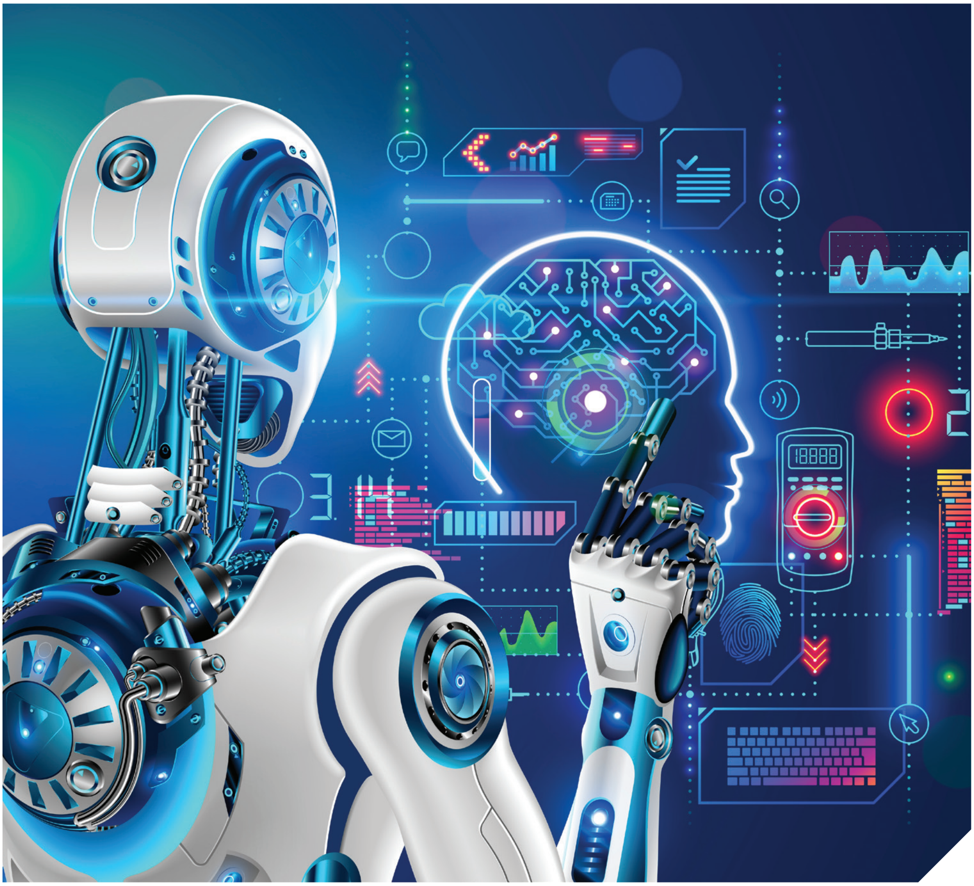


אדם, מכונה, מדינה

לקראת אסדרה של בינה מלאכותית

עמיר כהנא | תהילה שוורץ אלטשולר

עיון

המכון הישראלי לדמוקרטיה

המכון הישראלי
לדמוקרטיה

אדם, מכונה, מדינה
לקראת אסדרה של בינה מלאכותית

עמיר כהנא | תהילה שוורץ אלטשולר

Human, Machine, State:
Toward the Regulation of Artificial Intelligence
Amir Cahane | Tehilla Shwartz Altshuler

עיצוב הסדרה והעטיפה: סטודיו Alfabees
ביצוע גרפי: נדב שטכמן פולישוק
עיצוב התרשימים: דנה ברגר
הדפסה: דפוס מאור ולך, ירושלים

התצלום על העטיפה: Shutterstock

מסת"ב ISBN 978-965-519-433-3

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), 2023
נדפס בישראל, תשפ"ג/2023

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

להזמנת ספרים:
החנות המקוונת: www.idi.org.il/books
דוא"ל: orders@idi.org.il
טל': 02-5300800

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 המתפרסמים בספר זה אינם משקפים בהכרח את עמדת המכון הישראלי לדמוקרטיה.

תוכן העניינים

תקציר 7

מבוא 17

פרק ראשון
מהי בינה מלאכותית? 25

פרק שני
שימושים שונים של בינה מלאכותית 57

פרק שלישי
מדיניות אתיקה של בינה מלאכותית 95

פרק רביעי
רגולציה של בינה מלאכותית: סקירה השוואתית 119

פרק חמישי
בינה מלאכותית וזכויות אדם: חזית חדשה במדיניות טכנולוגיה 161

פרק שישי
"אחרי רבים להטות": הטיות אלגוריתמיות 179

פרק שביעי
שקיפות אלגוריתמית 203

פרק שמיני
פיקוח מוסדי על בינה מלאכותית 219

פרק תשיעי
יסודות לרגולציה מכוונת זכויות של בינה מלאכותית בישראל:
עקרונות כלליים 231

פרק עשירי
יסודות לרגולציה מכוונת זכויות של בינה מלאכותית בישראל:
ארגז כלים 257

פרק אחד עשר
המלצות לגבי מוסד מאסדר לבינה מלאכותית בישראל 279

פרק שנים עשר
סיכום 305

Abstract iii

תקציר

בשנים האחרונות מערכות בינה מלאכותית הולכות ונשזרות ברקמת חיי היום־יום. הן ממליצות על נתיבי נסיעה או על השיר הבא שיושמע, תומכות באבחון רפואי, ולאחרונה אף משתתפות באופן פעיל בהכנת שיעורי הבית. גופים ציבוריים ברחבי העולם מטמיעים מערכות אלגוריתמיות שמקבלות החלטות מינהליות הנוגעות להקצאת משאבים, לתכנון, לחיזוי פשיעה או להגנה על המרחב הציבורי, או תומכות בהחלטות כאלה. מעוזרים אישיים דיגיטליים ועד מכוניות אוטונומיות, מרובוטים המבצעים משימות פשוטות ועד מערכות מעקב, זיהוי וחיזוי.

ואולם על אף היתרונות הגלומים במערכות אלגוריתמיות – ללא בקרה על השימוש בהן, על פיתוחן ועל פרישׂתן הן עלולות לסכן זכויות אדם וחירויות יסוד. סכנות אלו יכולות לבוא לידי ביטוי בכל שרשרת הערך של פיתוחן והשימוש בהן – החל בשלב הגדרת תכליתה של המערכת, עבור בהסתמכות על מידע חלקי, שגוי, מזוהם או מוטה, וכלה באי־החלה של בקרות מאוחרות על תוצרי המערכת. יתר על כן, ככל שהבינה המלאכותית הולכת ומתפתחת כך שמערכותיה נוטות להציג יכולות חדשות שהמפתחים שלהן לא התכוונו להן או לא חזו אותן. מקצת היכולות מעוררות השראה ומקצתן בעלות פוטנציאל נזק, כגון היכולת לבצע פעולות סייבר התקפיות, לתמרן אנשים באמצעות שיחה או להפיץ מידע מלאכותי מוטעה ומטעה. היכולת לזהות יכולות אלו ולהגביל את הסיכונים שהן מביאות איתן נעשתה אפוא אתגר חשוב מאין כמוהו.

