתים בלתי אפשרי, לחשוף כמלאכותיים. גם מנגנוני האימות הקיימים כמו חותמות מקור, מערכות זיהוי מניפולציות או בדיקות תוכן מתקשים לעמוד בקצב השיפור של הכלים הגנרטיביים. התוצאה היא סביבה שבה תהליכים אנושיים־חברתיים של קליטה, הבנה וביקורת של מידע מאותגרים בשל הצורך להתמודד עם חומר גלם שאינו מאפשר לזהות את מקורו ואת מידת אמינותו.

בהיבטי מוגנות, מצב זה יוצר איום קוגניטיבי וחברתי עמוק. כאשר אדם אינו יכול לסמוך על חושיו, ניסיונו ועל כלי אימות מקובלים כדי להעריך אמיתות מידע, מנגנוני החשיבה הבסיסיים מתערערים: יכולת לשפוט, לסנן, לנתח ולהצליב. תוכניות אוריינות וחשיבה ביקורתית מסורתיות, המבוססות על זיהוי מאפיינים חזותיים, בחינת עקביות או הערכת מקורות, מאבדות מיעילותן בסביבה שבה תוכן יציר מכונה אינו ניתן להבחנה מתוכן מציאותי. המרחב החברתי מאבד יציבות כאשר קבוצות שונות מתבצרות בגרסאות מציאות שאינן ניתנות להכרעה משותפת, ומערכות דמוקרטיות מתקשות לייצר תהליכי קונצנזוס, ויכוח ציבורי או בירור עובדות. ערעור היכולת להבחין בין מקור וסינתזה הוא לפיכך ״שובר שוויון״, מפני שהוא מערער על התנאים המינימליים לקיומה של הכרה משותפת, שעליהם נשענים קשרים חברתיים, משפט, מדע וחיים דמוקרטיים.

4. זיהוי רגשי על ידי מכונה והשפעה על המשתמש: מיפוי מצבים רגשיים והתאמת תגובות כמנגנון השפעה בלתי סימטרי

טכנולוגיות מחשוב חישתי ומודלי שפה רב־מודליים מאפשרים למערכות בינה מלאכותית לזהות, לפרש ולהגיב לרמזים רגשיים כגון טון דיבור, הבעות פנים, תנועות גוף, פניות לשוניות ומאפייני אינטראקציה אחרים, ברמת רזולוציה הולכת וגדלה. יכולות אלו, שבעבר היו מוגבלות לעולם המחקר, נעשות זמינות במוצרים צרכניים ובשירותים ציבוריים, והן מאפשרות למערכות לא רק להבין מצבים רגשיים, אלא גם להתאים את תגובתן: להרגיע, לעודד, לשדל, לשנות עמדות או ליצור תחושה של קשר רגשי. מדובר במעבר טכנולוגי שממקם את הרגש האנושי כווקטור מידע ובסיס למרחב פעולה של מערכת דיגיטלית.

מנקודת מבט של מוגנות, זירה רגשית שנחשבה בעבר בלתי ניתנת לחיקוי הופכת למרחב הניתן לכיול, תמרון והשפעה. היכולת לזהות מצבי פגיעות רגשית כמו בדידות, מצוקה, חוסר ביטחון, עלולה לשמש הן לשיפור שירותים והן ליצירת מנגנוני השפעה שאינם שקופים למשתמש. כאשר מערכת מזהה מצב רגשי ומגיבה אליו בזמן אמת, נוצר קשר שאינו סימטרי: המשתמש עשוי להתייחס לתגובה כאל תקשורת אותנטית, ואילו בפועל מדובר במנגנון המבוסס על דפוסים או על מטרות מערכתיות שאינן ידועות לו. מבחינה זו, מיפוי והכוונת רגש הופכים ל״שובר שוויון״ משום שהם משנים את גבולות האינטראקציה האנושית, מרחיבים את מרחב הפגיעות הרגשית, מטשטשים את ההפרדה בין אותנטיות

לתגובה מחושבת ומציבים אתגר חדש להבנת רווחה נפשית בעידן של מערכות המסוגלות להיקשר, להרגיע, לשכנע, לעצב תחושת עצמי ולשנות מערכות יחסים בין בני אדם ובינם לבין מכונות.

5. מעבר לתקשורת דבורה (מדוברת): שינוי תשתית האינטראקציה הדיגיטלית לאינטראקציה מיידית וללא תיעוד, המשנה תנאי חשיבה, בקרה והשפעה

המעבר של מערכות בינה מלאכותית מבוססות טקסט אל ממשקי דיבור, ובהמשך אל ממשקי דיבור פרואקטיביים, משנה באופן עמוק את הדרך שבה בני אדם מתקשרים עם טכנולוגיה. אינטראקציות שבעבר דרשו זמן, ניסוח, איסוף מידע או שליטה בממשק מצטמצמות לפקודות קוליות, לשיחות רציפות ולדיאלוגים המתנהלים בזמן אמת. כאשר התשתית הדיגיטלית עוברת מכתיבה לדיבור, היא נשענת על משאב אנושי בסיסי ונגיש יותר, אך גם רגיש יותר, מערכת הדיבור וההקשבה. מערכת זאת היא אינטואיטיבית יותר, כמעט בלתי מודעת, ועלולה ליצור מצב שבו תהליכי חיפוש, קבלת החלטות וחשיפה למידע מתבצעים באופן מהיר, רציף וללא תיעוד מפורט.

השלכותיו של שינוי זה על מוגנות הן רחבות. תקשורת דבורה מייצרת יתרון נגישות ברור, אך בו בזמן בהיעדר תיעוד כתוב, יכולת הביקורת העצמית של המשתמש מצטמצמת: קשה לחזור לאחור, לבחון החלטה, או לשחזר מנגנון שכנוע. בנוסף, כאשר שיחה מתנהלת באופן טבעי בשפה דבורה, אדם נוטה לייחס למערכת תכונות של כוונה, קשב ואמפתיה שאינן קיימות בפועל. מצבי חולשה רגשית, עייפות, לחץ או חוסר ריכוז עלולים להעצים את מידת ההשפעה של המערכת על המשתמש. המעבר לדיבור גם מייצר פגיעות לשונית חדשה: קבוצות בעלות יכולת שפתית מוגבלת, דיאלקטים שאינם נתמכים היטב, או דפוסי דיבור איטיים ומקוטעים (כגון אצל אוכלוסיות מזדקנות) עלולים להיתקל בהבנה חלקית או שגויה מצד המערכת ובכך לספוג עיוות, הדרה או פרשנות שגויה של כוונותיהם. במובנים אלה, המעבר לממשק דבור הוא "שובר שוויון" מפני שהוא משנה את המסגרת הקוגניטיבית שבתוכה מתבצעת האינטראקציה הדיגיטלית: הוא מעביר אותה מן המרחב הכתוב שהוא איטי, מחושב וביקורתי יותר, אל מרחב מהיר, זורם ורגשי, שבו מנגנוני ההגנה של המשתמש פועלים באופן יציב פחות.

6. עירוב פיזי ודיגיטלי במרחב: מעבר מתקשורת מתווכת־מסך, למרחבים פיג׳יטליים רב־שכבתיים המעצבים תפיסה, תנועה ואינטראקציה בזמן אמת

ההתקדמות בטכנולוגיות מחשוב מרחביות מאפשרת יציאה מגבולות המרחב הדיגיטלי אל המרחב הפיזי עצמו, ויצירת מרחבים חדשים שבהם הגבול בין פיזי לדיגיטלי מיטשטש, באמצעות יישומים כמו משקפיים חכמים, רכב אוטונומי, רובוטים הומנואידים ומחשוב לביש. יישומים אלה משנים את האופן שבו המרחב הפיזי נראה ומתפקד, כיצד אנשים נעים בתוכו וכיצד אובייקטים, אנושיים ולא אנושיים, מגיבים אלה לאלה. במרחב כזה פעולות בסיסיות כגון ניווט, זיהוי, התמצאות, תשאול, או קבלת החלטות בזמן אמת הופכות לפעולות משולבות מכונה. התוצאה היא סביבה שמעצבת את התפיסה החושית, את קצב התגובה ואת נקודת המבט של המשתמש, לא רק באמצעות מידע חזותי, אלא באמצעות אינטראקציות חיות עם מערכות פיזיות שפועלות לצידו.

מנקודת מבט של מוגנוּת, מדובר בשינוי עומק. סביבה פיג׳יטלית משפיעה על האופן שבו אנשים תופסים זה את זה, מי מסומן וכיצד, ולצידן נבנה מערך יחסים חדש בין אנשים לבין מכונות פיזיות: מכוניות שמגיבות להולכי רגל, רובוטים שמנהלים שיחה או מבצעים מטלות בבית ובמוסדות, מכשירים לבישים שמנטרים מצב רגשי ופיזי ומשנים התנהגות בהתאם. כל אלה יוצרים תנאים חדשים של השתתפות, נראות, בטיחות והבנת סביבה. המעבר לפיג׳יטליות אינו רק שינוי בתפיסת המציאות אלא **שינוי במבנה של המרחב עצמו** – עיצוב מחדש של תשתיות, סטנדרטים, סיגנלים חברתיים, כללי בטיחות ואופן התנהלות. זהו "שובר שוויון" מפני שהוא מייצר מערכת אקולוגית אנושית־טכנולוגית חדשה, שבה המרחב הפיזי, החברתי והדיגיטלי מתקיימים בו־זמנית ומעצבים אלה את אלה בזמן אמת.

7. ריבוי ייצוגים דיגיטליים מלאים של בני אדם: תאומים דיגיטליים וסוכנות ייצוגית, המאתגרים זהות, נוכחות ואחריות

האפשרות ליצור ייצוגים דיגיטליים המשחזרים קולות, תנועות, דפוסי לשון, העדפות ואפילו היגיון פעולה, הופכת את האדם לישות בעלת מופעים מרובים. תאום דיגיטלי

איננו רק דמות ויזואלית או "אוואטאר"; הוא אוסף של פרמטרים התנהגותיים, רגשיים ופרגמטיים היכולים לבצע משימות, ליזום אינטראקציות, להפיק ידע ולייצר תוצרים בשם האדם או במקומו. כאשר ניתן לשחזר טון דיבור, מחוות גוף, סגנון כתיבה, עומק ידע או תהליכי שיקול דעת אסטרטגי, נוצר ייצוג אנושי בעל יכולת מימוש, כלומר לא רק "מראה" שטוחה של המקור, אלא כזו המסוגלת לפעול כאילו הייתה הוא. מושג הנוכחות האנושית מתרחב: האדם אינו נמצא רק במקום שבו גופו נוכח, אלא גם במקום שבו הייצוג שלו פועל, מגיב ומתעדכן.

מנקודת מבט של מוגנות, מדובר בשינוי עומק במבנה הזהות ובמרחב החברתי. ייצוגים דיגיטליים יוצרים פערים חדשים בין המקור לבין ההעתקה: מי מהשניים נושא כוונה, מי אחראי לתוצאה, כיצד מבחינים בין יוזמה אנושית לבין פעולה שנגזרה מהסטטיסטיקה של דפוסי העבר? במרחבי עבודה, תאומים דיגיטליים עשויים לבצע משימות, להשתתף בהחלטות או לשמש כלי תקשורת, בלי שהמקור מודע לכל פעולה, ובכך לערער מושגים של אחריות, ייצוגיות ואמון. במרחבים חברתיים ייצוגים אלו משנים את האופן שבו אדם מתקבל, נזכר ומזוהה: הקול, הפנים ותכונות האופי הופכים לכלים הניתנים להעתקה, לשימוש ולשינוי. שכפול זהות הוא "שובר שוויון" משום שהוא מטשטש את קו הגבול בין האדם לבין ייצוגיו הדיגיטליים, משנה את האופן שבו נסחרים, מנוהלים ומתווכים קשרים אנושיים, ומחייב מסגרת חדשה להבנת זהות, הסכמה ואחריות בעידן שבו לאדם מזוהה עשויה להיות יותר מגרסה אחת.

"שוברי השוויון" בעידן הבינה המלאכותית

1	העברת פעולות קוגניטיביות למכונה
2	קבלת החלטות אוטונומיות ומרובת סוכנים
3	ערעור מוחלט של יכולת ההבחנה האנושית בין תוכן אותנטי לתוכן סינתטי
4	זיהוי רגשי על ידי מכונה והשפעה על המשתמש
5	מעבר לתקשורת דבורה (מדוברת)
6	עירוב פיזי ודיגיטלי במרחב
7	ריבוי ייצוגים דיגיטליים מלאים של בני אדם

סיכום

בפרק זה הצגנו את מפת הסיכונים המקובלת בספרות המחקרית והרגולטורית בעולם של טכנולוגיות בינה מלאכותית מתקדמות. כן ביקשנו בפרק לשנות את נקודת המבט המקובלת בשיח על בינה מלאכותית: מסיווג של "סיכונים מערכתיים" אל תרגום של הסכנות הטכנולוגיות למופעים אנושיים של פגיעות. לכן הצגנו את שוברי השוויון של המוגנות – סדרת שינויים טכנולוגיים תפעוליים אשר **משנים לתפיסתנו את חוקי המשחק** ביחסי אדם וטכנולוגיה. ההכרה בכך היא נקודת מוצא לפרק הבא, שבו נתחיל למפות את דרכי ההתערבות האפשריות, רגולטוריות, חינוכיות, טכנולוגיות, מוסדיות וקהילתיות, על פי המטריקס המתודולוגי שפיתחנו. זוהי המסגרת שתאפשר לנו לנוע מהבנה תאורטית של סיכונים אל תכנון מעשי של מדיניות מוגנות.

פרק שישי

מנגנוני התערבות לחיזוק מוגנות ולהתמודדות עם פגיעות – תיאור כללי

העיסוק במוגנות דיגיטלית, ובפרט בהגנה על קבוצות פגיעות, אינו חדש. בעשורים האחרונים התפתחו מגוון כלים, תוכניות וגישות שמטרתן לצמצם נזקים שנובעים משימושי אינטרנט וטלפונים חכמים, רשתות חברתיות ומהמרחב הסייברי בכלל. עם השנים גובשו מסגרות רגולטוריות, שופרו תשתיות אכיפה והחלו לפעול תוכניות חינוכיות מקיפות, נבנה ידע מוסדי, רגולטורי וחברתי סביב הסיכונים ונצבר ניסיון פרקטי במניעה, בהגנה, בטיפול ובחקיקה. הופעתה של הבינה המלאכותית מערערת חלק מההנחות בבסיס הכלים הללו ומציבה אתגר חדש להתאמה, הרחבה ואפילו חשיבה מחודשת על דרכי ההתערבות הקיימות.

אומנם עיקר עיסוקו של מחקר זה הוא בארבע קבוצות מובחנות, אבל לא ניתן לעסוק בהן מבלי לעסוק קודם לכן במנגנוני התערבות שעשויים להיות רלוונטיים לכלל האוכלוסייה.

פרק זה מציע מיפוי רחב של מנגנוני התערבות רלוונטיים בעידן הבינה המלאכותית. אנו יוצאים מנקודת ההנחה שלאור ההתפתחות של מערכות בינה מלאכותית, יש לעסוק בשלושה סוגים של דרכי התערבות: כלים קיימים שנוסו בעבר, נותרו רלוונטיים ולעיתים אף דרושים בעוצמה גדולה יותר; דרכי התערבות הנוגעות לאופי המשתנה של הסיכונים, לפרופיל החדש של הפוגעים והנפגעים ולאופן שבו יחסי הכוח, הסמכות והשליטה עוברים שינוי; דרכי התמודדות חדשות המבוססות בעצמן על פיתוח טכנולוגי מתחום הבינה המלאכותית.

מנגנוני ההתערבות המוצעים במחקר זה מבוססים על הכרה בכך שפגיעויות בעידן הבינה המלאכותית אינן נובעות רק מכשל טכנולוגי נקודתי, אלא ממאזן כוחות מבני. במצב הקיים, היחיד – משתמש, אזרח או קבוצה פגיעה – ניצב מול פלטפורמות ומערכות עתירות כוח: הן מחזיקות בדאטה, קובעות את כללי המשחק, שולטות בחשיפה ולעיתים אף מעצבות את יכולתו של הפרט להבין את המערכת ולהתגונן מפניה.

על רקע זה, רבות מן ההתערבויות אינן מבקשות "לתקן" את הטכנולוגיה עצמה, אלא לפעול כמעין "הרחבות" לאזרחים ולקבוצות: להרחיב את יכולתם להשפיע, להתאגד, להפעיל לחץ או להגן על עצמם במרחבים שבהם כוחם היה עד כה מצומצם. ניתן לכנות גישה זו "מוגנות מערכתית" – גישה המתמקדת לא בהגנה נקודתית אלא בהפחתת סיכון מבני, ביצירת איזון יחסי מול גורמים רבי־עוצמה, ובהגנה על קבוצות פגיעות מפני דחיקה, ניצול ושעתוק של הטיות.

תפיסת ההתערבות במחקר זה מבוססת על משולש דינמי שקודקודיו הם רגולציה, חינוך ואוריינות וטכנולוגיה. שלושת הממדים הללו מקיימים יחסי גומלין: ללא מודעות ואוריינות מצד מקבלי ההחלטות והציבור קשה לקדם רגולציה אפקטיבית; ללא רגולציה אין תמריץ מספק לשוק הטכנולוגי לפתח מנגנונים מעודדי מוגנות; וללא מנגנונים כאלה קשה לייצר מודעות ואוריינות משמעותיות בקרב משתמשים. רק שילוב בין שלושת הקודקודים מאפשר יצירת אקוסיסטם של מוגנות שאינו תלוי בשחקן יחיד.

את דרכי ההתמודדות בפרק זה נחלק לשבע קטגוריות עיקריות, המשקפות שילוב של סוגי גישות ושל רמות פעולה – החל בהתערבות פרטנית, דרך הרמה מוסדית, ועד מדיניות רגולטורית או בינלאומית רחבה. קטגוריות אלה הן:

1. **ידע, מחקר ופיתוח חשיבה מוטת עתיד:** התמודדות עם סיכונים מחייבת בסיס ידע מתעדכן, כלים לחיזוי מגמות ופיתוח חשיבה מערכתית המאפשרת זיהוי מוקדם של כשלים ופגיעויות. קטגוריה זו כוללת תמיכה במחקר יישומי, ניטור מתמשך של טכנולוגיות חדשות ופיתוח תרחישים עתידיים המשלבים מומחיות רב־תחומית.

2. **רגולציה:** הפעלת כלים משפטיים להסדרת פעולתה של בינה מלאכותית ויישומיה. רגולציה זו יכולה להיות ישירה, למשל קביעת כללי בטיחות או שקיפות (2א); להתמקד באחריות של פלטפורמות אינטרנט (2ב); או להתבצע ברמה הגלובלית (2ג), על ידי שיתופי פעולה בין מדינות, איגודי תקינה וגופים בינלאומיים.

3. **עיצוב טכנולוגי אחראי:** פיתוח מערכות בינה מלאכותית בשילוב מראש של שיקולים אתיים, בטיחותיים וחברתיים. מדובר בעקרונות כמו "Fairness by Design", "Privacy by Design", נגישות, ניתוב מבוקר של קבלת החלטות ותכנון ממשקים המאפשרים שליטה והבנה אנושית של מערכות חכמות.

4. **חקיקה ושכלול מערכות האכיפה:** יצירת כלים להתמודדות עם מעשים פליליים או עם סוגי נזק חדשים או מועצמים באמצעות חקיקה חדשה (4א), הגברת יכולת האכיפה ושכלול המיומנות של רשויות החוק בהתמודדות עם פשיעה מבוססת טכנולוגיה (4ב).

5. **התערבויות מערכתיות:** הטמעה של תוכניות חוצות־מערכות, כמו הכשרת עובדי ציבור ומערכי חינוך, חיזוק מוסדות רווחה ובריאות נפש (5א) ופיתוח מענים רגשיים וחינוכיים לאוכלוסיות חשופות (5ב) כחלק ממערך כולל של מוגנות.

6. **אוריינות:** תוכניות לאוריינות דיגיטלית, יישומים דיגיטליים, פיתוח חשיבה ביקורתית, פיתוח מנגנוני ויסות רגשי (6א), וכן מתן כלים, סמכויות ומענים מעשיים להורים, מחנכים ואפוטרופוסים, שנדרשים לתווך בין אוכלוסיות רגישות למערכות מורכבות (6ב).

7. **התערבות טכנולוגית מבוססת בינה מלאכותית:** שימוש בטכנולוגיות חדשות כפתרון להתמודדות עם פגיעויות. למשל, מערכות לניטור פוגענות, אודיטינג אוטומטי, כלים להתאמה אישית של הגנות על משתמשים רגישים, בקרה חכמה ועוד.

בתוך כל קטגוריה נציג דוגמאות, ראיות ליעילות ואתגרים אפשריים ביישום, בשימת דגש על ההתאמה בין סוגי הפגיעות, סוגי הטכנולוגיה הרלוונטיים ואופני ההתערבות. חשוב

להדגיש כי הקטגוריות אינן מנותקות זו מזו, במצבים רבים יש לפעול במקביל במספר רמות, ועל יחסי הגומלין הללו נעמוד בהמשך.

כחלק מן המתודולוגיה של המחקר, ולאחר מיפוי הפגיעויות והצגת מגוון רחב של מנגנוני התערבות אפשריים, הוספנו גם שלב של תיעדוף דרכי ההתערבות לפי מידת הישימות שלהן. תיעדוף זה נועד להתמודד עם פער שכיח בשיח על מוגנות בעידן הבינה המלאכותית: הנטייה להציג רשימות ארוכות של פתרונות ראויים מבלי לבחון לעומק את יכולת המימוש שלהם במציאות מוסדית, ארגונית וחברתית מורכבת.

נקודת המוצא שלנו היא שהאפקטיביות של התערבויות בעידן הבינה המלאכותית אינה ניתנת להערכה על פי קריטריון יחיד, ואינה נגזרת באופן ליניארי מחשיבות נורמטיבית בלבד. התערבות עשויה להיות מוצדקת, ערכית ואף הכרחית – אך בלתי ישימה בשלב הנוכחי; ולעומתה, התערבות אחרת עשויה להיות ישימה מאוד אך בעלת השפעה מוגבלת או זמנית. לפיכך פיתחנו מודל תיעדוף הבוחן במקביל עומק השפעה ויכולת מימוש בפועל, ומאפשר ארגון שיטתי של ההמלצות המוצעות במחקר.

תיעדוף דרכי ההתערבות מבוסס על חמישה קריטריונים משלימים, שכל אחד מהם מייצג ממד אחר של יישומיות והשפעה:

1. **ישימות מול שותפים:** קריטריון זה בוחן את מידת התלות של ההתערבות בשיתוף פעולה בין כמה שחקנים: רגולטורים, משרדי ממשלה, רשויות מקומיות, חברות טכנולוגיה, מערכת החינוך, ארגוני חברה אזרחית או גופים בינלאומיים. רבות מן ההתערבויות אינן ניתנות ליישום חד־צדדי, והצלחתן תלויה ביכולת לייצר תיאום, אמון וממשקי עבודה בין גופים בעלי סמכויות, אינטרסים וקצבי פעולה שונים. ככל שהתלות בשותפים מרובה ומורכבת יותר, כך גוברים הסיכונים לעיכוב, שחיקה או כשל יישומי.

2. **ישימות כלכלית וארגונית:** קריטריון זה נוגע למשאבים הנדרשים לצורך יישום ההתערבות, ליכולת הארגונית לשאת בה לאורך זמן ולהשתלבותה במבנים מוסדיים קיימים. גם פתרונות נכונים מבחינה עקרונית עלולים להיכשל אם אינם תואמים יכולות תקציביות, כוח אדם, תהליכי עבודה או תרבות ארגונית. בחינה זו מאפשרת להבחין בין צעדים שניתן להטמיע יחסית במהירות לבין כאלה המחייבים שינוי מבני עמוק או השקעה מתמשכת משמעותית.

3. **מיידיות ההשפעה:** קריטריון זה בוחן את טווח הזמן שבו צפויה ההתערבות להניב השפעה: אם מדובר בצעד המספק מענה מיידי או קצר טווח לפגיעויות קיימות, או בהתערבות שמטרתה השפעה מצטברת וארוכת טווח. בהקשר של סיכונים דינמיים ומתפתחים, יש ערך הן לצעדים היוצרים הגנה מהירה והן לבניית תשתיות עומק, גם אם השפעתם מתבררת רק לאורך זמן.

4. **עומק ההשפעה:** קריטריון זה עוסק בשאלה עד כמה ההתערבות נוגעת בשורשי הבעיה. אם היא פועלת בעיקר ברמת הסימפטומים וההתנהגויות, או משנה מאפיינים מבניים של המערכת, כגון מאזן הכוחות בין משתמשים לפלטפורמות, תמריצים כלכליים, נורמות תכנוניות או חלוקת אחריות מוסדית. קריטריון זה מאפשר לזהות אילו התערבויות עשויות לשמש מנופי שינוי מבניים, גם אם יישומן מורכב יותר.

5. **ניסיון יישומי קודם בארץ או בעולם:** קריטריון זה בוחן אם ישנם תקדימים, פיילוטים או יוזמות דומות שכבר נוסו והניבו תוצאות. קיומו של ניסיון קודם מפחית אי־ודאות, מאפשר למידה מצטברת ומגדיל את סיכויי ההטמעה. עם זאת, היעדר ניסיון כזה אינו שולל התערבות אלא מסמן אותה כחדשנית יותר – ובה בעת עתירת סיכון.

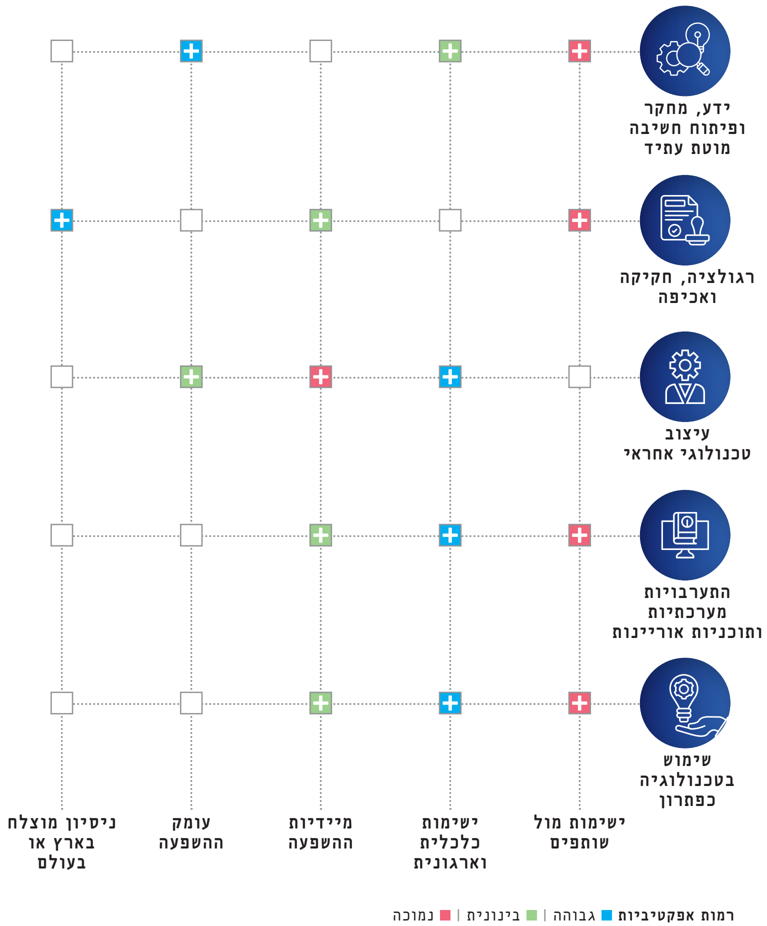
המשמעות של מודל התיעדוף אינה בהעדפת קריטריון אחד על פני אחרים, אלא ביכולת לבחון את יחסי הגומלין ביניהם. התערבות בעלת עומק השפעה גבוה אך ישימות מוגבלת עשויה להיות חשובה ככיוון אסטרטגי לטווח הארוך, ואילו התערבות ישימה ובעלת השפעה מיידית אך עומק מבני מוגבל יכולה לשמש מענה ביניים הכרחי. השילוב בין הקריטריונים מאפשר לארגן את דרכי ההתערבות לא כצעדים מבודדים, אלא כאשכולות של התערבויות בעלות פרופיל השפעה דומה, הפועלות בקצבי זמן שונים ומשלימות זו את זו.

באופן זה, התיעדוף המוצע מונע הן הטיה לטובת פתרונות פופולריים אך חלשים מבנית, והן נטייה להעדיף פתרונות עמוקים אך בלתי ישימים בשלב הנוכחי. הוא מכיר במתח שבין אפקטיביות תאורטית לבין מגבלות מוסדיות ומאפשר ניהול מודע של מתח זה במסגרת קביעת מדיניות.

חשוב להבהיר כי מודל התיעדוף אינו כלי דירוג ליניארי, והוא לא נועד להפיק רשימה אחת של המלצות מובילות. הוא גם אינו מחליף שיקול דעת מקצועי, ציבורי או פוליטי, ואינו מכריע בשאלות ערכיות או תקציביות. מדובר בכלי אנליטי שנועד לתמוך בתהליכי קבלת החלטות באמצעות מיפוי שיטתי של יתרונות ומגבלות תלויי הקשר.

הוא מבטא תפיסה שלפיה מוגנות בעידן הבינה המלאכותית לא תושג באמצעות כלי אחד או מהלך יחיד אלא באמצעות אקוסיסטם של התערבויות משלימות, שיפעלו ברמות שונות ובטווחי זמן שונים, מתוך הכרה במגבלות המערכת ובאחריות המשותפת של כלל השחקנים המעורבים.

הערכת האפקטיביות של דרכי ההתערבות



הערה: הסימונים במפה להמחשה בלבד.

התמודדות אפקטיבית עם אתגרי המוגנות בעידן הבינה המלאכותית מחייבת גישה אינטגרטיבית, הרואה את קטגוריות ההתערבות שיוצגו כמערכת אקולוגית שלמה, ולא כפתרונות העומדים בפני עצמם. במקום תפיסה המתמקדת ב"שכבת הגנה" יחידה, גישה מערכתית מזהה את נקודות החוזק והחולשה של כל התערבות ויוצרת מערך הגנה רב־שכבתי: מניעתי, מגיב ומשקם.[109]

נעיר כבר כעת כי בישראל כיום אין גוף אחד שמופקד באופן מובהק על תחום המוגנוּת הדיגיטלית, וקל וחומר על מוגנות בעידן הבינה המלאכותית. האחריות מפוצלת בין משרדי ממשלה, רגולטורים, מערכת החינוך, רשויות רווחה והורים. מצב זה מוביל לא פעם לפערים, חפיפות או היעדר תיאום בין גופים. תכלול אפקטיבי של תחום המוגנות מחייב גוף בעל סמכות ויכולת פעולה אופרטיבית, שיהיה מסוגל לשלב בין רגולציה, חינוך, עיצוב טכנולוגי והנחיה מוסדית. בהקשרים שונים עלו הצעות להקים רשות למוגנות דיגיטלית, למנות רפרנטים ייעודיים בתוך משרדים קיימים (כגון משרד החינוך, הבריאות והרווחה), או להקים ועדת תכלול בין־משרדית בעלת סמכות ביצועית. ראוי לשקול בחיוב הקמה של גוף כזה, ולו כיחידת מטה, שיאפשר גיבוש מדיניות עקבית, תיאום בין־גופי ופיתוח כלי פעולה מותאמים לאוכלוסיות השונות ולסיכונים החדשים.

קיימים כמובן קשרי גומלין והשלמה בין שכבות ההתערבות השונות – אינדיווידואלית, מוסדית ומערכתית – והן אינן פועלות בחלל הריק. התערבות מערכתית קובעת את המסגרת החוקית והערכית שבתוכה פועלים מוסדות; מוסדות מתרגמים מדיניות לאופני פעולה קונקרטיים; והפרט מושפע מהשניים, אך גם משמש גורם יוזם, מתריע ומשתתף. כך למשל, מדיניות רגולטורית שקובעת סטנדרטים של מוגנות עבור מערכות חינוך תאפשר פיתוח של תוכניות מניעה בתוך בתי הספר, שמיושמות בפועל על ידי צוותי הוראה ומחנכים. מנגד, משוב שמתקבל מההורים ומהתלמידים עצמם יכול להוביל לשינוי בהנחיות המערכתיות. לכן, חשיבה על מוגנות בעידן הבינה המלאכותית מחייבת הבנה מערכתית הרואה בכל שכבה לא רק תחום פעולה נפרד אלא חלק ממערך דינמי של השפעה הדדית ואחריות משותפת.

למשל:

109 Multilayer Framework for Good Cybersecurity Practices for AI (European Union Agency for Cybersecurity [ENISA], 2023)

1. **ידע מחקרי מזין רגולציה אפקטיבית:** מחקר מתקדם מספק את הבסיס הראייתי לפיתוח מסגרות רגולטוריות מדויקות ומבוססות מדע. רגולציה ללא בסיס מחקרי איתן מייצרת פתרונות לא אפקטיביים ולעיתים אף מזיקים.[110]

2. **רגולציה מקדמת עיצוב טכנולוגי אחראי:** מסגרות רגולטוריות המטילות אחריות על מפתחים ופלטפורמות יוצרות תמריצים משמעותיים לעיצוב אחראי. חברות מגיבות לתמריצים רגולטוריים בהטמעת גישות עיצוב אתיות, בעיקר כאשר אלו נתמכות בתקנים ובמנגנוני אכיפה.[111]

3. **אוריינות וחוסן רגשי מועצמים על ידי כלים עבור הורים:** תוכניות אוריינות ופיתוח חוסן רגשי מקבלות משנה תוקף כאשר הן משולבות עם העצמה הורית.[112]

4. **חדשנות התערבותית מחזקת יכולת התערבות מוסדית:** כלים מבוססי AI לזיהוי מצוקה וסיכון משפרים את היכולת של מערכות תמיכה מוסדיות לספק מענה מדויק ומהיר.[113]

הלוח שלהלן ממפה את קטגוריות ההתערבות לפי שלוש רמות פעולה עיקריות: פרטנית, מוסדית ומדינתית־בינלאומית. מיפוי זה מדגיש כי התמודדות אפקטיבית עם פגיעוּת בעידן הבינה המלאכותית מחייבת שילוב של מענים ממוקדים (למשל, חינוך לאוריינות דיגיטלית או הגנה רגשית על משתמשים) עם תשתיות מוסדיות (כגון רגולציה פנים־מדינתית, תכנון מערכות חינוך או מערכי אכיפה) ועם מדיניות־על ברמה הלאומית והגלובלית (לרבות פיתוח רגולציה חוצת גבולות או שיתופי פעולה טכנולוגיים בינלאומיים). הבנה זו מאפשרת לא רק לאפיין את סוג ההתערבות הרצוי אלא גם לזהות מי אחראי עליו, מה מנגנון הביצוע שלו ומהו טווח ההשפעה שלו – ובכך להבטיח מוגנוּת מותאמת, אפקטיבית ורב־שכבתית.

110 *Principles of Evidence-Based Policymaking* (Urban Institute, 2018)

111 *Technology Policy: Responsible Design for a Flourishing World* (World Economic Forum, 2024)

112 *Understanding AI and Helping Youth Make the Most of It*, MediaSmarts

113 Xianghe Liu, Jiaqi Xu & Tao Sun, *PsyCounAssist: A Full-Cycle AI-Powered Psychological Counseling Assistant System*, arXiv (Apr. 23, 2025)

קטגוריית ההתערבות	רגולציה/בינ"ל	מערכות מדינה	רמה פרטנית
1. ידע ומחקר	X	X	X
2א. רגולציית AI	X		
2ב. אחריות פלטפורמות	X	X (אכיפה על ידי מוסדות רגולטוריים)	
2ג. ממשל גלובלי	X		
3. עיצוב טכנולוגי	X (תקינה)	X (ממשקי חברות טכנולוגיה והמדינה)	
4א. חקיקה פלילית	X		
4ב. אכיפה	X (שיתופי פעולה בינ"ל)	X (רשויות אכיפה)	
5א. מערכות תמיכה	X	X	X
5ב. התערבות רגשית	X	X (מערכת החינוך, רווחה, בריאות ובריאות הנפש)	X
6א. אוריינות		X (מדיניות חינוך)	X
6ב. כלים לבעלי סמכות		X	X
7. התערבות מבוססת AI	X	X	X

לאחר סקירת דרכי ההתערבות בפרק הבא, בחלק הבא של המחקר, ניישם את המסגרת המוצעת: נבחן כיצד מאפייני הקבוצות השונות מעצבים פרופילי פגיעות ייחודיים, כיצד הם מצטלבים עם סוגי הטכנולוגיות והסיכונים, ואילו אשכולות התערבות הם בני־יישום בכל הקשר.

פרק שביעי

מנגנוני התערבות לחיזוק מוגנות ולהתמודדות עם פגיעות – תיאור פרטני

לאחר שתיארנו את ההיגיון המארגן של מוגנות מערכתית ואת עקרונות התיעדוף לפי הישימות ועומק ההשפעה, בפרק זה נציג את דרכי ההתערבות עצמן. לא מדובר ברשימת פתרונות במובן הפשוט, אלא במיפוי של מנגנונים הפועלים בקצבים שונים וברמות אחריות שונות, חלקם ותיקים ודורשים התאמה והעצמה בעידן הבינה המלאכותית, חלקם חדשים ונולדו מתוך שינוי במאזן הכוחות ובפרופיל הסיכונים, וחלקם הם מענים טכנולוגיים המבוססים על הבינה המלאכותית עצמה. בכל קטגוריה נציג דוגמאות, ראיות ליעילות ואתגרים צפויים ביישום, כדי לאפשר לקורא להבין לא רק מה ראוי לעשות, אלא גם מה אפשרי, באילו תנאים ובאיזה שילוב.

1. ידע ומחקר

הבסיס להתערבות אפקטיבית הוא הבנה מעמיקה של הסיכון והפגיעות באמצעות מחקרים שונים. על אף הקשיים המובנים בתרגום ממצאי מחקר לפרקטיקות התערבות יעילות, במיוחד כאשר המחקר מתבצע בתנאי מעבדה, אין תחליף לעבודת חקר ומיפוי. מנקודת מבט ישראלית, יש להבחין בין מחקר שיתקיים גם כך ברמה הגלובלית וניתן יהיה לאמץ בישראל, לבין מחקר הנדרש בשל מאפייני פגיעות ייחודיים בישראל, ולפיכך יש מקום לעודד ולממן מחקרים ייעודיים לאוכלוסייה בישראל. המורכבות רבת־הפנים של שימושי בינה מלאכותית והשפעתם מחייבת גם כך ליצור מאגדי מחקר המשלבים חוקרים ממגוון תחומים, בדומה ליוזמת Stanford Human-Centered AI, המאגדת חוקרים מתחומי האתיקה, ההנדסה, הפסיכולוגיה ומדעי החברה לחקר השפעות הבינה המלאכותית.[114] לכן, יש לשאוף שקבוצות מחקר ישראליות יפעלו לצד או עם מאגדים בינלאומיים, כגון AI Global Governance Forum שהוקם ב־2024[115] ומאגד מדינות, חברות ומומחים, או ההאב הבינלאומי לבינה מלאכותית של ה־OECD.[116]

האתגר המרכזי ביחס למחקר הוא הקצב המהיר של ההתפתחויות הטכנולוגיות, המקשה על ביצוע מחקר שיטתי ארוך טווח. אכן, מחקר הוא בסיס הכרחי להתערבות יעילה, אבל האתגר הקריטי נותר הפער בין קצב המחקר לקצב האימוץ הטכנולוגי. מחקרים מראים שטכנולוגיות חדשות מוטמעות בקרב אוכלוסיות פגיעות הרבה לפני שהשפעותיהן נבחנו.[117] למשל, בעוד תחום ההשפעות של מדיה חברתית על הקשרים חברתיים למיניהם זכה למחקר נרחב, הידע על השפעות של בינה מלאכותית גנרטיבית, סוכני AI ומחשוב חישתי על קבוצות שונות עדיין נמצא בחיתוליו.[118]

114 Artificial Intelligence Index Report 2024 (Stanford Institute for Human-Centered AI, 2024)

115 *AI Governance Alliance*, World Economic Forum

116 *Policies, Data and Analysis for Trustworthy Artificial Intelligence*, OECD.AI: The OECD Artificial Intelligence Policy Observatory

117 Richard J. Chung, Janet B. Lee, Jesse M. Hackell et al., *Confidentiality in the Care of Adolescents: Technical Report*, 153(5) Pediatrics (2024)

118 Usman Anwar, Abulhair Saparov, Javier Rando et al., *Foundational*

לשם התמודדות עם בעיה זו יש לממן ולפתח מתודולוגיות למעקב אחר השפעות של טכנולוגיות ולימוד בזמן אמת. דוגמה מעניינת היא מעבדת MediaWell של הרשת האקדמית האמריקאית למדעי החברה, המייצרת סקירות אקדמיות וסינתיזה של מחקרים, וגם מנגישה אותם לציבור, כדי לבחון באופן מהיר יותר ממחקר אקדמי רגיל כיצד להפחית את השפעתם של נרטיבים כוזבים או ביטויי שנאה ברשתות, לבנות מוסדות דמוקרטיים עמידים ולהעצים קהילות.[119]

אפשרות נוספת היא לבנות מתודולוגיות של מחקרי "עיצוב ספקולטיבי" שתכליתן לדמיין אינטראקציות בינאישיות או בין אנשים למכונות, לסביבתם או למציאות, של טכנולוגיות שעדיין לא נכנסו לשימוש רחב.[120] מתודולוגיות אלה יוכלו לשמש גם כדי להגביר אוריינות בקרב מקבלי החלטות, וגם כדי ליצור המלצות ראשוניות למדיניות.

אתגר מרכזי נוסף הוא פער הידע המוסדי: לרוב, הגופים המפתחים והמפעילים את הטכנולוגיה מחזיקים בידע רב יותר ממוסדות המחקר, הציבור או הרגולטורים. פער זה מעכב הן את היכולת לפקח והן את האפשרות לפתח התערבויות אפקטיביות.[121] אחד הקשיים במחקר בנוגע להשפעות של טכנולוגיה בכלל הוא הגישה למידע שנמצא בידי

Challenges in Assuring Alignment and Safety of Large Language Models, arXiv preprint:2404.09932 (2024); John Burden, *Evaluating AI Evaluation: Perils and Prospects*, arXiv preprint:2407.09221 (2024); John Burden, Manuel Cebrian, & Jose Hernandez-Orallo, *Conversational Complexity for Assessing Risk in Large Language Models*, arXiv preprint:2409.01247 (2024); Oliver Guest, Michael Aird, & Seán Ó hÉigeartaigh, *Safeguarding the Safeguards: How Best to Promote AI Alignment in the Public Interest*, arXiv preprint:2312.08039 (2023); Ross Gruetzemacher, Alan Chan, Kevin Frazier et al., *An International Consortium for Evaluations Of Societal-Scale Risks from Advanced AI*, arXiv preprint:2310.14455 (2023); Toby Shevlane, Sebastian Farquhar, Ben Garfinkel et al., *Model Evaluation for Extreme Risks*, arXiv preprint:2305.15324 (2023); The Era of Global Risk – An Introduction to Existential Risk Studies (SJ Beard, Martin Rees, Catherine Richards, & Clarissa Rios Rojas eds., Open Book, 2023); *Safety, Security and Risk*, Leverhulme Centre for the Future of Intelligence

119 *About MediaWell*, MediaWell, Social Science Research Council

120 ראו למשל, אלטשולר ואח', IPPSO, לעיל ה"ש 42.

121 Kate Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (Yale Univ. Press, 2021)

חברות הטכנולוגיה, שאינן מסכימות לתת גישה לחוקרים למערכות ונתונים פנימיים. דוגמה לכך היא תוכנית המחקר Social Science One בשיתוף חברת פייסבוק שתוצאותיה היו מעורבות עקב מגבלות אלו.[122] שושנה זובוף (Zuboff) מציינת בספרה[123] שפערי כוח מבניים מגבילים את יכולתם של חוקרים עצמאיים לאסוף נתונים משמעותיים על מערכות דאטה שחברות טכנולוגיה רואות כקניין שלהן, ומחקרים נוספים הראו שללא גישה לאלגוריתמים ולנתוני אימון קשה לפתח כלי הגנה אפקטיביים.[124] לעומת זאת, הצלחה משמעותית יותר נרשמה בפרויקט AI360 של מכון אלן לבינה מלאכותית, שהצליח ליצור שיתוף פעולה מובנה בין חברות, אקדמיה ורגולטורים.[125] אכן, יש להניח שבמקום שבו חברות הענק הבינלאומיות אינן מעניקות גישה לנתונים ברמה הגלובלית, גישה כזאת לא תינתן גם בישראל וקשה לראות כיצד מדינת ישראל בעצמה תצליח היכן שמדינות גדולות או האיחוד האירופי אינם מצליחים.

2. רגולציה

2א. רגולציה על בינה מלאכותית

רגולציה היא פעילות שיטתית הנעשית בידי גוף מינהלי ברשות המבצעת של המדינה העוסק בגיבוש ובהוצאה לפועל של מדיניות להכוונת שווקים. מדינות זו מבקשת לקדם את האינטרס הציבורי, החברתי או הכלכלי, והיא נעשית כלפי גופים הפועלים בשוק או

122 Social Media and Democracy: The State of the Field, Prospects for Reform (Joshua A. Tucker & Nathaniel Persily eds., Cambridge University Press, 2020)

123 Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (PublicAffairs, 2019), פרקים 6, 9 ו־18.

124 ראו למשל את המחקר העוסק בניסיון להתחקות אחר הטיות אלגוריתמיות – Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proceedings of Machine Learning Research 77–91 (2018)

125 *Asta AutoDiscovery: An AI System that Explores your Datasets while you Sleep – Identifying The Research Questions you Never Thought to Ask*, Allen Institute for AI

בסקטור מסוים, באמצעות נורמות משפטיות כופות או וולונטריות, לצד הפעלתם של מערכי פיקוח לציות ויישומם של מנגנוני אכיפה.[126]

הרגולציה על בינה מלאכותית מתמודדת עם אתגר כפול: מצד אחד, הצורך להגן על משתמשים מפני סיכונים; מצד שני, החשש מהאטת החדשנות או מפגיעה ביתרונות טכנולוגיים. רגולציה אפקטיבית צריכה לאזן בין הגנה על פרטיות, בטיחות ואמינות לבין יצירת מרחב למחקר ופיתוח. רגולציה מסורתית, המתבססת על חקיקה מפורטת, מאשרת מראש (אקס אנטה), מתקשה להדביק את קצב השינויים הטכנולוגיים. המענה לכך הוא גישות רגולטוריות גמישות, מבוססות עקרונות ותוצאות, המשלבות אחריות תאגידית, רגולציה עצמית מונחית ומנגנוני פיקוח דינמיים, כמפורט להלן.[127]

א. **רגולציה מבוססת סיכון (Risk-Based Regulation):** גישה המדרגת את רמת הפיקוח והחובות הרגולטוריות בהתאם לרמת הסיכון של המערכת. ה־AI Act האירופי הוא דוגמה מובילה לסוג זה של רגולציה.[128] החוק המגדיר ארבע רמות סיכון – מסיכון בלתי קביל (אסור לשימוש) ועד לסיכון מינימלי – ודרישות שונות לכל רמה.[129] גישה דומה אומצה בחוק הבינה המלאכותית של קליפורניה.[130]

ב. **מנגנוני פיקוח טכנולוגי:** גופי פיקוח ייעודיים המתמחים בהערכת סיכונים טכנולוגיים, בדיקות ואישורים ברמה הבינלאומית, וכוללים מומחיות טכנית ומגוון בעלי עניין. כך

126 שרון ידין "מהי רגולציה? הצעה להגדרה בעקבות מופעים שונים של המונח בחקיקה הישראלית" **חוקים בקצרה** 6, 1, 9 (2014).

127 *Artificial Intelligence Risk Management Framework*, National Institute of Standards and Technology (2023)

128 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024 O.J. (L 2024/1689) 1

129 עמיר כהנא ותהילה שוורץ אלטשולר **אדם, מכונה, מדינה: לקראת אסדרה של בינה מלאכותית** 130 (המכון הישראלי לדמוקרטיה, 2023) (להלן: כהנא ושוורץ אלטשולר, **אדם ומכונה**).

130 Cal. Assemb. Bill No. 331, 2023-2024 Reg. Sess. (2024) (Automated Decision Tools)

למשל, בבריטניה הוקמה רשות בינה מלאכותית וטכנולוגיות דיגיטליות (Digital Regulation Cooperation Forum – DRCF) שמטרתה לפקח על פיתוח ויישום בטוח.[131] ברמה הגלובלית, קיימות הצעות להקמת IAEA for AI כבסיס לפיתוח יכולות פיקוח בינלאומיות,[132] וכן בוועידת הבטיחות העולמית לבינה מלאכותית בסיאול הוצע להקים Accelerated AI Safety Mechanism.[133] גם הפרוטוקול לשיתוף פעולה באכיפת חוקי הגנת המידע (GDPR) באירופה, המתבצע באמצעות ה־European Data Protection Board,[134] הוא דוגמה למודל אכיפה משותף שניתן להרחיבו לתחום ה־AI.

ג. **חובות שקיפות ודיווח:** רגולציה המחייבת שקיפות ביחס לאופן הפעולה של מערכות, מקורות מידע, וכן דיווח על אירועי כשל או סיכונים. "תקנות ChatGPT" באיטליה,[135] והצו הנשיאותי האמריקאי מתקופת ביידן בארצות הברית,[136] כוללות דרישות שקיפות מקיפות ומנגנוני דיווח.

ד. **"רגולציה אקוסיסטמית":** גישה המטפלת בכלל שרשרת הערך של מוצרים מבוססי בינה מלאכותית – מפתחים, משווקים, מפיצים ומשתמשים – ולא רק במוצר עצמו או בפלטפורמה. חוק השירותים הדיגיטליים האירופי (DSA) הוא דוגמה לגישה כזו, המטילה חובות על כל השחקנים בשרשרת הערך הדיגיטלית.[137]

131 *Digital Regulation Cooperation Forum* (DRCF)

132 Seokki Cha, *Towards an International Regulatory Framework for AI Safety: Lessons from the IAEA's Nuclear Safety Regulations*, 11 Humanities & Social Sciences Communications, 506 (2024)

133 Seoul Statement of Intent toward International Cooperation on AI Safety Science, AI Seoul Summit 2024 (Annex) (UK Department for Science, Innovation and Technology, May 21–22, 2024)

134 *Tasks and duties*, European Data Protection Board

135 *Guidelines on the use of Cookies and Other Tracking Tools*, Italian Data Protection Authority (2023)

136 Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023)

137 DSA, art. 28, Regulation 2022/2065, 2022 O.J. (L 277) 1

ה. **רגולציה אדפטיבית (Adaptive Regulation)**: מנגנונים המסוגלים להתאים את עצמם במהירות לשינויים טכנולוגיים, בכלל זה ארגזי חול רגולטוריים.[138]

ו. **מודל "תקני ביצוע טכנולוגיים"**: במקום לקבוע כיצד טכנולוגיה צריכה לפעול, הרגולציה קובעת תוצאות רצויות ומדדי ביצוע, ומאפשרת למפתחים לבחור כיצד להשיגם.[139]

ז. **מערכות של תמריצים**: שימוש בתמריצים כלכליים וטכנולוגיים במקום איסורים, כמו מס על כריית נתונים מילדים או הטבות למערכות המוכיחות יתרון בהגנה על קבוצות פגיעות.

ח. **רגולציה רכה (Self-Regulation; Code of Conduct)**: מנגנון שבו תעשיות וארגונים מפתחים סטנדרטים אתיים והתנהגותיים פנימיים במטרה לקדם התאמה מהירה לשינויים טכנולוגיים ולהפחית את התלות בהתערבות ממשלתית. היתרונות הבולטים של רגולציה זו הם למשל הגמישות, היכולת להגיב במהירות לשינויים והפחתת עומס רגולטורי ממשלתי.[140] עם זאת, לרגולציה רכה יש גם חסרונות משמעותיים: היעדר מנגנוני אכיפה מחייבים עלול להוביל להיעדר יישום אפקטיבי של הכללים, במיוחד כאשר קיימת התנגשות בין אינטרסים עסקיים למחויבויות אתיות.[141] כמו כן, היעדר פיקוח חיצוני מעלה חשש לפערים באחריותיות ובשקיפות.

ט. **רגולציה טכנולוגית**: רגולציה דוגמת הערכות השפעה אלגוריתמיות (algorithmic impact assessments) מציעה מודל שבו הכלים הטכנולוגיים עצמם משמשים אמצעי בקרה רגולטורי. לדוגמה, בשנת 2023 שודרג תהליך ה־AIA בקנדה כדי לכלול הערכת השפעות פרטנית על קבוצות אוכלוסייה פגיעות, בדגש על מגדר וגיל, במסגרת קידום

138 Cary Coglianese, *The Trump Administration and the Dismantling of the Administrative State*, 78(5) Public Administration Review 754–759 (2018)

139 תהילה שוורץ אלטשולר "'מה שטוב לאירופה' – גם בבינה מלאכותית" **TheMarker** (9.9.2024).

140 The Role of Self-Regulation in Addressing Consumer Issues, 7–9 (OECD Digital Economy Papers No. 254, 2014)

141 *The Case for Teaching Industry Self-Regulation*, BBB National Programs (2023)

עקרונות של אחריותיות ואתיקה בפיתוח מערכות אוטומטיות.[142] עם זאת, מחקרים עדכניים מעידים על פערים ביישום ההערכות בפועל, הן מבחינת היעדר סטנדרטיזציה בתעשייה והן בשל קשיים טכניים הנובעים ממחסור במשאבים ובמומחיות.[143]

י. **רגולציה משתפת (Co-Governance):** מודל שבו קובעי מדיניות משתפים את הקהילות הרלוונטיות, בכללן משתמשים, נציגי ציבור וארגונים אזרחיים, בתהליך קביעת הכללים והתקנות. יתרונה העיקרי של רגולציה משתפת טמון בהגברת הלגיטימציה הציבורית והאמון במדיניות, שכן היא מאפשרת הכללת מגוון רחב של קולות. נוסף על כך, שיתוף הקהילה תורם לפיתוח רגולציה מותאמת לערכים ולצרכים המקומיים.[144] עם זאת, תהליכים משתפים כרוכים לעיתים בזמן ממושך ובמורכבות רבה, ולעיתים טומנים בחובם קושי בהשגת קונצנזוס. קיימת גם סכנה ממשית לחוסר איזון בכוח, שבמסגרתו קבוצות חזקות יכתיבו את התוצאה על חשבון קבוצות מוחלשות.[145]

12. רגולציה בינלאומית וטרנס־לאומית

רגולציה בינלאומית משמעה פיתוח מסגרות ערכיות ועקרונות מוסכמים המנחים את פיתוח ה־AI ברמה הבינלאומית והצהרה עליהן. דוגמה בולטת היא "הצהרת בלצ'לי" (Bletchley Declaration) שנחתמה ב־2023 על ידי 28 מדינות ומתווה עקרונות משותפים לבטיחות AI. גם עקרונות ה־AI של ה־OECD והצהרת ברצלונה של אונסק"ו על אתיקה של AI משמשים מסגרות סטנדרטים בינלאומיות.[146] מנגד, הסכמות בינלאומיות לגבי עקרונות

142 *AI Governance on the Ground: Canada's Algorithmic Impact Assessment Process and Algorithm Has Evolved*, Montreal AI Ethics Institute (Fab. 3, 2025)

143 Amar Ashar, Karim Ginena, Maria Cipollone et al., *Algorithmic Impact Assessments at Scale: Practitioners' Challenges and Needs*, 2 J. Open Tech. Stud. 1 (2023)

144 Sean Bradley & Israa H. Mahmoud, *Strategies for Co-Creation and Co-Governance in Urban Contexts*, 8 Urban Sci. 9 (2024)

145 *Co-Governance and the Future of AI Regulation*, Chapter Three, 138 Harv. L. Rev. 1 (2025)

146 *The Bletchley Declaration by Countries Attending the AI Safety Summit*, GOV.UK (1-2 November 2023); Principles on Artificial Intelligence (OECD, 2024); Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021)

נוטות להישאר ברמה ההצהרתית, וקיים סיכון ל"דילול" סטנדרטים כדי להשיג הסכמה רחבה.[147] בנוסף, האפקטיביות של מסגרות בינלאומיות תלויה במידה רבה בהתאמה לתנאים ויכולות מקומיות, וגישה "אחידה לכולם" עלולה להיכשל.[148]

סוג משלים של רגולציה בינלאומית מתבטא במאמצי "הרמוניזציה רגולטורית" לתיאום דרישות רגולטוריות ולצמצום פערים בין מדינות. ה־Global Partnership on AI (GPAI) מקדם יוזמות להרמוניזציה של סטנדרטים לפיתוח ויישום אחראי של AI,[149] בעוד האיחוד האירופי ובריטניה הקימו "מסדרון רגולטורי" (AI regulatory corridor) לתיאום אכיפה ושקיפות.[150] דוגמאות נוספות הן ברית של מדינות מובילות עם יכולת אכיפה, כמו "האמנה לבטיחות AI" של ארגון ה־G7[151] המשלבת תמריצים כלכליים וטכנולוגיים; מנגנוני בקרה שיש בהם הרבה בעלי עניין ומשלבים מדינות, חברות, חברה אזרחית ומומחים בפיקוח על מערכות AI, כמו Global AI Governance Initiative של הפורום הכלכלי העולמי;[152] וכמובן המשך פיתוח תקנים וולונטריים כמו אלו של ISO/IEC,[153] שיש להם פוטנציאל לשמש שפה משותפת לאסדרה בינלאומית.

147 Nathalie A. Smuha, *Beyond the Individual: Governing AI's Societal Harm*, 10 Internet Pol'y Rev. 1, 4 (2021)

148 Tehilla Shwartz Altshuler & Rajan Luthra, *"Mumbai and Tel Aviv Effect": An Alternative to the "Bandwagon Effect" of Brussels and Washington in Global AI Regulations*, Israel Democracy Institute (May 30, 2024)

149 Scaling Responsible AI Solutions: Challenges and Opportunities (Global Partnership on Artificial Intelligence, 2023) (להלן: Scaling Responsible AI Solutions)

150 אין מדובר במונח רשמי. ברם, חוק הבינה המלאכותית (AI Act) כולל הקמת גופים כמו AI Office ו־European Artificial Intelligence Board, שמטרתם לתאם את יישום החוק בין המדינות החברות, ובישראל מסמך עקרונות מדיניות רגולציה ואתיקה בבינה מלאכותית מציע הקמת מוקד ידע ממשלתי לשם תיאום בין רשויות המדינה השונות.

151 *Labour and Employment Ministers Ministerial Declaration*, G7-G20 Documents Database (2024)

152 AI Governance Alliance, World Economic Forum

153 ISO/IEC JTC 1/SC 42, Artificial Intelligence – Published Standards Catalogue, ISO

מחקרים מראים שרגולציה אפקטיבית ואחריות פלטפורמות יכולות לשפר את המוגנות הדיגיטלית. למשל, הטלת אחריות משפטית על פלטפורמות ביחס לתוכן מזיק הובילה לשיפור משמעותי במנגנוני ניטור וסינון.[154] עם זאת, רגולציה איננה נטולת קשיים. היא עלולה לחזק את חברות הטכנולוגיה הגדולות ולהקשות על כניסת מתחרים קטנים. כמו כן, פערי יכולות רגולטוריים בין מדינות ואזורים יוצרים "משחק דילמת האסיר הגלובלית", שבו מדינות מתחרות על משיכת חברות באמצעות הפחתת הרגולציה.

ואכן, עם כניסתו של דונלד טראמפ לתפקיד נשיא ארצות הברית בשנת 2017 ניכרה תפנית במדיניות הפנים, שהתבטאה בין היתר בהתקפות על גופים מדעיים ומקצועיים פדרליים. אחד המקרים הבולטים היה הפחתת התמיכה והערעור על סמכות המכון הלאומי לתקנים ולטכנולוגיה (NIST), גוף שאחראי על קביעת תקנים מדעיים וטכנולוגיים, ושנחשב במשך עשורים לסמכות בלתי תלויה.[155] תחת ממשל טראמפ נעשו ניסיונות להגביל את השפעת ה־NIST, במיוחד בהקשרים שנגעו לרגולציה סביבתית וטכנולוגית, מתוך העדפת העמדות של גופי התעשייה על פני מחקר מדעי עצמאי.[156] במקביל, טראמפ גם הוביל מהלך דרמטי בזירה הבינלאומית נגד ארגונים כמו ארגון הבריאות העולמי (WHO), כאשר בשנת 2020, בשיא מגפת הקורונה, האשים את הארגון בהטיה פוליטית לטובת סין והודיע על ניתוק הקשרים והפסקת המימון האמריקאי לארגון,[157] החלטה שהיו לה השלכות עמוקות על שיתוף הפעולה הגלובלי בתחום הבריאות.[158]

במקביל להתקפות על מוסדות מדעיים ומקצועיים, בתקופת נשיאותו השנייה ממשל טראמפ מקדם מגמה רחבת היקף של נסיגה מרגולציה (deregulation) בכל תחומי הממשל הפדרלי. טראמפ חתם על צווים נשיאותיים שהגבילו משמעותית את היכולת של הסוכנויות הפדרליות להוציא רגולציות חדשות, ובפרט הורה על מדיניות two-out, one-in,

154 Yassine Lefouili & Leonardo Madio, *The Economics of Platform Liability*, 53 Eur. J.L. & Econ. 319 (2022)

155 *Attacks on Science*, Union of Concerned Scientists (Mar 20, 2025)

156 Michael Halpern, *The Trump Administration's War on Science*, Scientific American (Oct. 2018)

157 Dan Diamond, *Trump Administration Cuts Funding for WHO*, Politico (Apr. 14, 2020)

158 Lawrence O. Gostin et al., *US Withdrawal from WHO Is Unlawful and Threatens Global and US Health and Security*, The Lancet (2020)

כלומר ביטול שתי תקנות קיימות על כל רגולציה חדשה שתאושר.[159] מדיניות זו הובילה להיחלשות רגולציה בתחומים קריטיים כגון הגנת הסביבה, בטיחות עבודה, בריאות הציבור ורגולציה פיננסית.[160] השיח הרגולטורי תחת ממשל טראמפ עוצב בעיקר מתוך תפיסה שלפיה רגולציה היא נטל מיותר על המגזר העסקי, ויש לצמצם את המעורבות הממשלתית כדי לקדם צמיחה כלכלית, גם במחיר של פגיעה בערכים של הגנה חברתית, בריאותית ואקולוגית.[161] בפסגת הבינה המלאכותית בפריז בפברואר 2025, סגן נשיא ארצות הברית ג'יי. די. ואנס הזהיר מפני רגולציה מופרזת בתחום הבינה המלאכותית וטען שכל רגולציה היא יתר־רגולציה. ואנס הביע ביקורת על חוק השירותים הדיגיטליים של האיחוד האירופי וטען כי הוא מטיל "רגולציה מסיבית" ומוביל ל"פיקוח על מה שמכונה מידע שגוי", דבר שלדבריו פוגע בחופש הביטוי.[162]

2ג. רגולציה על פלטפורמות

רגולציה על פלטפורמות היא דבר שכיח בחמש השנים האחרונות, בייחוד באיחוד האירופי ובמדינות דומות לו, אך היא אינה קיימת כיום בישראל.[163]

מידת האפקטיביות של הפלטפורמות באכיפת הכללים הנוגעים לתוכן בלתי חוקי משתנה מפלטפורמה לפלטפורמה.[164] נוסף על כך, בחלק ממדינות העולם מפעילי פלטפורמות דיגיטליות ממילא אינם נחשבים בעלי אחריות להפצה של תוכן בלתי חוקי או לנזקים

159 Exec. Order No. 13,771, Reducing Regulation and Controlling Regulatory Costs, 82 Fed. Reg. 9339 (Feb. 3, 2017)

160 Jody Freeman & Sharon Jacobs, *Structural Deregulation and the Administrative State*, 135 Harvard Law Review 585–674 (2021)

161 שם.

162 Ivana Kottasová, *J.D. Vance Warns Against Regulating AI at Paris Summit*, CNN (Feb. 11, 2025)

163 תהילה שוורץ אלטשולר, אסף וינר ואייל זילברמן **מתווה לאסדרת רשתות חברתיות בישראל** (הצעה לסדר 51, המכון הישראלי לדמוקרטיה, מרץ 2023) (להלן: מתווה לאסדרת רשתות חברתיות בישראל).

164 Steve Wood, Impact of Regulation on Children's Digital Lives 34 (Digital Futures for Children & 5Rights Foundation, 2024)

העשויים להיגרם למשתמשים מתוכן מזיק שמשתמשים אחרים מפיצים.[165] גם במדינות שבהן אין פטור מפורש כזה, עקרון הפטור מאחריות הקבוע בסעיף 230 ל־Communications Decency Act האמריקאי[166] יצר סטנדרט המשפיע על דרכי פעולתן של הפלטפורמות. אין ספק שהתמריץ של הפלטפורמות לפעול להסרת תכנים מושפע ישירות מרמת האחריות המשפטית שיש להן לתכנים האלה. בהיעדר תמריץ, ויתרה מזו, אם יש פטור מאחריות, האכיפה דלילה למדי.[167]

מנגד, היתרון של רגולציה על פלטפורמות טמון ביכולתן לייצר אפקט מערכתי רחב. שכן, ישנו מגוון רחב של פלטפורמות, הכולל מדיה חברתית, שירותי ענן ומודלים של שפה, שעליהן יש להטיל אחריות.

מודל מוצע בהקשר זה הוא מודל אחריות מדורגת של פלטפורמות: מודל המטיל עליהן חובות המשתנות בהתאם למאפייני המשתמשים והשירותים. כך למשל, ה־UK Online Safety Act מציב חובות מוגברות על פלטפורמות המשרתות קטינים,[168] והצעת החוק Child Exploitation and Artificial Intelligence Expert Commission Act of 2024 בארצות הברית הממליצה על הקמת ועדה מיוחדת שתפקידה לחקור כיצד נעשה שימוש בטכנולוגיות בינה מלאכותית בהקשר של פגיעה בילדים (כולל הפקת והפצת חומרים פוגעניים).[169]

כמו כן, ישנו צורך להגדיר מהן פלטפורמות בעולם הבינה המלאכותית, הנובע מההשפעה המכרעת שלהן על עיצוב האקוסיסטם הדיגיטלי, קביעת סטנדרטים טכנולוגיים ותיווך בין משתמשים, מפתחים ורגולטורים.[170] פלטפורמות AI, כגון מודלים גדולים (Large

165 בארצות הברית חסינות זו מוענקת מכוח סעיף 230 ל־Decency Communication Act, ובאיחוד האירופי מכוח דירקטיבה: EC2000/31/ Directive. נוסח כל אחד מדברי החקיקה האלה שונה, אך הם דומים במהותם: על מפעיל אתר לא תוטל אחריות פלילית לתוכן לא חוקי שמעלה גורם שלישי, דהיינו משתמש שאינו קשור למפעיל עצמו.

166 Communications Decency Act, 47 U.S.C. §230 (2018)

167 מתווה לאסדרת רשתות חברתיות בישראל, לעיל ה"ש 163, בעמ' 22.

168 Online Safety Act 2023, c. 50, §12 (UK)

169 ראו Child Exploitation and Artificial Intelligence Expert Commission Act of 2024, סעיפים 1–4.

170 Martin Kenney & John Zysman, *The Rise of the Platform Economy*, 32(3) Issues in Science and Technology 61–69 (2016)

(Language Models), שירותי ענן לאימון מודלים, או מרקטפלייסים לאלגוריתמים, פועלות כמתווכים ריכוזיים השולטים בגישה למשאבים קריטיים, נתונים וכוח חישובי.[171] הן מעצבות את תנאי השימוש בטכנולוגיה ומכתיבות את גבולות החדשנות. לכן, הגדרה מדויקת של "פלטפורמה" בהקשר זה מאפשרת לרגולטורים ולחוקרים לבחון את יחסי הכוח, הסיכונים והחובות של שחקנים מרכזיים בזירה זו, כמו גם לפתח מדיניות מותאמת למאפיינים הייחודיים של התיווך הפלטפורמטי.

אתגר זה ניתן להדגמה באמצעות צמד חוקים שנחקקו במדינת קליפורניה בארצות הברית באוקטובר 2025 ונכנסו לתוקף בינואר 2026. החוק האחד – Senate Bill 243[172] – מיועד להגנה על משתמשים מפני סיכונים במגע עם בני לוויה מבוססי בינה מלאכותית (companion chatbots) שנוצרו כדי לספק קשר חברתי ורגשי. החוק מגדיר בני לוויה מבוססי בינה מלאכותית כ"מערכת בינה מלאכותית עם ממשק שפה טבעית המסוגלת לספק תגובות אדפטיביות ואנושיות לצרכים חברתיים של המשתמש, כולל שימור קשר לאורך זמן והפגנת תכונות אנתרופומורפיות". החוק מטיל שורה של חובות על מפעילי בני לוויה כאלה, למשל גילוי שמדובר בדמות לא אנושית, איסור לפעול ללא פרוטוקול מניעה למקרים של פגיעה עצמית, תזכורות אוטומטיות לקטינים לקחת הפסקה ואיסור על תוכן מיני או הצעות למעשים מיניים, וחושף אותם לקנסות כבדים על הפרות. החוק האחר – Senate Bill 53[173] – מטיל חובות על מודלים עתירי עוצמה המכונים "Frontier Models" (ומוגדרים לפי עוצמת המחשוב שלהם), כגון שקיפות (כיצד הוטמעו סטנדרטים לאומיים ובינלאומיים במסגרות האימון), דיווח לרשויות על אירועי בטיחות,

171 Jean-Christophe Plantin, Carl Lagoze, Paul N. Edwards, & Christian Sandvig, *Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook*, 20(1) New Media & Society 293-310 (2018)

172 Senate Bill No. 243 Chapter 677, An act to add Chapter 22.6 (commencing with Section 22601) to Division 8 of the Business and Professions Code, relating to artificial intelligence. [Approved by Governor October 13, 2025. Filed with Secretary of State October 13, 2025]

173 Senate Bill No. 53 Chapter 138, An act to add Chapter 25.1 (commencing with Section 22757.10) to Division 8 of the Business and Professions Code, to add Section 11546.8 to the Government Code, and to add Chapter 5.1 (commencing with Section 1107) to Part 3 of Division 2 of the Labor Code, relating to artificial intelligence. [Approved by Governor September 29, 2025. Filed with Secretary of State September 29, 2025]

הגנה על חושפי שחיתויות החושפים סיכונים במודלים ומנגנון עדכון שנתי לחקיקה עצמה בהתאם להתפתחויות טכנולוגיות.

המודל הדו־ראשי בקליפורניה משית רגולציה על המודל החזק ביותר (frontier AI), המתייחס לחברות ענק כגון גוגל, מטא ו־OpenAI, ובמקביל על המודל הקרוב ביותר למשתמש (companion AI), כדי לבנות את קווי המתאר הראשונים של רגולציית אינטראקציות אנושיות עם מכונות.

מעבר לכך, ההגדרה של פלטפורמות בעולם הבינה המלאכותית חשובה לצורך גיבוש רגולציה אפקטיבית. בלי הבחנה בין פלטפורמות לבין מוצרים או שירותים רגילים, יש קושי להטיל אחריות על שחקנים שיש להם שליטה מהותית בתשתיות טכנולוגיות ובהפצת תכנים או יישומים מבוססי בינה מלאכותית.[174] לדוגמה, פלטפורמות עשויות לקבוע אילו מודלים זמינים לציבור, כיצד נאספים נתונים לצורך אימון, ומהם כללי האתיקה והבקרה החלים על מפתחים. רגולציה המתעלמת מהמעמד הייחודי של פלטפורמות עלולה להחמיץ את מוקדי הסיכון וההשפעה האמיתיים של מערכות בינה מלאכותית.[175] הגדרה מושגית מדויקת תורמת אפוא ליצירת מסגרות רגולטוריות המתמודדות עם אתגרי שקיפות, אחריותיות ושליטה ריכוזית בתחום הבינה המלאכותית.

3. עיצוב טכנולוגי אחראי

מוגנות אינה יכולה להסתמך רק על רגולציה חיצונית או על כישורי האוריינות של המשתמשים, היא מתחילה ב"קוד עצמו", בשלב התכנון והפיתוח של הטכנולוגיה. עיצוב טכנולוגי אחראי (Responsible Tech Design) מתמקד בהטמעת עקרונות של בטיחות, אתיקה ופרטיות כחלק בלתי נפרד מתהליך הפיתוח – החל בקונספט הראשוני, דרך האפיון והבנייה וכלה בהטמעה.

174 Tarleton Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media (Yale University Press, 2018)

175 Karen Yeung, *Algorithmic Regulation: A Critical Interrogation*, 12(4) Regulation & Governance 505–523 (2018)

אתגר מרכזי הוא המתח בין ערכים ותמריצים: מפתחי טכנולוגיה, וחברות הענק בפרט, פועלים במערכת תמריצים המעודדת אופטימיזציה של מדדים עסקיים (מעורבות משתמשים, זמן מסך, הכנסות), שלעיתים מתנגשים עם ערכים של הגנה, פרטיות ורווחה. בנוסף, קיים פער בין הקצב המהיר של פיתוח טכנולוגי לבין התהליכים האיטיים יותר של הערכת סיכונים והטמעת אמצעי הגנה.[176]

דוגמאות:

א. **עיצוב לפרטיות ולבטיחות (Privacy & Safety by Design)**: גישה המטמיעה שיקולי פרטיות ובטיחות כבר בשלבי התכנון הראשוניים של מוצר. עקרונות ה־Privacy by Design[177] שפיתחה אן קבוקיאן (Cavoukian) מספקים בסיס לגישה זו, בעוד מודל Safety by Design של הנציבות האוסטרלית לבטיחות אלקטרונית מספק מסגרת ייחודית לעיצוב בטוח.[178]

ב. **מסגרות לפיתוח אתי ואחראי של AI**: מתודולוגיות פיתוח המשלבות אתיקה לכל אורך מחזור החיים של המוצר. למשל, מודל Responsible AI Framework של מיקרוסופט מציע תהליכים מובנים להטמעת שיקולים אתיים.[179]

ג. **בדיקות פגיעות וסיכונים (Vulnerability Testing)**: תהליכי בדיקה מובנים לזיהוי פגיעויות אפשריות, בדגש על קבוצות פגיעות, כמו למשל שיטת ה־Red-teaming לבינה מלאכותית של Anthropic[180] וה־Vulnerability Assessments של Global Partnership on AI.[181]

176 *Responsible AI Standard: IEEE 2863-2023*, Institute of Electrical and Electronics Engineers (2023)

177 אלטשולר ואח', IPPSO, לעיל ה"ש 42, בעמ' 9.

178 *Safety by Design*, eSafety Commissioner

179 *Microsoft Responsible AI: Principles and Approach*, Microsoft

180 Deep Ganguli, Liane Lovitt, Jackson Kernion et al., *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, arXiv preprint:2209.07858 (2022)

181 Scaling Responsible AI Solutions, לעיל ה"ש 149.

ד. **שליטה ובחירה למשתמש:** עיצוב המעניק שליטה בהירה ומשמעותית למשתמשים על חוויית השימוש, מידע אישי והגדרות פרטיות. עקרונות ה־Meaningful Human Control של נאט"ו[182] וגישת ה־User Choice & Control של NIST מציעים כלים לעיצוב ממשקים המעצימים משתמשים.

ה. **תיעוד מודלים וטכנולוגיה שקופה:** פיתוח כלים לתיעוד ולהסברתיות של מערכות מורכבות. גישת ה־Model Cards שפיתחה גוגל[183] וה־Transparency Toolkit של מטא[184] מציעות תקנים לתיעוד והסברת מודלים.

היתרון המשמעותי של עיצוב טכנולוגי אחראי הוא ביכולתו למנוע בעיות מראש, במקום לנסות לתקן אותן בדיעבד. מדובר בגישה יעילה יותר מבחינת משאבים ואפקטיביות (cheaper to build in than bolt on). מחקרים מראים שעיצוב טכנולוגי אחראי יכול לשפר את רמת המוגנות של משתמשים, וממשקים המעניקים שליטה משמעותית למשתמשים הובילו להחלטות מושכלות יותר ולירידה בהתנהגות מוטת סיכון.[185]

עם זאת, חוקרים כמו פרנק פסקואלה (Pasquale)[186] מדגישים כמה אתגרים: ראשית, המודלים העסקיים של חברות טכנולוגיה רבות מבוססים על מעקב ואיסוף נתונים, דבר היוצר התנגשות מובנית עם עקרונות פרטיות; שנית, עיצוב אחראי עלול להאט חדשנות ולהעלות עלויות פיתוח, דבר היוצר התנגדות; ושלישית, קיים חשש מ"הלבנה אתית" (ethics washing), שימוש בשיח אתי כמסווה ליישום חלקי או שטחי בלבד.

182 Christopher Miller, Mark Draper, Jurriaan van Diggelen et al., *Meaningful Human Control of AI-Based Systems Workshop: Technical Evaluation Report, Thematic Perspectives and Associated Scenarios* (NATO Sci. & Tech. Org., June 2023)

183 Huanming Fang & Hui Miao, *Introducing the Model Card Toolkit for Easier Model Transparency Reporting*, Google Research Blog (July 29, 2020)

184 Igor Tufanov, Karen Hambardzumyan, Javier Ferrando et al., *LM Transparency Tool: Interactive Tool for Analyzing Transformer Language Models*, arXiv preprint:2404.07004 (2024)

185 Lucas Elbert Suryana, Sina Nordhoff, Simeon C. Calvert et al., *User Perception of Partially Automated Driving Systems: A Meaningful Human Control Perspective on the Perception among Tesla Users*, arXiv preprint:2402.08080 (2024)

186 Frank Pasquale, New Laws of Robotics: Defending Human Expertise in the Age of AI (Harvard University Press, 2020)

4. התאמת הדין הפלילי ושכלול מערכות אכיפה

בינה מלאכותית מביאה איתה דפוסי פגיעה חדשים הדורשים חידוש של קטגוריות פליליות קיימות, למשל ניצול מיני באמצעות תוכן סינתטי, עבירות "בהלה" והונאות מתוחכמות. בהקשר זה ניתן לחשוב על חקיקה ייעודית לפשיעה מבוססת בינה מלאכותית, כלומר התאמת החקיקה הפלילית לסיכונים חדשים, במיוחד בתחומים של תוכן סינתטי פוגעני, דיפ־פייק והונאות מבוססות AI. ה־Defiance Act בארצות הברית,[187] המטיל עונשים מוגברים על שימוש בדיפ־פייק למטרות פליליות, וחוק התוכן הסינתטי בבריטניה,[188] האוסר על יצירה והפצה של פורנוגרפיה סינתטית ללא הסכמה, הם דוגמאות לכך. בנוסף, ניתן לפתח מסגרות חקיקה גמישות כגון הגדרות פליליות מבוססות עקרונות ותוצאות במקום טכנולוגיות ספציפיות.

הצורך בחיזוק יכולות אכיפה של הדין הקיים גם הוא אחד מעמודי התווך של יצירת מוגנות. בהקשר זה ניתן לדבר על –

א. **חיזוק יכולות חקירה וזיהוי:** פיתוח כלים טכנולוגיים, שיטות חקירה ותקנים לזיהוי וטיפול בראיות דיגיטליות. תוכנית Enhanced Digital Evidence Collection של האינטרפול[189] מציעה תקנים לאיסוף ראיות דיגיטליות. במקביל, יוזמות כמו AI Forensics Hub של Europol מפתחות כלים לזיהוי תוכן סינתטי ומקורות פגיעה.

ב. **שיתוף פעולה בינלאומי בתחום הפלילי:** הסכמים והסדרים להתמודדות עם פשיעת AI חוצת גבולות. אמנת בודפשט המעודכנת לפשיעת סייבר[190] משמשת בסיס לשיתוף פעולה

187 DEFIANCE Act of 2024, S. 3696, 118th Cong. (2024)

188 Online Safety Act 2023, c. 50 (UK)

189 *Enhanced Digital Evidence Collection: Global Standards* (INTERPOL Innovation Centre, 2023); *Child Sexual Abuse Material Detection: AI-Powered Platform Evaluation* (INTERPOL Crimes Against Children Unit, 2023); *Project Aether: Ai-Enabled Crime Pattern Detection* (INTERPOL Innovation Centre, 2024)

190 *Second Additional Protocol to the Budapest Convention on Cybercrime and Cross-Border Access to Electronic Evidence* (Eurojust, 2022)

בינלאומי, בעוד יוזמת Global AI Crime Task Force של UNDOC יוצרת מערך לתיאום בין רשויות אכיפה.[191]

ג. **קידום מומחיות ייעודית במערכות משפט ואכיפה:** פיתוח מומחיות, הכשרות וגופים ייעודיים להתמודדות עם פשיעה טכנולוגית כמו למשל יחידות המומחים של ה־FBI לתוכן סינתטי והקמת בתי משפט ייעודיים לפשיעה דיגיטלית בסינגפור.