מהי בינה מלאכותית, מהם יתרונותיה ואילו חששות היא מעוררת, בעיקר כשהיא משמשת את רשויות השלטון? שאלות אלו הן עניינו של ספר זה.

מקבלי ההחלטות, התעשייה, האקדמיה וארגוני החברה האזרחית ברחבי העולם ובישראל זיהו שהבינה המלאכותית היא טכנולוגיה מתפרצת שיש להיערך אליה באמצעות אסטרטגיה לאומית ומדיניות רגולטורית. בשלהי העשור הקודם פורסמו עשרות מסמכי אתיקה שעסקו בבינה מלאכותית וניסו להניח עקרונות לפיתוח מערכות אלגוריתמיות, לשימוש בהן ולהטמעתן. את הערכים האתיים המוצעים בהם אפשר להעמיד על שבעה עקרונות: שקיפות; הוגנות; מניעת נזקים ובטיחות; אחריות ואחריותיות; פרטיות; קידום הטוב וערך האדם במרכז; חירות ואוטונומיה.

ואולם לא די בעקרונות אתיים. כדי להבטיח שמירה על זכויות אדם וחירויות יסוד יש לעגן עקרונות אלו בחקיקה. ואכן, ברחבי העולם אנו רואים ניצני חקיקה המבקשת לאסדר את השימוש בטכנולוגיות בינה מלאכותית ואת פיתוחן: הן על פי מודל של חקיקה רוחבית החלה על מערכות בינה מלאכותית באשר הן, כדוגמת תקנות הבינה המלאכותית האירופיות, הן כשמיכת טלאים רגולטורית המאסדרת מגזרים ושימושים קונקרטיים של בינה מלאכותית, כדוגמת הדינים הייעודים שתכליתם להתמודד עם אפליה אלגוריתמית במערכות לגיוס ולקידום עובדים.

ספר זה מציע עקרונות מנחים ליצירת מדיניות בינה מלאכותית מכוונת זכויות וארגז כלים לאסדרת בינה מלאכותית.

המלצות עיקריות

עקרונות מנחים ליצירת מדיניות בינה מלאכותית מכוונת זכויות

האדם במרכז. התכלית המרכזית של הפיתוח של בינה מלאכותית ומערכות לומדות צריכה להיות שירות המין האנושי – הן כקולקטיב הן כפרטים – באופן המיטיב עימו. עיקרון זה של חירות האדם והאוטונומיה שלו מצטרף לעקרון קידום הטוב ועקרון מניעת הנזק ומבסס ביתר שאת את החשיבות של עקרון האדם במרכז, חירות ואוטונומיה. המשמעות המעשית של העמדת האדם במרכז היא שפיתוחן של מערכות נבונות, פרישׂתן והשימוש בהן ייעשה בפועל על יסוד התפיסה הרואה בהגנה על זכויות יסוד ועל חירות האזרח עיקרון ראשון במעלה, ולא מס שפתיים שנועד להכשיר העדפת עקרונות כגון "קידום חדשנות", אינטרסים כלכליים כגון עידוד מגזר ההייטק או "הפיכת ישראל למובילה טכנולוגית גלובלית", ואף לא ייעול תהליכים במגזר הציבורי.

הדמוקרטיה במרכז. למערכות מבוססות בינה מלאכותית יש פוטנציאל רב לפגוע בדמוקרטיה במובנה הרחב, הן באמצעות השפעה על השיח הציבורי ופיזור של רעיונות, הן באמצעות מכשור לשליטה, למעקב, לזיהוי ולמשטור, הן באמצעות היכולת שלהן לשמש לזריעת ספק ולערעור עצם היכולת לברר את המציאות ולהבחין בין מקור לזיוף ובין אמת לשקר. לכן יש לתת משקל רב לעיקרון "הדמוקרטיה במרכז" גם אם הדבר פוגע לעיתים בהתקדמות טכנולוגיות או בחדשנות.

אוריינות דיגיטלית של מקבלי החלטות. אוריינות דיגיטלית משמעה היכולת לנתח את השוק ולהבין לאן מתפתחת הטכנולוגיה, לפחות בטווח הקצר. למשל: היכן נמצאים כספי המחקר והפיתוח של חברות הענק ומהם הפטנטים שהן רושמות כדי לעגן פיתוחים טכנולוגיים חדשים. היא כוללת גם הבנה של הדרכים המסחריות והרגולטוריות להכוונת פיתוח טכנולוגי, ומכאן גם הבנה של האחריות המוטלת על מעצבי מדיניות להשפיע על התפתחות הטכנולוגיה ולא רק להתבונן בה מן הצד. נדרש רובד ביניים בין הבנת הטכנולוגיה לבין יצירת

מדיניות בעניינה, מסגרת להבנת המשמעויות של מערכות טכנולוגיות, ליכולת לדמיין את האפשרויות החדשות שהן מביאות איתן ולעמוד על השלכותיהן על המוסר החברתי ועל שלד השיטה המשפטית. כיוון שמסגרת זו חסרה לעיתים קרובות, נוצרים פערי הבנה – בייחוד בנושאים בעלי השלכות רחבות כמו בינה מלאכותית.

המשגה חדשה של מערכות ויכולות בתחום הבינה המלאכותית. המושג בינה מלאכותית הוא מטפורה פוליטית־טכנולוגית רבת עוצמה. ההשוואה בין מערכת בינה מלאכותית למוח האנושי יוצרת קרבה ודמיון, והללו מובילים להטמעה חברתית של הרעיון שהמכונות עובדות כמו מוח אנושי, מבצעות פעולות אנושיות בדרך שבה אנשים מבצעים אותן, ולמעשה מתחרות בבני האנוש. אנו מציעים להמשיג את פעולות המכונות ואת התכונות המיוחסות להן בדרך שאינה תלויה בהשוואה זו.

פיתוח זכויות למושאי החלטות של בינה מלאכותית. נדרשת חשיבה מחודשת על זכויות יסוד משתי בחינות. ראשית, יש לצקת משמעות חדשה לתאוריה החוקתית של זכויות האדם; שנית, יש ליצור זכויות דיגיטליות חדשות, שלא היה צורך בהן בעבר, ובעיקר זכויות של פרטים בשעה שהם באים במגע עם מערכות ממוכנות מבוססות אלגוריתמים.