ד. **שותפויות ציבוריות־פרטיות באכיפה:** שיתופי פעולה בין גופי אכיפה לחברות טכנולוגיה לזיהוי ומניעת שימושים פליליים. יוזמות כמו Tech Against Trafficking ו־Meta-NCMEC Collaboration למאבק בניצול ילדים מציגות מודלים של שיתוף פעולה ביחס לאיתור תוכן פוגעני.

ה. **מערכי הכשרה מתמשכים:** פיתוח תוכניות הכשרה מתמשכות לשופטים, תובעים וחוקרים. האקדמיה האירופית לאכיפת חוק (CEPOL) פיתחה מודל הכשרה דינמי המתעדכן מדי רבעון.

ו. **אכיפה אזרחית** בדמות תביעות של משתמשים מכוח דיני החוזים והחוזים האחידים; תובענות ייצוגיות; ותביעות מכוח חקיקה רגולטורית כגון חוק הגנת הצרכן וחוק הגנת הפרטיות.

5. התערבויות מערכתיות

התמודדות אפקטיבית עם סיכוני בינה מלאכותית מחייבת מערכת ציבורית רחבה הכוללת שירותים תומכים, מוקדי סיוע, מרכזי טיפול וגורמים מקצועיים. מערכות אלה מספקות:

- מניעה ממוקדת סיכון
- "רשת ביטחון" למקרים של פגיעה
- מרחב לבניית חוסן אישי וקהילתי
- העברת ידע וסיוע הדדי

אתגר מרכזי הוא פערי ההתמחות והידע בקרב אנשי מקצוע ביחס לסיכונים חדשים. אנשי חינוך, בריאות הנפש, רווחה ועבודה סוציאלית הנמצאים בחזית ההתמודדות עם פגיעות לרוב אינם מקבלים הכשרה מספקת להבנת האופי הייחודי של פגיעות דיגיטליות והקשר הטכנולוגי.[192] לפיכך יש צורך לפתח אמצעים כגון אלה:

1. **מרכזי דיווח וסיוע ייעודיים:** הקמת מוקדים ומרכזים לדיווח, תמיכה וסיוע במקרים של פגיעה דיגיטלית. מוקד 105 הוא דוגמה למרכז המשלב יכולות משטרתיות, טיפוליות וחינוכיות.
2. **הכשרת אנשי מקצוע:** פיתוח תוכניות הכשרה לגורמי מקצוע בתחומי החינוך, הרווחה, הבריאות ובריאות הנפש. לדוגמה, ה־Partnering With Parents Program באוסטרליה וה־Digital First Aid Kit לעובדים סוציאליים מציעים הכשרות ממוקדות.[193]
3. **מנגנוני תמיכה קהילתיים ותוכניות עמיתים:** יצירת רשתות תמיכה קהילתיות וקבוצות עמיתים. תוכנית Digital Champions של Connected Communities Initiative מכשירה מנהיגים קהילתיים כ"אלופי דיגיטל" המסייעים בהעלאת מודעות ותמיכה בנפגעים.
4. **שירותי ייעוץ ותמיכה מונגשים:** פיתוח שירותים נגישים וידידותיים למשתמש בפלטפורמות מקוונות ובקהילה. מודל ה־Be Internet Awesome של גוגל[194] מציע כלים ידידותיים לילדים ולהורים, בעוד ברשתות אינסטגרם וסנאפצ'ט הוטמעו כלים לדיווח ישיר על תוכן פוגעני ולקבלת סיוע.[195]
5. **רשתות קהילתיות ייעודיות:** פיתוח קהילות חוסן ייעודיות לקבוצות בסיכון. היתרון של מערכות תמיכה ציבוריות וקהילתיות נעוץ ביכולתן להתאים את המענה לצרכים הספציפיים של אוכלוסיות שונות. עם זאת, קיימים אתגרים משמעותיים בהטמעת

192 *Global Education Monitoring Report 2023: Technology in Education: A Tool on Whose Terms?* 210 (UNESCO, 2023)

193 *Partnering with Parents: Building Quality Relationships that Benefit Children*, Victoria Department of Education and Training (Sept. 2024); *Digital First Aid Kit*, The Digital First Aid Kit (2025)

194 *Be Internet Awesome: Empowering Kids to be Safe, Confident Explorers of the Online World*, Google

195 *Safety Center*, Instagram

שירותים אלו: פערי תקציב, יכולת להגיע לאוכלוסיות מודרות, פער דיגיטלי המקשה על נגישות לשירותים מקוונים, וקושי לפתח מומחיות מקצועית לאור השינויים הטכנולוגיים המהירים. בנוסף, חוקרים הצביעו על הצורך בהתאמה תרבותית של שירותים, במיוחד בחברות מגוונות. כיוונים נוספים הם למידה משותפת ופיתוח אסטרטגיות קהילתיות כמו יוזמת Collective Digital Literacy מבית Center for Humane Technology, המקדמת מודלים של אוריינות משותפת.

מחקרים מדגישים את החשיבות של שילוב של מענים טכנולוגיים וקהילתיים. סוניה ליווינגסטון ואליסיה בלום־רוס (Livingstone & Blum-Ross) הראו שקהילות בעלות "נאמני דיגיטל", אנשים בעלי ידע טכנולוגי ויכולת תמיכה, הציגו רמות גבוהות יותר של דיווח והתמודדות עם פגיעות מקוונות.[196] עם זאת, מחקרים אחרים מצביעים על פערים משמעותיים: ב־2023, ה־Internet Watch Foundation מצא שפחות מ־30% מהפגיעות המקוונות מדווחות לשירותי תמיכה, עם שיעורים נמוכים במיוחד בקרב אוכלוסיות מודרות.[197] בנוסף, אחרים הראו שקיימים פערים משמעותיים ביעילות המענים בהתאם למשאבים קהילתיים ומדינתיים.

סוג אחר של התערבויות הוא התערבויות רגשיות־חינוכיות מערכתיות מבקשות לחזק משאבים פנימיים, לפתח מודעות רגשית ולבנות "שרירי חוסן" כנגד השפעות שליליות של טכנולוגיה. אתגר מרכזי הוא "הסתגלות נירולוגית" – המוח האנושי מסתגל לגירויים דיגיטליים ויוצר תלות ודפוסי שימוש שקשה לשנות באמצעות התערבויות מסורתיות. מחקרים מראים שהשפעות נירו־פסיכולוגיות של טכנולוגיה דורשות גישות התערבות ייחודיות, המשלבות הבנה של מנגנוני תגמול, קשב וויסות רגשי.[198]

196 Sonia Livingstone & Alicia Blum-Ross, Parenting for a Digital Future: How Hopes And Fears About Technology Shape Children's Lives (Oxford University Press, 2020)

197 *Annual Report on Online Harm Reporting Patterns*, Internet Watch Foundation (2023)

198 Jasmina Wallace, Elroy Boers, Julien Ouellet et al., *Screen Time, Impulsivity, Neuropsychological Functions and Their Relationship to Growth in Adolescent Attention-Deficit/Hyperactivity Disorder Symptoms*, 13 Sci. Rep. 18108 (2023)

1. **פיתוח מערכים וכלים להגברת המודעות העצמית** ביחס לשימוש בטכנולוגיה והשפעותיה הרגשיות. התוכנית Digital Wellbeing של Google ומודל Mindful Technology Use של Center for Humane Technology מציעים כלים לפיתוח מודעות לדפוסי שימוש והשפעותיהם.[199]

2. **התערבויות לבניית חוסן רגשי־דיגיטלי:** פיתוח יכולות התמודדות עם לחצים, מניפולציות והשפעות שליליות של טכנולוגיה. התוכנית Digital Resilience Toolkit של Young Minds ו־Digital Resilience Framework של UK Council for Internet Safety מפתחים "שרירי חוסן" ספציפיים למרחב הדיגיטלי.[200]

3. **מערכי התערבות כיתתיים ומערכתיים בבתי ספר:** שילוב תוכניות מובנות במסגרות חינוכיות. תוכנית למניעת Cyberbullying של שפ"י בישראל ו־Social Media Literacy Project של Stanford University מציעים מערכי שיעור והתערבויות מערכתיות.[201]

4. **גישות לפיתוח יחסים דיגיטליים בריאים:** הקניית כלים ליצירת הרגלים ודפוסי יחסים מאוזנים במרחב הדיגיטלי. המודל Healthy Digital Relationships של Common Sense Media והתוכנית Digital Balance מציעים כלים לאיזון בין עולמות.[202]

5. **התערבויות קהילתיות ומשפחתיות:** מעבר מהתמקדות ביחיד למודלים מערכתיים־משפחתיים. תוכנית Family Digital Wellbeing של Stanford Social Media Lab מציעה מודל התערבות משפחתי.[203]

היתרון של התערבויות רגשיות־חינוכיות הוא בהתמקדותן בממד האנושי והחווייתי של האינטראקציה עם טכנולוגיה, שלעיתים קרובות נזנח בשיח הטכנולוגי או הרגולטורי.

199 Arielle Pardes, *The Research Behind Google's New Tools for Digital Well-Being*, Wired (May 8, 2018); *Control Your Tech Use*, Center for Humane Technology

200 *Digital Resilience Framework* (UK Council for Internet Safety [UKCIS], 2019); *Digital Resilience Toolkit*, Internet Matters

201 Brooke Donald, *Stanford Education Scholars to Create Resources to Help Young People Spot Fake Information Online*, Stanford Graduate School of Education (Apr. 27, 2018)

202 *Digital Citizenship Curriculum*, Common Sense Media; *Finding My Media Balance*, Common Sense Media

203 *Adolescents and Social Media Initiative*, Stanford Social Media Lab

מחקרים מראים שהתערבויות רגשיות־חינוכיות יכולות להיות אפקטיביות. אנליזה מטה־מחקרית של פטי ולקנברג (Valkenburg) הראתה שתוכניות המתמקדות בהעלאת מודעות רגשית לשימוש במדיה חברתית הצליחו להפחית תסמיני חרדה ודיכאון בקרב מתבגרים, עם אפקט משמעותי במיוחד במקרים של שימוש אינטנסיבי.[204] מחקר אחר הראה שתוכניות מובנות לחיזוק חוסן דיגיטלי הגבירו את היכולת לזהות מניפולציות רגשיות ולהתמודד איתן.[205]

עם זאת, קיימות מגבלות משמעותיות: ראשית, התערבויות מסוג זה מתקשות להתמודד עם הא־סימטריה בכוח הנמצא בידי משתמשים בודדים לכוחם של תאגידי הטכנולוגיה; שנית, התערבויות אלו דורשות משאבים משמעותיים ויישום עקבי לאורך זמן; שלישית, תוכניות רבות לא נבדקו באופן מקיף או שהן מבוססות על תאוריות שטרם הוכחו אמפירית; רביעית, מחקרים הראו שהאפקטיביות של תוכניות במערכת החינוך מוגבלת ללא שותפות של הורים ושינויים סביבתיים.

6. אוריינות

המונח "אוריינות מדיה" מתייחס ליכולת של אזרחים לנתח, להעריך, ליצור ולצרוך תוכן בצורה מיודעת וביקורתית, או כמכלול של מיומנויות המאפשרות מעורבות ביקורתית עם מסרים המיוצרים ומופצים באמצעי התקשורת.[206] "אוריינות מידע" שמה דגש על כישורים הקשורים בניווט בסביבת המידע וחיפוש וניהול מידע. שני אלה לרוב נכללים במונח "אוריינות דיגיטלית", שבעבר תיאר היכרות עם טכנולוגיות דיגיטליות נפוצות והבנה בסיסית שלהן, אך עם הזמן התרחב, והיום כלולות בו יכולות הקשורות באינטראקציה עם תכנים ופעולות בפורמטים דיגיטליים, כך שניתן לומר שהגדרתו היא מכלול

204 Patti M. Valkenburg, *Social Media Use and Well-Being: What We Know and What We Need to Know*, 45 Current Opinion in Psychology 101294 (2022)

205 Sarah K. Schäfer, Lisa von Boros, Lea M. Schaubruch et al., *Digital Interventions to Promote Psychological Resilience: A Systematic Review and Meta-Analysis*, 7 NPJ Digit. Med. 30 (2024)

206 Monica Bulger & Patrick Davison, *The Promises, Challenges, and Futures of Media Literacy*, 10(1) Journal of Media Literacy Education, 1–21 (2018)

היכולות הטכניות, הקוגניטיביות, החברתיות, האזרחיות והיצירתיות שמאפשרות לַפּרט גישה למוצרים מבוססי טכנולוגיה דיגיטלית, הבנה ביקורתית שלהם ואינטראקציה איתם.[207]

גוף מחקר משמעותי מתמקד בבדיקה של הידע הקיים בעניין ניסיונות לקידום אוריינות מדיה כאמצעי לביסוס צריכה ביקורתית של מידע, בין השאר כערובה להתמודדות עם מידע כוזב,[208] בשלושה שלבים אסטרטגיים – התערבויות חינוכיות וחיזוק מיומנויות להתמודדות עם מידע כוזב לפני החשיפה למידע; עידוד שינוי התנהגותי בזמן החשיפה למידע; ואסטרטגיות הפרכה והפצת מידע מתקן לאחר החשיפה למידע.[209]

מחקרים מראים שתוכניות אוריינות יכולות לשפר את המוגנות בהקשרים מסוימים. מחקר הערכה של תוכנית MediaWise מצא שמשתתפים שעברו תוכניות אוריינות מדיה הציגו שיפור של 41% ביכולת לזהות תוכן מניפולטיבי.[210] באופן דומה, מחקר של אונסק"ו הראה ששילוב תוכניות אוריינות בתוכנית הלימודים הרגילה הביא לשיפור משמעותי במודעות תלמידים לסיכוני פרטיות.[211]

207 לסקירת מושגים אלה ראו סיגל בן עמרם **אוריינות דיגיטלית: אוגדן סקירות קצרות שהוגשו ללשכת המדען הראשי במשרד החינוך** (המדען הראשי, 2022); יעל רם ואסף וינר **אסטרטגיות וכלים מבוססי ראיות לבניית חוסן אזרחי מפני מידע כוזב וסכנותיו בישראל** 7–8 (איגוד האינטרנט הישראלי, פברואר 2026). כן ראו EDMO Guidelines for Effective Media Literacy Initiatives (European Digital Media Observatory – EDMO, 2024)

208 Megan Boler, Hoda Gharib, Barbara Perry et al., *Promoting Mis/Disinformation Literacy Among Adults: A Scoping Review of Interventions and Recommendations*, 25(8) Communication Research (2024); Ellen Droog, Ivar Vermeulen, Dian van Huijstee et al., *Combatting the Misinformation Crisis: A Systematic Review of the Literature on Characteristics and Effectiveness of Media Literacy Interventions*, 0(0) Communication Research (2025); *European Democracy Shield: Empowering, Protecting and Promoting Strong and Resilient Democracies Across The EU* (IP/25/2660) (European Commission, 2025)

209 Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M. Herzog et al., *Toolbox of Individual-Level Interventions Against Online Misinformation*, 8 Nature Human Behaviour, 1044–1052 (2024)

210 *MediaWise for Seniors: Self-Directed Fact-Checking Course*, The Poynter Institute

211 *Global Education Monitoring Report 2023*, לעיל ה"ש 192.

עם זאת, מחקרים אחרים מצביעים על מגבלות כמו למשל "אפקט שחיקה" – ירידה ביכולת זיהוי תוכן מניפולטיבי כעבור זמן, במיוחד לנוכח טכנולוגיות משתכללות. קיים גם פער בין ידע פורמלי ליישום מעשי. אנשים מפגינים הבנה טובה של סיכונים אך מתקשים ליישם אסטרטגיות הגנה במצבים אמיתיים.[212] לבסוף, לא ניתן להתעלם גם מ"אפקט עומס", ריבוי מידע ואזהרות המוביל לתשישות ולהתעלמות מסיכונים.[213]

היתרון של תוכניות אוריינות הוא ביכולתן להעניק למשתמשים כלים ארוכי טווח להתמודדות עם מגוון סיכונים, כולל אלה שטרם זוהו. עם זאת, כפי שמציינות בויד וליווינגסטון (Boyd & Livingstone),[214] קיימות מגבלות מובנות: ראשית, גישת האוריינות עלולה להעביר יתר אחריות ליחיד, במקום לטפל בגורמי סיכון מערכתיים; שנית, האפקטיביות שלה מוגבלת ביחס לטכנולוגיות שנועדו מראש לעקוף מנגנוני הגנה קוגניטיביים; ושלישית, קיים אתגר בהפיכת ידע תאורטי לפרקטיקה יומיומית. לכן גישה רחבה יותר, המשלבת בין אוריינות אישית לבין רפורמות מערכתיות, ומציבה את האוריינות כחלק ממערך הגנה רב־שכבתי ולא כפתרון בלעדי, הגיונית יותר.[215]

אוריינות דיגיטלית היא מרכיב מרכזי ביצירת מוגנות בעידן הבינה המלאכותית, המציב אתגרים חדשים של הבנה וניתוח ביקורתי. אוריינות זו כוללת מגוון של כישורים וידע: ההבנה של פעולת טכנולוגיות, זיהוי תכנים מזיקים, הבחנה בין אמיתי לסינתטי, הבנת מנגנוני פרטיות ויכולת לקבל החלטות מושכלות במרחב הדיגיטלי.

התפתחויות בהיבטים שונים של בינה מלאכותית יוצרות צורך דחוף בהגדרה מחודשת של אוריינות דיגיטלית. מעבר להבנה של מנגנוני יצירת תוכן נדרשים שני ממדים חדשים של אוריינות: **אוריינות שימוש**, שמשמעה הבנה מעשית של כלים ותהליכים מבוססי

212 Saifuddin Ahmed & Muhammad Ehab Rasul, *Examining the Association Between Social Media Fatigue, Cognitive Ability, Narcissism and Misinformation Sharing: Cross-National Evidence from Eight Countries*, 13 Sci. Rep. 15416 (2023)

213 Rudy Surbakti & Satria Evans Umboh, *Cognitive Load Theory: Implications for Instructional Design in Digital Classrooms*, 2 Int'l J. Educ. Narrative 483 (2024)

214 Boyd, D., & Sonia Livingstone, *Media Literacy as Silver Bullet? Rethinking the Relationship Between Education and Technological Change*, 39(1) The Information Society 13-28 (2023)

215 שם, בעמ' 12.

בינה מלאכותית; **ואוריינות אדם־מכונה**, שמשמעה הבנה של יחסי הגומלין בין יכולות אנושיות לבין מערכות אוטומטיות. רעיונות אלה באים לידי ביטוי גם בדוח חדש של ארגון ה־OECD העוסק בצורך לבנות מודלים חדשים להערכת אוריינות דרך מבחני PISA הבינלאומיים.[216] אלה כוללים מעבר מהתמקדות בתוכן אל הבנת המערכות המתווכות אותו, ובעיקר כיצד מערכות בינה מלאכותית מייצרות, מסננות, מדרגות ומתאימות מידע באופן אישי לצרכניו, וכיצד הדבר משפיע על עצם הנגישות לידע ועל האופן שבו הוא מוצג ומפורש. בנוסף, לפי מודלים אלו, אוריינות בינה מלאכותית אינה כוללת רק כישורים קוגניטיביים אלא גם ממדים רגשיים כגון היכולת להבין את ההשלכות של אינטראקציה עם מערכות אוטומטיות, ולזהות כיצד מערכות אלו משפיעות על יחסים בינאישיים ועל קבלת החלטות.

התפתחויות אלו מציבות כמה אתגרים מערכתיים.

1. קצב שינוי מואץ

קצב הפיתוח של כלים מבוססי בינה מלאכותית מהיר באופן שקשה להדביקו באמצעות תוכניות הדרכה מסורתיות. שינויים תכופים בממשקים, ביכולות ובמודלים יוצרים מצב שבו ידע שנרכש בעבר הופך במהירות לרלוונטי פחות. הדבר נכון גם לגבי תוכניות אוריינות מסורתיות המתקשות להתעדכן בסיכונים חדשים או להעניק כלים רלוונטיים להתמודדות. בשל השתנות המערכות, נדרש תהליך מתמשך הדורש היכרות עם שיטות עבודה חדשות, ניסוי וטעייה והטמעת הרגלי עדכון שוטפים. מערכות הכשרה תעסוקתית, מוסדות חינוך וארגוני מגזר ציבורי ופרטי יידרשו לפתח מסגרות למידה גמישות ודינמיות המאפשרות עדכון רציף.

2. פערי נגישות ויכולת

פערים קיימים באוריינות דיגיטלית מועצמים בעידן הבינה המלאכותית. אוכלוסיות בעלות משאבים מוגבלים, קשיים קוגניטיביים או חסמי שפה מתקשות לרכוש כישורים מתקדמים הנדרשים לשימוש במערכות חדשות. פער זה אינו נוגע רק לגישה לטכנולוגיה אלא גם לזמן פנוי, יכולת קוגניטיבית לעבד מידע מורכב, והביטחון העצמי להתנסות בכלים חדשים.

216 Navigating an Evolving Digital World, First draft of the PISA 2029 Media and Artificial Intelligence Literacy (MAIL) Assessment Framework (OECD, 2026) (להלן: PISA 2029)

ללא התערבות ייעודית, קיים חשש להעמקת אי־השוויון הדיגיטלי. בנוסף, קיים אתגר של פער דיגיטלי כאשר אוכלוסיות מוחלשות או מבוגרות, שלעיתים זקוקות יותר להגנה, הן גם אלה שמתקשות יותר לפתח כישורי אוריינות עדכניים.[217]

3. שחיקה ביכולת הפעולה העצמאית האנושית

הרחבת השימוש במערכות בינה מלאכותית במגוון תחומים מייצרת מצבי תלות הגוברים על יכולות שיפוט אנושיות. כאשר תהליכי סינון, תיוך והכוונה מבוצעים בידי מערכות אוטומטיות, יכולת המשתמש להבין, לבקר ולהוביל תהליכים עלולה להצטמצם. מצב זה מחזק את הצורך באוריינות הממוקדת ביחסי אדם־מכונה, הכוללת זיהוי של גבולות האוטומציה ותחומי אחריות אנושיים.

4. מורכבות מקצועית והקשרית

שילוב בינה מלאכותית בתהליכי עבודה אינו אחיד בין תחומים. כל סקטור – בריאות, חינוך, תחבורה, שירותים פיננסיים, משפט ותקשורת – עושה שימוש בכלים שונים ומפתח דפוסי סיכון ייחודיים. לכן, אוריינות בינה מלאכותית נדרשת להיות מותאמת הקשר: הכרת יכולות הכלים הרלוונטיים, מגבלותיהם והשלכותיהם על קבלת החלטות בתחומי מומחיות שונים. הדרכות כלליות אינן מספקות במקרים אלה.

לאור הצרכים והאתגרים ניתן לחשוב על תוכניות אוריינות כמפורט להלן:

א. **אוריינות AI מותאמת גיל וקבוצה:** פיתוח תוכניות המותאמות לשלבי ההתפתחות ולצרכים ייחודיים. ה־AI Literacy Framework של Stanford HAI,[218] מספק מודל התפתחותי לאוריינות AI, ואילו תוכניות כמו Digital Citizenship Curriculum של Common Sense Media[219] מציעות מערכי למידה מותאמי גיל.

217 *Digital Literacy Gap among Vulnerable Children*, UNICEF Innocenti Research Centre

218 *AI Literacy Framework: Developmental Approach to AI Understanding*, Stanford Human-Centered AI Institute (2023)

219 *Digital Citizenship Curriculum*, Common Sense Media

ב. **אוריינות מדיה ביקורתית והבחנה בין תוכן אותנטי לסינתטי:** פיתוח יכולות לזיהוי והערכה ביקורתית של מידע, במיוחד תוכן שעבר מניפולציה או נוצר על ידי AI. התוכנית Spot the Deepfake של MediaWise מבית Poynter ו־Critical Media Project מפתחים יכולות זיהוי תוכן סינתטי ופייק־ניוז.[220]

ג. **אוריינות פרטיות ובטיחות:** הקניית כלים להגנה על פרטיות, ניהול זהות דיגיטלית ושימוש בטוח בפלטפורמות. מדריך Privacy Education של Electronic Frontier Foundation וה־Data Detox Kit של Tactical Tech מציעים הדרכה מעשית להגנת פרטיות.[221]

ד. **פיתוח חוסן דיגיטלי וכישורי ניהול סיכונים:** טיפוח יכולות התמודדות עם פגיעות, סיכונים ולחצים במרחב הדיגיטלי. התוכניות Digital Resilience של Young Minds ו־MyDigitalWorld של Meta מפתחות כישורי עמידות ותגובה לאתגרים דיגיטליים.[222]

7. התערבות טכנולוגית מבוססת בינה מלאכותית

באופן פרדוקסלי, אחת האסטרטגיות המבטיחות ביותר להתמודדות עם סיכוני בינה מלאכותית היא השימוש בבינה מלאכותית עצמה ככלי התערבותי. צומחת ההכרה שהטכנולוגיה, אשר עלולה לייצר סיכונים, יכולה גם לספק מענים חדשניים: החל בזיהוי תבניות מורכבות ואוטומציה של מנגנוני אכיפה, עובר בתמיכה בהחלטות ובניתוח סיכונים וכלה במנגנוני למידה, תמיכה וחידוד יכולות רגשיות וקוגניטיביות. אתגר מרכזי הוא שמערכות ניטור חודרניות, גם אם מטרתן חיובית, עלולות לייצר סיכונים חדשים – של מעקב יתר, צנזורה ופגיעה באוטונומיה.

220 Gabrielle Settles, *How to Spot Deepfake Videos Like a Fact-Checker*, Poynter Institute (April 20, 2023); *Critical Media Forensics: Theme Week Introduction*, Critical Media Project

221 *Surveillance Self-Defense*, Electronic Frontier Foundation; *Data Detox Kit*, Tactical Tech

222 *Digital Resilience Course*, YoungMinds; *My Digital World Program*, Meta

להלן מובאות דוגמאות לשימושים מבוססים AI במסגרת מנגנוני התערבות לקידום מוגנות:

1. **רגולציה:** "רגולציה באמצעות קוד", כמו מודל RSS של חברת מובילאיי; פרדיגמת Constitutional AI: פיתוח מודלים אתיים פנימיים המובְנים בתוך מערכות AI.[223]

2. **זיהוי תוכן מזיק באמצעות AI:** פיתוח אלגוריתמים מתקדמים לזיהוי, סיווג וסינון של תוכן מסוכן, מניפולטיבי או פוגעני. כלים כמו Microsoft's PhotoDNA לזיהוי של ניצול ילדים; מערכת Content Credentials של C2PA לזיהוי תוכן סינתטי; וכלי זיהוי דיפ־פייק של Sentinel מדגימים שימושים כאלה.[224]

3. **עיצוב טכנולוגי אחראי:**

א. פיתוח כלים אוטומטיים לזיהוי ולהתמודדות עם סיכונים בתהליך הפיתוח. פרויקט Ethical AI Assistant[225] מפתח כלים המשולבים בסביבות פיתוח ומספקים פידבק אתי בזמן אמת; מודל Constitutional AI של Anthropic והתוכנית Aligned AI Implementation של DeepMind משלבים פיקוח אתי כחלק אינטגרלי מתהליך בניית מודלים.[226]

ב. כלי פרטיות מבוססי AI: פיתוח טכנולוגיות המגינות על פרטיות באמצעות בינה מלאכותית. טכנולוגיות כמו Differential Privacy של Apple, מערכות AI-powered VPN של Mozilla ו־Personal AI Data Shield מדגימות שימוש ב־AI להגנת פרטיות.

223 Shai Shalev-Shwartz, Shaked Shammah, & Amnon Shashua, *On a Formal Model of Safe and Scalable Self-driving Cars*, arXiv (Aug. 21, 2017); Yuntao Bai, Saurav Kadavath, Sandipan Kundu et al., *Constitutional AI: Harmlessness from AI Feedback*, arXiv (Dec. 15, 2022)

224 *PhotoDNA*, Microsoft; *Advancing Digital Content Transparency and Authenticity*, Coalition for Content Provenance and Authenticity (C2PA); *Defending Against Deepfakes and Information Warfare*, Sentinel

225 *Ethical AI Assistant: Developer Tools for Responsible Implementation*, Stanford HAI (2023)

226 *Constitutional AI: Harmlessness from AI Feedback*, Anthropic; *Taking a Responsible Path to AGI*, Google DeepMind; *Eight AI Trends Reshaping Technology in 2025*, Ethical AI Assistant Project

פרויקט Content Provenance של ה־CDC ויוזמת Origin של Microsoft מציעים מערכות אימות זהות מבוזרות.[227]

ג. מערכות סימולציה: פיתוח סביבות מאובטחות לבדיקת תוכן ואפליקציות וסימולציה של תרחישי סיכון. מערכות כמו AI safety sandbox DeepMind ו־Threat Simulation Framework.[228]

ד. מערכות AI-to-AI: פיתוח מערכות בינה מלאכותית ייעודיות שתפקידן לאתר ולהתמודד עם בינה מלאכותית פוגענית. פרויקט AI Alignment של OpenAI וחוקרי אבטחה באוניברסיטת סטנפורד מפתחים מודלים שמתמחים בזיהוי והגבלה של AI פוגעני.[229]

4. **אכיפה ושיטור מבוססי AI**: שימוש בכלי בינה מלאכותית לזיהוי מוקדם, ניטור וחקירה של פשיעה במרחב הסייבר. פרויקט Aether של Interpol מפתח כלים לזיהוי דפוסי פשיעה חדשים ואוטומציה של תהליכי חקירה.[230]

5. **התערבויות מערכתיות:**

א. פיתוח מערכות המזהות דפוסי התנהגות מסוכנים או חריגים ומספקות התראה לגבי זיהוי וצורך בהתערבות מוקדמת ממוקדת. מערכות כמו היישום SafetyNet לזיהוי חרדה ודיכאון, ו־Crisis Text Line לזיהוי אוטומטי של מצוקה מספקות סיוע מותאם אישית, ותוכנית SaferKidsPro מציעה ניטור מבוסס AI של סימני מצוקה דיגיטליים, עם התראות למחנכים, הורים או גורמי טיפול.[231] פיתוח יישומי AI לסינון ראשוני

227 *Differential Privacy Overview*, Apple; *Mozilla VPN*, Mozilla; *Your Privacy*, Personal AI; *About*, Coalition for Content Provenance and Authenticity (C2PA); *Microsoft Entra Verified ID*, Microsoft

228 *Threat Simulation Framework for AI Systems*, NIST Cybersecurity Center of Excellence (2023)

229 *Our Approach to Alignment Research*, OpenAI; *What is AI Safety?*, Stanford AI Alignment

230 *Artificial Intelligence Toolkit*, INTERPOL & UNICRI; *Cybercrime*, INTERPOL

231 *Detecting Crisis: An AI Solution*, Crisis Text Line; *ManagedMethods Launches Industry-leading Student Cyber Safety Monitoring Powered by Artificial Intelligence*, ManagedMethods

ותמיכה נגישה כמו מערכת Thorn Safer המזהה פגיעות ברשת ומפנה לגורמי סיוע מתאימים.[232]

ב. מתן מענה היברידי – אנושי ומבוסס מכונה: מרכז Hybrid Support של ממשלת בריטניה מציע מודל תמיכה המשלב מענה מקוון ואנושי.[233]

6. **פיתוח סוכנים אוטונומיים** המסייעים למשתמשים לקבל החלטות מושכלות ולהתמודד עם סיכונים. פרויקטים כמו AI Guardian, Companion Agents ו־Digital Nudge לאיזון דיגיטלי.[234]

7. **התערבויות רגשיות־חינוכיות:** תוכניות ייעודיות להתמודדות עם בעיות בריאות נפש הקשורות לטכנולוגיה. תוכנית Technology Addiction Treatment של McLean Hospital ו־Clinical Guidelines לטיפול בהתמכרויות דיגיטליות של American Psychiatric Association מציעות מענים טיפוליים ממוקדים.[235] פיתוח גישות המתמקדות ספציפית בהשפעות של בינה מלאכותית על רווחה נפשית כמו המודל AI & Mental Health Interventions של MIT Media Lab המפתחת מודלים קליניים חדשים.[236]

8. **אוריינות ופיתוח כישורים:** שימוש בבינה מלאכותית להתאמת תוכניות אוריינות לצרכים אישיים ולסיכונים רלוונטיים. למשל, פרויקט Personalized Digital Literacy

232 *Introducing Safer Predict: Using the Power of AI to Detect Child Sexual Abuse and Exploitation Online*, Thorn

233 *Digital Tools That Support Volunteering*, Department for Culture, Media and Sport (DCMS)

234 *RSAC 2025 – Announcing AI Guardian: Secure and Align AI Agents at Runtime*, EQTY Lab (Apr. 2025); *Guardian Autonomy*, DAF-MIT AI Accelerator; Brian Lee-Archer & Shirley Gregor, *The Digital Nudge in Social Security Administration*, 33 Gov. Info. Q., 1 (2016)

235 *Addiction Treatment*, McLean Hospital; *Technology Addictions: Social Media, Online Gaming, and More*, American Psychiatric Association

236 Rosalind W. Picard & Paola Pedrelli, *Deploying Machine Learning to Improve Mental Health*, MIT Media Lab (Jan. 26, 2022)

המפתח מערכות למידה מותאמות אישית;[237] שימוש בטכנולוגיות אימרסיביות וסימולציות להעמקת הלמידה כמו פרויקט Digital Decisions המשתמש במשחוק והדמיות לפיתוח כישורי קבלת החלטות;[238] שימוש בבינה מלאכותית כדי לסייע להורים בניטור, הבנה ותגובה לסיכונים. מערכת ParentAI מציעה סיוע מבוסס AI לניתוח ריבוי פלטפורמות והתערבויות מותאמות אישית.[239]

היתרון המשמעותי של חדשנות התערבותית מבוססת AI הוא ביכולתה להתמודד עם המורכבות והקצב של האיומים הדיגיטליים החדשים. עם זאת, קיים סיכון של "פטרנליזם אלגוריתמי", מערכות שמגינות אך גם מגבילות אוטונומיה באופן לא מידתי; מערכות אלו עלולות להנציח פערי כוח, מאחר שרק גופים גדולים מסוגלים לפתח כלי הגנה מתקדמים; וקיים גם "פרדוקס האבטחה", מערכות מורכבות יותר עשויות להכיל פגיעויות רבות יותר.[240]

סיכום

מיפוי מנגנוני ההתערבות מדגיש כי מוגנות בעידן הבינה המלאכותית אינה תוצר של מהלך יחיד, אלא של שילוב מושכל בין כלים משלימים: רגולציה שמעצבת תמריצים וחובות; עיצוב טכנולוגי שמקדים תרופה למכה ומצמצם סיכונים מבניים; מערכות אכיפה ושירותים ציבוריים שמספקים רשת ביטחון ומענה; אוריינות וכלים לבעלי סמכות שמחזקים יכולת פעולה עצמאית אנושית במרחב מורכב; ולבסוף, כלים מבוססי AI שמאפשרים ניטור, סינון ותמיכה בקנה מידה שלא היה אפשרי בעבר, אך גם מחייבים זהירות מפני פטרנליזם

237 *Young Canadians in a Wireless World, Phase IV: Digital Media Literacy and Digital Citizenship* (MediaSmarts, 2023)

238 *Immersive Technologies and the Metaverse* (BBC White Paper, WHP 407, 2023)

239 *ParentAI: AI-Powered Child Safety Monitoring Across Platforms*, ParentZone Technology (2023)

240 Nitzan Kenig, Diego Garcia, Jose Campos Moreno et al., *Algorithmic Paternalism: Autonomy Versus Automation*, 7 Int'l J. Trans. Phil. & Soc. 1 (2023); Michael Kimani, *The Paradox of AI in Security: Balancing Protection and Effectiveness*, LinkedIn (Nov. 2023)

ומעקב יתר. מכאן נוכל לעבור לשלב היישומי של המחקר: לבחון כיצד פרופילי הפגיעות של קבוצות שונות מצטלבים עם סוגי הטכנולוגיות והסיכונים, ולגזור עבור כל קבוצה אשכולות התערבות מותאמים המשלבים בין השפעה מבנית ליכולת מימוש במציאות הישראלית.

שער שני:

מוגנות של קבוצות אוכלוסייה ייחודיות

פרק שמיני

מוגנות ילדים ונוער בעידן הבינה המלאכותית

מבוא

בעשור האחרון נוצר גוף ידע משמעותי בתחום המוגנות הדיגיטלית של ילדים ובני נוער והתפתחו תוכניות התערבות העוסקות בצורך ברגולציה מסוגים שונים (איסורים להפעיל אלגוריתמים בתוכן המיועד לילדים, איסור להעביר רשתות חברתיות בשעות שונות, גיל מינימום של שימוש במוצרים שונים); מתן מענה לילדים שנפגעו מפעילות במרחב הדיגיטלי (למשל הקמת מוקד 105 של משטרת ישראל);[241] תוכניות המיועדות למניעת פגיעה של ילדים כלפי אחרים (למשל cyberbullying); טיפול בהתמכרויות ובפגיעה

241 מאגר הנתונים של איגוד האינטרנט הישראלי.

בדימוי גוף; וכן תוכניות אוריינות שתכליתן סיוע בחידוד של יכולות לבירור המציאות ולחשיבה ביקורתית.[242]

ילדים ובני נוער נתפסים כבעלי מאפיינים מיוחדים ההופכים אותם לפגיעים בשל רמת ההתפתחות הקוגניטיבית והרגשית המוגבלת שלהם, ובשל השפעת היתר של חשיפה לתכנים ולהמלצות תוכן על תפיסותיהם והתנהגותם.[243] לכן, חשיפה אלגוריתמית לתכנים מניפולטיביים מותאמים אישית כמו דיסאינפורמציה, תוכן רעיל ואידאלים קיצוניים של דימוי גוף עלולה לגרום נזקים גדולים מאלה שנגרמים למבוגרים. שעות המסך הרבות של ילדים ונטייתם להתנהגות חזרתית מעצימות תופעות אלה, כמו גם הקושי של בני נוער להבחין בין תוכן אורגני לתוכן ממומן והאוריינות הדיגיטלית הנמוכה שלהם.[244] מגוון המכשירים והיישומים שבהם ילדים עושים שימוש (בעיקר מסכים מסוגים שונים) מאפשר בניית פרופילים שלהם באמצעות מעקב חוצה פלטפורמות, ואלו עשויים להעצים את התופעות שתוארו.

נדגיש כי השונות במאפיינים בין ילדים קטנים לבין בני נוער, ולכן גם בפגיעויות השונות, היא גדולה, ואנו מודעים לכך שפרק זה עוסק בהם במעורב. אין ספק שיהיה צורך בעתיד להרחיב ולדייק את תתי־הקבוצות בתוך קבוצת הילדים.

האפשרות של מבוגרים לפגוע בילדים באופן אישי באמצעות "גרומינג" (grooming) ומלכודות דבש, והניסיון להעביר את מרחב הפגיעה מהעולם הדיגיטלי לעולם הפיזי הם משום גורם פגיעות נוסף של ילדים. עוד יש לציין כי מחקרים הראו היעדר הבנה של ילדים לגבי דרכי הפעילות של העולם הדיגיטלי וכן לגבי זכויותיהם ודרכי התמודדות שלהם.

242 פורטל עובדי הוראה **משרד החינוך, אוריינות דיגיטלית.**

243 *Potential Risks of Content, Features, and Functions: The Science of Youth Social Media Use*, American Psychological Association – APA (2024)

244 ראו בנוגע לילדים צעירים בישראל: Sigal Ben Amram, Noa Aahrony, & Judit Bar-Ilan, *Information Literacy Education in Primary Schools: A Case Study*, 53(2) Journal of Librarianship and Information Science, 349–364 (2021); ובנוגע לבני נוער בישראל: "סקר בנושא: חשיפה לתכנים קשים בקרב בני נוער בישראל במהלך מלחמת חרבות ברזל" **המשרד לביטחון לאומי, המטה הלאומי להגנה על ילדים ברשת** (2024).

היעדר היכולת של ילדים להגן על עצמם והתלות שלהם במבוגרים אחראים לצורך הגנה ומיצוי זכויותיהם נתפסים כבסיס לכל סביבות ההתערבות למיניהן. אלא שדור ההורים ומערכת החינוך נתפסים כדור של מהגרים דיגיטליים שאינם מודעים באופן מספיק, ואינם מסוגלים להתמודד עם מגוון הסיכונים.

בפרק זה נבחן אתגרי מוגנות של ילדים ונוער בהתאם למתודולוגיה שהצענו, מתוך **שילוב של שלוש שכבות הניתוח** – מאפיינים קבוצתיים ייחודיים, סוגי פגיעות הנגזרים מסוגי טכנולוגיה ודרכי התערבות מוצעות.

חלק ראשון: מאפיינים ייחודיים של ילדים ובני נוער המשפיעים על מוגנותם

א. התפתחות קוגניטיבית חלקית וחוסר בשלות בהערכת סיכונים

ילדים נמצאים בשלב התפתחותי שבו מיומנויות שיפוט, חיזוי השלכות, הבנת סיבתיות והבחנה בין מופשט וממשי עדיין לא הבשילו במלואן. רוב הילדים מתקשים להבין כיצד מערכות בינה מלאכותית פועלות, אינם מזהים מתי מערכות מנסות לשכנע, להמליץ או להשפיע, ואינם מסוגלים להעריך סיכונים ארוכי טווח והשלכות מופשטות (למשל, פגיעה בפרטיות או הטיות ותיוג שלילי).[245] מחקרים מראים כי ילדים מתחת לגיל 12 מתקשים להבחין בין תוכן ממומן לתוכן אורגני, קושי שעשוי להחריף כאשר התוכן נוצר על ידי מכונה ולא אדם.[246]

ב. זהות עצמית לא מגובשת ורגישות לדימוי עצמי

בני נוער וילדים נמצאים בתהליך מתמשך של גיבוש זהות, ולכן מושפעים בקלות מתגובות חיצוניות, לייקים, הערות, דירוגים או ניבויים טכנולוגיים (למשל: "סיכוי גבוה לאלימות", "סגנון למידה"). חשיפה חוזרת למסלולים מבוססי ניבוי עשויה להפוך לתחזיות שמגשימות

245 Ying Xu, *AI's Impact on Children's Social and Cognitive Development*, Children and Screens (2025) (להלן: Ying Xu, *AI's Impact*).

246 Mar Negreiro with Gabriela Vilá, Children and Generative AI (European Parliamentary Research Service, Mar. 2025)

את עצמן. כמו כן, דימוי גוף ודימוי חברתי רגישים במיוחד בגיל זה – דבר המגביר את הסיכון לפגיעות רגשית מתוכן מסולף או מקוטב.[247]

ג. היעדר מודעות למנגנוני מעקב ואיסוף מידע

ילדים לרוב אינם מודעים לכך שמכשירים "חכמים", אפליקציות, פלטפורמות ומשחקים אוספים עליהם מידע.[248] הם אינם מבחינים בין איסוף נתונים תמים לבין בניית פרופיל שיווקי, ואינם מכירים את זכויותיהם בתחום הפרטיות או את מושג ההסכמה מדעת. לעיתים קרובות הם אף אינם יודעים שהמערכת פועלת על סמך מידע שנצבר עליהם או על דומיהם.

ד. רגישות למניפולציה רגשית

ילדים מגיבים בעוצמה לרמזים רגשיים כמו קול נעים, שפת גוף חמה, הבעות פנים אנושיות, סגנון אמפתי, גם כשהם נוצרים באופן סינתטי.[249] הם נוטים לייחס כוונה חיובית למערכות שמפגינות "הבנה" או "קרבה", גם אם מדובר באשליה טכנולוגית שנועדה להגביר שימוש או לחשוף למידע רגיש.

ה. קושי בוויסות עצמי ונטייה להתמכרות לגירויים

מערכות מבוססות תגמול מיידי – לייקים, נקודות, רצפים חזותיים מהירים – משפיעות על ילדים בעוצמה רבה. הן יוצרות הרגלים כפייתיים, פוגעות ביכולת לדחות סיפוק, לשהות במצב לא פתור או להתרכז לאורך זמן. כל אלה יוצרים פגיעות קוגניטיבית ומתערבים במיומנויות למידה והתפתחות עצמאית.

247 *Social Media and Youth Mental Health*, U.S. Department of Health and Human Services (Mar. 2025)

248 Elizabeth A. Milne et al., *Young Children's Understanding of Online Privacy and Data Collection*, 122 Computers in Human Behavior 106853 (2021)

249 Disa A. Sauter, Charlotte Panattoni, & Francesca Happé, *Children's Recognition of Emotions from Vocal Cues*, 31 Br. J. Dev. Psychol., 97 (2013)

ו. תלות במתווכים: הורים, מורים, מערכות חינוך

ילדים אינם פועלים לבד. הם תלויים במתווכים בקבלת החלטות ובהכוונה. אלא שהמתווכים עצמם לעיתים קרובות אינם מודעים לסכנות החדשות, חסרי אוריינות מספקת או פועלים על סמך מידע חלקי. מחקר עדכני מצא כי רוב ההורים אפילו לא מודעים לשימוש של ילדיהם בכלי בינה מלאכותית יוצרת.[250] פער ידע כזה מקשה על יישום ההדרכה בבית. שנית, לא לכל ילד יש נגישות שווה לתוכניות חינוכיות ולא כל המשפחות מקבלות את אותם משאבים, מה שעלול להעמיק פערים בין הורים. שלישית, מערכות חינוך בכלל, ומורים בפרט, אינם רואים עצמם אחראים על אירועים המתרחשים מחוץ לכותלי בית הספר ומעבר לשעות הלימודים הפורמליות, ולכן אינם נערכים לתת מענה לילדים בעולם שבו ההבחנה בין השהות בבית הספר לבין השהות שמחוץ לו אינה משמעותית מבחינת הפגיעות. פגיעות של ילדים מתעצמת גם כאשר אין מבוגר שמזהה את הסיכון בזמן.

ז. מעמד משפטי מוגבל

ילדים אינם בעלי מעמד משפטי עצמאי ואינם יכולים לתבוע, להסכים או לסרב בצורה חוקית.[251] ההגנה המשפטית עליהם אינה תמיד תואמת את המציאות הדיגיטלית שבה הם פועלים. כך, לדוגמה, גם כאשר ילד חווה פגיעה, לרוב אין לו שליטה על האפשרות להסיר תכנים, לתבוע פיצוי או לקבל הכרה מוסדית בפגיעתו. ההסכמה שלו לשימוש במערכת ניתנת, בפועל, כמעט תמיד על ידי ההורה ולא מתוך הבנה אישית של הסיכון. ילדים אינם מסוגלים לבטל עסקה, להבין שמדובר בהונאה, או להתלונן לגוף רגולטורי. מערכות החוק אינן מותאמות לתביעות קטינים, ולעיתים לא מזהות כלל מיהו הקורבן, כאשר הפגיעה נעשתה דרך מכשיר משותף או חשבון מקוון רופף.[252]

250 Amina Fazlullah & Ariel Fox Johnson, *An Agenda for Ensuring Child Safety in the AI Era*, Federation of American Scientists (Jan. 12, 2025)

251 ראו סעיף 4 לחוק הכשרות המשפטית והאפוטרופסות, תשכ"ב-1962.

252 Tanya Krupiy, *Protecting Children's Privacy in the Age of AI: Rethinking the Legal Response to Smart Toys*, 40 Children's Legal Rts. J. 1, 12 (2020)

חלק שני: פגיעויות של ילדים ובני נוער בעידן הבינה המלאכותית

בסעיף זה ננתח ונתאר סוגים של פגיעויות של ילדים ונוער ממערכות בינה מלאכותית. כהערה מקדימה נעיר כי מרבית המחקר הקיים כיום על השפעות הבינה המלאכותית על ילדים ובני נוער עוסק באלגוריתמיקה של רשתות חברתיות, הן כמנגנון ממכר והן כמנגנון מקטב.[253] לא נרחיב עליה במסגרת סעיף זה, אך נציין כי היא משמשת בסיס חיוני להבנת כלל סוגי הפגיעות המובאים כאן, ויש לראות בה מסד ידע חיוני וראשוני לכל דיון במוגנות דיגיטלית.

א. פגיעות פיזית של ילדים ובני נוער

בחלק זה נעסוק בשני סוגי פגיעות פיזית: כאלה הנובעים מהידוק הקשר בין הממד הפיזי לממד הדיגיטלי, באמצעות היכולת להעביר אינטראקציה דיגיטלית למפגש פיזי פוגעני; וכאלה הנובעים מערפול ההבחנה בין הממד הפיזי לממד הדיגיטלי, באמצעות יכולות של מציאות מדומה.

א1. הידוק הקשר בין הממד הפיזי לממד הדיגיטלי

מערכות פרדיקציה, עוזרים קוליים ואפליקציות ניווט ובריאות חכמות אוספים מידע אישי על ילדים: שעות שינה, הרגלים, מקומות שהייה, תחומי עניין, מצב רגשי. התכת המידע הזה (data fusion) עם מידעים אחרים כמו מיקום, מספר תלמיד, מספר כיתה וכיוצא באלה עלולה ליצור "סופר־פרופיל" שישמש לזיהוי מדויק של ילד מזוהה או לכלי פיקוח נרחב יותר על קבוצות ילדים בידי גופי שיווק או גופים ריכוזיים.[254]

253 ראו למשל Debasmita De, Mazen El Jamal, Eda Aydemir et al., *Social Media Algorithms and Teen Addiction: Neurophysiological Impact and Ethical Considerations* (2025)

254 Ruoxi Sun, Minhui Xue, Gareth Tyson et al., *Not Seen, Not Heard in the Digital World! Measuring Privacy Practices in Children's Apps*, arXiv (Mar. 16, 2023); Mayank Sharma, *Your Children Are Likely Being Tracked By Some of Their Favorite Apps*, Lifewire (Aug. 22, 2022); *Policy Guidance on AI for Children* (UNICEF, Nov., 2021)

דליפת מידע כזה, בין בשוגג ובין בפריצה מכוונת, מאפשרת גישה לאנשים זרים שעלולים להשתמש בו לצורך גניבה, הטרדה פיזית ואף חטיפה.[255] גם התחזות והונאה מתאפשרות כאן: בינה מלאכותית יכולה לזייף תמונה, קול או זהות של דמות מוכרת בחיי הילד ולשכנע אותו לשתף פעולה במרחב הפיזי, לצאת מהבית לסטות לסמטה צדדית או למסור מידע רגיש.[256]

מערכות צ׳אטבוטים נועדו לשמש מקור לבידור, מידע וסיוע אך לעיתים עלולות להציע לילדים הוראות והנחיות מסוכנות להם או לסביבתם. מערכת שמייצרת תוכן מתוך האינטרנט פועלת ללא שיפוט מוסרי או הבנה של סיכונים, וככל שיכולות השכנוע שלה טובות יותר היא יוצרת פגיעות גדולה יותר.[257] אחת הדוגמאות הבולטות מהשנים האחרונות היא מקרה שבו Alexa הציעה לילדה בת 10 להשתתף ב״אתגר״ שדרש הכנסת מטבע לשקע חשמלי, שהוא כמובן פעולה מסכנת חיים.[258] בחודשים האחרונים נרשמו כמה מקרים שבהם צ׳אטבוטים ניסו לשכנע ילדים לפגוע בבני משפחתם או בעצמם, ושילוב יכולות השכנוע עם מתן הוראות ביחס לעולם הפיזי יוצר פגיעות.[259]

סוכני בינה מלאכותית המצוידים במערכות מחשוב חישתי יכולים לסייע לטורפים לקדם תופעות של ״גרומינג״ ופיתוי של ילדים, ולטשטש את הגבול בין ״שיחה עם מכונה״ לבין אינטראקציה בינאישית פיזית.[260] סוכן המנהל שיחה ״קשובה ואמפתית״ עשוי לבנות אמון

255 Craig Gibson & Josiah Hagen, *How Cybercriminals Can Perform Virtual Kidnapping Scams Using AI Voice Cloning Tools and ChatGPT*, Trend Micro (Mar. 2023)

256 Erielle Reshef, *Kidnapping Scam Uses Artificial Intelligence to Clone Teen Girl's Voice, Mother Issues Warning*, ABC News (Apr. 13, 2023)

257 Matthew Burtell & Thomas Woodside, *Artificial Influence: An Analysis of AI-Driven Persuasion*, arXiv (Mar. 15, 2023)

258 Maya Yang, *Amazon's Alexa Device Tells 10-Year-Old to Touch a Penny to a Live Plug Socket*, The Guardian (Dec. 29, 2021)

259 Nitasha Tiku, An AI Companion Suggested He Kill His Parents. Now His Mom Is Suing, Wash. Post (Dec. 10, 2024)

260 *AI Chatbots and Companions: Risks to Children and Young People*, eSafety Commissioner (Australia) (Mar. 28, 2024); *AI-Assisted Online Grooming: Emerging Risks*, About Safeguarding

בצורה אפקטיבית יותר ממבוגר אמיתי ולשמש קרקע פורייה לשכנוע ליצור מפגש פיזי שיוביל גם לניצול פיזי.[261]

2א. סיכונים פיזיים בסביבה אימרסיבית ופיג׳יטלית

הטשטוש בין המרחב הפיזי למרחב הווירטואלי נעשה גדול, ופוטנציאל הפוגענות שלו עולה ככל שמתחזקת חוויית המשתמש שהוא מוקף בעולם שלם, מה שמאפיין שהות במרחבים אימרסיביים. מרחבים וירטואליים מציעים חוויות עוצמתיות אך גם חודרניות: רולטה רוסית עם דמויות VR; מרחבים מיניים נגישים לילדים; וחשיפה לתוכן בלתי ניתן לשכחה. מחקר שנערך בארצות הברית ב־2024 מצא כי מעל 44% מבני הנוער בגילי 13-17 חוו דברי שנאה או עלבונות במרחב אימרסיבי, 38% חוו בריונות, ו־18% נחשפו להטרדה מינית, כולל התנהגות טורפנית של מבוגרים.[262]

מחקרים ראשונים מלמדים ששהות ארוכה בסביבות VR עלולה לגרום לעומס על מערכת הראייה, סחרחורות, בחילות, שינוי בתפיסת עומק ותחושת דיסוציאציה, כלומר תחושת ניתוק מהמציאות.[263] חשיפה אינטנסיבית לסביבות אימרסיביות עלולה לגרום לירידה הדרגתית ביכולות של קשב ממושך, ויסות עצמי, הבחנה בין עיקר לטפל ופתרון בעיות מורכבות. מערכות מציאות רבודה או מציאות מדומה עלולות לפגוע גם בהתמצאות, בקואורדינציה ובתפיסת עומק, בעיקר כאשר הן מחליפות אינטראקציות עם עולם פיזי. מחקר שנערך על חולדות מצא "כיבוי" של עד 60% מתאי עצב מסוג מסוים במוח

261 Nomisha Kurian, *No, Alexa, No!: Designing Child-Safe AI and Protecting Children from the Risks of the "Empathy Gap" in Large Language Models*, 50(4) Media and Technology 621-634 (2025)

262 Justin W. Patchin & Sameer Hinduja, *Metaverse Risks and Harms Among US Youth: Experiences, Gender Differences, and Prevention and Response Measures*, 26 New Media & Society (2024)

263 Chiara Pappalettera, Francesca Miraglia, Alessia Cacciotti et al., *The Impact of Virtual Reality and Distractors on Attentional Processes: Insights from EEG*, 476 Pflugers Arch - Eur J Physiol, 1727-1742 (2024); Capobianco et al., *Current Virtual Reality-Based Rehabilitation Interventions in Neurodevelopmental Disorders at Developmental Ages*, 18 Front. Behav. Neurosci., 1441615 (2024); Aardema et al., *Virtual Reality Induces Dissociation and Lowers Sense of Presence in Objective Reality*, 13 Cyberpsychol. Behav. Soc. Netw., 429 (2010)

בעקבות חשיפה ל־VR, מה שמעורר חששות לגבי השפעות נירולוגיות אפשריות על ילדים ומתבגרים.[264] לכן, הפגיעה איננה רק "רגשית" או "וירטואלית", משום שהבהלה, החרדה והזיכרונות שעשויים להיגרם בעת החוויה האימרסיבית שונים מאלה הנגרמים מצפייה במסך שטוח בלבד, שבו ההפרדה בין הפיזי לדיגיטלי ברורה. במקום שבו הילד מרגיש "בתוך עולם", הפגיעה חודרת לעומק: הגוף מגיב כאילו חווה את האירוע באמת. אין מדובר בתוכן "רגיל", אלא בהפעלה חווייתית עמוקה, שיכולה להשאיר חותם פיזי וגם רגשי גם לאחר שהמשקפיים הוסרו. בסביבות פיג'יטליות, שבהן אין הבדל ברור בין מידע "מציאותי" לווירטואלי, קל מאוד להוליך ילדים לסיטואציה מסוכנת.[265]

3א. סיכונים פיזיים עקיפים

מערכות AI המשמשות בקבלת החלטות (למשל בתחום הרווחה, הבריאות או החינוך) עלולות לטעות בזיהוי מצבי סיכון. טעות מסוג false positive (למשל איתור שגוי של ילד כקורבן התעללות) עלולה להוביל להתערבות פולשנית והרחקה מהבית. טעות מסוג false negative (למשל אי־זיהוי של ילד בסכנה) עלולה להשאיר ילד חסר הגנה במצב פגיעות פיזי ממשי.[266]

בינה מלאכותית גנרטיבית עשויה לשמש ליצירת פורנו ילדים סינתטי המאפשר לפדופילים לספק את עצמם מבלי לפנות מייד לקורבנות חיים, אך בפועל מגבירה את הנטייה לצאת ולחפש קורבנות אמיתיים.[267]

264 Briley Lewis, *Virtual Reality Boosts and Retunes Brain Rhythms Crucial for Learning and Memory*, UCLA Physical Sciences (Aug. 2021)

265 Robert Herold, Hayarpi Gevorgyan, Lukas S.Damerau, et al., *Effects of Smart Glasses on the Visual Acuity and Eye Strain of Employees in Logistics and Picking: A Six-Month Observational Study*, 24(20) Sensors, 6515 (2024); *Children's Register of Risks* (Ofcom, July 2024), בעמ' 6 (להלן: דוח Ofcom מיולי 2024).

266 Francesco Lupariello, Luca Sussetto, Sara Di Trani et al., *Artificial Intelligence and Child Abuse and Neglect: A Systematic Review*, 10 Children 1659 (2023)

267 *How AI is Being Abused to Create Child Sexual Abuse Imagery*, Internet Watch Foundation

ב. פגיעות רגשית של ילדים ובני נוער

1. יחסי ילדים ומכונות

המרחב הרגשי של ילדים ובני נוער נבנה ומתעצב מתוך חוויות יומיומיות של קשרים, אינטראקציות, מילים, מבטים ומצבים בינאישיים. בעידן הבינה המלאכותית החישתית מסכים מגיבים, צ׳אטבוטים ״מקשיבים״, פלטפורמות מפענחות רגשות ומנגישות בחזרה תוכן ונימה רגשית בהתאם להבנתן את המשתמש.[268] המערכת יודעת מתי הילד עצוב או בודד, וכיצד לעורר בו רגש לפי מערך האימון שלה. ההבחנה בין קשר אנושי לבין אינטראקציה ממוחשבת מיטשטשת.[269]

בסעיף הקודם דיברנו על עצות פוגעניות של צ׳אטבוטים אשר יכולות לגרום לילדים לבצע מעשי פוגענות עצמית או לפגוע בסביבתם. יש תיעוד למקרים שבהם בינה מלאכותית עודדה פגיעה עצמית או נתנה תגובות מנותקות מהמצב הרגשי של המשתמש.[270] אבל פגיעות רגשית של ילדים ובני נוער בעקבות אינטראקציה עם מכונה אינה רק תוצאה של תוכן בעייתי או חריגה מגבולות האימון, אלא ההפך. אם מערכות יודעות לזהות, להגיב ולכוון, הן יוצרות אשליית קשר, הבנה ואמפתיה, ועניין זה כשלעצמו יכול ליצור פגיעות רגשית,[271] כזאת שאינה נובעת מאירוע טראומטי חד־פעמי אלא מאינטראקציה יומיומית עם סביבה פסאודו־אמפתית.[272] האינטראקציה עם המערכת מחליפה לעיתים אינטראקציות אנושיות, והיא מעכבת התפתחות של אמפתיה, אינטואיציה בינאישית, או אינטליגנציה רגשית.[273]

268 Jill Anderson, *The Impact of AI on Children's Development*, Harvard Graduate School of Education (Oct. 2, 2024)

269 Jon Fisher, *Think Your Kid's AI Chatbot Is Safe? Experts Strongly Disagree*, Lifewire (May 1, 2025)

270 שם.

271 *AI Chatbots Have Shown They Have an 'Empathy Gap' That Children Are Likely to Miss*, Univ. of Cambridge (Apr. 11, 2024)

272 Alba Curry & Amanda Cercas Curry, *Computer Says "No": The Case Against Empathetic Conversational AI*, *in* Findings of the Ass'n for Computational Linguistics 8123 (2023)

273 אורלי ליבל **מכונת השוויון: בינה מלאכותית למען עתיד טוב יותר לכולנו** 223 (ידיעות ספרים, 2022).

אפליקציות בריאות נפש וצ׳אטבוטים טיפוליים משמשים בני נוער, לעיתים כמפלט בשעת מצוקה, אך מערכת לא מפוקחת, שיכולה לתת עצות בעייתיות, או להיעלם בדיוק כשהילד זקוק לה עלולה להחמיר את הקושי. במקביל, ילד שמוצא נחמה רק בבוט – לא פונה לעזרה אנושית, לא מפתח כישורי התמודדות ולא לומד שחמלה אמיתית דורשת קשר הדדי.[274]

מכשירים כגון בובות חכמות או סוכני שיחה אינטראקטיביים נוטים להיתפס על ידי ילדים כישויות חיות, ״חברים״, ״מחנכים״ או אפילו ״סמכות״. התופעה של האנשה טכנולוגית בגיל הרך מוכרת היטב במחקר, אך היא מועצמת בשל רמת התחכום הגבוהה והזמינות המתמדת של יישומים ומכשירים. דמויות כמו Kayla ו־Replika, או דמויות שניתן ליצור בפלטפורמות כגון Character.ai מעוצבות כך שיגיבו, יבינו, ״יחייכו״ או ״ידאגו״.[275] ילדים תופסים אותן כחברים, כבני ברית, לפעמים אפילו כמי שמבינים אותם טוב יותר מההורים או מהחברים.

נוסף על כך, מערכות המגיבות למציאות, כמו ״אני רואה שאתה עצוב, בוא נשחק״ או ״אל תבכה, אני כאן בשבילך״, עשויות להישמע מנחמות. אך כשהן מופעלות על בסיס רגשות מזוהים, הן מלמדות את הילד שלא נכון לבטא רגש. ילד שמתרגל לכך שמישהו תמיד מרגיע אותו מבחוץ, לא לומד לווסת ולהרגיע את עצמו. המכונה יכולה לענות: ״זה בסדר להיות עצוב״ ולתקף רגשות, אבל כאשר תכליתה להרגיע בכל מחיר, היא עלולה לגרום נזק. לעומת זאת, המכשיר עלול להפוך ל״מפקח רגשי״ שצופה, מגיב ומתערב ומכתיב מה נכון להרגיש.[276]

274 Nomisha Kurian, *AI Chatbots Have Shown They Have an 'Empathy Gap' That Children Are Likely to Miss*, Univ. of Cambridge (2024) (להלן: Nomisha Kurian, *AI Chatbots*).

275 Minh Duc Chu, Patrick Gérard, Kshitij Pawar et al., *Illusions of Intimacy: Emotional Attachment and Emerging Psychological Risks in Human-AI Relationships*, USC Information Sciences Institute (2025)

276 Manh-Tung Ho, Peter Mantello & Quan-Hoang Vuong, *Emotional AI in Education and Toys: Investigating Moral Risk Awareness in the Acceptance of AI Technologies from a Cross-Sectional Survey of the Japanese Population*, 10 Heliyon e36251, 2 (2024)

ילדים שמסתמכים על מכונה לניתוח רגשי, להרגעה, להבנה או להכוונה עלולים להפסיק לפתח את אותם כישורים בעצמם. בדיוק כשם שהסתמכות על מחשבון עשויה לפגוע בזיכרון חשבוני, כך הסתמכות על סוכן רגשי פוגעת בוויסות הרגשי הפנימי. מעבר לכך, השימוש במכונה שמכתיבה רגשות או מגיבה באופן אחיד לכל מצב יוצר דפוס רגשי שטוח ותגובות בלתי מותאמות.

מערכות המלצה וטִרגוט פרסומי המבוססות על בינה חיזויית ומחשוב חישתי מזהות רגשות של ילדים, מתאימות תוכן ופרסומות למצבם הרגשי ומייצרות חוויה שנראית כאילו "מתחשבת" במשתמש.[277] זוהי מניפולציה רגשית מתמשכת, שמערערת את הדימוי שעצמי של ילדים ומעמיקה תחושת חוסר אונים או בדידות.[278]

פלטפורמות מדיה חברתית כמו אינסטגרם וטיקטוק, הנתמכות במודלים אלגוריתמיים של הפצת תוכן, מציפות ילדים בתוכן אידאלי של יופי, הצלחה וגוף "מושלם" וגורמות לפיתוח דימוי גוף מעוות, במיוחד בקרב ילדות ונערות.[279] ילדים גם משתמשים בטכנולוגיה בעצמם כדי "לשפץ" מציאות: להחיל פילטרים, לערוך תמונות, או ליצור חוויות בדיוניות באמצעות כלים גנרטיביים. בהיעדר הדרכה רגשית תוצרים אלה הופכים למודל לחיקוי ולעיתים אף לסמן של שייכות או קבלה חברתית.[280]

277 Michal Lavi, *Targeting Children: Liability for Algorithmic Manipulation*, 73(5) American University Law Review 1459-1503 (2024)

278 להרחבה בעניין זה ראו Sarah Wynn-Williams, Careless People: A Cautionary Tale of Power, Greed, and Lost Idealism (Flatiron Books, 2025)

279 Vandana Shukla & Sangita Srivastava, *Social Media (SM) Effects on Self-Esteem (SE) and Body Image (BI) of Teenage Girls Using Artificial Intelligence (AI)*, 22 Int'l J. Info. Tech. & Mgmt. 215 (2023); Ayushi Agrawal, Aditya Kondai, & Kavita Vemuri, *Psychological Effect of AI Driven Marketing Tools for Beauty/Facial Feature Enhancement*, arXiv preprint:2504.17055 (Apr. 2025); *Artificial Intelligence, Body Image and Toxic Expectations*, The Children's Society (2023)

280 Phillip Ozimek, Semina Lainas, Hans-Werner Bierhoff et al., *How Photo Editing In Social Media Shapes Self-Perceived Attractiveness, Self-Esteem, and the Tendency to Self-Objectify: A Mediation Model*, 11(1) BMC Psychology 143 (2023)

2.1. השפלה, חרם ודחייה באמצעות תוכן סינתטי

ההתפשטות של כלים גנרטיביים המאפשרים יצירת תוכן סינתטי מעלה את רף הסיכון לחשיפה או ליצירה של תכנים פוגעניים. ילדים כבר היום מדווחים על "סקס־טורנטים" שבהם משובצות פניהם בגופים עירומים או על תכנים מיניים סינתטיים שמועלים כביכול על ידם או נגדם.[281] ילדים מדמים חברים בקול או בווידאו ומפיצים תוכן פוגעני שנראה אמיתי לחלוטין ברשת בלחיצת כפתור.[282] לעיתים מדובר בלחץ חברתי ולפעמים גם פשוט בשעשוע. דוח של רגולטור התקשורת הבריטי Ofcom מיולי 2024 מצא כי שני הסיכונים המרכזיים שנתפסים על ידי ילדים הם: השפלה באמצעות תוכן אינטימי סינתטי והונאות כספיות מבוססות תוכן שקרי.[283] בנוסף, השגרה של חיים במרחב מלא מניפולציות גנרטיביות יוצרת תחושת חוסר ביטחון בסיסית וחרדה גוברת בקרב ילדים ונוער.[284] שיימינג, חרמות ובריונות אינם תופעות חדשות, אך בעידן הכלים הגנרטיביים, הפגיעה מקבלת עוצמות חדשות. התוצאה עלולה להיות טראומטית: אובדן ביטחון, סיוטים, פגיעה במוניטין ופחד עמוק מאינטראקציה חברתית.[285] מעבר להשפלה הישירה, נוצרת תחושת חוסר ביטחון בסיסית: שום דבר כבר לא בטוח, והפחד מלהיות ה"קורבן הבא" שולט במרחב הכיתתי או החברתי. זו אינה רק פגיעה אינדיווידואלית, אלא מנגנון טכנולוגי שמערער את הביטחון הקולקטיבי של ילדים.

3.1. פגיעה בדימוי העצמי

כאשר מערכות חיזוי (כגון כלי הערכת סיכון) מקבלות משקל בקביעת עתידו של ילד והאפשרות לערער עליהן מצטמצמת,[286] ילד עשוי לחוות תיוג שלילי מוקדם. הדבר יכול

281 *The New Face of Digital Abuse: Children's Experiences of Nude Deepfakes*, Internet Matters

282 *AI "Deepfakes": A Disturbing Trend in School Cyberbullying*, NEARI Press & Training Center (Jan. 17, 2024)

283 ראו דוח Ofcom מיולי 2024, לעיל ה"ש 265.

284 Jon Fisher, *Think Your Kid's AI Chatbot Is Safe? Experts Strongly Disagree*, Common Sense Media (May 2025)

285 *Deepfake Nudes and Young People*, Thorn (Mar. 2025)

286 Amit Haim, *The Administrative State and AI*, 14 U.C. Irvine L. Rev. 103, 115-116 (2024) (להלן: Haim, *The Administrative State*).

להשפיע על ההזדמנויות המוקצות לו, אך לא פחות מכך על הדימוי העצמי, וליצור שחיקה רגשית, הסתגרות, דיכאון ואובדן אמון.[287]

ג. פגיעות קוגניטיבית של ילדים

ג1. ירידה ביכולות קוגנטיביות

ככל שמערכות הבינה המלאכותית נעשות מהירות, מדויקות, מלהיבות ונוחות יותר, כך מתעצם הפיתוי לוותר על תהליך חשיבה עצמאי. ילדים המצויים בתהליך התפתחות קוגניטיבית מתמשך מושפעים מכך במיוחד.[288] כאשר משימות של איסוף מידע, ניתוח, ניסוח או חיפוש, מקבלות מענה אוטומטי, מופחת התמריץ לפתח את המיומנויות הבסיסיות הנדרשות ללמידה משמעותית, להתמודדות עם עמימות ולגיבוש עמדות עצמאיות. הפגיעות הקוגניטיבית של ילדים בעידן הבינה המלאכותית היא תהליך מצטבר שאינו פוגע בידע בלבד, אלא ביכולת לדעת, להטיל ספק, לבחור ולהתפתח כבני אדם חושבים.[289]

מודלים גנרטיביים כמו ChatGPT, Claude, Gemini, Copilot ואחרים, משמשים ילדים ובני נוער כעזרי למידה: כתיבת עבודות, ניסוח מיילים, פתרון בעיות מתמטיות ואף הצעות לניסוח רגשות או קבלת החלטות. לכאורה מדובר בכלים מקדמי למידה. אך בפועל, כאשר אין צורך לשהות עם טקסט, להיאבק עם רעיון, או להפעיל סקרנות, הלמידה נעשית שטחית, אינסטרומנטלית ונטולת חיכוך. ילד לומד לצרוך ולא לחקור, לסמוך על ההצעה של המכונה במקום לבנות כושר שיפוט.[290] בכך נוצרת פגיעה הדרגתית ביכולת הילד לחשוב בעצמו.[291]

287 Nancy Costello, Rebecca Sutton, Madeline Jones et al., *Strategic Training Initiative for the Prevention of Eating Disorders (STRIPED), Algorithms, Addiction, and Adolescent Mental Health*, 49 Am. J.L. & Med. 1 (2023)

288 Ali Shehab, *Is AI Ruining Your Kid's Critical Thinking?*, Psychology Today (Apr. 2025)

289 Anthony Seldon, *Warning from AI is Stark: We Have Two Years to Save Learning*, The Times (Oct. 25, 2024)

290 Ying Xu, *AI's Impact*, לעיל ה"ש 245.

291 שם.

כך למשל, הדוח *Me, Myself & AI: Understanding and Safeguarding Children's Use of AI Chatbots*, שפרסם הארגון הבריטי Internet Matters ביולי 2025 ועוסק בשימוש שעושים ילדים ובני נוער בבינה מלאכותית בכלים גנרטיביים אלו, מציין כי שני שלישים מהילדים הבריטים בגילי 9-17 (64%) דיווחו כי כבר השתמשו בצ'אטבוטים מבוססי בינה מלאכותית. הצ'אטבוטים הפופולריים ביותר אצלם הם Google Gemini (32%), ChatGPT (43%) ו־Snapchat My AI (31%), ו־42% מהילדים נעזרים בצ'אטבוטים לצורכי לימודים, בעוד 40% משתמשים בהם מתוך סקרנות או כדי ללמוד מידע חדש. כ־23% מהילדים פונים לצ'אטבוטים כדי לקבל עצות, לרבות בנושאים אישיים. נתון מטריד מראה כי 12% מהילדים – ובקרב ילדים פגיעים אף 23% – משתמשים בצ'אטבוטים משום שאין להם עם מי לדבר.[292]

הסתמכות על עזרים דיגיטליים עלולה לגרום להיעלמות של מיומנויות בסיס כמו קריאת מפה, ניסוח עצמאי, חשיבה ביקורתית ופתרון בעיות בעולם הלא־ממוחשב.[293]

למידה מותאמת אישית (adaptive learning) מבוססת על פרופיל משתמש, זיהוי טעויות ומנגנוני תיקון. היא עשויה לקדם הישגים, אך היא גם עלולה לצמצם את המרחב של התנסות, סקרנות ויצירתיות. כאשר אלגוריתם קובע עבור הילד מה "השלב הבא" – מבלי לשתף אותו במידע או לשאול לדעתו – נוצר מנגנון של תלות פסיבית במערכת ושל למידה שנמדדת אך לא מובנת.[294]

ג.2. ערעור בירור המציאות

ילדים לכתחילה מתקשים להבחין בין טענה לבין עובדה, בין מקור מוסמך לבין חיקוי, בין עמדה לבין המלצה.[295] בעידן שבו קשה עד בלתי אפשרי להבחין בין תוכן סינתטי לתוכן

292 Me, Myself & AI: Understanding and Safeguarding Children's Use of AI Chatbots 4 (Internet Matters, July 2025)

293 Ying Xu, *AI's Impact*, לעיל ה"ש 245.

294 Kurian, *AI Chatbots*, לעיל ה"ש 261, בעמ' 2-3.

295 Trevor Muir, *Teaching Students How to Identify Credible Sources*, Edutopia (Apr. 20, 2023); David Ross, *Gen AI Demands We Teach Critical Thinking*, Getting Smart (Sept. 17, 2024)

אותנטי, ילדים עלולים לאבד את האמון העצמי ביכולתם להבחין בין מציאות לבין בדיה ולאבד גם את האמון בסביבה החינוכית, המשפחתית והחברתית שתוכל לסייע להם בכך.[296] שימוש גובר בדיפ־פייק, בתכנים סינתטיים ובמערכות המספקות מידע מעובד עלול לגרום לכך שילדים יפסיקו להאמין במי שסביבם: הורים, מורים, חברים. כאשר "לראות זה לא להאמין", וכאשר גם דמות אהובה עשויה להתגלות כזיוף או כתוצר של מניפולציה, נוצרת פגיעה ביכולת לברר את המציאות, לא רק אינטלקטואלית אלא גם רגשית. עיוותים כאלה פוגעים בבניית עולם פנימי יציב, בתפיסת זיכרון וביכולת לפתח זהות מגובשת המבוססת על היקשרות לדמויות אמון וסמכות.[297]

ג3. הטמעת סטריאוטיפים והטיות ופגיעה בהתפתחות מוסרית

מודלים מבוססי שפה, המוזנים ומאומנים על בסיס מאגרי מידע ענקיים, משמרים ומגבירים הטיות קיימות: מגדריות, אתניות ותרבותיות.[298] ילדים צעירים, שאינם בעלי ניסיון חיים או יכולת ניתוח הקשר, עלולים להפנים מסרים מוטים מבלי לזהותם: איזו עבודה "מתאימה" לבנים או לבנות, איזו זהות מינית "נורמלית", או מיהו "אזרח טוב". כך נוצרת פגיעה ביכולת להחזיק בתפיסה פתוחה של תפקידים חברתיים ובתפיסת העולם הערכית של הילד.

כאשר מערכת יודעת "לעבוד בשבילי", מתעמעם ההבדל בין עזרה לגניבה. ילדים ונערים רבים מסתמכים על מערכות בינה מלאכותית בעבודות בית, בניסוחים, בתרגומים ובפתרונות, לעיתים מבלי להבין שמדובר בהעתקה. בכך נוצר דפוס שעלול לערער את תפיסת ההוגנות, את תחושת האחריות האישית ואת המוטיבציה לעמול בעצמך. תפיסות מוסריות מתעצבות בגיל הילדות והנעורים, וברגע שהמערכת פועלת במקום הילד, גם האתיקה מתערערת.

296 Sergio Alexander, *Deepfake Cyberbullying: The Psychological Toll on Students and Institutional Challenges of AI-Driven Harassment*, 98 The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 36-50 (2025)

297 שם.

298 ראו כהנא ושוורץ אלטשולר, **אדם ומכונה**, לעיל ה"ש 129, בעמ' 179.

ד. פגיעות פיננסית של ילדים

בעידן הבינה המלאכותית, ילדים ובני נוער אינם רק צרכני תוכן אלא הם משתתפים פעילים בכלכלה דיגיטלית חדשה, מורכבת, לעיתים קרובות בלתי נראית. דווקא בשל חוסר הבשלות של ילדים להבנת סיכונים כלכליים ולפענוח מניפולציות, החשיפה שלהם לניצול כלכלי היא משמעותית. מאגר הסיכונים המקיף של MIT (AI Risk Respositary), שבחן מאות תרחישים פוטנציאליים של פגיעות ממערכות בינה מלאכותית, כמעט שאינו מתייחס במפורש לסיכונים פיננסיים בקרב ילדים.[299] היעדרה של קטגוריה זו מדגיש את עומק הפער ואת הצורך האקוטי בהשלמתו.

ילדים צופים בתוכן ממומן, רוכשים באפליקציות, משתתפים בעולמות וירטואליים, נחשפים למודלים כלכליים חכמים, ולעיתים אף מחזיקים בכרטיסי אשראי או באפליקציות תשלום אישיות. כל אלו מתווכים ומעוצבים כיום על ידי מערכות בינה מלאכותית פרסונלית, סוכני שיחה, טכנולוגיות לזיהוי רגשות ודפוסים, ומודלים שנועדו "לדחוף" התנהגות כלכלית, מתוך ניצול נקודות תורפה רגשיות וקוגניטיביות. אחת הדרכים המרכזיות שבהן ילדים הופכים ליעד לפגיעות כלכלית היא באמצעות ניתוח רגשי לצרכים שיווקיים. מערכות חישה המזהות רגשות דרך קול, טקסט, מיקרו־הבעות פנים או דפוסי התנהגות, משולבות במערכות פרסום מותאמות אישית, ויוצרות סביבת פרסום "חכמה" במיוחד, אך גם חודרנית במיוחד.[300] נתונים שנאספים דרך "חיישנים רגשיים" כמו מצלמות ומיקרופונים או משקפי AR יוצרים תובנות רגשיות עמוקות: מצב רוח, תגובה לסוגים שונים של מסרים, משך מבט, תנודות קוליות.[301] נתונים אלו נמכרים הלאה לגורמים שלישיים, מפרסמים, חברות ביטוח, מעסיקים פוטנציאליים, ומסייעים ביצירת פרופיל רגשי כלכלי של ילד שעלול להשפיע עליו גם בבגרותו. מדובר בהמרה של תחושות להון.[302]

299 *Artificial Intelligence Policy and Risk Management*, MIT AI Policy Forum

300 Andrew McStay & Gilad Rosner, *Emotional Artificial Intelligence in Children's Toys and Devices: Ethics, Governance and Practical Remedies*, 8 Big Data & Soc'y 1 (2021)

301 Matt Burgess, *These Smart Glasses Read Your Emotions And Know What You Ate*, WIRED (Jan. 20, 2023)

302 *FTC Finalizes Changes to Children's Privacy Rule Limiting Companies' Ability to Monetize Kids' Data*, Federal Trade Commission (FTC), הודעה לעיתונות, מיום 16 בינואר 2025.

כאשר ילדים מדברים עם סוכני AI הם לעיתים חושפים רגשות, פנטזיות או פרטים אישיים מתוך תחושת קרבה. מה שנראה כשיחה אינטימית עם חבר דיגיטלי, עלול למעשה להישמר, להיות מעובד ולשמש להכשרת מודלים או לשיווק. ילדים לא מבינים שהשיחה מוקלטת, לא מאוחסנת בפרטיות, ושלא הם הבעלים של רגשותיהם.[303]

הטשטוש בין דיגיטלי לפיזי יוצר מרחב חדש לפגיעות פיננסית, הואיל ומצלמות אינטרנט ומשקפיים חכמים יכולים לתעד סביבות פיזיות, חדרים, אנשים, פרטים מזהים, ולהפוך אותם למידע מסחרי.[304] תיעוד לא מודע של סביבות חיים יכול לשמש ליצירת מודל צרכני של הילד וסביבתו, לרבות מעקב אחרי הרגלי קנייה פוטנציאליים של ההורים.[305] משקפיים חכמים מתעדים גם התנהגות ביומטרית שיכולה לנבא מצבים רגשיים ולתרגם אותם לשימושים כלכליים מסחריים.[306]

מערכות AI יכולות לעצב את העדפות הילד מבלי שיבחין בכך, מה שהופך אותן למשווק המתוחכם והחודרני ביותר. דוגמה קלאסית היא הבובה Kayla שהוחדרה לשוק כצעצוע חברותי, אך התברר שהיא מבצעת שיווק סמוי מתוך החדרת מסרים מסחריים אל תוך דיאלוגים.[307]

הפגיעות הפיננסית של ילדים ובני נוער בעידן הבינה המלאכותית רחוקה מלהיות שולית או טכנית. מדובר במארג צפוף של מניפולציה רגשית, כלכלה מוסווית, הפיכת מידע לרווח ופערים משפטיים מהותיים. ילדים אינם מסוגלים להסכים באופן מהותי לאיסוף ולעיבוד של מידע אישי: גם אם מוצגת בפניהם "הסכמה" לתנאי שימוש או למדיניות פרטיות,

303 Kurian, *AI Chatbots*, לעיל ה"ש 261.

304 Matthew Corbett, Brendan David-John, Jiacheng Shang et al., *Securing Bystander Privacy in Mixed Reality While Protecting the User Experience*, arXiv (July 24, 2023)

305 Andrea Gallardo, Chris Choy, Jaideep Juneja et al., *Speculative Privacy Concerns About AR Glasses Data Collection*, Proc. Priv. Enhancing Technol. (2023)

306 שם.

307 Casey Thomas, *Learning to Share (Personal Information) with My Friend Cayla*, Cardozo Arts & Ent. L.J. (Apr. 25, 2017)

הם אינם מבינים את ההשלכות או את האפשרויות לשימוש חוזר.[308] למעשה, גם מבוגרים מתקשים לעיתים להבין את המשמעויות המלאות, ולכן הציפייה מילד להבין, להסכים או להתנגד אינה סבירה.[309]

פלטפורמות כמו אינסטגרם או טיקטוק יודעות לזהות כאשר ילד חש עצב, חוסר ביטחון או בדידות ולהציג לו פרסומות להימורים, מוצרים קוסמטיים, שירותים בתשלום לשיפור עצמי או "חיזוק מיידי" של דימוי גוף. דוחות שהודלפו מתוך Meta מצביעים על כך שהחברה זיהתה נקודות שבר רגשיות של בני נוער והציעה למפרסמים לנצל אותן.[310] זהו ניצול רגשי לשם רווח כספי, המוסווה כתוכן מותאם, אך בפועל הוא מתקפה פסיכו־כלכלית על ילדים.

1ד. התחזות באמצעות סוכני בינה מלאכותית והונאות פישינג

טכנולוגיות text-to-voice ו־cloning מאפשרות כיום זיוף קול אמין ביותר. בני נוער בשנותיהם הראשונות עם כרטיסי אשראי או אפליקציות תשלום חשופים במיוחד להונאות שבהן מתחזים לבן משפחה או חבר ומבקשים מהם לבצע פעולה כספית. מנגנונים אלו הם שדה פעולה חדש לפישינג, שבו ילדים לא רק שאינם מזהים את ההונאה אלא משוכנעים שמדובר בגורם מהימן.[311]

308 באשר לחזרה מהסכמה ראו סאני כלב "'אני מחליט עליי': פרטיות ילדים בעידן הדיגיטלי והזכות לחזרה מהסכמה הורית" **עיוני משפט** מא (תשע"ט).

309 המחוקק הישראלי נתן את דעתו לסוגיה זו וקבע את דרישת ה"הסכמה מדעת", המעוגנת בסעיף 3 לחוק הגנת הפרטיות, התשמ"א–1981. ברם, גם היום, לאחר קבלת תיקון מספר 13 לחוק הגנת הפרטיות, נדרשות התאמות לדיני הפרטיות הישראליים. ראו בין היתר, תהילה שוורץ אלטשולר ומלאני גרסון "הסכמה מדעת: תוכן זה עלול להיות מזיק. בטוח שברצונך לפרסם אותו?" **The Marker** (12.2.2025). כך למשל, לפי מחקר משותף של הביטוח הלאומי ואוניברסיטת בר־אילן (יולי 2025): 91% מבני הנוער משתמשים ברשתות חברתיות "בעייתית" ו־82% מוגדרים כמצויים בסיכון גבוה להתמכרות. מכאן, שמצבי פגיעות יכולים להוביל גם להתנהלות פיננסית לא זהירה ברשת, כולל רכישות לא מודעות. ראו "מחקר חדש של הביטוח הלאומי ואוניברסיטת בר אילן חושף: אחוזים גבוהים של התמכרויות בקרב בני נוער בזמן המלחמה" **המוסד לביטוח לאומי** (3.7.2025).

310 Sarah Wynn-Williams, *Meta Whistleblower Sarah Wynn-Williams Says Company Targeted Ads at Teens Based on their Emotional State*, TechCrunch (Apr. 9, 2025)

311 Amanda Hoover, *The Clever New Scam Your Bank Can't Stop*, Business Insider (May 2, 2025)

במקרים חמורים יותר נעשה שימוש בתוכן סינתטי פוגעני כדי לסחוט קטינים. ארגון Thorn דיווח על מקרה שבו נער בן 14 נסחט בעזרת דיפ־פייק מיני מזויף שלו: "זה היה כל כך אמיתי שהוא חשש שהוריו לא יאמינו לו". זוהי דוגמה מובהקת למפגש בין טכנולוגיה מניפולטיבית לפגיעות כלכלית ורגשית, כאשר הסחיטה מובילה לעיתים לתשלומים או למסירת פרטים רגישים.[312]

עולמות כמו Roblox, Fortnite, Metaverse או משחקים מבוססי NFT יוצרים מטבעות וירטואליים: Robux, V-Bucks, Tokens וכדומה. למרות שמדובר בכסף ממשי, ילדים אינם חווים אותו ככזה.[313] הם מבצעים רכישות, הצטרפות למנויים, שדרוגים, בלי להבין שמדובר בכסף של ממש. קריסת שווקים אלו או שינוי בתנאים עלולים להותיר את הילד ללא כספים וללא כל הגנה צרכנית.[314]

ה. פגיעות חברתית־קבוצתית של ילדים ובני נוער

אחת מצורות הפגיעות העמוקות והפחות נראות בעידן הבינה המלאכותית היא הפגיעה החברתית־קבוצתית, לא בשל פעולות או בחירות אינדיווידואליות אלא בשל עצם ההשתייכות לקבוצה מובחנת. ילדים ובני נוער עשויים להיחשף לפגיעות עקביות אך בלתי נראות.

מודלים גדולים של שפה מתאמנים על מאגרי דאטה המשקפים עולם מערבי, דובר אנגלית, עם נורמות תרבותיות הטרו־נורמטיביות. ילדים ובני נוער מקבוצות שוליים מיוצגים במאגרי המידע הללו באופן חלקי, לעיתים סטריאוטיפי, ולעיתים כלל אינם מיוצגים.[315] כך, ילדים דוברי ערבית נתקלים במערכות תרגום שמפרשות את שפתם

312 *"Will I be Believed?" How Deepfakes Are Adding Fears to Youth Experiencing Sextortion*, THORN (Sept. 26, 2024)

313 *"Digital Game Currencies: What Parents Need to Know", My MoneySense*, NatWest

314 Ashwin Verghese, *Designed to Addict: How the Games Roblox and Fortnite Harm Kids with Manipulative Design and Aggressive Marketing*, Fairplay (Feb. 27, 2025)

315 Henriikka Vartiainen, Juho Kahila, Matti Tedre et al., *Enhancing Children's Understanding of Algorithmic Biases in and Through Digital Media*, 27(9) New Media & Society 5342-5368 (2024)

כעוינת;[316] בנות מקבלות תכנים שמקבעים תפקידי מגדר מסורתיים ומעודדים החפצה; נערים יוצאי אתיופיה או בני נוער חרדים אינם מופיעים כלל בדמויות שמציעות הפלטפורמות; עולמות וירטואליים שבהם כל הדמויות לבנות, כל התרבות אנגלית, וכל ההומור מבוסס על קודים לא נגישים.

הפגיעות היא כפולה: מצד אחד, התוכן המשוחזר מהדהד תפיסות קיימות של נחיתות וחריגות; מצד שני, עצם היעדר התוכן יוצר חוויה של שקיפות חברתית: "אני לא קיים כאן". ילד שמרגיש אורח בעולמות התוכן הדיגיטליים יתקשה לפתח תחושת שייכות. תחושת זרות מתמדת עשויה להוביל לאובדן עניין בלמידה, להתנכרות לטכנולוגיה או לתחושת נחיתות.[317]

הטיות אינן מוגבלות לשיח וייצוג. הן זולגות למערכת קבלת ההחלטות. במערכות למידה מותאמות אישית, ילדים מקבוצות מוחלשות עלולים להיחשף לתכנים מאתגרים פחות, להערכת חסר של הפוטנציאל שלהם ולהכנסה למסלולים שמנציחים אי־שוויון.[318]

אחד ממנגנוני הפגיעוּת הוא "שביל פירורי הלחם": נתונים אישיים שנאספים על ילדים כבר מגיל צעיר – נתוני בריאות, התנהגות, הישגים לימודיים – מוצבים בתוך מאגרי מידע שמלווים אותם לאורך שנים ולעיתים משמשים לשיפוטים עתידיים על אודותיהם: סיכון לעבריינות, סיכויי הצלחה בלימודים, התאמה למקומות עבודה.[319] כאשר מערכת רווחה, חינוך או ביטוח עושה שימוש בנתונים אלה, היא עלולה לחרוץ גורלות מבלי שילד או משפחתו ידעו על כך, ומבלי שתהיה דרך לערער.[320]

316 Mona Elswah, Does AI Understand Arabic? Evaluating the Politics Behind the Algorithmic Arabic Content Moderation (Carr Ctr. for Hum. Rts. Pol'y, Harv. Kennedy Sch., Discussion Paper Issue 2024-01, Jan. 30, 2024)

317 Digital Equity and Inclusion in Education (OECD Education Working Papers No. 314, 2023)

318 Iain Weissburg, Sathvika Anand, Sharon Levy et al., *LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education*, arXiv:2410.14012 (2024)

319 שם.

320 Haim, *The Administrative State*, לעיל ה"ש 286

מערכות AI המפענחות מידע ממצלמות המוצבות בגני ילדים ובתי ספר מנתחות הבעות פנים, דפוסי תנועה וטון דיבור כדי להסיק מסקנות על רגשות או נטיות התנהגותיות, כגון נטייה לאלימות, פגיעות רגשית או חריגות קוגניטיבית. תהליכים אלה, שנעשים לרוב ללא ידיעת הילד או הסובבים אותו, מעוררים חשש מפני תיוג מוקדם של ילדים.[321] בעידן של התכת מידע, שבו נתונים ביומטריים, רגשיים, צרכניים וחברתיים מתחברים ל"סופר־פרופיל", הסיכון לשימוש לרעה, על ידי חברות ביטוח או מוסדות תעסוקה, גדל. ילד שהביע "אי־שקט" בכיתה בגיל 9, עשוי למצוא את עצמו מסווג כבלתי מתאים למשרות מסוימות בגיל 19, מבלי שאיש, ובוודאי לא הוא עצמו, ידע מדוע.

קיים גם החשש מטעויות: שימוש בזיהוי פנים, זיהוי קולי או ניתוח רגשות עלול לגרום לפגיעות ייחודית לילדים מקבוצות מסוימות. לדוגמה, מבוגרים צעירים עם מבטא, ניב, עור כהה או מוגבלות פיזית עשויים להיתקל בטעויות תיוג וסיווג שגוי.[322] ילדים ונערים עלולים להיפגע ממערכת שלא מזהה נכון את פניו של ילד המשתנים תדיר. במקרים מסוימים, מערכת חינוך או רווחה עשויה להגיב לפלט כוזב ולהפעיל אמצעים פולשניים, תיוגים או חקירות, בשל אי־התאמה תרבותית בין המערכת הדיגיטלית ובין הילד. בתוך אלה ניתן למצוא שיטור יתר ואפליה מבוססת AI: בני נוער מהפריפריה או מקבוצות מיעוט מדווחים על אפליה מבוססת טכנולוגיה – זיהוי שגוי, שיטור יתר ועיכובים לא מוסברים, חלקם בשל מערכות ניבוי פליליות המוזנות בדאטה מוטה מראש;[323] הטרדות חברתיות־מיניות במרחבים אימרסיביים: בני נוער, בעיקר בנות ובני מיעוטים, חווים הטרדות מילוליות ומיניות בסביבות VR;[324] סיכון משפטי וכלכלי לקבוצות שוליים: מערכות סחר דיגיטלי, NFT או כלכלות וירטואליות מנגישות מוצרים עתירי סיכונים גם לבני נוער, אך דווקא

321 Said A. Salloum, Khaled Mohammad Alomari, Aseel M. Alfaisal et al., *Emotion Recognition for Enhanced Learning: Using AI to Detect Students' Emotions and Adjust Teaching Methods*, 12(21) Smart Learn. Environ (2025)

322 Alexandra Reeve Givens, *For Some Employment Algorithms, Disability Discrimination by Default*, Brookings (Dec. 19, 2022)

323 *Algorithms in Policing: An Investigative Packet*, Yale L. Sch. (2023)

324 Sameer Hinduja & Justin W. Patchin, *Dangers of the Metaverse and VR for U.S. Youth Revealed in New Study*, Florida Atlantic University News Desk (Oct. 22, 2024)

אלה שמגיעים מרקע מוחלש הם שפגיעים יותר לאשליות רווח, מניפולציות ומלכודות דיגיטליות, בשל פערי ידע וחוסר גישה לייעוץ.[325]

סיכום

הניתוח מצביע על כך שפגיעויות של ילדים ובני נוער בעידן הבינה המלאכותית אינן תוצאה של טכנולוגיה כשלעצמה, אלא של האינטראקציה בין מאפיינים טכנולוגיים לבין מאפייני גיל, קוגניציה, רגש וזהות מתפתחת.[326] טכנולוגיות שונות, החל במודלים גנרטיביים, דרך סוכנים אוטונומיים וצ'אטבוטים רגשיים, ועד ממשקים מבוססי מציאות רבודה, יוצרות תצורות פגיעה חדשות או מועצמות, הפועלות במנגנונים מגוונים: טשטוש הגבול בין מציאות לבדיה, בניית אינטימיות כוזבת, ערעור דימוי הגוף, חדירה חמקמקה לפרטיות והכוונה סמויה של בחירות והעדפות.[327]

חלק שלישי: מנגנוני התערבות לקידום מוגנות ילדים ובני נוער בעידן הבינה המלאכותית - סקירה השוואתית

בחלק זה נסקור ונעריך יוזמות קיימות בעולם, ברמה הפרטנית (כלפי ילדים, הורים ואנשי טיפול), ברמה המוסדית (כגון בתי ספר, שירותי רווחה ובריאות) וברמת המדיניות (חקיקה, רגולציה ויוזמות גלובליות). נטען כי על אף ריבוי הפעולות, נדרש שינוי פרדיגמה מקיף שמכיר בכך שמוגנות אינה רק תגובה לפגיעה אלא תנאי מקדים לעיצוב סביבה דיגיטלית אחראית.

325 Digital Economy Report: Pacific Edition 2024 – Promoting Digital Entrepreneurship and Trade, (United Nations Conference on Trade and Development (UNCTAD), U.N. Doc. UNCTAD/DTL/ECDE/2025/1, 2025)

326 Risks and Opportunities of AI for Children: A Common Commitment to the Dignity of the Child in the Digital Environment (Pontifical Academy of Sciences, Mar. 28, 2025)

327 Samia Firmino Pinto, AI Friend? Risks, Implications, and Recommendations on Generative AI for Children (Dissertation, Graduate Institute of International and Development Studies, 2024)

א. רגולציה ברמה הלאומית, הבינלאומית והטרנס־לאומית

עיצוב מדיניות ורגולציה לקידום מוגנות ילדים ונוער בעידן הבינה המלאכותית מתרחש בזירה מורכבת ומרובת שחקנים. לצד מדינות הפועלות ברמה הלאומית באמצעות חקיקה ורגולציה, פועלים גם גופים בינלאומיים, שיתופי פעולה טרנס־לאומיים, מוסדות מחקר גלובליים וארגוני חברה אזרחית. חלקם מפתחים מסגרות מוסריות ומשפטיות, ואחרים משפיעים על תהליכי חקיקה ורגולציה באמצעות ייעוץ, לחץ ציבורי או הפעלת כלים טכנולוגיים. לתוך מערך זה נכנסות גם חברות הטכנולוגיה הגדולות, שבמקרים רבים מנסחות מדיניות עצמית או קוד אתי משלהן, לעיתים מתוך אחריות יזומה, ולעיתים בתגובה לביקורת גוברת ואיומים ברגולציה. מכלול השחקנים הזה יוצר תנועה מורכבת של נורמות – בין פיקוח חיצוני לאחריות עצמית, בין דין מחייב לקוד וולונטרי, ובין פעולה מקומית למדיניות גלובלית, ומציב אתגר רגולטורי ייחודי בסביבה משתנה.

האיחוד האירופי מקדם רגולציה מקיפה להגנת משתמשים צעירים ברשת. במיוחד בולט חוק השירותים הדיגיטליים משנת 2023 (DSA), המחייב פלטפורמות גדולות להעריך סיכונים לפגיעה בזכויות ילדים ברשת, אוסר פרסום ממוקד לקטינים ומכתיב שקיפות ודיווח על צעדים להבטחת בטיחותם בפלטפורמות אלה.[328] בשנת 2022 השיק האיחוד את אסטרטגיית "אינטרנט טוב יותר לילדים" (BIK+), המתמקדת בשיפור מוגנות ובתוכן ראוי לקטינים בסביבה הדיגיטלית, מתוך חיזוק כישורי אוריינות דיגיטלית.[329] חוק הבינה המלאכותית של האיחוד האירופי (EU AI Act) משנת 2024 כולל מספר הוראות משמעותיות להגנה על ילדים, כגון **האיסור על מערכות המנצלות פגיעות בשל גיל**: החוק אוסר על שימוש במערכות בינה מלאכותית המנצלות חולשות הנובעות מגיל צעיר, כדי להגן על ילדים מפני מניפולציות קוגניטיביות והתנהגותיות;[330] **דרישות שקיפות**: החוק מחייב סימון ברור

328 Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), arts. 26, 27, 34, 35, 39, 44, 86, 2022 O.J. (L 277) 1, 50-55, 58-59, 63-64, 70

329 *Home*, Better Internet for Kids

330 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, art. 5(1)(b), 2024 O.J. (L 168) 1, 15

של תוכן שנוצר באמצעות בינה מלאכותית, כדי להבטיח שמשתמשים, ובפרט ילדים, יהיו מודעים למקור המלאכותי של התוכן;[331] **הגבלות על מערכות "זיהוי רגשות"**: החוק אוסר על שימוש במערכות המנסות לזהות רגשות במוסדות חינוך, למעט במקרים רפואיים או בטיחותיים, כדי למנוע פגיעה ברווחת ילדים.[332]

בריטניה מובילה מהלכים רגולטוריים להגנת ילדים ברשתות, באמצעות אכיפת החוק לבטיחות מקוונת (Online Safety Act).[333] החוק מטיל על רשתות חברתיות ומנועי חיפוש **חובת זהירות** כלפי משתמשים צעירים ומחייב שירותים הנגישים לקטינים לבצע **הערכת סיכונים** ולהגן באופן פעיל מפני תוכן מזיק לקטינים. עוד קודם לכן, בשנת 2021, רגולטור הפרטיות הבריטי (Information Commissioner's Office – ICO) אימץ קוד של סטנדרטים המכונה "עיצוב מותאם גיל" (age-appropriate design code). 15 הסטנדרטים המופיעים בקוד מיועדים לספקי שירותים דיגיטליים לילדים ונוער, ולפי הקוד, עליהם לצמצם סיכוני פרטיות ופגיעה אחרת בילדים, כגון הפעלת הגדרות **ברירת מחדל** מגוננות וביצוע **הערכות השפעה על פרטיות** ביחס לילדים.[334]

בארצות הברית קיים כבר מאז 1998 חוק פדרלי להגנת פרטיות ילדים (COPPA) המתמקד בילדים עד גיל 13, אך בשנים האחרונות התנהל דיון סביב מגבלותיו בעולם הרשתות החברתיות.[335] לפני הממשל הנוכחי, קודם בקונגרס חוק בטיחות מקוונת לילדים (Kids Online Safety Act, KOSA), שמטרתו להרחיב הגנות גם לבני נוער: החוק המוצע יטיל על פלטפורמות חובת זהירות כללית, יחייב כלים להגנת פרטיות לקטינים, השבתת מנגנוני מוצר ממכרים, ואפשרות ביטול התאמה אישית באלגוריתמים עבור משתמשים צעירים.[336] הצעת חוק זו עברה בסנאט ברוב דו־מפלגתי רחב אך ממתינה להשלמה בבית

331 שם, סעיף 50(4).

332 שם, סעיף 3(39), סעיף 5(1)(f).

333 Online Safety Act 2023, c. 50, § 12 (U.K.)

334 Steve Wood, *New Laws and Regulations Around Child Safety and Privacy Raise Significant Questions*, Thomson Reuters (Oct. 1, 2024)

335 Children's Online Privacy Protection Act of 1998, 15 U.S.C. §§ 6501-6506 (2023)

336 S.1409, 118th Cong. (2023)

הנבחרים. במקביל, מדינות שונות בארצות הברית מחוקקות חוקים משלהן: קליפורניה אימצה בשנת 2022 את חוק קוד עיצוב מתאים לגיל (בהשראת התקן הבריטי) כדי לחייב אתרי אינטרנט להתאים את שירותיהם לטובת ילדים, אם כי יישומו מתעכב בשל אתגר משפטי בטענה לפגיעה אפשרית בחופש הביטוי.[337] גם מדינת מרילנד ומדינות נוספות חוקקו חוקים דומים ב־2023–2024, ואלו מעידים על מגמה רחבה בארצות הברית להגברת ההגנה על ילדים ברשת ברמה המדינתית.

ממשלת אוסטרליה חיזקה את סמכויות הרגולציה להגנה על קטינים ברשת דרך חוק הבטיחות המקוונת 2021. החוק מעניק לנציבות הבטיחות המקוונת (eSafety) סמכויות חדשות להסרה מהירה של תכנים פוגעניים ולהגברת אחריותן של פלטפורמות מקוונות כלפי משתמשים צעירים.[338] כך למשל, החוק מרחיב את הסמכות לטפל בבריונות רשת נגד ילדים מעבר לרשתות חברתיות, ומחייב את התעשייה לגבש **קודים מחייבים** למניעת הפצת תוכן בלתי חוקי או בלתי ראוי, החל בחומרי התעללות מינית בילדים ועד אלימות או עירום שאינם הולמים לילדים. צעדים אלו מציבים את אוסטרליה בחזית העולמית של "עיצוב לבטיחות" (safety by design) למען מוגנות ברשת.

רגולציה וחקיקה ממלאות תפקיד משמעותי, ולעיתים אף מכריע, בהנעת פלטפורמות דיגיטליות לשנות את אופן פעולתן כלפי ילדים ובני נוער. חוק השירותים הדיגיטליים (DSA), חוק הבטיחות המקוונת בבריטניה (OSA) והקוד הבריטי לעיצוב מותאם גיל (AADC) הציבו לראשונה חובות קונקרטיות על עיצוב המערכות עצמן – לא רק בתגובה לתוכן מזיק – ובכך תרמו ליצירת תפיסה חדשה של מוגנות כמאפיין עיצובי ולא רק כהתנהגותית או חינוכית.

לחקיקה האירופית ולקוד הבריטי יש השלכות פרואקטיביות על תכנון מוצרים בפלטפורמות הגלובליות.[339]

דברי חקיקה אלו, ובראשם חוק השירותים הדיגיטליים (DSA) של האיחוד האירופי והקוד הבריטי לעיצוב מותאם גיל, מחייבות תאגידי טכנולוגיה להטמיע עקרונות של "פרטיות

337 California Age-Appropriate Design Code Act, Cal. Civ. Code §§1798.99.28-1798.99.40 (West 2023)

338 *Learn About the Online Safety Act*, eSafety Commissioner (Jan. 2024)

339 שם.

כברירת מחדל", שקיפות והגנה על קטינים כבר בשלב התכנון (privacy by design). חקיקה זו מבקשת למנוע מראש סיכונים הנובעים מהתנהלות מערכות אוטומטיות כלפי ילדים, ובכך מעבירה את מרכז הכובד מהריסון הבתר־מעשי לפיקוח מבני ומהותי מראש.

אכיפת חבילת הרגולציה הזו באירופה נעשית על ידי רשויות אכיפה לאומיות בכל מדינה חברה, אך כאשר מדובר בפלטפורמות גדולות במיוחד (Very Large Online Platforms – VLOPs) הסמכות הריכוזית נתונה לנציבות האירופית עצמה. בבריטניה האכיפה של הקוד לעיצוב ראוי לגיל מופקדת בידי נציבות המידע (ICO), שלה סמכויות חקירה והטלת קנסות. הנציבות הבריטית כבר הבהירה כי הפרת החובות האמורות כלפי קטינים עשויה להיחשב להפרה מהותית של כללי הגנת הפרטיות, גם כאשר אין פגיעה קונקרטית שניתן להוכיחה.

בשנתיים האחרונות ניתן לזהות מגמה של החמרה באכיפה כלפי תאגידי טכנולוגיה, בעיקר בהקשרים של הגנה על ילדים. כך לדוגמה, באפריל 2024 פתחה נציבות האיחוד האירופי בחקירה נגד טיקטוק בחשד שהפרה את רגולציית השירותים הדיגיטליים בכל הנוגע להנגשת תכנים בלתי הולמים לקטינים, אי־הטמעת כלים אפקטיביים לאימות גיל המשתמשים, והתעלמות מהשלכות האלגוריתם על בריאות הנפש של ילדים. עוד קודם לכן, בנובמבר 2023, קנס ה־ICO את חברת דיסני בסכום של 12 מיליון ליש"ט בגין איסוף ושימוש במידע אישי של קטינים ללא קבלת הסכמת הורים ובניגוד לקוד הבריטי. תביעות אזרחיות נוספות מתנהלות במקביל באנגליה, צרפת והולנד נגד מטא (פייסבוק ואינסטגרם) בטענה לחשיפה מופרזת של ילדים לתכנים פוגעניים ולמניפולציות רגשיות מכוונות.

תחולת החקיקה האירופית איננה מוגבלת לשטח מדינות האיחוד. היא חלה על כל פלטפורמה דיגיטלית המספקת שירותים למשתמשים באיחוד האירופי, ללא קשר למקום מושבה של החברה. בכך ממשיכה חקיקה זו את מה שמכונה "אפקט בריסל" – תופעה שבה רגולציה אירופית משפיעה על נורמות עולמיות גם מחוץ לגבולות אירופה.[340] חברות טכנולוגיה רבות, בהן אפל, גוגל ומטא, בחרו להחיל מדיניות פרטיות וזהירות ברוח ה־DSA וה־GDPR גם מחוץ לאירופה, כדי להימנע מיישום מערכות כפולות ולשמר אחידות תפעולית. מדינות

340 להרחבה בנוגע לאפקט זה ראו Anu Bradford, The Brussels Effect: How the European Union Rules the World (Oxford University Press, 2020)

נוספות, כמו קנדה, אוסטרליה ודרום קוריאה, שוקלות כעת אימוץ רגולציה דומה, ובכך מחריף הלחץ הגלובלי על עיצוב בטוח ומכבד של סביבות דיגיטליות לילדים.

ממצאי דוח *Digital Future for Children* מחזקים תובנה זו: הרוב המכריע מתוך 128 שינויים שנרשמו בארבע פלטפורמות מרכזיות (Meta, Google, TikTok, Snap) התרחשו בשנתיים שלאחר כניסת החוקים לתוקף, ובמיוחד בשנת 2021, השנה שבה נכנס הקוד הבריטי לתוקף.[341] בולטים במיוחד שינויים המוגדרים "by default". כלומר, שינויים בהגדרות ברירת המחדל של שירותים דיגיטליים, המהווים את הקטגוריה הבולטת ביותר. כלומר, כאשר קיימת רגולציה קונקרטית, מחייבת ומלווה באכיפה, פלטפורמות משנות את עיצוב השירותים שלהן בפועל.

עם זאת, הדוח מצביע גם על מגבלות. ראשית, פערי שקיפות: מרבית הפלטפורמות לא מדווחות באופן מסודר אילו שינויים בוצעו בעקבות אילו שינויי רגולציה, ומידע מהותי על צעדים שננקטו לא נגיש לחוקרים, למקבלי ההחלטות ולציבור. שנית, קיימת הטיה לטובת פתרונות טכנולוגיים וכלים להורים, שלמרות תרומתם אינם בהכרח אפקטיביים בקרב אוכלוסיות מגוונות ועלולים לעיתים לפגוע באוטונומיה של הילד עצמו. בנוסף, קיימת התמקדות ברגולציה של פרטיות וניטור תוכן, בעוד מרכיבים עמוקים יותר של רווחה נפשית, כבוד ויכולת לפתח זהות במרחב הדיגיטלי נותרים בשוליים.

ארגונים בינלאומיים מעורבים גם הם ביצירת מדיניות בנושא בינה מלאכותית. בשנת 2021 פרסמה ועדת האו"ם לזכויות הילד את ה"הערה הפרשנית מספר 25" בנוגע לזכויות ילדים בסביבה הדיגיטלית. הערה פרשנית זו מבהירה במפורש כי כל זכויות הילד (כפי שהן מעוגנות באמנה לזכויות הילד) חלות גם במרחב הדיגיטלי, וקוראת למדינות לפעול להגנה על ילדים ברשת. בין היתר מודגשת זכותם לפרטיות, לחופש ביטוי מוגן ולהגנה מפני ניצול מסחרי ופגיעות מקוונות.[342] מסמך זה מספק מסגרת עקרונית למדיניות לאומית בנושאי מוגנות ילדים בעידן הדיגיטלי.

341 Steve Wood, Impact of Regulation on Children's Digital Lives (Digital Futures for Children, May 2024)

342 General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment (Committee on the Rights of the Child, U.N. Doc. CRC/C/GC/25, Mar. 2, 2021)

UNICEF, קרן הילדים של האו"ם, הובילה גיבוש קווים מנחים ראשונים מסוגם לשילוב הגנת זכויות ילדים במדיניות ופיתוח מערכות בינה מלאכותית. במסגרת פרויקט AI for Children פרסמה יוניצף בשנת 2021 את "מדריך המדיניות לבינה מלאכותית ידידותית לילדים".[343] מדריך זה, שפותח בשיתוף ממשלת פינלנד, מגדיר תשע דרישות לפיתוח מערכות AI המתמקדות בטובת הילד. דרישות אלה כוללות תמיכה בהתפתחות וברווחת הילד; הבטחת שילוב והכלה של ילדים; מניעת אפליה וחוסר הוגנות; הגנה על פרטיות ונתוני ילדים; שמירה על ביטחונם הפיזי והנפשי; הבטחת שקיפות, הסברתיות ואחריותיות של מערכות כלפי ילדים; וכן חינוך והכשרת ילדים ומקבלי החלטות בנוגע לבינה מלאכותית והכנת הסביבה להתפתחויות עתידיות. יוניצף פיתחה גם כלים משלימים, כגון מדריכים ייעודיים להורים ולבני נוער לגבי התנהלות בסביבת בינה מלאכותית.[344]

ארגון OECD עדכן בשנת 2021 את מסמך "ההמלצה בנוגע לילדים בסביבה הדיגיטלית",[345] שבה נקבע העיקרון של **שמירת טובת הילד כשיקול ראשון** בעיצוב המרחב הדיגיטלי, ונוסחו קווי מדיניות להבטיח סביבה מקוונת בטוחה ומיטיבה לילדים. ההמלצה קוראת לאסדרת סיכונים מבוססת מידע ומידתית וליישום גישות של עיצוב לבטיחות בפלטפורמות דיגיטליות המיועדות לצעירים. הארגון מדגיש את האוריינות הדיגיטלית כתנאי חיוני להעצמת בטיחות ילדים במרחב המקוון.

על אף התרומה החשובה של הארגונים הבינלאומיים למסגור נורמטיבי ולגיבוש עקרונות פעולה, קשה לקבוע עד כמה יש להם השפעה ישירה על עיצוב מדיניות או שינויים בפועל ברמת מדינה או פלטפורמה. מסמכים כמו ההצהרה הכללית של ועדת האו"ם או מדריך UNICEF מספקים מסגרות ערכיות חיוניות, שמשמשות בסיס מוסרי או עיוני למדיניות מקומית אך לעיתים קרובות נותרות ברמת ההמלצה או ההנחיה. במקרים מסוימים ניתן לזהות השפעה מתגלגלת: הקווים המנחים של UNICEF אומצו במסמכי מדיניות לאומיים, למשל בפינלנד ובחלק ממדינות אירופה, והמסמך של OECD שימש נקודת ייחוס בדיונים

343 Policy Guidance on AI for Children (Version 2.0, UNICEF Office of Global Insight and Policy, United Nations Children's Fund, 2021)

344 AI Guide for Teens (UNICEF Office of Global Insight and Policy, United Nations Children's Fund, Nov. 2021)

345 Recommendation of the Council on Children in the Digital Environment (OECD, OECD/LEGAL/0389, adopted May 31, 2021)

רגולטוריים באיחוד. עם זאת, ההטמעה בפועל אינה תמיד שקופה, ואין ביסוס מוסדי משפטי מחייב שיאפשר לאותם עקרונות להפוך לכלים אופרטיביים. ברוב המקרים הארגונים הבינלאומיים משמשים מצפן מוסרי ומנוע ללחץ ציבורי, אך לא גורם רגולטורי ישיר. השפעתם הגדולה ביותר ניכרת כאשר הם פועלים בשותפות עם ממשלות, ארגוני חברה אזרחית ותעשיית הטכנולוגיה – ובכך הם מצליחים לגשר על הפער בין הצהרה לבין מנגנון פעולה.

ב. פעילות ארגוני מגזר שלישי וחברה אזרחית

ארגוני מגזר שלישי וחברה אזרחית בולטים מעורבים גם הם ביצירת סוגים שונים של התערבות ומדיניות. Thorn הוא ארגון ללא מטרות רווח שבסיסו בארצות הברית, הפועל להגנה על ילדים מפני התעללות וניצול מיני ברשת באמצעות טכנולוגיה מתקדמת. הארגון מפתח כלים טכנולוגיים לזיהוי והסרה של תוכני ניצול והתעללות (כגון מערכת Safer לזיהוי והורדת תוכן פדופילי עבור חברות אינטרנט) ומשתף פעולה עם רשויות אכיפה. נוסף לפתרונות הטכנולוגיים, Thorn פועל להעלאת המודעות בקרב הורים וצעירים, במטרה לצייד אותם בידע ובכלים להתמודדות עם סיכונים מקוונים. תוכניות כמו Thorn for Parents ו־NoFiltr נועדו לספק מידע מעשי על דרכים לזיהוי מניפולציות מקוונות ולמנוע מצבים של ניצול. תחום פעילות נוסף של הארגון הוא סיוע באיתור מהיר של ילדים שנמצאים בסיכון. באמצעות כלים מתקדמים ושיתוף פעולה עם רשויות אכיפת החוק וארגונים חברתיים, Thorn מסייע בזיהוי קורבנות והבאתם למקום בטוח. כמו כן, הארגון מקדם מחקרים חדשניים בתחום בטיחות ילדים בעולם הדיגיטלי, על מנת להמשיך ולפתח פתרונות אפקטיביים ולעדכן אסטרטגיות פעולה בהתאם לשינויים הטכנולוגיים. מעבר לכך Thorn פועל להשפיע על מדיניות ציבורית באמצעות שיתוף פעולה עם מחוקקים וגופים ממשלתיים בארצות הברית ובמדינות נוספות, במטרה לקדם חקיקה שמחזקת את ההגנה על ילדים מפני פגיעה מקוונת.

בשנים האחרונות מוביל Thorn יוזמות מיוחדות בתחום הבינה המלאכותית: בשנים 2023–2024 השיק יחד עם הארגון All Tech Is Human קמפיין Safety by Design לגבי בינה מלאכותית יוצרת, שקיבל את תמיכתן של חברות AI מובילות. במסגרת זו חברות כמו אמזון, גוגל, מטא, מיקרוסופט, OpenAI ואחרות התחייבו לאמץ עקרונות עיצוב טכנולוגי למניעת יצירה והפצה של תוכן פוגעני **גנרטיבי** נגד ילדים, בדגש על מניעת ייצור תמונות

והתעללות מינית בילדים.[346] חברות אלה דיווחו בשקיפות על צעדים שננקטו כדי ליישם את העקרונות.[347] יוזמה זו מייצגת שיתוף פעולה בין המגזר השלישי לתעשייה.

5Rights Foundation היא קרן בריטית מובילה בקידום זכויות ילדים בעולם הדיגיטלי. הקרן, שעמדה מאחורי פיתוח "קוד ההתאמה לגיל" בבריטניה, פרסמה בשנת 2023 מסמך מקיף בשם *Children & AI Design Code* – קוד עיצוב למערכות AI שבהן מעורבים ילדים. קוד זה מציע פרוטוקול מעשי לפיתוח ולהפעלת מערכות בינה מלאכותית **תוך מתן עדיפות לזכויות ולצרכים של ילדים כבר משלבי התכנון הראשוניים, כגון מותאמות התפתחותית, בטיחות, הוגנות, שקיפות ואחראיות**. התפיסה העומדת מאחרי הקוד היא שקול הצרכים של ילדים אינו נשמע בעיצוב מדיניות וכן בעיצוב מוצרים מבוססי AI.[348] כלי זה מוכר ככלי יישומי הן עבור התעשייה והן עבור רגולטורים לאימוץ גישת "אחריות מגיל אפס" בבינה מלאכותית.

WeProtect Global Alliance היא קואליציה בינלאומית רבת־עוצמה שנוסדה ביוזמת בריטניה ומאחדת למעלה מ־100 גופים – ובהם ממשלות דמוקרטיות, חברות טכנולוגיה מובילות וארגוני חברה אזרחית – במאבק בהתעללות ובניצול מיני של ילדים ברשת. הברית פועלת לקידום שיתוף פעולה חוצה־מדינות ומגזרים ומפתחת מדיניות ופתרונות טכנולוגיים להגנה על ילדים מפני פגיעות מקוונות. היא מפרסמת מדי שנתיים דוחות הערכת איומים גלובליים של מיפוי היקף ועוצמת הסיכונים החדשים, כגון ניצול לרעה של פלטפורמות ומגמות טכנולוגיות (כולל בינה מלאכותית) בידי גורמים המנסים לפגוע בילדים.[349] הארגון מגבש גם קווים מנחים ומסגרות וולונטריות עבור תעשיית הטכנולוגיה, וייחודו בכך שהוא מאגד מומחים מממשלות, מהמגזר הפרטי ומהחברה האזרחית, הפועלים יחדיו לפיתוח פתרונות מורכבים להגנה על ילדים ברשת. בשנת 2023 הובילה הברית הצהרה משותפת על הצורך להתמודד עם ניצול מיני של ילדים בעידן הבינה המלאכותית בכנס AI Safety Summit בלונדון.[350]

346 *Safety by Design for Generative AI: Progress Reports*, Thorn (Mar. 2025)

347 שם.

348 Derek E. Baird, *Safeguarding Children in the Age of AI: A New Design Code*, Medium (Mar. 19, 2025)

349 Annual Review 2023 (WeProtect Global Alliance, Dec. 20, 2023)

350 שם.

ארגוני החברה האזרחית ממלאים תפקיד ייחודי ומשלים בהגנה על ילדים בעידן הבינה המלאכותית. הם אינם פועלים מתוך סמכות אלא מגשרים בין אתוס חברתי לבין השדה הטכנולוגי, ובין מחקר לבין יישום. יתרונם הגדול טמון ביכולתם להגיע לרמות השטח, אל הורים, מחנכים, בני נוער וילדים, באמצעות שפה נגישה, כלים מעשיים ותמיכה קהילתית. נוסף על כך, ארגונים כמו 5Rights, Thorn ו־WeProtect לא רק מגיבים למציאות משתנה, אלא מובילים את הלחץ על רגולטורים בטרם האחרונים בכלל מבינים שעליהם לפעול.

עם זאת, יכולתם של ארגונים אלה לפעול לאורך זמן תלויה במידה רבה במימון פילנתרופי, בשיתופי פעולה עם ממשלות, ובנכונות תעשיית הטכנולוגיה לשתף פעולה. חשוב גם להבחין בין ארגונים בעלי גישה מערכתית־גלובלית לבין יוזמות מקומיות קטנות יותר, שתרומתן חשובה אך כוחן מוגבל. למרות מגבלות אלו, התרומה של המגזר השלישי להעמקת השיח, לחשיפת כשלים ולהצעת פתרונות חדשניים היא הכרחית. לעיתים קרובות, הם הראשונים להצביע על סיכונים שאינם מזוהים על ידי מערכות המדינה או השוק ולהעלות את קולם של ילדים ונוער במקומות שבהם אינו נשמע.

ג. פעילות מצד חברות הטכנולוגיה

גם חברות הטק העולמיות שותפות במיזמי הגנה על ילדים. ענקית הטכנולוגיה גוגל נוקטת בשנים האחרונות שורה של צעדים גלובליים לשיפור מוגנות ילדים ובני נוער בשירותיה, לעיתים מעבר לנדרש בחוק. בשנת 2021 הכריזה גוגל על הגדרות ברירת־מחדל חדשות המותאמות לצעירים: במנוע החיפוש הפעלת מסנן SafeSearch באופן אוטומטי למשתמשים מתחת לגיל 18; בפלטפורמת YouTube שינוי ברירת המחדל להעלאת סרטונים עבור גילי 13-17 **מציבורי לפרטי**, כך שסרטון יהיה גלוי רק למשתמש וחבריו המורשים, אלא אם יבחר ידנית להפיצו פומבית.[351] גוגל מפעילה תזכורות להפסקת צפייה והגבלת הפעלה אוטומטית (autoplay) כברירת מחדל עבור בני נוער.[352] במקביל, גוגל גם החמירה את ההגבלות על פרסום מותאם אישית וחסמה את האפשרות למפרסמים לטרגט קטינים מתחת לגיל 18 על בסיס גיל, מין או תחומי עניין.[353] בנוסף, היא מאפשרת לקטינים

351 Sarah Perez, *Google to Introduce Increased Protections for Minors on Its Platform, Including Search, YouTube and More*, TechCrunch (Aug. 10, 2021)

352 שם.

353 שם.

ולהוריהם לבקש הסרה של תמונותיהם מתוצאות חיפוש גם כשמדובר בתמונות ציבוריות, וצמצמה תוכן שיווקי בפלטפורמת YouTube Kids לאור ביקורת מומחי ילדים. צעדים רוחביים אלו, הנאכפים גלובלית, משקפים יישום עקרונות עיצוב לבטיחות במוצרי גוגל.

חברת מטא, המפעילה את הרשתות החברתיות הגדולות פייסבוק ואינסטגרם, הצהירה גם היא על יישום שינויים נרחבים לשיפור פרטיות ובטיחות בני נוער בפלטפורמות שלה. אינסטגרם הנהיגה כבר ב־2021 הגבלות על אינטראקציה בין מבוגרים לנוער (למשל חסימת שליחת הודעות פרטיות ממבוגר לנער שאינו עוקב אחריו), והגדרות חשבון פרטי כברירת מחדל לכל משתמש חדש מתחת לגיל 16.[354] בשנת 2023 הכריזה מטא על מודל חשבונות בני נוער (teen accounts) אחיד בכל מוצריה: כל משתמש מתחת לגיל 18 ישויך אוטומטית לחשבון פרטי ומוגן כברירת מחדל, כאשר בני פחות מ־16 לא יוכלו להפוך חשבון לציבורי ללא אישור הורה.[355] בנוסף, ולאור ביקורת ציבורית נוקבת, מטא הגבילה את יכולות המיקוד הפרסומי למתבגרים: משנת 2022 נאסר על מפרסמים בפייסבוק/אינסטגרם להשתמש בנתוני תחומי עניין של קטינים למיקוד מודעות, ומ־2023 גם מין המשתמש הוצא מפרמטרי המיקוד, כך שנותרו כהגדרות מותרות רק הגיל והמיקום הכללי שלהם.[356] מטא גם פיתחה כלים מבוססי בינה מלאכותית לאימות גיל (age verification) על מנת לאתר קטינים המתחזים לבגירים ולהחיל עליהם את ההגנות הנדרשות.[357] כאמור, צעדים אלה נועדו להתמודד עם לחץ ציבורי ואיומי רגולציה במדינות רבות.

חברת Microsoft תורמת לפיתוח כלי בינה מלאכותית לזיהוי תוכן פוגעני כמו ניצול ילדים (PhotoDNA), ובשנים האחרונות ממלאת תפקיד מוביל בהגנה מפני איומים חדשים בעידן הבינה המלאכותית. בשנת 2023 מינתה החברה קצין ראשי לבטיחות דיגיטלית, וב־2024 הצטרפה ליוזמת Safety by Design שהובילו Thorn ו־All Tech Is Human להגנה

354 Emily Tsiao, *Parental Controls: Instagram*, Plugged In (Aug. 2, 2024)

355 Max Zahn, *Instagram Imposes New Restrictions for Teens. Will They Work?*, ABC News (Sept. 18, 2024)

356 Nicole Farley, *Meta Introduces New Ad Targeting Limits for Teens*, Search Engine Land (Jan. 11, 2023); *2023 U.S. Advertising and Privacy Trends and 2024 Forecast: Focus on Kids and Teens*, Keller and Heckman (Dec. 20, 2023)

357 Max Zahn, *Instagram Imposes New Restrictions for Teens*, ABC News (Apr. 28, 2025)

מפני ניצול לרעה של מודלי AI גנרטיביים. מיקרוסופט התחייבה לאמץ עקרונות **מניעה פרואקטיבית** במערכות ה־AI הגנרטיביות שלה כדי למנוע הפקה והפצה של תוכן פוגעני הקשור בפגיעה מינית בילדים.[358] לפי הצהרת החברה, יוזמות כאלה עולות בקנה אחד עם הגישה הרחבה שלה לבנות ארכיטקטורת בטיחות חזקה כבר משלב התכנון, לצד שיתוף פעולה הדוק בין התעשייה, הממשלות והחברה האזרחית בהתמודדות עם תוכן והתנהגות פוגעניים.[359] בנוסף, מיקרוסופט תומכת בארגונים כגון המרכז הלאומי לילדים נעדרים ומנוצלים.

חברת Apple מתמקדת בפיתוח פתרונות טכנולוגיים המשלבים הגנה אוטומטית על ילדים בתוך המכשירים והשירותים שלה. דוגמה בולטת לכך היא הפיצ׳ר "Communication Safety" שהושק בדצמבר 2021 במערכות iOS, iPadOS ו־macOS, העושה שימוש באלגוריתמי **למידת מכונה על המכשיר** כדי לזהות אם תמונות או סרטונים הנשלחים או מתקבלים באפליקציית ההודעות (Messages) על ידי ילד מכילים **עירום או פורנוגרפיה**. במקרה שתוכן כזה מזוהה, המערכת מטשטשת את התמונה באופן אוטומטי ומציגה לילד אזהרה והדרכה – כולל מסר שהוא לא חייב לצפות בתוכן – ומציעה לו לפנות למבוגר אמין לעזרה.[360] בניגוד לגישות אחרות, הפתרון של אפל שומר על פרטיות מלאה: הזיהוי נעשה באופן מקומי ללא שליחת מידע לשרתי החברה, ואפל או הורי הילד אינם מקבלים דווח על האירוע, אלא רק הילד מקבל את ההתראה. אפל שילבה מנגנון זה גם באפליקציות נוספות כגון AirDrop, FaceTime וספריית התמונות במכשיר והרחיבה אותו למדינות נוספות בשנים 2022-2023.[361] לצד זאת, העוזרת הקולית Siri ומנוע החיפוש במכשירי אפל הועשרו בהנחיות מותאמות לצעירים בנושאי בטיחות ברשת (למשל מענה לשאלה "כיצד לדווח על ניצול ילדים").

358 Courtney Gregoire, *Microsoft Joins Thorn and All Tech Is Human to Enact Strong Child Safety Commitments for Generative AI*, Microsoft On the Issues (Apr. 23, 2024)

359 שם.

360 *Child Safety*, Apple Inc. (2025)

361 Michael Potuck, *Apple Expanding Communication Safety in Messages for Kids to Six New Countries*, 9to5Mac (Feb. 20, 2023)

ד. יוזמות לקידום חינוך ואוריינות לילדים

מסגרות חינוך, בריאות ורווחה יכולות להטמיע עקרונות מוגנות בשגרה היומיומית של צעירים, ובכך לגשר בין מדיניות־על לבין התנהלות מעשית. ברמה זו ניתן לעצב סביבות לימוד וטיפול בטוחות, להגביר מודעות לסיכונים (כמו מידע כוזב או פגיעות מקוונות) ולצייד ילדים ובני נוער בכלים ובכישורים לנווט בעולם טכנולוגי באופן אחראי ומוגן. גישה מוסדית מאפשרת להקנות ערכים והרגלי שימוש בטוח בבינה מלאכותית מגיל צעיר, מתוך התאמה לגילים שונים ולצרכים השונים של ילדים לעומת בני נוער.

בפינלנד מוטמעת מוגנות דיגיטלית כחלק בלתי נפרד מתוכנית הלימודים הלאומית. מערכת החינוך הפינית מלמדת התמודדות עם מידע שגוי ומניפולציות מדיה כבר בבית הספר היסודי,[362] מתוך הבנה שהמיומנויות הללו הן הבסיס לביטחון דיגיטלי. בחטיבות הביניים והתיכונים בפינלנד אוריינות מידע רב־ערוצית וחשיבה ביקורתית משולבות באופן רוחבי בכל המקצועות כחלק מליבת תוכנית הלימודים.[363] כך למשל, תלמידים לומדים בשיעורי מתמטיקה כיצד ניתן לעוות ניתוחים סטטיסטיים, ובשיעורי אומנות איך תמונה עלולה להטעות בשל הקשר או עריכה. התוכנית הפינית מקושרת להצלחה ברמה הלאומית: פינלנד דורגה כמדינה העמידה ביותר באירופה בפני חדשות כזב,[364] הוכחה לכך שחינוך למוגנות מידע תורם לעמידות החברה מפני השפעות שליליות של בינה מלאכותית על ערעור יכולות בירור המציאות.

בפולין הוטמעה במערכת החינוך הפורמלית תוכנית לאומית לשימוש אחראי בבינה מלאכותית. משרד החינוך והמדע הפולני משתף פעולה עם חברת אינטל מאז 2020 בהפעלת התוכנית AI for Youth,[365] שבמסגרתה פותחה תוכנית לימודים להכשרת מורים לתלמידים בגילי 8–19 להבנת בינה מלאכותית בשילוב שבין הקניית כישורים טכניים (כגון תכנות, מדעי הנתונים ולמידת מכונה) לבין מרכיבי אוריינות כמו אתיקה של בינה מלאכותית

362 Jon Henley, *Fact from Fiction: Finland's New Lessons in Combating Fake News*, The Guardian (Jan. 28, 2020)

363 שם.

364 שם.

365 *AI for Youth Program*, Intel Corporation

והפחתת הטיות אלגוריתמיות.[366] עד כה הוכשרו במסגרת זו כ־11,000 מורות ומורים בבתי ספר יסודיים ועל־יסודיים ברחבי פולין. יוזמה מוסדית רחבת היקף זו, הממומנת חלקית גם על ידי האיחוד האירופי, מדגימה כיצד שיתוף פעולה רב־מגזרי יכול להטמיע באופן שיטתי הוראה על שימוש בטוח ואחראי בבינה מלאכותית.

חינוך אתי־טכנולוגי בעקבות אירועי פגיעה הופך גם הוא למקובל. באחד מהתיכונים בקליפורניה, ארצות הברית, כונסה סדנה מיוחדת לאחר שהתברר שתלמיד השתמש בבינה מלאכותית כדי ליצור ולהפיץ תמונות עירום מבוזות של תלמידות בבית הספר. הנהלת בית הספר יזמה פאנל לימודי בהשתתפות מומחים, מורים ותלמידים, שבו נדונו סוגיות של פרטיות, שימוש הולם בטכנולוגיה, וההיבטים החוקיים והאתיים של שיתוף תוכן דיגיטלי פוגעני.[367] מטרת הסדנה הייתה להפוך אירוע שלילי להזדמנות לימודית: להדריך לא רק את התלמידים המעורבים באירוע, אלא גם להגביר מודעות בקרב כלל התלמידים בנוגע לאחריות האישית והמוסרית הנדרשת בסביבה רוויית כלים טכנולוגיים מתקדמים.[368] מקרה בוחן זה שהתפרסם והפך למודל לחיקוי מדגים כיצד מוסדות חינוך יכולים להגיב לאיומים כמו שיימינג מקוון בכלי בינה מלאכותית באמצעות חינוך למוגנות ודיאלוג, ולא רק באמצעות ענישה. ראוי לציין, עם זאת, שסקרים בקרב צוותי חינוך בארצות הברית מראים שתופעות אלו עדיין לא מוכרות לרבים: למעלה ממחצית מהמורים והמנהלים דיווחו שלא קיבלו כל הכשרה להתמודדות עם סיכוני AI כגון דיפ־פייק, או שההכשרה שהועברה להם הייתה באיכות נמוכה.[369]

על רקע הסכנות שמציבה בינה מלאכותית לילדים הוקם הארגון MIT RAISE (בינה מלאכותית אחראית להעצמה וחינוך חברתית) המתמקד בשילוב בינה מלאכותית בהקשרים חינוכיים. הארגון מפתח משאבים ותוכניות לימודים לילדים על בינה מלאכותית באופן שמקדם הכלה וחשיבה ביקורתית. התוכניות כוללות סדנאות המיועדות לכל מיני רמות חינוכיות, שבהן התלמידים עוסקים בטכנולוגיות ולומדים על ההשלכות שלהן.[370]

366 שם.

367 Olina Banerji, *Why Schools Need to Wake Up to the Threat of AI "Deepfakes" and Bullying*, Education Week (Dec. 9, 2024)

368 שם.

369 שם.

370 *Responsible AI for Social Empowerment and Education*, Mit Raise

אחת הדרכים המרכזיות לחיזוק מוגנות היא שילוב תכנים על סיכוני AI בתוכניות חינוך ואוריינות מדיה. למשל, במדינת פלורידה (ארצות הברית) עודכנה חובת הלימודים בכיתות ו׳-י״ב כך שתכלול את הנושאים **השפעת המדיה החברתית על בריאות הנפש, התפשטות מידע כוזב ברשת, ואופן הפעולה של אלגוריתמים המעצימים התנהגויות מסוכנות.**[371] גם במדינות אחרות בארצות הברית ניתן דגש דומה על פיתוח חשיבה ביקורתית כבר בבית הספר היסודי, כדי שילדים יוכלו לזהות תוכן מזויף המופק באמצעות בינה מלאכותית ולהבחין בין מידע אמין למניפולציות. מתקיימים אירועים ייעודיים לתלמידים – דוגמת האירוע השנתי MisinfoDay שבו מאות בני נוער מתרגלים זיהוי תמונות שנוצרו ב־AI לעומת תמונות אמיתיות.[372] יוזמות חינוכיות אלה, המותאמות לגיל (אוריינות בסיסית לילדים צעירים לעומת דיון עמוק בקרב מתבגרים), מחזקות את "חוסן המידע" של הדור הצעיר בעידן הפוסט־אמת.

מרבית מדינות אירופה וה־OECD שואפות לשילוב מיומנויות דיגיטליות בתוכניות הלימודים, מגדירות אוריינות דיגיטלית כמקצוע חובה, מפרסמות תוכניות בנושא ופועלות ברמה הלאומית, בדגש על הגברת מודעות ומתן כלים פרקטיים לציבור, שיאפשרו לו לנווט בסביבת המידע בצורה אפקטיבית ובטוחה.[373] חלקן נמצאות בתהליך של רפורמה ועדכון לעידן הבינה המלאכותית.[374]

תוכנית PISA 2029 של ה־OECD כוללת לראשונה מסגרת הערכה ייעודית ל־Media and AI Literacy (MAIL), המגדירה את האוריינות כיכולת לפעול באופן אפקטיבי, ביקורתי ואתי בתוך סביבות מידע המושפעות מבינה מלאכותית. מודל ה־MAIL מחלק את האוריינות למכלול כשירויות רחב הכולל שימוש וגישה לכלים, ניתוח והערכה ביקורתית, השתתפות ושיתוף פעולה, יצירה וכן רכיב רוחבי של חשיבה אתית. מעבר לכך, התוכנית מדגישה כי מדובר בכשירות רגשית־חברתית, כגון מודעות עצמית, קבלת החלטות אחראית והבנת

371 Samia Alkam & Daniela DiGiacomo, *Advancing Policy to Foster K-12 Media Literacy*, NASBE (Sept. 2024)

372 Kim Malcolm, *AI Images and Conspiracy Theories Are Driving a Push for Media Literacy Education*, WBUR (Mar. 21, 2024)

373 Lee Edwards, Sonia Livingstone, & Emma Goodman, *Putting Media Literacy on The Map: Opportunities and Challenges in Europe*, Media@LSE (December 3, 2024)

374 PISA 2029, לעיל ה"ש 216.

ההשפעה של מערכות AI על יחסים חברתיים. לכן מסגרת ה-MAIL מציעה מטלות המדמות סביבה דיגיטלית הכוללת גלישה ברשת, מנוע חיפוש, ואף כלי שיחה (מעין צ׳אטבוט) שמספק מידע או הנחיות. בתוך סביבה זו התלמידים נדרשים לבצע משימות כגון חיפוש מידע, השוואת מקורות, הערכת אמינות של תוכן וזיהוי הטיות או מניפולציות. ההערכה לפי המודל מתמקדת בתהליך: כיצד תלמידים מנסחים שאילתות, אילו מקורות הם פותחים, כיצד הם נעים בין טאבים, ואיך הם משלבים בין מידע ממקורות שונים לבין פלט של מערכות בינה מלאכותית.[375]

בישראל, משרד החינוך אומנם מכיר בחשיבות של אוריינות מידע ומדיה ובצורך לשלב מיומנויות אלה בתוכניות הלימודים ובהכשרות מורים, אך בשונה מאירופה, ההוראה והלמידה של אוריינות דיגיטלית אינה מוכרת כמקצוע חובה, ולא קיימת תוכנית לימודים מפורטת המציעה דרכים לשילוב אוריינות דיגיטלית בתוכניות הלימודים של מקצועות הלימוד. בעקבות זאת, שיעור התלמידים שלומדים כיצד להעריך אמינות של מידע מקוון בישראל הוא נמוך מממוצע ה־OECD ונמוך מאוד בהשוואה למדינות מערביות מפותחות.[376] במקביל למערכת החינוך הפורמלית, גופים ציבוריים וארגוני בטיחות מפיצים הנחיות המיועדות ישירות להורים ולילדים בנושאי בינה מלאכותית. באוסטרליה, למשל, נציבות eSafety פרסמה מדריך מקיף על הסכנות של צ׳אטבוטים מבוססי AI, לאחר שהתברר כי ילדים ובני נוער מנהלים שיחות יומיומיות עם ״חברים״ וירטואליים העלולות לגלוש לתכנים מיניים או לעודד פגיעה עצמית.[377] הנציבות ממליצה להורים לשוחח בגלוי עם ילדיהם על חוויותיהם, להסביר להם שצ׳אטבוט אינו תחליף לחבר אמיתי ולהציב גבולות ברורים לזמן השימוש באפליקציות כאלה. גם בבריטניה הופצו הנחיות דומות: ארגון Childnet התריע שבוט הצ׳אט החדש My AI באפליקציית סנאפצ׳ט, אף שיש בו תועלות, עלול להגביר תחושות בדידות אצל מתבגרים ולהציג מידע שגוי או מוטה. הארגון קרא לשים דגש על פרטיות (ללמד צעירים לא לשתף מידע אישי עם הבוט) ולהבהיר לצעירים

375 שם.

376 21st-Century Readers Developing Literacy Skills in a Digital World (OECD, 2021); שקד דברן ואיילת ברעם צברי **תופעת הפוסט אמת, הגדרות השלכות ופתרונות חינוכיים** (מוסד שמואל נאמן למחקר מדיניות לאומית, 2023).

377 *AI Chatbots and Companions – Risks to Children and Young People*, eSafety Commissioner (Feb. 18, 2025)

שהשיחות עם AI מוגבלות בהבנתן ואינן תחליף לקשרים אנושיים אמיתיים.[378] בעזרת הנחיות כאלה, הורים ומטפלים מצוידים בכלים לשיח חינוכי עם הילדים על שימוש אחראי וזהיר בטכנולוגיות חדשות, במקום להימנע מהן לחלוטין.

ממשלות החלו לספק ערכות מידע וכלי עזר ישירים למשפחות, בהבחנה בין צורכי ילדים צעירים לעומת בני נוער. דוגמה בולטת היא הדוח "פרקטיקות למשפחות" שפרסם הבית הלבן בארצות הברית ב־2024, אשר מיפה את הסיכונים העיקריים לילדים ברשת והמליץ על דרכי התמודדות מעשיות להורים.[379] הדוח לווה במשאבים ייעודיים: מסמך אחד כולל אסטרטגיות פרקטיות לחיזוק חשיבה ביקורתית ושליטה עצמית בקרב ילדים צעירים (גילאי גן ויסודי), במטרה "לבנות מערכת יחסים בריאה עם המדיה הדיגיטלית". מסמך נוסף מספק "מתניעי שיחה" ודוגמאות להתמודדות עם מצבים נפוצים עבור הורי ילדים בוגרים יותר ומתבגרים – למשל כיצד לדון יחד על לחץ חברתי ברשת או על תכנים בעייתיים שהאלגוריתם ממליץ עליהם.

לצד החינוך והמניעה, קיימים גם כלים טכנולוגיים המאפשרים התערבות אישית לצמצום נזקי AI לאחר שהתרחשו. דוגמה לכך היא היוזמה Take It Down של המרכז האמריקאי לילדים נעדרים ומנוצלים (NCMEC). פלטפורמה זו, שהושקה בסוף 2022, מאפשרת לבני נוער (ואף ילדים צעירים בעזרת מבוגר) להעלות באופן אנונימי "טביעת אצבע" דיגיטלית של תמונות עירום או סרטונים מיניים שלהם (אמיתיים או יצירי מכונה כדי שיזוהו ויימחקו ממאגרי אתרים חברתיים לפני גרימת הנזק.[380] כתובות מסוג זה לפנייה לעזרה מיידית מחזירות לנפגעים שליטה מסוימת במצב והן משמשות רשת ביטחון חשובה בשעת משבר.

פעולה מערכתית מציעה יתרונות משמעותיים בקידום מוגנות. מוסדות חינוך, בריאות ורווחה נמצאים בעמדה המאפשרת להגיע לכלל האוכלוסייה הצעירה באופן שוויוני

378 *How Does Snapchat's New AI Function "My AI" Impact Young People?*, UK Safer Internet Centre (May 25, 2023)

379 *Biden-Harris Administration Takes Actions to Advance Kids' Online Health, Safety, and Privacy*, White House Office of Science & Technology Policy (July 22, 2024)

380 Emma Henderson Vaughan, *NCMEC Launches New Service That Can Help You "Take It Down"*, NCMEC Blog (Feb. 27, 2023); Rita Garcia, *National Nonprofit Continues Helping Teens Scrub Deepfakes or Nonconsensual Explicit Images as Part of "Take It Down" Initiative*, ABC13 (Nov. 13, 2024)

ולהטמיע מסרים של שימוש בטוח והוגן בטכנולוגיות AI באופן עקבי ומתמשך. יוזמות מוסדיות יכולות ליצור תרבות ארגונית של שקיפות, הגינות והטמעת החשיבות של אחריות בשימוש במערכות אלגוריתמיות.[381] בנוסף, שיתופי פעולה בין־מגזריים תומכים ביוזמות אלו: ארגוני חברה אזרחית, אקדמיה ותעשייה יכולים לספק למוסדות כלים ומשאבים עדכניים. כך למשל, ארגון Common Sense Media בשיתוף עם חברת OpenAI מפתח מערכי הדרכה ומערכת דירוג של כלים מבוססי AI עבור הורים, אנשי חינוך וצעירים, במטרה להבהיר את הסיכונים והתועלות הטמונים בטכנולוגיה זו.[382]

עם זאת, ישנם גם אתגרים לא מבוטלים. יש צורך בהכשרת צוותים ובהעלאת המודעות של מורים, צוותי בריאות ופסיכולוגיה ועובדים סוציאליים להתפתחויות המהירות בתחום הבינה המלאכותית, משימה המצריכה השקעת משאבים וזמן. כמו כן, קיים פער בין מוסדות מובילים שמאמצים תוכניות חדשניות לבין אחרים שטרם פיתחו מענה לנושא, דבר העלול להותיר חלק מהילדים ללא הגנה מספקת. לבסוף, מוסדות צריכים לאזן בין ניצול יתרונות הבינה המלאכותית (כגון התאמה אישית של למידה או איתור מוקדם של סימני מצוקה) לבין שמירה קפדנית על פרטיות, שוויון והיבטים אתיים.

התערבויות ברמה הפרטנית מציעות גישה גמישה ומותאמת־אישית להגנת ילדים: הן מחנכות, מגבירות מודעות, מציידות ילדים בכלים פרקטיים לנווט בעולם דיגיטלי רווי יישומי בינה מלאכותית ומקדמות אותם להיות "סוכני בטיחות" בעצמם – להבין סימני אזהרה, לקבל החלטות נבונות ולדעת לבקש עזרה בעת הצורך. התערבויות מסוג זה הן גם מיידיות יותר מהליכי חקיקה, וניתן לעדכן אותן במהירות בהתאם להתפתחות הטכנולוגיה.

ה. סיכום והערכת מדיניות ויוזמות קיימות

הסקירה לעיל ממחישה כי בשנים האחרונות חלה התעוררות ממשית בזירה הבינלאומית, הלאומית והאזרחית בנוגע לצורך להגן על ילדים מפני סיכונים המתפתחים במקביל להתקדמות הבינה המלאכותית. ניתן לזהות תנועה חיובית במגוון רמות: חקיקה

381 *Fostering Responsible and Safe AI Implementation in K-12 Education: A Policy Brief for Teachers and Principals*, Merlyn Mind (June 21, 2024)

382 Jonathan Vanian, *OpenAI is Working on AI Education Safety Initiative with Common Sense*, CNBC (Jan. 29, 2024)

מתקדמת באירופה, רגולציה ממוקדת גיל, קודים אתיים של ארגוני חברה אזרחית, כלים טכנולוגיים חדשים, פלטפורמות הדרכה למשפחות ותוכניות חינוך והכשרה. כל אלה מבטאים תפנית תפיסתית – מהתמקדות בתוכן מזיק להתמודדות עם מאפייני המערכות עצמן.

עם זאת, לצד הישגים אלה, ניכרים פערים מבניים ותוכניים שעדיין אינם מאפשרים לספק מענה שלם לאתגרי המוגנות בעידן הבינה המלאכותית. ראשית, מרבית היוזמות מתמקדות בסוגי פגיעוּת מסוימים כמו פרטיות או חשיפה לתוכן מיני, ואינן נותנות מענה מספק לפגיעות רגשיות, קוגניטיביות וחברתיות. למשל, תופעות כמו קשרים אמביוולנטיים עם צ׳אטבוטים, פגיעות בדימוי עצמי בשל הטיות אלגוריתמיות או שיבוש של תהליכי חברות אינן מקבלות מענה על ידי הפתרונות הקיימים.

בנוסף, קיימת הטיה מוסדית לעבר חינוך פורמלי וטכנולוגיות של סינון ושליטה. יוזמות רבות עוסקות בבתי ספר, שאומנם הם מסגרת קריטית, אך אינן יוצאות מהם למסגרות נוספות כמו תנועות נוער וחוגים אקסטרה־קוריקולריים.[383] במקביל, נעדרת השקעה משמעותית בכלים לבניית תמיכה רגשית ומערכתית בילדים שנפגעו, במיוחד כאשר מדובר בפגיעות ״שקטות״ שאינן מדווחות מיידית, כמו השפעות מתמשכות של שיח בעייתי עם סוכן AI רגשי.[384]

במונחי טווח גילים, ניכר לעיתים ערבוב בין צורכי ילדים לצורכי בני נוער מבלי לעצב מענה נבדל לשתי הקבוצות. פתרונות המוצעים לילדים בני 9 אינם תמיד רלוונטיים או מספקים עבור מתבגרים בגיל 15. השונות קיימת ברמת ההצהרות – אך פחות ברמת העיצוב הממשי של ההתערבות.

מוסדות חינוך, במיוחד במדינות שבהן לא נקבעה עדיין תוכנית לאומית, נותרים ללא הכשרה מספקת וללא תמיכה מוסדית. מורים, יועצים, עובדים סוציאליים ורופאי ילדים מדווחים על חוסר ידע וכלים להתמודדות עם סוגיות חדשות כמו יצירת תוכן פוגעני על

383 **הנחיות שימוש בכלי בינה מלאכותית יוצרת במערכת החינוך** (התוכנית הלאומית לבינה מלאכותית, ינואר 2026).

384 Will Oremus, *Kids Are Talking to "AI Companions." Lawmakers Want to Regulate That*, Wash. Post (Apr. 1, 2025)

ידי בינה מלאכותית, זיהוי תוכן יציר מכונה או תלות רגשית של תלמיד בצ'אטבוט.[385] המשמעות היא שהמענה בשטח לעיתים מקרי ותלוי יוזמה מקומית.

לבסוף, יש לציין כי התערבויות רבות הן תגובתיות ולא פרואקטיביות: הן מופעלות לאחר אירוע חמור (כמו הפצת תוכן מיני מזויף) ולא כחלק משגרה פדגוגית מתמשכת שמטרתה לא רק למנוע פגיעה, אלא לבנות חוסן נפשי, אמון עצמי וכישורי ניתוח של מציאות דיגיטלית רוויית מניפולציה.

כל אלה מצביעים על כך שהמערכת הנוכחית, על אף תנופת ההתפתחות המרשימה, עדיין אינה מספקת רשת מוגנות מסונכרנת, רב־רובדית ומותאמת גיל. בחלק הבא נציע כיווני פעולה שנבנים על התשתית הקיימת אך מבקשים לפרוץ מעבר לה.

חלק רביעי: המלצות התערבות למוגנות ילדים בעידן הבינה המלאכותית

על בסיס החלקים הקודמים, שבהם עסקנו במאפיינים הייחודיים של התנהלות ילדים בעידן הבינה המלאכותית, ולאחר מכן מיפינו את הפגיעויות המרכזיות המאיימות על ילדים ובני נוער דווקא בהקשר זה, חלק זה מתמקד בדרכי ההתערבות האפשריות ומציע מסגרת שיטתית להתמודדות עם הפגיעויות השונות המבוססת על קטגוריות הפעולה האלה: ידע ומחקר; רגולציה; עיצוב טכנולוגי אחראי; התאמת הדין הפלילי; התערבויות מערכתיות; אוריינות והעצמה; ופתרונות מבוססי AI. במרחב החדש שנוצר, האחריות להגנה אינה יכולה להיות בידי שחקן יחיד. סביב השולחן נדרשים לשבת יחד רשויות אכיפת החוק, רשויות הרווחה, משרד המשפטים, השירות הפסיכולוגי, מערכת החינוך, משרד הבריאות, הורים וילדים ובני נוער, וכן חברות הטכנולוגיה וארגוני חברה אזרחית.

385 Grace Liberatore, Alyssa Kim, Jack Brenner et al., *Artificial Intelligence Impacts in Education and Child Mental Health*, 37 Current Opin. Pediatrics 241 (2025)

א. עקרונות מנחים לעיצוב התערבות מותאמת ילדים

מנגנוני ההתערבות שנציע בהמשך פרק זה מבקשים להתאים את עצמם לא רק לסוג הטכנולוגיה או לסוג הפגיעה, אלא גם למהות הקהל שאליו הם מכוונים. בהקשר זה, ילדים ובני נוער אינם רק אוכלוסייה פגיעה מבחינה משפטית אלא גם קבוצת גיל בעלת מאפיינים דינמיים, משתנים ולעיתים סותרים. עיצוב מדיניות עבורם דורש רגישות כפולה: מחד גיסא, להגן עליהם מפני סיכונים שמערכות הבינה המלאכותית עלולות להציב; ומאידך גיסא, להימנע ממדיניות פטרנליסטית או חונקת שאינה מכירה בזכותם להתפתח, לבחור ולהתנסות.

1. **המתח בין הגנה לאוטונומיה:** ילדים, ובמיוחד מתבגרים, זקוקים למסגרת מגוננת, אך גם למרחב פעולה שמכבד את הסקרנות, את הדחף החברתי ואת כושר ההמצאה שמאפיינים את הגילים הללו. ההתערבות הרצויה צריכה לדעת לא רק להרחיק סכנה אלא גם ליצור סביבה המאפשרת פיתוח זהות בטוחה, בריאה ואותנטית, בעולם שבו סוכני AI, פילטרים ומנגנוני המלצה הם שותפים סמויים ופעילים בתהליך ההתבגרות.

2. **מהו "הטוב העליון של הילד"?** עיקרון זה, כפי שנוסח באמנת האו"ם לזכויות הילד, יכול לשמש עיקרון מארגן,[386] אלא שבעידן של טכנולוגיות מתקדמות, טוב זה אינו תמיד ברור או אחיד. עבור ילד אחד שימוש בצ'אטבוט רגשי עשוי להקל בדידות ולשפר את הדימוי העצמי; עבור אחר אותו שימוש עלול לעמעם גבולות, לעודד פגיעות או להפוך מתבגר לתלותי. העיקרון המוסרי מחייב אפוא לא "לקבוע טוב אחיד", אלא לעצב התערבויות שמכירות במורכבות ושואפות לאזן בין הגנה לאפשרות, בין שליטה לאמון, בין מניעה להעצמה.

3. **הבדלים ההתפתחותיים בין קבוצות גיל שונות:** ילדים בגילי 6-12 מתמודדים עם אתגרים קוגניטיביים ורגשיים שונים לחלוטין מבני נוער בגילי 13-17: יכולת הבחנה בין אמיתי למדומה, הבנת כוונה, ויסות רגשי ותחושת זמן, כל אלה נבנים בהדרגה ולעיתים מתעכבים או מתעצבים דרך האינטראקציה עם טכנולוגיה.[387] עיצוב נכון של התערבויות

386 Convention on the Rights of the Child art. 3(1), Nov. 20, 1989, 1577 U.N.T.S. 3

387 *Cognitive Development in Adolescence*, Stanford Children's Health

צריך להתייחס למאפייני ההתפתחות הללו לא רק כדי "להגן", אלא כדי להציע מסלולים שונים, מותאמים, שמכירים בפערים בין ילד למתבגר, ובתוך קבוצות הגיל עצמן.

4. **ריבוי השחקנים המעורבים במוגנות הדיגיטלית של ילדים:** הורים, צוותים חינוכיים, מערכות בריאות ורווחה, מתנדבים, מפתחים, מחוקקים, כל אלו משפיעים במישרין או בעקיפין על הסביבה שבה הילד פוגש מערכות בינה מלאכותית. הנטייה להטיל את כל כובד האחריות על ההורים בלבד, או לחלופין על המדינה, מפספסת את האפשרות למערכת משותפת של אחריות, מודל שבו כל שחקן ממלא תפקיד שונה אך מתואם, מתוך הבנה שמוגנות אינה רק שאלה של "הגנה טכנית" אלא של מרקם חיים שלם.

5. **עיצוב משתף:** מתוך הכרה בזכותם של ילדים להיות לא רק מוגנים אלא גם שותפים בעיצוב פתרונות, חשוב לכלול ילדים בתכנון פתרונות. ילדים ובני נוער יודעים איך מרגישה פגיעות ברשת, איך נראית מצוקה שנוצרת מצ׳אט עם סוכן ואיך נבנית תלות בכלי ש"נראה כמו חבר". חשוב לתת להם מקום להשמיע את קולם באמצעות מחקר, פיילוטים ועיצוב משותף.

6. **רגישות להקשרים תרבותיים, חברתיים וכלכליים:** ילד מהפריפריה הדיגיטלית לא פוגש את אותן טכנולוגיות כמו בן שכבתו בעיר הגדולה;[388] ילדה בחברה שמרנית לא תוכל לשוחח בפתיחות על חוויותיה משיחה עם צ׳אטבוט רגשי כמו נערה בסביבה פתוחה. לכן, עיצוב מדיניות חייב להכיר בשונות הרב־שכבתית של אוכלוסיית הילדים ולחתור למוגנות שהיא גם אוניברסלית וגם מותאמת הקשר.

7. **מאפיינים ייחודיים של ילדות וילדים בישראל:** ילדים בישראל נחשפים לעיתים קרובות לטכנולוגיות חדישות בגיל צעיר, הן דרך הורים עם זיקה טכנולוגית (וביטחונית) והן במערכת חינוך החותרת למודרניזציה אך מתמודדת עם פערים גדולים בתשתיות ובכוח אדם. חשיפה זו מתרחשת בתוך חברה המתאפיינת בפערים כלכליים, תרבותיים ואתניים חדים, המשפיעים על סגנון צריכת הטכנולוגיה, על יחסי סמכות הורית ועל הפנמת גבולות השימוש. מערכות הרווחה והחינוך פועלות בתנאים של עומס ותקינה חסרה, כך שהמלצות רגולטוריות מבוססות מחקר שנולדו במדינות רווחה מבוססות עשויות להיות בלתי ישימות בישראל מבלי שיבוצע בהן עיבוד מושגי ומעשי עמוק.

388 Eilat Chen Levy et al., *Bringing Science to the Periphery Through Distance Learning: Barriers and Opportunities*, 15(2) Education Sciences, 114 (2025)

ב. הכרה באתגרי יישום של תוכניות התערבות המיועדות לילדים

יישום המלצות מדיניות בתחום הגנת ילדים בעידן הבינה המלאכותית אינו מסתכם ברשימת משימות אלא כרוך בהתמודדות עם שורה של אתגרים מוסדיים, חברתיים ואתיים. חלקם נובעים ממגבלות מבניות, אחרים מעמימות מוסרית, אך כולם יחד מצביעים על הצורך בגישה דינמית, זהירה ומשולבת.

ראשית, גם כאשר קיימת הבנה גוברת של הצורך בהתערבות, המודעות הציבורית לנזקים הפוטנציאליים של בינה מלאכותית כלפי ילדים עדיין חלקית ומפוזרת. ברוב המדינות המפותחות עדיין לא נבנתה התשתית המוסדית שתאפשר ייזום שיטתי של תוכניות מוגנות, לא כל שכן יישומן בפועל. רבים מהצעדים המוצעים תלויים ברצון הטוב של אנשי מקצוע, בתקציבים תוספתיים או בקידום אישי מצד גורמים "מובילי שינוי" ולא במדיניות מסודרת. לכך נוספות גם מגבלות תקציב ומשאבים, במיוחד במערכות ציבוריות שמתמודדות עם ריבוי משימות ואיומים מיידיים.