אין להשלים עם מצב שבו ישראל היא חצר אחורית. בשנים האחרונות מקודמות הצעות לאסדרת בינה מלאכותית במדינות מובילות, וגם באיחוד האירופי, ויש להניח שמדובר בתחילתה של מגמה עולמית. גם אם יש שוני ערכי בין האסדרות במקומות שונים – אם בבחירת סוג המערכות המוגדרות מסוכנות ואם במטרות המוצהרות של החקיקה עצמה – יש להן גם הרבה מן המשותף. לכן אין להשלים עם מצב שבו במדינת ישראל לא תהיה חקיקה שתהיה בהרמוניה עם החקיקה המקובלת בעולם.

הרמוניה איננה בהכרח זהות. כאשר הפיתוח מכוון לייצוא ממילא נדרש הסטנדרט הרגולטורי לעמוד לכל הפחות בסטנדרטים הזרים, שהם לרוב גבוהים יותר מהסטנדרטים בישראל. אבל גם כאשר הפיתוח מכוון לשוק המקומי, ומושאי ההחלטות של מערכות הבינה המלאכותית אינם אזרחי חוץ אלא אזרחי המדינה, אין להנמיך את הסטנדרטים, שהרי אזרחי המדינה ראויים אף הם להגנה. אומנם

סטנדרט רגולטורי נמוך מהמקובל בעולם יכול לעודד חדשנות, אך זו לא בהכרח חדשנות רצויה: ישראל עלולה להפוך לחצר אחורית טכנולוגית, כלומר למקום שבו מפתחים מערכות שאסור לפתח אותן, להפיץ אותן או להשתמש בהן לפי הדין הזר.

רגולציה חסינת עתיד. תהליכי חקיקה מפגרים תמיד אחר המציאות. הטכנולוגיה מתקדמת בקצב מהיר, ובמדינות כמו ישראל, שבהן תהליכי החקיקה הם איטיים מאוד, נוצר פער גדול במיוחד. על כן אין לאסדר טכנולוגיה ספציפית, שכן זה מתכון בטוח להתיישנות הרגולציה, אלא להשתדל לקבוע עקרונות מנחים והגדרות כלליות שיקנו גמישות באכיפה עתידית.

מסגרת אסדרה היברידית: עקרונות, זכויות וחקיקה. מסגרת מבוססת עקרונות מציגה מערך עקרונות ליבה אתיים, מסגרת מבוססת זכויות מתמקדת בהגנה על זכויות האדם ועל החירויות של מי שמושפעים מיישומי הטכנולוגיות המבוססות על בינה מלאכותית, ומסגרת מבוססת חקיקה מאפשרת שלא להסתמך אך ורק על אסדרה וולונטרית, המבוססת על הרצון הטוב של השחקנים הכלכליים בשוק. שלוש המסגרות אינן סותרות ויש לשלב ביניהן במסגרת היברידית, המשלבת אסדרה רכה (עקרונות אתיים) עם תפיסת ניהול סיכונים שתתבטא בהוראות החוק ובכללים רגולטוריים נוקשים.

גמישות באשר למועד ההתערבות האסדרתית. במקרים מסוימים יש יתרונות ברורים לרגולציה מוקדמת, כלומר לפני שמוצרים שמבוססים על טכנולוגיה מסוימת חודרים לשוק. אם טכנולוגיה נתפסת מסוכנת במיוחד – פיזית (כמו רכב אוטונומי) או מוסרית (למשל יצירה מלאכותית של תוכן המסית לביצוע מעשי טרור) – הגיוני לאסדר מראש. גם במקרים קיצוניים פחות התערבות מוקדמת יכולה להועיל במה שנוגע לעיצוב כיווני המחקר ולתכנון השקעת המשאבים בפיתוח. יתר על כן, מאחר שבשלב זה הושקעו משאבים מעטים יחסית והעלויות השקועות נמוכות יותר ייתכן שהתערבות רגולטורית תיתקל בהתנגדות מצומצמת יותר מצד בעלי עניין. מנגד, ייתכנו מקרים שבהם עדיף להמתין ולהתמודד עם בעיות כאשר הן מתעוררות במקום לנסות לצפות אותן.

אסדרה ענפית ואסדרה רוחבית. אסדרה רוחבית משיגה מטרות של משילות וודאות מצד השלטון המרכזי, יוצרת הרמוניה רגולטורית ולפיכך יכולה להגביר

את אמון הציבור ואת הוודאות הרגולטורית בעבור התעשייה. מנגד, אסדרה ענפית מאפשרת שימוש ברגולטורים קיימים ובסמכויות האכיפה שלהם, איננה מחייבת להקים מסגרות מוסדיות חדשות, מאפשרת יצירת הסדרים ואמצעי אכיפה המותאמים בצורה מדויקת יותר לענף הרלוונטי, מגבירה את הבהירות והוודאות הרגולטורית ומאפשרת גם לבעלי עניין בתוך כל מגזר להשתתף בחשיבה על ההסדרים האלה. ואולם הקושי באסדרה ענפית הוא בסתירה אפשרית בין ענפים, ביצירת סטנדרטים לא אחידים, בהעמקת הפערים בין הרגולטורים וכן בהותרת מרחבים לא מאוסדרים הנופלים בין הכיסאות. אשר על כן מוצע לנקוט שילוב של אסדרה רוחבית ואסדרה ענפית, דוגמת מודל של מאסדר־על שיש לו סמכויות הנחיה וייעוץ למאסדרים המגזריים.

ארגז כלים לאסדרת בינה מלאכותית

הבנת "מעגל החיים" של מערכות לומדות צריכה לשמש בסיס לאסדרתן. כדי ליצור אסדרה אפקטיבית של מערכות לומדות יש להביא בחשבון את כל רכיבי מעגל החיים שלהן. הואיל ועקרונות האסדרה, כגון הוגנות, פרטיות, שקיפות, אחריותיות וניהול סיכונים, מתבטאים בהקשרים שונים בכל אחד מרכיבי מעגל החיים, היעדר התייחסות לרכיבים אלו עלול ליצור מצב של אסדרת יתר של רכיבים מסוימים והתעלמות מרכיבים אחרים, ולכן לצמצם את האפקטיביות של האסדרה.

לעומת זאת, תפיסה אינטגרטיבית מביאה בחשבון את מכלול רכיבי מעגל החיים ונדרשת גם לקשרי הגומלין ביניהם. למשל: תכלית המערכת ומסגור הבעיה צריכים להשפיע על בחירת המודל (אם לבחור במודל שיש בו עכירות רבה יותר באשר לדרכים שבו הוא מקבל החלטות או שלא לאפשר זאת); תוצאות הערכה בשלב בניית המודל מזינות תהליכים להערכת סיכונים, והם בתורם מחייבים קבלת החלטות הנוגעות לאימון המודל, לפרישׂתו או לאבטחתו; תכלית המודל משפיעה על הבחירה בממשקי המשתמש (האם מערכת המעניקה ייעוץ רפואי אמורה להציג את האפשרות שהיא טועה? האם נכון לספר למשתמש שהוא מתקשר עם מערכת מלאכותית ולא אנושית?).