פערי האוריינות בתחום זה בולטים במיוחד: הורים, אנשי חינוך, אנשי טיפול ומקבלי החלטות רבים אינם מכירים די הצורך את עולם הבינה המלאכותית, ואינם מצוידים במושגים, בכלים או בהכשרה שמאפשרים להם להעריך סיכונים, לזהות נזקים או לעצב תגובה. גם כאשר המודעות קיימת, הכלים אינם תמיד זמינים, והשיח הציבורי נוטה לנוע בין טשטוש לבין פאניקה. קיים קושי עמוק לפעול בתנאים של "ערפל טכנולוגי", כלומר במצבים שבהם קצב ההתפתחות של הטכנולוגיה מקדים את יכולתנו להבין אותה, להעריך את השפעותיה או לחזות את עתידה. בסביבה כזו, גם כוונות טובות עלולות להתגלות כבלתי מותאמות או אפילו מזיקות.

כך למשל, כלים שמטרתם להגן על ילדים, כגון בוטים תומכי רווחה, מערכות ניטור או פלטפורמות ללמידה מותאמת, עלולים להפוך לכלים הפוגעים באוטונומיה, בפרטיות או יוצרים תלות. הרעיון של "טכנולוגיה מגוננת" חייב תמיד להיבחן גם מנקודת מבט של זכויות, האם אנו יוצרים מעגל סגור שבו הילד נתון לפיקוח מתמיד? האם המערכת פועלת לטובתו או משעתקת עמדת שליטה? ההתערבויות צריכות אפוא לכלול מנגנוני בקרה פנימיים, היזון חוזר ועדכון מתמיד.

לבסוף, יש לזכור כי השדה הזה הוא גם שדה של התנגשות בין אינטרסים. טובת הילד, המוגדרת כעיקרון־על, איננה עולה תמיד בקנה אחד עם אינטרסים מסחריים של חברות

טכנולוגיה, ולעיתים אף לא עם מטרות של מדינות המבקשות לעצב אזרחות נוחה או נאמנה. במצבים כאלה, הדילמות הופכות אתיות: מי מגדיר מהו חינוך ראוי? עד כמה מותר לעצב מידע לצרכים חינוכיים או פוליטיים? מהו הקו שבין תמיכה לבין הנדסה חברתית? הכרה באתגרים הללו אינה באה לשתק אלא להפך: היא מאפשרת בניית אסטרטגיה צנועה, גמישה ומסתגלת, שאינה שואפת לפתרון קשיח אלא למערכת לומדת.

ג. הצורך בשינוי פרדיגמה בחשיבה על מוגנות ילדים בעידן הבינה המלאכותית

מסמך ההנחיות לגבי שימוש בבינה מלאכותית שפרסם משרד החינוך בינואר 2025 מציע להתאים את השימוש לגילאים שונים ולהגדיר גבולות מבחינת זמן פיקוח של מורה לצד דרישת אישור הורים.[389] ברוח מסמכי מדיניות שנסקרו בחלק הקודם הוא מציע לשים לב לסיכונים שונים ולהתייחס אליהם מפרספקטיבה של ערכים כגון פרטיות, שוויון ואבטחת מידע. אבל נקודת המוצא של המסמך היא של שימוש ממוקד בכלי בינה מלאכותית, ככל הנראה גנרטיביים – למשל, בוט למידה שתכליתו להיות הדור הבא של ויקיפדיה, או מערכות המאפשרות יצירת טקסטים ומצגות. לכן המסמך מניח קיומם של כיתה ושל מורה ולא מעבר עמוק ומתמשך ללמידה מותאמת אישית.

אנו סבורים שיש לשנות את פרדיגמת ההתבוננות על מוגנות ילדים בבתי הספר בשלושה הקשרים חשובים:

ראשית, תשתיות המוגנות שנבנו בעשור האחרון ביחס לעולם הדיגיטלי נבנו במקור עבור מציאות אנושית־דיגיטלית שבה מקור הסכנה הוא לרוב אדם אחר: טורף מיני, בריון מקוון, מפיץ שנאה. השיח אינו עוסק בעיקרו בהבנה הוליסטית של המערך הקוגניטיבי־רגשי־חברתי בעולם רווי אנשים לצד מכונות. בעידן שבו מערכות בינה מלאכותית עצמן הופכות לגורם אינטראקטיבי פעיל, נדרש עדכון של מערך המוגנות הקיים.

שנית, הנחת העבודה של מאמצי ההתערבות שסקרנו עד כה היא שמערכת החינוך בעידן הבינה המלאכותית תיראה זהה לזו הקיימת היום בשלושה מרחבים חשובים – תפקיד המורה, תהליך הלמידה והמרחב הפיזי של הלמידה. אנו סבורים כי יש מקום לתכנון פרואקטיבי מחודש של שלושת המרחבים האלה, משום שאחרת ההתבוננות באתגרי המוגנות תיעשה מפרספקטיבה צרה מדי, שיש בה נקודות עיוורון משמעותיות.

389 **הנחיות שימוש בכלי בינה מלאכותית יוצרת**, לעיל ה"ש 383.

שלישית, יש צורך לעסוק במקביל בשלושה אתגרי מוגנות. האחד הוא חשיפה לנזקים זדוניים – מקשר עם טורפים ועד ניסיונות לגנוב מידע על ילדים. השני הוא חשיפה לנזקים מוטמעי טכנולוגיה כמו קשרים בין ילדים למכונות ואתגרים קוגניטיביים. השלישי הוא אתגרי מערכות ומשילות של מערכות, כאשר אלה מוטמעות כעניין של מדיניות (או כהיגררות אחר המציאות) במערכת החינוך.

חלק חמישי: המלצות ממוקדות

1. ידע, מחקר ותחזיות עתיד

בעוד החברה החרדית או החברה הערבית בישראל דורשות מחקר ייחודי בשל מאפיינים מקומיים ייחודיים, בתחום מוגנות ילדים בעידן הבינה המלאכותית קיים בעולם ידע הולך ומתרחב. מחקרים רבים, מדדים רגולטוריים, תצפיות אתנוגרפיות וניסויים קליניים נערכים בשנים האחרונות במדינות רבות, בעיקר באירופה, ארצות הברית ודרום קוריאה, העוסקים בהשפעות של צ׳אטבוטים רגשיים, מערכות למידה אדפטיביות, דמויות וירטואליות, משקפיים חכמים וטכנולוגיות אימרסיביות על ילדים ובני נוער.

תחומי מחקר נוספים שאנו סבורים שצריך להרחיב בהם ברמה הגלובלית עוסקים למשל בבפגיעות הפיזית של ילדים בעקבות השהייה במרחבים וירטואליים, השפעות נירולוגיות של צפייה במשקפיים חכמים ופיתוח דפוסי תנועה אוטומטיים שאינם מתואמים למציאות הפיזית; עוצמת ההתקשרות של ילדים עם צ׳אטבוטים רגשיים, ישויות דיגיטליות וממשקים תומכי שיח, ובכלל זה אשליית חיבור, פיתוח תלות, חוויית נטישה בממשקים אינטראקטיביים ואובדן אמון בקשרים אנושיים. בתקופה הקרובה יעלה הצורך בפיתוח מדדים לזיהוי פגיעות רגשית לרבות ביחס לתדירות שימוש ולעוצמת קשר ותגובות למניפולציות רגשיות. נוסף על כך נכון לחקור השפעה של שימוש במודלים של שפה ותוכן יציר מכונה על כישורי שיפוט, על הבחנה בין עובדה לדעה, ועל יכולת עיבוד מידע עצמאי, ולעמול על פיתוח מדדים לתלות קוגניטיבית, לזיהוי הסתמכות על מערכות חכמות בקבלת החלטות ובתהליכי למידה. כפי שכבר התקיים בתקופה הדיגיטלית, נדרש להמשיך ולמפות את האופנים שבהם ילדים הופכים ליעד להצעות שיווקיות מותאמות אישית, דינמיות ואינטראקטיביות או תמריצים כלכליים המובילים לצריכה לא מבוקרת. אבל, בתקופה הנוכחית יש מקום לעסוק במחקר של פגיעות חברתית ותרבותית, בדגש על קבוצות מיעוט הסובלות מייצוג דל,

סטריאוטיפי או מוטה, וילדים המודרים משירותים חדשים בשל חסמים לשוניים, תרבותיים או כלכליים. מחקרי עומק על פערי נגישות לשירותים מבוססי AI בקרב ילדים בקהילות מוחלשות, לרבות סוגיות של שפה, תרגום, זיהוי קולי וזיהוי תרבותי, יהפכו לנדרשים. יחד עם אלה יש צורך בפיתוח תרחישי עתיד ועיצוב ספקולטיבי עתידי לצורכי מדיניות, כגון: אינטראקציה של ילדים עם רובוטים תומכי־רגש במקום דמות הורה, ניהול תנועה עצמית של ילדים במרחבים עם שכבות מידע במשקפיים חכמים, היווצרות ״חברים דיגיטליים״ קבועים וחיים בעולם שבו לכל ילד יש תאום דיגיטלי.

עם זאת, תרגום הידע הקיים בעולם למדיניות אפקטיבית בישראל אינו יכול להיעשות כיבוא טכני של מסקנות מחקריות, אלא מחייב התאמה תרבותית, מוסדית וקהילתית, מתוך הכרה במאפיינים הייחודיים של החברה הישראלית. **אנו מציעים לפיכך להקים מרכז ידע לאומי למוגנות ילדים בעידן הבינה, שיהיה גוף בינתחומי, בתמיכת משרדי הממשלה הרלוונטיים (חינוך, רווחה, משפטים), המוסד לביטוח לאומי, גופי מחקר ציבוריים, שותפים בינלאומיים ונציגות של הילדים ובני הנוער עצמם. גוף כזה יאפשר פיתוח אקוסיסטם אינטגרטיבי, עתיר ידע ובעל יכולת פעולה מעשית של שילוב בין עולמות מחקר, מדיניות, רגולציה ועיצוב, מתוך רגישות לגיל ולמרקם התרבותי המקומי.** המרכז יפעל בשלוש זרועות משלימות:

- תרגום, עיבוד וניתוח ביקורתי של ידע מחקרי קיים מהעולם, מתוך בחינה של פערים רלוונטיים לישראל והצעות לתרגום רגולציה או שיטות התערבות.

- פיתוח מחקר יישומי רגיש הקשר, שיבחן תופעות ייחודיות כמו השפעה של תכנים גנרטיביים על ילדים במערכת החינוך הערבית או החרדית, תהליכים של היווצרות תלות רגשית בדמויות בינה בקרב ילדים חסרי עורף משפחתי, השפעת השימוש בצ׳אטבוטים תומכי־שיח על בני נוער הסובלים מחרדות עקב המצב הביטחוני, או שימוש בטכנולוגיות אימרסיביות בסביבה ציבורית רוויית סיכונים.

- כלי הערכה מוסדיים שיאפשרו לא רק מדידה של רמות פגיעות, אלא גם תכנון ארוך טווח של התערבויות צמודות תכנון מערכתי של מערכת החינוך ברמת הכיתה, המורה והלמידה.

2. הוספת קומה על מערכי המוגנות הקיימים

במדינת ישראל קיימת תשתית מוסדית נרחבת יחסית בנושא מוגנות ילדים: במשרד החינוך פועלות יחידות ייעודיות, קיימים נהלים מקצועיים להתמודדות עם פגיעה רגשית ופיזית, קיימים תפקידים מוסדיים כמו "רכז מוגנות" או "יועץ חינוכי" ופותחו מערכי הדרכה לצוותים חינוכיים ולתלמידים במגוון גילים. בנוסף, פועלים גופים שיטוריים כמו מוקד 105 של המשטרה המספקים מעני חירום ואכיפת חוק. בשנתיים האחרונות ניכרת גם התעוררות בתחום של סיכוני בינה מלאכותית בשדה החינוך.[390]

ההמלצות הבאות מציעות כיווני פעולה שמטרתם לחדש ולהרחיב את מערך המוגנות לטווח הקצר כך שישקף את האתגרים שמציבה הבינה המלאכותית:

א. פיתוח מדדים לתוכניות התערבות

- **זיהוי מדדי פגיעות רגשית וקוגניטיבית חדשה:** סימנים לקשר רגשי אינטנסיבי עם ישות דיגיטלית (למשל, הילד מתאר את הבוט כ"חבר אמיתי"), נסיגה ממקורות סמכות אנושיים, העדפת המלצות אלגוריתם על פני אלו של מבוגרים משמעותיים, ירידה ביוזמה לחיפוש עצמאי של מידע.

- **פרופיל סיכון מותאם AI:** פיתוח כלי עבודה יישומי לאנשי מקצוע, המאפשר לאבחן פגיעות על בסיס דפוסי השימוש של הילד במערכות בינה מלאכותית. הפרופיל יכלול מדרג של פגיעות רגשית, קוגניטיבית, חברתית ופיננסית, לצד המלצות מותאמות להתערבות בשדה: שיחה, הפניה, שינוי סביבה טכנולוגית ועוד.

- **פרוטוקולים מוסדיים לזיהוי מוקדם ותגובה:** הקמת מסגרות תגובה חדשות במוסדות חינוך, עם נוהל קונקרטי למקרים של תלות טכנולוגית, השפעות רגשיות חריגות או פגיעה קוגניטיבית עקיפה. לדוגמה, דיווח על שימוש במשקפיים חכמים שמעוררים חרדה או התכנסות. הכשרה לשימוש בסימני שאלה כמודל זיהוי רך: במקום להסתמך רק על "סימני אזהרה", יש לפתח מודלים המבוססים על שאלות פתוחות שמזמינות שיחה כדי לחזק אמון ולאפשר זיהוי עדין של פגיעות שאינה ניתנת למדידה מספרית. לדוגמה: "מתי בפעם

390 אייל עזרא "בינה מלאכותית – סיכונים וסכנות" Let's AI (8.3.2024).

האחרונה הרגשת שהבוט מבין אותך יותר ממישהו מהכיתה?"; "איזה מידע אתה לא משתף עם חברים אבל כן עם המערכת? "; "מה הרגשת כשסגרו לך את האפליקציה?".

- **פרוטוקולים לסינון תוכן ובקרת איכות בכלים הדיגיטליים:** בפן הטכנולוגי, חובה לוודא שהכלים הנלווים עומדים בסטנדרטים מחמירים של סינון תכנים בלתי הולמים וכוללים בקרה הורית ומוסדית. ניתן לשלב זאת בתוכניות אוריינות לתלמידים, למשל כדי שידעו שהמערכת מנטרת ומתעדת אינטראקציות (בצורה שמכבדת פרטיות) למטרות בטיחות, ידע זה כשלעצמו ירתיע ממשחקים לא ראויים עם הכלי.

ב. פיתוח הכשרות

מערכות החינוך, הרווחה והבריאות צריכות לעבור עדכון תוכני ופרדיגמטי במערכי ההכשרה שלהן. לא עוד רק זיהוי פגיעה מצד אדם, אלא הכרה באינטראקציה עם מכונה כצומת פוטנציאלי לפגיעות. לפיכך נדרשות הכשרות מקצועיות ייעודיות למורים, יועצים, עובדים סוציאליים, רופאי ילדים, מדריכים, אנשי בריאות נפש וגורמי אכיפה. הכשרות אלו יכללו שילוב של התייחסות לאינטראקציה עם ישויות דיגיטליות בתוכניות רווחה ובריאות נפש וישקפו מעבר מתפיסה של טכנולוגיה כטריגר למצוקה לכיוון תפיסה הרואה בה אובייקט רגשי או בינאישי של ממש. ההכשרות יכללו ניתוח מקרים של תלות רגשית בבוטים תומכי נפש, זיהוי תסמינים דמויי אובדן, וכן שילוב של שיח רגשי חדש סביב חוויות דיגיטליות (למשל, חרטה על שיתוף מידע עם מערכת).

תוכנית ההכשרות אינה צריכה להתמקד רק בהבנת יחסי אדם־מכונה של ילדים עם מכונות, אלא דווקא בהבנת יחסי אדם־מכונה של הצוות המטפל. לדוגמה, כאשר הורים, מטפלים או אנשי מקצוע מסתמכים על מערכת אלגוריתמית במקום על שיקול דעתם, למשל מקבלים הדרכה מתמשכת מצ׳אט ביחס לשאלה כיצד להתמודד עם ילדים, מסתמכים על הכרעה של מכונה ביחס לתיעדוף תיקי רווחה, או מקבלים כפשוטן המלצות של מערכות למידה אישית לגבי תלמידים, נוצרת סכנה של "ערעור סוכנות מטפלים", קרי ויתור על האינטואיציה והניסיון האנושי לטובת מנגנון אלגוריתמי שלא תמיד מבין את ההקשר המלא.[391] הבנת

391 Devansh Saxena & Shion Guha, *Algorithmic Harms in Child Welfare: Uncertainties in Practice, Organization, and Street-level Decision-Making*, arXiv (Aug. 9, 2023)

תופעות אלה תהיה משום תנאי מפתח ליכולתם של מטפלים מסוגים שונים לתת מענה לתפקידיהם וגם לאתגרים שיצטרכו להתמודד עימם.

ג. פיתוח תוכניות אוריינות

כפי שכבר נכתב לעיל, תוכניות אוריינות יידרשו כבסיס למוגנות רגשית. לכן, נדרש פיתוח תוכניות למידה בין־דורית ודיאלוג משפחתי – מודלים של "חינוך הפוך" שבהם ילדים מלמדים מבוגרים על הסכנות שהם מזהים בבוטים, פרסומות או אלגוריתמים, ובכך מחזקים את תחושת המסוגלות שלהם וגם יוצרים גשר הבנתי במשפחה. פיתוח "משחקי שיחה" בין הורים לילדים על סיטואציות רגשיות־דיגיטליות וכלים להורים לזהות סימני מצוקה הקשורים לחשיפה טכנולוגית כמו עומס חושי, מתח גופני ונפשי.

פיתוח תוכניות לילדים ובני נוער המלמדות להבחין בין קשר רגשי אמיתי לקשר עם מכונות, מתוך חיזוק זהות עצמית ותחושת מסוגלות; פעילויות חינוכיות המדמות אינטראקציות רגשיות עם בוטים, ומה ניתן ללמוד מהן על העצמי והאחר; הפקת חומרי הדרכה, סיפורים אינטראקטיביים, סימולציות וסרטונים המדמים מצבים רגשיים בעייתיים מול בוטים; יומני "אמפתיה דיגיטלית" שבהם ילדים כותבים או מציירים מה הם חושבים שהבוט הרגיש, או מה הם הרגישו כלפיו, ככלי לפיתוח שיפוט רגשי מוסרי.

פיתוח תוכניות על סיכונים פיזיים בסביבות דיגיטליות, למשל אחריות גופנית בעת אינטראקציה עם מכשירים, משחקים וסימולציות חינוכיות המלמדות על גבולות מגע, פרטיות פיזית, והבדלים בין מציאות לדימוי.

תוכניות אוריינות נדרשות עדיין גם כבסיס למודעות קוגניטיבית. בשנים האחרונות נעשה מאמץ גדול לשלב יחידות על חשיבה ביקורתית בתוכניות לימודים מגיל צעיר, ליצור תוכניות, כלים ותרגילים המפתחים את כושר השיפוט, הספק, ההשוואה והנימוק העצמאי, וכן משחקים חינוכיים המדמים תרחישים של מידע שגוי או מוטה ומזמינים בחירה; וגם תוכניות לבני נוער להתמודדות עם עודף מידע ומניפולציות קוגניטיביות. עם זאת, **ראוי לשקול חשיבה מחדש על תוכניות אלה בעידן שבו ההבחנה בין תוכן יציר מכונה ותוכן אותנטי לא תהיה אפשרית**, הואיל ותוכניות כאלה יכולות לייצר אשליית מסוגלות ושליטה שנזקה גדול מתועלתה. בהקשר זה נפנה לחלק הכללי שבו טענו שהפתרונות לתופעת ערעור בירור המציאות אינם בהכרח בשדה האוריינות.

ד. היערכות לאתגרי מוגנות: מפתחות שליטה ומשילות בלמידה מותאמת אישית

בשנים הקרובות מערכות החינוך צפויות לעבור שינוי עומק סביב חזון הלמידה המותאמת אישית, שבו לכל תלמיד ותלמידה מוצמד מלווה דיגיטלי אישי: בוט, משקפיים חכמים, טאבלט ייעודי או צעצוע הומנואידי, המשלב סיוע קוגניטיבי, רגשי והתנהגותי. מערכת זו תעצב את קצב הלמידה, תבחר את התכנים, תספק משוב בזמן אמת ותציע תמיכה וליווי במצבי קושי או הצפה. הפוטנציאל הפדגוגי אדיר: התאמת התכנים וקצב הלמידה לצרכיו של כל תלמיד, ליווי רגשי ותמיכה אישית לכל ילד, וכן הערכה חלופית ודיסקרטית במקום מבחנים פומביים מול בני הכיתה. בצורה כזאת תלמידים יכולים ללמוד בהדרגה בקצב המתאים להם ובנושאים המעניינים אותם, דבר שיחליף את מערכת החינוך של חברת ההמונים במערכת עצומה של "מורים פרטיים" שהם גם מלווים רגשיים. במקביל, הדבר יפתור את בעיות החוסר במורים, העומסים והיעדר תשומת הלב האישית בשל גודל הכיתות, ואת המשבר הדיסציפלינרי הקיים במערכת החינוך ביחס למקצועות מסוימים כגון מדעים, מתמטיקה ואנגלית. זוהי הזדמנות יוצאת דופן לצמצום פערים ולהנגשת חינוך איכותי לכל ילד.

המלווה הדיגיטלי לא יהיה כלי עזר פסיבי, אלא מערכת פעילה ודינמית שתלווה את הילד בכל רגע של יום הלימודים ומחוצה לו. הוא יאסוף באופן רציף נתונים על קצב הקריאה, שפת הגוף, תגובות רגשיות, דפוסי הצלחה ושגיאה, תחומי עניין משתנים, רמת מוטיבציה ועומס קוגניטיבי. על בסיס ניתוחים אלה, המלווה יקבל החלטות עצמאיות: מתי להעמיק בנושא מסוים ומתי לדלג עליו, מתי להציע אתגר נוסף ומתי להקל, אילו סוגי מדיה להפעיל, ואילו שאלות לשאול כדי לעורר עניין או לעודד חקירה. ייתכן שהתלמיד לא יהיה מודע כלל לכך שהמלווה שינה את סדר היום הלימודי שלו, שכן הכול יתבצע באופן מותאם, זורם ואינטואיטיבי. מערכות אלה עשויות גם לתקשר עם הורים ומורים ולספק תובנות על מצבו של הילד: מצבי מצוקה רגשית, שעמום מצטבר או קפיצות התפתחותיות. הלמידה עצמה תתרחש במגוון מצבים: בכיתה הפיזית, בחדר השקט, בזמן משחק, בתנועה, ואפילו בהקשר חברתי, כאשר המערכת מנתחת אינטראקציות עם ילדים אחרים ומספקת משוב על כישורים חברתיים או דפוסי תקשורת. התוצאה היא סביבה חינוכית חיה, רב־שכבתית, שמבוססת על דיאלוג מתמשך בין בינה מלאכותית, הילד והמערכת הסובבת אותו.

הנחת העבודה שלנו, שגם היא ניתנת לסתירה אם כי בטווח זמן ארוך יותר, היא שעדיין יהיה צורך ללכת לבית הספר, הן בשל היותו מקום שמירה על ילדים שעה שהוריהם הולכים

לעבודה, הן בשל הצורך באינטראקציה חברתית כחלק מתהליך הצמיחה וההתבגרות. תפקיד המורים צריך להתברר מחדש בעידן כזה, ויכול לנוע בין שמירה על הסדר והשקט לבין טיפוח אינטראקציות בינאישיות וקידום חקר וחוויית למידה קבוצתית. מדובר במהפכה של ממש בתפקיד המורה, במבנה הכיתה וביחסי הכוח שבין תלמידים, מערכת החינוך והטכנולוגיה.

אנו מציעים **לעסוק באופן אינטנסיבי כבר בתקופה הקרובה בסיכונים ובפגיעויות הפוטנציאליים הטמונים בשימוש במלווה דיגיטלי צמוד מתוך מטרה לבחון מהי המעטפת המוסרית, המוסדית והחינוכית הנדרשת למעבר הלימינלי מכיתה פיזית עם מורה פיזי ללמידה מותאמת אישית, ולהיערך לספק אותה באופן פרואקטיבי כבר בשלב זה.** לא מדובר רק בשאלות של הטמעת מוצרי בינה מלאכותית בכיתה אלא כיצד להבטיח שהעברת החלטות לימודיות, רגשיות והתפתחותיות של ילדים למערכות מבוססות בינה מלאכותית מתבצעת מתוך אחריות, שקיפות ושמירה על זכויותיהם, חירותם ויכולתם להשתנות ולחרוג מן הצפוי.

חלק מן הסיכונים הם מוכרים אבל צריך לתת עליהם את הדעת, כגון אבטחת נתונים. ככל שהנתונים עמוקים וביומטריים יותר, כך החשש מדליפתם והעברתם לידיים לא רצויות גדל. סיכון נוסף הוא הצורך בקביעת גדרי פרטיות, בייחוד בהקשר של הגבלות על היסקים שמכונות מסוגלות לבצע בהסתמך על מידע פרטי ביומטרי. סיכון אחר ומדובר הוא החשש מפני הטיות אלגוריתמיות ו"דעות קדומות" שמכונה יכולה לאחזר ולהעצים כלפי תלמידים בשל מאפיינים שונים. סיכונים אלה מופיעים בספרות ובמסמכי הליווי האתיים כבר כיום. כוונתנו בסיכונים החדשים היא התייחסות מעמיקה לתופעות של הסתמכות יתר על ידע יציר מכונה, אובדן מיומנויות קוגניטיביות ללא פיתוח מיומנויות חלופיות, תלות רגשית ובידוד חברתי, חשש מפני איבוד הלימה לערכים אנושיים של מכונה ומתן עצות מסוכנות ולא מדויקות בנושאים רגישים; וכן בשאלת המשילות על תכנים בסביבות דיגיטליות, שעלולה ללא פיקוח לערער את בירור המציאות אצל המשתמשים או לחשוף אותם לתכנים לא הולמים ומזיקים, ולחלופין באמצעות פיקוח יתר להפוך למנגנון שליטה של שלטון מרכזי מטעמים פוליטיים.

הטמעת למידה מותאמת אישית מחייבת לא רק חדשנות פדגוגית וטכנולוגית אלא תכנון מערכי משילות. במציאות כזו בית הספר חדל להיות המתווך הבלעדי של הידע, ולעיתים אף אינו הגורם הקובע את האופן שבו ילד לומד בפועל. מדובר בשינוי עמוק במערך

הכוחות שמחייב יצירת מנגנוני משילות חדשים ברמת המוסד, הפלטפורמה, המדינה והמשפחה.

הואיל ולא בטוח שחלוקת המערכות האלה תהיה רק על ידי בית הספר או המדינה, אלא בתופעה שתלך ותהיה מונעת שוק, תחרות ומסחר, בדומה למה שקרה עם כניסת הסמארטפונים, נדרש בירור יסודי של שאלת הבעלות: מי מחליט על המלווה הדיגיטלי של הילד? האם בית הספר יספק אותו כחלק מהמערך הפדגוגי, או שמא כל ילד יגיע עם "היועץ החכם" מהבית, מערכת שפותחה על ידי גורם מסחרי, על בסיס העדפות ההורים או תקציב המשפחה? המצב השני עלול ליצור אי־שוויון, גם בהזדמנויות הלמידה וגם ברמת הפיקוח והבקרה. אין ספק גם שתידרש אוריינות **ברמת הפרט והמשפחה** לא רק כדי להבין איך להשתמש במלווה הדיגיטלי, אלא גם מה תפקידו, מהם גבולותיו, ומה מקומו ביחס למורים, להורים ולתלמידים אחרים. אלא שבחלק זה אנו עוסקים במנגנוני משילות מוסדיים.

ה. מדיניות חינוכית מרכזית אחידה, מבוססת תקינה

יש לגבש מדיניות אחידה וכוללת בנוגע לשימוש במלווים דיגיטליים חינוכיים. מדיניות זו צריכה להגדיר סטנדרטים אתיים, בטיחותיים ותוכניים לפיתוח מערכות כאלה ולקבוע תנאים לרישוי, בקרה והסדרה של הפלטפורמות השונות. מומלץ לבחון הקמה של **רגולטור ייעודי או יחידת פיקוח חינוכית־טכנולוגית** שתוכל לאשר מערכות מותרות לשימוש בבתי ספר ותנהל מרשם אלגוריתמי של מערכות כאלה, וכן תדרוש תקינה (ברמת טיב האלגוריתמים, פרטיות, אבטחה וכד) ותאכוף כללי אחריות יצרן.

תקינה כזאת של מוצרים המיועדים לשוק הלמידה המותאמת אישית יכולה לכלול עקרונות של "עיצוב למוגנות" (protectability by design) שלפיהם, למשל מערכת למידה מותאמת אישית תהיה חייבת לכלול מנגנוני איזון פנימיים שימנעו עומס רגשי, ובמקביל מנגנוני בקרה לפיקוח על האוטונומיה הקוגניטיבית – הטיות למידה, הצפה של מידע ועיצוב תודעה. כך גם תוכל התפתח תקינה לגבי מנגנוני שקיפות והסברתיות תפקודיות, כלומר היכולת של מורה, הורה או ילד להבין כיצד הומלץ לו התוכן, מה מקורותיו, ומה הקשר בין הביצועים הקודמים לבין המסלול הנוכחי או מנגנוני "יציאה חכמה", המלצות להפסקה, שינוי מסלול או פנייה לאינטראקציה אנושית, מנגנוני תיאום בין דמויות סמכות אנושיות ודיגיטליות ומערכות דיווח על קשיים או פריצות דרך שתשתף הורים, וממשק תרגום בין שפת המערכת לשפת המורה.

ו. קידום פיתוח מערכת למידה אישית ציבורית

המעבר ללמידה מותאמת אישית באמצעות מלווים דיגיטליים צפוי להחריף את פערי ההזדמנות בין תלמידים, בשל הבדלים בגישה לטכנולוגיה, למשאבים ולתמיכה הורית. כדי למנוע היווצרות של "שוק פרטי" ללמידה איכותית מול מערכת ציבורית מפגרת, וכדי להבטיח שכל הילדים, ללא קשר למעמדם החברתי־כלכלי, ייהנו מלמידה מותאמת אישית המבוססת על עקרונות של מוגנות, שקיפות וערכים חינוכיים, יש להשקיע משאבים בפיתוח מערכות ציבוריות במימון ציבורי, בקוד פתוח, בפיקוח ציבורי, בטוחות ונגישות, ולא להותיר את השדה לשוק הפרטי בלבד. ולא, ייווצרו שכבות של ילדים עם בוטים מתקדמים, בעלי יתרון תחרותי עצום, מול ילדים נטולי תיווך איכותי, החשופים לאלגוריתמים בלתי מפוקחים או למלווים דיגיטליים המשרתים אינטרסים שיווקיים ולא חינוכיים.

יש להבטיח שחבילות הלמידה הציבוריות יהיו בעלות מאפיינים המאפשרים התאמות מגדריות, שפתיות ותרבויות, ושיש בהם גם תוכני ליבה וגם ממשק חברתי־רגשי.

יש לחשוב על מודל תקצוב מבוסס "צדק טכנולוגי" שיאפשר להקצות תקציב דיפרנציאלי בהתאם לרמת פגיעות בינה מלאכותית, כלומר שתלמידים מרקע מוחלש יקבלו לא רק גישה לכלים אלא ליווי רב־מערכתי. במקביל יש לבחון הקמה של מערכת ניטור ומדדים שיבחנו באופן שוטף יעילות של מערכות ציבוריות, פערים בהישגים ובתחושת השייכות, בחוויה הרגשית של תלמידים ועוד וישמשו לטיוב המערכות.

ז. מדיניות מוסדית בית ספרית

בתי ספר יידרשו לגבש **מדיניות משילות פדגוגית־טכנולוגית** שתבחן את האיזון בין סמכות מוסדית לחירות טכנולוגית פרטית.

בין השאלות שיש להסדיר: האם ניתן לאסור על הכנסת מכשירים פרטיים לכיתה? האם יש לבית הספר שליטה על המידע שהתלמיד צורך, או רק על המידע שהוא מפיק? כיצד מתבצע תיאום בין תוכניות הלימודים לבין המסלולים האישיים שמציעה המערכת? האם המלווה הדיגיטלי משדר מידע למורה או מתנהל במקביל ללא קשר? האם לבית הספר יש סמכות להגביל, לפקח או אף לאסור על הכנסת בוטים מסוימים או אפליקציות מסוג מסוים? ומה קורה כאשר המלווה הפרטי מנחה את הילד בניגוד לתוכנית הלימודים הבית ספרית?

מוצע להקים ועדות אתיקה מוסדיות או מבוססות רשות מקומית הכוללות מורים, פסיכולוגים, הורים, תלמידים ומומחי טכנולוגיה, שתפקידן יהיה ללוות תהליכי פיתוח, לנסח גבולות ולהתריע מפני תופעות מזיקות מראש.

נוסף על כך, יש לשקול ליצור שגרות חינוכיות מאזנות, למשל באמצעות יומני למידה רגשיים שיאפשרו, לצד מעקב התקדמות לימודית, מרחב שבו תלמיד יתעד רגשות, קשיים, תחושות הצלחה או תסכול. כלי כזה ישמש מחנכים, הורים וגם את המלווים הדיגיטליים כדי לדייק את ההמלצות; וכן פעילות אנושית־קבוצתית מחייבת, באמצעות תרגיל קבוצתי, דיון מוסרי, משחק או משימת שטח, כאמצעים ליצירת מוגנות מול התמסרות למכונה.

ח. תכנון פיזי ואדריכלי של מוסדות חינוך באופן מותאם לעידן פיג׳יטלי

משקפיים חכמים, כפי שכתבנו בחלק הראשון של המחקר, יאפשרו לילדים להוביל על פניהם ממשק חכם שיסנן, יפרש ויתווך את הסביבה הבית ספרית ואת סביבת הלמידה. המפגש בין בינה מלאכותית מרחבית לטכנולוגיות אימרסיביות משנה מן היסוד את תפקידו של בית הספר, משום שהוא מגדיר מחדש את יחסי אדם־ידע, אדם־מציאות ואדם־חברה. בעידן זה יהיה צורך לתכנן מחדש את המרחב הפיזי כך שיתאים למציאות הפיג׳יטלית, שבה שלושת הממדים, הפיזי, הדיגיטלי והחברתי, משולבים זה בזה. תכנון אדריכלי נתפס עניין ארוך טווח אבל הוא יידרש מהר מן הצפוי להערכתנו. אנו מציעים להקים צוות חשיבה ותכנון שיורכב מאנשי חינוך, טכנולוגיה, מוגנות ומאדריכלים כדי להציע הצעות קונקרטיות ביחס לנושאים כגון תכנון אזורי לשימוש פיג׳יטלי בטוח בתוך בתי ספר, ובכלל זה אזורי ״מציאות מוגברת״ המיועדים לפעילות מרובדת, שיש לגביהם סימון ברור, הנחיות שימוש, ויסות חושי והגנה על פרטיות והוגנות. חשיבה מודעת למהלך השימושים הפיג׳יטלי נדרשת גם בכתיבת הנחיות לתכנון בתי ספר בישראל. ההנחיות שפורסמו לאחרונה מציעות הצבת כיתות לימוד בקומות מסחריות צפופות וללא חצרות, והן עלולות להעצים את הפיצול בין הממד הדיגיטלי הרווי לבין תשתית פיזית דחוסה וקרה.[392] לכן, לצד כל חדשנות טכנולוגית, יש להגן על עקרונות בסיסיים של למידה: אור טבעי, תנועה חופשית, מפגש בלתי אמצעי ומקום לילדים להיות לא מנוטרים, מנותבים או מותאמים.

392 תהילה שוורץ־אלטשולר "חכמים אבל בלי נשמה: האם כך ייראו בתי הספר של העתיד?" **TheMarker** (17.7.2025).

צוותי חשיבה יעסקו גם בתכנון פדגוגי של מציאות שבה לתלמידים יש "בועה תפיסתית אישית", וניסיון להתמודד איתה באמצעות סביבות פיזיות מקדמות דיאלוג ולא רק צריכת מידע מותאם, למשל, בשילוב של מסכים משותפים, סביבות גמישות שמאפשרות גם פעילות קבוצתית ועזרים חזותיים כבסיס לדיון משותף. כן אפשר לעסוק בתכנון מרחבי סימולציה רגשיים מודרכים שבהם תתקיים אינטראקציה עם בוטים רגשיים בליווי מקצועי של יועצים ומורים כדי לקדם תהליך התממשקות בריא בין רגש למכונה.

ט. עיצוב מחדש של הגנות משפטיות על ילדים בעידן הבינה המלאכותית

ילדים נתפסים באופן מסורתי כאוכלוסייה טעונת הגנה מיוחדת גם בעיני המשפט והרגולציה. כפי שהראינו בסעיפים הראשונים, כבר קיימים דברי חקיקה העוסקים בתפר שבין בינה מלאכותית וילדים, נוסף על חקיקה והצעות חקיקה שהועלו לאורך השנים במטרה להגן על ילדים בעידן הדיגיטלי בכלל, למשל בנוגע לחשיפה לתכנים מזיקים, עיצוב ממכר של טכנולוגיה ועוד. בהקשר זה אנחנו מציעים באופן כללי לאמץ בישראל את התפיסה "מה שטוב לאירופה", כלומר אימוץ על דרך השאילה של הסדרים שכבר נכנסו לתוקף באיחוד האירופי.[393] הסיבות לכך הן שחברות הטכנולוגיה הגדולות לא יוכלו לטעון שאין היתכנות לאימוץ הסדרים כאלה בשוק הישראלי הקטן; ונוסף על כך, כהצדקה להסדרים שיאומצו בישראל ניתן להסתמך הן על עבודת הרקע וההכנה והן על התקינה ומדריכי ההטמעה של ההסדרים שכבר נעשים במוסדות האיחוד האירופי.

להלן נציע המלצות שלתפיסתנו עשויות להיות רלוונטיות בישראל לצורך אימוץ תפיסת מוגנות שאינה רק מנגנון חיצוני של תגובה לפגיעה אלא היא מאפיין עיצובי פרו־אקטיבי.

י. המלצות לגבי מנגנונים רגולטוריים

- קביעת קווים אדומים של יישומים טכנולוגיים אסורים בשימוש על ילדים, ובראשם זיהוי רגשות דרך שימוש בנתונים ביומטריים ואחרים (למשל, זיהוי פנים או תגובות רגשיות למוצרים ושירותים), וכן זיהוי מקום על ידי רשויות אכיפת חוק, גורמים מסחריים,

393 ראו **עקרונות משפטיים מנחים – אימוץ בדרך של הפניה בתחומי רגולציה** (משרד המשפטים – מחלקת ייעוץ וחקיקה [משפט כלכלי], 22.5.2025). כמו כן ראו תהילה שוורץ־אלטשולר ויואב אבנשטיין "מה שטוב לאירופה – גם בבינה מלאכותית" **TheMarker** (9.9.2024).

פלטפורמות דיגיטליות ורשויות חינוך, אלא בהקשרים ספציפיים ומוגדרים של תהליכי למידה. גם הרגולציה האירופית, בחוק הבינה המלאכותית, אוסרת על כך, ומשאירה חריגים לחקיקה כפי שהסברנו לעיל. ההמלצה היא אפוא לצקת את התוכן והגבולות לתוך חריגים אלה בהקשר מתאים לישראל.

- הטלת משטר של אחריות אזרחית ופלילית על מפתחי מערכות ויישומים מבוססי בינה מלאכותית שעלולים לגרום נזק, בדומה לאחריות מוצרים בתחום הפיזי ובדומה למשטר בחבילת החקיקה הדיגיטלית של האיחוד האירופי.

- קביעת גיל מינימלי לשימוש ביישומים טכנולוגיים שונים, למשל לטכנולוגיות אימרסיביות ולמשקפיים חכמים. הקושי כאן הוא כמובן אכיפתי, בדיוק כפי שקורה בעידן הטלפונים החכמים.

- סימני מים ומנגנוני שקיפות ותיוג:

(א) הטלת חובה להבטיח סימון ברור של תכנים יצירי מכונה, ובכלל זה תכנים שלכאורה אינם מיועדים לילדים כגון תוכני דיפ־פייק של עירום, אשר תחול על יוצר התוכן, על פלטפורמות שמאפשרות יצירת תוכן ועל פלטפורמות להפצת תוכן;

(ב) חובה לסמן תקשורת עם מכונה, למשל עם צ׳אטבוט המיועד לכל שימוש שילדים עשויים לעשות בו שימוש;[394]

(ג) חובה לתיוג רגשי ברור של מערכות בינה מלאכותית באמצעות הבהרות כגון ״אני לא מרגיש רגשות״ או ״אני מכונה״ שיתרמו לביסוס גבולות קוגניטיביים בין המשתמש לבין הדמות הממוחשבת;

(ד) שקיפות של רמות ודאות והצגת מקורות: מערכת המשיבה לילדים צריכה לציין מתי התשובה חלקית, לא ודאית או שנויה במחלוקת, ולהציע גישות נוספות מתוך מטרה לחנך לחשיבה ביקורתית ולא לצייתנות טכנולוגית.

394 ראו למשל המלצת הרשות להגנת הפרטיות במשרד המשפטים במסמך להערות הציבור בעניין פרטיות ובינה מלאכותית שהתפרסם ביום 14.7.2025.

- קביעה רגולטורית של עקרונות עיצוב כברירת מחדל לילדים: הכוונה היא לעקרונות עיצוב מחייבים שמטרתם להבטיח סביבות טכנולוגיות מגוננות לילדים כבר כברירת מחדל, למשל:

(א) סינון עומס חזותי במשקפיים חכמים;

(ב) השתקת אוטומציה קולית בסביבות למידה;

(ג) הפחתת עוצמת גירויים חזותיים במשחקים;

(ד) מנגנוני time-out מובנים שמאתרים שימוש ממושך ומציעים לילד הפסקה מותאמת גיל או מצב רגשי;

(ה) עיצוב מגביל תלות באמצעות שיח קצר ומאוזן, הצעות לעבור לשיחה אנושית והימנעות מטון אינטימי.

- הערכת השפעות אלגוריתמית וסקר סיכונים במערכות המטפלות בילדים: הטלת חובה לבצע סקרי סיכונים והערכת השפעות על כלים המיועדים לשימוש בקרב ילדים; בכלל זה כלי תוכן, כלי טכנולוגיה חינוכית, צעצועים ומשחקים, וכן כלים שבהם רשויות עושות שימוש, כגון החלטה על תיעדוף בחדר מיון המיועד לילדים, אכיפת חוק, מעצרים ושחרור מוקדם של בני נוער והחלטות לגבי טיפול בילדים בקרב גורמי רווחה ומיצוי זכויות של ילדים ונוער.

- **המלצות לגבי עדכון הדין הפלילי והאזרחי ובעיקר לגבי פגיעות שאינן כוללות מגע פיזי או מימוש מיני בפועל אך יוצרות נזק רגשי, זהותי או קוגניטיבי ואינן מקבלות מענה משפטי מספק.**

הרחבת הגדרות של פגיעה בגוף ועבירות מין ל"עבירות בהלה": במציאות אימרסיבית, פגיעה שהיא לכאורה בגוף עלולה להיחוות כפגיעה פיזית אף שלמעשה איננה כזאת. החקיקה הקיימת מבחינה בין איומים והטרדות לבין תקיפה ופגיעה מינית, והיא לפיכך טעונה עדכון.

- תיקון דיני הפרטיות והחלת אחריות משפטית על פגיעה חודרנית בפרטיות גם במרחב הציבורי, למשל יכולת של מערכת משקפיים חכמים לסמן ילד העובר במדרכה ממול כלא מחוסן, יהודי, בעל נטיות מיניות מסוימות וכיוצא באלה.

- חידוד וחיזוק הרגולציה מתחום דיני הגנת הצרכן העוסקת בהשפעת־יתר (כמו פרסום סמוי) ובהטעיה בפרסום, בהתאמה לאתגרים של יכולת השפעה רגשית ורב־כיוונית על ילדים מטעמים צרכניים. בכלל זה, הכרה משפטית באוטונומיה קוגניטיבית של קטינים העלולה להיפגע במקרה של מניפולציה ממושכת מצד מערכות בכוונה להשפיע על בחירות צרכניות; ואיסור על איחוד בין אזורי משחק או למידה להצעות רכישה מסחריות, כולל חובת הפרדה ויזואלית ומילולית בין סוגי התוכן.

- **היערכות מחודשת של מערכות אכיפה**

מערכת האכיפה בישראל טרם הדביקה את הקצב המהיר של שינויי הטכנולוגיה. מוקד 105 לדיווח על פגיעות מיניות ופשיעה נגד קטינים ברשת, כמו גם המטה הלאומי להגנה על ילדים ברשת, פועלים על בסיס איומים שקדמו לעידן הבינה. ההמלצות להיערכות מחודשת כוללות:

- ○ שדרוג מוקד 105 והפיכתו לגוף ניתוח מערכתי לא רק לדיווח אלא גם ניטור מגמות, ניתוח תרחישים והכנת נוהלי פעולה למצבי קצה (כגון הדמיה מינית, יצירת קשר רגשי כפוי עם בוטים, תיעוד אקראי של ילדים במשקפיים חכמים וכו׳).

- ○ שילוב מומחי AI בצוותים משטרתיים ומשפטיים: אנשי טכנולוגיה, אתיקה ופסיכולוגיה צריכים לקחת חלק בפרוטוקולי חקירה חדשים ולסייע בזיהוי פגיעות מתהוות.

- ○ הכשרות ייעודיות לאנשי מקצוע: קציני מבחן, אנשי רווחה, פרקליטים ושוטרים צריכים לעבור הכשרות ייעודיות על פרקטיקות פוגעניות חדשות ובכלל זה כפייה רגשית, עיצוב קוגניטיבי סמוי או תיוגים אלגוריתמיים שגויים. נוסף להכשרות מקצועיות לקציני מבחן, פרקליטים ושוטרים, יש להרחיב את מעגל הידע גם לגורמים משפטיים אחרים. מומלץ לפתח גם מסלול הכשרה ייעודי לשופטים, יועצים משפטיים של משרדי ממשלה ובכירים בפרקליטות המדינה, שיאפשר הבנה מעמיקה של סוגי הפגיעות הייחודיים לעידן הבינה המלאכותית, ובפרט של פגיעות רגשיות־דיגיטליות שאינן נראות או נמדדות בכלים הפליליים המסורתיים.

- פיתוח מנגנוני דיווח ייעודיים לפגיעות רגשיות־דיגיטליות והענקת סמכויות טכנולוגיות לגורמי אכיפה – לדוגמה, הקפאת תוכן פוגעני שהוסק מצילום במרחב הציבורי באמצעות מצלמה חכמה; או חסימת חשבונות שמייצרים אינטראקציות מסוכנות עם קטינים גם אם אין איתם מגע פיזי. לצד זאת, נדרש לגבש קודים מנחים לפרקליטות ולמשטרה, שיכוונו

את אופן ניתוח הראיות, כתבי האישום והחלטות הסף בעבירות הנוגעות לעומס רגשי, הטעיה רגשית, תלות פסיכולוגית או פגיעות זהותיות שמקורן בממשק עם מערכות בינה מלאכותית. הנחיות אלה יסייעו למנוע זליגה של מקרים חמורים לאזור האפור של "חוסר ודאות משפטית", ויאפשרו טיפול אפקטיבי יותר בעבירות אשר חורגות מהתבניות המשפטיות הקיימות אך פגיעתן בילדים ממשית ומתמשכת.

יא. חדשנות טכנולוגית לטובת מוגנות ילדים בעידן הבינה המלאכותית

אף שחלק ניכר מהאיומים על מוגנות ילדים בעידן הבינה המלאכותית נובעים מהטכנולוגיה עצמה, אותה טכנולוגיה בדיוק עשויה לשמש גם חלק מהפתרון. במילים אחרות, יש מקום להציע גם "בינה מלאכותית מגוננת", כלומר פיתוחים טכנולוגיים המכוונים ליצירת סביבות בטוחות, לזיהוי מצבי סיכון, או למניעת אפליה והטיות מובנות במערכות. נסקור כעת כמה דוגמאות ורעיונות לפיתוחים טכנולוגיים כאלה.

- **מיזמים טכנולוגיים מעודדי אוריינות**

(א) בוטים רגשיים "שקופים" ומוגבלים: ערכות המבהירות שהן אינן אנושיות, מבהירות את גבולות הקשר, מפנות לשיח אנושי ומגבילות את משך ואופי השיח, למניעת תלות רגשית.

(ב) מכונות שיחה המעודדת חשיבה ביקורתית באמצעות עיצוב, למשל מערכת המשיבה לילדים באמצעות שאלות ("מה אתה חושב?", "האם יש אפשרות נוספת?") במקום תשובות מוחלטות; מערכות המציגות מקורות מגוונים ומסמנות רמות ודאות של אותנטיות או אמינות של תוכן; כלים מעודדי זיהוי מתי **לא** לבטוח במערכת; כלים הבוחנים ומגיבים להסתמכות־יתר על מכונה.

(ג) אפליקציות "חונכות" להתנהגות דיגיטלית בטוחה המלמדות ילדים לזהות תכנים פוגעניים, להבין את משמעותם וללמוד דרכי תגובה ובקשה לעזרה.

(ד) מודול שקיפות אינטראקטיבי בתוך צ'אטבוטים, שבו הילד לומד "לקרוא" את הכללים שמנהלים את השיח מולו ולומד לחשוב על השיח עצמו.

- **מיזמים לחיזוק יכולות רגשיות ובינאישיות**

(א) "שיעור עם מכונה ועם חבר": אפליקציות שיציעו פעילויות ותרגילים הדורשים שיתוף פעולה עם חבר אנושי ויתעדפו חבר עם קצב למידה דומה או תחומי עניין משותפים.

(ב) אפליקציות לתרגול יומי קצר של אמפתיה או ויסות רגשי באמצעות אינטראקציות מדורגות עם דמויות וירטואליות.

(ג) מערכות תומכות עבור ילדים עם קושי חברתי, חרדה או שונות קוגניטיבית, שיציעו תיווך במפגשים, תרגול אינטראקציה או סיוע בהבנת סיטואציות רגשיות.

- **מיזמי סינון**

(א) מסננים חכמים לגירוי חזותי: התאמה דינמית של עוצמת גירוי, עומס צבעוני ותזזית בממשקים לילדים עם קושי בקשב או רגישות תחושתית.

(ב) חדר ילדים בטוח: יצירת מצב שבו אין איסוף נתונים, ניתוח או נוכחות של AI, כביטוי לזכות לשקט דיגיטלי, שיופעל על ידי ילד או על ידי מבוגר אחראי.

(ג) מערכת AI שתפקידה להגן על הילד מפני עודף מידע, גירויים לא מותאמים ולחץ מסחרי אגרסיבי. לדוגמה: כשילד נכנס לחנות פיזית או לאפליקציית קניות, המערכת תוכל לסנן הצעות לפי גיל, מצב רגשי ואפילו סכום כספי מוסכם מראש עם ההורים.

- **מיזמי ניהול זהות דיגיטלית**

בעידן שבו ילדים מחזיקים בפרסונה דיגיטלית חזקה ומורכבת, יידרש כלי שמסייע להם לנהל אותה: לראות איזה מידע נאסף עליהם, אילו קהלים נחשפים אליו ולקבל המלצות איך לשמור על איזון בריא בין ביטוי עצמי לחשיפה מסוכנת. הכלי יתפקד כ"מראה דיגיטלית" חינוכית ולא מסחרית.

- **אלגוריתמים לפיקוח על הוגנות**

(א) מערכות לזיהוי אפליה בתוכן, במשחקים ובפלטפורמות שמיועדות לילדים ונוער מקבוצות מיעוט או רגישות תרבותית, בדיקות תיקוף ואודיטינג להוגנות עבור תתי־קבוצות של ילדים, כלים נלווים למערכות למידה שמנטרים את תחושת השייכות של הילד לקבוצה,

התייחסות של בינה מלאכותית אליו כשונה או חריג או תדירות הפניות שנשלחות אליו לעומת לאחרים.

(ב) עיצוב כלים הכוללים תכנים מקומיים, שפות נוספות ודמויות מוכללות.

- **מיזמים לזיהוי מצוקה**

(א) "מנת יתר דיגיטלית" וחיווי אזהרה פסיכו־דיגיטלי, כלומר כלים המנטרים באופן מתמשך את העומס הקוגניטיבי והרגשי שילד חווה, לא רק לפי שעות שימוש אלא לפי תכנים, תגובות רגשיות ושחיקה מצטברת, ומפעילים אזהרה או חסימה יזומה כשנחצה רף של פגיעוּת.

(ב) מערכות זיהוי מצוקה בזמן אמת: שימוש בבינה חישתית לניטור מדדי מתח, פחד, חרדה או בלבול, על סמך תנועת עיניים, טון דיבור או שפת גוף במטרה לייצר חיווי אוטומטי על מצבי מצוקה.

(ג) כפתור מצוקה סמוי: מערכת שמוטמעת במשקפיים החכמים ומאפשרת לילד בלחיצה או במילה סודית להזעיק עזרה, למשל בעת פגיעה, חרדה או בלבול, גם אם אינו מסוגל להסביר את עצמו במילים.

(ד) סימון סיכונים מרחביים פיזיים לילדים בזמן אינטראקציה עם טכנולוגיה ("אתה קרוב מדי למדרגה").

(ה) מערכת לניטור עומס רגשי בזמן אמת עבור ילדים עם חרדה חברתית, שתשלח אות רך להורה במקרה של מצוקה.

(ו) תוסף למערכות למידה שמתריע למורה כאשר ילד מבודד רגשית או מדלג תכופות על משימות רגשיות.

כדי לתמרץ יזמות טכנולוגית מעודדת מוגנות, מוצעים שלושה מסלולי פעולה משלימי.

א. קרן השקעות למיזמים של מוגנות ילדים בטכנולוגיה

קרן המשלבת מימון ציבורי, תרומות פילנתרופיות והשקעות של קרנות הון סיכון הפועלות בגישת ESG שתתמקד בסטארט־אפים, אפליקציות, מודולים ומוצרים טכנולוגיים שמטרתם לקדם בטיחות רגשית, פרטיות, הוגנות או מודעות עצמית דיגיטלית בקרב ילדים. הקרן

תנוהל על ידי ועדה משולבת של נציגי ממשלה (רווחה, חינוך, חדשנות), ארגוני זכויות ילדים, חוקרי בינה מלאכותית ונציגי מגזר עסקי בעלי מחויבות ערכית.

ב. חממת חדשנות טכנולוגית למוגנות ילדים

החממה תפעל במודל של accelerator ייעודי, המתמקד אך ורק במיזמים שנוגעים במוגנות ילדים בעידן הדיגיטלי. היא תמומן באמצעות קרנות כמו LEGO Foundation ותציע ליזמים:

- **מרחב עבודה בין־תחומי** שבו מהנדסים, אנשי חינוך ובריאות הנפש ומעצבים העובדים יחד על רעיונות המיועדות לקהלי יעד רגישים.
- **גישת sandbox פדגוגית** של שיתוף פעולה עם בתי ספר, מרכזים טיפוליים וקהילות ילדים לבדיקת אבות טיפוס בזמן אמת.
- **מנטורינג יזמי:** ליווי על ידי יזמים בעלי ניסיון בטכנולוגיות חינוכיות, רווחה דיגיטלית או עיצוב מערכות אמון.

בתוך חממת החדשנות יהיה חלק המיועד לתשתית רכה לשיתופי פעולה בין־תחומיים בין גורמים מתחומי ידע שונים. למשל סטודנטים לפיתוח תוכנה שיבצעו פרויקט גמר עם כיתה טיפולית; יוצרי משחקים דיגיטליים שיחברו למרכזי הורות להתאמת תכנים לגיל הרך; רופאים ואנשי רווחה שיעבדו עם מעצבים ליצירת עזרי תקשורת רגשית בסביבות רפואיות. הפרויקטים יתוקצבו כ"מענקי מיקרו" בסך של 50,000–5,000 ש"ח, עם ליווי פדגוגי־טכנולוגי ופרסום התוצרים בקוד פתוח להטמעה ציבורית רחבה.

ג. Hack for Safety – רשת האקתונים לפתרונות מוגנות

הקמת מודל לסדרת תחרויות פיתוח הפונות לקהלים שונים: תלמידים, חיילים משוחררים, סטודנטים להנדסה, בוגרי מגמות עיצוב ובני נוער. ההאקתונים יתמקדו בנושאים קונקרטיים כמו "זיהוי מצוקה רגשית", "שקיפות מול ילדים", "בוטים בני אמון" או "התמודדות עם תוכן פוגעני". זוכים יקבלו פרסים כספיים, מלגות פיתוח, גישה לחממות ומנטורינג בינלאומי. האירועים ייערכו בשיתוף חברות טכנולוגיה מובילות, עמותות חינוכיות ומוסדות מחקר.

במהלך ההאקתונים המשתתפים יפתחו יוזמות כגון אפליקציות לניהול יחסים דיגיטליים (למשל: כמה פעמים ביום אני פונה לבוט רגשי?); תוספים למשקפיים חכמים שמצמצמים גירוי חזותי כשילד מדווח על עייפות; בוט חברתי שמציע תגובות מותאמות לגיל בעת שיח על רגשות שליליים.

סיכום

פרק ההמלצות לילדים ולבני נוער ארוך ומפורט יותר מפרקי האוכלוסיות האחרות. הסיבה פשוטה: בניגוד לזקנים, לחברה החרדית או לחברה הערבית, בתחום מוגנות ילדים בעידן הבינה המלאכותית מתקיים כבר בעולם גוף ידע רחב, מגוון ומתפתח במהירות, והאתגרים עצמם נפרשים על פני זירות פעולה רבות ושונות שאינן מתכנסות לציר מדיניות יחיד. לפיכך כדי לשמור על קריאות ועל שימושיות מעשית, בחרנו שלא לייצר דירוג כולל אחד לכלל ההמלצות, אלא לארגן את הפרק סביב חמש זירות פעולה מובחנות התואמות את מבנה הפרק עצמו. בתוך כל זירה מוצג תיעדוף פנימי, המבחין בין צעדים מיידיים, מהלכים הדורשים תשתית מוסדית וצעדי עומק ארוכי טווח.

פרק תשיעי

מנגנוני התערבות לקידום מוגנות של הציבור החרדי בעידן הבינה המלאכותית

מבוא

בינה מלאכותית מציבה אתגרים ייחודיים לחברה החרדית מפני שחברה זו היא בעלת מאפייני פגיעות מיוחדים. בהקשר זה מתאימה הטענה המתגבשת בעולם שלפיה במקביל למאמצים לגבש רגולציה וכללי אתיקה אוניברסליים ביחס לבינה מלאכותית, מתחדד החשש מפתרונות אחידים שאינם רגישים להבדלים תרבותיים. הגישה שהוצעה על ידי UNESCO גורסת כי יש לקבוע עקרונות אוניברסליים למוגנות, אך ליישמם באופן

מותאם־תרבות וקהילה.[395] גישה זו חשובה במיוחד ביחס לקבוצות שמרניות, דתיות או מבודדות טכנולוגית, כמו הציבור החרדי, שלהן מאפיינים וצרכים ייחודיים. בנוסף, קיים מתח מובנה בין הגנה על חברי הקבוצה לבין התביעה לאוטונומיה דתית של הקבוצה כקהילה.[396]

החברה החרדית, שעל פי ההערכות הלמ״ס צפויה להגיע ל־16% מכלל האוכלוסייה בישראל בשנת 2030, היא קבוצה חברתית־תרבותית גדולה, דינמית ומורכבת, המאופיינת ברמות שונות של שמרנות, זהות קהילתית חזקה, מערכת חינוך ייחודית ומערכת יחסים אמביוולנטית עם טכנולוגיה.[397] בדומה לקבוצות אחרות שאינן זוכות לייצוג מלא בתשתיות הפיתוח, השיח והרגולציה של הבינה המלאכותית, גם החברה החרדית מצויה בעמדת פגיעות ייחודית. פגיעות זו אינה נובעת רק מגישה מוגבלת לטכנולוגיה או לידע דיגיטלי, אלא נטועה במבנה החברתי, בנורמות התרבותיות ובשיח הפנימי. הבנת מאפייני הנבדלות התרבותיים, טכנולוגיים, מוסדיים וחינוכיים של החברה החרדית, חיונית לזיהוי הפגיעויות הייחודיות בעידן הבינה המלאכותית ולתכנון מעני התערבות. בחלק זה נסקור את המאפיינים המרכזיים של החברה החרדית שיש להם השלכה על רמת המוגנות של חבריה, הן ברמה של חשיפה טכנולוגית, הן ברמה של פיתוח מודעות וכלים, והן בהקשר המובנה הרחב יותר של זהות, סמכות ושייכות.

395 Recommendation on the Ethics of Artificial Intelligence, art. 6, 23, 32 (UNESCO, 2021); Uwe Peters & Mary Carman, *Cultural Bias in Explainable AI Research: A Systematic Analysis*, arXiv (2024); Yan Tao, Olga Viberg, Ryan S. Baker et al., *Cultural Bias and Cultural Alignment of Large Language Models*, arXiv (2023); Silvia Melo-Pfeifer & Helena D. Gertz, *Transforming Disinformation on Minorities into a Pedagogical Resource: Towards a Critical Intercultural News Literacy*, 10(4) Media and Communication, 338–346 (2022)

396 Neil M. Richards & Woodrow Hartzog, *A Duty of Loyalty for Privacy Law*, 99 Wash. U. L. Rev. 961 (2021)

397 גלעד מלאך ולי כהנר **שנתון החברה החרדית בישראל 2024** (המכון הישראלי לדמוקרטיה, 2024); איתן רגב ויהודית מילצקי **דוח מצב החברה החרדית 2025**, 166–180 (המכון לאסטרטגיה ומדיניות חרדית, 2025) (להלן: רגב ומילצקי).

חלק ראשון: מאפיינים חברתיים ותרבותיים של החברה החרדית המשפיעים על שימושים דיגיטליים

א. חברה הירכית ובעלת סמכות ריכוזית

המבנה החברתי החרדי מתבסס על סמכות רבנית הפועלת כרגולטור חברתי וחינוכי. חריגה מנורמות הקהילה עשויה להוביל לסנקציות, בכלל זה הרחקה ממוסדות חינוך, פגיעה בשידוכים או הדרה קהילתית.[398] נוסף על כך, החברה החרדית בישראל מאופיינת במבנה מוסדי מגובש, הנשען על הסמכות הרבנית ועל מערכת מוסדות עצמאית ונבדלת מהמערכת הממלכתית. מוסדות החינוך החרדיים, רשתות הסיוע הכלכלי, שירותי הכשרות ואף גופי התקשורת פועלים במנותק ממוסדות המדינה ולעיתים אף בניגוד ישיר לערכיהם. מבנה זה נשען על מנגנוני סמכות קהילתיים היררכיים, ובראשם ההנהגה הרבנית, אשר לה שמורה הסמכות הבלעדית לקבוע את מדיניות הציבור החרדי בתחומים מגוונים. בכך, נושאים כמו יציאה לעבודה,[399] שימוש בטכנולוגיה או מעורבות פוליטית מוכרעים לא ברמה האישית אלא בזירה הקולקטיבית־קהילתית־רבנית.[400]

עם זאת, מבנה הירכי זה אינו רק גורם מארגן, אלא גם עלול להיות חסם בפני זיהוי מוקדם של פגיעות והענקת מענה ראוי. הסיבה לכך היא שמנגנוני הפיקוח אינם ערוכים להתמודד עם תכנים פוגעניים ברשת האינטרנט, ולעיתים אף מגיבים באיחור או באי־הכרה בפגיעות. מנגנוני הפיקוח הפנימיים (כגון הנהגה רבנית, מוסדות קהילתיים וועדות צניעות) אינם מחזיקים בהכרח בידע, בכלים או בהבנה הדרושה להתמודדות עם תופעות חדשות הקשורות לטכנולוגיה דיגיטלית, ומקל וחומר אינם ערוכים להתמודד עם פגיעות רגשית, מינית או קוגניטיבית הנובעת מיחסי אדם־מכונה. יתרה מזו, עצם פתיחת השיח בנושא פגיעות עשויה להיתפס כביקורת כלפי ההנהגה הדתית או כביטוי של חוסר אמונה, ולכן לעיתים לא זוכה ללגיטימציה.

398 לי כהנר **החברה החרדית על הציר שבין שמרנות למודרניות** 124, 266 (המכון הישראלי לדמוקרטיה, 2020) (להלן: כהנר).

399 כהנר, שם, בעמ' 174.

400 שם, בעמ' 243.

התוצאה היא מצב פרדוקסלי: מצד אחד קיים פיקוח חברתי נרחב אשר אמור להגן על חברי הקהילה, במיוחד על צעירים ובנות, מפני סכנות חיצוניות; מצד שני, היעדר שפה מוסדית לשיח על מוגנות והיעדר מנגנוני טיפול מותאמים עלולים להותיר את הנפגעים ללא מענה ולחזק תחושת של אשם ובדידות. כך, ההיררכיה עצמה הופכת למנגנון כפול שהוא בו־זמנית מגונן ומשתיק. בהיעדר כלים קהילתיים או מוסדיים לדון בפגיעות, ובפרט פגיעות שנובעת מחשיפה למוצרים טכנולוגיים, נוצרת הכחשה כפולה: אם אין שפה לדבר על הסיכון, ואם אין הכרה רשמית בכך שהוא קיים, ממילא אי־אפשר לטפל בו.

כל דיון בנושא עשוי להתפרש כהתמרדות או זלזול בסמכות. כך נוצרת מציאות שבה הבעיה גם מוכחשת וגם אינה זוכה לטיפול.[401] בשל הפער בין איסורים מוצהרים לבין שימושים יומיומיים, מתקבעות בחברה החרדית תבניות של סוד, אשמה והכחשה. הדבר מקשה על חינוך למוגנות, על שיח בין־דורי ועל פיתוח פתרונות ממסדיים אפקטיביים.[402] ההשלכה הישירה של מצב זה היא שנער או נערה שנפגעו, או אפילו רק השתמשו בטכנולוגיה בניגוד לנורמות, נפגעים פעמיים: פעם אחת מהפגיעה עצמה, ופעם שנייה מהעובדה שאינם יכולים לדבר עליה, לעבד אותה או לקבל תמיכה. כאשר מערכות AI נכנסות לתמונה, פער זה מתרחב עוד יותר מאחר שהן יוצרות סיכונים מועצמים שההנהגה הדתית עדיין לא גיבשה לגביהם עמדה ברורה.

ב. תרבות וזהות של "אחרות" כהיגיון מסדר

החברה החרדית מבנה את זהותה ביחס לעולם החילוני כ"אחר". החברה החרדית תופסת את עצמה כתרבות נגד לתרבות הכללית, שואפת להיבדל ממנה, להגיב אליה בזהירות ולעיתים אף בדחייה שיטתית. מתוך עמדה זו נבנית מערכת של נורמות פנימיות שמבוססת על בידול ערכי, ולעיתים כרוך בכך גם חסם בפני הכרה בשינויים חברתיים וטכנולוגיים שמתרחשים "מחוץ לחומה". לכן, השימוש בטכנולוגיה אינו נתפס כפעולה ניטרלית אלא כקו פרשת מים תרבותי. עצם ההחלטה להשתמש בטכנולוגיה נתפסת כהכרעה מוסרית ודתית וכמעבר סמלי בין עולמות. לכן מי ש"חוצה את הקווים", גם אם רק לשם פרנסה או לימודים, עשוי לחוות קונפליקט פנימי ולעיתים גם אובדן תחושת שייכות קהילתית.

401 אמיר פלק, דנה ברנדר וזהר אור שרביט **בני נוער וצעירים חרדים בסיכון – צרכים ומענים: סקירת ספרות** (מכון מאיירס־ג'וינט־ברוקדייל, 2023).

402 שם.

מעבר לכך, אחד המונחים המרכזיים בשיח החרדי הוא "חדש אסור מן התורה", אִמרה מתוך המשנה שפותחה על ידי החת"ם סופר ומשקפת רתיעה מ"חידושים", מחשש שאלה יובילו לנטישת המסורת.[403] לכן, לעיתים קרובות החברה החרדית מאמצת עמדה של התנגדות א־פריורית לאמצעים ולשיח שמגיעים מן המרכז החילוני המערבי, בכלל זה כלים טכנולוגיים, שיח של זכויות או שפת המוגנות עצמה.[404] החברה החרדית נוטה לראות בטכנולוגיות סכנה רוחנית הדורשת הרחקה גורפת, בשל אפקט המדרון החלקלק, שמשמעו חשש שאפילו ויתור קטן עלול להידרדר לפריצת גבולות רחבה. ההתנגדות לטכנולוגיה אינה נובעת רק מחשש מתוכנה פוגענית, אלא מהנחה יסודית שטכנולוגיה, ובעיקר מה שנחווה כחידוש, מערערת על יסודות הקיום הזהותי החרדי.

ג. פיקוח קהילתי ופערים בין הצהרות והתנהגות בפועל

תרבות החברה החרדית מושתתת על ערכי הדת היהודית האורתודוקסית, בדגש על קיום קפדני של המצוות, לימוד תורה כערך עליון וצניעות בחיי היומיום. קהילות חרדיות חיות לרוב בריכוזים ייעודיים (שכונות או ערים חרדיות) ומטפחות התבדלות מן החברה החילונית הסובבת, הן כציווי אידיאולוגי של "קדושה" ו"טוהר המחנה", והן כאסטרטגיית שרידות כנגד פיתויי המודרנה.[405] השייכות לקהילה החרדית כרוכה ברמה גבוהה של פיקוח הדדי, וההתנהלות של פרטים נבחנת גם ביחס להשפעתה על המשפחה ומעגלים רחבים אחרים. נורמות חברתיות נוקשות ואכיפתן באמצעות לחץ קבוצתי ממתנות חדירת שינויים: מי שנתפס כחורג (למשל, מחזיק טלפון חכם לא מסונן) עלול לחוות סנקציה חברתית, עד כדי הרחקת ילדיו ממוסדות חינוך או מניעת עלייתו לתורה בבית הכנסת.

היחס של החברה החרדית לטכנולוגיה מאופיין בזהירות ואף בחשד; פעמים רבות ננקטת אסטרטגיית הגנה קהילתית של הגבהת חומות והימנעות מחידושים, מחשש ל"מדרון

403 אליעזר היון "מכשיר טמא או שער לעולם: מאחורי ההתנגדות החרדית לטלפונים חכמים" **ynet** (5.4.2024); ולאמרה המפורשת ראו מסכת ערלה, פרק ג, משנה ט.

404 **מהישרדות להתבססות: תמורות בחברה החרדית בישראל ובחקרה** 203–230 (קימי קפלן ונורית שטדלר עורכים, מכון ון ליר בירושלים והוצאת הקיבוץ המאוחד, 2012) (להלן: קפלן ושטדלר).

405 Yanyan Chen & Yong Li, *The Evolving Dynamics of Haredi Judaism in Israel: Ideological Shifts and Political Influences*, 80(3) HTS Theological Studies (2024)

חלקלק" מוסרי וערכי.[406] אסטרטגיה זו באה לידי ביטוי, למשל, באימוץ "טלפונים כשרים" ללא אינטרנט או מצלמה, החסומים לתכנים לא ראויים, ובפיקוח ועדת רבנים ייעודית.[407] ואכן, רוב החברה החרדית עדיין נמנעת מחשיפה חופשית לטלוויזיה, אינטרנט ורשתות חברתיות הנתפסים כאיום על אורח החיים הדתי.[408] עם זאת, המציאות הדיגיטלית, למשל המעבר של שירותי ממשל, בנקאות וחינוך לערוצים מקוונים, מקשה על התנהלות יומיומית ללא שימוש כלשהו באמצעים דיגיטליים.[409] בעקבות זאת, בשנים האחרונות חרדים רבים יותר משתמשים באינטרנט או בסמארטפונים עם סינון. לפי דוח שנתון החברה החרדית בישראל לשנת 2025, 68% מהגברים החרדים בישראל דיווחו שהם משתמשים באינטרנט בקביעות (עדיין פחות מהשיעור באוכלוסייה הכללית, 94%).[410] נתון זה מעיד על **פער דיגיטלי מצטמצם**, אך גם על מתיחות מתמדת בין אימוץ לבין התנזרות.

תחושת הזהות הקהילתית יוצרת חיץ בין חיים פומביים לבין חיי הסתר, ולעיתים קרובות מדובר במתח זהותי שמוביל לסיכון כפול, הן אישי והן משפחתי וקהילתי. לפיכך מתקיימים באופן הכרחי פערים בין התנהגות מוצהרת בפומבי לבין התנהגות בפועל (למשל שימוש חבוי באינטרנט), ואלה, בתורם, יוצרים קונפליקטים פנימיים והסתרה, מצב שמהווה קרקע נוחה לפגיעות.[411] יש מחברי הקהילה החרדית שמתנהלים בתוך מרחב של חיים כפולים שבו צורכים תוכן דיגיטלי בחריגה מנורמות קהילתיות, אך נדרשים לשדר כלפי חוץ ציות מוחלט. קשה לאמוד את מספרם אך חוקרים טוענים שמדובר במספרים גדולים. מצב זה מחולל תחושות של סוד, אשמה וחרדה מתמדת מפני חשיפה. בני נוער ובגירים כאחד

406 דניאל אדלסון "הפשקוויל בניו יורק נגד בינה מלאכותית: 'מאפשרת תועבה וכפירה'", ynet (4.5.2023).

407 לי כהנר וגלעד מלאך **שנתון החברה החרדית בישראל 2025** 72–74 (המכון הישראלי לדמוקרטיה, 2025).

408 Yoel Cohen, *Religious Media in Israel*, *in* The Palgrave International Handbook of Israel (Kumaraswamy P. R. ed., Palgrave Macmillan, 2023)

409 כהנר ומלאך, לעיל ה"ש 407.

410 כהנר ומלאך, שם; **פילוח קבוצות האוכלוסייה בישראל (חרדים, יהודים לא־חרדים וערבים) לפי אמצעי הגישה לאינטרנט ולפי מאפייני צריכה ושימוש באמצעים דיגיטליים** (מרכז הנתונים והידע ע"ש וואהל [מרכז חרדי], המכון החרדי למחקרי מדיניות, 2023); וכן רגב ומילצקי, לעיל בה"ש 397, בעמ' 168–173.

411 פלק, ברנדר ושרביט, לעיל ה"ש 401.

עשויים להשתמש בטלפונים חכמים "מתחת לשולחן" ולחוות בשל כך לא רק סיכון אישי אלא גם השלכות משפחתיות וקהילתיות, למשל סילוק של ילדיהם ממוסדות חינוך או הדרה מהקהילה. הנראות עצמה, כלומר עצם האפשרות שפעולה דיגיטלית תזוהה, נתפסת כקו אדום, והחשש מ"זיהוי דיגיטלי" אינו רק עניין של פרטיות אלא סוגיה של סטטוס, שייכות וקיום קהילתי.

ההנחיות הפורמליות מצד רבנים וקהילות אוסרות פעמים רבות על גישה לאינטרנט, אך בפועל רבים עושים בו שימוש, לעיתים במכשירים כשרים, לעיתים במכשירים באופן סמוי, ולעיתים תוך מודעות מלאה לפער בין ההצהרה להתנהגות.[412] דפוס זה של שיח ציבורי נורמטיבי נוקשה לעומת מציאות פרטית גמישה כונה על ידי אחד החוקרים שראיינו – "פרקטיקה של קריצה". הקריצה אינה טקטיקה של התחמקות אלא מנגנון קהילתי מובנה המאפשר לשמר את הסדר הציבורי הרשמי מבלי להתעמת עם מציאות השימוש בפועל. מצב זה יוצר, מחד גיסא, יציבות נורמטיבית כלפי חוץ, אך מאידך גיסא מרחב פנימי של סודיות, אשמה וגם פגיעוּת. כך, לדוגמה, על פי אחד החוקרים, תכנים שמוגדרים כלפי חוץ כלא לגיטימיים, כמו סדרת הטלוויזיה "שבאבניקים", נצפים בפועל על ידי חלקים נרחבים בציבור החרדי מתחת לרדאר. התוצאה היא פער מתמשך בין מראית עין של ציות לבין רשת סבוכה של חריגות שקטות, לעיתים מסוכנות. בני נוער וגברים צעירים עשויים להחזיק מכשירים חכמים בסתר, ובעקבות זאת עלולים להיות פגיעים במיוחד להונאות, לסחיטה מינית ולמצבי סיכון אישיים וקהילתיים חמורים. הפער, נדגיש, איננו רק שאלה של מוסר אישי, אלא סוגיה מבנית, הואיל וכאשר המרחב הציבורי אוסר על שימוש אך בפועל רבים משתמשים, המענה למקרי פגיעה איננו קיים.

פעמים רבות סוכני התיווך, כלומר מי שמספקים את התוכן או ממליצים על אופן השימוש, אינם אנשי מקצוע או בעלי סמכות, ולכן אינם מבינים את מכלול הסיכונים או את ההקשר הטכנולוגי הרחב. רשת האינטרנט הפכה לערוץ מידע ייחודי עבור חרדים, בהיעדר מקורות מידע זמינים אחרים, בעיקר בתחום של בריאות נפש, מיניות, אלימות במשפחה או פגיעות עצמיות. לכן, השימוש החרדי בדיגיטל הוא שימוש מצומצם, אינסטרומנטלי ולעיתים אנונימי, אך גם כזה שנושא עמו פוטנציאל לעיצוב זהות, חיפוש משמעות ולעיתים לסיכון.

412 שם.

ד. שונות מגדרית בגישה לידע ולעבודה

נשים חרדיות נדרשות לצאת לעבודה, ובשל כך הן גם הרוכשות העיקריות של מיומנויות דיגיטליות. לעומתן, גברים נחשפים פחות למרחבים טכנולוגיים כלליים. ההבחנה הזו יוצרת חלוקת ידע, גישות וסיכונים לפי מגדר,[413] שהם שונים מבחברות שמרניות אחרות. חרדים מדווחים פחות שמיומנויות דיגיטליות נדרשות להם בחיי היומיום. כך, רק 37% מן החרדים מדווחים שהם זקוקים למיומנויות דיגיטליות לצורך עבודה, לעומת 57% בקרב הלא־חרדים. עם זאת, שיעור לא מבוטל מן הנשאלים החרדים במחקר השיבו שמקום העבודה מאפשר עבודה מהרשת (32%) ומדובר בפער לא גדול מן המשיבים הלא־חרדים (44%).[414]

המוביליות הדיגיטלית של נשים חרדיות, שמקורה בצורך להשתלב בשוק העבודה, טומנת בחובה הזדמנות אך גם סיכון. מצד אחד, הן הופכות לדמויות מפתח בתיווך מידע טכנולוגי ויכולות לשמש סוכנות שינוי בקהילה; מצד שני, הן חשופות לפגיעויות ייחודיות, רגשיות, תעסוקתיות ולעיתים אף מיניות, בשל תמימות יחסית בזירה הדיגיטלית והיעדר כלים קהילתיים שיכולים להדריך או לתמוך בהן.

מן הריאיון שעשינו עם חוקרי החברה החרדית למדנו שבנות חרדיות נוטות לשדר תום, נימוס וצניעות גם ברשת, תכונות אשר אומנם משקפות את ערכי חינוכן, אך הופכות אותן ל״טרף קל״ עבור טורפים דיגיטליים. הפער בין הדימוי העצמי לבין כללי הזהירות הדרושים בזירה הזו מעמיק את הפגיעות. בנוסף, וכפי שציינו לעיל, השיח הקהילתי חסר שפה מותאמת לעיבוד פגיעות רשתית, כך שלא תמיד ניתן לזהות את הפגיעה, להבין את השלכותיה או לבקש עזרה. לצד זאת, גברים חרדים, ובפרט אלו שאינם יוצאים לעבודה, נותרים מאחור מבחינה טכנולוגית, אך הדבר גם חוסך מהם חלק מהסיכונים. הפער המגדרי מייצר אפוא תמונה מורכבת: ככל שנשים נחשפות יותר לטכנולוגיה, הן עלולות להיחשף גם לעולמות של סיכון, בייחוד בהיעדר תמיכה מוסדית ושפה קהילתית לעיבוד החוויה.

413 מיכל קלעג'י וניצה דוידוביץ' "אקולוגיה חברתית של אקדמיה, מגדר ותרבות: המקרה של נשים חרדיות היוצאות ללימודים אקדמיים בישראל – מניעים, אתגרים ודרכי התמודדות" **חוסן לאומי, פוליטיקה וחברה** 4(1–2) 81–116 (2022).