חלק חשוב של הבנת מעגל החיים של מערכות לומדות נוגע לצורך לנטר אותן לאחר שיושמו בפועל (post deployment) בעולם האמיתי (למשל הבניה של המערכת בתוך מוצר או בתוך ממשק). זאת משום שמערכת לומדת, שלא כמו מוצרים אחרים (תרופות למשל), יכולה מעצם טיבה להשתנות גם לאחר יישומה בשל המשוב שהיא מקבלת מן המשתמשים.

פיתוח מתודולוגיות לניהול סיכונים. הצורך ליישם מתודולוגיות ניהול סיכונים על מערכות אלגוריתמיות עדיין נמצא בחיתוליו, למרות חיוניותו. אנו מציעים מודל של ניהול סיכונים שלפיו כדי להעריך מסוכנות של מערכת נדרשת התבוננות כפולה – תחילה יש להעריך את פוטנציאל המסוכנות של המערכת כפי שתוכננה; אחר כך יש להעריך את רמת הקשר בין המשימה לתוצאה (alignment), כלומר את האפשרות של המערכת לממש פוטנציאל מסוכנות מחוץ לתפקיד שייעדו לה מתכנניה.

הערכה כפולה זו תהיה הבסיס לקבלת ההחלטות וכדי שיהיה אפשר להוציאה לפועל יהיה צורך לנסח כללי משילות ובטיחות, כגון כללים לאימון אחראי (אם לאמן מודל חדש שמראה סימנים מוקדמים של סיכון – וכיצד); וכללים ליישום אחראי (אם, מתי וכיצד ליישם מודלים שעלולים להיות מסוכנים); ולקבוע אילו רמות של שקיפות ותיעוד נדרשות במקרה של מודלים שעלול להיות בהם סיכון קיצוני ואילו בקרות ומערכות אבטחת מידע יש ליישם בעניינם.

תיעוד נתונים, משילות נתונים והליכי בקרה (auditing) בדיעבד. המורכבות של מערכות בינה מלאכותית בכלל, ושל מערכות לומדות בפרט, מערימה קשיים מיוחדים על קובעי המדיניות והרגולטורים בבואם לנסח כללי אחריות ולזהות בפועל את השרשרת הסיבתית שהובילה לפגיעה בזכויות, בייחוד בהתחשב בשונות בין מגזר למגזר ובין יישום ליישום.

המשותף לכל אלו הוא שבלי תיעוד ראוי הם בלתי אפשריים. הבסיס לכל בחינה עובדתית של כשל קונקרטי במערכות בינה מלאכותית הוא משילות נתונים ותיעוד קפדני של הליכי עבודה, מקורות מידע, תיוגים, מודלים, תהליכי עיצוב קוד, הערכת סיכונים ובסיסי נתונים, וזיהוי הפערים בכל אלו. עיצוב טוב של התיעוד הוא גם אינטרס של יזמים ומפתחים, שכן הוא מאפשר להם הן לתחקר

בדיעבד כשלים ותופעות שהם לא צפו הן לעמוד בחובות תיעוד רגולטוריות ממקורות אחרים.

פיתוח כלים להתמודדות עם הטיות ו"הנדסת הוגנות". אף שאי־אפשר למנוע את בעיית ההטיות האלגוריתמיות, בייחוד כשמקור ההטיה הוא במציאות עצמה, המגולמת בנתונים, אפשר לזהות אותה ולצמצם את היקפה. מוצע לנקוט כמה אסטרטגיות לצמצום הטיות אלגוריתמיות, ובהן קיום הליך הוגן מבחינה סטטיסטית, גיוון ההון האנושי בקרב מפתחים של מערכות בינה מלאכותית ויישום הליכי בקרה בדיעבד (auditing).

התמודדות עם אתגרי השקיפות האלגוריתמית: איתנות מדעית תהליכית. מוצע לאמץ מודל המבוסס על התפיסה הקלסית של שקיפות, אבל כולל חלופה שמותאמת למגבלות הטכנולוגיות של מערכות אלגוריתמיות שאינן מסוגלות לספק הסבר רציונלי לפלט נקודתי. המודל מיועד למקרים שבהם אין אפשרות לספק פלט, אבל יש נחיצות חברתית לשמור על חובות של שקיפות כדי שלא למנוע פיתוח בטכנולוגיות מסוימות ושימוש בהן.

המלצות לגבי מוסד מאסדר לבינה מלאכותית בישראל

מוצע להקים מוסד מאסדר לבינה מלאכותית בישראל, שתפקידיו יהיו לקדם ולתאם אסדרה של מערכות נבונות בישראל, לרבות תהליך של גיבוש חקיקה רוחבית מתוך מעקב אחר התפתחויות גלובליות מתאימות. המוסד המאסדר יתווה מדיניות בעניין פיתוח מערכות נבונות בישראל, הטמעה שלהן ושימוש בהן; יספק הנחיה מקצועית לרגולטורים ענפיים כדי להבטיח הרמוניזציה של הכללים החלים על פיתוח מערכות אלו, פרישׂה שלהן ושימוש בהן; וישמש גורם מנחה שיורי, כלומר יהיה ממונה על אסדרת מערכות נבונות פרטיות בתחומים שאין בהם מאסדר מגזרי.

הרגולטור יספק הנחיה מקצועית בתחום הבינה המלאכותית גם לרשויות המדינה. בין השאר הוא ייתן חוות דעת על מסמכי מכרזים ממשלתיים של מערכות בינה מלאכותית כדי להבטיח שהמדינה מצטיידת במערכות בינה מלאכותית העולות בקנה אחד עם הסטנדרטים הישראליים והעולמיים בתחום, וכן כדי לייצר בעבור השוק הפרטי תמריצים "רכים" לעבוד בהתאם לסטנדרטים.

הרגולטור יהיה ממונה גם על ייעוץ בנושאי משפט הבינה המלאכותית, ומוצע שתינתן לו הסמכות להציג עמדה עצמאית בתחומים אלו לכנסת ולבתי המשפט. נוסף על התפקידים הנזכרים, הרגולטור המוצע יהיה מוקד ידע, הדרכה ושיתופי פעולה ויקדם את האוריינות הדיגיטלית בתחומים אלו בקרב גורמי הממשלה. הוא אף ידון עם בעלי עניין מקומיים ובינלאומיים מתחומי התעשייה, הממשל והאקדמיה על ההשפעות האפשריות של טכנולוגיות אלו על החברה.

בעת הזאת יהיה נכון למקום את רגולטור הבינה המלאכותית כיחידה בתוך רשות הרגולציה, שכן תפקידיה הולמים מאוד את העולם המתפתח של אסדרת הבינה המלאכותית – לפחות בשנים הקרובות, עד התייצבותו. ההתמודדות עם אתגרי הגמישות, עם הסביבה הרגולטורית העמוסה ועם הטכנולוגיה מצריכים הקצאת משאבים ראויה, שתאפשר ליצור תקנים הן למומחים טכניים הן למומחים במשפט ובמדיניות. מוצע כי גם בהיעדר חוק בינה מלאכותית שלם תתוקצב הרשות בהחלטת ממשלה כדי שתוכל להקים את היחידה של רגולטור הבינה המלאכותית.

המלצות לתקופת הביניים עד לאסדרה של התחום

חקיקה משלימה ועדכוני חקיקה. גם בהיעדר חוק בינה מלאכותית כללי וייעודי יש לחייב כבר בעת הזאת את מקבלי ההחלטות והרגולטורים הייעודיים לעדכן את החקיקה הקיימת ולחוקק חקיקה משלימה. מדובר בעיקר בחוקים כגון חוק התחרות וההגבלים העסקיים, חוק זכויות יוצרים, חוק הגנת הפרטיות, פקודת הראיות וחוק הרכש הממשלתי.

יצירת "אקוסיסטם קדם־אסדרתי". ראוי ליצור אקוסיסטם קדם־אסדרתי שבמסגרתו יפרסמו הרגולטור הייעודי, רגולטורים ענפיים ורגולטורים משלימים (כגון הרשות לפרטיות והרשות לתחרות) הנחיות וגילויי דעת; והם בתורם ישמשו את התעשייה ושחקנים קונקרטיים בה, וגם את בתי המשפט. מסמכי הנחיות אלו יוכלו לשמש את התעשייה כדי לקבוע סטנדרטים של אסדרה עצמית, מתוך הנחה שרגולציה עתידית תהיה דומה להנחיות. הם יוכלו לשמש גם את חברות התקינה כדי ליצור את מסגרות התקינה.

סביב האקוסיסטם הקדם־אסדרתי יתפתחו תהליכים של חידוד דפוסי הבקרה (auditing) בהקשרים שונים, ייערכו תהליכי חשיבה בעניין סטנדרטים של זהירות ראויה וייעשה מאמץ סביר לקדם אחריות בפיתוח וביישום של מערכות לומדות; ייבנו ארגזי חול וייערכו ניסויים רגולטוריים שונים; ויקודם מאגר של מומחים שיוכלו לשמש את התעשייה, את הרגולטורים ואת בתי המשפט בהתמודדות עם סוגיות חדשות ומורכבות. במקביל, התעשייה והשחקנים יוכלו לתת משוב לגופים הרגולטוריים השונים וכך לטייב את ההנחיות. נוסף על כך, בהיעדר אסדרה, בתי משפט יוכלו להשתמש בהנחיות כהשראה לפרשנות בסכסוכים המובאים לפניהם. פסקי הדין יוכלו גם הם להעשיר את גוף הידע ולאפשר לשחקנים וגם לרגולטורים לטייב ולשייף את הנחיותיהם לקראת גיבוש אסדרה.

מבוא

מערכות בינה מלאכותית נעשו בשנים האחרונות חלק בלתי נפרד מן החיים. מעוזרים אישיים דיגיטליים ועד מכוניות אוטונומיות, מרובוטים המבצעים משימות פשוטות ועד מערכות מעקב, זיהוי וחיזוי. בינה מלאכותית ואלגוריתמים מבוססי למידה נבונה משמשים בין השאר במערכות פיננסיות ממוחשבות,[1] בשירות הסוכנויות לאכיפת החוק והביטחון,[2] ברפואה,[3] וגם בזירה המשפטית. לצד כלים

1 ראו **להלן** בסעיף 1.9.

2 ראו **להלן** בסעיף 1.8.2.

3 ראו **להלן** בסעיף 1.8.2.

חכמים התומכים בעבודתם של משפטנים[4] יש דיונים תאורטיים על האפשרות שבעתיד יוכלו מערכות אלגוריתמיות "לתפור" חקיקה בהתאמה אישית.[5]

ככל שמערכות הבינה המלאכותית מתפתחות כך הן נוטות להציג יכולות חדשות שהמפתחים שלהן לא התכוונו להן או חזו אותן. יש יכולות שמעוררות השראה ויש יכולות בעלות פוטנציאל להזיק, כגון היכולת לבצע פעולות סייבר התקפיות, לתמרן אנשים באמצעות שיחה או לספק הוראות מעשיות לביצוע פעולות טרור. היכולת לזהות יכולות אלו ולהגביל את הסיכונים הצפונים בהן נעשות אפוא אתגר משמעותי.

ההוצאות של ממשלות ועסקים ברחבי העולם על מוצרים המבוססים על בינה מלאכותית יגיעו ב־2023 ל־500 מיליארד דולר.[6] עם השימושים האלה מגיעים גם חששות מסוגים שונים: נזקים, אפליה, הוגנות, הטיות, גיוון ופרטיות. על שולחן הדיונים כבר מונחת הצעת חוק של האיחוד האירופי בעניין בינה מלאכותית, הצעה של רשות הסחר ההוגן בארצות הברית, והרבה מסמכי מדיניות והצעות חקיקה במדינות אחרות. גם בישראל מתפרסמים מסמכים שונים בעניין זה.

מהי בינה מלאכותית, מהם יתרונותיה ואילו חששות היא מעוררת, בעיקר כשהיא משמשת את רשויות השלטון? שאלות אלו הן עניינו של ספר זה.

ספר זה נועד לשמש בראש ובראשונה מבוא למקבלי החלטות בכלל ולעוסקים ברגולציה של בינה מלאכותית בפרט. אבל הוא מיועד גם לכל מי שעולם הבינה המלאכותית מעניין אותו ומבקש להבין אותו ואת סביבתו. תחילה נסביר מהי בינה מלאכותית ונסקור יישומים שונים שלה. אחר כך נידרש לאתגרים הייחודיים שהבינה המלאכותית מעמידה לפני הרגולטור – הטיות אלגוריתמיות והסברתיות (explainability), ואף לחשש שמערכות בינה מלאכותית יפגעו בזכויות אדם. ולבסוף נדון בתשתית המשפטית והמוסדית הנדרשת לבינה מלאכותית בטוחה. תשתית כזאת צריכה לתת את הדעת על האדם, מושא ההחלטות האלגוריתמיות,

4 ראו **להלן** בסעיף 1.8.2.

5 ראו **להלן** בסעיף 1.8.3.

6 Worldwide Semiannual Artificial Intelligence Tracker, IDC

ועל המשטר הדמוקרטי. מאחר שהן בזירה הטכנולוגית הן בזירה הרגולטורית העולמית תחום אסדרת הבינה המלאכותית הוא דינמי ומתפתח, בחרנו שלא להציע מתווה אסדרה פרטני אלא להמליץ על קווי מתאר, עקרונות וכלים שרצוי לאמץ.

לנוכח רוחב היריעה של הנושא יש להדגיש באילו נושאים ספר זה אינו עוסק.