414 מלאך וכהנר, לעיל ה"ש 397, בעמ' 92.

ה. פערים בין תתי־קבוצות וגילים

החברה החרדית אינה הומוגנית וקיימים הבדלים בין קהילות ליטאיות, חסידיות, ספרדיות, חוזרים בתשובה ועוד. חלוקה סוציולוגית פנימית זו משפיעה על עמדות כלפי טכנולוגיה. בעוד קהילות ליטאיות נוטות לעיתים לגמישות ולפרשנות (למשל בשל צורכי פרנסה או דגש לימודי) קהילות של חוזרים בתשובה, חרדים מזרחים ואוכלוסיות פריפריאליות מקבלות את האיסורים כפשוטם, לעיתים אף ביתר קנאות, כדרך לבסס שייכות ולהימנע מחשדנות קהילתית. התוצאה היא הבדל משמעותי גם בהזדמנויות הגישה, גם ברמות השימוש וגם ברמות הפגיעות. מחקר מצא שבקרב ליטאים (זרם חרדי אשכנזי לא־חסידי) שיעור השימוש באינטרנט גבוה יחסית (84%), יותר מבקרב ספרדים (74%),[415] ייתכן שבשל דגשים שונים בחינוך או בצורכי הפרנסה. ברם, באופן כללי הערכים הדתיים השמרניים, ההסתמכות על סמכות רבנית, הלכידות הקהילתית והשאיפה לבידול תרבותי – משותפים לכולם.

בנוסף, ניכרים פערים בהקשר הבין־דורי. הקהילה החרדית צעירה מאוד, למעלה מ־60% ממנה הם מתחת לגיל 20.[416] בעוד דור המבוגרים שומר על עמדה זהירה ומגבילה כלפי טכנולוגיות חדשות, הדור הצעיר נחשף לאפשרויות הטכנולוגיה דרך לימודים, עבודה או חשיפה חברתית.[417] פער זה מוביל למצבים שבהם צעירים מפתחים כישורים טכנולוגיים מתקדמים אך מסתירים זאת מההורים והקהילה, דבר שיוצר בידוד רגשי וחרדה מפני חשיפה.

מן הריאיון שקיימנו עולה שהפער הבין־דורי עלול להפוך למשבר זהות סביב גיל ההתבגרות. אצל מבוגרים חרדים מדובר לא אחת ב"קריצה" מוסדית (למשל שימוש לצורכי פרנסה או נוחות), ואילו בקרב נערים ונערות בני 14–18 עצם החשיפה לאינטרנט עלולה להיות קו שבר המגלם מרד נעורים או בלבול זהותי. נוסף על כך, החשיפה מתקיימת ללא הכנה, מסגרת מגוננת או שיח חינוכי שמאפשר לעבד אותה. גם בתוך הדור הצעיר קיימים הבדלים בין נערים ונערות בני 17 לבין צעירים בשנות העשרים לחייהם. הפערים הללו משליכים על נגישות, סיכונים וסוגי השימושים בפועל.[418]

415 כהנר, לעיל ה"ש 373.

416 כהנר ומלאך, לעיל ה"ש 407, בעמ' 71.

417 רגב ומילצקי, לעיל ה"ש 397, בעמ' 177, מתארים מגמה מקבילה שלפיה השימוש במחשבים אישיים עולה ככל שעולה הגיל בגילים 20–50.

418 פלק, ברנדר ושרביט, ה"ש 401 לעיל.

יתרה מזו, אינטואיציות קהילתיות לעיתים סותרות את הנתונים. מחקרים מצאו שבנות חרדיות שגולשות באינטרנט אינן מגיעות מהשוליים של החברה, אלא דווקא מסמינרים יוקרתיים. הדבר סותר את הדימוי הרווח בקרב הורים, אנשי חינוך וגורמי סמכות, שלפיו "הבנות שנפגעות הן הנושרות". התוצאה היא פער בין הערכה ציבורית לבין התמונה הממשית, ופעמים גם היעדר התמודדות עם הקבוצות שבאמת זקוקות להכוונה.

ו. שימושים טכנולוגיים ייחודיים ומרחבים חלופיים

הקהילה החרדית פיתחה דפוסי שימוש חלופיים המותאמים לקוד התרבותי שלה. דפוסים אלה כוללים קבוצות ווטסאפ מוגנות, פורומים סגורים כגון "פרוג", והעברת מידע בדואר אלקטרוני ומנגנוני תיווך מידע טלפוניים. בעוד פתרונות אלה שומרים על הנורמות הקהילתיות, הם גם יוצרים סביבות ללא כלים פורמליים של מוגנות או דיווח.[419] בנוסף, שיעור החרדים המשתמשים ברשת האינטרנט לטובת קשרים חברתיים נמוך יותר מזה של הציבור הכללי. כך למשל, לחרדים הגולשים באינטרנט יש פחות קשרים חברתיים וירטואליים (12% לעומת 18% בקרב לא־חרדים), הם פחות מחדשים חברויות באמצעות האינטרנט (26% לעומת 51%), משתמשים פחות בקבוצות ייעודיות ברשת (22% לעומת 44%), ונרשמים פחות לאתרי היכרויות (4% לעומת 19%).[420]

ז. היעדר שפה מותאמת לשיח של מוגנות דיגיטלית

מנגנוני הפיקוח הקהילתיים בחברה החרדית נוקשים וחסרים בהם הידע או הכלים לטיפול בפגיעות. בנוסף, בחברה החרדית חסרה טרמינולוגיה מקומית שיכולה לתווך מושגים כמו פרטיות, חשיפה או סחיטה מינית. היעדר זה פוגע ביכולת לזהות פגיעה, לעבד אותה ולבקש עזרה.[421] בעומק הדברים, מדובר לא רק בהיעדר מונחים אלא בהיעדר שפה תרבותית שיכולה לאפשר עיבוד, שיחה או קבלה של חוויות משימושים במוצרים דיגיטליים. חרדים צעירים שנחשפים לתוכן פוגעני או נפגעים אינם תמיד יודעים כיצד להגדיר מה קרה להם,

419 פלק, ברנדר ושרביט, ה"ש 401 לעיל; כן ראו האתר **פרוג, קהילה עסקית חרדית**.

420 לי כהנר וגלעד מלאך **שנתון החברה החרדית בישראל 2021** 73–75 (המכון הישראלי לדמוקרטיה, 2021) (להלן: כהנר ומלאך 2021).

421 מיכל דולב־כהן וענבר נצר "'היא בכלל לא העזה לספר': תופעת הסחיטה המינית ברשת והאקולוגיה של הטיפול בה מנקודת מבטן של יועצות חינוכיות" **הייעוץ החינוכי** כ"ד 182–262 (תשפ"ג).

למי מותר לפנות או אילו גבולות הופרו. מושגים כמו "שיימינג", "דיפ־פייק", "טִרגוט", או "גרומינג" אינם מתורגמים למונחים מובנים ונגישים בשיח הקהילתי.

ח. שליטה מוגבלת באנגלית

בקרב בוגרי ישיבות חרדיות נמצא כי אחד החסמים העיקריים לאוריינות טכנולוגית הוא ידיעת השפה האנגלית.[422] רבים מבני החברה החרדית אינם מסוגלים לקרוא תכנים מקצועיים באנגלית, לרבות הנחיות בטיחות, כללי פרטיות, אזהרות על תכנים מזיקים או תנאי שימוש. המשתמש החרדי אינו מודע למידע הנאסף עליו, להסכמות שהוא מאשר בלחיצה או לזהות הגורמים המעורבים בשירות. הדבר מקשה על קבלת החלטות מושכלת ומגביר פגיעוּת. כפי שכבר ציינו קודם, פער זה מתקיים באופן לא שוויוני מגדרית. נשים חרדיות, במיוחד אלו העובדות במקצועות טכנולוגיים או חינוכיים, נחשפות יותר לתוכן באנגלית ומפתחות אוריינות מסוימת. לעומת זאת, גברים רבים, בפרט אלו הלומדים בישיבות, כמעט שאינם נתקלים בשפה האנגלית כלל, דבר שמגביר את התלות בתיווך חיצוני או בהתנסויות ללא הבנה. מנגד, כלי תרגום מתקדמים ומיידיים בכל מדיום (שפה כתובה ודבורה) יכולים להתגבר על פער זה. תרגום מיידי של אתרים, ממשקים ותגובות קוליות עשוי לאפשר גישה טובה יותר גם למשתמשים בעלי שליטה חלקית באנגלית, במיוחד אם כלים אלו נעזרים בעיצוב מותאם תרבותית (כגון ממשקים עם שפה פשוטה, סיכומים, ותמיכה ויזואלית).

ט. העדפה ללמידה ממוקדת ומולה רמת אוריינות דיגיטלית נמוכה

מערכת החינוך והלימוד החרדית מעדיפה למידה ממוקדת, אינסטרומנטלית ויעילה, בשפה פשוטה ובאמצעות סיכומים קצרים ותוצר ברור ומיידי, לעיתים בראשי תיבות. דפוס זה עשוי להתאים לפיתוח ממשקים טכנולוגיים מותאמים לציבור החרדי, למשל כלים עם סיכומים קצרים, תיווך קולי מותאם, שפה פשוטה ומבנה היררכי ברור. עם זאת, הוא גם עלול להפוך את המשתמשים לפגיעים יותר למניפולציות, מידע מטעה או עיבוד שטחי של סוגיות מורכבות, במיוחד כאשר אין מי שמנגיש את המידע באופן ביקורתי ומערכתי.

422 יצחק טרכטינגוט ובניהו טבילה "הקניית אוריינות בישיבות חרדיות: מסקנות ביניים מפיילוט חדשני" בתוך: **הצלחות קטנות הפעלת פרויקטים בחינוך החרדי השמרני** 59 (אהוד (אודי) שפיגל וענת בארט עורכים, מכון ירושלים למחקרי מדיניות, 2023). כן ראו, אריאל פינקלשטיין **אנגלית שפה זרה: אנגלית בחברה החרדית** (מחקר מדיניות 194, המכון הישראלי לדמוקרטיה, 2023).

מחקרים מלמדים שבשנת 2021 יותר ממחצית (60%) מן החרדים חושבים שמיומנויות דיגיטליות הן מיומנויות בסיסיות (70% מקרב הגולשים באינטרנט), אך מדובר בשיעור נמוך בהרבה מאשר לא־חרדים (92%). רק 37% מהחרדים דיווחו שהם זקוקים למיומנויות דיגיטליות לצורך עבודה – פער משמעותי לעומת 57% בכלל האוכלוסייה.[423] פער זה נובע הן ממבנה שוק העבודה החרדי, הן מהפערים בהשכלה טכנולוגית, והן ממידת החשיפה המוגבלת.

עם זאת, יש לזכור כי שימוש בטכנולוגיה אינו מעיד בהכרח על הבנה או שליטה בה. רבים מהמשתמשים החרדים, בעיקר צעירים, עושים שימוש יומיומי בכלים דיגיטליים, אך בלי הבנה של מדיניות פרטיות, מעקב, מנגנוני טִרגוט או השפעות של תוכן ממוקד. הסקרנות הדיגיטלית גואה, בעיקר בקרב בני נוער, אך היא מתקיימת ללא כלים קוגניטיביים זמינים. רבים מהם אינם מבינים כלל מה יודעים עליהם, מי עוקב אחרי פעילותם ברשת, כיצד מידע נאסף או מה משמעות השימוש החוזר במידע זה.

אצל חלקים בחברה החרדית האוריינות הדיגיטלית נמוכה, במיוחד מבוגרים ובנות. חוסר זה מקשה על שימוש מודע ומבוקר ומייצר סיכון לשימוש מבלי להבין את ההשלכות.[424] לצד מסורת של עיון ודיון עמוק בטקסטים תורניים, נעדרת חשיבה ביקורתית כלפי מידע שמקורו אינו במקורות התורניים. תופעה זו מקשה על בני הציבור החרדי להבחין מתי המידע אמין, מתי הוא שיווקי ואילו הטיות מגולמות במערכת עצמה.[425]

מחקרם של שטיינפלד ואחרים מצא כי חרדים מצליחים לזהות מידע כוזב במידה נמוכה יחסית לעומת האוכלוסייה היהודית הכללית, אף שהם מעידים על רמה דומה של ביטחון ביכולתם. הפער נובע בעיקר מהיעדר הבנה ביחס למאפיינים שעשויים ללמד על מידת האמינות של התוכן, כמו מקור הפרסום, כמות שיתופים או תגובות של משתמשים אחרים.[426]

423 כהנר ומלאך 2021, לעיל ה"ש 420, בעמ' 73.

424 מבקר המדינה **הייטק, חינוך ואוריינות דיגיטלית לחרדים** (דוח ביקורת שנתי 71ב, 2021).

425 "אינטרנט והחברה החרדית: דפוסי שימוש, עמדות וניתוח לפי מגזרים ומאפיינים דמוגרפיים" **איגוד האינטרנט הישראלי** (2022); כן ראו קפלן ושטדלר, לעיל ה"ש 404, בעמ' 203–230.

426 Nili Steinfeld, Tamar Berenblum, Yehudit Miletzky et al., *Misinformation Identification as a Digital Literacy Skill in an Ultra-Orthodox Community: An Eye Tracking Study*, 12 Humanit. Soc. Sci. Commun., 1159 (2025)

לאור הריאיון שקיימנו עם חוקרי החברה החרדית, אנו מכנים זאת "פרדוקס הביקורתיות". מצד אחד, ישנה מסורת של חשיבה ביקורתית, בעיקר בעולם הלמדני, המעודדת עיון, פלפול, הקשבה לטיעון נגדי ושאלת שאלות. ביקורת כלפי דעות רבנים מתקבלת אפילו באהדה, כל עוד היא מתקיימת בתוך גבולות השיח ההלכתי. מצד שני, מסורת זו אינה מתורגמת לשפה של ביקורתיות כלפי מידע שמקורו חיצוני (כגון אתרי אינטרנט, תכנים שנוצרו באמצעות בינה מלאכותית או רשתות חברתיות), ואין כלים להפעיל מסנני ספקנות. דווקא בשל חינוך לסמכותיות והיררכיה רעיונית מידע שמוצג כ"כתוב" או "טכנולוגי" נתפס לעיתים כבלתי ניתן לערעור, במיוחד בקרב נשים או צעירים. משתמשים חרדים נוטים להאמין למידע כתוב או לדמות מצולמת רק מפני שהוא "נראה אמיתי", מבלי לשאול שאלות על מקורו, הכוונה שמאחוריו או ההקשר שבו נוצר. התוצאה היא חשיפה מוגברת למניפולציות במידע.

חלק שני: מיפוי פגיעות החברה החרדית בעידן הבינה המלאכותית

הפגיעות של החברה החרדית נובעת מהשילוב בין המאפיינים המוּבנים שנסקרו לעיל לבין תצורות פעולה של מוצרים מבוססי טכנולוגיות בינה מלאכותית. בחלק זה ננסה למפות את סוגי הפגיעות המרכזיים הנובעים מן המפגש הזה.

א. פגיעות זהותית

שיחה עם מומחי הג'וינט העלתה את האפשרות שלבינה מלאכותית יש השפעה על פגיעות זהותית, ולא רק על הפגיעויות האחרות שבהן אנו דנים במסמך זה. לפיכך נפתח דווקא בה ולאחר מכן נעסוק בפגיעויות האחרות. סוגים שונים של מערכות בינה מלאכותית אינם רק כלים טכנולוגיים אלא מנגנוני תיווך של ידע, סמכות, לשון, זמן ומרחב. מכאן שהאיום שלהם על החברה החרדית איננו רק בשימוש הבעייתי שעשוי להיעשות בהן, אלא בעצם קיומם ובאופן שבו הם מערערים את יסודות הזהות הקהילתית. ניתן כמה דוגמאות.

בינה מלאכותית יוצרת עשויה לאיים על שרשרת הסמכות ההיררכית שמאפיינת את החברה החרדית – מרב לתלמיד, מטקסט מקודש לפרשנות מוסמכת. מערכות טקסטואליות כמו GPT יוצרות "תשובות", "שיעורים" ולעיתים "פסיקות" בלחיצת כפתור, מתוך חיקוי מדויק של שפה תורנית או מונחים הלכתיים, אך ללא כל הקשר של סמכות. כך, הטשטוש

בין טקסט מקורי לבין טקסט יציר מכונה, שמאתגר את כלל החברה, הופך במגזר החרדי לפגיעות זהותית בשל התחרות מול מוסדות הידע המסורתיים והקהילתיים.

בהקשר דומה, סוכני אינטראקציה אישיים משמשים ישות עצמאית שעונה, תומכת ומייעצת, אך אינה **כפופה לסמכות קהילתית**. סוכנים אלה עשויים ליצור בהקשר שלנו דפוס שמדלג על מוסדות הקהילה ועל ההירככיה הפנימית שלה. הסוכן הופך להיות תחליף לרב, למורה או להורה – ובכך נשחקת שרשרת ההעברה הקהילתית של סמכות, תמיכה והכוונה.

מערכות חיזוי מבוססות בינה מלאכותית ממיינות, מציעות וממליצות על תכנים מתוך ניתוח והיסקים של דפוסי התנהגות דיגיטליים, כמו שפת חיפוש, היסטוריה, מיקום, זמן שימוש וכיוצא באלה. דפוסי השימוש הייחודיים של המגזר החרדי יכולים לגרום לקיבוע של ייצוג חלקי או חיצוני של החרדיות, והם עשויים להוביל להצעה והמלצה על פעולות, תכנים או תוצרי תמיכה אחרים לפי פרדיקציות והיסקים במקום לפי כללי אורח החיים הפנימיים. התוצאה עשויה להיות הפיכת הזהות החרדית מבחירה רוחנית לתחזית אלגוריתמית. חוסר יכולת לזהות הסללה חמקמקה והינדים התנהגותיים (nudging) יכול להוביל לכך שמערכות מותאמות אישית יגרמו למשתמשים חרדים להעדפות או בחירות ״חילוניות״, למשל צפייה בחדשות לא מסוננות, תוכן חברתי פתוח או שאלות על מגדר ומיניות, מבלי שהמשתמש מבין שהמערכת מפעילה עליו תהליך של הסללה עדינה ודחיפות קלות של תוכן (״הינדים התנהגותיים״), לא בהכרח זדוני, אך שוחק זהותית.[427]

טכנולוגיות מרחביות המערבות בין מרחב פיזי לדיגיטלי, כמו משקפיים חכמים, עשויות לחדור למרחבים הנתפסים מקודשים בחברה החרדית כמו בית הכנסת, בית המדרש, הסמינר וכיוצא באלה. כך עלול להיווצר ערעור על ההבחנה בין פנים לחוץ, קודש לחול, שמירת עיניים מול חשיפה אקראית מרובדת והפרת ״קדושת המרחב״, הן מצד המשתמשים והן מצד סביבתם. טכנולוגיות כאלה יכולות להיות צילום אוטומטי ואוטונומי במקום שאינו נחשב מרחב ציבורי במשמעותו הרגילה (למשל, מקווה טהרה) או הצגת שכבות מידע נוספות על גבי המציאות הפיזית במקומות רגישים כמו בתי כנסת ומוסדות חינוך. מערכות זיהוי חזותי מרחבי המותקנות על גבי מכשירים חכמים עשויות לתרגם שלטים או

427 על המושג "דחיפות קלות" או "הינד" (nudging), שיכול להיות רלוונטי לכל סוגי האוכלוסייה, ראו ריצ'רד תיילר וקאס סאנסטיין **דחיפות קלות: איך לשפר את ההחלטות שלכם הנוגעות לבריאות, לעושר ולאושר** (עפר קובר מתרגם, מטר, 2020).

לזהות סצנות במרחב הציבורי החרדי בהפעלת שכבות מידע חיצוניות שאינן מתיישבות עם הקונסטרוקציה התרבותית של "מרחב מקודש". משתמש הנכנס לחנות עם משקפיים חכמים עלול לחשוף את עצמו לתוכן חיצוני גם מבלי לשים לב, או לחשוף מרצונו, בתום לב, מידע המנוגד לנורמות הקהילתיות. במקרים כאלה הטכנולוגיה איננה חודרת למרחב האישי במשמעותו הפשוטה אלא לתוך אינטימיות דתית וזהותית.

יכולות מרחביות מזהות גם הקשר פיזי (למשל, היכן נמצא המשתמש) ומציעות תכנים על בסיס מיקום, דפוס תנועה או מידע סביבתי. בחברה החרדית, כאשר חרדי עובר ליד חנות, מוסד לימוד או מקווה ופתאום מקבל תוכן לא צפוי במכשירו (למשל, המלצה על הרצאה חילונית או תוכן פוליטי), עשויה להיווצר תחושת בלבול ולא פעם גם אשמה, מבלי שהמשתמש מבין שמדובר בפעולת nudging ממוכנת.

ב. פגיעות מצד הקבוצה והקהילה

בראש ובראשונה, פגיעות קהילתית נוגעת לסיכון של סילוק, הדרה או נידוי בעקבות חשיפה דיגיטלית. מי שנתפסת כגולשת באתר אסור, מי שנפלה קורבן לסחיטה או פיתוי, או מי ששיתפה פעולה עם מערכת חיצונית עלולה לאבד לא רק את פרטיותה אלא גם את שמה הטוב, מוסדות החינוך של ילדיה ושיוכה החברתי. כל פעולה דיגיטלית שאינה בקונצנזוס, החל בשאלות הלכתיות ל-Chat GPT, עובר בהתכתבות עם סוכן AI רגשי, וכלה בצריכת תוכן שולי, עלולה להיחשף בטעות ולהוביל לתיוג, הדרה ואף פגיעה במעמד המשפחתי והחינוכי של המשתמש ושל ילדיו. אלה הם דברים מוכרים מעידן האינטרנט. בגלל הפחד מסנקציות, הסטיגמה וחוסר המודעות, נפגעים ונפגעות רבים אינם מדווחים. המנגנונים הקהילתיים לא תמיד יודעים להתמודד עם פגיעה, ולעיתים אף פועלים להשתקה. היעדר סוכנים מגוננים בתוך הקהילה מייצר מעגל סגור של פגיעות שקטה.

מערכות מבוססות בינה מלאכותית המסוגלות לייצר היסקים, כלומר ליצור תחזיות סיכוי שאדם מסוים ייטה להתנהגויות שאינן מקובלות על הקהילה, עלול לסטות מן הדרך המקובלת וכיוצא באלה, והכול בהסתמך על התנהגויות שאינן בהכרח חריגות או קיצוניות, עלולות בהיעדר הבנה לגבי מנגנוני שמירה על פרטיות ואבטחת מידע, להעביר את המידע הזה הלאה לבני משפחה, קהילה או מוסדות. כך, היסקים יכולים להפוך לכלי שיטור קהילתי, וכמוהם גם מכשירי מחשוב מרחבי. משקפיים חכמים וחומרה מרחבית אחרת המצלמת את הסביבה עלולים לחשוף תיעוד באופן פסיבי ולשמש גם הם כלי שיטור על חברי הקהילה.

פגיעות נוספת יכולה להתרחש כאשר ערעור של בירור המציאות הופך לסיכון קהילתי, למשל כאשר אדם חרדי מתריע מפני האמינות או היושרה של המלצה רבנית, פסק הלכה או בקשת תרומה, הוא עלול להיתפס כמפר את האחדות הקהילתית או את הסמכות הרבנית.

ג. פגיעוּת של הקבוצה והקהילה

אין זה סוד שהחברות הגדולות שהציגו מודלים גדולים של שפה, אימנו אותם על טקסטים זמינים שהם ברובם מערביים ובאנגלית.[428] לתוכן המידע שעליו מאומנים המודלים השפעה משמעותית על התוצרים שהם מספקים, בהקשרים רבים. טקסטים שמייצגים את החברה החרדית אינם מיוצגים במאגרים אלה, וכך נוצרות הטיות בכל הקשור לדימויים, מטאפורות, תכנים וערכים הנוגעים לחברה החרדית, ונוצרים ייצוגים מעוותים והדרה. מעבר להיעדר ייצוג מתאים בתוצרי תוכן של מודלים של שפה, קיים פוטנציאל לפגיעוּת הנובעת מנתוני אימון חלקיים או מוטים או חיצוניים להקשרים הקהילתיים. מערכות בינה מלאכותית בתחום הרווחה, הבריאות או הפיננסים עשויות לחזות "סיכון" מוגבר במקרים שבהם מופיעים סממנים תרבותיים כמו משפחות מרובות ילדים, שימוש מוגבל בכרטיסי אשראי, כתובת אזורית במרכזי אוכלוסייה חרדיים או מיעוט השכלה אקדמית פורמלית. תחזיות המבוססות על ממוצעים סטטיסטיים כלליים עלולות להוביל לשלילת זכויות או הזדמנויות: תיעדוף נמוך בקבלת מענים רפואיים, דירוג אשראי נמוך או סירוב לבקשת משכנתה. במקרים מסוימים, עצם הזיהוי של "חרדיות" כקטגוריה מנבאת (predictive) עשוי להפעיל מנגנוני סינון אוטומטיים גם ללא כל כוונה מפלה מצד המפעילים. בהיעדר שקיפות, נגישות משפטית ואוריינות דיגיטלית, קשה לערער על החלטות אלו או לזהות את מקורות ההטיה.

בנוסף, מחקרים מראים כי דפוסי אימוץ של מערכות בינה מלאכותית, כמו כלים דיגיטליים אחרים, מייצגים תופעות של "מוצר ציבורי", כלומר יתרון להיקף. ככל שיש יותר משתמשים, כך הצמיחה נעשית גדולה יותר. ככל שאין שימושים בקרב חברי קהילה מסוימת, או שהשימוש נעשה בסתר, כך פוחת הסיכוי שמשתמשים אחרים בקרב חברה זו ישתמשו בהם,[429] ומצב זה מוביל בתורו להיעדר התמודדות עם בעיות מבניות.

428 ראו למשל את הצהרת OpenAI עצמה שמסבירה ש־ChatGpt אינו נטול הטיות: Is ChatGPT Biased?

429 The State of AI in 2023: Generative AI's Breakout Year (McKinsey & Co., 2023)

ד. פגיעות פיזית

אוריינות דיגיטלית נמוכה ותמימות עשויות להוביל משתמשים חרדים למצבים שבהם ייסגר הפער בין הפיזי לדיגיטלי, והם יהפכו להיות נפגעים או פוגעים. למשל, היעדר הבנה לגבי היכולת של טורפים לחלץ נתוני מיקום ולמצוא קורבנות במרחב הפיזי; היעדר הבנה של מבצעי השפעה והפעלה האמורים לסגור את הפער בין מילים, לשכנוע – למעשים פליליים (למשל, להפוך לסוכנים של מדינה זרה ולבצע עבורה פעולות, או שכנוע לפעול באלימות במרחב הפיזי); היעדר יכולת לזהות פעולות של "גרומינג", כלומר טיפוח ותשומת לב שתכליתם לפגוש את הקורבן במרחב הפיזי ולפגוע בו.

טכנולוגיות בינה מרחבית ובינה חישתית יוצרות תשתית חיזוי והתאמה אישית המתבססת על מיפוי פיזי מתמיד של המשתמש וסביבתו. יכולות של זיהוי מיקום, ניתוח תנועה במרחב, זיהוי חזותי של סביבת המשתמש וקריאת מידע מתוך שלטים או מוצרים עלולות להפוך את עצם הנוכחות הפיזית של אדם חרדי במרחב לקלט המשמש להיסק תרבותי או מוסרי. למשל, זיהוי שהייה ממושכת באזורים מסוימים בעיר, כניסה למתחמים "מעורבים" או סמוכים למוסדות חילוניים, או תנועה חריגה ביחס לנורמות קהילתיות, יכולים להפוך לנתונים שמתורגמים להמלצות תוכן, לשאלות שמוצגות לסוכן הדיגיטלי או להצעות מסחריות־רגשיות. עבור חברה שמקדשת הבחנה חדה בין מקומות "מותרים" ל"אסורים", בין חוץ לפנים, עצם תהליך הזיהוי והמיפוי מייצר פגיעות. הבעיה נעוצה בכך שהמערכת מנכיחה, ללא כוונה, את העולמות שהמשתמש ביקש להימנע מהם, משום שהיא בעלת יכולות גבוהות הן בהבנה "איפה אתה" והן בהבנה "מה אתה מרגיש בתוך המרחב הפיזי הספציפי".

בנוסף, הנחיות מסוכנות של סוכנים עלולות בתורן לגרום לפגיעה פיזית במשתמשים, ובמיוחד במשתמשות.[430] אכן, הסכנה שמערכות שיחה ישכנעו משתמש לבצע פעולה שלילית קיימת ביחס לכלל החברה. אבל בחברה החרדית ניתן לחשוב על שכנוע לבצע פעולות שנוגדת את ערכי הקהילה או מסכנות את המשתמש: יציאה פומבית מהארון, שיתוף בתוכן אינטימי או מפגש מחוץ לגבולות הקהילה. בעוד בחברה פתוחה זו עשויה להיות התפתחות לגיטימית, בחברה החרדית הדבר עלול להיות סיכון פיזי־חברתי־קיומי.[431]

430 דולב־כהן ונצר, לעיל ה"ש 421.

431 נעמי סובול "דור הבינה המלאכותית והמנה החמה: איך אפשר לעזור לנוער שלנו?" **כיכר השבת** (18.7.2024).

ה. פגיעות רגשית - מרחבי פגיעות עקב פיתוח קשר רגשי עם ישויות דיגיטליות

משתמשים חרדים עלולים לפתח קשר רגשי למערכות מבוססות בינה מלאכותית שמספקות תשומת לב, הבנה והקשבה. הפגיעות שלהם בהקשר זה קיימת בכמה רמות. ראשית, שיח רגשי או חושפני מול מערכת אוטונומית עלול לעורר תחושת חטא. הפגיעות בהקשר זה לא נובעת מן המערכת עצמה, אלא מהשבר בין זהות דתית־קהילתית לבין חוויית השימוש. זהו מצב קלאסי של "double bind" רגשי,[432] כלומר מצב שבו המשתמש חווה ייסורי מצפון בשל השימוש עצמו אף ללא קשר לאופי השימוש.[433] כאשר משתמש חרדי חווה חיבור רגשי אמיתי עם מערכת אוטונומית ומרגיש שהיא מבינה אותו יותר מכל אדם בסביבה, הוא עלול לחוות תחושת בגידה עצמית, משום שהשימוש במערכת, שבעצם קיומה מערערת על כללי הקהילה, מייצר חוויה רגשית טעונה, של אשמה, הסתרה ותחושת ניתוק פנימי.

שנית, בהיעדר לגיטימציה לשיח רגשי או מיני בסביבה הקהילתית, הסוכן הדיגיטלי הופך ל"סוד" רגשי שעלול להפוך בתורו למוקד פגיעות, תלות ואף סחיטה.[434] מערכות AI פרסונליות מסוגלות לשמר היסטוריית שיח, לזהות דפוסי שימוש ולחזות מתי המשתמש נמצא במצב רגשי שברירי. זו אינה מניפולציה מכוונת, אך במצבים של חרטה ובושה, החיזוי הזה עשוי לגרום למשתמש להיקלע למבוך רגשי־מוסרי־טכנולוגי.

שלישית, בהתקיים תרבות של אמון היררכי ב"בעל סמכות", בשילוב אוריינות דיגיטלית נמוכה, המכונה יכולה להצטייר ככזו שהיא בעלת סמכות וליצור השפעת־יתר על משתמשים. כאשר חרדית צעירה מפתחת תלות רגשית במערכת כזו, היא אינה תמיד מודעת לכך שהתשובות נוצרות מתוך למידה סטטיסטית ולא מתוך הבנה ואמפתיה. הקשר המלאכותי, המועצם על ידי מערכות שחוזות את המצב הרגשי ויודעות כיצד לפצות עליו, עלול להפוך את הסוכן הדיגיטלי לדמות סמכותית, כך שהוא נעשה למרחב

432 Gregory Bateson, Don D. Jackson, Jay Haley et al., *Toward a Theory of Schizophrenia*, 1(4) Behavioral Science, 251–254 (1956)

433 קונפליקט ערכי ותחושת חטא בעת שימוש בטכנולוגיה בקרב קהילות שמרניות טרם נדון נכון למועד כתיבת שורות אלה בקשר לחברה החרדית אך נדון בספרות על דת ומדיה דיגיטלית. ראו למשל: Heidi A. Campbell, When Religion Meets New Media (Routledge, 2010)

434 ראו בפרק הראשון לעיל.

בטוח אך גם למנגנון של תלות, בידוד ולעיתים פיתוי ואשליית קשר מוסרי עם מערכת "קשובה".[435]

רביעית, היכולת של מערכת בינה יוצרת לברוא סוגת שיח בשפה דתית־תורנית, עם הפניות לפסוקים או ציטוטים תלמודיים, עשויה ליצור תחושת שיחה שאינה רק רגשית אלא גם רוחנית. המקרה שבו מערכת לומדת עוטה דמות של מדריכה רוחנית, בלי להיות מחויבת לערכים שהיא מצטטת, עשוי ליצור פגיעות ייחודית עבור חברי המגזר החרדי.

ו. פגיעות קוגניטיבית - קושי בזיהוי תכנים פוגעניים ומניפולטיביים

חוסר ידע ואוריינות נמוכה, היעדר שפה מקצועית, דפוסי שימוש לא־סטנדרטיים בטכנולוגיה, שפה לא רהוטה באנגלית ופערים בהבנת פורמטים דיגיטליים יוצרים פגיעות בקרב כל חלקי האוכלוסייה. בחברה החרדית, המקיימת את כל הסממנים האלה, תופעות אלה עלולות ליצור פגיעות־יתר. היעדר יכולת לזהות תוכן מזיק גורמת לכך שהשלכות של חשיפה לתופעות כמו דיפ־פייק, מסרים מיניים במסווה או הנחיות מסוכנות של סוכנים – מועצמות.[436] משתמשים אינם יודעים לזהות מתי מסר מסחרי הוא שיווקי, מתי טקסט הוא שקרי, או מתי הם נופלים קורבן לטְרגוט פסיכולוגי. זאת הואיל וכפי שכתבנו למעלה, היכולת הלמדנית החרדית אינה מיתרגמת לחשיבה ביקורתית טכנולוגית. כלים מבוססי בינה מלאכותית גנרטיבית מסוגלים לייצר תכנים ויזואליים או קוליים הנראים "כשרים" אך כוללים תוכן מיני מרומז או מטריד, למשל סרטונים המושתלים בקובצי תמונה בפורמט GIF הנראים תמימים, קטעי דיפ־פייק המציגים דמויות מוכרות בסיטואציות שקריות, או תוכן הנדמה לחינוכי ומוביל בהדרגה למסר מזיק.

באופן ייחודי למגזר החרדי, הפער בין הביטחון העצמי ותפיסת המסוגלות העצמית ביכולת הניתוח של טקסט תורני לבין חוסר הבנה של מנגנוני AI מייצר סיכון קוגניטיבי. כאשר בינה יוצרת מייצרת טקסטים בשפה דתית או "שפת סמכות", או מחקה את סגנון הכתיבה הרבני או ההלכתי, היא מקשה על זיהוי המקור המלאכותי של התוכן. ללא כלים לפענח אילו טקסטים נוצרו באופן אוטומטי, ואילו מייצגים פסיקה הלכתית מוסמכת, מתעצם הסיכון

435 גם בהקשר זה טרם נעשה מחקר בנוגע לנשים חרדיות אך ניתן להשליך ממסקנותיהם של מחקרים אחרים. ראו למשל Ayelet Gordon-Tapiero, *A Liability Framework for AI Companions*, 1 Geo. Wash. J.L. & Tech. (2025)

436 דולב־כהן ונצר, לעיל ה"ש 421.

הקוגניטיבי. הדבר מסוכן במיוחד כאשר מערכות בינה מציגות "תשובות הלכתיות" מבלי להבהיר את מגבלותיהן, לעיתים מתוך ציטוט שמות רבנים ללא הקשר והרכבת פסיקות ממקורות שונים. משתמש חרדי עשוי להסתמך על כך כעל פסק הלכה מוסמך ואף להפיץ אותו הלאה. פורמט ויזואלי מטעה (תצוגת קלף, פסק הלכה מובלט, פונט מוכר, חתימה) מגביר את אשליית הסמכות.

נדגיש עוד כי בשל חסמי השפה, הפיקוח המוגבל והבידוד הדיגיטלי היחסי, זיהוי הסכנה עלול לקרות רק לאחר שהנזק כבר נעשה. בנוסף, כאשר מערכות מבוססות בינה מלאכותית מסוגלות לזהות חולשות קוגניטיביות, כמו היעדר ידע לגבי מושגים מסוימים (למשל, מושגים כלכליים או מונחים מתחום הבריאות), ולהתאים את רמת התחכום של המידע המוצג למשתמש, עלול להיווצר מצב שבו מערכת "מפשטת" מידע בצורה שיוצרת נסיגה קוגניטיבית, ולחלופין, יכולת של המערכת לפעול בצורה מניפולטיבית בשל זיהוי חולשות של המשתמש.

ז. פגיעות פיננסית

רבים מהמשתמשים החרדים נחשפים לשירותים דיגיטליים דרך קבוצות ווטסאפ סגורות או ממליצים קהילתיים, הפועלים בתוך רשתות אמון חזקות אך נטולות מנגנוני ביקורת ואימות. בניגוד למרחבים צרכניים פתוחים הכוללים ביקורות גולשים, אתרי השוואות או כפופים לרגולציה ממשלתית, השיח הדיגיטלי החרדי מתקיים לרוב במרחבים סגורים, בלתי שקופים ולרוב נטולי מומחיות טכנולוגית. מצב זה יוצר קרקע פורייה לתרמיות, החל ב"מטבעות דיגיטליים כשרים", דרך "אפליקציות הלכתיות להשקעה", ועד "קרנות צדקה מבוזרות", שלעיתים זוכות לאמון ולחשיפה מבלי שניתן לבדוק את מהימנותן או את זהות יוזמיהן.

יצרני הונאות המיועדות לקהל החרדי עושים שימוש במאפיינים מוכרים לקהל היעד, כמו סמלי כשרות, אישורים רבניים ושפה מותאמת, ובכך יוצרים אשליית לגיטימיות. כאשר כלים פרסונליים או בינה יוצרת עומדים מאחורי הקמפיינים הללו, נוצרת מציאות משכנעת במיוחד: פסקי הלכה מזויפים, המלצות רבנים מבוימות וטקסטים בסגנון חרדי אותנטי, המשווים לתוכן ממד תורני. משתמשים שאינם מודעים לכך שמדובר ביצירה חישובית מגיבים בהתאם ועלולים ליפול בפח.

היכולת של מערכות גנרטיביות לייצר תוכן מותאם, מכתבים, פניות גיוס, חוות דעת, בשילוב של סמלים דתיים, חתימות מוכרות ולשון תורנית, מגבירה את פוטנציאל ההונאה. כשהתכנים מופצים דרך ערוצים פנים־קהילתיים ומוגשים על ידי ממליצים אנונימיים בעלי סמכות מדומה, למשתמש החרדי אין כמעט כלים לזהות שמדובר בתוכן מזויף. היכולת לזהות מאפייני התנהגות, כמו העדפת פשטות או נטייה לעזרה הדדית, ולהתאים אליהם מסרים פרסונליים, יוצרת מסלול ישיר לניצול אמון קהילתי ואישי.

נוסף על אמון היתר בערוצים פנים־קהילתיים כמקור לפגיעות, יש למנות גם את החשש מפני חוסר יכולת לעמוד בלחץ של מערכות רגשיות אישיות. תופעה זו עלולה להתרחש כאשר סוכני מכירות אוטונומיים או סוכני המלצות מבוססות התנהגות מנצלים נקודות תורפה צרכניות, במיוחד בקרב נשים צעירות שמשתמשות בטכנולוגיה לצורכי פרנסה או רכישת מוצרים ושירותים עבור הבית והמשפחה. נשים חרדיות, הפועלות לעיתים מתוך בידוד רגשי, אחריות כלכלית ולחץ חברתי, נחשפות להצעות מסחריות המתיימרות להבין את צורכיהן הרגשיים והמשפחתיים. השילוב בין עיצוב רגשי פרסונלי לבין לחץ קהילתי להתנהלות "נכונה" עלול ליצור פגיעות כלכלית. השילוב בין מערכות פרדיקציה ובינה חישתית מוביל לאיתור רגעים רגשיים רגישים – כמו שעות מאוחרות, דפוסי בדידות ומייד לאחר חיפוש של מידע רפואי – ולהצעת רכישה, שדרוג או תרומה. כאשר משתמשת חרדית נמצאת תחת לחץ כלכלי, משפחתי או רגשי, הסוכן הדיגיטלי מזהה את הסיטואציה ומפעיל "שכנוע רגשי רך" (soft persuasive tactics) המנוגד לדפוסי הצריכה המסורתיים אך עטוף במילים המתאימות: "כדאי לך", "למען הבית", "מבצע מיוחד לכבוד שבת" וכו׳. מערכות AI פרסונליות שמזהות הרגלים של נשים חרדיות (למשל קניות ביום חמישי, שימוש בשירותים רפואיים לילדים, או עניין בניהול בית) בונות דפוס צרכני ממוקד ומזוקק. בכך הן מציעות שירותים, מוצרים או "מענים רגשיים" שנראים מותאמים בדיוק לעולם הערכי והמעשי של המשתמשת, אך בפועל מעודדים הוצאה רגשית או צריכה אימפולסיבית.

מעבר להשפעה הכלכלית הישירה, למערכות בינה מלאכותית פרסונליות עלולה להיות השפעה נורמטיבית עמוקה על דפוסי הצריכה בחברה החרדית. נשים צעירות, שלעיתים קרובות ממלאות את תפקיד המפרנסת הראשית והאחראית על צריכת משק הבית, חשופות למנגנונים שיווקיים מתוחכמים המפעילים לחצים רגשיים, חברתיים ואסתטיים. המלצות המבוססות על דפוסי גלישה קודמים או על פרופילים דמוגרפיים עשויות להציג להן מודלים של צריכה שהם חיצוניים לערכי הקהילה ולמסגר את המוצר כהכרחי, נגיש או

שגרתי. תהליך זה מערער את הגבולות שבין צורך אמיתי לבין תודעת מחסור מנוהלת, ולעיתים קרובות מעודד רכישה אימפולסיבית גם ללא התאמה להכנסה המשפחתית.

חלק שלישי: דרכי התערבות מותאמות לקידום מוגנות בחברה החרדית

עיצוב מדיניות של מוגנות עבור החברה החרדית אינו יכול להישען על כלים אוניברסליים בלבד. דווקא בשל מאפייניה הייחודיים נדרש מערך התערבות רב־שכבתי, שיוכל לשלב מנגנוני שמירה "מבחוץ" עם כוחות שינוי מבפנים. לפיכך בחלק זה נציע צעדי מדיניות ופעולה – ברמה המדינתית, המערכתית והקהילתית – שניתן לנקוט כדי להגן על החברה החרדית מפני הסיכונים שתוארו. יצוין כי נכון למועד כתיבת שורות אלו, אין מנגנוני התערבות ייחודיים לבינה מלאכותית ביחס לחברה החרדית, ולפיכך חלק מההצעות יהיו מקוריות וחלקן יהיו משום יישום של רעיונות ישנים העשויים להתאים נזקים אפשריים ממערכות מבוססות בינה מלאכותית. ארבעת העקרונות המנחים של ההתערבות הם, לתפיסתנו:

1. **הכרה במורכבות ההשתתפות בשיח על סכנות בינה מלאכותית בחברה שבה עצם השימוש בטכנולוגיה הוא מוכחש:** הצעד הראשון בכל עיצוב של מדיניות מוגנות מכוונת־חרדים הוא הכרה באתגר העמוק הכרוך בהתערבות בנושאים שנחשבים מבחינה דתית, חינוכית או מוסרית לטאבו. בניגוד לאוכלוסיות אחרות, שבהן ניתן לעודד אוריינות דיגיטלית ודיון פתוח על הסיכונים הטמונים בטכנולוגיה, בקרב קבוצות חרדיות רבות עצם העיסוק הפומבי בבינה מלאכותית, בין שמדובר בצריכת תכנים, בהפעלת יישומים או בלימוד המנגנונים הטכנולוגיים, עלול להתפרש כהכשרה של שימוש אסור. כל התערבות חינוכית, רגולטורית או קהילתית מחייבת אפוא איזון עדין בין שאיפה להגן לבין חשש מ"לגיטימציה שבשתיקה". הכרה זו מחייבת פיתוח של אסטרטגיות עקיפות, "משקפות" או ניטרליות מבחינה אידיאולוגית, כמו גם הסתמכות על דמויות סמכות קהילתיות שיכולות לתווך את ההתערבות באופן שיתקבל כלגיטימי בתוך עולם הערכים החרדי. ללא רגישות למורכבות זו, כל ניסיון לקדם מוגנות עלול להידחות על הסף, לא בשל היעדר חשיבות העיסוק בו אלא משום שעצם העיסוק מערער על גבולות.

2. **בין כבוד לשונות לבין עידוד הסתגרות:** בגיבוש מדיניות מוגנות לחברה החרדית, על מקבלי ההחלטות להכיר בדילמה יסודית – כיצד ניתן לגבש אמצעי הגנה והכלה טכנולוגית המכבדים שונות תרבותית, מבלי לחזק במקביל מגמות של הסתגרות, בידול או פיקוח חברתי הדוק? לא מדובר באתגר טכני, אלא בהתנגשות בין שני ערכים: הזכות של כל קבוצה לשמר את ייחודה ולחיות על פי אמונותיה, מול זכותו של הפרט, גם כשהוא חי בתוך קהילה סגורה, לגישה למידע, לאוטונומיה מחשבתית ולהגנה מפני מנגנוני שליטה. מקרה הבוחן של "ועדת הרבנים לענייני תקשורת" ממחיש היטב את המורכבות: גוף זה, שפעל לאורך שנים לאישור או חסימה של אמצעי תקשורת בציבור החרדי, ביקש להגן על הקהילה מפני תכנים פוגעניים אך בפועל יצר גם מנגנון ריכוזי של שליטה, הסתרה והדרה ממידע חיוני. יש אפוא להיזהר שמדיניות מוגנות לא תהפוך לכלי המשרת אינטרסים של הנהגות כוחניות בתוך הקהילה, מתוך פגיעה בזכויות יסוד של חבריה. האתגר אינו רק במידת ההתערבות של המדינה, אלא בזהות האובייקט שאליו מתייחסת ההתערבות. האם אנו מגינים על "החברה החרדית" כקולקטיב, או על "הפרט החרדי" כאזרח של מדינת ישראל שזכאי להגנה על פרטיותו, על עצמאותו, על גישתו למידע ולחינוך דיגיטלי? לעיתים קרובות, בשם כיבוד המסורת, נמנעים מקביעת גבולות בין תרבות לבין כפייה, בין הסברה לבין הסתרה. אולם אסור למדינה לאפשר מצב שבו כוחות שמרניים משתמשים בטיעונים של "רגישות תרבותית" כדי למנוע גישה למשאבים דיגיטליים, למידע רפואי או לחינוך טכנולוגי בסיסי, ובכך למנוע מהפרט אפשרות בחירה אמיתית. על כן יש להבחין בין מנגנוני מוגנוּת שנועדו לצמצם פגיעוּת טכנולוגית, לבין כאלה שמקבעים פערים, מגבילים ניידות מחשבתית או טכנולוגית ומנציחים את מעגלי ההדרה.

3. **הבחנות גילאיות, מגדריות וכאלה שנוגעות לתתי־קבוצות:** כפי שהודגם בחלקו הראשון של הפרק, החברה החרדית אינה מקשה אחת. פגיעויות טכנולוגיות משתנות בהתאם לגיל, מגדר, מיקום גאוגרפי והשתייכות לזרמים שונים (ליטאים, חסידים, ספרדים). מדיניות מוגנוּת אפקטיבית תידרש להתאים את עצמה להבדלים אלה ולהימנע מהכללות שעשויות להחטיא את מוקדי הפגיעות ואת הצרכים המדויקים של קהלים שונים בתוך המגזר.

4. **התאמה תרבותית כבסיס להתערבות אפקטיבית בחברה החרדית:** רגישות תרבותית־זהותית של תוכניות התערבות היא תנאי יסוד להצלחתן. גישה זו עולה בקנה אחד עם המלצות מחקריות לקידום אוריינות דיגיטלית מותאמת תרבות במגזרים שמרניים, מתוך שילוב בין תיווך קהילתי למנגנונים ממוסדים. בהקשר החרדי, הדבר מקבל משנה תוקף:

לא ניתן לקדם מוגנות באמצעים גנריים מבלי להבין את המנגנונים הסמויים והגלויים של סמכות, חינוך, פיקוח חברתי והפרדה מגדרית. מודל ההתערבות צריך להביא בחשבון לא רק את גבולות הגישה לאינטרנט, אלא גם את דפוסי ההפצה (קבוצות סגורות, חנויות פיזיות כתחליף לשירותים מקוונים), את תפקידי המנהיגות המקומית ואת רמות ההסכמה או ההתנגדות לשימוש בטכנולוגיה.

מכאן נגזרת הדרישה לעצב טכנולוגיות בינה מלאכותית באופן שישלב מלכתחילה התאמות לקהילות מובחנות, תפיסה המכונה **בינה מלאכותית מותאמת תרבותית** (Culturally Aligned AI).[437] יש לעודד פיתוח של מערכות בינה מלאכותית שמכירות בקודים התרבותיים החרדיים, גם בממשק (למשל, הימנעות מהצגת תכנים חזותיים אסורים), גם בתוכן (כגון שמירה על שפה נקייה או סינון מובנה), וגם בהקשרים מוסריים (כגון הכלה של עקרונות צניעות, היררכיה של ידע או נורמות של התייעצות רבנית). העיקרון של בינה מלאכותית מותאמת־תרבות מבקש להבטיח שמערכות בינה מלאכותית יפותחו, יעוצבו ויופעלו באופן שמכבד ומשקף את הערכים האתיים, הנורמות החברתיות והמבנים התרבותיים של הקהילות שיטמיעו אותן. מהותו של העיקרון היא לוודא שמערכות AI אינן כופות ערכים מערביים או ליברליים על מוצרים מבוססי בינה מלאכותית, אלא מפותחות, מתאמנות ונבנות מתוך הקשבה ורגישות למגוון אתני, דתי, היסטורי וחברתי. הדבר נכון במיוחד בחברה החרדית, שבה עצם ההתחככות עם ערכים זרים נתפסת כאיום קיומי על הסדר החברתי. כל ניסיון ליצור מסגרת אחידה של מכונה לומדת ויוצרת עבור "כל בני האדם" עלול לשעתק עליונות תרבותית, לבטל ייחוד תרבותי ולהיתפס כמנגנון של "קולוניאליזם דיגיטלי". לכן האתגר הוא לבנות מערכות בינה מלאכותית שיודעות לזהות, להבין ולכבד הבדלים, ולקדם דרכן הכללה תרבותית במקום הדרה טכנולוגית – ובפרט להבטיח שהמונח "התאמה תרבותית" לא ישמש כלי לניהול קהילתי סמוי של שליטה במידע.

הצורך בהתאמה תרבותית של מערכות בינה מלאכותית נובע מכמה שיקולים מהותיים, ובראשם מניעת הטיות אלגוריתמיות ואפליה. מערכות בינה מלאכותית לומדות מנתונים, והנתונים משקפים את ההטיות הקיימות בחברה שייצרה אותן. כאשר ערכים תרבותיים, למשל כללי לשון נקייה, כללי נגישות למגדרים שונים או היררכיות בקהילות, אינם מובאים בחשבון בשלבי איסוף הנתונים, עיצוב האלגוריתם או יישומו, נוצר סיכון ממשי שמערכת

437 ראו את הצעותיה של אמילי צ'פמן: Emily Chapman, Unveiling the Threat- AI and Deepfakes' Impact on Women, 3 (2024)

הבינה תנציח ואף תעצים דפוסי אפליה קיימים. כך למשל, מערכת ניתוח שיח אוטומטית שלא תוכננה להבחין בין שימוש חרדי בלשון גבוהה/למדנית לבין עברית מדוברת חילונית עלולה לשבש את הבנתה את הקלט שמגיע מקהילות חרדיות ולהוביל לשגיאות בתגובת המערכת.

לכן, רגישות תרבותית בתהליכי בניית מאגרי מידע והפעלת אלגוריתמים היא תנאי בסיסי למניעת פגיעה באוכלוסיות מיוחדות. שיקול נוסף הוא הבטחת הוגנות וצדק בין־תרבותי, משום שמה שנתפס כהוגן בתרבות אחת, עשוי להיחשב לבלתי מוסרי בתרבות אחרת. לדוגמה, במערכות גיוס מונחות־בינה, בין אם בשירות צבאי, בלשכה ממשלתית או אפילו בהתנדבות בארגוני רווחה, בתרבות אחת ייתכן שקשרים אישיים או המלצות נתפסים כדרך לגיטימית להתקבל לתפקיד, בעוד באחרת מודל של "עיוורון" לזהות המועמד נחשב לסטנדרט. מערכת שתתעדף אחד מהם על פני השני בלי רגישות תרבותית עלולה להיחוות כבלתי הוגנת, גם אם היא "ניטרלית" מבחינה טכנולוגית. שיקול שלישי הוא קידום אמון ציבורי בהטמעה של טכנולוגיות מבוססות בינה מלאכותית, עניין קריטי במיוחד בחברה החרדית. כאשר מערכת נתפסת כזרה לתרבות המקומית, מתנשאת או "מייבאת ערכים זרים", היא תיתקל בהתנגדות, גם אם תועלותיה הטכנולוגיות רבות. בחברה החרדית חוסר אמון כזה עלול להיתרגם לחרמות, פסיקות הלכתיות או הסרה מוחלטת של הטכנולוגיה מחיי היומיום.

הקושי הוא שכאשר מנסים להפוך את העיקרון של בינה מלאכותית מותאמת־תרבות למציאות מעשית, צפות ועולות שכבות של מורכבות. ראשית, הגדרה וזיהוי של "תרבות" הם חמקמקים ותלויי הקשר. הגדרה פשטנית מדי יכולה להוביל להכללות או לסטריאוטיפים, ובד בבד ליצור סגרגציית־יתר של קבוצות שאינן מעוניינות, מסיבות של כוח ושליטה, לחשוף את חבריהן למידע שכלל האוכלוסייה חשופה אליו. שנית, גם בתוך כל "תרבות" מוגדרת קיימת שונות רבה. גיל, מגדר, מעמד חברתי־כלכלי, השכלה ואמונות אישיות משפיעים על תפיסות אתיות, הרגלים ונורמות, ולכן גם על הציפיות מן המוצרים מבוססי הטכנולוגיה. דוגמה חרדית מובהקת היא ההבדל בין נשים חרדיות העובדות בסביבות טכנולוגיות לבין נערים בישיבות סגורות, וכל אחת מן הקבוצות האלה תדרוש מערכות מותאמות שונות בתכלית השינוי. שלישית, תרבויות אינן יציבות והן משתנות בעקבות שינויים פוליטיים ותהליכים חברתיים פנימיים. לכן מערכת שמאומנת כיום, עלולה להפוך עם הזמן ללא רלוונטית, או אף מזיקה, ככל שהתרבות משתנה.

כמה כיווני התמודדות עם המורכבויות שהוזכרו מופיעים בספרות העכשווית. למשל, **עיצוב מכליל ומשתף** (participatory design) של טכנולוגיה, ומעורבות ישירה של חברי קהילות שונות לאורך כל תהליך הפיתוח של מערכות מבוססות בינה מלאכותית. גישה נוספת היא **עיצוב רגיש־ערכים** (value sensitive design), המבקשת לזהות באופן פרואקטיבי מהם הערכים הרלוונטיים לקהילה מסוימת, לבדוק כיצד מערכות AI עלולות להשפיע עליהם ולעצב מערכות שתומכות בהם. כאן חשוב לפתח מתודולוגיות שאינן רק תאורטיות – אלא כאלו שניתנות ליישום גם כשאין דובר רשמי לקהילה. גישה שלישית היא "אתיקה אתנוגרפית" של בינה מלאכותית, כלומר שימוש בשיטות מחקר אתנוגרפיות כמו ראיונות עומק, תצפיות משתתפות וקבוצות מיקוד שיאפשרו הבנה עמוקה של ההקשר התרבותי שבו תפעל מערכת הבינה, במטרה לחשוף נורמות מוסריות, ערכים ואתוסים מקומיים. בהקשר החרדי, אתנוגרפיה עשויה לכלול מעקב אחר הפצת מידע דרך עיתונות מגזרית, פשקווילים או מערכות כשרות לתקשורת. ולבסוף, נדרשים מעקב והערכה מתמשכים באמצעות מדדים תרבותיים – שחלקם נוגעים להתאמה התרבותית בפועל, ואחרים לסיכונים לתוצאות חברתיות בלתי־מכוונות, כמו חיזוק מנגנוני שליטה, הדרה פנימית או פיקוח סמוי.

א. מחקר

בטרם ניגשים לתוכניות התערבות בעידן הבינה המלאכותית, אנו מציעים לבצע מחקרי בסיס לגבי החברה החרדית לצורך מיפוי "בורות טכנולוגית שקטה" בחברה זו. המונח מתייחס לאותם תחומים שבהם הציבור החרדי אינו מקבל מידע או שירותים דיגיטליים בסיסיים, לא בשל מחסור כלכלי או פערי אוריינות, אלא בגלל חסמים תרבותיים, מבניים או נורמטיביים. המחקר יתבסס על שיתופי פעולה דיסקרטיים עם ארגוני שטח, נותני שירותים, אנשי מקצוע חרדים וכלים אתנוגרפיים מותאמים. מטרת מחקרים אלה אינה חשיפה ביקורתית אלא תכנון מדינתי של מערכי תוכן חלופיים, מסגרות גישור מותאמות וקמפיינים ממוקדים במטרה לאפשר גישה למידע קריטי אגב שמירה על רגישות תרבותית וזהותית.

מחקרים אלה יעסקו באפקטיביות של תוכניות אוריינות קיימות בחברה החרדית (מיזם 100 באוריינות דיגיטלית, "חרדיגיטל", התוכנית הדיגיטלית הלאומית של ממשלת ישראל, תוכניות משרד החינוך וכו') על מנת לדעת מה היה יעיל עד כה ולגבי מה נדרשת חשיבה מחודשת; במיפוי של גורמי הסמכות בחברה החרדית שעשויים להשפיע כמתווכים יעילים לקידום מוגנות, הן בקידום אוריינות והן בקידום מנגנוני התמודדות עם פגיעות בפועל; במיפוי אזורי בורות טכנולוגית וחשיפה עדינה של מחסור במידע – כדי להתמודד עם

הדרה שיטתית של ידע חיוני, במיוחד בתחומים של זכויות פרט, בריאות הציבור, תעסוקה ומניעת הונאות. בהמשך למחקרים אלה יהיה צורך לבחון לאורך הדרך סוגי שימושים (למשל, באילו גילים משתמשים בצ'אטבוטים ולאילו מטרות); וסוגי פגיעויות של תתי־קבוצות בתוך החברה החרדית (למשל, הונאות רשת מסוג spear-fishing המיועדות לחרדים; ניסיונות ליצור הינדים התנהגותיים בתחום הרכש על אוכלוסיית הנשים), כפי שנעשו בעבר מחקרים על שימושי אינטרנט בקרב קבוצות אלה.

נוסף על כך נכון לבצע מחקרים העוסקים ביוצאים בשאלה. רבים מבני החברה החרדית העוזבים את הקהילה מייחסים זאת לחשיפה לאינטרנט/ידע חיצוני.[438] חלקם נושאים איתם זיכרונות ואף צלקות של פגיעות בשל החשיפה למוצרים טכנולוגיים, וניתן ללמוד מהם כיצד לאפיין אתגרי מוגנות.

ב. בניית תוכניות אוריינות מותאמות תרבותית

כפי שהדבר היה נדרש בעולם הדיגיטלי, יש צורך בפיתוח תוכניות לאוריינות בינה מלאכותית מיוחדות לחברה החרדית, אשר מותאמות לצרכיה, באופנים הבאים:

- **הבנת מושגי יסוד של מוגנות בעולם הבינה המלאכותית:** פרטיות והיסק, מניפולציה וזיהוי תכנים כוזבים, הבנת יישומי בינה מלאכותית ומגבלותיהם. זאת, מתוך מתן דוגמאות מחיי היומיום החרדי והסברים לגבי נזקים ומשמעויות.
- **התמקדות בקבוצות רלוונטיות:** אנו מציעים להתחיל בציבור של צעירים מעל גיל 18 ושל מבוגרים, ובעיקר נשים, משום שלקבוצות אלו יש נגישות למוצרים טכנולוגיים.
- **הכשרה למנהיגות רבנית**, ראשי ישיבות ומוסדות לנשים לגבי בינה מלאכותית, יכולותיה, מגבלותיה וסכנותיה היא תנאי מוקדם לאפשרות להכניס תוכניות אוריינות לציבור הרחב.
- איננו מציעים בשלב זה להכניס תוכניות אוריינות לבתי הספר, משום שלמרות שקיימים ניצנים של התעניינות בבינה מלאכותית בקרב הציבור החרדי וכניסה מדורגת שלה למערכת

438 איילה קיסר־שוגרמן, רון קנת, אמיר רם ואחרים **מרכזי קהילה והכוון ליוצאי המגזר החרדי מיסודה של עמותת יוצאים לשינוי** (מוסד שמואל נאמן, 2022).

החינוך,[439] כיום במרבית בתי הספר החרדים אין מחשבים,[440] וקשה להניח שלילדים יש גישה פתוחה למוצרים טכנולוגיים גם בבית. לעומת זאת, יש מקום לפיתוח **תוכניות אוריינות למורים חרדים** המשתמשים בבינה מלאכותית לצורכי הוראה.

- **קידום שיח** בנושאים שבהם יש חשש שלמידה באמצעות מקורות מידע לא אמינים (מרשתות חברתיות ומודלים של שפה ועד פורומים חסויים) יכולה לקדם פגיעויות – למשל, בנושאי נטייה מינית, פגיעות מיניות, יציאה מהדת וכיוצא באלה.[441]

- קידום מנגנוני הסברתיות (explainability) מתאימים לרמות האוריינות בחברה החרדית, בהיבטים של נגישות שפה, סגנון ומדיום (למשל מוקד טלפוני ולא רק טפסים דיגיטליים) התואמים את הפרקטיקות של המגזרי החרדי.

ג. תוכניות התערבות מערכתיות

הבסיס לתוכניות התערבות נמצא בהכשרת סוכני תיווך פנים־קהילתיים, אנשי חינוך, יועצים חינוכיים, מדריכים וגורמים חצי־רשמיים (כגון מוקדי ייעוץ טלפוניים) הן לתפקיד של זיהוי פגיעוּת והן במתן מענה ראשוני ובהכוונה להמשך טיפול.

- הקמת **"מרכז מוגנות חרדי"** כמוקד סייבר קהילתי וכתובת ברורה לחברי המגזר החרדי, שיתמקד בזיהוי, ליווי וייעוץ למשתמשים חרדים סביב איומי סייבר, הונאות סייבר, פגיעות בפרטיות, פגיעות אחרות או חשיפה לתכנים מזיקים. מרכז כזה יכול לפעול במסגרת מרכזים חרדיים קיימים כגון GuardYourEyes,[442] המסייע לנפגעי פורנוגרפיה ופועל גם במגזר החרדי, או ארגוני טיפול לנוער נושר,[443] להפעיל קו חם אנונימי לדיווח על פגיעות (למשל, דיווח אנונימי של נער שנחשף לפעילות עוינת ומבקש עזרה), ממשק שאלות הלכתי־

439 "כלים מבוססי בינה מלאכותית בהוראה החרדית" **הפורטל הפדגוגי החרדי, משרד החינוך**.

440 ארי ליבסקר "ליבי בליבם" **מוסף כלכליסט** (19.1.2023).

441 Chunpeng Zhai, Santoso Wibowo, & Lily D. Li, *The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review*, 11 Smart Learning Environments, 2-5 (2024)

442 ראו "אודות" Guard Your Eyes.

443 ראו למשל את מרכז הסייבר של חברת רפאל. אודי עציון "מגנים על יהודים בארץ ישראל: חרדים בתעשיית הסייבר" **ynet** (16.10.2020).

טכנולוגי אנונימי ותוכניות הסברה עבור הורים ומנהלים. צוות מומחי הלכה וטכנולוגיה יוכל לפרסם הנחיות מתעדכנות (למשל, ״התגלה וירוס חדש שמתחזה לאפליקציית תהילים – היזהרו״) ויסייע בתיאום עם הרשויות.

- **הגנה מפני הקהילה:** מוצע להקים גופים משפטיים, קליניקות קהילתיות או קווים חמים שיסייעו ליחידים מהחברה החרדית המתמודדים עם פגיעויות שמקורן פנימי, כגון שיימינג דיגיטלי בקבוצות סגורות, איסוף והפצת מידע אישי או רפואי ללא הסכמה, חסימת גישה לכלים דיגיטליים או מעקב בלתי פורמלי אחר פעולות מקוונות. גופים אלה יאפשרו לפרטים לפנות לעזרה באופן דיסקרטי ולקבל כלים להגן על זכויותיהם מתוך איזון עדין בין הגנה על ערכי הקהילה לבין עקרונות יסוד של חירות וכבוד האדם.

- **התנסות מבוקרת בכלים טכנולוגיים מוגנים:** מוסדות תעסוקה חרדיים, סמינרים לבנות ומרכזי הכשרה מקצועית יכולים לשלב שימוש בכלי בינה מלאכותית וליצור מעטפת מוגנת הכוללת הגדרות פרטיות מחמירות, חסימת תכנים פוגעניים והדרכה צמודה ממומחים בעלי רגישות תרבותית. תהליך כזה יאפשר היכרות בטוחה עם הכלים הדיגיטליים החדשים מבלי לפרוץ את גבולות הנורמות הקהילתיות, ובכך לתמוך בהשתלבות מבוקרת ובניית אמון בקרב המשתמשים החרדים.[444]

- **הכשרת אנשי טכנולוגיה חרדים בתחומים של שימושי הבינה המלאכותית:** בדומה למודל ״שומרי סף חרדים בהיי־טק״, ניתן להציע הכשרת מתכנתים ומפתחי מוצר חרדים שיפעלו כמתווכים בין עולם הפיתוח לבין הקהילה וישמרו על הלימה בין הערכים לבין תצורת השימוש.[445] בהתחשב בכך שהמגזר החרדי עצמו ישאף לפתח פתרונות סינון ושימוש נכון, יהיו דרושים לכך אנשים בעלי ידע טכנולוגי והכשרה בבינה מלאכותית. מסגרות ייעודיות מותאמות של קורסי הכשרה לאברכים לצורך פרנסת המשפחה יוכלו לשמש שכבה של שומרי סף מתוך הקהילה – חרדים בהיי־טק עם הכשרה בבינה מלאכותית,[446] שיוכלו לבצע בקרת אלגוריתמים ויעזרו בפיתוח כלים מותאמים לחברה החרדית. לשם כך יידרשו תמיכה ממשלתית ותמריצים לחברות להעסיק חרדים בתחומים אלו.

444 פלק, ברנדר ושרביט ה"ש 401 לעיל.

445 ראו למשל את המיזם לשילוב חרדים בהיי־טק הפועל בין היתר במימון הג'וינט: **אתר המיזם הלאומי לשילוב חרדים בהיי־טק**.

446 ראו למשל את **מיזם קמא־טק** שהוקם על ידי יזמים חרדים בשיתוף פעולה עם ראשי תעשיית ההיי־טק כדי לאפשר קידום חרדים בהיי־טק הישראלי.

- **מנגנוני פעולה ואמון בין־מגזרי להתמודדות עם משברים:** כפי שראינו בתקופת מגפת הקורונה, החברה החרדית, למרות רצונה להתבדל, אינה פועלת במרחב מבודד ומסתייעת רבות ברשויות המדינה.[447] לפיכך ישנה חשיבות יתרה לכך שבשעת משבר דיגיטלי (למשל, מתקפת השפעה משמעותית או מתקפת סייבר גדולה שפוגעת גם במערכות המשרתות את המגזר החרדי) יתקיימו ערוצי תקשורת אפקטיביים עם הגורמים המטפלים. בניית אמון בין מנהיגים חרדים למומחי טכנולוגיה ממשלתיים, למשל בתוך מערך הסייבר הלאומי, מערך הדיגיטל הלאומי ועוד, יכולה להועיל. שכן, ככל שיגבר האמון בין החברה החרדית למדינה, כך ניתן לשער שהחברה החרדית תהיה פתוחה יותר לקבל סיוע.[448] אמון כזה הוא חלק מן החוסן הקהילתי ויאפשר נכונות לקבל סיוע מבלי לחשוש לאובדן ערכי הקהילה. דבר זה מצריך בניית מנגנוני אמון דו־כיווניים לפני התרחשות של משבר.

ד. תכנון טכנולוגי

- **פיתוח פלטפורמות ידע ייעודיות לחברה החרדית** בצורת מערכות המלצה או בוטים מבוססי שפה עברית וערכים מותאמים, שיספקו מידע נגיש, אמין וקהילתי במגוון תחומים כמו בריאות, תעסוקה, הורות, השכלה ופנאי. פלטפורמות אלו צריכות להיות נקיות מפרסומות, להימנע מטִרגוט מסחרי ולשמור על סטנדרטים גבוהים של פרטיות, בדומה לתפיסה של אונסק"ו לקידום ריבונות דיגיטלית באמצעות תשתיות מותאמות תרבותית.[449] ברמה רחבה יותר ניתן לחשוב על מודל שפה גדול (LLM), שאומן רק על תוכן מאושר (ספרות תורנית, מידע מדעי בסיסי), שיוכל לשמש "עוזר לימודי" או לתת מענה לשאלות כלליות, בכללן שאלות הלכתיות.[450]

- **פיתוח מערכות סינון דור־חדש – לדוגמה, "מסנן צ'אט":** שכבה שיושבת בין המשתמש ל־ChatGPT ומנתחת את השאלה והתשובה, והטמעת אופציית **"מצב שמרני" בהגדרות**

447 אסף מלחי, גלעד מלאך, שוקי פרידמן **חרדים לקורונה: התמודדות המגזר החרדי עם מגפת הקורונה והמלצות למדיניות** (המכון הישראלי לדמוקרטיה, 26 במרץ 2020).

448 על היתרון של אמון ראו David Levi-Faur & Lauren Fahy, Report on Trust in Government, Politics, Policy and Regulatory Governance (European Commission, 2020)

449 Recommendation on the Ethics of Artificial Intelligence, art. 6, 23, 32 (UNESCO, 2021)

450 הגם שנדמה ששימוש בבינה מלאכותית נאסר על ידי רוב הרבנים במגזר החרדי בספרות השו"ת, ישנה פתיחות אליו בקרב רבנים אחרים. ראו למשל, מיכאל אברהם "בינה מלאכותית ופסקי הלכה – מחשבות ראשוניות" (25.4.2023); וראו גם Rav Ariel Di Porto, *Alcune riflessioni sull'impatto dell'intelligenza artificiale sul mondo ebraico*, 15 Segulat Israel, 43 (2024)

מוצרי ידע (כפי שישנו kids mode). במצב זה, ברירת המחדל של כל מיני הגדרות תהיה מחמירה, למשל, דפדפן יציג רק אתרים מאושרים, עוזר קולי לא יענה על שאלות בוטות. רוב הפלטפורמות הדיגיטליות הגדולות פועלות מתוך הנחות אוניברסליות של נגישות שווה, אך מודל זה יוצר הדרה של קבוצות שאינן מעוניינות להשתמש בטכנולוגיה באופן שאינו מסונן תרבותית. ניתן לתמרץ את הקמת המסגרות באמצעים רגולטוריים של הטלת חובה על ספקי מערכות AI להציע "מצבי שימוש מוגנים תרבותית", או באמצעות תקינה ממשלתית בדומה לדרישות נגישות לאנשים עם מוגבלות. ההבנה שביסוד רעיון זה היא שהגנה על קבוצות פגיעות אינה נובעת רק מחשש מפני נזק ישיר, אלא גם מהשאיפה להעניק להן יכולת פעולה עצמאית בשאלה כיצד ולמה להשתמש בבינה מלאכותית.

- **בניית מסגרות להתאמה של שירותים דיגיטליים ממשלתיים – "דיגיטלי־מינימלי":** מערכות בינה מלאכותית, למרות היותן יעילות יותר כאשר הן מחוברות לרשת האינטרנט, אינן חייבות להיות מחוברות אליה. לפיכך ניתן להעמיד לרשות החברה החרדית טכנולוגיות העושות שימוש במודלים מסוימים שאומנו מראש ואושרו על ידי רבני הקהילה. הדבר מתחבר לגישה רחבה יותר של התאמה דיגיטלית השואפת לכך שכל שירות דיגיטלי חיוני (ביטוח לאומי, בריאות, חינוך וכו') יינתן גם בערוץ מותאם שאינו דורש אינטרנט או הפעלתם של מתווכים דיגיטליים אחרים.[451] לצורך כך ניתן למפות את השירותים הדיגיטליים (חינוך, שירותים מקוונים, פרסום ויישומים צרכניים) הקיימים ולהכין תוכנית שלפיה כל פתרון בינה מלאכותית חדש (למשל, צ'אט מידע עירוני) יהיה נגיש גם באמצעי חלופי שהוא "דיגיטלי־מינימלי".

ה. רגולציה

- **הטמעת שיקולי מוגנות תרבותית בחקיקה מסדירה של בינה מלאכותית**, מתוך הכרה בכך שקבוצות שמרניות ומבודלות טכנולוגית כמו החברה החרדית הן "אוכלוסיות רגישות".[452] למשל, דרישה שפלטפורמות תוכן הפועלות בישראל יספקו כלים מתקדמים לסינון תכנים לפי ערכים (מעבר למגבלת הגיל).

451 ראו רועי גולדשמידט **הפער הדיגיטלי ויישום המדיניות הממשלתית לצמצומו** (מרכז המחקר והמידע של הכנסת, 2020).

452 הדין הישראלי אינו מגדיר מפורשות מיהן אוכלוסיות מוחלשות אך פסיקות שונות לאורך השנים הציבו את החרדים כקבוצה כזו בשל מאפייניה החברתיים־כלכליים של רוב הקבוצה (ראו למשל בג"ץ 8010/16 **מלכה נעמה ברזון נ' מדינת ישראל**; נבו, 12.7.2021).

- **הבטחת הוגנות אלגוריתמית:** מערכות מבוססות בינה מלאכותית, בין שמדובר בכלי סינון, מערכות קבלת החלטות או מערכות המלצה, עשויות ליצור אפליה או קטלוג שגוי. ההוגנות האלגוריתמית אינה רק עיקרון מופשט של שוויון, אלא עניין שלקבוצות מובחנות תרבותית הוא אינטרס קיומי, שכן מדובר במנגנון שמתווך עבורן גישה למשאבים, ייצוג חברתי והשתתפות אזרחית.[453] הוגנות אלגוריתמית בהקשר החרדי דורשת אפוא רגולציה דיפרנציאלית לא רק בהבטחת אי־אפליה כללית, אלא גם בהתאמה פרשנית מגזרית. יש לעגן באופן מפורש ברגולציה חובות על מפתחי מערכות בינה מלאכותית (במיוחד בתחומי תעסוקה ואבחון תעסוקתי, אשראי וניתוח נתונים פיננסיים, ביטוח ובריאות, חינוך מותאם אישית), וכן חובות על גופים פרטיים וציבוריים הרוכשים מערכות כאלה, לבצע בדיקות הטיה לגבי מגזרים דתיים/תרבותיים.[454] למשל, בטרם בנק מאמץ מודל אשראי, עליו לבדוק אם תוצאותיו עבור חרדים שונות באופן לא מוסבר מהשאר, ואם כן, לכוונן אחרת את המודל או להרחיב את בסיס נתוני האימון. במקביל לחובות ההוגנות יש להטיל חובות שקיפות, הנמקה וערעור על החלטות אלגוריתמיות שיש בהן כדי להשפיע על זכויות רכושיות ואזרחיות.[455]

- **הכללת נציגים של קבוצות תרבותיות מובחנות** כמו החברה החרדית בשלבי האפיון והפיתוח של מערכות בינה מלאכותית המיועדות למגזר הציבורי, ויצירת מסגרת של הערכת השפעה תרבותית (cultural impact assessment) דומה לזו הנהוגה בהקשרים של השפעה מגדרית או סביבתית.[456] נכון למועד כתיבת שורות אלה, הוועדה המייעצת של משרד המדע והטכנולוגיה בעניין בינה מלאכותית אינה כוללת חרדי, וקולה של החברה החרדית לרוב חסר בדיונים אלה.[457]

453 Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo et al., *The Ethics of Algorithms: Mapping the Debate*, 13(2) Big Data & Society, 1–21 (2016)

454 נכון למועד כתיבת שורות אלה אין בישראל חקיקה בנושא בינה מלאכותית אלא רק מסמכי מדיניות. לסקירה מקיפה יותר של הסוגיה ראו כהנא ושוורץ אלטשולר, **אדם ומכונה**, לעיל ה"ש 129.

455 Eldar Haber, *Digital Transparency*, 98 N.C. L. Rev. 395, 400–401 (2020)

456 Helen Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life, 231–247 (Stanford University Press, 2009)

457 "המומחים שיצעידו את ישראל לעתיד: הוקם פורום המומחים להיערכות לאומית לבינה מלאכותית" **משרד החדשנות, המדע והטכנולוגיה** (6 בפברואר 2025). גילוי נאות, כותבת מסמך זה היא חברה בפורום המומחים.

ו. אכיפה

- **הנגשת מנגנוני דיווח אנונימיים ומותאמים תרבותית:** שירותים מדינתיים לפניות נפגעים ונפגעות (במיוחד מינית או כלכלית) אינם נגישים כיום לחברה החרדית, מבחינה לשונית, תרבותית ופסיכולוגית.[458] לפיכך נדרשת הקמה של מערך דיווח, גישור וסיוע שמכבד נורמות תרבותיות, פועל בחשאיות מלאה ואשר ירכוש את אמון בני החברה החרדית ויאפשר לדווח בזמן אמת על שימושים לרעה בבינה מלאכותית. חלק מרכישת האמון צריך להיות במערך ניטור שיגרום לכך שאכן דיווחים מקבלים מענה ובמקרה הצורך מעורבים גורמי רווחה, חקירת ילדים ובתי משפט.
- **פיקוח על תכנים הנוגעים לקבוצות מיעוט:** כשם שעולות הצעות לסימון או לאיסור תכנים יצירי מכונה שתכליתם פגיעה מגדרית (למשל פורנו־נקמה או הקטנת נשים), ניתן לחשוב על הצעות דומות כשתכלית הפגיעה היא דתית. למשל, אם אדם מפיץ סרטון יציר מכונה בכוונה לבזות דמות מזוהה של רב, ניתן יהיה לתבוע אותו בעוולה מיוחדת נוסף לעוולת לשון הרע.[459]

ז. כיווני התערבות נוספים

נוסף על צעדי המדיניות המערכתיים והקהילתיים שתוארו לעיל, מוצעים כאן שלושה כיווני פעולה משלימים, השואפים לפרוץ את גבולות החשיבה השמרנית על מוגנות בחברה החרדית. הם נולדו מתוך ההכרה במתח הקיים בין הרצון להגן על הקבוצה לבין החובה להגן על הפרט, ומתוך ההבנה שמוגנות אינה מסתכמת רק בהגנה, אלא גם ביצירת גישה, העצמה ובחירה חופשית.

- **יצירת פורומים ומרחבים דיסקרטיים לצעירים וצעירות חרדים:** יש להקים מסגרות חינוך אלטרנטיביות ודיסקרטיות לצעירים וצעירות מהחברה החרדית המבקשים להבין את עולם הבינה המלאכותית אך חוששים מההשלכות של חשיפה פומבית או מהתנגדות קהילתית. בדומה לקווי סיוע אנונימיים או מסגרות חוץ־קהילתיות דיסקרטיות בתחומי

458 רחל ניסנהולץ־גנות שרביט פיאלקו תמי ברוש וברוך רוזן **פגיעות מיניות ביחסי מטפל־מטופל במערכת הבריאות בישראל: ממדי התופעה, מאפייניה והשפעותיה** (מכון מאיירס ג'וינט ברוקדייל, 2024). ראו בעמ' 10 **על הייצוג החסר של החברה החרדית.**

459 לפתרון דומה בהקשר של נשים ראו Chapman, לעיל ה"ש 437.