ספר זה אינו על בינה מלאכותית שבאה לחסל את כולנו, אלא הוא מתבונן במבט מפוכח בצורך לגבש בעשור הקרוב מדיניות טכנולוגיה. התרבות הפופולרית מרבה לתאר תרחישי אימים של מכונות המפתחות תודעה אנושית או על־אנושית ופונות כנגד מפתחיהן האנושיים.[7] הרעיון של יצורים מלאכותיים, מעשי ידי אדם, בעלי תבונה ואוטונומיה, גירה את הדמיון האנושי לאורך ההיסטוריה. האגדה על הגולם מפראג, שנתנה השראה לשמם של המחשבים הראשונים במכון ויצמן,[8] מתארת יצור דמוי אדם עשוי חימר או אדמה שעל מצחו כתוב "אמת" והוא ניעור לחיים באמצעות קוד של צירופי אותיות מיסטיים. המפלצת שברא ויקטור פרנקנשטיין, ששבה לרדוף אותו, היא עוד דוגמה לנרטיב הגולם – שבמקרה זה אכן קם על יוצרו. ואילו בתרבות הפופולרית בת זמננו התגלגל המיתוס בתרחיש האימים של תבונה ממוחשבת מעשה ידי אדם שמשתלטת על הפלנטה ומשעבדת את המין האנושי.[9]

התסריטים האפוקליפטיים האלה נוטים להדהד יתר על המידה באזהרותיהם של מדענים[10] ואנשי טכנולוגיה[11] בני זמננו. בסקר שנערך בשנת 2022 בקרב

7 לדוגמאות ראו **להלן** ה"ש 9.

8 "מחשב אלקטרוני 'גלם' הופעל במכון וייצמן" **דבר** (30.12.1963).

9 ארתור ס' קלארק "חייג פ' ותקבל את פרנקנשטיין" **הרוח הנושבת מן השמש** 90 (מתרגם יורם רפפורט, 1978); הרלן אליסון "אין לי פה ואני מוכרח לזעוק" **מדע בדיוני: הטוב שבטוב** כרך 4, 34 (עורך איזק אסימוב, מתרגמת אילנה דגני בינג, 1981); ראו גם הסרטים **אודיסיאה בחלל: 2001** (סטנלי קובריק במאי ומפיק 1968); **שליחות קטלנית 2: יום הדין** (ג'יימס קמרון במאי ומפיק 1991); **מטריקס** (במאיות לילי ולאנה ואצ'ובסקי 1999).

10 סטיבן הוקינג, למשל, הזהיר מפני פיתוח לא מבוקר של אינטליגנציה מלאכותית. ראו רפאלה גויכמן "סטיבן הוקינג מזהיר בפני הסכנות האמיתיות של המכונות" **ynet** (15.10.2015).

11 Peter Holley, "Elon Musk's Nightmarish Warning: AI Could Become 'an Immortal Dictator from which We Would Never Escape,' " The Washington Post (6.4.2018)

חוקרי בינה מלאכותית, 36% מהנשאלים השיבו כי יש סבירות שמערכות בינה מלאכותית יוכלו "לגרום במאה זו לקטסטרופה גרועה לפחות כמו מלחמה גרעינית כוללת".[12] ואכן, לאחרונה חתמה שורה של מומחים טכנולוגיים על הצהרה המתריעה על הצורך למתן את הסיכון של הכחדת המין האנושי בשל בינה מלאכותית,[13] וקדמה לה עצומה של מומחים לבינה מלאכותית ומובילי תעשייה שקראה להשהות לשישה חודשים פיתוח של מערכות בינה מלאכותית מתקדמות מחשש שאלו מסכנות את המין האנושי ועלולות להביא ל"אובדן השליטה על הציוויליזציה שלנו".[14]

אכן, קיימים אינטרסים שונים המובילים לאזהרות אלה. החל בניסיון להעביר אחריות לרגולטור על החלטות עסקיות מסוכנות, עובר ברצון לקנות זמן כדי להצליח להתמודד בשוק תחרותי או רצון להשאיר כוח בידי פלטפורמות ולא בידי יחידים, וכלה בחשש כן מפני השלכות חברתיות מסוכנות של שימושים שונים בטכנולוגיה.

ואולם הדימוי הפופולרי של בינה מלאכותית כאיום קיומי על המין האנושי עלול לעוות את הדיון בהשפעת מערכות הבינה המלאכותית בטווח הקרוב והבינוני – בסיכויים הטמונים בה ובסכנות הנשקפות ממנה. יש הטוענים כי הצפת השיח הציבורי בקריאות לרגולציה (לרבות הקמת סוכנות בינלאומית לפיקוח על בינה מלאכותית, לפי הדגם של סבא"א)[15] שתגן עלינו מפני אפוקליפסת בינה מלאכותית נועדה להסיח את הדעת מהבעיות המיידיות שנוגעות לשימוש, פיתוח

12 Julian Michael et al., *What Do NLP Researchers Believe? Results of the NLP Community Metasurvey* (26.8.2022), available at arXiv

13 אושרי אלקסלסי "סאם אלטמן ובכירים בענף ה־AI מזהירים מפני 'הכחדה' של המין האנושי" **גיקטיים** (30.5.2023); "*Statement on AI Risk: AI Experts and Public Figures Express their Concern about AI risk*," Center for AI Safety (30.5.2023)

14 *Pause Giant AI Experiments: An Open Letter*, Future of Life Institute (22.3.2023)

15 יובל מן "OpenAI מציעה להקים סוכנות בינלאומית לפיקוח על בינה מלאכותית" **ynet** (23.5.2023).

ופרישה של מערכות אלו,[16] שהרי כבר עתה מערכות נבונות הן חלק ממארג חיי היום־יום של כולנו.

ספר זה אינו על בינה מלאכותית כאישיות משפטית. בשנת 2022 טען בלייק למוין (Lemoine), מהנדס תוכנה בחברת גוגל, שמודל השפה המתקדם של החברה פיתח מודעות עצמית.[17] חברת גוגל הכחישה זאת,[18] אבל השאלה אם מערכות מחשב יוכלו לפתח תודעה עצמאית שתרומם אותן מדרגת חפץ דומם לדרגה של יצור תבוני בשר ודם (או מעבר לו) מוסיפה לגרות את הדמיון האנושי.

מבחינה משפטית שאלה זו מעלה סוגיית משנה: האם מערכות מחשב תבוניות ובעלות תודעה ראויות להיחשב לאישיות משפטית?[19] אנלי ניוויץ טוענת שייחוס אישיות משפטית למערכות בינה מלאכותית היא מסך שמגן על מי שייתכן שהם האחראים האמיתיים לפגמים בהתנהלותן של מערכות אלו; על פי עמדתה של ניוויץ, אנשי הפיתוח והמהנדסים הם שלא השכילו לנפות מראש הטיות אלגוריתמיות.[20] שאלת האישיות המשפטית מתעוררת לעיתים בהקשרים של זכויות יוצרים – האם ניתן להקנות למערכת בינה מלאכותית זכות בקניין רוחני

16 ראו למשל Matteo Wong, *AI Doomerism is a Decoy*, The Atlantic (2.6.2023)

17 "מהנדס תוכנה של גוגל טען שהבינה המלאכותית של החברה פיתחה תודעה – והושעה" **הארץ** (13.6.2022).