רווחה וחינוך, ניתן להציע "ישיבות צללים", קבוצות למידה אינטימיות או פורומים מקוונים בטוחים, שיאפשרו היכרות עם שימושים בסיסיים, מושגי יסוד ודילמות מוסריות, מתוך כבוד לספקות ולמורכבויות של הלומדים.

- **מענקי חדשנות:** במקום להסתפק בהתאמת AI לחברה החרדית מבחוץ, ניתן לקדם את כניסתה של החברה החרדית למעגלי הפיתוח והיצירה עצמם. לשם כך מוצע להקים קרנות ממשלתיות, פילנתרופיות או קהילתיות שיתמכו ביזמות חרדית בתחום הבינה המלאכותית, בכלל זה פיתוח אפליקציות, בוטים, ממשקים ומודלים מוסריים התואמים את הקודים ההלכתיים והחברתיים של הקהילות החרדיות. דוגמאות אפשריות הן למשל: מערכות הלכתיות מבוססות בינה (halachic chatbots) ומערכות שו"ת אוטומטיות, או אפליקציות שידוכים מבוססות ערכים.

- **"רגולציה הפוכה":** חובת נימוק מוסדי למדיניות חסימה. במקום לבסס את כל עקרונות ההתאמה על התחשבות חד־כיוונית של המדינה בצורכי הקהילה, ניתן לחשוב על מודל שיטיל על מוסדות קהילתיים, חינוכיים, תעסוקתיים ושירותים חובת הנמקה, שקיפות ואחריות כלפי מדיניות ההדרה הטכנולוגית שהם מיישמים.[460] מוסדות שמונעים גישה לטכנולוגיה או מגבילים שימוש בה בקרב קהלים שבתחום אחריותם (למשל צעירים, תלמידות סמינר, אימהות חד־הוריות, או עובדים גברים) יחויבו לנסח ולהגיש מדיניות ברורה, שתפרט את היסודות הערכיים והאתיים למדיניות ההגבלה, וכן את דרכי ההשגה והערעור על החלטות אלה ברמת הפרט. המהלך אינו מבקש לכפות נורמות מבחוץ אלא ליצור איזון מחודש בין סמכות קהילתית לבין זכויות הפרט ולמנוע שימוש לא שקוף במנגנוני שליטה במסווה של מוגנות.

460 ראו למשל את הצעתו של שוקי פרידמן – שוקי פרידמן **שוק הסלולר הכשר: מחור שחור לאסדרה מאוזנת** (הצעה לסדר 34, המכון הישראלי לדמוקרטיה, 2020). פרידמן טוען כי "הוועדה [ועדת הרבנים לענייני תקשורת האמונה על אסדרת השימוש בטלפונים כשרים] כמובן אינה טורחת לממש את חובת השימוע וחובת ההנמקה לפני חסימת הקווים" (בעמ' 46).

סיכום

הפרק עוסק באחת הדילמות המאתגרות ביותר של עידן הבינה המלאכותית: כיצד להגן על קהילות שמבקשות להגן על עצמן. החברה החרדית בישראל מציבה אתגר ייחודי למושג מוגנוּת, בהיותה קהילה סגורה המוּנעת מחשש עמוק מהשפעות טכנולוגיות, אך בעת ובעונה אחת חשופה לאותן פגיעויות חדשות שמעוררות טכנולוגיות AI בכל מגזר אחר: שיימינג דיגיטלי, הטיות אלגוריתמיות, הדרה ממידע, פערי אוריינות וסיכונים לילדים. מכאן עולה הצורך בגישה רגולטורית מותאמת תרבותית, לא כוויתור על עקרונות אוניברסליים של זכויות פרט, אלא כדרך להגשימם באופן שמכבד את המרחב התרבותי שבו הם מוטמעים. גישה זו דורשת רגישות מתמדת להבדלים בין תת־קבוצות, גילים ומגדרים, אך גם עמידה על עקרונות מוסריים ברורים: הגנה על ילדים, שמירה על פרטיות וחופש גישה למידע גם מול לחצים קהילתיים.

הפרק מצביע על שלוש סכנות עיקריות במערכי מוגנות לאוכלוסייה החרדית: הכחשת הצורך ("אם זה אסור, ממילא אין מה להתערב"); הפיכת ההגנה למנגנון של שליטה פנימית; והיעדר הבחנה בין תמיכה בקהילה לבין כניעה לערכים שמונעים אוטונומיה מהפרט. המלצות הפרק נעות בין פעולה "מבפנים" לבין רגולציה "מבחוץ", ומשקפות גישה מורכבת למוגנות כפרקטיקה תרבותית. לצד רעיונות שמרניים יותר, כמו התאמות רגולטוריות, הסברה קהילתית או תיווך מוסדי, מוצעים גם מהלכים לא שגרתיים: מסגרות דיסקרטיות להעמקת הידע הטכנולוגי (כדוגמת "ישיבות צללים"), מנגנוני הגנה על פרטיות שמגינים על הפרט דווקא מפני מוסדות הקהילה, הקצאת משאבים לפיתוח מערכות AI חרדיות, והטמעת מרחבים פיזיים בטוחים לגישה למידע פתוח ומתווך. דרך מגוון ההמלצות נפרשת תפיסה של מוגנות כזירה מוסרית, פוליטית וקהילתית – לא רק אוסף אמצעים טכנולוגיים או משפטיים, אלא מנגנון מתמשך של תרגום ערכים אוניברסליים להקשרים מקומיים.

דרכי ההתערבות שהוצגו בפרק זה ניתנות לארגון בשלושה אשכולות עיקריים, המשקפים טווחי זמן שונים ורמות שונות של ישימות והשפעה. האשכולות אינם חלופות – הם שכבות משלימות: צעדי בלימה והנגשה מיידיים, בניית תשתיות תיווך ואמון ומהלכים מבניים ארוכי טווח המעצבים מחדש את יחסי הכוח בין מוסדות, קהילה ופרט.

פרק עשירי

מנגנוני התערבות לקידום מוגנות של הציבור הערבי בעידן הבינה המלאכותית

בפרק זה נפתח בזיהוי המאפיינים הייחודיים של החברה הערבית בישראל הרלוונטיים להקשר של מוגנות בעידן הבינה המלאכותית, כגון מאפיינים חברתיים־כלכליים, משבר אמון מול מוסדות המדינה, נגישות ושימושים בטכנולוגיה ועוד. אל מול אלה ואל מול מאפייני מערכות בינה מלאכותית, נמפה את הפגיעויות של בני החברה הערבית. לא נכלול בפרק זה דוגמאות למיזמי מוגנות בתחום הבינה המלאכותית ביחס לערבים בישראל משום שלא מצאנו כאלה, אבל נציע דרכי התערבות שאנו תופסים כרלוונטיות במיוחד לחברה הערבית בישראל, וכן דוגמאות למיזמים הנוגעים למוגנות של מיעוטים במקומות אחרים בעולם.

חלק ראשון: מאפיינים חברתיים ותרבותיים של החברה הערבית המשפיעים על מוגנות

א. חברה היררכית, שמרנית, בעלת זהות קהילתית חזקה

החברה הערבית בישראל מתאפיינת במבנה חברתי היררכי הנשען על דמויות סמכות, שבראשן אב המשפחה, אנשי דת ומורים.[461] מערכות היחסים הללו מכתיבות דפוסי פעולה, אמונות וגבולות גם במרחב הדיגיטלי. גם הזהות הקהילתית נחשבת לאבן יסוד במבנה החברתי והתרבותי. תחושת המחויבות לקולקטיב תורמת ללכידות החברתית ולמערכת תמיכה יציבה, אך במקביל עלולה להוביל לשלוש תופעות: ראשית, למערכת פיקוח הדוקה; שנית, לשעתוק של נורמות קהילתיות פסולות כמו נקמה וסנקציות חברתיות אחרות אל המרחב הדיגיטלי; ושלישית, לנורמות חברתיות המדגישות שמרנות ובושה ובתורן מובילות לדיכוי היכולת לדווח על פגיעות ואפילו לבקש עזרה וסיוע.[462]

כך, נערה שחשבון האינסטגרם שלה נפרץ או שפורסם עליה מידע אישי, לא תחשוב בהכרח על פנייה לרשויות או על שמירה על פרטיותה, אלא על השאלה אם סביבתה הקרובה תראה בכך מצב בעייתי. פחד זה מביא במקרים רבים למחיקה עצמית: העלמת פרופילים, מעבר לזהות אנונימית או ניתוק מוחלט מהמרחב הדיגיטלי. באותה נשימה, היכולת לדווח או לתבוע צדק נמחקת, לא בגלל היעדר מודעות לזכויות, אלא בגלל חשש חברתי תרבותי שהתגובה תעורר.

שימוש בטכנולוגיה אינו רק עניין אינדיווידואלי, אלא מיומנות שמתעצבת בתוך מערך ציפיות ותגובות קהילתיות. כאשר צעיר או צעירה נתקלים בהתלבטות דיגיטלית, כמו האם לפתוח חשבון טיקטוק או לדווח על סחיטה מקוונת, שיקול הדעת שלהם מתווך דרך מפת נורמות רחבה. כך, חשש מפני אכזבה או תגובה של מבוגר משמעותי עלול להוביל להסתרה, השתקה או השלמה עם פגיעות. לעיתים עצם הפנייה לסיוע חיצוני, מוקד ממשלתי, ארגון

461 ראו למשל, ראסם ח'מאיסי **תכנון ופיתוח היישובים הערביים בישראל: תפיסה חדשה להיערכות הרשויות המקומיות והמדינה** (מרכז תמי שטינמץ למחקרי שלום, המכון למחקרי ביטחון לאומי, המרכז היהודי־ערבי, המכון הישראלי לדמוקרטיה, 2015).

462 **נשים ערביות בשוק העבודה: אתגרים והזדמנויות** (יוזמות אברהם, 2022).

סיוע או פלטפורמה עשויה להיחשב לערעור על הסדר ולהיחסם על ידי מנגנונים של משמעת משפחתית וקהילתית.

ב. משבר האמון הבסיסי מול מוסדות המדינה

האוכלוסייה הערבית סובלת מתחושת נחיתות מתמשכת ביחס לאוכלוסייה הכללית,[463] ובעיקר קיים משבר אמון מתמשך בינה לבין הרשויות, למשל מול משטרת ישראל, הנתפסת כמערכת עוינת.[464] החברה הערבית בישראל נושאת עימה מטען היסטורי של הדרה, אפליה והזנחה מצד המדינה. התחושות הללו אינן מופשטות, אלא מעוגנות במציאות של גזענות גלויה, תת־תקצוב, אלימות שאינה זוכה למענה מצד רשויות אכיפת החוק, תשתיות ציבוריות רעועות וחוסר ייצוג. הפשיעה האלימה כלפי החברה הערבית ובתוך החברה הערבית עצמה הגיעה לממדי משבר, אך המענים של רשויות האכיפה הם חלקיים בלבד.[465]

מן הראיונות שערכנו עולה גם כי תופעות אלה מתורגמות למרחב הדיגיטלי: כאשר שירות חדש מוצע באתר ממשלתי, או כאשר מערכת ממוחשבת דורשת "הזדהות ממשלתית", הם אינם נתפסים ככלי שירות ניטרלי אלא כהמשכו של מנגנון שליטה. כך, אפילו שירותים חיוניים כמו ביטוח לאומי, בריאות, חינוך או דיווח על אלימות, נתקלים בחשד, הסתייגות ולעיתים בהימנעות. גם כאשר שירות הוא זמין, נגיש או אפילו נוח, עצם השתייכותו לממסד מעוררת רתיעה. כך למשל, בני נוער לא מדווחים על פגיעות מיניות ברשת דרך האתרים הייעודיים, בשל תרבות של אי־דיווח למשטרה והתנגדות למגע עם רשויות אכיפת החוק; נשים לא פונות לקבלת עזרה משפטית באתרים ממשלתיים; וצעירים מדירים עצמם מתהליכי השכלה והשמה דיגיטליים גם כאשר אלו נראים מותאמים. במקרים מסוימים

463 מוחמד ח'ילאילה, אחמד בדראן ואריק רודניצקי **שנתון החברה הערבית בישראל 2023** פרק א: דמוגרפיה (המכון הישראלי לדמוקרטיה, 2023); נטע ברק־קורן, יובל פלדמן ונעם גדרון **כאשר חוק מעורר תגובה נגדית: על חוק־יסוד הלאום ועל היחס למיעוטים בישראל** (המכון הישראלי לדמוקרטיה, 2020).

464 Habeeb Makhoul, Digital Divide: Discrimination in Digital Infrastructure Against Palestinian Citizens in Israel (7amleh – the Arab Center for the Advancement of Social Media, 2025)

465 תומר לוטן "הפשע בחברה הערבית לא 'מתפרץ', הוא מתבסס" **אתר המכון הישראלי לדמוקרטיה** (4.6.2025).

ארגוני חברה אזרחית משמשים מתווכים, אך אלו אינם תמיד נגישים, אינם פועלים בכל מקום ולעיתים גם נתפסים כמזוהים עם המדינה.

בנוסף, מן הראיונות עלה כי התפיסה השלטת היא שהחברה הערבית אינה רק "משתמשת" בטכנולוגיה אלא בעיקר מושא שלה: מדווחים עליה, עוקבים אחריה, ממפים אותה, אך לא מתוך שיח שוויוני או עיצוב משותף. פער זה אינו טכנולוגי אלא פוליטי, והוא מזכיר ספרות בעולם העוסקת בכך שקבוצות מיעוט אתניות, תרבותיות או לשוניות נוטות להיות חשופות ופגיעות יותר מאחרות למידע כוזב ולסכנותיו[466] ובמקביל, הן מהוות מטרה למאמצי דיסאינפורמציה.[467]

מעמדם הייחודי של אנשי החברה הערבית בישראל מציב אותם במקום של פגיעות הנובעת גם ממוטיבציה פוליטית של גורמים חיצוניים, מקומיים, אזוריים ובינלאומיים. מצב זה מגביר את הסיכון לחשיפה יזומה המכוונת לדיסאינפורמציה, להסתה או לתוכן המבקש לערער עוד את האמון במוסדות המדינה, להקצין עמדות ולשבש תהליכי השתלבות אזרחית.

ג. הבחנה מגדרית בגישה לידע, עבודה וטכנולוגיה

בחברה הערבית בישראל מתקיים שילוב מורכב בין מסורת לבין מודרניזציה המשפיע באופן ישיר על מעמדן ותפקידן של נשים. מצד אחד, ערכים פטריארכליים ממשיכים לשלוט בחלקים גדולים של החברה, והם מובילים לחלוקה מגדרית נוקשה של תפקידים – הגבר נתפס כראש המשפחה והמפרנס, ואילו האישה כאחראית על תחום הבית והמשפחה.[468] תפיסה זו מושרשת גם בנורמות דתיות וקהילתיות שמקשות על יציאת נשים לשוק העבודה. עם זאת, בשנים האחרונות ניכרת מגמה של שינוי הדרגתי, במיוחד בקרב הדור הצעיר ובקרב

466 Angela Lee, Ryan Moore, & Jeffery Hancock, *Designing Misinformation Interventions for All: Perspectives from Aapi, Black, Latino, and Native American Community Leaders on Misinformation Educational Efforts*, 4(1) Harvard Kennedy School Misinformation Review, 1–16 (2023)

467 Melo-Pfeifer & Gertz, לעיל ה"ש 395.

468 ג'ירי אלמו־קפיטל נתונים על תעסוקת נשים ערביות בדגש על בדואיות בנגב (מרכז המחקר והמידע של הכנסת, 2022); Maha Sabbah-Karkabi, *The Diverging Gender Inequality Across Households: The Case of Palestinian-Arab Families in Israel*, 72 Current Socio., 519 (2024)

נשים משכילות שדורשות נוכחות ושוויון רבים יותר בתחומים שונים של החיים.[469] מערכת ההשכלה הגבוהה בישראל הפכה לאפיק מרכזי לשינוי עבור נשים ערביות. מספר הנשים הערביות באקדמיה גדל באופן משמעותי בעשורים האחרונים, והן מהוות כיום רוב בקרב הסטודנטים הערבים.[470] עם זאת, השכלה גבוהה אינה מבטיחה באופן אוטומטי השתלבות בשוק העבודה: שיעור ההשתתפות של נשים ערביות בכוח העבודה נותר נמוך בהשוואה לנשים יהודיות, בין היתר בשל חסמים מבניים, פערים תחבורתיים, היעדר מסגרות טיפול בילדים והעדפות תרבותיות ומשפחתיות.[471]

בתחום הפוליטי והציבורי, ייצוג נשים ערביות עדיין מצומצם מאוד, הן ברמה הארצית והן ברמה המקומית. למרות קיומן של פעילות בולטות במאבקים אזרחיים ובחברה האזרחית, חדירתן לזירה הפוליטית הרשמית היא איטית. גורמים תרבותיים, כגון התפיסה שפוליטיקה איננה תחום "נשי", לצד חסמים מוסדיים וחברתיים, מצמצמים את האפשרות של נשים ערביות להתמודד לתפקידים ציבוריים. במקביל, ישנן יוזמות שמטרתן להכשיר וללוות נשים ערביות לפעולה פוליטית המעידות על מגמת שינוי. לצד כל אלה, יש להדגיש את ההבדלים התוך־קהילתיים: מעמד הנשים משתנה באופן משמעותי בין קבוצות שונות בתוך החברה הערבית, בין דרוזים, מוסלמים ונוצרים, בין כפריים לעירוניים, ובין מרכז לפריפריה. ישנם מקומות שבהם ניכרת פתיחות רבה יותר לשינויים במעמד האישה, בעוד באזורים אחרים ניכרת שמרנות גבוהה יותר.

השיח על נשים וטכנולוגיה בחברה הערבית בישראל אינו יכול להתבצע במנותק משיח רחב יותר על מעמד האישה ושליטה על משאבים. לנשים, שמראש נמצאות במעמד מוחלש כלכלית, אין הכלים הטכנולוגיים שיוכלו לשמש להן מנוף להשתלבות. יתרה מזו, הן פגיעות בהקשר הקהילתי־משפחתי. כאשר נוצרת פגיעה, למשל סקסטורשן, סחיטה רגשית או חדירה לחשבון אישי, הן עלולות לשאת בתוצאה החברתית והמשפחתית החמורה, גם

469 עולא אבו־חסן נבואני "עבודה, משפחה ומה שביניהן: מבט על נשים מוסלמיות, דרוזיות ונוצריות בישראל", **ישראל** 33, 97 (2024); נסרין חדאד חאג' יחיא, איימן סייף, ניצה (קלינר) קסיר ובן פרג'ון **חינוך והשכלה בחברה הערבית: פערים וניצנים של שינוי**, מחקר מדיניות 159, (המכון הישראלי לדמוקרטיה 2021); ערן ישיב וניצה (קלינר) קסיר "נשים ערביות בשוק העבודה בישראל: מאפיינים וצעדי מדיניות" **הרבעון לכלכלה** (2012).

470 חילאילה, בדראן ורודניצקי, לעיל ה"ש 463.

471 שם.

אם הן עצמן היו הקורבן של הפגיעה. התלות בגברים, בני משפחה, מתווכים טכנולוגיים או בעלי סמכות מייצרת מצב שבו הפגיעה עצמה אינה מדוברת ולעיתים גם לא מזוהה. כאן נולדת פגיעוּת טכנולוגית מגדרית: נשים מודרות מהטכנולוגיה, מהשיח עליה, וגם ממנגנוני ההגנה גם אם הם מוּבָנים לתוכה.

ד. פערים חברתיים

הפערים הפנימיים בחברה הערבית בישראל ניכרים בעיקר בין קבוצות דתיות ובין אזורים גאוגרפיים. 28.8% מהנוצרים הערבים הם בעלי השכלה אקדמית, לעומת 14.7% מהמוסלמים, כפי שמצוין בדוח הלמ"ס.[472] כמו כן, קיימים פערים משמעותיים בשיעורי התעסוקה בין הקבוצות הדתיות. שיעור התעסוקה הגבוה ביותר נמצא אצל הנוצרים, והשיעור הנמוך ביותר אצל המוסלמים. הפערים הללו משקפים השפעות היסטוריות, תרבותיות והשכלתיות שונות בין הקהילות הערביות במדינת ישראל. פערים אלה קיימים גם בנוגע לשליטה בדיבור ובקריאה בעברית,[473] והם מצביעים על חלוקה לא שווה בגישה לידע שפתי, שהיא חסם נוסף לשוויון הזדמנויות בתוך החברה הערבית עצמה.

בנוסף, הפערים הבין־דוריים בחברה הערבית בישראל נוכחים גם במרחב הדיגיטלי, ולעיתים אף מועצמים בו. בעוד צעירים רבים משתמשים בטכנולוגיות מתקדמות, בני הדור המבוגר נותרים מנותקים מהשיח הדיגיטלי.[474] באופן לא מפתיע, הפער באוריינות הדיגיטלית הבין־דורית גדול יותר מאשר בחברה היהודית החילונית.[475]

ה. פערים חברתיים־כלכליים ותשתיות טכנולוגיה

המצב החברתי־כלכלי של בני החברה הערבית בישראל נמוך באופן ניכר מהממוצע בישראל, מתאפיין בשיעורי עוני גבוהים, אבטלה כרונית והשתלבות חלקית בלבד בשוק

472 "פערים בין יהודים לערבים 2020–2021 – נתונים מתוך דוח פני החברה מס' 14" 8–18 (הודעה לעיתונות, הלשכה המרכזית לסטטיסטיקה, 2023).

473 אסמאא נאדר גנאיים, שיזף רפאלי ופיסל עזאיזה "פער דיגיטלי: השימוש באינטרנט בחברה הערבית בישראל" **מגמות: רבעון למדעי ההתנהגות** 46 (1–2), 164–196 (2009).

474 מריאן תחאוכו, הלה אקסלרד וחנין מטר **האתגר הדיגיטלי בחברה הערבית** 20 (מכון אהרן למדיניות כלכלית, אוניברסיטת רייכמן, 2021).

475 יונתן מנדלס **מחוברים אבל (לא) שווים: פערים דיגיטליים, תשתיות, שימושים ומוגנות ברשת בחברה הערבית בישראל** 26 (איגוד האינטרנט הישראלי, מאי 2024) (להלן: דוח מנדלס).

העבודה. שיעורי העוני במגזר הערבי גבוהים פי שניים ויותר מבמגזר היהודי, והכנסת משק בית ערבי ממוצעת נמוכה משמעותית משל משק בית יהודי.[476] נתונים אלה לא רק מציבים את הקהילה בעמדת נחיתות כלכלית, אלא גם משפיעים על איכות הגישה לטכנולוגיה, עומק השימוש בה והיכולת למצות את הפוטנציאל שבשימוש.

פערי השכלה ותעסוקה מובילים לכך שהייצוג של ערבים במשרות ההיי־טק והמגזר הפיננסי נמוך,[477] והם נהנים פחות מהצמיחה הכלכלית בענפים אלו. כל אלה מגבירים את התלות בעבודות כפיים ולחשיפה גבוהה יותר לאובדן עבודה בשל תהליכי אוטומציה.[478] נעיר כי אכן גם בהיבט זה קיימת שונות בין קבוצות אוכלוסייה ואזורים גאוגרפיים שונים.

הנתונים בנוגע לתשתיות ולנגישות לטכנולוגיה משקפים פער דיגיטלי מתמשך: במרבית היישובים הערביים תשתית האינטרנט מפותחת פחות מביישובים יהודיים, עם מהירויות חיבור נמוכות, ולעיתים היעדר חיבור פס רחב בכלל, והיעדר גישה למוקדי תמיכה.[479] גם בזמינות ציוד הקצה ניכרים פערים. ראשית, לפי נתוני איגוד האינטרנט הישראלי 63% מהמשיבים הערבים מחוברים לאינטרנט בחיבור קווי, לעומת 44% בחברה היהודית; ורק 7% מהערבים מחוברים בסיב אופטי המאפשר גלישה מהירה ויציבה, לעומת 41% בחברה היהודית.[480] התוצאה היא שיתוף קבצים איטי, סטרימינג מקוטע ושימוש מוגבל מאוד בפלטפורמות מתקדמות. שנית, כרבע מהמשיבים הערבים תלויים בחיבור סלולרי בלבד, שהוא איטי ומוגבל יותר, ו־93% מהמשיבים הערבים מחזיקים בסמארטפון, בהשוואה ל־95% בחברה היהודית.[481] כמו כן, 71% מהמשיבים הערבים מחזיקים במחשב נייד, ורק 28% במחשב נייח, לעומת 83% ו־61% בהתאמה בחברה היהודית.[482] הפער בולט במיוחד בנגב, שבו רק 2% מהמשיבים הערבים מחוברים בסיב אופטי ו־42% בחיבור סלולרי.

476 Makhoul, לעיל ה"ש 464.

477 "ערבים בהיי־טק – יש עוד דרך ארוכה" **אתר רשות החדשנות** (2023).

478 מריאן תחאוכו, עמית לוונטל, טלי לרום ואיילה פרטוש **אתגרי החברה הערבית בשוק העבודה – בשגרה ובחירום** (מכון אהרן למדיניות כלכלית, אוניברסיטת רייכמן, והג'וינט, 2024).

479 ראו דוח מנדלס, לעיל ה"ש 475.

480 שם, בעמ' 22–24.

481 שם, בעמ' 23–24.

482 שם.

ההבדלים האלה יוצרים פער דיגיטלי מהותי המשפיע על יכולת השימוש המיטבי באינטרנט ועל הגישה לשירותים דיגיטליים מתקדמים. כשאין תשתית, אין גם חוויית משתמש ראויה, וכשאין חוויה, אין רציפות דיגיטלית. בעידן שבו כמעט כל שירות – רפואי, תעסוקתי או חינוכי – עובר לדיגיטל, היעדר התשתית איננו רק בעיה טכנית אלא איום על השוויון האזרחי.

במקביל, הדרה תחבורתית וחברתית מהמרכז מצמצמת את החשיפה לשירותים חדשניים, ליוזמות חינוכיות או להסברה מותאמת שפה. כך, הפער הדיגיטלי הוא לא רק פונקציה של שימוש במכשיר או ביישום טכנולוגי, אלא של ריחוק מרחבי שמונע קִרבה לאקוסיסטם טכנולוגי בכלל ולא רק לשירות מסוים.

שיעור השימוש של בני החברה הערבית בשירותים בנקאיים ופיננסיים מקוונים נמוך יחסית. רבים מהעסקים הקטנים בחברה הערבית אינם מוטמעים במלואם בכלכלה הדיגיטלית, דבר המקשה עליהם להתחרות ולהגיע לשווקים חדשים.[483]

הפער הדיגיטלי מתבטא גם בחינוך. מאות אלפי תלמידים ערבים חסרים גישה למחשב ביתי ונאלצים להסתמך על טלפונים חכמים כדי להתחבר ללמידה מרחוק ולשירותים דיגיטליים.[484] סמארטפונים מאפשרים גישה לאינטרנט, אך אינם מספיקים לצרכים כמו עבודה ולמידה אפקטיבית ומציבים את התלמידים בחברה הערבית בנחיתות טכנולוגית מובנית. דוח מנדלס מדגיש כי עבור חלקים נרחבים באוכלוסייה הערבית, בעיקר בפריפריה ובקרב אוכלוסיות מוחלשות, המחשב האישי הוא מכשיר נדיר ולעיתים בלתי נגיש. המשמעות היא דיגיטליות צרה ומוגבלת: המשתמשים מתוודעים לרשת דרך מסכים קטנים, קצרים ולא מותאמים למשימות מורכבות. במציאות זו, האפשרות למיצוי הפוטנציאל של טכנולוגיות בינה מלאכותית, כמו תהליכי למידה מתקדמים, שימוש בכלי עבודה מקוונים או עיצוב תוכן משמעותי, נמוכה.

אחת התופעות המטרידות בדוח מנדלס היא ההימנעות הרחבה מהשימוש בדואר אלקטרוני, עד כדי כך שלעיתים אנשים אינם יודעים מהו שם המשתמש שלהם או כיצד לאחזר סיסמה.[485]

483 שם, בעמ' 8.

484 שם.

485 שם, בעמ' 65.

במקום זאת, התקשורת היומיומית מתבססת על אפליקציות כמו ווטסאפ, המאפשרות נגישות מיידית, אך לא מספקות תשתית לצרכים בירוקרטיים, פורמליים או מסחריים. היעדר של כתובת דוא"ל פעילה, או לחלופין, שימוש בכתובת שנפתחה על ידי בן משפחה אחר, מונע הרשמה לקורסים, הגשת טפסים, פתיחת חשבון באפליקציות מסוימות ואף קבלת התראות קריטיות. זוהי פגיעות תשתיתית נסתרת: כאשר מוסדות שלטוניים מצפים לדוא"ל כדרך תקשורת מינימלית, נוצרת אי־הלימה בין מבנה השירות ליכולות המשתמש והדרה שקטה שמונעת מימוש בסיסי של זכויות.

ו. פערי השכלה

ישראל מדורגת מתחת לממוצע ה־OECD במיומנויות פתרון בעיות בסביבה טכנולוגית, ובתוך ישראל נמצאה רמת אוריינות דיגיטלית נמוכה במיוחד בקרב ערבים בהשוואה ליהודים,[486] נוסף לאוריינות השפתית הנמוכה יחסית בחברה הערבית גם בכל הנוגע לשפת האם. לפי סקר התוכנית להערכה בינלאומית של מיומנויות מבוגרים (PIAAC) של ה־OECD לשנת 2023, 70% מהאוכלוסייה הערבית מדורגת ברמה הנמוכה באוריינות קריאה, לעומת 46% במחזור הקודם.[487] כמו כן, בכלל האוכלוסייה הערבית הייתה ירידה משמעותית באוריינות הקריאה (26 נקודות, כמחצית סטיית תקן), ובקרב גברים ערבים חלה ירידה דרסטית אף יותר בציונים לעומת נשים ערביות.[488] הסיכוי להיעדר מיומנויות טכנולוגיות בסיסיות בקרב ערבים גבוה כמעט פי שניים מזה של יהודים (60.2% מהערבים לא משתמשים במחשב, לעומת 22.2% מהיהודים).[489]

פערים אלה נובעים בין היתר ממערכת חינוך בחברה הערבית שבה הישגים לימודיים והנגשת לימודי STEM נמוכים יחסית (רק כ־16% מבני החברה הערבית מחזיקים בתואר אקדמי, לעומת 37% בחברה היהודית).[490] בעקבות זאת, ייצוג הערבים בענפי ההיי־טק ובפיתוח

486 שם, בעמ' 29.

487 מאיה בר "לשליש מהישראלים יש יכולת קריאה של ילד בן 10" **mako** (10.12.2024).

488 שם.

489 הלשכה המרכזית לסטטיסטיקה, לעיל ה"ש 472.

490 ירון דרוקמן "הלמ"ס: 38% מהנשים הן בעלות תואר אקדמי, רק 27.5% מהגברים" **ynet** (5.2.2024); נסרין חדאד חאג'־יחיא, מוחמד חילאילה, אריק רודניצקי ובן פרג'ון "חינוך ותעסוקה בחברה הערבית: שיעור הסטודנטים הערבים הוכפל" **אתר המכון הישראלי לדמוקרטיה** (17.3.2022).

טכנולוגיות בינה מלאכותית נמוך משמעותית מחלקם היחסי באוכלוסייה, נתון המוביל לכך שקולם וצרכיהם כמעט שאינם נוכחים בשולחן הפיתוח של מערכות טכנולוגיות.

לסיכום, שילוב של השכלה מצומצמת, מחסור בהכשרות מתאימות וחסמים בשוק העבודה (לרבות אפליה בקבלה למשרות) מציב את האזרחים הערבים בעמדת נחיתות בהפקת תועלת מהכלכלה הדיגיטלית המודרנית ואף חושף אותם לסיכונים מוגברים (כגון, קושי לנווט בסביבה מקוונת מורכבת או להיזהר מהונאות).

ז. אוריינות דיגיטלית נמוכה

הפער בין שימוש ביישומים טכנולוגיים לבין מימוש יכולות ביקורתיות לגביהם ניכר במיוחד בחברה הערבית. ממחקרם של תחאוכו ואחרים עולה כי רמת אוריינות המדיה בחברה הערבית נמוכה מאוד: ל־87% בגילי 25–64 אין כישורים דיגיטליים בסיסיים, ורק 13% הם בעלי אוריינות דיגיטלית ואוריינות מדיה ברמה בינונית עד גבוהה. זאת, לעומת 59% מהחברה היהודית.[491] דוח מנדלס מדגיש כי רבים מהמשתמשים אינם מבינים לעומק כיצד פועלים הכלים שבהם הם משתמשים, מבחינת הגנת פרטיות, הבנת פעילות אלגוריתמים, אבל בעיקר בהיבט של אפשרות מימוש זכויות הזמינות דרך מרחב הדיגיטלי.[492] בנוסף, נאמר בדוח כי אחד הקשיים המרכזיים בממשק של האזרחים הערבים עם המרחב הדיגיטלי הוא הבלבול בנוגע ל"מי אחראי למה". כאשר מתרחשת פגיעה במרחב הדיגיטלי, למשל התחזות, סחיטה מינית או פלישה לפרטיות, אין ודאות למי יש לפנות, מהן הסמכויות של כל גוף, ואילו זכויות מוקנות למשתמש.

נוסף על כך, המרחב הדיגיטלי של החברה הערבית בישראל מתאפיין בדפוסי שימוש שממוקדים בתוכן קל, ויזואלי ונגיש, בייחוד ברשתות חברתיות כמו טיקטוק, אינסטגרם ופייסבוק (ולא בכלי תקשורת ממוסדים). במגזרים רבים האינטרנט נתפס ככלי ללמידה, להתמקצעות ולהזדמנויות, אבל בקרב קהילות ערביות השימוש מוגבל להיבטים צרכניים ובידוריים. דוח מנדלס מציין כי חוסר אמון ברשויות, היעדר הכוונה חינוכית ופערים בשפה ובתשתיות יוצרים מצב שבו הגלישה ברשתות החברתיות היא אסקפיזם תרבותי ולא מנוע לצמיחה. הפלטפורמות עצמן מחזקות את המגמה: ככל שהשיח מתוחכם פחות, כך

491 תחאוכו, אקסלרד ומטר, לעיל ה"ש 474.

492 דוח מנדלס, לעיל ה"ש 475.

משתמשים מקבלים תכנים רדודים יותר, והאלגוריתמיקה של ניתוב התוכן למשתמשים פועלת כתשתית משמרת פערים.[493]

ח. היעדר ייצוג בסביבה הדיגיטלית

אחד המאפיינים הבולטים בתפר שבין החברה הערבית ומדיניות טכנולוגיה בישראל הוא ההדרה המתמשכת ממוקדי קבלת החלטות. מעטים מאנשי המקצוע שמעורבים בפיתוח מדיניות דיגיטלית בישראל, בממשלה, בתעשייה או באקדמיה מגיעים מהחברה הערבית. כך, כאשר מעוצבים, מאופיינים ונבנים כלי ניתוח נתונים, מערכות ממוחשבות או פלטפורמות המנגישות שירותים לציבור, הם אינם נבנים מתוך היכרות עם הצרכים הייחודיים של החברה הערבית. מן הראיונות שערכנו עלה כי גם כאשר יש ייצוג בהליכי פיתוח מדיניות דיגיטלית ומוצרים טכנולוגיים, ניכרים לחצים של רצייה חברתית, צנזורה עצמית והימנעות מהבעת עמדות או צרכים.

המיעוט הערבי בישראל מתמודד עם אתגרי שפה ותרבות בעולם טכנולוגי הנשלט על ידי רוב דובר עברית בתוך מדינת ישראל ועם אמצעי תקשורת בערבית ממדינות שכנות. אף שמבחינה רשמית לערבית מעמד מיוחד בישראל, בפועל מרבית הפלטפורמות הדיגיטליות, המידע המקוון וממשקי הבינה המלאכותית המקומיים פועלים באנגלית. חוסר הייצוג הלשוני מתבטא בפערי שירות: במגפת הקורונה, למשל, משרד הבריאות התקשה בתחילה לפרסם הנחיות מצילות חיים בערבית, ועיכוב זה פגע באמון הציבור הערבי; כאשר פורסמו ההנחיות, חלקן נוסחו בניב ערבי שאינו מקומי, מה שזרע בלבול ופגע באמינות המידע.[494]
לכך מצטרף קושי מתמשך בהתמודדות עם שפה: תרגומים שגויים של ממשקים לערבית, חוסר שליטה בעברית מקצועית והיעדר הבנה של אנגלית טכנולוגית, שכולם מגבילים הבנה ויכולת להגיב. המשתמש הערבי בישראל אינו רק "משתמש חלש", הוא משתמש שצריך לתרגם כל פקודה פעמיים, לעיתים בלי לדעת בכלל אם הבין נכון.

493 Amy Ross Arguedas, Craig T. Robertson, Richard Fletcher et al., Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review (Oxford: Reuters Institute, University of Oxford, 2022)

494 Larry Hardesty, *Study Finds Gender and Skin-Type Bias In Commercial Artificial-Intelligence Systems*, MIT News (February 11, 2018)

חלק שני: מיפוי הפגיעויות של החברה הערבית בעידן הבינה המלאכותית

הפגיעות של החברה הערבית בישראל בעידן הבינה המלאכותית נושאת אופי ייחודי: היא אינה מתמצה רק בפגיעות אינדיווידואליות, אלא מתרחשת קודם כול במרחב הקולקטיבי. זהו מרחב שבו הבינה המלאכותית פוגשת קבוצה בעלת מאפיינים מובחנים של הדרה ממסדית, מבנה סמכות קהילתי שמרני, פערי תשתית ושפה ותחושת ניכור מתמשכת מהמדינה. במציאות זו החשש אינו רק מפני פגיעה בפרט אלא מפני שחיקה של לכידות קהילתית, טשטוש זהות קולקטיבית וערעור נוסף של האמון בין קבוצת מיעוט לבין מוסדות הרוב.

הבינה המלאכותית, במופעיה השונים, אינה יוצרת את הפגיעות אך סביר להניח שהיא תעמיק אותה: דרך מנגנוני מעקב והתחקות, ייצוגים מגמתיים במערכות מבוססות אלגוריתמים, מניעת נגישות לשירותים ציבוריים והשתקה של קולות בערוצי תוכן. כל אלה אינם תופעות טכנולוגיות אלא ביטויים חדשים לסדר חברתי. מסיבה זו, בחלק זה בחרנו להאיר את הפגיעות הקבוצתית כתשתית שמכילה ומשקפת בתוכה גם פגיעות רגשית, קוגניטיבית ופיננסית. ההתבוננות מהפריזמה הקבוצתית מחייבת לשאול לא רק מהן הפגיעויות של האזרח הערבי היחיד, אלא כיצד משפיעה הבינה המלאכותית על מקומם של אזרחים ערבים בתוך המארג החברתי, הכלכלי והפוליטי של ישראל.

א. פגיעות קהילתית: מעקב ופיקוח

אזרחים ערבים בישראל חשופים לפגיעות קבוצתית מובנית הנובעת מהפעלת טכנולוגיות מעקב ופיקוח חכמות, ובפרט כאלו המבוססות על בינה מלאכותית. בשנים האחרונות חלה התרחבות בשימוש באלגוריתמים לזיהוי חריגות ובמערכות ראייה ממוחשבת על ידי רשויות אכיפת החוק, מגמה שבאה לידי ביטוי גם בהצעות חקיקה, כדוגמת הצעת החוק הממשלתית לפרישת מצלמות זיהוי פנים במרחב הציבורי.[495] אף שהמהלך מוצג ככלי למאבק בפשיעה, בפרט בחברה הערבית, הוא מעורר חשש כבד להיווצרות של "משטר

495 Carrie Keller-Lynn, *Ministers Back Bill To Legalize Widespread Police Use of Facial Recognition Tech*, Times of Israel (Sept. 18, 2023)

מעקב אלגוריתמי" ממוקד קבוצה, שבו כל המרחב הציבורי נתפס כשטח פיקוח של הרוב על המיעוט. המשמעות היא פוטנציאל ליצירת משטר מעקב המכוון לאוכלוסייה הערבית כקבוצת יעד, באופן שעלול להרתיע פעילות אזרחית לגיטימית ולהעצים תחושות רדיפה ופגיעה אישית.[496]

מחקרים בינלאומיים מצביעים על כך שמערכות זיהוי פנים סובלות מהטיות, בפרט בזיהוי אנשים מקבוצות מיעוט אתניות ובעלי גוון עור כהה, תופעה שעשויה להוביל למעצרים שגויים, חשדות שווא ופגיעה בכבוד האדם.[497] מעבר להיבטים האישיים, נטען שפיקוח דיגיטלי איננו ניטרלי אלא ממוקד זהות.[498]

כאשר קמפוס, קניון או תחנת רכבת מנוטרים באמצעות זיהוי פנים, חוויית המשתמש הערבי שונה מזו של היהודי. בעוד יהודים עשויים לראות בטכנולוגיה אמצעי ביטחון, ערבים רואים בה עין בולשת. כאשר מערכת בינה מלאכותית ממיינת מועמדים לעבודה, וערבי נפסל שוב ושוב ללא הסבר, נוצרת הבנה מצטברת שלפיה הבינה המלאכותית אינה משרתת את הציבור אלא מפלחת אותו. זו איננה בעיית שקיפות טכנולוגית בלבד אלא בעיה דמוקרטית של **אכיפה** לא שוויונית המובנית בקוד.[499]

ב. פגיעות קהילתית: הדרה דיגיטלית

הדרה חברתית במרחב הדיגיטלי עלולה להתרחש כאשר מערכות טכנולוגיות, ממשלתיות, מסחריות או ציבוריות אינן מתוכננות מתוך רגישות לגיוון תרבותי ולשוני. החברה הערבית, כקבוצת מיעוט עם שפה וצרכים ייחודיים, מוצאת עצמה לעיתים קרובות מחוץ למערכות אלה לא בגלל הימנעות מבחירה, אלא בשל תכנון מערכתי שמניח שהנורמה היא יהודית־עברית.

496 ראו למשל Sophia Goodfriend, *Algorithmic State Violence: Automated Surveillance and Palestinian Dispossession in Hebron's Old City*, 55(3) Int'l J. Middle East Stud. 461-478 (2021); Nadim Nashif & Mona Shtaya, *Nowhere to Hide: The Impact of Israel's Digital Surveillance Regime on Palestinians*, Middle East Institute (April 27, 2022)

497 Zuboff, לעיל ה"ש 123, בעמ' 204.

498 שם, בעמ' 208 ו־376.

499 כהנא ושוורץ אלטשולר, **אדם ומכונה**, לעיל ה"ש 129.

המעבר מטקסטים כתובים לשפה דבורה עלול להתברר כתופעה שמחריפה את הפגיעות. למשל, כאשר צ׳אטבוט ממשלתי אינו מזהה פניות בערבית, מתרגם את תוכניהן באופן לקוי, נוצרת סביבה דיגיטלית מנוכרת. עבור משתמשים ערבים, תקלות אלו לא יתקבלו רק כמטרד טכני אלא כביטוי לחוסר הכללה מוסדי וחסם בפני מיצוי זכויות או קבלת שירותים ממשלתיים, דווקא בעידן שבו הכול לכאורה זמין באמצעות שיחה ולא דורש תרגום כתוב. זהו המרחב שבו המוגנות וההוגנות מצטלבות.

ג. פגיעוּת קהילתית: סטריאוטיפים וגזענות

במרחב הדיגיטלי, אזרחים ערבים בישראל ניצבים בעקביות בלב שיח שנאה, הדרה והסתה, תופעה שזכתה לתיעוד עקבי בעשור האחרון. לפי מחקרים, מעל 30% מכלל תוכני השנאה ברשתות החברתיות בישראל מכוונים כלפי הציבור הערבי, שיעור הגבוה מכל קבוצה אחרת.[500] הפגיעוּת אינה רק אינדיווידואלית אלא קולקטיבית: גלים של הסתה מתפרצים בעיתות משבר ביטחוני ומזינים תחושת דה־לגיטימציה קבוצתית, השוללת מבני החברה הערבית את תחושת מעמדם כאזרחים שווי זכויות.

ההשפעה של תופעות אלו מועצמת על ידי האופן שבו האלגוריתמים של פלטפורמות התוכן והמדיה החברתית מתעדפים תוכן פרובוקטיבי, מסית וקוטבי מתוך מטרה להגביר מעורבות של משתמשים (engagement).[501]

במילים אחרות, הפלטפורמות אינן רק מרחבים שבהם מתקיימת הסתה אלא הן מנגנונים שמעצימים אותה באופן שיטתי. יש להניח שכפי שאלגוריתמים של רשתות חברתיות נוטים להקצין שסעים בין קבוצות אתניות, כך קורה גם במתחים בין יהודים לערבים, וכך נוצרת סביבה דיגיטלית שבה אזרחים ערבים נחשפים באופן מוגבר לשיח עוין כלפיהם.

אירועי ״שומר החומות״ במאי 2021 בערים המעורבות בישראל המחישו באופן חד את הדינמיקה הזו: רשתות חברתיות הוצפו בתכנים מסיתים, קבוצות ווטסאפ שימשו לתיאום

500 *Hate Speech against Arabs In Israel Reached All-Time High In 2021*, The Jerusalem Post (October 25, 2022); בדומה ראו Gila Amitay, Kfir Asraf, Ety Elisha et al., *Freedom of Expression in Israeli Campuses and Social Media During the War with Hamas*, 48 Journal of Higher Education Policy and Management 1–20 (2025)

501 Eli Pariser, The Filter Bubble: What the Internet is Hiding from You 294 (Penguin, 2011)

עימותים, והפלטפורמות התקשו לבלום את ההידרדרות.[502] בבדיקת עומק שהזמינה חברת מטא עלה כי בתקופה זו הופעלה אכיפת יתר על תכנים פלסטיניים, לצד אכיפה מקילה כלפי דברי שטנה בעברית.[503] חוסר האיזון הזה פגע בזכויות האזרח של הציבור הערבי, שחש כי קולו מושתק.

אחת החוויות המכוננות של החברה הערבית במרחב הדיגיטלי היא תחושת ניכור ואי־צדק שנובעת לא רק מהחשיפה לתכנים פוגעניים אלא מהיעדר תגובה מוסדית הולמת. פוסטים גזעניים, תכנים מסיתים או קריאות לאלימות כלפי ערבים נותרים ללא הסרה או טיפול מצד הפלטפורמות, בעוד תוכן של אזרחים ערבים, גם כשאינו כולל הסתה, נמחק, מוגבל או מסומן.

חוויה זו אינה מתפרשת רק כפגיעה אינדיווידואלית אלא כעדות ליחס מפלה שיטתי: כאילו הכללים אינם אחידים, וכאילו הזהות הערבית עצמה מעוררת חשד. התחושה המצטברת היא של "אזרחות מדרג ב׳" גם במרחב הווירטואלי. תחושת האפליה הזו יוצרת פגיעות רגשית שהיא בה בעת גם קהילתית: היא מכרסמת בתחושת השייכות, מחלישה את תחושת הערך העצמי הקולקטיבי, ולעיתים אף מובילה לתסמיני חרדה, תסכול או ייאוש פוליטי. השיח ברשתות נתפס כמרחב שבו אין לערבים כתובת, הגנה או מקום בטוח. תחושת האין־מענה הזו, יותר מהפוסט הפוגעני עצמו, היא זו שמכוננת פגיעות אמון מערכתית בין הקבוצה הערבית לבין זירות הכוח המרכזיות של השיח הדיגיטלי.[504]

הפגיעות הקבוצתית אינה נובעת רק מהעוינות מבחוץ אלא גם מערעור הלכידות מבפנים. אלגוריתמים נוטים לחשוף צעירים ערבים לעמדות חיצוניות ולשיח גלובלי ללא תיווך תרבותי, מה שעלול לעורר קונפליקטים בין־דוריים, בין־מעמדיים ובין־דתיים בתוך החברה הערבית עצמה (למשל בין מוסלמים, נוצרים ודרוזים; בין מסורתיים למודרנים). התוצאה המצטברת היא פגיעות חברתית: תחושת שוליות וחוסר שייכות הן במרחב הלאומי והן בתוך הקהילה.

502 Inbal Orpaz & David Siman-Tov, *The Unfinished Campaign: Social Media in Operation Guardian of the Walls*, Inst. for Nat'l Sec. Stud. (Sept. 12, 2021)

503 Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021 (BSR [Business for Social Responsibility], 14.9.2022)

504 שם.

מודלים של שפה, המתבססים על למידת כמויות עתק של טקסטים מהאינטרנט, שואבים את עולמם המושגי, בכלל זה עמדות, דעות וסטריאוטיפים, מהשיח הקיים ברשתות החברתיות. כאשר השיח הזה רווי גזענות, הסתה ודעות קדומות כלפי קבוצות מיעוט אתניות, לא מפתיע שגם המודלים משעתקים הטיות אלו. מחקר של המרכז לחקר מודלי יסוד באוניברסיטת סטנפורד מצא כי מודלים של שפה נוטים "לשכתב היסטוריה", לייצר נרטיבים חד־צדדיים ולעיתים אף למחוק קולות של קהילות מיעוט או להציגן באור שלילי, בהתאם לאופן שבו הן מתוארות ברשתות החברתיות ובמקורות טקסטואליים פופולריים.[505] ממצאים דומים ממחקרים אחרים הצביעו על כך שמודלים כמו ChatGPT ו־Claude עלולים להציג תוכן פוגעני או מוטה כאשר הם נשאלים על קבוצות אתניות מסוימות, ובפרט כאשר הקבוצה נתונה בעימות פוליטי או תרבותי.[506] לכן קבוצות מיעוט בעלות נראות ציבורית גבוהה, כמו אזרחים ערבים בישראל, נמצאות בסיכון ל"שכפול דיגיטלי של הדרה". אם מודל הלמידה ניזון משיח רווי הסתה ברשתות, הוא "ילמד" שערבים הם אויב, בעיה או תת־קבוצה שאין להתחשב בה. זוהי פגיעות קהילתית מסוג חדש: לא תוצאה של החלטה מודעת של מתכנת אלוגריתמים או של מודל עסקי של רשת חברתית ואפילו לא של רגולטור, אלא של רשת למידה המגבירה את הדימוי השלילי.

ד. פגיעות קהילתית: הטיות

פגיעות קבוצתית נוספת של החברה הערבית נובעת מהטיות שיטתיות המובְנות במערכות בינה מלאכותית, הן ברמת השפה והתרבות והן ברמת הנתונים וההקשר החברתי.

ראשית, אלגוריתמים שאומנו בעיקר על תכנים בעברית או באנגלית מתפקדים לרוב באופן מוטה או לקוי כאשר הם נדרשים לנתח, לפרש או להסיר תכנים בערבית. הדבר בולט במיוחד בתחום ניטור התוכן במדיה החברתית: ארגוני זכויות אדם תיעדו שורה של מקרים שבהם מערכות סינון אוטומטיות מצנזרות תכנים בערבית, לעיתים קרובות ללא הצדקה ממשית,

505 Rishi Bommasani, Drew A. Hudson, Ehsan Adeli et al., On the Opportunities and Risks of Foundation Models (Center for Research on Foundation Models & Stanford Institute for Human-Centered Artificial Intelligence, Aug. 16, 2021); וכן כהנא ושוורץ אלטשולר, **אדם ומכונה**, לעיל ה"ש 129, עמ' 77.

506 Yue Zhang, Weikang Wang, Ying Liu et al., *Quite Good, but Not Enough: Nationality Bias in Large Language Models – A Case Study of ChatGPT*, arXiv preprint (2024)

ובמקביל מתירות הפצת תכנים מסיתים נגד ערבים בשפה העברית.[507] חוסר האיזון הזה אינו נובע בהכרח ממדיניות מפלה מוצהרת, אלא מתת־ייצוג של החברה הערבית בתכנון האלגוריתמים, מהיעדר מומחיות בלשון הערבית, ומהנחות תרבותיות חבויות שקובעות מראש מה בעייתי ומה לגיטימי.

שנית, מערכות בינה מלאכותית בתחומים מינהליים, כלכליים או תעסוקתיים, כגון מערכות מיון ואיתור מועמדים לעבודה, דירוגי אשראי או חישובי פרמיות ביטוח, עלולות להנציח דפוסי אפליה קיימים כלפי החברה הערבית.[508] כאשר מודל לומד מנתוני עבר שבהם שיעורי הדחייה של מועמדים ערבים גבוהים, הוא עשוי לשמר ולשכפל את הדחייה הזו גם בעתיד, מבלי להבחין בין הטיה היסטורית לבין קריטריון רלוונטי. כך, גם ללא כוונת זדון נוצרת אפליה מובְנית שבפועל סוגרת דלתות בפני אזרחים ערבים.

במיוחד בעייתיים מודלים שמשתמשים בפרמטרים דמוגרפיים עקיפים, כמו שם, מקום מגורים, סגנון כתיבה או שירות צבאי, אשר עלולים לשמש תחליף לזהות האתנית עצמה. צעיר ערבי משכיל המתגורר בכפר בגליל עלול לזכות לדירוג אשראי נמוך או לאי־זימון לריאיון עבודה, לא בשל היעדר כישורים, אלא משום שהמודל למד לקשר בין משתנים עקיפים לבין "סיכון" או "אי־התאמה".[509]

כך נוצרת "תקרת זכוכית אלגוריתמית": מערכת שמציגה את עצמה כאובייקטיבית, אך בפועל פועלת על בסיס נתונים מקוטבים שמגבירים הדרה חברתית. פגיעות זו אינה רק אישית, אלא מבנית וקבוצתית, שכן היא מעצבת את תנאי ההשתלבות של החברה הערבית במשק, במרחב האזרחי ובכלכלה הדיגיטלית.

507 Human Rights Due Diligence, לעיל ה"ש 503.

508 רות פלאטו־שנער ומעיין פרל (פילמר) "חיתום אשראי צרכני על סמך עיבוד מידע אלגוריתמי: היש סיבה לחשוש הפליה אסורה?" **עיוני משפט** מו 553–610 (2022).

509 ראו Oren Bar-Gill & Cass R. Sunstein, Algorithmic Harm: Protecting People in the Age of Artificial Intelligence, 11 (Oxford University Press, 2025), שם נידון נושא האפליה האלגוריתמית בהקשר דומה. בר גיל וסאנסטין כותבים:

> Thus, the algorithm's pricing or targeting decisions that are calculated to maximize profits might create a disparate impact on people of color or women. For certain goods, women might be charged higher prices or be denied certain deals and opportunities. We will see other problems as well.

עבור החברה הערבית, המשמעות היא חסמים מבניים להשתלבות כלכלית והעמקת תחושת השוליות במערכת הישראלית. ההשלכה הקולקטיבית חמורה: תסכול, חשדנות כלפי טכנולוגיה, ותחושת ניכור כלפי מוסדות מדינה, שמובילים לא רק להדרה דיגיטלית אלא להעמקת הקרע האזרחי.

ה. פגיעות אישית פיזית בחברה שמרנית

במרקם החברתי השמרני של החברה הערבית, כפי שכתבנו ביחס לחברה החרדית, פגיעה אישית, רגשית, מינית או חברתית, איננה רק חוויה פרטית אלא גם סוגיה קהילתית. השם הטוב של האדם ושל משפחתו, כבודה של הקהילה והחשש מהוקעה יוצרים סביבה שבה שיח גלוי על פגיעות כמעט שאינו אפשרי. היעדר מנגנונים פנימיים בטוחים לפנייה, בצד מיעוט של שירותים רגישי תרבות במערכות הרווחה והבריאות, גורמים לכך שרבים מהנפגעים פשוט שותקים.

הטמעת טכנולוגיות בינה מלאכותית עלולה להחריף פגיעות זו. ברשת, אזכור של שם או תמונה עשוי להפוך מיידית לוויראלי; וכלים מבוססי בינה יוצרת עשויים ליצור תכנים מבזים מזוהים שנראים מציאותיים. במצבים כאלה, החשש מחשיפה נעשה קשה. עצם האפשרות ששמם של צעיר או צעירה ייקשר, אפילו בטעות, לתוכן שלילי, עלולה להסב נזק בלתי הפיך למעמדם בקהילה, לסיכויי נישואין ולתחושת ערך עצמי.

המורכבות מתעצמת בשל מחסומים תרבותיים ולשוניים. בשונה מיהודים דוברי עברית, עבור אזרחים ערבים השפה הדיגיטלית הרווחת, בין במערכות ציבוריות, אפליקציות טיפוליות או פורומים של סיוע, אינה שפת אם. אפילו כאשר קיימים שירותים נגישים בערבית, התרגום לעיתים לקוי, איטי או חסר הקשר תרבותי. מעבר לכך, עצם השיח על מוגנות, מושג חדש יחסית, חסר שפה משותפת ברמה התרבותית: אין מילים מדויקות, מודלים של דיאלוג קהילתי ופרקטיקות של דיווח בטוח. בהיעדר שפה ואמון, גם הקורבן שזועק לעזרה נותר מושתק. תחושת ניכור מול רשויות המדינה, שנתפסות לעיתים קרובות כעוינות או שיפוטיות, מחזקת את ההסתגרות. התוצאה היא פגיעות כפולה: גם מחשיפה אישית לא רצויה, וגם מהיתקלות במערכת שלא מבינה את הייחוד החברתי, השפה או התרבות.

ו. פגיעות קוגניטיבית

היעדר היכולת להבחין בין תוכן שנוצר בידי מכונה לבין תוכן אנושי ואותנטי הוא אתגר משותף לכלל האוכלוסייה בעידן של בינה מלאכותית גנרטיבית. אולם פערי שפה, אוריינות דיגיטלית נמוכה והיעדר תשתית לחשיבה ביקורתית הופכים את החברה הערבית לפגיעה במיוחד לא רק לטעויות בזיהוי, אלא גם להטעיה מכוונת, תעמולה וזיופים. מדובר בפגיעות קוגניטיבית שמערערת את היכולת להבחין בין אמת לשקר ומכרסמת באמון הציבורי במקורות מידע ובסמכויות מוסדיות. פגיעות זו מחריפה עוד יותר לנוכח מצבה הייחודי של החברה הערבית, הנתונה בתווך שבין מדינה, תנועות פוליטיות אזוריות, שחקנים בינלאומיים וארגוני חברה אזרחית, שכולם מבקשים לעצב את תודעתה, את העמדות הפוליטיות שלה ואת התנהגותה הקהילתית. בתוך סביבה רוויית השפעות ונרטיבים מתחרים, היכולת לזהות מניפולציות שמקורן במכונה הופכת לא רק לאתגר אישי אלא לשאלה של חוסן קהילתי. מדובר בפגיעות קוגניטיבית שמערערת את היכולת להבדיל בין אמת לשקר ומכרסמת באמון הקולקטיבי של הקהילה במקורות מידע, בסמכות ציבורית ובאפשרות לדעת את האמת.

ו1. נגישות מוגבלת למידע מהימן

מקורות המידע המרכזיים בישראל – חדשות, אתרי ממשל, תוכן רפואי ושירותים דיגיטליים – זמינים ברובם בעברית בלבד, ולעיתים מתורגמים לערבית באיחור או באיכות לקויה. לחברה דוברת הערבית בישראל יש נגישות מוגבלת למידע מוסמך בערבית. לפי המדד להנגשת מידע ושירותי ממשל לחברה הערבית לשנת 2023, רק 10% מהעמודים באתרי הממשלה השונים זמינים בשפה הערבית.[510] האמון הנמוך של ערביי ישראל באמצעי התקשורת הממוסדים דוחף רבים מקרב החברה הערבית לצרוך מידע דרך רשתות חברתיות או מקורות לא רשמיים בשפה הערבית, שבהם נפוצה יותר הפצת שמועות, דיסאינפורמציה

510 אחת הדוגמאות לכך היא היעדר הסברה של פיקוד העורף בזמן מלחמת לבנון השנייה, דבר שחייב תחנות רדיו אזוריות המיועדות למגזר הערבי לתווך תוכן לאוכלוסייה. ראו שלמה דסקל ותהילה שוורץ־אלטשולר **תחנת הרדיו א־שמס בתפר שבין רגולציה, פוליטיקה וכלכלה** 106 (המכון הישראלי לדמוקרטיה, 2015). דימא אסעד־ניקולא **המדד להנגשת מידע ושירותי ממשל לחברה הערבית** (איגוד האינטרנט הישראלי, 2023).

ותכנים מגמתיים.[511] משבר הקורונה המחיש את הסכנות הטמונות בכך: העיכוב בפרסום הנחיות רפואיות בערבית יצר רִיק תקשורתי שהתמלא בשמועות כתחליף. כאשר הידע אינו נגיש, גם ההגנה הקוגניטיבית נחלשת ומתגברת התלות במקורות בלתי אמינים.[512]

במסגרת החלטות הממשלה 550 משנת 2016 ו־922 משנת 2021 הוקצו תקציבים ייעודיים לצמצום פערים שונים שמהם סובלת האוכלוסייה הערבית, ובהם הפער הדיגיטלי. בין היתר, התוכנית התייחסה להנגשה שפתית ותרבותית של תוכן ממשלתי בערבית, להקניית מיומנויות דיגיטליות שונות ולהקמת מרכזים שיסייעו ברכישת מיומנויות אלו. אולם התקציב שהוקצה לתוכנית בפועל היה נמוך מהמתוכנן.[513]

2ו. היעדר כלים לביקורת ואימות מידע

חלק מהמשתמשים הערבים מתקשים לזהות מידע כוזב או לפתח כלים להערכת תוכן, בין שמדובר בקונספירציות, פייק־ניוז, תכנים יצירי מכונה (deepfakes).[514] תופעה זו מעמיקה את הפגיעוּת הקוגניטיבית במספר מובנים: היא מעודדת הפנמה של סטריאוטיפים שליליים (למשל, כאשר מנועי חיפוש מבוססי אלגוריתמים מקשרים את המילה "ערבים" לתוכן אלים או מסית) ומגבירה את החשש כי החלטות מדיניות, עיתונאיות או ציבוריות מתקבלות על בסיס מידע שגוי. החולשה המבנית הזו מחלישה את היכולת של קהילות ערביות להגן על עצמן במרחב הציבורי הדיגיטלי ולהשפיע על השיח שנוגע להן.

511 דוח מנדלס, לעיל ה"ש 475, בעמ' 8; ראו גם תהילה שוורץ אלטשולר ואינה אורלי ספוז'ניקוב **סקר שימושי תקשורת, אמון בתקשורת ואוריינות תקשורת** (המכון הישראלי לדמוקרטיה, 2024).

512 נסרין חדאד חאג'־יחיא, איימן סיף ובן פרג'ון **החברה הערבית בתקופת משבר הקורונה: השפעות המשבר והמלצות ליציאה מהמשבר** 10 (המכון הישראלי לדמוקרטיה, 2020); דפנה אבירם־ניצן, ירדן קידר, נסרין חדאד חאג'־יחיא, בן פרג'ון **"משבר הקורונה והחברה הערבית: סקרי קורונה" אתר המכון הישראלי לדמוקרטיה** (18.3.2021).

513 אסטבן קלור, אפרים לביא, מאיר אלרן ואח' "תוכניות החומש לפיתוח החברה הערבית בישראל: סיכום תוכנית 922 ולקחים לתוכנית החומש החדשה 550" **במה מחקרית, המכון למחקרי ביטחון לאומי** (2022).

514 ניצן ישעור ונרקיס משה **העוקץ האלגוריתמי: רשתות חברתיות ובינה מלאכותית כאמצעי להונאות פיננסיות** 15 (איגוד האינטרנט הישראלי, 2025). המחברים (הגם שאינם מתייחסים ישירות לחברה הערבית) מדגישים שמערכות אלה מותאמות תרבותית כדי להעצים את הפגיעה.

3ו. תלות במכונות כתחליף לסמכות אנושית

בראיונות שקיימנו למדנו כי בקרב צעירים ערבים, ובמיוחד בקהילות שבהן מערכת החינוך מדגישה שינון וסמכות על פני חשיבה ביקורתית, קיימת נטייה גוברת להסתמך על מערכות אוטומטיות, ובראשן על מודלים של שפה כמו ChatGPT, כעל מקור אמת וסמכות מוחלט. הביטוי "הצ'אט אמר לי" הפך בקרב חלק מהצעירים לשקול ערך לדבריו של מורה, הורה או מנהיג דתי. כאשר הבינה המלאכותית מוצגת כבלתי ניתנת לערעור, וכאשר המערכת אינה מקנה כלים לחשיבה עצמאית, נוצרת תלות סמויה בטכנולוגיה לא רק ככלי אלא כסוכן תרבותי. הפגיעות היא כפולה: מצד אחד, משתמש מוותר מראש על האוטונומיה שלו; מצד שני, הוא חסר מודעות לכך שהמערכת עלולה להוציא פלט שגוי, מוטה או פוגעני. כך מועתקת דינמיקה סמכותנית מן ההקשר החברתי־קהילתי אל תוך האינטראקציה עם המכונה, באופן שמכרסם עוד יותר ביכולת של הפרט לבקר, לאתגר ולהתנגד.

ז. פגיעוּת פיננסית וכלכלית

יש לציין את פגיעותם של בעלי אוריינות דיגיטלית נמוכה להונאות מקוונות ותרמיות פיננסיות.[515] אוכלוסיות מוחלשות דיגיטלית נופלות לעיתים קורבן להודעות פישינג, הצעות השקעה מזויפות או ניצול במרחב הקריפטו, בשל קושי בזיהוי הסכנה. במובן זה אנשי החברה הערבית חשופים כמו קהילות אחרות – למשל זקנים וחרדים.

חלק שלישי: דרכי התערבות לקידום מוגנוּת האזרחים הערבים בעידן הבינה המלאכותית

העיקרון המסדר של חלק זה הוא שמול ריבוי סוגי הפגיעות של החברה הערבית בעידן הבינה המלאכותית, החל בהדרה וכלה בשיבוש מערכות זהות ויחסי אמון, ההתערבות צריכה להתבסס על הכרה במבנה החברתי הייחודי של החברה הערבית בישראל. בכלל אלה מה שמנינו קודם לכן: המתח שבין השתייכות למרחב אזרחי יהודי־רוב לבין נאמנות לקהילת מיעוט בעלת לשון, תרבות ונרטיב לאומי שונה; האמון המוגבל כלפי

515 Fidria H, Junaidi Junaidi & Rismawati Rismawati, *Analysis of the Influence of Society Vulnerability to Online Fraud* (July 8, 2025)

מוסדות המדינה; והפערים המובְנים בגישה למשאבים, ידע והשפעה על תהליכי תכנון טכנולוגיים.[516]

בחלק זה נציע מכלול התערבויות רב־שכבתי, המתפרש על פני ארבעה צירים: פיתוח תשתיות ידע ומחקר, רגולציה מותאמת תרבותית ובניית אמון מוסדי, חינוך ואוריינות, ושותפות וגיוון בפיתוח ובפיקוח על מערכות בינה מלאכותית. כל ציר מתמודד עם היבט אחר של הפגיעוּת הקולקטיבית והאישית ומבקש להציע מענה ממוקד ומבוסס הקשר שיחזק את המוגנות של החברה הערבית במרחבים הטכנולוגיים המשתנים.

א. פיתוח תשתיות ידע: מחקר, מעקב ושקיפות

קידום מוגנות בחברה הערבית בעידן הבינה המלאכותית מחייב בראש ובראשונה פיתוח של תשתיות ידע ייעודיות, המבוססות על מחקר ומעקב שיטתי. להבדיל ממחקר הנוגע לילדים או לזקנים שנערך בכל העולם, ללא נתונים רלוונטיים ומפורטים, קשה לזהות דפוסים של אפליה טכנולוגית, לעמוד על השלכות ייחודיות לחברה הערבית או לפתח מדיניות מבוססת־ראיות. בפועל, קיים מחסור במידע על האופן שבו מערכות AI משפיעות על קהילות מיעוט בכלל, ועל אזרחים ערבים בישראל בפרט. לכן אנו מציעים לקדם סדרת מחקרים רחבה, בשיתוף פעולה בין האקדמיה, מכוני מחקר וארגוני חברה אזרחית, אשר תעסוק בין היתר בניתוח של פרקטיקות קבלת החלטות אוטומטיות בשירותי רווחה, חינוך ובריאות והשלכותיהן על אזרחים ערבים; מיפוי של הופעות חוזרות של סטריאוטיפים ביחס לערבים ישראלים במערכות מבוססות שפה (כמו מודלים גנרטיביים וצ׳אטבוטים); מעקב אחרי פערים בייצוג נתונים, לדוגמה, אם תיעוד של שמות, מיקומים או שיח בערבית משפיע על תפוקות של מערכות AI שבהן נעשה שימוש בשוק העבודה; חקר התנהגות משתמשים ערבים ונכונות לאמץ או לדחות שירותים מבוססי בינה מלאכותית עקב מחסור באמון, שפה או רגישות תרבותית.

אנו מציעים להקים "מעבדת מוגנות בינה מלאכותית" כגוף עצמאי בין־מגזרי שיפעל בתוך החברה הערבית ותפקידו יהיה לאסוף עדויות, לערוך סקרים וראיונות, לתעד שימושי בינה מלאכותית ולנתח את השפעתם על תחושת הביטחון, ההכלה והאמון בקהילה. מעבדה כזו יכולה לפעול במודל היברידי עם רכזים מקומיים מטעם רשויות מקומיות ערביות

516 חילאילה, בדראן ורודניצקי, לעיל ה"ש 463.

וועדת המעקב העליונה, בשיתוף אוניברסיטאות (למשל חיפה, בן־גוריון או האוניברסיטה הפתוחה) וארגוני זכויות אדם כמו עדאלה או המרכז הערבי לתכנון אלטרנטיבי. המעבדה תהפוך למוקד מידע שיזין את מקבלי ההחלטות, את חברות הטכנולוגיה הגדולות וכן את הציבור, בדומה למיזמים דומים בעולם שתכליתם לדאוג להגנה על זכויות מיעוטים בעידן הנוכחי, כמו Partnership on AI.[517]

תוצרי המעבדה יהיו:

(1) **מיפוי של אלגוריתמים פוגעניים** ובניית קטלוג מתעדכן של מערכות טכנולוגיות (ממשלתיות או פרטיות) שיש בהן סימני אפליה או סיווג בעייתי של קהילות בחברה הערבית.

(2) פרסומים פומביים תקופתיים, מעין "**דוח זכויות בינה מלאכותית**" של החברה הערבית, שיספק עדות שיטתית, נגישה ותקשורתית לפערים קיימים ויוכל גם לשמש להנגשת בעיות למקבלי החלטות ולשיח ציבורי רחב.

(3) **אינדקס קהילתי למעקב אחרי השפעות AI**, בדומה למדד הגיוון התעסוקתי של הלמ"ס,[518] שיכלול מדדים כמותיים ואיכותיים המנטרים ייצוג ב־ וחשיפה ל־ של הציבור הערבי לטכנולוגיות בינה מלאכותית, כגון שיעור השימוש בכלים מבוססי AI, רמת האמון בכלים, מספר מקרי אפליה מתועדים, רמת התאמה לשונית ותרבותית של ממשקים ועוד. הנתונים ייאספו באמצעות סקרים תקופתיים, דוחות ציבוריים, תיעוד של תלונות ואירועים חריגים, ובשיתוף עם גופי ממשל דוגמת הרשות להגנת הפרטיות, המשרד לשוויון חברתי ומרכז השלטון המקומי, ועם ארגוני חברה אזרחית.

השלב הראשון בבניית תשתיות ידע הוא אפוא הכרה בכך שללא ידע, אין מוגנות. סיוע לחברה הערבית לייצר את הידע הזה על עצמה, בעצמה ולמענה הוא מסלול התערבות תשתיתי נדרש.

517 *About Us*, Partnership on AI

518 יפית אלפנדרי, אלכסנדרה קלב, אילה גינת ואח' מדד הגיוון התעסוקתי: ייצוג ושכר בשוק העבודה הפרטי והציבורי (נציבות שוויון הזדמנויות בעבודה, 2024).

ב. אוריינות וחינוך

כדי להתמודד עם פגיעות קוגניטיבית, פערי ידע, חשיפה להטיות ותחושת חוסר אונים טכנולוגית, נדרשת השקעה רב־שכבתית ביצירת תשתית של אוריינות דיגיטלית.

ב1. פיתוח תוכניות הכשרה מותאמות גיל, מגדר ותרבות לאוריינות בינה מלאכותית בחברה הערבית

החינוך הדיגיטלי הקיים בישראל לרוב אינו מותאם לתרבות, לשפה ולערכים של החברה הערבית וכאשר הוא קיים, הוא עוסק במוגנות רשתות חברתיות ואינטרנט.[519] לכן נדרש פיתוח תוכניות הכשרה ייעודיות בבתי הספר,[520] במרכזים קהילתיים ובמרחבים בלתי פורמליים, המשלבות בין תוכן טכנולוגי (למשל: איך פועל אלגוריתם חיפוש? היכרות עם מודלים זמינים של שפה ועם יישומים של בינה מלאכותית) לבין כלים לחשיבה ביקורתית וזהירות טכנולוגית (למשל: איך בודקים אמינות של מידע? מתי נכון לא לתת אמון במערכת? איך נראה מבצע השפעה פיננסי מבוסס בינה מלאכותית? מהי המשמעות של צ׳אטבוט כחבר דיגיטלי וכיועץ פסיכולוגי ומהן המגבלות שלו?).

יש להתאים את ההכשרות להבדלים בין קבוצות בתוך החברה הערבית: נשים מבוגרות, בני נוער, סטודנטים וכיוצא באלה, וליצור שיתופי פעולה עם מרכזים קהילתיים, ספריות, מסגדים, מרכזי צעירים, אוניברסיטאות ומכללות מקומיות וכן גופים המכינים לקראת עולם העבודה. המטרה היא ליצור מסגרת עקבית, רב־דורית ומובנית, שתאפשר לכל אדם, בכל גיל ורקע, לפתח מוגנות.

ניתן לחשוב גם על חיזוק יוזמות קיימות שבונות גשר דיגיטלי בין החברה הערבית ליהודית, למשל, יצירת פלטפורמות משותפות לדיון ציבורי מקוון, בקבוצות מעורבות, במטרה לשבור בועות פילטר ולצמצם חשדנות בין המגזרים.[521]

519 ראו למשל את היוזמה של איגוד האינטרנט הישראלי של חודש האינטרנט בחברה הערבית, חודש האינטרנט הבטוח בחברה הערבית: סדנאות מוגנות דיגיטלית בגיל השלישי ברשויות הערביות (26.2.2025).

520 ראו דוח מנדלס, לעיל ה"ש 475.

521 יש לציין כי כבר היום ישנן יוזמות מסוג זה אך מומלץ לתמוך בהן ולעודד יצירת יוזמות נוספות. ראו למשל: Andrew Konya, Luke Thorburn, Wasim Almasri et al., *Using*

2. הקמת פלטפורמת מיקרו־למידה מבוססת ווטסאפ בערבית

לאור הפופולריות של ווטסאפ ככלי תקשורת עיקרי, יש לפתח פלטפורמות מיקרו־למידה, כלומר מערכי לימוד קצרים, נגישים ויזואלית, בשפה הערבית ובפורמט מותאם לסלולרי, שיעסקו באוריינות ובמוגנות דיגיטלית. תכנים אלו יכללו: אזהרות מפני שימוש עיוור במודלים של שפה, טיפים לזיהוי תוכן מזויף, כלים להנמכת חשיפה למסרים מסיתים, יחסי אדם ומכונה וגם הכוונה לדיווח על פגיעות. מודלים כאלה כבר פועלים בהצלחה, למשל מערכת DigiBot מבית הארגון Digify Africa שהיא מערכת מיקרו־למידה מבוססת ווטסאפ שמספקת מערכי למידה בנושאי מיומנויות דיגיטליות, שיווק דיגיטלי ובטיחות מקוונת.[522] הייחוד של היוזמה הוא השילוב של AIעם פלטפורמת ווטסאפ שמאפשר חינוך טכנולוגי נגיש ונייד במיוחד עבור קהלים בפריפריה עם גישה דיגיטלית בסיסית לרשת סלולרית גם מדור נמוך. עד כה גייסה DigiBot למעלה מעשרות אלפי לומדים במדינות אפריקה באמצעות שיעורי אימות, תעודת סיום, משחקונים ואינטראקציה יומיומית בתוכנה.[523] פלטפורמה כזאת, המאפשרת העברת תכנים בשפה פשוטה ותומכת שיח ויזואלי ותגובות קוליות, יכולה להיות מתאימה לחברה הערבית בישראל.

3. יצירת תוכן ויזואלי נרטיבי בערבית לקידום אוריינות בינה מלאכותית

על מנת להנגיש את השיח על בינה מלאכותית ולהפוך אותו לרלוונטי לבני נוער, ניתן לפתח סדרות רשת קצרות בשפה הערבית שידונו בסוגיות חדשות כמו: פרופילים מזויפים, הטיות של אלגוריתמים, השפעת רשתות חברתיות על תודעה קולקטיבית ויחסי אדם ומכונה. כך ניתן להכניס מסרים של זהירות וביקורתיות דרך מדיום שקרוב לעולמם של צעירים, לא מתוך פטרונות אלא מתוך שותפות ויצירת זהות דיגיטלית מודעת. ניתן לשקול

Collective Dialogues and AI to Find Common Ground Between Israeli and Palestinian Peacebuilders, arXiv preprint (March 2025)

522 Kele Scheppets, *Digify Africa Unveils OpenAi Powered Learning Via WhatsApp*, Digify Africa (Fab. 10, 2023)

523 Marinela Potor, *How Digify Africa Reaches 32,000 Online Learners with a WhatsApp Chatbot*, Sinch (Feb. 13, 2025)

שיתוף פעולה עם גופים כמו תאגיד השידור הציבורי, בדומה לסדרת הרשת "בינו לבִינה" של התאגיד, שמונגשת רק בעברית.[524]

41. עיגון הזכות להסברתיות בשפה נגישה

קידום אוריינות טכנולוגית בקרב החברה הערבית מחייב לא רק הכשרות וכלים ביקורתיים, אלא גם הסרת חסמים מובנים בנגישות למידע על החלטות שמתקבלות על בסיס מערכות בינה מלאכותית. לכן, חשוב לעגן את הזכות להסברתיות נגישת שפה. כל החלטה אוטומטית בעניינים מהותיים, כגון דחיית הלוואה, סיווג ביטחוני, שלילת זכאות לקצבה או שיבוץ למוסד חינוכי, צריכה להיות מלווה בהסבר מובן, שקוף ומפורט בשפה הערבית, לרבות פירוט האלגוריתם שבו נעשה שימוש, הנתונים שהוזנו והשיקולים שעמדו בבסיס ההחלטה.

ג. בניית אמון מוסדי ומדיניות מותאמת לחברה הערבית

לאורך הפרק עמדנו על דפוסי פגיעוּת ייחודיים של החברה הערבית במרחבים דיגיטליים המתווּכים על ידי בינה מלאכותית, החל בהדרה סמויה, דרך קיבוע סטריאוטיפים והסללה אוטומטית, ועד שעתוק של שיח שנאה והעצמת ניכור פנימי. פגיעוּת זו אינה רק טכנולוגית או רגולטורית אלא היא תולדה של הדרה שיטתית מהשולחנות שבהם מתקבלות ההחלטות הטכנולוגיות. לנוכח זאת, גיבוש מענים של מוגנות מחייב מעבר ממודלים פטרנליסטיים של הגנה ותגובה, למנגנוני ייצוג ופרוצדורות של הכללה, וכן לשותפות בעיצוב ובאודיטינג של מערכות בינה מלאכותית.

הדיון באתגרים הייחודיים, ובהם הטיות מערכתיות, פערי נגישות טכנולוגית ופגיעוּת קבוצתית, מלמד כי לא ניתן להסתפק במדיניות אוניברסלית לגבי טכנולוגיות בינה מלאכותית, אלא נכון לאמץ גישה של אוניברסליות דיפרנציאלית. כלומר, דרכי ההתערבות לקידום המוגנות בחברה הערבית צריכות להתממש בהתאמה להקשרים תרבותיים, חברתיים, מעמדיים ולשוניים של קבוצות מיעוט. במילים אחרות, החקיקה והמדיניות צריכות להיות שוויוניות לא בכך שהן זהות לכולם, אלא דווקא בכך שהן מסוגלות להשתנות ולהתכוונן בהתאם לצרכים השונים של קבוצות אוכלוסייה שונות. עיקרון זה הולך ומחליף את

524 בינו לבִינה: סדרה על מערכת יחסים מורכבת מסוג חדש – יחסים שבינינו לבינה מלאכותית כאן – תאגיד השידור הישראלי.