18 ראו **להלן** ה"ש 37. מחקרים מאוחרים יותר גורסים שאומנם למודלים מתקדמים של שפה יש מאפיינים של מה שנקרא בספרות הפסיכולוגית "תאוריה של תודעה" (Theory of Mind), כלומר היכולת להבין שלזולת יש מחשבות ואמונות השונות משלך, אך מדובר ביכולת מוגבלת שנשענת על הסתמכות יתר על היריסטיקות ולא על שיקול דעת כללי וחסין. ראו Natalie Shapira et al., *Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models* (24.5.2023), available at arXiv

19 ראו לדוגמה הדיון אצל Visa Kurki, A Theory of Legal Personhood 175–189 (2019); Simon Chesterman, We, the Robots? Regulating Artificial Intelligence and the Limits of the Law 114–142 (2021)

20 Annalee Newitz, *The Curious Case of the AI and the Lawyer*, 255 New Scientist 28 (23.7.2022)

שהיא חוללה?[21] בית המשפט הפדרלי לערעורים בארצות הברית ענה לאחרונה בשלילה על שאלה זו,[22] אך הסוגיה עודנה מוסיפה להעסיק ערכאות אחרות.[23]

יוער כי יש המתנגדים לשימוש במונח בינה מלאכותית בטענה שהוא מעורר קונוטציות למכונות המפתחות תודעה ורגשות אנושיים ומעודד השוואות שגויות. לדבריהם, לא מדובר אלא באוסף של טכניקות אלגוריתמיות סטטיסטיות ולמידת מכונה.[24] כך או כך, כפי שנסביר בהמשך החיבור, נדרשת המשגה חדשה שתוכל להרחיק את הדיון מהשוואות שגויות אלו. למשל, החלפת המונח "בינה מלאכותית" ב"למידה חישובית".

ספר זה אינו על בינה מלאכותית במשפט הפרטי, אלא על סוגיות מעולם המשפט הציבורי והמעין ציבורי. מקצת הסוגיות נוגעות לתהליכי הפיתוח של מערכות בינה מלאכותית, הניזונות ממידע רב, ולשאלה מי בעל זכויות היוצרים על המידע,[25] או מה מידת הפגיעה בפרטיות הכרוכה בשימוש בו, ומקצתן יכולות להשיק לשאלת האישיות המשפטית ונוגעות לשאלות של אחריות במקום שבו נגרם נזק בשל פעילות של מערכת בינה מלאכותית.[26] למשל, מי אשם במקרה של

21 לדיון ראו Aviv H. Gaon, The Future of Copyright in the Age of Artificial Intelligence (2021). ראו גם Ryan Abbott and Elizabeth Shubov, *The Revolution Has Arrived: AI Authorship and Copyright Law*, Florida Law Review, forthcoming (8.8.2022); Anke Moerland, *AI and Intellectual Property Law*, *in* The Cambridge Handbook of Private Law and Artificial Intelligence (Philip Morgan and Ernest Lim eds., forthcoming 2023)

22 Thaler v. Perlmutter, No. 1:20-cv-00903-LMB-TCB (E.D. Va. Aug. 5, 2022)

23 Thaler v. Comptroller-General of Patents, Designs and Trade Marks, UKSC [2022] (Decision granting permission to appeal, 12.8.2022)

24 Luc Julia, There Is No Such Thing as Artificial Intelligence (2019); Stefano Quintarelli, *Let's Forget the Term AI. Let's Call Them Systematic Approaches to Learning Algorithms and Machine Inferences (SALAMI)*, Quinta's Weblog (24.11.2019)

25 לדוגמה ראו משרד המשפטים, **שימושים בתכנים מוגנים בזכויות יוצרים לצורך למידת מכונה** (18.12.2022).

26 לדוגמה מן העת האחרונה ראו את הצעת דירקטיבת האחריות האזרחית בעניין בינה מלאכותית. הדירקטיבה מציעה כללים של גילוי ראיות בעניין מערכות בינה מלאכותית בסיכון גבוה (ראו **להלן** בסעיף 4.1.1.2) במסגרת תובענות אזרחיות שאינן על בסיס חוזי, וכללי נטל ההוכחה בעניין תובענות אזרחיות שאינן על בסיס חוזי בשל עילות נזיקיות

תאונה של רכב אוטונומי וכיצד קובעים כי התקיימה רשלנות?[27] מי אחראי אם יש נפילה בבורסה בשל שגיאה במערכת האלגו־טריידינג, הסוחרת במניות במהירות הבזק, או בשל מניפולציות בתמחור מניות, הדורשת יסוד נפשי של כוונה?[28]

ספר זה אינו על הסכנה הנשקפת לתעסוקה מבינה מלאכותית. לצד היתרונות הכרוכים באוטומציה של תהליכים ואיכות החיים הנלווים לה מתעורר גם חשש שהקִדמה תצמצם את הביקוש לידיים עובדות, תייתר מקצועות והתמחויות מסוימים ותגדיל את האבטלה. בסוף הרבעון הראשון של 2023 בחן בית ההשקעות גולדמן סאקס 900 מקצועות ב־22 תעשיות והעריך כי רבע ממשימות העבודה הנוכחיות בתעשיות אלו יכולות להתבצע באופן אוטומטי על ידי בינה מלאכותית.[29] הדוח מעריך כי בשל מבנה שוק העבודה הישראלי, בישראל מדובר בכמעט 30% משוק העבודה.

כיום ברור שלא כמו בראשית המהפכה התעשייתית ובתגובת הנגד הלודיטית, המקצועות שנשקפת להם סכנה רבה יותר אינם מקצועות שכרוכים בעבודת כפיים ("צווארון כחול"), אלא מקצועות כגון כתיבה ורעיונאות (קופירייטינג), רפואה ואבחון, ניתוח מידע ונתונים, עיצוב ואומנות, ראיית חשבון ומשפטים, וגם תכנות.[30] יש להודות ביושר כי בחלק מן המשימות הנדרשות במקצועות אלו יוכלו מכונות לעלות בביצועיהן על בני אנוש. יתר על כן, ייתכן שהיעלמות מקצועות מסוימים אינה נובעת רק משימוש במערכות אלגוריתמיות, אלא גם

שמקורן במערכות בינה מלאכותית. European Commission, Proposal For a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to artificial intelligence (AI Liability Directive) COM/2022/496 (28.9.2022) (להלן: הצעה לדירקטיבה אירופית לאחריות בתחום הבינה המלאכותית).