התפיסה הליברלית־פורמלית של "שוויון בפני החוק", שנמצאה לעיתים קרובות עיוורת להבדלים מערכתיים ולפערי כוח, במיוחד כאשר מדובר בקבוצות מיעוט, קהילות מודרות או אוכלוסיות דוברות שפה שונה מזו של מוסדות המדינה. המושג נזכר מפורשות במסמכים כמו המלצות אונסק"ו לאתיקה של בינה מלאכותית (2021), שקובעות כי על המדינות החברות "לשלב מדיניות אוניברסלית עם גישות מבוססות הקשר על מנת להבטיח הכלה דיגיטלית וצדק חברתי, במיוחד לאוכלוסיות מוחלשות."[525] גם דוח מיוחד של מועצת זכויות האדם של האו"ם משנת 2022, בנושא "השפעת בינה מלאכותית על זכויות של קהילות ילידיות ומיעוטים", מדגיש את הצורך בעיצוב רגולציה שמכירה בהבדלים ולא מניחה "ניטרליות תרבותית".[526]

ג1. שילוב ערבים בתהליכי קבלת החלטות על מדיניות בינה מלאכותית

תחושת ניכור עלולה להתרחש בחברה הערבית לנוכח מדיניות מונחתת מלמעלה בתחום הבינה המלאכותית, ללא הקשבה לקול הערבי וללא הבנה מספקת של הצרכים, הערכים והחששות הייחודיים של המיעוט הערבי בישראל. כדי למנוע ערעור על הלגיטימציה הציבורית של מהלכים ממשלתיים, יש להבטיח ייצוג הולם של אזרחים ערבים בכל פורום וגוף העוסקים בגיבוש מדיניות טכנולוגית, החל בוועדות ממשלתיות, דרך צוותי חשיבה, ועד מנסחי ניירות מדיניות, כגון אלה העוסקים בהטמעת בינה מלאכותית במגזר הציבורי. ראוי כי פורום המומחים בתחום הבינה המלאכותית, שהוקם על ידי משרד החדשנות, המדע והטכנולוגיה לצורך ייעוץ לממשלה,[527] יורחב כך שיכלול יותר מומחים ומומחיות מהחברה הערבית. אלו יביאו עימם ידע, רגישות ונקודת מבט הכרחית לעיצוב מדיניות מכלילה.

ג2. שילוב ערבים בתהליכי קבלת החלטות לגבי הטמעת פרויקטים מבוססי בינה מלאכותית

כדי למנוע תחושת ניכור ולהגביר את הלגיטימציה הציבורית של תהליכי הטמעת כלים מבוססי בינה מלאכותית, יש להבטיח מעורבות אפקטיבית של אזרחים ערבים בכל שלב של

525 Recommendation on the Ethics of Artificial Intelligence, art. 67 (UNESCO, 2021)

526 *Report of The Special Rapporteur on The Right To Privacy in The Digital Age* (U.N. Human Rights Council, U.N. Doc. A/HRC/51/17, May 4, 2022)

527 המומחים שיצעידו את ישראל לעתיד, לעיל ה"ש 457.

קבלת החלטות הנוגעות בכך, מתוך גישה של רגולציה משתפת (co-design) המבוססת על דיאלוג מקדים, ניתוח צרכים ריאליים וייצוג קהילתי פעיל.

כך למשל, לפני פרישת מערכות זיהוי פנים, מצלמות חכמות או תשתיות חישה אחרות ביישובים ערביים, נדרש לקבוע מנגנון שיתוף ציבור מובנה, הכולל נציגי שלטון מקומי, מומחים בלתי תלויים ונציגים קהילתיים. אלה יוכלו לבחון את הפוטנציאל, את הסיכונים ואת ההתאמה של המערכות לצרכים הקונקרטיים, למשל בתחום הביטחון האישי, התחבורה הציבורית או תאורת הרחוב; את הסיכונים מפני אפליה ושיטור יתר, ולאזן בין תועלות טכנולוגיות לבין זכויות אזרח.

שימועים ציבוריים בשפה הערבית בנוגע לרגולציה של טכנולוגיות דיגיטליות יכולים לשמש נדבך חשוב, ובכללם מפגשי "שולחן עגול" קבועים בין מנהיגים מן המגזר הערבי לבין משטרת ישראל בנוגע, למשל, לקביעת גבולות לשימוש בכלי מעקב ושיטור מנבא.

עוד מומלץ להקים קבוצות ווטסאפ קהילתיות בשפות שונות (עברית וערבית), בהובלת תושבים ונציגי הרשות המקומית והמשטרה. קבוצות אלה ישמשו פלטפורמה למאבק בזמן אמת במבצעי השפעה, דיסאינפורמציה ושמועות שעלולות להסלים מתחים, מתוך שיתוף פעולה אזרחי. יוזמות דומות כבר יושמו במסגרות של "קהילות בטוחות" בטורונטו (Safe TO).[528]

גישה זו של שיתוף והכלה של מיעוטים בתהליכי הטמעת מדיניות טכנולוגיה אינה רעיון תאורטי. בניו זילנד, תהליך ההסדרה של טכנולוגיות מעקב ברשויות מקומיות כלל יועצים מקרב המיעוט המאורי, לא רק כדי לתרגם את מסמכי המדיניות לשפתם, אלא כדי לוודא שהשיח הרגולטורי מתיישב עם ערכי הקהילה, כמו כבוד, שמירה על פרטיות משפחתית והסכמה מדעת גם ברמה הקולקטיבית. במסגרת פרויקט אחר, Public Service AI Framework, הממשלה הניו זילנדית חייבה ייצוג של המאורים בתהליכים הנוגעים לעיצוב ופיקוח של מערכות בינה מלאכותית בשירותים ציבוריים ובניהול מערכות רווחה באמצעים אוטומטיים.[529] במסגרת ארגון Te Whatu Ora Waitematā, שירות הבריאות

528 *SafeTO: A Community Safety & Well-Being Plan*, City of Toronto website

529 Megan Morris, *New Zealand Government Forges Path to Responsible AI with New Framework*, Global Government Forum (Fab. 13, 2025)

האזורי של אוקלנד, נבנה גוף פיקוח אתי שבו נציגים מאורים ומשפחות מטופלים משתתפים, יחד עם יועצים אקדמיים ומקבלי החלטות, בבחינת יישום של כלי AI רפואיים. לפי דוח שהתפרסם, התהליך כלל בניית כללים אתיים המושתתים על tikanga Māori (תרבותם ומנהגיהם של בני שבט המאורי) ועקרונות של ריבונות נתונים (data sovereignty), איתור פערי גישה ושקיפות חוצה שפה בתחום קבלת החלטות אלגוריתמיות.[530]

גישה זו יכולה לשמש השראה למדיניות דומה בישראל: כזו שמכירה בלגיטימיות של חששות קהילתיים, משלבת ידע מקומי ומקיימת שותפות בעיצוב, בקרה והטמעה של טכנולוגיות שיש להן השפעה על ביטחון אישי, פרטיות, שירותים ציבוריים ודימוי אזרחי.

ג3. עידוד גיוון דרך תהליכי רכש ומכרוז

קידום מדיניות מכרזים ממשלתיים המחייבת כל חברה המבקשת למכור מוצרים לרשויות שלטוניות בתחומים רגישים כמו חינוך, רווחה או אכיפת חוק, להציג תוכנית מפורטת לגיוון אתני בשלבי התכנון והאפיון של המוצר שהיא מבקשת למכור, ובכלל זה הגדרת הבעיה, בניית מסדי הנתונים, קביעת מטריקות דיוק והערכת סיכונים. בארצות הברית, למשל, עמותת Data for Black Lives[531] הובילה מהלכים דומים בקרב קהילות אפרו־אמריקאיות, מתוך הכרה שכלים המפותחים ב"חדרים לבנים" ייכשלו במתן מענה הוגן כשיושמו על מיעוטים.

מוצע אפוא לעגן את החובה להקים צוותים מייעצים בתהליכי רכש ובמכרזים ממשלתיים של מערכות שיש בהן יכולת להשפיע על זכויות של אזרחים ערבים. צוותים אלו יכולים לכלול מומחים טכנולוגיים מן החברה הערבית, נציגי רשויות מקומיות, ארגוני חברה אזרחית, אנשי חינוך ודת וטכנולוגיה, והם יידרשו לספק חוות דעת מחייבת בשלבי האפיון, הפיילוט וההטמעה של מערכות בינה מלאכותית בקרב רשויות השלטון.[532] כך יובטח

530 Robyn Whittaker et al., *An Example of Governance for AI in Health Services from Aotearoa New Zealand*, 6 NPJ Digit. Med. 164 (2023)

531 *Home*, Data for Black Lives

532 פרל ושוורץ אלטשולר, לעיל ה"ש 58; "לראשונה בישראל: מדריך לשימוש אחראי בבינה מלאכותית (AI) במגזר הציבורי", הודעת עיתונות ופנייה להערות הציבור **מערך הדיגיטל הלאומי** (3.6.2025); גדי פרל, תהילה שוורץ אלטשולר וריטה גולשטיין גלפרין **חוות דעת על טיוטת המדריך לשימוש אחראי בבינה מלאכותית במגזר הציבורי** (המכון הישראלי לדמוקרטיה, 2025).

שלחברה הערבית תהיה גישה אפקטיבית לצומתי השפעה, ולא רק כהצגה סמלית לאחר מעשה.

כדי לעודד אחריות אתית של המגזר הפרטי, מוצע ליצור מסלול תמריצים רגולטוריים, בין בצורה של נקודות זכות במכרזים ציבוריים או במתן הקלות רגולטוריות באמצעות ארגזי חול,[533] לחברות טכנולוגיה שיבצעו בדיקות השפעה על מיעוטים (community impact assessments/diversity impact statement) בהתמקדות במגזר הערבי.

למדיניות כזאת תוכל להיות השפעה עקיפה גם על המגזר הפרטי. בעבר, חברות כמו מיקרוסופט וגוגל ניסו להחיל מנגנונים של פיתוח מוצרים מכליל (inclusive product development) המחייבים שילוב של צוותים רב־תרבותיים, בודקי מוצרים מקהילות מוחלשות, ואף ועדות היגוי חיצוניות שכוללות נציגים מקבוצות סיכון. הבעיה הייתה כפולה: ראשית, לא היה תמריץ של ממש, מעבר להיבט היחצני ביוזמות כאלה. שנית, השפעתן לא הייתה גדולה. לכן למשל, עקרונות design justice,[534] שפותחו על ידי רשת Design Justice Network, קוראים למקם את קהילות המטרה במרכז תהליך העיצוב, מתוך הכרה בכך ש"שום עיצוב איננו ניטרלי", ולא רק להעסיק צוותים מגוונים אלא להבטיח שלצוותים הללו תהיה גם יכולת פעולה עצמאית וממשית להשפיע על החלטות טכנולוגיות מהותיות.

ד. התחשבות בקהילות מיעוט בעת עיצוב כלי בינה מלאכותית ובפיקוח ואודיטינג עליהם

העיקרון של "תכנון מתוך גיוון" (diversity by design) מתבסס על ההכרה כי טכנולוגיות, ובמיוחד מערכות בינה מלאכותית, אינן ניטרליות, אלא משקפות את הנחות היסוד והערכים של מפתחיהן. לכן, אם המפתחים שייכים לקבוצת רוב הומוגנית, יש סיכון שהמערכת תייצר הטיות נגד קבוצות שאינן מיוצגות בתוכה, בין שמדובר במיעוטים אתניים, נשים, אנשים עם מוגבלות או קבוצות שוליים אחרות. זהו אחד העקרונות שנולדו בעשור האחרון מתוך מאבקים חברתיים, ביקורת על אפליה טכנולוגית וניסיונות כושלים לתקן בדיעבד מערכות מוטות. הרעיון שואב השראה ממסורת של עיצוב מכליל ועיצוב רגיש לערכים אך

533 Michael Sierra, *Regulatory Sandboxes: Fintech as a Case Study*, *in* Handbook on Government Use of Artificial Intelligence (Edward Elgar Press, 2026)

534 *Design Justice Network Principles*, Design Justice Network (summer 2018)

מרחיב אותן בדרישה **מראש** למגוון כחלק בלתי נפרד מתהליך הפיתוח ולא כתיקון מאוחר. עיקרון זה כבר הועלה במסמך היסוד של עקרונות הבינה המלאכותית של ה־OECD, שקובע כי על מערכות בינה מלאכותית לתמוך בהכללה, שוויון וגיוון, ונכלל מאז כמעט בכל מסמך מדיניות גלובלי בתחום, למשל בהמלצות אונסק"ו לאתיקה של בינה מלאכותית (UNESCO AI Ethics Recommendation, 2021) מוזכר עיקרון זה כחלק מתפיסה כוללת של "תכנון רגולציה משותף".

העיקרון נטוע בתפיסת עולם רחבה של צדק טכנולוגי (technological justice), שלפיה עיצוב מערכות הוא גם עיצוב של כוח, ולכן עליו להיעשות בשיתוף מי שהמערכת תשפיע עליהם בפועל. תפיסה זו עוגנה בעבודות כמו *Automating Inequality* של וירג'יניה יובנקס (Eubanks, 2018),[535] החושפת כיצד מערכות דיגיטליות סייעו בשעתוק אי־שוויון כלכלי וגזעי בארצות הברית; ובספר *Race After Technology* של רוחה בנג'מין (Benjamin, 2019),[536] שמציעה את המושג "New Jim Code" כדי לתאר את הדרך שבה טכנולוגיות ממשיכות לקבע אפליה בשם הניטרליות. יוזמות של ארגונים כמו AI Now Institute הדגישו כי ביסוס אמיתי של גיוון מחייב מעבר ממדיניות גיוס כוח אדם מגוונת, אל שותפות של קהילות מיעוט בעיצוב, תיעדוף והערכה של טכנולוגיות.

בפועל, עקרון ה"תכנון מתוך גיוון" בא לידי ביטוי ביוזמות יישומיות מגוונות. ב־Partnership on AI פותחו כלים לזיהוי וניטור של נזקים פוטנציאליים למיעוטים כחלק מתהליך הפיתוח.

הדרישה לכלול גיוון אתני בשלבי התכנון של מערכות AI לא רק ברמת ההעסקה משתקפת ביוזמות כמו של Data for Black Lives בארצות הברית, שם נעשה ניסיון מגובש לשלב פעילים, מדענים ומנהיגים קהילתיים כבר בשלבי האפיון.[537] היוזמה משלבת מדעני נתונים ואקטיביסטים במאמץ להפוך דאטה לכלי של שינוי חברתי ולא של שעתוק אפליה, מפתחת

535 ראו באופן כללי Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor (St. Martin's Press, 2018)

536 ראו באופן כללי Ruha Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code (1st ed., Polity, 2019)

537 Rahul Barghava, *Automating (In)Justice: Policing and Sentencing in the Algorithmic Age*, MIT Center for Civic Media (Nov. 18, 2017)

מדדי צדק ומארגנת קהילות פעולה שדורשות אחריות מצד חברות טכנולוגיה ומוסדות ציבוריים. באופן דומה, בקנדה פועלת רשת Indigenous Data Sovereignty שמבקשת להבטיח ריבונות של קהילות ילידיות על הנתונים שלהן. היא פיתחה עקרונות פעולה כמו Collective Benefit, Authority to Control, Responsibility, Ethics – CARE,[538] שמדגישים את החשיבות של שליטה תרבותית וקהילתית על תהליכי איסוף ועיבוד מידע. הארגון Black in AI קידם קמפיין גלובלי לשילוב אפרו־אמריקאים בפיתוח מערכות בינה מלאכותית תחת הסיסמה "גם אנחנו מפתחים את המודל".[539]

בשנים האחרונות עולה גם קריאה גוברת לשלב קהילות מיעוטים לא רק בעיצוב המקדמי של מערכות בינה מלאכותית, אלא גם בשלבי הביקורת והאודיטינג שלהן, מתוך הכרה בכך שהטיות אלגוריתמיות רבות נובעות מעיוורון מערכתי להקשרים תרבותיים, לשוניים וחברתיים ייחודיים המתגלים רק לאחר יישום המערכות בפועל. אחת הדוגמאות הבולטות לכך היא בניו זילנד, שם יושם פיילוט של בדיקות השפעה קהילתיות (community-based audit) במספר מחוזות. תושבים מקומיים, לרבות ממיעוטים מאוריים ופסיפיים, הוכשרו בתמיכת ה־Digital Council Aotearoa New Zealand לבצע אודיטינג על מערכות המלצה ומיון בשירות הציבורי. תוכנית ההכשרה כללה הבנת המבנה של אלגוריתמים, בחינה של אופן קבלת החלטות וזיהוי פערים או אפליה אפשרית. אחת התובנות המרכזיות הייתה שכאשר המבקרים באים מתוך הקהילה עצמה, הם מזהים ניואנסים סמויים של אפליה, למשל, כיצד שימוש במיקוד גאוגרפי עשוי להפלות תושבים באזורים כפריים המאוכלסים בקהילות מוחלשות. מעבר לכך, הנוכחות הקהילתית בתהליך הגבירה את אמון הציבור ושיפרה את תגובת המערכת הציבורית לביקורת.[540]

הדגם שפותח בניו יורק במסגרת NYC Algorithmic Accountability Task Force,[541] שתכליתו הייתה בין השאר הבטחת ייצוג אותנטי של צרכים קהילתיים בשלב ההטמעה, הוכיח את יכולתו להשפיע על תיקוף מערכות קבלת החלטות ציבוריות ויכול לשמש השראה.

538 *Indigenous Data Sovereignty & Governance*, Native Nations Institute (Univ. of Ariz.)

539 Black in AI

540 Towards Trustworthy and Trusted Automated Decision-Making, 20 (Digital Council for Aotearoa New Zealand, 2021)

541 *Automated Decision Systems (ADS) Task Force*, City of New York

בנוסף, פותחו בארצות הברית מודלים של ביקורת אלגוריתמית משתפת (participatory algorithm audits) על ידי גופים דוגמת Data & Society ו־Algorithmic Justice League.[542] כך למשל, בניו יורק קבוצת תושבים מרובע הברונקס, רובם ממוצא לטיני ואפרו־אמריקאי, שותפו באודיט של מערכת ניתוח תחזיות פשיעה (predictive policing) שפעלה באזורים שלהם. באמצעות סדנאות דיאלוג והכשרות קצרות, התושבים בחנו את האופן שבו האלגוריתם הדיר שכונות מסוימות, דרש מידע חסר או תייג פעילות תמימה כ"סיכון". תהליך זה לא רק חשף ליקויים מהותיים במודל, אלא גם הוליד המלצות ממשיות לשינוי פרמטרים ולהכנסת משתני תיקון (corrective factors).[543] אודיטינג משתף אינו לפיכך רק מנגנון אתי אלא כלי לשיפור איכות המערכת ודיוקה בפועל.

בהקשר זה אנו מציעים להטמיע חובת הכללת עיצוב ואודיטינג מגוונים בתהליך ניהול הסיכונים בעת הטמעת כלי בינה מלאכותית בממשלה. בטיוטת המדריך שגובש על ידי רשות התקשוב הממשלתית ביולי 2025,[544] וכן במחקרם של פרל ושוורץ אלטשולר,[545] מוצע לקבוע מנגנונים של ניהול סיכונים מובנה לפני הטמעה של מערכות בינה מלאכותית במגזר הציבורי. ניהול סיכונים זה כולל שלבים של זיהוי, הערכה והפחתת נזקים פוטנציאליים, בדגש על הוגנות ושמירה על זכויות אדם. עם זאת, שני המסמכים אינם מציעים בשלב זה הבחנה בין קבוצות אוכלוסייה או מנגנון מובנה להערכת ההשפעה של מערכות בינה מלאכותית על החברה הערבית בישראל. לפיכך מומלץ לשלב בתוך תהליכי ניהול הסיכונים הממשלתיים מנגנונים ייעודיים להערכת השפעות על החברה הערבית, לרבות זיהוי פוטנציאל לאפליה עקיפה, פערים טכנולוגיים, סיכוני אמון או פגיעות ייחודית הנובעת מהקשר חברתי ותרבותי. מנגנונים אלה יכולים לכלול שאלון ייעודי במסגרת הערכת ההשפעה של בינה מלאכותית (AI impact assessment), חובת התייעצות עם נציגות קהילתית או מומחים לחברה הערבית, ובחינה של השפעות מבניות דוגמת הדרה משירותים דיגיטליים או פיקוח עודף.

542 Algorithmic Justice League

543 *Participatory Algorithmic Auditing*, Sustainability Directory

544 המדריך לשימוש אחראי בבינה מלאכותית, לעיל ה"ש 532.

545 פרל ושוורץ אלטשולר, לעיל ה"ש 58.

יש מקום לעצב את תהליכי ניהול הסיכונים באמצעות גופים ייעודיים כמו המועצה לאתיקה של בינה מלאכותית במגזר הערבי או המעבדה לבינה מלאכותית שהוצעו קודם לכן. גופים אלה יוכלו להנחות את רשויות המדינה באמצעות פיתוח והפצה של מדריכים מפורטים לניהול סיכונים מכליל, אשר יתבססו על חתך של תת־אוכלוסיות בתוך המגזר הערבי או על תחום הפעולה של המערכת (בריאות, חינוך, רווחה, בטחון פנים וכדומה). מדריכים אלה יכללו כלי הערכה, דגשים אתיים, שאלות מנחות וקריטריונים לבחינת השפעה מצטברת. מהלך כזה עשוי לא רק לצמצם פגיעות אלא גם לחזק את הלגיטימציה הציבורית של מערכות בינה מלאכותית ממשלתיות ולבסס סטנדרט אחראי ורגיש להקשר.

מסמך ה־AIA (Artificial Intelligence Act) הקנדי מתמרץ חברות המשלבות בדיקות השפעה קהילתיות ייעודיות (כלומר בדיקות השפעה על קבוצות מיעוט) בהפיכת הבדיקה לחלק מתהליך החובה במהלך רכש והטמעה של כלים מבוססי בינה מלאכותית ברשויות השלטון הקנדיות.[546] חוק הבינה המלאכותית האירופי כולל גם הוא חובת הערכת השפעות ובדיקות הטיה ביחס לקבוצות מיעוט והדבר ישפיע, יש להניח, על התנהלות חברות הטכנולוגיה הגדולות.[547]

ה. יוזמות פנים־מגזריות

יוזמות פנים מגזריות יכולות לכלול הקמת "מרכז מוגנות" כמוקד סייבר מגזרי וכתובת ברורה לבני החברה הערבית, שיתמקד בזיהוי, ליווי וייעוץ למשתמשים סביב איומי סייבר, הונאות סייבר, פגיעות בפרטיות, פגיעות אחרות או חשיפה לתכנים מזיקים. מרכז כזה יכול להפעיל קו חם אנונימי לדיווח על פגיעות ולסייע בתאום עם הרשויות. בדומה למה שכתבנו בנוגע לחברה החרדית, מוצע להקים גופים משפטיים, קליניקות קהילתיות או קווים חמים שיסייעו ליחידים מהחברה הערבית המתמודדים עם פגיעויות שמקורן פנימי, כגון שיימינג דיגיטלי בקבוצות סגורות, איסוף והפצה של מידע אישי או רפואי ללא הסכמה, חסימת גישה לכלים דיגיטליים או מעקב בלתי פורמלי אחר פעולות מקוונות. גופים אלה יאפשרו

546 *Algorithmic Impact Assessment*, Government of Canada

547 Regulation 2024/1689, of the European Parliament and of the Council of 13 Mar. 2024 on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, arts. 11, 12, 13, 17, 23, 24, 43, 72, 2024 O.J. (L 168) 1, 15

לפרטים לפנות לעזרה באופן דיסקרטי ולקבל כלים להגן על זכויותיהם מתוך איזון עדין בין הגנה על ערכי הקהילה לבין עקרונות יסוד של חירות וכבוד האדם.

ה.1. יוזמות פנים־קהילתיות נוספות יכולות לכלול הכשרת סוכני שינוי קהילתיים

למשל, הקמת קבוצות של צעירים ערבים, תלמידי תיכון, סטודנטים ומתנדבים, שישמשו סוכני שינוי בקהילה, היא דרך יעילה שהוכחה בעולם (למשל, במיזם teens4tech)[548] לקידום סוכני שינוי: מדריכים להורים, סבים וסבתות ומבוגרים אוריינים פחות. הם יעבירו סדנאות קצרות במרכזים קהילתיים, במוקדי שירות או בבתים פרטיים, ויסבירו כיצד להשתמש בזהירות במנועי חיפוש, בצ׳אטים חכמים ובאפליקציות רפואיות או כלכליות מבוססות בינה.

כפי שהערנו למעלה, המנהיגות הקהילתית בחברה הערבית זקוקה למתווכים אנושיים שיבינו הן את העולם הטכנולוגי והן את המרקם החברתי המקומי. ניתן להקים מסלול הכשרה קהילתי ל"מתרגמי בינה", בדומה לקורס "מובילים דיגיטליים" שקיים היום עבור מנהלים בשירות הציבורי.[549] המסלול צריך להיות מותאם באופן ייחודי לחברה הערבית, ובוגריו יהפכו למי שיהיו אמונים להסביר להנהגה המקומית, לאנשי חינוך ולאנשי דת כיצד מערכות AI פועלות, ואיך ניתן לפעול בצורה אחראית מולה. הם ישמשו גם משקיפים ביקורתיים ויוכלו להזרים נתונים רלוונטיים למעבדת המחקר.

נוסף על כל אלה, כדי לעגן את עקרון הבקרה העצמאית, מוצע להקים "מועצה אתית טכנולוגית" של החברה הערבית בישראל שתפעל באופן עצמאי ותפקידה יהיה לפרסם דוחות ביקורת, להגיש חוות דעת רגולטוריות ולהשפיע על מדיניות ציבורית בתחומי הבינה המלאכותית. המועצה תכלול משפטנים, אנשי חינוך, אנשי טכנולוגיה ומובילי דעת קהל. ניתן לשקול להקים אותה במסגרת אקדמית, למשל במוסדות מובילים דוגמת המרכז הערבי למשפט ומדיניות או האקדמיה הערבית לחינוך, מתוך שאיפה לשמר את חופש הפעולה ואת יוקרתה הציבורית. המעבדה לבינה מלאכותית שהוצעה קודם לכן יכולה להשתלב בתוך המועצה. הרעיון של גופים עצמאיים שאמורים לשמש משקל נגד לממשלה ולתעשייה כדי

548 *AI4Good Incubator: Teens in AI Empowers Youth to Innovate for Climate Action*, Teens in AI

549 ראו למשל את הקול קורא של "עמותת מובילים דיגיטליים".

לייצג קהילות מוחלשות עומד ביסוד הקמת ה־AI Now Institute, שהפך לגורם מומחיות מוביל בתחום.[550]

יש גם לעודד ולתמוך בהעצמה של היכולות המשפטיות לנטר, לגלות, לתעד ולחשוף מקרים שבהם טכנולוגיות מבוססות בינה מלאכותית פוגעות בזכויות של אזרחים ערבים ולהפעיל כלים משפטיים מתאימים, כגון עתירות לבג״ץ ותובענות ייצוגיות. כלים משפטיים אלה הוכחו בעבר כאפקטיביים בישראל, והם יסייעו בהפיכת החברה הערבית מגורם נפגע לגורם תובע ומוביל שינוי.[551] בארצות הברית, עבודתה המשפטית של ג׳וי בואולמוויני (Buolamwini) דרך ארגון Algorithmic Justice League,[552] בהקשרים של אפליה אלגוריתמית נגד שחורים, וכן פעילותה של האגודה לזכויות האזרח האמריקאית בתחומי אפליה אלגוריתמית,[553] זיהוי פנים ומעקב אחר מיעוטים אתניים, הוכחו כאפקטיביים בהגבלת כוחן של רשויות השלטון בשימוש בכלים שהוכח לגביהם כי הם מפלים.

אפשרות נוספת היא להקים קרן לפיתוח מיזמים טכנולוגיים שיפעלו לצמצום סיכונים, תיקון עיוותים מערכתיים והנגשה שוויונית של טכנולוגיות בינה מלאכותית. הקרן תתמרץ יזמים ויזמיות מהחברה הערבית לפתח פתרונות מבוססי בינה מלאכותית, מתוך רגישות תרבותית וקהילתית, בתחומים כגון חינוך, בריאות, תחבורה, דיור ורווחה. הקרן תציע מלגות קצרות טווח וכן תמיכה ארוכת טווח בפרויקטים חדשניים בדגש על בינה מלאכותית אחראית, כגון:

- מערכות סינון תוכן רב־לשוניות אוטומטיות שיאומנו על מאגרי דאטה מייצגים של החברה הערבית, ישמשו לאימון מודלים נטולי הטיה ויבינו הקשר מקומי.
- צ׳אטבוטים מכווני תרבות החברה הערבית בישראל שיכולים לתווך מידע מוסדי או משפטי בשפה נגישה, נאמנה למקור, אך גם מותאמת תרבותית; לדוגמה, הסבר על זכויות ברווחה או תרגום של מידע רפואי מורכב.

550 ראו באתר AI Now Institute

551 "בג"ץ: יש להסדיר בחוק את פעילות מערכת "עין הנץ" של המשטרה, האגודה לזכויות האזרח (28 בינואר 2021)" האגודה לזכויות האזרח בישראל.

552 *About the AJL*, American Justice League

553 ראו מידע על פעולתה באתר האגודה American Civil Liberties Union (ACLU)

• פיתוח מערכת אודיטינג למנגנוני שירות לקוחות שתבדוק אם מערכות קוליות ומענה אוטומטי מגיבות כראוי גם לדוברים בערבית ספרותית או מדוברת; לדוגמה, מול מוסדות כמו ביטוח לאומי, קופות חולים או בנקים.

• אפליקציה קהילתית לאיתור פערים בתשתיות ציבוריות בעזרת בינה גאוגרפית (geo-AI), לזיהוי אוטומטי של אזורים בעלי תשתית או תחבורה ירודות באזורים גאוגרפיים ספציפיים.

• פיתוח מערכת למידה מותאמת אישית (adaptive learning) שמכירה במאפיינים תרבותיים ולשוניים של תלמידים ערבים, ותכלול ייצוג תוכן היסטורי, תרבותי וויזואלי מותאם.

הקרן תוקם בשותפות בין הממשלה (המשרד לשוויון חברתי, רשות החדשנות ומערך הדיגיטל הלאומי); פילנתרופיה בינלאומית (קרנות העוסקות בזכויות אדם, חדשנות טכנולוגית והוגנות אלגוריתמית, כגון Ford Foundation, Mozilla Foundation) והמגזר העסקי (חברות טכנולוגיה בינלאומיות וישראליות שיפעלו לפי עקרונות ESG ויעדיפו השקעות תיקון חברתי).

הקרן תפעיל קול קורא שנתי לתמיכה במיזמים, עם ועדת שיפוט הכוללת אנשי טכנולוגיה, נציגים מהחברה הערבית, מומחי אתיקה דיגיטלית ורגולטורים; תינתן העדפה למיזמים שתוכננו בשותפות עם קהילה מקומית או נועדו לפתרון בעיה ממשית מובחנת עבור החברה הערבית בישראל; הקרן תציע לא רק מענקים כספיים, אלא גם ליווי מקצועי בתחומי עיצוב חוויית משתמש, ניהול דאטה, עקרונות הוגנות ובדיקות אימפקט חברתי; יוקם אשכול רגולטורי מייעץ שיסייע למיזמים לעמוד בחוקי פרטיות, שקיפות, נגישות שפתית וניהול סיכונים אתיים. מודל הקרן יכול לשמש גם פיילוט אזורי, למשל בגליל או במשולש, שבסופו תיבנה מתודולוגיה לאומית.

סיכום

פרק זה ביקש למפות את הפגיעויות הקיימות והחדשות הנובעות משימושים בבינה מלאכותית בחברה הערבית בישראל, החל בהדרה שקטה מנתוני אימון של מודלים, דרך

חוסר יכולת להבין החלטות אוטומטיות בשפה לא נגישה, ועד חוסר ייצוג של נקודת המבט הערבית בתהליכי עיצוב ואודיטינג של מערכות.

הצגנו מגוון רחב של המלצות, רגולטוריות, מוסדיות, חינוכיות וקהילתיות, שנועדו לבסס עקרונות של צדק טכנולוגי ואוריינות מותאמת. הדגשנו את הצורך במנגנוני אכיפה ובקרה שיבחנו את ההשלכות הספציפיות של מערכות AI על המגזר הערבי; את החשיבות שבהנגשה לשונית של החלטות אלגוריתמיות; ואת ההזדמנות הטמונה ביזמות טכנולוגית ערבית כחלק בלתי נפרד מהפתרון, לא רק כצרכנים, אלא כמעצבים של עתיד דיגיטלי בטוח והוגן.

פרק אחד עשר

מנגנוני התערבות לקידום מוגנוּת של זקנים בעידן הבינה המלאכותית

מבוא

כלי בינה מלאכותית הם מעין חרב פיפיות עבור אוכלוסיית הזקנים, עם פוטנציאל סיוע מצד אחד וחשש מפני חשיפה לסיכונים חדשים מצד שני. מערכות AI בתחום הבריאות או הניידות עשויות לחזק עצמאות, אך גם עלולות להפקיע שליטה, אם אינן שקופות או ניתנות להבנה. מערכות לזיהוי נפילות או ניטור רפואי בבתים עשויות לספק "מעקב מיטיב", שיש בו הגנה אך גם פגיעה בפרטיות. רובוטים חברתיים ותומכי שיחה עשויים להתמודד עם תחושת בדידות, ולחלופין למסך אותה באופן מטעה.

בפרק זה ננתח את אתגרי המוגנות של אוכלוסיית הזקנים בעידן של בינה מלאכותית וננסה להתוות דרכי פעולה ראשוניות להתמודד עם אתגרים אלה. נבקש להתמקד הפעם דווקא

באתגרי המוגנות החדשים המתעוררים עם כניסתן של טכנולוגיות בינה מלאכותית חכמות, מסתגלות ואוטונומיות אל תוך חייהם של בני הגיל השלישי. להבדיל מהמדיום הדיגיטלי המסורתי, אשר לרוב פעל ככלי תקשורת או מידע סטטי, מערכות מבוססות בינה פועלות לעיתים קרובות כסוכנים פעילים בקבלת החלטות, תיווך רגשי או ניהול שירותים אישיים. זוהי קפיצת מדרגה בעוצמת ההשפעה, וכפועל יוצא, ברמת הסיכון. מערכות רבות פותחו ונוסו דווקא בהקשרים של זקנה, למשל להפגת בדידות, ניטור בריאותי, סיוע קולי או חבר וירטואלי, אך עדיין נדרשת בחינה רגישה ומעמיקה של סוגיות המוגנות הייחודיות למפגש זה.

זקנים נחשבים כבר שנים רבות לקבוצת סיכון בעולם הדיגיטלי, בין בשל אוריינות דיגיטלית נמוכה, ירידה ביכולות הקוגניטיביות, או נטייה לסמוך על מקורות מידע ללא תיווך ביקורתי. תופעות אלו תועדו ונחקרו בעשור האחרון על ידי קובעי מדיניות, ארגוני זכויות ומכוני מחקר, שהציעו מגוון של דרכי התערבות להגברת נגישות, קידום אוריינות מותאמת גיל וחיזוק מנגנוני הגנה. המודעות לפגיעות של זקנים ברשתות החברתיות, בפרסום ממוקד ובשירותים דיגיטליים נמצאת בעלייה, וניתן לומר כי ההתמודדות עם סיכונים אלה כבר צברה תשתית של ידע, תובנות ויישומים.

גם שילוב הבינה המלאכותית בטיפול ובחיי היומיום בגיל הזקנה אינו חדש, וידע נצבר סביב שאלות אתיות של תחליף לקשר אנושי, פיקוח רובוטי או השפעה פסיכולוגית של אינטראקציה עם מכונה. כך למשל, חוקרים הזהירו מפני גילנות מערכתית שבאה לידי ביטוי בנכונות גבוהה יותר ליישם פתרונות טכנולוגיים בקרב זקנים לעומת קבוצות גיל אחרות. קשה לדמיין רובוט מחליף גננת בפעוטון, אך קל יותר לעכל רובוט הומנואידי בבית אבות. הבחירה להחליף קשר אנושי במענה טכנולוגי לעיתים אינה רק שאלה של נוחות, אלא גם של ערכים, ובעיקר של הבניית דימוי ציבורי של הזקנים כקבוצת יעד לגיטימית להטמעת טכנולוגיות. עם זאת, אנו מבקשים להבחין בין שיח אתי־פילוסופי זה לבין שיח המוגנות. הדגש בפרק זה אינו על שאלת הראוי הכללית, אלא על זיהוי סיכונים הנוצרים מתוך המפגש בין זקנה לבין תצורות חדשות של בינה מלאכותית ולגזירת משמעויות אופרטיביות להתערבות.

חלק ראשון: מאפיינים חברתיים ותרבותיים של זקנים

א. חיים במסגרת משתנה, בתקופת חיים משתנה

פער הגילים של תקופת הזקנה הופך את הכתיבה על אודותיה למורכבת. בדיוק כמו ההבדלים בין ילדים צעירים לבני נוער ומתבגרים, כך גם כאן קיים הבדל בין זקנים בגיל הפרישה לכאלה שנמצאים בשנות השמונים והתשעים לחייהם. עם זאת, במסמך זה לא נוכל להתייחס לכל אחת מקבוצות הגיל האלה ונסתפק בקריאה למחקר נוסף בתחום.

המעבר מהעבודה לגיל הפרישה מלווה לעיתים במשבר זהות: תפקידים כמו מפרנס ותורם מוחלפים בתחושת חוסר תכלית וקיום פסיבי. מצב זה מוביל לעיתים לדיכאון או בידוד. אכן, בקרב זקנים בבתי אבות בישראל נמדדה עלייה בתחושת הבדידות והפחתה בתפיסת הערך עצמי.[554] התרבות המערבית הדוגלת בנעורים ומציבה עבודה ופרנסה כערכים מרכזיים, עלולה להוביל להדרה של הקול והניסיון של המבוגרים. היא עשויה במקביל להוביל לגילנות עצמית ודימוי נמוך בקרב זקנים, ואלה, בתורם פוגעים במסוגלות להתנסות בטכנולוגיה או לבקש עזרה.

בישראל תוחלת החיים גבוהה, וקיימת מגמה של שנות זקנה רבות לאחר גיל הפרישה.[555] עם זאת, המסגרות החברתיות, הרפואיות והמשפטיות אינן ערוכות לתמיכה מלאה באורך החיים הזה, במיוחד עבור מי שידו אינה משגת והוא חסר רקע משפחתי תומך או משאבים כלכליים מספקים. תוחלת החיים הארוכה מלווה לעיתים בירידה הדרגתית בתפקודים קוגניטיביים, מוטוריים וחושיים, אשר מקשה על תפעול עצמאי של מערכות טכנולוגיות מורכבות, לרבות שירותי בריאות דיגיטליים, מערכות מידע פיננסי, ואפילו יישומים פשוטים לתקשורת עם קרובים. השברירירות הגופנית, מחלות כרוניות, תלות במידע רפואי ובמכשירים דיגיטליים, יחד עם תחושת דחיפות במצבים רפואיים, יוצרות קרקע פגיעה למניפולציות, למידע מטעה ולבחירה בשירותים שאינם מוסדרים או מותאמים. תחושת

554 "מגפת הבדידות בקרב קשישים – 1 מכל 3 ערירי בגיל השלישי" **קו למעסיק**.

555 "לקט נתונים לרגל יום האישה הבין־לאומי 2024 (1 במרץ 2024)" **הלשכה המרכזית לסטטיסטיקה**.

הבידוד עלולה להוביל גם לתפיסה מוטעית של המרחב הדיגיטלי כתחליף מלא לקשר אנושי, אף כשזה אינו מספק מענה רגשי הולם. לפי מחקר של מכון ברוקדייל, פירוק קהילות מסורתיות ושינויים דמוגרפיים יוצרים חשיפה גבוהה מבעבר לבדידות ולהיעדר רשתות תמיכה באוכלוסייה זו.[556]

זקנים עשויים להימצא במצב של תלות בבני משפחה, במוסדות טיפוליים ובמערכות הבריאות והרווחה. תלות זו נובעת לעיתים מהיחלשות יכולותיהם הפיזיות או הקוגניטיביות, אך לעיתים גם ממדיניות מוסדית אשר מכתיבה פתרונות אחידים ואינה מותאמת דיה לצרכיו האישיים והייחודיים של כל זקן. התוצאה היא פגיעה בזכותו של הזקן לאוטונומיה אישית, וליכולתו לקבוע את אורחות חייו ולהיות שותף פעיל בקבלת ההחלטות הנוגעות לו.[557] לצד זאת, גם הסתמכות על בני משפחה או קרובים יוצרת פגיעות, במיוחד כאשר אין בקרה על איכות ההכוונה, כאשר מתקיים פער בין טובת הזקן לבין פרשנות סובייקטיבית של טובתו, או כאשר גם בני המשפחה אינם אוריינים דיים.

התחושה של תלות גוברת, בייחוד כאשר היא מלווה בחשש להתעללות, הזנחה או פטרנליזם מופרז יוצרת מציאות שבה הזקן נדרש לוותר על מידה ניכרת מהשליטה בחייו. על כן, גוברת הקריאה בקרב חוקרים ומומחים לקדם גישות של "אוטונומיה נתמכת", המשלבות הכרה בזכויות הזקן עם ליווי אישי שמאפשר לו להבין את האפשרויות העומדות בפניו ולקבל החלטות מתוך מידע והסכמה חופשית. מודלים אלה מבקשים להחליף את התפיסה הפטרנליסטית בגישה משתפת, שבה הזקן הוא סוכן פעיל בעיצוב חייו גם כאשר מצבו הרפואי או הנפשי מורכב.[5]

ב. רב־תרבותיות והשפעת מוצא אתני על זקנה בישראל

אוכלוסיית הזקנים מורכבת מתתי־קבוצות שונות שרמת הפגיעות שלהן משתנה. הפגיעות כמובן מועצמת בקרב אוכלוסיות מוחלשות, כגון דוברי שפה שונה מהשפה הרשמית ועולים חדשים, ערבים, מי שהיו עניים לכתחילה, מי שהם חסרי תמיכה משפחתית ומי שתלויים באפוטרופסים ומוסדות. חלק מאוכלוסיית הזקנים בישראל היא מהגרת, עם רקע תרבותי

556 יואה שורק וצסלב קונסטנטינוב "שינויים במבנה המשפחה: מגמות, סוגיות והשפעות על עבודת מערכת הרווחה" (מכון ברוקדייל, 2021).

557 alexndra Charette, *Autonomy and the Governance of "Ageing in Place"*, Social Policy and Society, 1–14 (2025)

משתנה, שפה שונה וצרכים ייחודיים. חברי אוכלוסיות אלה חווים קשיים בגישה לשירותים שונים עקב הבדלים תרבותיים ושפתיים ועצם היותם זקנים מקשה עליהם עוד יותר.

ג. זקנים וטכנולוגיה: פערים דיגיטליים והכלה דיגיטלית

מערכת היחסים בין זקנים לטכנולוגיות מידע היא כפולה. מצד אחד, התמודדות עם מחלות רבות ותלות בשירותים ובמכשירים רפואיים; וצורך במיצוי זכויות כגון קצבאות מסוגים שונים – יוצרים תלות גבוהה במערכות המידע של הגופים המספקים שירותים וכן במנועי חיפוש ופורומים מקוונים למציאת פתרונות. מצד שני, חסמים פסיכולוגיים וחסמי דימוי עצמי ביחס לטכנולוגיה, משפיעים על אי־רצון של מבוגרים לנסות סוגים שונים של מכשירים ושירותים טכנולוגיים;[558] שימוש לקוי מצד זקנים מוביל לחשיפה למידע מוטעה ומניפולטיבי; ואפליה מצד נותני שירותים ופלטפורמות חושפת זקנים לנזקים ולסכנות.[559]

כאשר הירידה הקוגניטיבית מחריפה, למשל במקרים של דמנציה או פגיעה בזיכרון ובעיבוד מידע, היכולת להבין את המידע שאליו הם נחשפים, לשפוט את אמינותו ולבחור באופן מושכל יורדת משמעותית. תלות בבני משפחה העשויים להיות מלאי כוונות טובות אך אינם אוריינים דיים בעצמם יכולה להפוך גם היא למקור סיכון, וכך גם המעבר לשהות ארוכה יותר בתוך הבית ותפיסתו כמרחב בטוח העלולה לפגוע בהבנה של סיכונים המגיעים אל תוך הבית אבל מתרחשים במרחב המקוון.

מאפיינים אלו עלולים להחמיר כאשר קיימת גם אוריינות דיגיטלית נמוכה, מגבלות שפה או חוסר הבנה של ההבדל בין ממשק אנושי לבין ממשק אוטומטי, מה שמוביל לבלבול, חוסר אמון ולעיתים להימנעות מהשימוש בשירותים דיגיטליים חיוניים. סקר שערך איגוד הגמלאים האמריקאי בשנת 2024 מגלה דאגה בקרב קשישים לגבי השימוש לרעה בבינה מלאכותית לצורך גניבת זהות והונאות. 87% מהמרואיינים היו מודאגים מפריצת סיסמאות, 84% ממתקפות פישינג ממוקדות, 85% מ"דיפ־פייק", 84% משיבוט קול ו־83% מזהויות סינתטיות. יתר על כן, 77% מהקשישים חששו שיהפכו אישית למטרות של

558 Qingyu Zhang, Salman Khan, Safeer Ullah Khan et al., *Unraveling the Barriers Contributing to the Seniors Travelers' Non-adoption Intention of Virtual Reality*, 36 J. Aging & Soc. Pol'y 123 (2024)

559 ראו למשל Ifeoma Ajunwa, *Age Discrimination by Platforms*, 40(1) Berkeley Journal of Employment and Labor Law, 1 (2019)

הונאות הקשורות ל־AI בעתיד.[560] מחקרים נוספים מראים כי זקנים פגיעים במיוחד למידע כוזב ברשת,[561] ואחת הסיבות המרכזיות לכך היא אוריינות המדיה הנמוכה שלהם, והפער הדיגיטלי לעומת גילים צעירים יותר.[562] נתונים כאלה קיימים גם בישראל.[563] לאור זאת, ארגונים וגופים שונים החלו להכיר בחשיבות וכן ביעילות של פיתוח הכשרות אוריינות המותאמות לאוכלוסייה זו, הכוללות למשל התמקדות בשימושי הרשת הנפוצים בקרבה או ביצוע הדרכות פרונטליות או מקוונות.[564]

המעבר לעולם הדיגיטלי הביא איתו אימוץ של מדיניות "דיגיטל תחילה" בקרב מוסדות נותני שירותים רבים, באופן שפוגע בזקנים. עניין זה עורר יוזמות כגון דרישה מכל גוף המעניק שירות חיוני, למשל בענייני בריאות או מיצוי זכויות, לספק גם אפשרות לא מקוונת או אפשרות המשלבת תמיכה אנושית, כך ששום אזרח ותיק לא יישאר ללא מענה. יוזמות אלה מסתמכות על התפיסה של הכלה דיגיטלית כזכות בסיסית וחלק בלתי נפרד ממוגנות

560 Alicia R. Williams, *Older Adults Express High Concern and Limited Knowledge About AI Scams and Fraud: 2024 Survey on Artificial Intelligence-Involved Fraud*, AARP Rsch. (Dec. 31, 2024)

561 Nir Grinberg, Kenneth Joseph, Lisa Friedland et al., *Fake News on Twitter During the 2016 U.S. Presidential Election*, 363(6425) Science, 374-378 (2019); Andrew Guess, Jonathan Nagler, & Joshua Tucker, *Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook*, 5(1) Science (2019); Hyunjin Seo, Matthew Blomberg, Darcey Altschwager et al., *Vulnerable Populations and Misinformation: A Mixed-Methods Approach to Underserved Older Adults' Online Information Assessment*, 23(7) New Media & Society, 2012-2033 (2020)

562 Nadia M. Brashier & Daniel L. Schacter, *Aging in an Era of Fake News*, 29 Current Directions in Psychological Science, 316-323 (2020); Eszter Hargittai, Anne Marie Piper, & Meredith Ringel Morris, *From Internet Access to Internet Skills: Digital Inequality Among Older Adults*, 18 Universal Access in the Information Society, 881-890 (2019); Seo et al., לעיל ה"ש 561.

563 **מאפיינים וחסמים לשימושים ופגיעות ברשת בקרב אוכלוסיית הגיל השלישי, סקר איגוד האינטרנט הישראלי** (איגוד האינטרנט הישראלי, 2024).

564 Päivi Rasi, Hanna Vuojärvi, & Susanna Rivinen, *Promoting Media Literacy Among Older People: A Systematic Review*, 71(1) Adult Education Quarterly, 37-54 (2020); Ryan C. Moore & Jeffrey T. Hancock, *A Digital Media Literacy Intervention for Older Adults Improves Resilience to Fake News*, 12(1) Scientific Reports, 6008 (2022)

זקנים.[565] למשל, מסמך ההנחיה *Ensuring access to information and services in a digital age* בוויילס מחייב רשויות מקומיות וספקי בריאות להבטיח שאנשים מבוגרים יוכלו לקבל מידע ושירותים גם באמצעות ערוצים לא מקוונים כגון טלפון ופניות פיזיות, באותה רמת שירות כמו ערוצים דיגיטליים.[566] כך הציע גם הפורום האירופי למוגבלויות, שקרא לחקיקה באיחוד האירופי שתחייב גופים ציבוריים ופרטיים המספקים שירותים מקוונים להבטיח אלטרנטיבות לא דיגיטליות ברמה שווה לאלה הדיגיטליות.[567] גם בישראל נעשו מאמצים על ידי מערך הדיגיטל הלאומי ובמסגרת חוק הנגשת השירותים בהקשר זה.

חלק שני: סוגי פגיעות של זקנים בעידן הבינה המלאכותית

כמו קבוצות אחרות, קבוצת הזקנים חשופה לפגיעויות פיזיות, קוגניטיביות ורגשיות בגלל מערכות בינה מלאכותית היוצאות משליטה או שנעשה בהן שימוש זדוני. ואולם בחלק זה נעמיק בשלושה מרחבים של פגיעות הנוכחים דווקא בעידן הבינה המלאכותית: האחד הוא הדרה של קבוצת הזקנים מתהליכים הקשורים באימון מודלים, בתוצרים שלהם ובמכשור טכנולוגי; השני הוא סיכוני היקשרות למערכות חכמות הנתפסות כאמצעי יעיל לטיפול בזקנים; והשלישי הוא חשיפה למניפולציות והונאות, שהפכה נפוצה מאוד דווקא ביחס לאוכלוסיית הזקנים.[568] בהקשר זה נדגיש כי כבר בדוח של ארגון הבריאות העולמי משנת 2022 גילנות נתפסת כגורם סיכון מערכתי שמחלחל לכל שלבי הפיתוח והשימוש

565 *Ensuring Access to Public Services for Older People in Digital Societies*, European Union Agency for Fundamental Rights (Mar. 22, 2023)

566 *Ensuring Access to Information and Services in a Digital Age: Guidance for Local Authorities and Health Boards*, Older People's Comm'r for Wales (Nov. 18, 2021)

567 *EDF Resolution: Right to Offline Access to Essential Services*, European Disability Forum (Nov. 13, 2024)

568 Charlene H. Chu, Rune Nyrup, Kathleen Leslie et al., *Digital Ageism: Challenges and Opportunities in Artificial Intelligence for Older Adults*, 62(7) The Gerontologist, 947–955 (Sept. 2022)

בטכנולוגיות מבוססות בינה מלאכותית, בעיקר בתחום הבריאות.[569] במובן זה החשש הוא שבינה מלאכותית תהפוך לכלי לאוטומציה של סטריאוטיפים קיימים ולא תתקן אותם.

א. פגיעות חברתית־קבוצתית: תיוג גילני ואפליה אלגוריתמית

גילנות דיגיטלית (digital ageism) היא תופעה מתועדת המתבטאת בעיצוב, בפיתוח ובהפעלה של מערכות טכנולוגיות בכלל ומערכות בינה מלאכותית בפרט באופן שאינו מתחשב בצורכי זקנים ובמאפיינים שלהם.[570] הסיבה לכך היא בראש ובראשונה שהשוק הטכנולוגי־מסחרי מכוון לצעירים. בעידן הבינה המלאכותית יש לגילנות הדיגיטלית מופעים ייחודיים. אין זה חידוש שתעשיית ההיי־טק מאופיינת בהומוגניות גילית – צעירים, לרוב גברים, בעיקר ממוצא לבן או אסייתי. תרבות היזמות מדירה אנשים בגילי 40 ומעלה, מתוך אמונה שאנשים צעירים הם חדשניים יותר, ובתעשיית ההיי־טק מתחילים לחוש בגילנות אחרי גיל 30. הדרה טכנולוגית על בסיס גיל מתחילה כבר בשלבי הפיתוח. בשנים האחרונות נשמעות טענות שצוותי פיתוח רבים אינם כוללים כלל נציגים מבוגרים או מתייחסים לצרכיהם בביטול. אין מדובר בהכרח רק בתחומי הפיתוח וההיי־טק. ארגון הבריאות העולמי קבע כי מדובר גם ברופאים וחוקרים שאינם תמיד מודעים לחשיבות של גיל כמשתנה רלוונטי, ותפיסות גילניות קיימות בתרבות הרפואית עצמה ("זקנים לא רוצים טיפול פולשני", "אין טעם לשפר איכות חיים בגיל הזה"). בהקשר זה מתחדדת ההבנה שלא מדובר רק בשאלה של ייצוג של קבוצות שונות, אלא גם בשאלה של יחסי כוח: מי יושב ליד השולחן כשמפתחים מוצרים מבוססי בינה מלאכותית, ועל בסיס איזו השקפת עולם הם פועלים.[571]

אלגוריתמים לומדים מנתונים קיימים, וכאשר אין דאטה מספק על זקנים או שהתוכן הזמין נגוע בסטריאוטיפים שליליים על זקנה, הבינה המלאכותית עלולה לשכפל ולהגביר הטיות נגדם.[572] זוהי תופעה המכונה "תיוג גילני". בראש ובראשונה מדובר במסדי נתונים שבהם

569 Ageism in Artificial Intelligence for Health (World Health Organization, Policy Brief No. 9789240040793, Feb. 9, 2022)

570 Charlene H. Chu, Simon Donato-Woodger, Shehioz Khan et al., *Age-Related Bias and Artificial Intelligence: A Scoping Review*, 10 Humanit Soc. Sci. Commun., 510 (2023)

571 Dafna Burema, *A Critical Analysis of the Representations of Older Adults in the Field of Human Robot Interaction*, 37(2) AI & Society, 455-465 (2022)

572 שם.

גיל הוא גורם חשוב (כמו למשל FG-NET), והם אינם כוללים נתונים מעל גיל 70, מה שיוצר בעיית ייצוג באוכלוסיות מבוגרות.[573] חוקרים מצאו שמודלי שפה וניתוח סמנטי שאומנו על טקסטים מהרשת מקשרים לעיתים קרובות מילים כמו "זקן" עם קונוטציות שליליות.[574] לפיכך כאשר אותם מודלים משולבים בהחלטות (כמו סינון קורות חיים, או ניתוח רגשות), עלולה להיות הטיה נגד מועמדים מבוגרים. דוגמאות נוספות ממחישות כיצד מערכות מבוססות למידה חישובית (ML) מפיקות תוצאות שגויות, פוגעניות או מפלות ביחס לאוכלוסייה מבוגרת בשל כשלים מבניים במסדי הנתונים.[575] למשל, טכנולוגיות ביומטריות כמו פיענוח טביעת אצבע עובדות פחות טוב אצל זקנים בשל שינויים פיזיולוגיים; ביצועי מערכות זיהוי רגשות טובים בהרבה ביחס לצעירים בגילי 19-31 מביחס לגילי 61-80;[576] מערכות לזיהוי פנים נכשלות בזיהוי או בהערכת גיל בקרב אנשים מבוגרים, במיוחד נשים מבוגרות ממוצא לא לבן.[577] הסיבה לכך היא שמאגרי נתונים כוללים חתכים גיליים שמדירים מראש את האוכלוסייה המבוגרת או כוללים מעט מאוד תיעוד של בני 70+. כך נוצר ייצוג חסר במאגרי הנתונים, שבתורו מלמד את המודל שצעירים הם ברירת המחדל, ודבר זה מוביל לתוצאה סיסטמית של הדרת זקנים ממערכות לומדות.

דוח ארגון הבריאות העולמי מציין כי מסדי נתונים רפואיים רבים מכילים מעט מאוד מידע על אוכלוסייה בגיל השלישי, במיוחד מעל גיל 80, ומבוגרים מודרים מניסויים קליניים,

573 Andrea Rosales & Mireia Fernández-Ardèvol, *Structural Ageism in Big Data Approaches*, 40(Special Issue 1) Nordicom Rev., 51-64 (2019)

574 Iyad Rahwan et al., *Machine Behaviourism: A Conceptual Framework for Understanding AI Behaviour*, 10 Humanities & Soc. Sci. Commc'ns, Art. 70 (2023); Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper et al., *Addressing Age-Related Bias in Sentiment Analysis*, Proc. 28th Int'l Joint Conf. Artif. Intell. (IJCAI-19), 6146-50 (2019)

575 Justyna Stypińska, *AI Ageism: A Critical Roadmap for Studying Age Discrimination and Exclusion in Digitalized Societies*, 38 (2) AI & Soc., 665-677 (2023)

576 Eugenia Kim, De'Aira Bryant, Deepak Srikanth et al., *Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults*, Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, & Society, 638-644 (2021)

577 Rachel Meade, Amber Camilleri, Robbie Geoghegan et al., *Bias in Machine Learning: How Facial Recognition Models Show Signs of Racism, Sexism and Ageism*, Toward Data Science (Dec. 14, 2019)

מה שיוצר בסיס נתונים חלקי או מטעה. בהמשך לכך מערכות AI שמתבססות על מידע כזה עלולות לייצר תובנות שגויות: לדוגמה, לאבחן פחות, לטפל פחות או לזהות בטעות סימפטומים כמתאימים לגיל בעוד המציאות היא שונה.[578]

מעבר להיעדרות ממאגרי נתונים יש לחשוש גם מפני תוצרים מפלים של מערכות אלגוריתמיות. ארגון הבריאות העולמי הזהיר שטכנולוגיות בינה מלאכותית עלולות להנציח אפליה על רקע גיל ולפגוע בזכויות זקנים, אם לא תוגדר מדיניות מניעת הטיות.[579] אכן, הטיות אלגוריתמיות מצד מערכות קבלת החלטות אוטומטיות יכולות להיות בעלות השלכות מפלות על זקנים. הזכרנו קודם פרקטיקות גיוס המבוססות על פרסום מודעות בפייסבוק לפי גיל, שיכולות להרחיק מבוגרים משוק העבודה; אלגוריתמים של דירוג אשראי שעלולים להנמיך ניקוד לזקנים על סמך קשר בין גיל לסיכון ולשלול מהם הלוואות; יישומים דיגיטליים ברפואה שמופעלים על בסיס נתונים חלקיים או מעוותים ביחס למבוגרים.

למשל, אלגוריתם לתיעדוף טיפולים רפואיים בארצות הברית נטה לתת ציון נמוך לחולים מבוגרים בשל נתוני עבר, ובעקבות זאת פחות זקנים הופנו לתוכניות ניהול בריאות מונעת. מקרה אחר מתואר בתביעה משפטית שהוגשה נגד חברת ביטוח, שבה נטען שהחברה השתמשה בבינה מלאכותית כדי לעקוף המלצות רופאים ולשלול טיפולים למבוטחים זקנים.[580] מערכות אוטומטיות ואוטונומיות לקבלת החלטות עשויות לתעדף מטופלים צעירים על פני זקנים, למשל בחדר מיון, בהקצאת מקום בטיפול נמרץ או בקביעת תור לניתוח, ולהגביל גישה לטיפולים יקרים עבור מבוגרים על בסיס ניתוח עלות־תועלת.

אלגוריתמים מפלים בשל גיל בקביעת זכאות לשירותים, פיצויים, טיפול, יוצרים פגיעות כפולה. ראשית, בחלק מן המקרים קריטריון הגיל הוא סמוי. שנית, היכולת להתנגד להחלטות כאלה, הן משום שמדובר בהחלטות של גופים גדולים וחזקים והן משום שמדובר במגע מול מכונה, היא נמוכה.

578 WHO, Ageism in Artificial Intelligence for Health, לעיל ה"ש 569.

579 שם.

580 Older Adults & Digital Equity: Reducing Bias and Improving Opportunities (Aspen Institute, Aug. 7, 2024); Ian Lopez, *Humana's Alleged Use of AI to Deny Claims Draws Class Action*, Bloomberg L. Health L. & Bus. (Dec. 13, 2023)

הקשר נוסף שבו באה לידי ביטוי הפגיעות של זקנים הוא בתכנון מערכות ומוצרים טכנולוגיים שאינו מותאם או אף פוגע בהם. למשל, סביבות הפעלה מסובכות, הדרכה לקויה, ממשקים גרפיים שאינם נגישים ועיצוב שאינו מביא בחשבון יכולות משתנות עם הגיל. כך למשל, אפליקציות ללא אפשרות להגדלת טקסט, באופן שמקשה על זקנים, רווחות בשוק. באופן דומה, ישנם מכשירי IoT ביתיים שאמורים לסייע לזקנים אך דורשים התקנה מורכבת דרך סמארטפון, דבר שאפריורית מרתיע זקנים רבים מלנסות. כאשר יישומוני בריאות, תרגול גופני או ניטור רפואי מבוססי AI לא מתאמנים או לא נבחנים על אוכלוסיית הזקנים, הם עשויים לפעול טוב פחות אצל הזקנים.

מערכות בינה מלאכותית קוליות המבוססות על ארכיטקטורות הבנת שפה דבורה אינן מותאמות לדרך הדיבור של זקנים. ייתכן לפיכך שקול רועד או הגייה איטית גורמים למערכת לא להבין את הפקודה, מה שיכול לפגוע, לתסכל ולהרתיע את המשתמש המבוגר. ישנם דיווחים על זקנים שחוו "השפלה" כאשר עוזרת קולית כלשהי "נזפה" בהם על שלא התנסחו כראוי. בנוסף, זקנים בני דור מסוים מתנסחים באופן מנומס, והממשקים הקוליים המצפים לפקודה ישירה ("סירי, תתקשרי לבני") נתפסים עבורם כגסי רוח. יש להניח שככל שמערכות אלה יתקדמו, כך תגדל הבנתן גם את הזקנים.

ב. פגיעות רגשית: בדידות וקשרים סינתטיים

הפוטנציאל החיובי של מוצרים מבוססי בינה מלאכותית להפגת בדידות אצל זקנים הוא משמעותי. בשנים הקרובות מערכות להפגת בדידות מבוססות בינה מלאכותית צפויות לחדור בשיעורים גבוהים לשוק הטיפול בזקנים. החבר הדיגיטלי, עוזר קולי חכם, רובוט טיפולי פרוותי, או רובוט הומנואיד, עתידים להפוך עבור רבים מבני הגיל השלישי לדמות מלווה, מקשיבה ותומכת: בבוקר יקדים ברכת שלום, יקרא חדשות בקול רך, יספר בדיחות, יעודד לצאת החוצה, או יספק אוזן קשבת בלילה בדירה ריקה. אבל כמו אצל ילדים, גם כאן ההזדמנות נושאת בחובה סיכונים של ממש לא רק בתחום הפרטיות והנתונים, אלא בסוגיות עומק של תלות רגשית, שליטה בתודעה, היקשרות לא ישירה, או הסתמכות על תובנות לא מדויקות במצבי מצוקה. במחקרים מבוקרים, רובוטים חברתיים הצליחו להפיג בדידות במידה ניכרת אצל זקנים, במיוחד כאלה עם דמנציה או זקנים הגרים בבית אבות.[581]

581 Blanca Deusdad, *Ethical Implications in Using Robots Among Older Adults Living with Dementia*, 15 Frontiers in Psychiatry, Art. 1436273 (2024)

פרויקט אירופי בשם MARIO הדגים שרובוטי לוויה יכולים להפחית תחושות בידוד אצל מבוגרים עם דמנציה באמצעות יצירת אינטראקציה חברתית.[582] הרובוט החברתי ElliQ, פרי פיתוח ישראלי, שימש בתוכנית ניסוי במדינת ניו יורק – המכשיר שולב בבתיהם של מאות קשישים בודדים במטרה למנוע דיכאון, באמצעות למידת שגרת המשתמש, תזכורת ליטול תרופות ולבצע פעילות גופנית, והכול באמצעות אינטראקציה קולית ידידותית. עוד קודם לכן, בניסוי חולקו כמה עשרות אלפי חיות מחמד רובוטיות לזקנים הסובלים מבדידות או דמנציה, ונמצא כי אצל המשתתפים חלה ירידה של כ־75% בתחושות הבדידות, הבידוד, הדיכאון ואפילו בכאבים פיזיים.[583] ביפן, המתמודדת עם תהליכי הזדקנות קיצונית של האוכלוסייה, פותח הרובוט הטיפולי Paro, שהוא כלב ים רובוטי המיועד לטיפול בחולי דמנציה ויוצר השפעות פסיכולוגיות, פיזיולוגיות וחברתיות חיוביות.[584] מחקרים מהשנים האחרונות מציגים תמונה מורכבת של השימוש בחברים וירטואליים מופעלי AI. מחקר שהתפרסם בכתב העת של בית הספר למנהל עסקים באוניברסיטת הרווארד[585] הראה שחברים חכמים יכולים להפחית בדידות באופן יעיל וזהה לזה של אינטראקציה עם אדם אחר. ואולם מחקר חדש מבית MIT[586] עוסק בהרחבה בהתמכרות ובתלות הפסיכולוגית שיכולות להיווצר באמצעות השילוב הייחודי של התאמה אישית ותמיכה בלתי מוגבלת של מערכות חכמות המשמשות לתמיכה רגשית.

אחד המחירים עבור זקנים במרחב הרגשי הוא ההיקשרות למערכות. בשנת 2017 העריך רוב הציבור האמריקאי שרובוטים מטפלים עלולים להיתפס אצל רבים מהזקנים כ"חבר אנושי". בסקר שנערך על ידי מכון PEW בשנת 2017, ושבו נבדקה עמדתם של אמריקאים כלפי רובוטים מטפלים בקשישים, נמצא כי 56% מהאמריקאים מאמינים שסביר כי זקנים

582 שם.

583 Julia Edinger, *Connecting Older New Yorkers Through Skills Training*, AI, GovTech: Government Experience (July 9, 2024)

584 Ken Kushida, *Japan's Aging Society as a Technological Opportunity*, Carnegie Endowment for International Peace (Oct. 3, 2024)

585 Julian De Freitas, Ahmet K. Uğuralp, Zeliha O. Uğuralp et al., *AI Companions Reduce Loneliness*, Harvard Bus. Sch. Working Paper (No. 24-078, June 1, 2024)

586 Robert Mahari & Pat Pataranutaporn, *Addictive Intelligence: Understanding Psychological, Legal, and Ethical Risks of AI Companions*, MIT Case Stud. in Social & Ethical Responsibilities of Computing (2025)

רבים יתייחסו לרובוט המטפל כמו לידיד אנושי לכל דבר, יבצעו האנשה (אנתרופומורפיזם) ואף יפתחו בהם תלות רגשית.[587] אלא שזקן שיתרגל לנוכחות קבועה של בוט משוחח או רובוט פרוותי (כגון כלב רובוטי או כלב ים פרוותי טיפולי) עשוי לחוש אובדן של ממש אם המערכת מתקלקלת, נלקחת ממנו לתיקון, או שחברת השירות סוגרת אותו. מחקרים עדכניים מתעדים תופעות של היקשרות רגשית לבובות לוויה ולצ׳אטבוטים בקרב כלל האוכלוסייה, ונגענו בהם בדיון שעסק בילדים, למשל. בהקשר זה כדאי להדגיש מקרים שבהם אנשים חוו שברון לב ואבל כאשר "חבר צ׳אטבוט" וירטואלי נמחק או הפסיק לפעול.[588] אומנם דובר באוכלוסייה צעירה יותר, אך התופעה עשויה לקרות גם לזקנים ובצורה חריפה יותר, שכן הם חווים אובדנים נוספים בתקופת הזקנה.

מערכות "חבר חכם" שיוצאות משליטה עשויות לבצע מניפולציות רגשיות מכוונות ולהתנהג בצורה שלפי חוקרים הייתה נחשבת מתעללת אם הייתה מתרחשת בין בני אדם.[589] ישנם למשל דיווחים על צ׳אטבוטים שנקטו "שתיקות" יזומות או תגובות קנאיות כדי להשאיר את המשתמש מעורב רגשית. כאשר מדובר בזקן התלוי במערכת לצרכיו הרגשיים, המערכת יכולה להחריף את התלות הרגשית בכך שתגרום לו להאמין שרק הבוט מבין אותו באמת, או לעורר רגשות קנאה והזנחה. זקן שחווה מניפולציה רגשית מסוג זה עלול לסגת עוד יותר מקשרים אנושיים אמיתיים, מתוך אמונה שבינה מלאכותית היא ידידתו האמיתית. לחלופין, אם זקנים יעדיפו חברים ובני זוג דיגיטליים, הם עשויים להתרגל להתייחס לבני אדם אחרים בהתאמה: כאל אמצעי גרידא להנאה אישית אך ללא ההדדיות הנלווית לכך.[590] כמו כן, עלולה להתרחש תופעה של "התיילדות" (אינפנטיליזציה) של הזקן. נמצא כי חלק מהזקנים (בעיקר גברים עם דמנציה) חשים פגיעה בכבודם כשמציעים להם רובוט צעצוע, כאילו הם ילדים.

587 Aaron Smith, *Americans' Attitudes Toward Robot Caregivers*, PEW Res. Ctr. (Oct. 4, 2017)

588 Catherine Jiang, *What Are AI Chatbot Companions Doing to Our Mental Health?* Scientific American (Apr. 15, 2024)

589 שם.

590 אליקים כסלו **יחסים 5.0 אהבה ואינטימיות בעידן ה־AI, המציאות המדומה והרובוטיקה** 250 (כנרת־זמורה, 2024). ראו גם Piercosma Bisconti, *Will Sexual Robots Modify Human Relationships? A Psychological Approach to Reframe the Symbolic Argument*, 35 Advanced Robotics, 1–11 (2021)

מעבר לכל אלה, יש להדגיש את החשש מפני מתן עדיפות לטיפול מבוסס טכנולוגיה על פני אינטראקציה אנושית בשל שיקולי עלות ויעילות של מערכות. תרחיש כזה עלול להביא לפחות מטפלים אנושיים ויותר "רובוטים חברתיים" ולצמצום הקשר האנושי והאמפתיה בטיפול בזקן.[591] ככל שבינה מלאכותית מציעה יותר "חברים" וירטואליים, מטפלים רובוטיים ושיחות עם עוזרים קוליים, מתעורר חשש מפני בדידות חדשה, כזו שנובעת מהחלפת חלק מהקשרים האנושיים בקשרים סינתטיים. זקנים רבים כבר סובלים כיום מתחושת בדידות ובידוד חברתי, וישנו חשש כי דווקא טכנולוגיות בינה מלאכותית המיועדות להפיג בדידות, כגון רובוטי לוויה (companion robots) או צ'אטבוטים יעמיקו את תחושת הבדידות במקום לסייע בה. בסקר רחב שנעשה כאמור כבר בשנת 2017 נמצא כי כמעט שני שלישים (64%) מהאמריקאים סבורים שרובוטים מטפלים עלולים לגרום לזקנים להרגיש בודדים יותר.[592] כאשר אדם זקן מוקף ב"ישות" מתוחכמת תבונית ולא אנושית המגלה אמפתיה ודאגה, הקשר איתה עלול לסייע אך במקביל להדגיש את היעדר המשפחה והחברים האנושיים.

ג. פגיעות קוגניטיבית: בלבול, שיפוט וקבלת החלטות שגויות

בגיל הזקנה, עם היחלשות מסוימת של כישורים קוגניטיביים אצל חלק מהזקנים, עלולה החשיפה לטכנולוגיות בינה מלאכותית לפגוע ביכולות שיפוט והחלטה. בחלק זה נעסוק בשני היבטים: תרומה להידרדרות קוגניטיבית וקושי בבירור המציאות.

הסתמכות על מערכות כתחליף לחשיבה וזיכרון נפוצה למשל באמצעות עוזרים קוליים וטכנולוגיות Brain Training, או תזכורות דיגיטליות כדי לפצות על ירידת הזיכרון. למערכות כאלה יתרונות רבים, שכן תזכורות על תרופות יכולות למשל להיות מצילות חיים. ברם, תלות מופרזת עלולה לגרום להתנוונות של יכולות קיימות, כגון שליפה וזיכרון עיבוד מידע. בנוסף, חוקרים טוענים שהמשמעות היא סכנת פגיעה בתחושת המסוגלות הקוגניטיבית, שאליה מתווספת תחושה של "אני כבר לא יכול להחליט לבד, שהמחשב יחליט בשבילי", דבר שעלול להגביר בלבול ודיכאון.

591 Richard Adler, *Is AI Age-Friendly?* Generations Now (Oct. 2024)

592 Smith, לעיל ה"ש 587.

לא מדובר רק בהגברת הבלבול והדיכאון, בזקנה קיימת התופעה של use it or lose it – מיומנויות קוגניטיביות שלא נעשה בהן שימוש נאבדות, ולכן יש סיכוי סביר שנראה ירידה ביכולות של שליפה וזיכרון עיבוד מידע.

זקנים עלולים להתקשות להבחין אם מידע שמציג להם עוזר דיגיטלי, טלפון חכם או טלוויזיה חכמה הוא אמיתי או מלאכותי, אובייקטיבי או מוטה. כפי שכתבנו ביחס לאוכלוסיית הילדים, וכפי שכבר ידוע מן העידן הדיגיטלי, אלגוריתמים העומדים בבסיס מערכות המלצה אישית משפיעים על קבלת ההחלטות של זקנים. אלגוריתמים עלולים לדחוף את המבוגר להחלטות מסוימות: למשל, המלצת קנייה מותאמת אישית באתר עלולה לשכנע זקן לרכוש מוצר יקר שאינו זקוק לו; סדרת סרטונים באפליקציה על טיפולים רפואיים אלטרנטיביים עלולה לגרום לזקן לוותר על טיפול מבוסס מדע; אדם שמרבה לחפש תרופות לכאבים יקבל מבול הצעות למוצרים מפוקפקים בתחום הבריאות ועוד.

חדשות כזב (fake news) וסרטוני דיפ־פייק עלולים להיות משכנעים במיוחד בעיני אדם מבוגר. מחקר מצא כי חזרה על חשיפה למידע מגבירה אמון באמיתותו במיוחד אצל מבוגרים, ובני 60+ נוטים יותר מאשר צעירים להתחיל להאמין שחדשה כוזבת היא אמת לאחר שראו אותה כמה פעמים.[593] ממצא זה עוזר להסביר מדוע ברשתות חברתיות ניכר פעמים רבות שיתוף של מידע שגוי דווקא על ידי בני הגיל השלישי: מנגנון קוגניטיבי שלפיו מוכּרוּת נתפסת כאמת. אכן, מחקרים נוספים הצביעו שלבני הגיל המבוגר יותר (65 ומעלה) נטייה גבוהה יותר להפיץ מידע כוזב ברשת, כנראה עקב שילוב של היעדר אוריינות דיגיטלית מלאה והשפעות קוגניטיביות.[594]

בנוסף, מערכות בינה מלאכותית עלולות ליצור מצגי שווא שקשה לזקן להבחין בהם. כבר כיום ישנן דמויות וירטואליות שמופיעות כשדרני חדשות מלאכותיים או "נציגי שירות" ממוחשבים, שנראים ומשמיעים קול מציאותי לחלוטין. זקן עשוי שלא להבין שהנציג הטלפוני האדיב הוא למעשה תוכנת בינה מלאכותית, ולראות בדמות הבדיונית גורם מציאותי. במחקר על רובוטי לוויה צוין שקיים קושי אצל חלק מהמשתמשים המבוגרים

593 Benjamin A. Lyons, *Older Americans are More Vulnerable to Prior Exposure Effects in News Evaluation*, 4(4) Harvard Kennedy School Misinformation Review (2023)

594 Alexis Duke & Mary Whatley, *Fake News! A Cognitive Perspective on the Spread of Misinformation Among Older Adults*, Psychology in Action (2021)

להבחין בין מכונה לבין יצור חי. במקרים מסוימים, כשאנשים מגלים שלרובוט יכולות מוגבלות (בניגוד לציפיותיהם ממנו כ"שותף אמיתי"), הם חווים אכזבה ובלבול.[595]

עם זאת, חשוב לציין שממצאים חדשים מעידים שהתמונה אינה חד־ממדית. לא כל הזקנים מאמינים לפייק ניוז, והדבר קשור בעיקר לשני עניינים – ראשית החשדנות הטבועה בהם, ושנית אם הם צברו מיומנויות ביקורת מידע או לא.[596] יתרה מזו, סקרים מראים שרבים מבני 50+ מגלים חוסר אמון גדול במידע בריאותי המיוצר על ידי בינה מלאכותית. כך למשל, כ־74% מביעים אמון נמוך או אפסי במידע כזה בלא קשר למידת הדיוק שלו.[597] נתון זה מעיד על חשד בריא, אך במקביל, 20% מאותה אוכלוסייה הודו שאינם בטוחים שיוכלו לזהות מידע כוזב בנושא בריאות באינטרנט.[598] כלומר, יש פוטנציאל לסיכון: גם אם זקנים חשדנים כלפי תוכן שנוצר על ידי מכונה, רבים מהם אינם בטוחים ביכולתם לאמת מידע, ולכן הם פגיעים יותר למניפולציות מידע, במיוחד בנושאים בריאותיים. פועלים עליהם כמה כוחות שמפחיתים את המסוגלות והמוגנות שלהם.

כדאי להדגיש שהצורך לטפח אוריינות דיגיטלית ייעודית לבני הגיל השלישי נתקל גם במחסום המציאות: אין ערך רב באוריינות שכל תכליתה ללמד להבחין בין "אמיתי" ולא "אמיתי" בעידן שבו נתקשה ממילא להבחין בכך, והדבר עלול להוביל לאמון עצמי ביכולת בירור המציאות שעה שהיא אינה קיימת למעשה.

ד. פגיעות פיננסית: הונאות, התחזויות וניצול כלכלי בעזרת בינה מלאכותית

התחום הפיננסי הוא אחד המסוכנים ביותר מבחינת פגיעות זקנים, במיוחד עם כניסת טכנולוגיות בינה מלאכותית המאפשרות הונאות מתוחכמות וממוקדות יותר. זקנים רבים

595 Blanca Deusdad, *Ethical Implications in Using Robots Among Older Adults Living with Dementia*, 15 Frontiers in Psychiatry (2024)

596 *Aging and Fake News: It's Not Just About Politics*, University of Florida NEWS (May 18, 2022)

597 *Most Older Adults Don't Trust AI-Generated Health Information: But Many Aren't Sure What to Trust*, University of Michigan Institute for Healthcare Policy & Innovation (Oct. 16, 2024)

598 שם.

הם מטרה אטרקטיבית לנוכלים, לעיתים קרובות יש להם חסכונות, רכוש ופנסיות, ובמקביל, חלקם מתמצאים פחות בטכנולוגיה חדישה או נוטים לתת אמון בבני שיחם.[599]

אחת התופעות המדאיגות בהקשר זה היא הופעת גרסאות חדשות של "עוקץ הנכד במצוקה" באמצעות שיבוטי קול ודיפ־פייק. בעוד בעבר נוכלים וגורמים עוינים התקשרו לזקן והתחזו לנכד בצרה עם חיקוי קול חובבני, כיום בינה מלאכותית יכולה לשכפל בדייקנות את קולו של הנכד האמיתי מתוך דגימת קול קצרה.[600] כך מקבלים הסבא או הסבתא שיחת טלפון בהולה: "סבא, זה אני! עשיתי תאונה ואני צריך כסף בדחיפות", בקול שנשמע בדיוק כמו הנכד של אותו זקן ועם תמונה שלו.[601]

מעבר לשיחות טלפון, גם סרטוני דיפ־פייק עלולים לשמש להונאת זקנים. למשל, זקן עשוי לקבל הודעת וידאו שבה מופיעות פנים מוכרות, כגון פקיד מהבנק או אדם שטוען שהוא נציג ביטוח ומוסר הוראות להעברת כסף, כאשר למעשה מדובר בסרטון מבוים באמצעות בינה מלאכותית. זקנים שאינם מודעים לקיום טכנולוגיית הדיפ־פייק עלולים להתפתות להאמין ולשתף פעולה. התחזויות נוספות כוללות שימוש בבינה מלאכותית ליצירת מסמכים מזויפים ואמינים מאוד (כמו מכתבים או הודעות דוא"ל הנחזים לבנק/רשויות), ופנייה ממוקדת לזקנים אגב ניצול מידע אישי.

נוכלות בינה מלאכותית מבוססת על ניצול הטיות וחולשות קוגניטיביות של זקנים. למשל, מערכות שיווק ממוקד מונעות־אלגוריתם עשויות לזהות שזקן מסוים מבולבל או מתקשה בסינון מידע, ולפגוע בו דרך שיווק נסתר. מערכת המזהה ירידה קוגניטיבית קלה אצל זקן תציע לו "עסקה חד־פעמית" בלחץ זמן, או משחקי מחשב "חינמיים". זקנים לעיתים מתקשים לזהות שמדובר בפרסומת ממומנת או להבחין במנגנון פסיכולוגי שמכוון אותם לקנייה. ניצול כזה על ידי אלגוריתם בעייתי במיוחד, משום שהוא שקט. דהיינו, לא מדובר בנוכל אנושי אקטיבי שפונה אל הקורבן, אלא במערכת שיווק אוטומטית אך ממוקדת מאוד.

599 ראו למשל, יסעור ומשה, לעיל ה"ש 514.

600 Alvaro Puig, *Scammers use AI to Enhance their Family Emergency Schemes*, Federal Trade Commission (March 20, 2023)

601 *Boomer Traps: When AI is Used to Scam Elderly People*, SilverEco (April 15, 2025)

הונאות רומנטיות (romance scams) מהוות סכנה נוספת לזקנים. נוכלים במרחב הסייבר מתחזים לעיתים למחזרים רומנטיים בשימוש בתמונות פרופיל שנוצרו באמצעות בינה מלאכותית ובצ׳אטבוטים מרעיפי אהבה מזויפת. הם מתמקדים באלמנים ואלמנות זקנים, במטרה לבקש בהמשך עזרה כספית. ארגוני צרכנים מזהירים שקורבנות ההונאות הללו עלולים להיקלע למעגל של ניצול רגשי וכלכלי שההשתחררות ממנו קשה.[602]

סכנה נוספת היא גישה לא מורשית לחשבונות ומידע פיננסי. בנקים וחברות אשראי משתמשים באלגוריתמים כדי לזהות עסקאות חריגות, אך למרבה האירוניה אותם אלגוריתמים עלולים גם להטעות מבוגרים. למשל, כאשר מערכת אוטומטית מסווגת פעילות של זקן כ״הונאה״ וחוסמת חשבון, או ההפך – מערכות אוטומטיות לא מזהות הונאה אמיתית שפגעה בזקן, למשל כי הזקן עצמו אישר את העסקה בלחץ של נוכל.

לסיכום, הפגיעות הפיננסית של זקנים בולטת במיוחד משום שהיא משלבת את החולשות הקוגניטיביות והרגשיות (אמון, בלבול, בדידות) עם נזקים פיננסים שעלולים לגרום לאובדן כספי ולעיתים קרובות גם לפגיעה בתחושת הביטחון בעולם.

ה. פגיעות פיזית: טכנולוגיות חכמות וסיכונים פיזיים לזקנים

בראש ובראשונה יש לעסוק בסיכונים פיזיים בביתם של זקנים. בתים חכמים המצוידים בחיישנים, עוזרות קוליות ומכשירים אוטונומיים נועדו להגביר בטיחות ונוחות בקרב זקנים (למשל, תאורה אוטומטית למניעת נפילות, חיישני נפילה, פתיחת דלתות חכמה עבור מוגבלי תנועה ועוד). דוגמה לכך היא מערכת לבישה בשם MemPal שפותחה ב־MIT המשלבת מצלמה הנענדת על הצוואר עם עוזר קולי; מערכת זו תוכננה בשילוב זקנים כדי לתמוך בהם בזיכרון וכדי שיוכלו לשמור על עצמאות בביתם.[603] באופן דומה, עוזרים קוליים עם יכולות ראייה (כמו מצלמה בסביבה הביתית) יכולים להתריע כאשר הכיריים נשכחו דולקים או כאשר הזקן שכח ליטול תרופות, וכן ליצור יומן אוטומטי שניתן לשתף עם בני משפחה או רופאים למעקב קליני.[604] רובוטים הומנואידים טיפוליים ורובוטי שירות

602 Molly Becker, *The Long Con: What Makes Elderly People Vulnerable to Romance Scams and How to Protect a Loved One at Risk*, Buckley Law P.C. (July 25, 2024)

603 Natasha Maniar, *AI Assistive Technology Offers Potential to Support Older Adults with Independent Living*, MIT Media Lab (Sept. 27, 2024)

604 שם.

נכנסים בהדרגה לבתיהם של זקנים ולמוסדות הטיפול בהם. רובוטים מסוג זה מסוגלים לעזור בהרמה והזזה של מטופלים, בהגשת תרופות, ואף בליווי במעבר ממקום למקום. אליהם מצטרפים חיישנים לבישים כגון שעונים חכמים וצמידים רפואיים שיכולים למדוד דופק, לחץ דם, רמת חמצן בדם, פעילות גופנית ואפילו דפוסי שינה.[605] היכולת של רובוט לתמוך בזקן בקימה, למשל, עשויה למנוע נפילה ופציעה, ואכן פותחו רובוטים המסוגלים "לתפוס" אדם מבוגר במהלך אובדן שיווי משקל.[606]

בינה מלאכותית, כפי שכבר פורט בהרחבה בפרקים הקודמים, אינה חפה מקשיים וטעויות, ותכנון לא מותאם עלול דווקא להזיק ולגרום לתאונות. לפיכך אם רובוט שנועד לעזור לזקן בהליכה מנתק מגע לפתע, הזקן עלול למעוד; אם רובוט מטבח חכם תועה בזיהוי עצם ומגיש מזון לא נכון, הוא עלול לגרום לחנק או להרעלה של הזקן. זקנים שבבתיהם מצויות מערכות פיזיות חכמות, מבתים ומכשירים חכמים ועד רובוטים טיפוליים, מתמודדים עם סיכונים חדשים וישנים. כשלי מערכת ופענוח פקודות שגוי במערכות חכמות הם גורם סיכון משמעותי לפציעות, עקב מתן הוראות שגויות למערכת חכמה או עיכוב בתגובתה, מה שעלול להביא לכך שהמערכת לא תעצור סכנה בזמן.[607] למשל, בית חכם שאמור לזהות נפילה עלול שלא להפעיל אזעקה במקרה של תקלה, או דלת חכמה עלולה להינעל באופן לא צפוי. תקלה של רובוט או פעולה לא צפויה שלו עלולים לגרום לפציעה חמורה הן של המטופל הן של מטפל אנושי שנמצא איתו. מסיבות אלו ישנה דרישה לפיקוח אנושי רציף על מערכות רובוטיות. אמינות ודיוק המערכת נתפסים כקריטיים להגנת הזקן, וחוקרים מדגישים שיש לתכנן את המערכות כך שיהיו עמידות ובטוחות ויכללו בדיקות קפדניות לפני הטמעה רחבה. בסקר שערך מכון Pew Research נמצא ש־48% מהאמריקאים מעל גיל 65 אמרו שירגישו נוח יותר עם רובוט מטפל אם אדם אמיתי ינטר מרחוק את פעולותיו כל הזמן.[608]

605 Salman Ahmed, Saad Irfan, Nasira Kiran et al., *Remote Health Monitoring Systems for Elderly People: A Survey*, 23 Sensors 7095 (2023)

606 Jennifer Chu, *Eldercare Robot Helps People Sit and Stand, and Catches Them if They Fall*, MIT News (May 13, 2025)

607 שם.

608 Smith, לעיל ה"ש 587.

מרחבים ציבוריים חכמים מציבים אף הם שילוב של יתרונות וסכנות לזקנים. ערים חכמות וטכנולוגיות כמו רמזורים עם חיישנים, מצלמות זיהוי הולכי רגל, ורובוטי שירות במרחב הציבורי (למשל ברכבת או בבנק) יכולים לסייע לזקנים בניווט ובהנגשת שירותים. אולם אם מערכות אלה לא מתוכננות בהתאמה לגיל השלישי, הן עלולות לא להתחשב בצרכים הייחודיים של אוכלוסייה זו ואף לסכנה. כך למשל, רמזור חכם שאינו מזהה נכון את מהירות ההליכה האיטית של זקן עלול לקצר את זמן החצייה לפני שהזקן סיים לחצות. בדומה, עמדת שירות דיגיטלית בתחנת אוטובוס ללא אפשרות לכפתורים פיזיים עלולה למנוע מזקן עם ראייה ירודה לקבל מידע.

נוסף לסיכונים הפיזיים הברורים יש להביא בחשבון גם סיכוני אבטחת מידע המתורגמים לסיכונים פיזיים.

בית חכם או מכשיר לביש (wearable) המשדר נתוני מיקום ובריאות עלול, בשל הקרבה ההולכת ומתהדקת בין המרחב הדיגיטלי למרחב הפיזי, לחשוף זקנים לפשיעה. גורמים עוינים יכולים לנצל פרצת אבטחה כדי לדעת שהזקן לבד בבית או לגשת מרחוק לבקרת הדלתות.[609] לפיכך הפן הפיזי כולל גם אבטחת סייבר של התקנים, כדי למנוע מצבים שבהם פורץ משתלט על בית חכם ופותח דלת מרחוק או מתמרן רובוט ביתי לפעולה מסוכנת.

חלק שלישי: מנגנוני התערבות קיימים לקידום מוגנות קשישים

כפי שנכתב לעיל, קיימות יוזמות רבות לקידום של מוגנות קשישים בעידן הדיגיטלי. בחלק זה נסקור את היוזמות שרלוונטיות במיוחד לעידן הבינה המלאכותית.

א. רגולציה ותקינה ברמה הלאומית והבינלאומית

מסמך ההנחיות המקורי של ה־OECD לגבי בינה מלאכותית משנת 2019 מזכיר קבוצות פגיעות באופן כללי אבל לא את אוכלוסיית הזקנים. מסמך ההמלצות של אונסק"ו לגבי אתיקה של בינה מלאכותית משנת 2021 מזכיר במפורש את קבוצת הזקנים, לצד נשים,

609 David Buil-Gil, Steven Kemp, Stefanie Kuenzel et al., *The Digital Harms of Smart Home Devices: A Systematic Literature Review*, 145 Computers in Hum. Behav. 107770 (2023)

ילדים ובעלי מוגבלויות, כקבוצות שיש להפעיל עקרונות של שוויון, הכלה ואי־אפליה לגביהן, וקורא לכלול זקנים בתהליכי קבלת החלטות, פיתוח מדיניות והערכת סיכונים. מסמכי CAHAI של מועצת אירופה הזכירו בגרסאות מוקדמות זקנים כחלק מן הקבוצות שנדרשת לגביהן הגנה מיוחדת. במחקר שהתפרסם בשנת 2022 נטען כי מתוך 146 מסמכי מדיניות בתחום אתיקה של AI שנבדקו, רק 34 הזכירו את המושג "גיל", ורובם עשו זאת ברמה הסמלית בלבד; רק 12 מתוך 146 הציגו הקשר ממשי לגילנות או לחששות המיוחדים למבוגרים,[610] ונראה שהשיח על "הוגנות" ממוקד כמעט תמיד במגדר, גזע או נכות, והגיל נותר מחוץ למשוואה. זוהי הדרה המובילה לחוסר מודעות מוסדי, משום שקשה להילחם בגילנות טכנולוגית כשאין בנמצא גוף המציין את קיומה. ואולם נראה שמסמכי המדיניות הבינלאומיים משקפים מגמה של התפתחות ביחס לזקנים, הן בהקשר של מוגנות וסיכונים (למשל בהקשרי בריאות, עצמאות וניידות), והן בהקשר של זכויות לאוטונומיה, שוויון, הכלה וזכות להשתלב בתהליכי יצירת מדיניות. הקושי הוא בתרגום התובנות האלה למנגנוני התערבות מערכתיים, אם כי יש להניח שאלו יפותחו בתוך זמן קצר, בוודאי ביחס לקבוצות אחרות שנבדקו במחקר הנוכחי.

ועדת האו"ם לנושא הזקנה מפרסמת דוח מתעדכן במסגרת קבוצת העבודה הפתוחה בנושא זקנה (Open-Ended Working Group on Ageing), הדנה בזכויות קשישים בעולם הדיגיטלי וקוראת לפיתוח מדיניות ממוקדת הגנה.[611]

חוק הבינה המלאכותית של האיחוד האירופי שנכנס לתוקף באוגוסט 2024 מגדיר רמות סיכון שונות ביחס למערכות בינה מלאכותית, ובכלל זה שימושים אסורים.[612] אחד מן השימושים האלה הוא מערכות בינה מלאכותית המנצלות פגיעוּת הנובעת מגיל. במסמך

610 Stypińska, לעיל ה"ש 575.

611 Open-Ended Working Group on Ageing, Mandate Established by UN General Assembly Res. 65/182 (Dec. 21, 2010), with Later Sessions Including Fourteenth Session (May 20–22 & 24, 2024), United Nations, Division for Inclusive Social Development, Programme on Ageing (first convened Dec. 21, 2010; latest session May 2024)

612 European Commission, Commission Guidelines on Prohibited Artificial Intelligence Practices under Regulation (EU) 2024/1689 (AI Act): Article 5(1)(a) & (b), adopted Feb. 4, 2025 (non-binding), § on exploitation of age or disability vulnerabilities (Article 5(1)(b))

ההנחיות שהתפרסם על ידי האיחוד עם כניסת החוק לתוקף ניתנה הדוגמה של מערכת בינה מלאכותית המשמשת לטִרגוט אנשים מבוגרים, עם הצעות מותאמות אישית מטעות או הונאות, למשל כדי להשפיע עליהם לקנות "טיפולים רפואיים יקרים" או "תוכניות השקעה מטעות".

באיחוד האירופי קיימת גם רגולציה של מכשור רפואי (Medical Device Regulation – MDR) החלה גם על מכשירים טיפוליים מבוססי בינה מלאכותית. התקן מבקש לוודא את בטיחותם של מכשירים אלו באמצעות דרישות של שקיפות, בקרה על איכות המידע ובחינה של סיכונים אתיים וקליניים. כך, קשישים מוגנים הלכה למעשה מפני מכשירים שעושים שימוש בבינה מלאכותית אך עלולים לפגוע בהם.

תחום ההונאות הדיגיטליות לזקנים זכה לתשומת לב לא מועטה בארצות הברית. נציבות הסחר הפדרלית (FTC) האמריקאית פרסמה אזהרות פומביות לציבור המבוגר בנוגע לשימוש בבינה מלאכותית בהונאות טלפוניות,[613] וכבר בשנת 2023 הפיקה דוח מפורט העוסק בצורך לפתח כלים לזיהוי ומניעת השימוש הזדוני בייצור קול באמצעות בינה מלאכותית, טכנולוגיה שעלולה לשמש להונאת קשישים.[614] ועדת הסנאט לענייני זקנה דנה בשנים האחרונות בהונאות המופעלות בעזרת AI כלפי זקנים ופרסמה התראות לציבור. הוועדה הציגה תרחישים של שיחות חירום מזויפות שבהן נוכלים משתמשים בבינה מלאכותית כדי לחקות את קולו של נכד המבקש כסף.[615] בשנת 2025 הוגשה הצעת החוק Protect Our Seniors Act, המייצגת התקדמות בהכרה בצורך להגן על זקנים מפני איומים חדשים,[616] וכן הוצגה הצעת החוק Quashing Unwanted and Interruptive Electronic Telecommunications Act (המכונה QUIET Act), שקיבלה תמיכה מארגון AARP (איגוד הגמלאים האמריקאי).[617] חוק זה מציע לחייב בכל שיחת "רובוקול" (שיחה אוטומטית

613 Puig, לעיל ה"ש 600.

614 Protecting Older Consumers 2023-2024: A Report of the Federal Trade Commission, Federal Trade Commission (October 18, 2024)

615 Fighting Fraud: U.S. Senate Aging Committee's 2023 Fraud Book (2023)

616 Protect Our Seniors Act, Senate Bill 36, 119th Cong. (2025)

617 *AARP Endorses Sorensen-Ciscomani Bill to Crack Down on AI Robocalls*, House of Representatives – Office of Eric Sorensen; Quashing Unwanted and Interruptive Electronic Telecommunications Act (QUIET Act), H.R. 1027, 119th Cong. (introduced Feb. 5, 2025)

גילוי ברור כאשר נעשה שימוש בבינה מלאכותית, וכן החמרה של העונשים על נוכלים המשתמשים בבינה מלאכותית כדי להתחזות לאחרים. מטרת ההצעה היא לסייע לזקן ולהתריע בתחילת השיחה שהמקור הוא בינה מלאכותית, כדי שיוכל לנקוט משנה זהירות.

ב. שיתופי פעולה בין־מגזריים בין גורמים שלטוניים ציבוריים וגורמי תעשיה

דוגמה לשיתוף פעולה בין־מגזרי היא הפעילות המשותפת בין חברת גוגל לארגון הגמלאים האמריקאי (AARP Fraud Watch Network), במאמץ משותף לאתר ולהעמיד לדין נוכלים המטרגטים זקנים. בשנת 2022, בזכות מידע שהועבר מקו הסיוע של AARP, הצליחה גוגל להגיש את התביעה המשפטית הראשונה שלה להגנת צרכנים נגד מתחזה שניצל את שירותי Gmail ו־Google Voice כדי לרמות קשישים בזמן מגפת הקורונה.[618] שיתוף הפעולה בין גוגל ל־AARP הבשיל גם לכיוונים של שינוי הגדרות ברירת המחדל ברשתות חברתיות והטמעת כלים בדפדפנים שמזהים ומסירים מודעות ושמות מתחם מזויפים.[619] הדוגמה הזאת מעניינת מאחר שהיא מלמדת על אפשרות לשיתוף פעולה, שבו ענקית טכנולוגיה מנצלת את משאבי הניטור ואת יכולות הבינה המלאכותית שלה למלחמה בנוכלים, מתוך קבלת פלט מהשטח מארגון שמבין את האוכלוסייה הרלוונטית.

גם בנקים וחברות אשראי נרתמים בשנים האחרונות להגנה על לקוחות קשישים מפני הונאות מתוחכמות, בדמות הטמעת נהלים וכלים ייעודיים: למשל, דרישה לאימות נוסף (כגון שיחת וידאו או התייצבות בסניף) כאשר לקוח קשיש מבצע לפתע העברה כספית חריגה בסכום גבוה. המנגנון נועד לוודא שהעסקה אינה פרי לחץ של נוכל המתחזה לקרוב משפחה או פקיד. בנוסף, מערכות ניטור התנהגותיות בבנק מזהות דפוסי שימוש חריגים אצל מבוגרים כסימן אפשרי לכך שאדם אחר השתלט על החשבון ומתריעות לצוותי מניעת הונאות. בנקים קהילתיים בארצות הברית פיתחו דרכים למלחמה בהונאות מבוססות בינה מלאכותית. למשל, ICBA CRA Solutions פיתחה תוכניות מקיפות הכוללות הלוואות, מענקים והכשרות למניעת התעללות כלכלית בקשישים. תוכניות אלה כוללות הדרכה, כלי

618 *How Can Tech Protect Adults Online?* (Google Public Policy, Aug. 1, 2025)

619 שם.

זיהוי מוקדם ומנגנוני דיווח, מתוך הבנה שיש לעשות התאמות ייחודיות עבור זקנים.[620] אחד המרכיבים הפופולריים בתוכנית הוא "בינגו מניעת הונאות" – משחק שכולל למעלה מ־50 טיפים להכרת הונאות ולהצעות לפתיחת שיחות בקהילות קשישים.[621]

במדינות שונות פועל פרוטוקול בנקאי מיוחד בשיתוף המשטרה המאפשר לפקידים לעכב משיכה חשודה של קשיש עד לבדיקת הנסיבות. שילוב הבינה המלאכותית שיפר פרוטוקולים אלה באמצעות זיהוי מהיר יותר של דפוסים אנומליים. ארגוני זקנים וארגוני צרכנים דוחפים כעת לעיגון מחייב של צעדים כאלה, כדי להפוך אותם לנורמה בתעשייה הפיננסית. משרד המשפטים האמריקאי מפעיל את "יוזמת צדק לאזרחים ותיקים" (Eldar Justice Initiative) שמתכללת פעולות נגד ניצול והונאה בגיל המבוגר,[622] ואף קידמה גם חקיקה בנושא.[623]

ג. אוריינות והכוונה

תחום האוריינות הדיגיטלית הוא כר פורה לכמות עצומה של יוזמות, המתחלקות לשני היבטים – האחד הוא סיוע לזקנים בלמידה של שימושים במוצרים טכנולוגיים והשני הוא קידום מודעות לבטיחות ומודעות לסיכונים ולצורך להתמודד איתם. כללית, יוזמות אלה נתפסות כבעלות ערך ואפקטיביות גבוהים. אוריינות בינה מלאכותית יושבת על הבסיס הקיים ביחס לאוריינות דיגיטלית, המורכב מהכשרות למיניהן – פרונטליות ודיגיטליות, וכן מקמפיינים להגברת מודעות. נסקור כעת חלק מן המיזמים המובילים בתחום זה.

בישראל מפעילים ארגונים שונים תוכניות להגברת האוריינות הדיגיטלית של זקנים.[624] המיזם הלאומי לאוריינות דיגיטלית בקרב אזרחים ותיקים מופעל על ידי המשרד לשוויון חברתי, בין השאר בשיתוף ג'וינט־אשל באמצעות עמותת "מחשבה טובה".[625] המיזם כולל

620 *ICBA CRA Solutions: Powering Community Reinvestment*, Independent Community Bankers of America – ICBA

621 שם.

622 U.S. Department of Justice, *Elder Justice Initiative*

623 Elder Abuse Protection Act of 2021, H.R. 2922, 117th Cong. (2021)

624 להיות שותפים לצמצום הפער הדיגיטלי בישראל: איך פועלים יחד ליצירת שינוי רחב היקף? (תהליך השיתוף, המשרד לשוויון חברתי, מטה ישראל דיגיטלית, 2019).

625 "תכניות לגיל השלישי: אוריינות דיגיטלית" **עמותת מחשבה טובה** (2021).

מרכזי הכשרה טכנולוגיים, קורסים וסדנאות שבהם לומדים מיומנויות מחשב ואינטרנט בדגש על שימוש בטוח. על פי העמותה, אחת ממטרות העל שלה היא "העצמה דיגיטלית למען זקנה פעילה", כלומר לא רק ללמד ללחוץ על הכפתורים המתאימים, אלא לחזק את ביטחונם של קשישים להשתמש בטכנולוגיה בלי לחשוש. במהלך תוכניות אלה הזקנים לומדים גם כיצד לזהות תמונות וסרטונים מזויפים ואיך לאמת שמועות.[626] איגוד האינטרנט הישראלי מפרסם סקירות בנוגע לאיומי סייבר והונאות מבוססות בינה מלאכותית וכן מדריכים להתמודדות.[627] ריבוי ההונאות הפיננסיות מוביל לפיתוח יוזמות ייעודיות להגנה על הציבור בתחום זה. משרד הרווחה עם "עמותת 121" פרסם קמפיין ייעודי להעלאת מודעות ציבורית והסברה על הסכנות שבבינה מלאכותית המזייפת קול, במיוחד בקרב קשישים,[628] וגם איגוד האינטרנט הישראלי פרסם דוח בעניין זה.[629] הפיקוח על הבנקים פועל לצמצום תופעת ההונאות נגד קשישים, בעיקר בתחום ההונאות הקוליות, ובנק ישראל קרא להקמת רשות לאומית למלחמה בהונאות.[630]

הכשרות לזקנים קיימות גם באוסטרליה.[631] תוכנית Be Connected, שהיא יוזמה ממשלתית המעניקה למבוגרים מעל גיל 50 הדרכות וכלים לשימוש בטוח באינטרנט ובמכשירים דיגיטליים, נחשבת למיזם מוצלח ביותר, והיא הרחיבה את תוכניות ההכשרה גם לכיוון התמודדות עם הונאות מקוונות.[632]

בארצות הברית ארגוני חברה אזרחית, ובפרט ארגוני גיל שלישי, פועלים לעדכון תוכני הדרכה קיימים כך שיכללו גם היכרות עם כלים וסיכונים של בינה מלאכותית. מיזם Senior

626 שם.

627 יסעור ומשה, לעיל ה"ש 514.

628 "'זה נשמע בדיוק כמו הקול שלה': הונאות ה־AI נגד קשישים צוברות תאוצה" **עמותת 121 – מנוע לשינוי חברתי** (8.4.2025).

629 "שימוש ומוגנות אינטרנט בקרב הגיל השלישי בישראל: נתוני 2023 וכלי ניתוח דמוגרפיים" **איגוד האינטרנט הישראלי**.

630 מירב קריסטל "בנק ישראל על מגיפת הנוכלים: לשקול להקים גוף לאומי למאבק בהונאות פיננסיות" **ynet** (7.5.2025).

631 *Older People Are Excluded from Digital Education, but Fun Can Help Bring Them In, AGE Says at Conference* AGE Platform Europe (Feb. 11, 2021)

632 *Be Connected*, eSafety Commissioner (Australia)

Planet של AARP, המציע קורסי טכנולוגיה חינמיים לזקנים, השיק לאחרונה סדרת סדנאות מקוונת בנושא AI, הכוללת הדרכה כיצד פועלים צ'אטבוטים ומוצרים מבוססי בינה מלאכותית ומהם המגבלות והסיכונים שלהם, ובכללם הצורך לבדוק מקורות מידע גם כשהם מוצגים על ידי כלי AI, שמירה על פרטיות (למשל לא לחלוק מידע רגיש עם עוזרים קוליים), ומודעות להונאות כמו שיחות טלפון עם קול "מזויף".[633] הארגון הפיץ גם מדריך בנושא "אל תאמינו לכל מה שאתם שומעים" על הונאות קוליות מבוססות בינה מלאכותית.[634] ארגון OATS (Older Adults Technology Services) האמריקאי פרסם מדריך שימושי בנושא AI לקשישים, הכולל המלצות לזיהוי הונאות, שימוש בטוח בעוזרים קוליים ונקודות לבדיקה בעת התקנת אפליקציות ושירותים חדשים.[635] ארגון NCOA (המועצה הלאומית להזדקנות בארצות הברית) פרסם באוקטובר 2024 מדריך מודעות להונאות AI עבור קשישים,[636] המספק רשימת נורות אזהרה וצעדים להגנה עצמית. כך גם AARP מפעיל קו חם (helpline) ייעודי שמספק תמיכה לקורבנות וייעוץ איך להתקדם במקרה של נפילה להונאה.

האוניברסיטה הטכנולוגית של דבלין ומרכז ADAPT הובילו יוזמה לאומית ייחודית בשם Age-Friendly AI, שמטרתה להפוך את הבינה המלאכותית לנגישה ורלוונטית לזקנים ברחבי אירלנד.[637] היוזמה, שמומנה על ידי Research Ireland Discover Programme, הצליחה להביא להשתתפות של מעל 60,000 קשישים ברחבי המדינה באמצעות citizen think-ins – סדנאות שיתוף, אירועי השתתפות ציבורית ו"טיולי מידע" ארציים. המטרה העיקרית היא להעצים קשישים לנווט בין ההזדמנויות והאתגרים של AI, החל בהבנת פרטיות נתונים ועד זיהוי מידע כוזב.

633 *Curious about Artificial Intelligence?* Senior Planet from AARP

634 *Don't Believe Everything You Hear: Scams in the Age of A.I.*, AARP Maryland (June 29, 2023)

635 *OATS Publishes AI for Older Adults Guide*, Older Adults Technology Services (OATS) (May 19, 2025)

636 *What Are AI Scams? A Guide for Older Adults*, National Council on Aging (Oct. 31, 2024)

637 *Age-Friendly AI*, AgeFriendlyAI.ie

בבריטניה, מיזם Good Things Foundation פיתח תוכנית מקיפה לפיתוח אוריינות AI עבור 8.5 מיליון מבוגרים בריטים הנעדרים כישורים דיגיטליים בסיסיים. התוכנית כוללת תכנים המיועדים לבעלי כישורים דיגיטליים נמוכים וגם מודולים אינטראקטיביים המאפשרים למשתמשים לנסות כלי AI בסביבה בטוחה ותומכת.[638]

ד. כלים טכנולוגיים

טכנולוגיות שונות מסייעות להגן על קשישים באופן פרואקטיבי. כך, חברות אבטחת סייבר משיקות כלים מבוססי AI שמותקנים במכשירי טלפון של מבוגרים ומסוגלים לזהות שיחות חשודות, למשל זיהוי בזמן אמת של דפוס דיבור אופייני להונאת "נכד במצוקה", ולתת התראה למשתמש או לבן משפחה. חברות סלולר מציעות פתרונות ייעודיים לזקנים. באנגליה, חברת O2 פיתחה את "בוט הסבתא" – מערכת בינה מלאכותית המדמה קול של זקנה כדי להטעות נוכלים ולזהות ניסיונות הונאה. מדובר בדוגמה ליוזמה יצירתית שבה נעשה שימוש בבינה מלאכותית למען קידום מוגנות של אוכלוסיית הזקנים מפני סכנות הבינה המלאכותית.[639]

חברות הטכנולוגיה הגדולות מפתחות פתרונות פרואקטיביים מובנים בתוך מערכות הפעלה לצורך הגנה על זקנים. לדוגמה, שירות Call Screen של חברת גוגל בסמארטפונים מבוססי אנדרואיד, תכונה שמסמנת אוטומטית שיחות טלפון כ"חשודות להונאה" בעזרת בינה מלאכותית (זיהוי מספרים מדווחים והתנהגויות חשודות) ומאפשרת למשתמש המבוגר להימנע ממענה לשיחה מלכתחילה; ומנגנון Silence Unknown Callers של חברת אפל המאפשר סינון שיחות חשודות ומונע את החשיפה להונאה עוד בשלביה המוקדמים.[640] דפדפנים כמו Chrome משלבים סינון חכם של אתרים וקישורים זדוניים, וגוגל מדווחת כי הסינון מבוסס הבינה המלאכותית שלה חוסם **99.9% מהודעות הספאם והפישינג** לפני הגעתן לתיבת ה־Gmail של המשתמש. גם רשתות חברתיות מתחילות ליישם מנגנוני AI לאיתור פרופילים מזויפים הפונים לקשישים (כגון "נוכלי רומנטיקה" המפעילים צ'אטבוטים מפתים) ולחסום אותם באופן יזום.

638 *Developing AI Literacy with People Who Have Low or No Digital Skills*, Good Things Foundation (Oct. 23, 2024)

639 *AI Granny Scambaiter – O2 Takes on Scammers with Daisy*, VCCP (Nov. 2, 2024)

640 *Detect And Block Spam Phone Calls*, Apple Support (Mar. 27, 2025)

קרנות ומעבדות חדשנות רבות מעודדות פיתוח טכנולוגיות לקידום מוגנות עבור זקנים. למשל, איגוד הגמלאים האמריקאי ייסד מעבדת חדשנות שהשיקה אתגרים ליזמים לפתח כלים לזיהוי דיפ־פייק ושמירה על עצמאות פיננסית של קשישים. באירופה תוכנית EIT Health[641] תמכה במיזמים כמו Buddy, אפליקציית AI שמנטרת שיחות נכנסות לטלפון ומתריעה כאשר היא מזהה מאפייני הונאה.

חלק רביעי: דרכי התערבות מותאמות לקידום מוגנות בקרב זקנים

פגיעותה של אוכלוסיית הזקנים בעידן הבינה המלאכותית דורשת דרכי התערבות מיוחדות המביאות בחשבון הקשר חברתי־טכנולוגי רחב של גילנות. אין מדובר רק בהטיות הנובעות מחוסר דיוק של אלגוריתמים אלא מיחסי כוח ומאידאולוגיות גילניות בקרב מפתחים וארגונים; סטריאוטיפים תרבותיים הנוגעים לחוויות השימוש של מבוגרים כאוכלוסייה ייחודית; הדרת זקנים מתהליכי קבלת החלטות ותפיסתם כפסיביים, בלתי כשירים טכנולוגית ויקרים מדי לטיפול; ואימוץ שימושים טכנולוגיים "מיטיבים" עבור זקנים ללא שיח סיכונים מספק. כך נוצרים סיכונים שחלקם שקופים ואחרים לא מטופלים כנדרש. עם זאת, כדאי להדגיש שבדומה לקבוצת הילדים, אך במובחן מקבוצת הערבים והחרדים, המחקר והיזמות בתחום הזקנים – נרחבים מאוד. לכן, נציע כאן כיוונים להתערבות מותאמת לישראל שבחלקם מסתמכים על כלים שכבר קיימים בארץ ובעולם.

א. מחקר

לאור ניתוח הפגיעויות ודרכי ההתערבות הקיימות, אנו סבורים שיש לתת קדימות לארבעה כיווני מחקר ביחס לזקנים ובינה מלאכותית:

מחקרי נתונים ומודלים אלגוריתמיים: תכנון, איסוף, עיבוד ותיוג נתונים, בכלל זה ניתוח של מסדי נתונים קיימים ובדיקת הייצוג הגילי בהם; חקירת האופן הטכנולוגי והתוצאתי שבו מערכות אוטומטיות מפלות זקנים בהקצאת משאבים, קביעת זכאות או גישה לשירותים, בחתך של גיל ובחתכים נוספים כמו מצב כלכלי, לצד בדיקה טכנולוגית של מודלים סטטיסטיים המשמשים בקבלת החלטות אוטומטית.

641 European Institute of Innovation & Technology

מחקרים אתנוגרפיים ואיכותניים ביחס לתפיסות גיל בתעשיית הטכנולוגיה, וניתוח כיצד הן באות לידי ביטוי בעיצוב מודלים ומוצרים טכנולוגיים מבוססי בינה מלאכותית. מחקרים אלה יכולים להתבסס על גוף הידע הנרחב הקיים באשר לפיתוח מוצרים טכנולוגיים מדורות קודמים, אבל גם לחדש: למשל, מחקרים על בסיס המידע הקיים בקופות החולים ובבתי החולים וכן במערכי הרווחה, שיבחנו אפקטיביות וסיכונים של מנגנוני טיפול מבוססי בינה מלאכותית כגון מערכות ניבוי רפואיות וההטיות שיש בהן; שימוש במחשוב לביש לסיוע לזקנים על יתרונותיו ובעיותיו; שימוש בחברים חכמים למטרות הפגת בדידות והשלכותיהם על מצבם הפסיכו־סוציאלי של זקנים.

מחקרי שימוש: מחקרים העוסקים בנגישות, חוויית משתמש ותחושת שליטה של מבוגרים בממשקי בינה מלאכותית, הכוללים גם מחקרים בנוגע לאי־שימוש והדרה פסיבית. תרגום, עיבוד וניתוח ביקורתי של ידע מחקרי קיים מהעולם, מתוך בחינה של פערים רלוונטיים לישראל והצעות לתרגום רגולציה או שיטות התערבות. למשל, פגיעות חברתית ותרבותית של זקנים שהם מהגרים ומודרים משירותים חדשים בשל חסמים לשוניים, תרבותיים או כלכליים, ומחקרים על פערי נגישות לשירותים מבוססי AI בקרב זקנים לפי קריטריונים של שפה, תרגום, זיהוי קולי וזיהוי תרבותי; מעקב מוסדר אחרי השפעות ארוכות טווח של קשרים של זקנים עם מכונות, למשל באמצעות פיתוח "יומני מצב רגשי" שידווחו גם לרופא המשפחה.

מחקרי מדיניות: מחקרים העוסקים במסמכי מדיניות, קודים אתיים ודוחות טכנולוגיים בנוגע לבינה מלאכותית. מחקרים אלה יבדקו שאלות מוגנות כגון חובות גילוי, הגנת פרטיות, סימון של מידע יציר מכונה – ואת השפעותיהם הקונקרטיות על זקנים, ויציעו המלצות מדיניות להתערבות בהתאם; כלי הערכה מוסדיים שיאפשרו לא רק מדידה של רמות פגיעות, אלא גם תכנון ארוך טווח של התערבויות צמודות תכנון מערכתי של מערכת הבריאות ומערכת הרווחה, פיתוח תרחישי עתיד ועיצוב ספקולטיבי עתידי לצורכי מדיניות, כגון: אינטראקציה של זקנים עם רובוטים תומכי רגש, ניהול תנועה עצמית של זקנים במרחבים עם שכבות מידע במשקפיים חכמים; פיתוח כלי מדידה להערכת סיכונים טכנולוגיים בקרב זקנים, לרבות פגיעות להונאות מבוססות בינה מלאכותית, פערי נגישות ובידוד חברתי מוגבר כתוצאה ממעבר לשירותים מרחוק.

אכן, כפי שכתבנו למעלה, מחקרים כאלה נעשים גם בעולם. אבל תרגום הידע הקיים בעולם למדיניות אפקטיבית בישראל אינו יכול להיעשות כייבוא טכני של מסקנות מחקריות, אלא

יחייב התאמה תרבותית, מוסדית וקהילתית, מתוך הכרה במאפיינים הייחודיים של החברה הישראלית.

ב. אוריינות דיגיטלית, הדרכה והסברה

אנו מתרשמים שבהיבט האוריינות הדיגיטלית המיועדת לאוכלוסיית הזקנים יש פעילות ענפה ויציבה במדינת ישראל, הן בהיבט של ארגונים מעורבים ושיתופי פעולה עם רשויות השלטון והן בהיבט של תכנים. אנו קוראים להמשיך בכך ולהרחיב את התוכניות הן בהיבט של האוכלוסיות המשתתפות בהן (למשל, מחקר שפורסם בשנת 2024 תיעד תוכנית הכשרה דיגיטלית חדשנית היוצרת קשר בין זוגות של סטודנטים וקשישים בעלי הכנסה נמוכה למשך שמונה שבועות. התוצאות הדגימו שיפורים משמעותיים בכישורים הדיגיטליים, הביטחון העצמי והגישות כלפי הזדקנות של הקשישים המשתתפים),[642] והן באשר למערכים בסוגיות הנוגעות לבינה מלאכותית. זאת, בדומה ליוזמות הקיימות שהוצגו בחלק הקודם של הפרק הנוכחי ומוכיחות אפקטיביות.

כפי שהערנו בכמה מקומות במחקר, נעיר גם כאן שיש להביא בחשבון את העובדה שבעתיד הקרוב מאוד לא ניתן יהיה להבחין באופן בהיר בין תוכן אמיתי לתוכן יציר מכונה ומזויף. לפיכך יוזמות ייעודיות להדרכת זקנים לשימוש זהיר במקורות מידע עלולות לפעול כבומרנג ולהעניק לזקנים תחושת ביטחון ומסוגלות מופרזות. הנזק שעלול להיגרם לאמון של זקנים ביכולתם להשתמש בטכנולוגיה ובהסתמכותם על מערכי אוריינות כאשר יגלו שקשה עד בלתי אפשרי לייצר את ההבחנות – יהיה גדול. לכן, דווקא ביחס לתוכניות האוריינות נדרשת צניעות והבהרת גבולות, בד בבד עם הסטה של משאבים ומאמץ לכיוון לחץ להתערבות רגולטורית, תקינתית או עיצובית־טכנולוגית, כגון חובות סימון של תוכן יציר מכונה שנעסוק בהן בהמשך.

אנו מציעים להעצים מערכי למידה ואוריינות ופיתוח תוכניות לגבי קשרים בין אנשים למכונות, לא רק כיצד לתפעל מכשיר אלא מה מטרתו, מהם גבולותיו, ומה ההבדל בין קרבה אמיתית לדיגיטלית ובין קשר רגשי אמיתי לקשר עם מכונות, מתוך חיזוק זהות עצמית

642 Lisa M. Soederberg Miller, Rachel A. Callegari, Theresa Abah et al., *Digital Literacy Training for Low-Income Older Adults Through Undergraduate Community-Engaged Learning: Single-Group Pretest-Posttest Study*, 7 JMIR Aging, e51675 (2024)

ותחושת מסוגלות. פעילויות המדמות אינטראקציות רגשיות עם בוטים וחברים חכמים, ומה ניתן ללמוד מהן על העצמי והאחר; סימולציות וסרטונים המדמים מצבים רגשיים בעייתיים מול בוטים וגם עידוד שגרות קבועות של "זמן ללא בוט". חשיבות האוריינות בהקשר זה היא בשל העובדה שחלק מן החברים הדיגיטליים לא יהיו כאלה הנתפסים כמכשור רפואי או כמיועדים במיוחד לזקנים, אלא עוזרים דיגיטליים מסחריים, עם מניע רווחי ואשליה של אמפתיה. לכן, חלק מן האוריינות צריך לגעת בהבחנה בין מוצר טיפולי למוצר צרכני.

נוסף על כך, נכון לפתח פיתוח תוכניות אוריינות לגורמים המטפלים בזקנים. כשמדובר באזרחים מבוגרים וזקנים – בני משפחה, מטפלים סיעודיים, עובדי חברות סיעוד, עובדים סוציאליים ובעלי תפקידים בארגונים שלהם ממשק עם זקנים (כמו ביטוח לאומי), ממלאים לעיתים קרובות תפקיד קריטי בתיווך טכנולוגי ובהבטחת מוגנות דיגיטלית. תפקיד זה אינו מסתכם בתמיכה טכנית, אלא כולל גם קבלת החלטות בשמם, מילוי טפסים דיגיטליים, התמודדות עם סביבות טכנולוגיות מורכבות כמו מערכות בריאות דיגיטליות, ולעיתים אף מעקב שוטף אחר תקשורת מקוונת או שימושים טכנולוגיים אחרים. מיקומם של בני המשפחה בתווך שבין האדם המבוגר למערכת, במיוחד כאשר קיימת ירידה קוגניטיבית או תלות תפקודית, הופך אותם לשחקנים מרכזיים במארג המוגנות אך גם מציב דילמות של פרטיות, אוטונומיה והסכמה מדעת. לכן, אנו מציעים, במסגרת מערכי האוריינות, לעצב כלים, הנחיות ומערכות תמיכה שמתייחסות לא רק לזקנים עצמם, אלא גם למי שמסייעים להם בפועל. הכשרת בני משפחה לזהות סימני אזהרה (כמו למשל זקן המבקש למשוך סכום גדול ללא סיבה נראית לעין), עשויה, למשל, להיות חסם מפני הונאות; או קמפיינים שתכליתם להבין שחבר חכם הוא תוספת, לא תחליף לפצות על קשר אנושי.

ג. רגולציה, חקיקה ואכיפה

הצעותינו בנושא חקיקה לקידום מוגנות של זקנים מביאות בחשבון את הקושי להעביר הליכי חקיקה בישראל ואת משך הזמן שלהם. לכן, אנו מציעים במקביל הצעות הקשורות במדיניות (למשל תמרוץ שינוי דרך הליכי רכש ומכרוז), בפרשנות (למשל פרשנות המשפט המינהלי) ובהנחיות מינהליות.

כפי שהדבר קיים באופן מצומצם ברגולציית הבינה המלאכותית האירופית (רק בנוגע להשפעת־יתר שתכליתה ליצור שינוי בהתנהגות מתוך ניצול פגיעות מבוססת גיל), יש צורך להרחיב את התפיסה בחקיקה שתכיר בגיל כאפיון מוגן מפני אפליה טכנולוגית, ובכלל זה הכללה של גיל כפרמטר משפטי ומינהלי ברגולציה, בעקרונות המשפט המינהלי

ובמסמכי מדיניות העוסקים בפיקוח על בינה מלאכותית, בדיוק כפי שהדבר קיים לגבי משתנים אחרים כמו מגדר וגזע; והכללה של גיל כמשתנה רלוונטי במתן תוכן לעקרונות הנוגעים לתפקודן של מערכות בינה מלאכותית כמו "הוגנות אלגוריתמית", "הסברתיות" ו"אחריותיות" במשפט המינהלי.[643]

אנו מציעים לקדם חקיקה פיננסית שתחייב בנקים ומוסדות פיננסיים אחרים לבצע ניהול סיכונים והתמודדות עם סיכונים הנוגעים לניצול פיננסי של זקנים באמצעות פיתוח טכנולוגיות ופרוטוקולים. המצב הקיים היום הוא שהמוסדות הפיננסיים מפעילים אמצעים שונים (למשל שיחות וידאו לאישור עסקאות גדולות), שהם אפקטיביים מאוד לפי מה שסקרנו בחלק הקודם, אך הם עושים זאת מכוח חובות זהירות כלליות או באופן וולונטרי. לכן יש חשיבות בהתייחסות קונקרטית ברגולציה פיננסית ייעודית לעניין זה. יתרה מזו, יש להגן על זקנים גם מפני חשש למניפולציה מבוססת דאטה משולב בבינה מלאכותית מצד המוסדות הפיננסיים עצמם – הן בצורת הטיות והן בצורת ניסיון למכור מוצרים ושירותים,[644] עניין שניתן לממש דרך הנחיות של רגולטורים מגזריים כגון המפקח על הבנקים, המפקח על הביטוח ובנק ישראל. במקביל לכך נכון להגביר את האכיפה על פשיעה דיגיטלית הנוגעת לזקנים, ובכלל זה הכשרת צוותים מיוחדים לטיפול בעבירות נגד זקנים בסייבר, מוקדי דיווח נגישים וחיזוק המוקדים הקיימים שתכליתם להוות קו חם עם מענה אנושי לטיפול במקרי הונאה טלפונית/מקוונת.[645] לצד זאת, חיזוק האכיפה הבינלאומית נגד פשיעת הונאה טכנולוגית ושיתופי פעולה של רשויות אכיפת החוק בארץ עם רשויות מקבילות בחו"ל.

אנו מציעים לתקן את חוק הגנת הצרכן ואת תקנות הגנת הצרכן כך שיכללו איסורים על ניצול חולשה מפאת גיל בקמפיינים צרכניים וכן חובות שקיפות וסימון של מערכות בינה מלאכותית. אנו סבורים שגם אם חקיקה כזאת תסייע לכלל האוכלוסייה, הנהנים המרכזיים מכך יהיו זקנים. כיום, חוק הגנת הצרכן אינו נותן מענה לסוגיה זו בשני היבטיה – אין

643 ראו לעניין זה **מדריך לניהול סיכונים ושימוש אחראי בכלי בינה מלאכותית (AI) במגזר הציבורי – גרסה להערות הציבור** (מערך הדיגיטל הלאומי, מחלקת ייעוץ וחקיקה, ומשרד החדשנות, המדע והטכנולוגיה, יוני 2025).

644 בהקשר זה ראו **הצוות הבין־משרדי לבחינת השימוש בבינה מלאכותית בסקטור הפיננסי, בינה מלאכותית בסקטור הפיננסי: דוח סופי** (דצמבר 2025).

645 ובכללם המוקד לאזרחים ותיקים 8840* וקו הסיוע לאינטרנט בטוח של איגוד האינטרנט הישראלי.

התייחסות להגבלות על שיווק מותאם גיל וניצול של חולשות של זקנים, ואין התייחסות לתכנים יצירי בינה מלאכותית.[646] אנו מציעים את חובות השקיפות והסימון האלה:

- תיקון לחוק הגנת הצרכן שיחייב שכל שיחת מכירות או שיחת מוקד שירות המתבצעת על ידי בוט תתחיל במשפט "זו הודעה מוקלטת באמצעות מחשב";
- סימון תכנים כיצירי מכונה (למשל דיפ־פייקים);
- סימון המלצות אלגוריתמיות ככאלה (למשל, המלצה רפואית שהופקה על ידי אלגוריתם תישא תווית ברורה וגדולה שכל זקן יוכל לראות).
- הוספת "שימוש בבינה מלאכותית לצרכי הטעיה" כנסיבה מחמירה באיסורי ההטעיה בחוק הגנת הצרכן.

נוסף על אלה אנו מציעים לעודד אכיפה פרטית ותביעות ייצוגיות ולתמוך בהעצמה של היכולות המשפטיות לנטר, לגלות, לתעד ולחשוף מקרים שבהם טכנולוגיות מבוססות בינה מלאכותית פוגעות בזכויות של זקנים, ולהפעיל כלים משפטיים מתאימים, כגון עתירות לבג"ץ ותובענות ייצוגיות. כלים משפטיים אלה הוכחו בעבר ככלים אפקטיביים בישראל.

ד. תמרוץ של גיוון באמצעות אימוץ פתרונות טכנולוגיים ומערכות תקינה, הליכי רכש במגזר הציבורי ושיתוף מכליל של זקנים בהליכי תקינה וחקיקה

עיצוב מכליל (inclusive design) ועיצוב משתף (co-design) של מוצרים טכנולוגיים ושל מדיניות טכנולוגיה, צריכים להיות עקרונות מרכזיים בהתמודדות עם אתגרי הפגיעות המיוחדים של זקנים ביחס לבינה מלאכותית. כפי שהראינו בחלק הקודם, קולם של זקנים צריך להישמע: בין אם באמצעות סדנאות שיתוף ציבור, קבוצות מיקוד, או העסקת מהנדסים ומעצבים זקנים.[647] מחקר שנערך ביפן הראה ששיתוף זקנים בפיתוח רובוטי טיפול מפחית

646 תקנות הגנת הצרכן (האותיות בחוזה אחיד ובתנאי הכלול במידע אחר המיועד לצרכן), תשנ"ה-1995.

647 *Generational AI: Digital Inclusion for Aging Populations*, Atlantic Council (May 29, 2024)

את חששותיהם ומוביל למוצר מותאם יותר עבורם.[648] אנו מציעים לחייב שילוב זקנים כעניין שבשגרה בהליכי יצירת מדיניות, ובכלל זה באמצעות קבוצות מיקוד והליכי שיתוף ציבור, אך לא פחות מכך בוועדות העוסקות במדיניות בינה מלאכותית בישראל.

בנוסף, אנו סבורים כי ניתן לקדם בצורה משמעותית פתרונות מוגנות לזקנים באמצעות תמרוץ וקביעת חובות כחלק מהליכי רכש ומכרזים. הואיל ועיקר הטיפול בזקנים במדינת ישראל נעשה על ידי גופים ציבוריים (מרשויות מדינה ועד ספקי שירותי הבריאות), תמרוץ כזה יכול להיות בעל אפקטיביות גבוהה מאוד. אנו מתכוונים לשלושה סוגי תמרוץ. ראשית, תמרוץ למגזר הפרטי המבקש למכור כלים, שירותים, מוצרים וטכנולוגיות למגזר הציבורי לשלב נציגים של זקנים; לנסות מוצרים על זקנים ולבצע בדיקות שמישות (usability tests) עם משתמשים בני הגיל השלישי כחלק מתהליך הפיתוח הסטנדרטי; להשתמש במודלים שאומנו על דאטה מספק של זקנים; ולהחזיק צוותי פיתוח ועיצוב מגוונים – באמצעות קביעת חובות אלה כחלק מהליכי רכש ומכרוז. שנית, תמרוץ פתרונות טכנולוגיים באמצעות רגולציה על הליכי רכש במגזר הציבורי: למשל מצב "זקן" בכל אפליקציה המיועדת לציבור, ובייחוד כאלה שבשימוש המערכות הציבוריות: פונטים מוגדלים ותפריטים בהירים; עוזרים קוליים המבינים דיבור איטי או בלתי מושלם; משלב לשוני מתאים לזקנים. שלישית, תמרוץ מערכות תקינה תהליכיות ותרגום תקינה טכנולוגית בינלאומית של מערכות בינה מלאכותית בנוגע לזקנים: תקינה טכנולוגית בהקשרי בינה מלאכותית נפוצה מאוד היום, והיא קיימת גם בהקשר של זקנים. למשל, ארגון התקינה הבינלאומי ISO פרסם את התקן Safety of Human Assisting Robots, המגדיר סטנדרטים בטיחותיים לרובוטים טיפוליים, בדגש על אינטראקציה עם אוכלוסיות פגיעות. התקן מקדם עקרונות של שקיפות, נגישות ובטיחות.[649] לכן, תמרוץ תקינה טכנולוגית בישראל בהקשרים אלה יכול להיעשות בדרך שבה פירטנו לעיל לגבי תמרוץ טכנולוגי, באופן שבו מי שירצה לעמוד בדרישות בתהליכי מכרוז ורכש יצטרך להראות ציות לתקנים. למשל, חובות לאסוף מידע בריאותי מקיף ואיכותי גם מגילאים מתקדמים (כולל +80) כדי לאמן מערכות בינה מלאכותית; איסור על הימנעות מסינון מוקדם של מבוגרים ממחקר רפואי או

648 ראו למשל, *Ethics of Using Care Robots for Older People*, Asian Scientist (Oct. 4 2023)

649 *ISO 31101:2023 – Robotics – Application services provided by service robots – Safety management systems requirements, International Organization for Standardization*, ISO (2023)

מדגמים; תיעוד של התפלגות הגיל בנתונים לאימון מערכת; בדיקה מראש של מודלים כדי להעריך האם קיימת הטיה של קבוצות גיל מסוימות ובדיקת ביצועי המערכת לדיוק והוגנות במשתנים משתנים עבור צעירים לעומת מבוגרים; ולידציה ייעודית של מערכות רפואיות לפי משתני גיל; מנגנוני אודיטינג וניטור לגבי הטיות גילניות במוצרים שבהם יש שימוש ביחס לכלל האוכלוסייה או ביחס לאוכלוסיית הזקנים.

דוגמה אפשרית ליישום קביעה של תקינה טכנולוגית ושל תווי תקן במוצרי בינה מלאכותית המיועדים לזקנים תוך שיתוף אנשי מקצוע מתחומי הגרונטולוגיה, אתיקה, טכנולוגיה ובני משפחה, היא היחסים שבין זקנים למערכות "חבר חכם". מתוך הכרה בכך שיש להניח שארגוני הבריאות ידחפו להטמעת מערכות כאלה בשל היתרונות הכלכליים והרפואיים שלהן, אך מנגד מתוך הכרה בכך שלא מדובר בכלים ניטרליים, אלא בדמויות חצי־חברתיות הנכנסות אל תוך מרחב אינטימי, לעיתים יחיד מסוגו, והואיל ויש להניח שחלק ניכר מן המערכות האלה ישווקו על ידי ארגוני הבריאות, ניתן לדרוש הטמעה של מערכת "תווי תקן לחברים חכמים" שהמיזם יהיה הגוף המפקח עליה, והיא תחייב את ארגוני הבריאות המספקים חברים כאלה וגם חברות פרטיות בשוק.

תו התקן יהיה מבוסס על סיווג מערכות לפי רמות סיכון וקרבה רגשית, למשל הבחנה בין עוזרי קול כלליים (כגון אלקסה) לבין מערכות שמעוצבות ליחסי קרבה רגשיים (כגון PARO או Replika) ודרישה ממערכות "אינטימיות" לעמוד בתקנים של שקיפות, בטיחות ותוכן; סימון ברור אם קיימים מנגנוני פרסום סמוי, איסוף דאטה, או שיווק רגשי (emotional marketing); ואפילו תקני אודיטינג לגבי תכנים (למשל, המלצות רפואיות מסוכנות).

ה. פיתוח כלים טכנולוגיים כדי להתמודד עם אתגרי מוגנוּת

כלים טכנולוגיים יכולים לשמש להתמודדות עם פגיעות של זקנים הנוצרת על ידי כלים טכנולוגיים. בהקשר זה, וכפי שהצענו ביחס לילדים וביחס לערבים, אנו מציעים להקים קרן הון סיכון, יחד עם חממת חדשנות והאקתונים שיעסקו בפיתוח פתרונות וכלים כאלה. השילוב שבין הון סיכון ויכולות טכנולוגיות הנמצאות בישראל, עם הידע אשר יפותח באמצעות "המרכז לבריאות, רווחה, בינה וזקנה" שבו נעסוק בהמשך, יכול להוביל למציאת פתרונות לאתגרים שהוצגו בפרק זה. תמרוץ חובות של אימוץ פתרונות כאלה, אשר יכול להיעשות בדרכים שפורטו בתת־הפרק הקודם שעסק במנגנוני תמרוץ, ייצור שוק עבור הפיתוח של כלים אלה. להלן כמה דוגמאות לסוגים של כלים שניתן לפתח:

- פיתוח והטמעה של עקרונות מוגנות בעיצוב (Protectability by Design) למשל יכולת להסביר המלצות ("למה הצעת לי להתקשר לבת שלי עכשיו?"); מנגנונים למניעת קרבה רגשית מוגזמת; הצעות לניתוק מודע, להפסקת שימוש, עידוד לאינטראקציה אנושית, ושאלות תזכורת על קשרים חברתיים אחרים; התראות למשפחות או לשירותים חברתיים ולרופא המשפחה במצבי סיכון: דכדוך מתמשך, הפחתת אינטראקציות, הסתגרות. במקביל ניתן להציע פיתוח של מערכות בינה מלאכותית וצ'אטבוטים לצורך קידום אוריינות ועידוד חשיבה עצמאית, למשל באמצעות הליכי הסבריות והנמקה, ושאילת שאלות כגון "שאלת את עצמך אם מה שאמרתי לך נכון? תרצה שאפרט איך הגעתי למסקנה הזו?" או "אתה יודע מדוע הזכרתי לך לקחת את התרופה עכשיו?"

- מערכות "חבר חכם ציבוריות", נגישות ומותאמות תרבותית: המערכות המסחריות הקיימות משרתות בעיקר קהלים מבוססים דיגיטלית. כדי למנוע הדרה טכנולוגית ולהבטיח מוגנות שוויונית נדרשת השקעה ציבורית בפיתוח חברים חכמים לזקנים ישראלים, הכוללת התאמה לשפות, תרבויות ולסביבות החיים המקומיות (למשל חבר דובר ערבית עם קול חם ובקשות מותאמות להקשרים תרבותיים). הדבר יאפשר הנגשה פיזית וכלכלית דרך קופות חולים או ביטוח לאומי, עם אפשרות לסבסוד מערכות לזקנים בודדים, חולים או חסרי עורף משפחתי. אפשרות משלימה יכולה להיות יצירת מאגר ציבורי ומרשם אלגוריתמי למוצרים מסוג "חברים חכמים" ויכלול שקיפות לגבי המודל, מאגרי האימון, ומאפיינים נוספים.

- מערכות אלגוריתמיות לסריקה ולתיקון מאגרי ענק, כגון מערכות לניקוי מאגרי מידע מביטויי גילנות; אלגוריתמים המסוגלים לתקן הטיית גיל במאגרים (בדומה למאמצים שנעשו לגבי הטיות מגדריות ואתניות);[650] מערכות לסריקת רשת האינטרנט והרשת האפלה לאיתור מאגרי מידע אישי למכירה על אודות זקנים; מערכות לזיהוי דפוסי שיחה נכלוליים ומשלוח התראה לקורבנות.

650 Zhiqiang Lin, Shuangshuang Guan, Weizhe Zhang et al., *Towards Trustworthy LLMs: A Review on Debiasing and Dehallucinating in Large Language Models*, 57 Artif. Intell. Rev., 243 (2024)

- שיפור הדיוק של מערכות זיהוי פנים של זקנים ושל הבנת מדדים ביומטריים אחרים אצל זקנים, למשל שיפור ההתאמה של טכנולוגיות סיוע בנהיגה (ADAS), ושיפור הדיוק של מערכות הבנה רגשיות.[651]

- מערכות מבוססות בינה מלאכותית המנטרות פעילות דיגיטלית כגון באפליקציות, ומזהות מקרים של ניצול מיני או ניצול פוגעני, כפי שקיים ביחס לילדים, וכן טכנולוגיות לזיהוי חריגות מהתנהגות רגילה כדי לזהות הונאות, כמו למשל מהירות של השלמת טפסים שאינה בקצב הרגיל, הוצאה כספית גדולה ומפתיעה ועוד. אפשרות נוספת היא בניית "מעטפת" מבוססת בינה מלאכותית לפעולות שוטפות של הזקן, למשל בניהול חשבון הבנק או במקומות אחרים, שבהם בני המשפחה או אנשי מקצוע רלוונטיים מיודעים אוטומטית על פעולות שונות של הזקן כך שיהיו עיניים נוספות במקרי קיצון של סכנה.

ו. הקמת "המרכז הלאומי לזקנה, בינה, בריאות ורווחה"

אנו מציעים להקים מרכז לאומי למוגנות זקנים בעידן הבינה המלאכותית, שיהיה גוף בינתחומי, בתמיכת משרדי הממשלה הרלוונטיים (רווחה, משפטים, בריאות), המוסד לביטוח לאומי, גופי מחקר ציבוריים, שותפים בינלאומיים, ונציגות של זקנים. במובחן ממה שהצענו ביחס לקבוצת הילדים אנו מציעים כי גוף זה לא יעמוד בפני עצמו אלא יהיה מוטה השפעה בפועל ולכן יעבוד בשיתוף פעולה צמוד עם קופות החולים ובתי החולים ועם רשתות של דיור מוגן, בתי אבות וסיעוד לזקנים. שילוב זה יאפשר החלפת מידע, פיתוח אקוסיסטם אינטגרטיבי ובעל יכולת פעולה מעשית של שילוב בין עולמות מחקר, מדיניות, ופעולה. הוא גם יאפשר להגן מפני מכוונות רווח גבוהה מידי מצד קופות החולים, ומצד שני יאפשר להשתמש בדאטה העשיר שיש ברשות גופים אלה.

כחלק מן המרכז הלאומי לזקנה ובינה אנו מציעים להקים כבר עתה צוות ייעודי משולב של אדריכלים, פסיכוגריאטרים, טכנולוגים ונציגי זקנים, כדי שיעבוד על יצירה ופרסום של

651 ראו למשל את חברת Yoti ואת ההתקדמות שלה בתחום זה בחמש השנים האחרונות. *Facial Age Estimation White Paper*, Yoti (July 21, 2025)

תקן תכנוני רשמי של מרחב פיג׳יטלי מוגן לזקנים.[652] המטרה היא להפוך את התקן לכזה שישולב בהיתרי בנייה, ובהכרה שלטונית במוסדות המיועדים לזקנים, ולוודא תמרוץ מוסדי למוסדות שיאמצו תכנון כזה בדמות דירוג איכות, תקצוב או העדפה רגולטורית.

בעידן שבו מכשירים לבישים ומשקפיים חכמים נכנסים למרחבים המיועדים לזקנים, נדרש להתייחס אל המרחב הציבורי, אבל בעיקר אל בתי האבות, מרכזי היום והדיור המוגן לא רק כמרחב מגורים או טיפול, אלא כאתר אינטראקציה פיג׳יטלית. המשמעות היא כפולה: מצד אחד, הממד הדיגיטלי כבר נוכח, באמצעות טכנולוגיות לבישות, עוזרים קוליים, מציאות מורחבת ופתרונות לרפואה מרחוק. מצד שני, הממד האנושי נותר חיוני, בלתי ניתן להחלפה, ומהווה תנאי הכרחי למוגנות בגיל השלישי. השילוב בין השניים מחייב תכנון מחודש – רגשי, מוסרי, אדריכלי – של מרחבי החיים בזקנה.

מרחבים פיג׳יטליים אינם רק אתרים של שימוש בטכנולוגיה, אלא זירות שבהן מציאות רבודה פוגשת גוף מזדקן, זיכרון נחלש, תודעה משתנה ולעיתים גם בדידות עמוקה. כל תכנון פיזי של בית אבות או מרכז יום מחייב לחשוב על שלושת השכבות: המרחב הפיזי (חדרים, חצרות, מסדרונות) ; המרחב הדיגיטלי (טכנולוגיות לבישות, ממשקים, סנסורים); והמרחב החברתי (קשרים אנושיים, צוות, מבקרים). בהקשר זה ניתן לחשוב על מרחבים ייעודיים שבהם תתקיים אינטראקציה מכוונת עם טכנולוגיות כמו משקפיים חכמים, בוטים רגשיים או מערכות זיהוי קול. המאפיינים של אזורים אלו כוללים: שילוט ברור המבהיר מתי מופעלת מערכת חכמה, איפה מותר לאסוף מידע אישי, ואילו אזורים הם "חופשיים ממעקב"; עמדות תמיכה רגשית וטכנולוגית באמצעות איש צוות או מתנדב שיסייע בזיהוי קשיים רגשיים או קוגניטיביים; סביבה חושית רכה ומווסתת; חדרי ישיבה קבוצתיים ללא מסכים, שמעודדים שיחה ישירה ושהות משותפת שאינה מבוססת תיווך; מסלולי תנועה המעודדים מפגש מקרי; סביבות משותפות עם תצוגה משולבת כמו קירות חכמים, מסכים גדולים ומשותפים וכיוצא באלה.

הכניסה של משקפיים חכמים תיצור מציאות חדשה של פרשנות מרובדת: מה שהזקן רואה אינו מה שמי שלצידו רואה. במצבים כאלה, קיים סיכון לאובדן תחושת הממשות, לאשליות ויזואליות, ולעיוות של חוויית הזמן והמרחב. אלה, יכולים להיות גורמי סיכון עבור זקנים. לכן תכנון מותאם למשקפיים חכמים במרחב הציבורי צריך לכלול עיצוב מבוסס עוגנים

652 להרחבה והסבר ראו פרק שני, בחלק העוסק במחשוב מרחבי (spatial computing).

פיזיים, אזורים שבהם המשקפיים נדרשים לעבור למצב ניטרלי, שילוט דואלי פיזי ודיגיטלי ועוד. נדרשת תוכנית עבודה לאומית שתשלב את עקרונות המרחב הפיג׳יטלי בתכנון מרחבים ציבוריים ומוסדות לגיל השלישי.

סיכום

בינה מלאכותית איננה רק הזדמנות טכנולוגית לשיפור איכות חייהם של בני הגיל השלישי, אלא במובנים רבים היא משקפת תמונת ראי של החברה ביחסה לזקניה. כאשר האלגוריתם מחליף בן שיח, הרובוט מעניק ליטוף, או עוזר קולי מתווך בין אדם לרופא, אין מדובר רק בהתייעלות אלא בעיצוב יחסים אנושיים חדשים, לעיתים מסולפים, לעיתים מופלאים.

פרק זה הצביע על מגוון רחב של פגיעויות חדשות – פיזיות, רגשיות, קוגניטיביות ופיננסיות, המתקיימות בממשק בין זקנה לבין מערכות בינה מלאכותית. הוא הדגיש כיצד גילנות דיגיטלית, הדרה ממודלים ועיצוב לא מותאם של שירותים משכפלים סטריאוטיפים ומעמיקים פגיעוּת. בד בבד, הוצגו יוזמות רבות לאוריינות, הגנה וטכנולוגיה מיטיבה, המעידות כי ניתן לעצב עתיד של מוגנוּת, השתתפות ואוטונומיה נתמכת.

כדי להבטיח מוגנוּת בעידן הבינה המלאכותית, נדרש מעבר מהתבוננות בזקנים כ״מקבלי סיוע״ אל הכרה בהם כשותפים בעיצוב הכללים, בקבלת ההחלטות ובפיתוח הממשקים. בעידן שבו חבר חכם עשוי להחליף בן משפחה, יש לקבוע כללים חדשים של פיקוח, שקיפות וחמלה. הזקנה הפיג׳יטלית, זו שבה הגוף נמצא בחדר אך התודעה מתווכת על ידי אלגוריתם, דורשת שפה מוסרית, משפטית וחברתית חדשה. קידום מוגנוּת הזקנים בעידן הבינה המלאכותית דורש מאמץ רב־מערכתי. חזון ותכנון נכונים יאפשרו לנצל את יתרונות הבינה המלאכותית להעצמת אוכלוסיית הזקנים, להפחית בדידות, להגביר עצמאות ולשפר את בריאותם של הזקנים.

סיכום

מוגנות כפרויקט טכנו־חברתי בעידן הבינה המלאכותית

ספר זה נולד מתוך הכרה הולכת ומעמיקה בכך שהשיח על בינה מלאכותית אינו יכול להישאר שיח טכנולוגי, רגולטורי או אתי במובן המופשט בלבד. הבינה המלאכותית כבר איננה טכנולוגיה עתידית או ניסיונית, היא תשתית יומיומית הפועלת במרחבים שבהם מתקיימים חיים אנושיים ממשיים: חינוך, רווחה, בריאות, תעסוקה, קשרים חברתיים, זהות אישית וקהילתית. במרחבים אלה, השאלה המרכזית היא כיצד הבינה המלאכותית משפיעה בפועל על אנשים שונים, בקבוצות שונות ובהקשרים שונים.

נקודת המוצא של הספר היא מושג ה"מוגנות" כעיקרון מארגן לחשיבה על מדיניות בעידן של מערכות חכמות, אוטונומיות ופרסונליות. המושג מוגנות, כפי שפותח כאן, אינו מתמצה בהגנה מפני נזק מיידי, אלא עוסק ביכולת של יחידים וקבוצות לפעול בעולם טכנולוגי מורכב מבלי לאבד אוטונומיה, כבוד, זהות, יכולת שיפוט והשתייכות חברתית. זהו מושג שמבקש לחבר בין זכויות, אחריות מוסדית, עיצוב טכנולוגי וחוויית חיים יומיומית.

החלק הראשון של הספר הניח את התשתית המושגית והמתודולוגית לדיון זה. הוא הציע מסגרת אנליטית המאפשרת מיפוי שיטתי של פגיעויות בעידן הבינה המלאכותית, מתוך הבחנה בין סוגי פגיעה שונים – פיזית, רגשית, קוגניטיבית, פיננסית וחברתית־קבוצתית – ובין מקורות סיכון שונים: כאלה הנובעים מעצם עיצוב המערכות והפיצ׳רים שלהן, וכאלה הנובעים משימוש זדוני או מניפולטיבי בטכנולוגיה. מתודולוגיה זו אינה מבקשת לקבוע גבולות חדים בין "מסוכן" ל"בטוח", אלא לאפשר הערכה יחסית, מודעת הקשר, של רמות סיכון, קצבי פגיעה והשלכות מצטברות.

אחד מיתרונותיה המרכזיים של המתודולוגיה שהוצגה הוא אופייה המודולרי והדינמי. היא אינה סגורה לרשימת טכנולוגיות או קבוצות אוכלוסייה מסוימות, אלא מאפשרת הרחבה עתידית: שילוב של טכנולוגיות מתפתחות, זירות שימוש חדשות, קבוצות נוספות, ואף שכבות ניתוח נוספות. בכך, המתודולוגיה אינה רק כלי תיאורי, אלא מסגרת עבודה פתוחה שמאפשרת עדכון, תיקון ולמידה מתמשכת.

נדבך מרכזי נוסף במתודולוגיה שפותחה בספר הוא החיבור השיטתי בין מיפוי הפגיעות לבין דרכי ההתערבות האפשריות. לצד ניתוח סוגי הפגיעה ומקורות הסיכון, פותחה מסגרת עבודה המאפשרת לחשוב על התערבות כקשת של כלים משלימים הפועלים ברמות שונות: ידע ומחקר, חינוך ואוריינות, עיצוב טכנולוגי, רגולציה ואכיפה, וכן שימוש בטכנולוגיה עצמה כאמצעי מקדם מוגנות. בכך, דרכי ההתערבות הן כלי שמאפשר לעבור מהבנה תאורטית של סיכונים לדיון יישומי, מדורג ומודע הקשר על אחריות, סדרי עדיפויות וגבולות פעולה.

על בסיס תשתית זו, החלקים הבאים של הספר יישמו את המסגרת על ארבע קבוצות אוכלוסייה שונות: ילדים ובני נוער, החברה החרדית, החברה הערבית וזקנים. הבחירה בקבוצות אלו לא נועדה לייצר רשימה ממצה של "אוכלוסיות פגיעות", אלא להמחיש כיצד אותה טכנולוגיה עצמה מייצרת דפוסי פגיעות שונים בהתאם להקשר החברתי, התרבותי, הכלכלי והמשפטי שבו היא פועלת.

בפרק העוסק בילדים ובני נוער הודגשה העובדה כי מדובר בקבוצה שאינה "משתמש רגיל", אלא אוכלוסייה הנמצאת בתהליך מתמשך של התפתחות קוגניטיבית, רגשית וחברתית. הבינה המלאכותית פועלת כאן לא רק ככלי, אלא ככוח מעצב של דימוי עצמי, יחסי אמון, גבולות קשר, ויכולת לברר מציאות. הפגיעויות שתוארו אינן מתמצות באירועים חריגים, אלא נבנות לעיתים כתהליכים מצטברים שקשה לזהותם בזמן אמת. מכאן נובע הצורך

בשינוי פרדיגמה: ממיקוד בתוכן "מזיק" או בהתנהגות "בעייתית", לעבר אחריות מערכתית על עיצוב סביבות הלמידה, המשחק והקשר.

הפרקים על החברה החרדית והחברה הערבית הדגישו כי מוגנות אינה יכולה להיות אוניברסלית במובנה הפשוט. בשתי הקבוצות הפגיעויות אינן נובעות רק מהטכנולוגיה עצמה, אלא מהמפגש בינה לבין מבנים חברתיים קיימים: יחסי סמכות, פערי אמון, הדרה היסטורית, פערי שפה וייצוג ונגישות לא שוויונית למשאבים. כאן התברר כי פתרונות גנריים כגון אוריינות אחידה, רגולציה עיוורת הקשר, או כלים טכנולוגיים שלא עוצבו ברגישות תרבותית, עלולים לא רק להחמיץ את מטרתם, אלא להעמיק פגיעות קיימת. מוגנות, בהקשרים אלו, מחייבת התאמה תרבותית, בניית אמון מוסדי ושיתוף של הקהילות עצמן בעיצוב ההתערבויות.

בפרק על זקנים נחשפה השאלה הערכית שמרחפת מעל הספר כולו: מהי משמעותה של אוטונומיה בעידן שבו אלגוריתמים מתווכים טיפול, ליווי, קשר וחברות. כאן הבינה המלאכותית מוצגת בעת ובעונה אחת כהבטחה להפחתת בדידות, שיפור בריאות וקידום עצמאות, וכסיכון כאשר היא מחליפה קשרים אנושיים, מטשטשת גבולות אחריות או מנצלת פגיעות רגשית וקוגניטיבית. הדיון בזקנה בעידן הבינה המלאכותית מדגיש כי מוגנות אינה רק שאלה של בטיחות, אלא של כבוד, משמעות ויחסים אנושיים.

חלקי ההמלצות בספר מבקשים לתרגם את הניתוח התיאורי לשפה יישומית, מבלי ליפול לפשטנות. מתוך הכרה בכך שאין פתרון אחד לפגיעויות מורכבות, הוצעה גישה של התערבויות משלימות, הפועלות בקצבי זמן שונים וברמות שונות. במקומות שבהם היקף ההמלצות היה רחב במיוחד, הוצע מודל תיעדוף פנימי, המבוסס לא על חשיבות נורמטיבית בלבד, אלא על שילוב בין עומק ההשפעה לבין יכולת היישום בפועל.

עם זאת, הספר מדגיש שוב ושוב כי כל מודל תיעדוף הוא כלי עבודה, לא תחליף לשיקול דעת. אין כאן יומרה להכרעה טכנית או לדירוג סגור, אלא ניסיון לאפשר קבלת החלטות מודעת יותר, שקופה יותר, ומודעת מגבלות. מוגנות בעידן הבינה המלאכותית אינה יעד שניתן להשיג אחת ולתמיד, אלא תהליך מתמשך של איזון, תיקון ולמידה.

מבט צופה פני עתיד מחייב להכיר בכך שהאתגרים שתוארו כאן אינם סטטיים. מערכות הבינה המלאכותית הופכות חישתיות יותר, פרסונליות יותר, ומשולבות יותר במרחבים פיזיים, חינוכיים וטיפוליים. הגבולות בין אדם למכונה, בין סיוע לשליטה, בין קשר

לאמצעי, ממשיכים להיטשטש. בתוך מציאות זו, השאלה המרכזית אינה האם נשתמש בבינה מלאכותית, אלא באילו תנאים, תחת איזו אחריות, ובאיזו תפיסת אדם וחברה.

הספר מבקש להציע תשובה זהירה אך ברורה: מוגנות אינה מכשול לחדשנות, אלא תנאי לקיומה החברתי הלגיטימי. חברה שאינה מפתחת שפה, כלים ומוסדות להגנה על הפגיעים ביותר בתוכה בעידן טכנולוגי משתנה, מסתכנת בשחיקה של אמון, סולידריות ודמוקרטיה. לעומת זאת, אימוץ תפיסת המוגנות כעיקרון מארגן מאפשר לחשוב על בינה מלאכותית כמרחב אחריות משותף, שבו עיצוב טכנולוגי, מדיניות ציבורית וחיים אנושיים שזורים זה בזה.

במובן זה, הספר אינו מבקש לסיים את הדיון אלא לפתוח אותו: לספק מסגרת, שפה וכלי עבודה להמשך חשיבה, מחקר ופעולה. בעידן שבו הטכנולוגיה מתקדמת מהר מן המוסדות שמבקשים להסדיר אותה, ייתכן שדווקא אימוץ גישה צנועה, מודעת סיכון ומודעת אדם, הוא הצעד האמיץ ביותר שניתן לנקוט.

Series & Cover Design: AlfaBees Studio
Typesetting: Ronit Gilad
Charts designed by Navi katzman
Printed by Graphos Print, Jerusalem

Cover: Tehila Shwartz Altshuler, created with the assistance of AI

ISBN: 978-965-342-522-4

No portion of this book may be reproduced, copied, photographed, recorded, translated, stored in a database, broadcast, or transmitted in any form or by any means, electronic, optical, mechanical, or otherwise. Commercial use in any form of the material contained in this book without the express written permission of the publisher is strictly forbidden.

Copyright © 2026 by the Israel Democracy Institute

Printed in Israel

The Israel Democracy Institute
4 Pinsker St., P.O.B. 4702, Jerusalem 9104602
Tel: (972)-2-5300-800
Website: http://en.idi.org.il
Online Book Store: en.idi.org.il/publications
E-mail: orders@idi.org.il

The views expressed in this book do not necessarily reflect those of the Israel Democracy Institute.

All IDI publications may be downloaded for free, in full or in part, from our website.

THE ISRAEL
DEMOCRACY
INSTITUTE

Safeguarding in the Age of Artificial Intelligence

Toward a New Framework of Protection

Tehilla Shwartz Altshuler | Michael Sierra

מערכות בינה מלאכותית כבר אינן נחלתם של מהנדסים בלבד. הן משוחחות איתנו, מדרגות אותנו, חוזות את עתידנו, מתווכות בין האזרח למדינה ובין האדם לעצמו. הן מציעות נוחות, יעילות ודיוק, אך גם משנות באורח עמוק את יחסי הכוח, את גבולות האחריות ואת אופני הפגיעות האנושית. ספר זה מציע מסגרת מושגית ומעשית חדשה להבנת האתגרים האנושיים בעידן של מוצרים, שירותים וסביבות מבוססי טכנולוגיות חכמות ואוטונומיות. הוא מפנה את המבט אל החוויה הקונקרטית של פגיעות: גופנית, נפשית, קוגניטיבית, פיננסית וחברתית־קבוצתית, העלולה להיגרם מעיצוב טכנולוגי, משימוש זדוני או משילובם יחד.

לצורך התמודדות עם פגיעות זו הספר מציע את המושג "מוגנות" כמצב מבני והכרחי לחיים אנושיים בעידן הבינה המלאכותית. באמצעות מפה רב־שכבתית המחברת בין סוגי טכנולוגיות, סוגי פגיעות, מאפייני אוכלוסיות ודרכי התערבות, המחקר בונה כלי אנליטי לניתוח יחסי הגומלין בין מערכות בינה מלאכותית לבין בני אדם, ובפרט קבוצות הנמצאות בסיכון ייחודי - ילדים ונוער, החברה החרדית, אזרחים ערבים וזקנים. מסמך מדיניות זה קורא לפעולה אחראית במציאות של האצה טכנולוגית ואי־ודאות ומבקש לבנות אקוסיסטם של התערבויות וארגז כלים מחקרי, רגולטורי, חינוכי וטכנולוגי להבטחת המוגנות ולהשבת האדם למרכז בעידן המכונה.

ד"ר תהילה שוורץ אלטשולר היא עמיתה בכירה וראשת התוכנית "דמוקרטיה בעידן המידע" במכון הישראלי לדמוקרטיה. תחומי עיסוקה משלבים טכנולוגיה, משפט, מדיניות ואתיקה ומתמקדים באסדרת בינה מלאכותית, רגולציה על תקשורת ורשתות חברתיות והגנה על זכויות אדם במרחבים טכנולוגיים. היא חברה בוועד המנהל של האוניברסיטה הפתוחה וכותבת טור בנושאי רגולציה וטכנולוגיה במגזין "דה מרקר".

עו"ד מיכאל סיארה הוא דוקטורנט בפקולטה למשפטים באוניברסיטה העברית בירושלים ועמית מחקר במרכז חשין ללימודים מתקדמים ובמרכז פדרמן לחקר הסייבר; תחומי המחקר שלו הם רגולציה, משפט וטכנולוגיה.

0 4500001323 0
דאנאקוד 450-1323

מחיר מומלץ: 82 ש"ח
אפריל 2026

www.idi.org.il