27 ראו לדוגמה Sven Nyholm and Jilles Smids, *The Ethics of Accident-Algorithms For Self-Driving Cars: An Applied Trolley Problem*, 19 Ethical Theory and Moral Practice 1275 (2016); Karni Chagal-Feferkorn, *The Reasonable Algorithm*, 2018 U. Ill. J.L. Tech. & Pol'y 111 (2018)

28 Gina-Gail S. Fletcher, *Deterring Algorithmic Manipulation,* 74 Vand. L. Rev. 259 (2021)

29 Jan Hatzius et al., *The Potentially Large Effects of Artificial Intelligence on Economic Growth,* Goldman Sachs Economic Research (26.3.2023)

30 Michael Webb, *The Impact of Artificial Intelligence on the Labor Market* (6.11.2019)

ממודלים כלכליים חדשים – דוגמת כלכלת החלטורות (the gig economy) וכלכלת השיתוף (the sharing economy) – שנעזרים בפלטפורמות מקוונות כדי לערער את המערכות המסחריות המסורתיות. כרגע נסתפק בהערכה שגם מקצועות שצופים כי ייעלמו – דוגמאות נפוצות הן הרדיולוגיה, שנעשית כיום באמצעות מערכות אוטומטיות,[31] או עיצוב, שנעשה באמצעות תוכנות גנרטיביות – גם מקצועות אלו יזדקקו למומחה אנושי שיחלוק את העבודה עם מכונה כשם שמטוסים בעלי טייס אוטומטי אינם ממריאים בלי קברניט בשר ודם.[32] לפיכך פיתוח היכולת לעבוד עם מכונות חכמות ולצידן ילך ויהפוך למיומנות עבודה חיונית ויתבטא במשימות שמכונות היום "לחישה" (שִׁכלול הנוסחה שתוזן למכונה כדי לקבל תוצר מיטבי – קוד, טקסט או דימוי חזותי); בהבנת תהליכי היסק והנמקה של מכונות כדי להשלים תהליכי חשיבה אנושיים;[33] ובפיקוח ברמת שונות (קביעת דרגת אוטונומיה של רכב, חוות דעת על אבחון רפואי ועוד).

בינה מלאכותית, לרבות האסדרה שלה, היא תחום דינמי שמשתנה תדיר. ספר זה נועד להיות טקסט מבואי ולהציג למי שאינם בעלי רקע טכנולוגי הן את התחום עצמו הן את אתגרי המדיניות והרגולציה שהוא מעורר. מחקר זה עדכני ליוני 2023 ואינו כולל שינויים או התפתחויות מאוחרים יותר.

אנו מבקשים להודות לפרופ׳ סוזי נבות, לפרופ׳ יובל שני, לד"ר רחל ארידור הרשקוביץ ולעו"ד גדי פרל על הערותיהם הטובות לאורך הדרך. לד"ר רועי צזנה ולשירה פטרק על עזרה מחקרית בשלבים הראשונים של המחקר. תודה לאנשי ההוצאה לאור של המכון הישראלי לדמוקרטיה.

31 על רדיולוגיה ראו Ramón Alvarado, *Should We Replace Radiologists with Deep Learning? Pigeons, Error and Trust in Medical AI*, 36 Bioethics 121 (2022)

32 על הסתגלות של מומחים בתחומי האבחון הרפואי ראו Saurabh Jha and Eric J. Topol, *Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists*, 316 JAMA 2353 (2016); לדיון בנושא טכנולוגיה ופרופסיות ראו Daniel Susskind and Richard Susskind, The Future of the Professions: how Technology Will Transform the Work of Human Experts (2015)

33 להרחבת רעיון זה ראו למשל איתי ברון ותהילה שוורץ אלטשולר "מקום ליד השולחן – האם ה־ChatGPT יכול 'לשבור את הקונספציה'?" **המכון לחקר המתודולוגיה של המודיעין** (1.5.2023).

פרק ראשון

מהי בינה מלאכותית?

הרעיון של ישויות מלאכותיות, אוטומטונים, שאפשר להנדס או להרכיב, נדון ברצינות לאורך השנים.[34] בשנות הארבעים של המאה העשרים טבע נורברט וינר את המונח קיברנטיקה כדי לתאר מערכות מבוססות משוב (feedback) והצביע על הדמיון בין פעולתם של מחשבים ובין תהליכים של זיהוי דפוסים במוח האנושי.[35] אלן טיורינג, בן תקופתו של וינר, שעסק בתאוריות מתמטיות ובמדעי

34 ראו למשל תומאס הובס **לויתן** 5 (מתרגם אהרן אמיר 2010).

35 Norbert Wiener, Cybernetics: or Control and Communication in the Animal and the Machine (1948)

המחשב, מוכר כמי שהציע מבחן היפותטי שנועד לתת מענה אפשרי לשאלה אם מכונה יכולה לחשוב כמו בן אדם.[36] תחת שאלה טעונה זו הציע טיורינג משחק שקרא לו משחק החיקוי ולימים נודע בשם מבחן טיורינג. למשחק נדרשים שני בני אדם ומכונה. אדם אחד, המשמש כחוקר, מנהל במקביל דיאלוג בכתב עם האדם השני ועם המכונה, בלי לדעת מי הוא מי. אם החוקר אינו מצליח להבחין בין האדם ובין המכונה אפשר לומר שהמכונה עברה את מבחן טיורינג.

הצירוף בינה מלאכותית נטבע בשנת 1956, בעקבות ההצעה לסדנת המחקר הראשונה בתחום, שנערכה באותה שנה באוניברסיטת דארטמות'. הסדנה התבססה על ההשערה כי "כל היבט של לימוד או כל מאפיין אחר של תבונה יכולים, עקרונית, להיות מתוארים באופן מדויק די הצורך שמכונה תוכל לחקות אותם".[37] חרף העניין המחקרי בתחום, לקראת אמצע שנות השבעים דעך העניין בבינה מלאכותית לנוכח האכזבה מההתקדמות בפועל.[38]

התקופה המכונה "החורף של הבינה המלאכותית" הסתיימה בראשית שנות השמונים, בעקבות פעפוע רעיונות של למידה על ידי חיזוקים (reinforcement learning) לענף מתחום הפסיכולוגיה והתפתחויות בתחום הרובוטיקה.[39] לאחר מחזור נוסף של "קיץ" ו"חורף" בשנות השמונים – תנופת השקעות בבינה מלאכותית ואחריה תקופה של אכזבה ומשיכת השקעות מהתחום – פותחו בשנות התשעים מערכות בינה מלאכותית שידעו להתמודד עם בעיות מתמטיות מורכבות (מסוג NP־שלמות). נקודת ציון נוספת היא התבוסה של השחמטאי

36 Alan Turing, *Computing Machines and Intelligence* 59 Mind 433 (1950)

37 J. McCarthy et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* (31.8.1955)

38 ראו למשל את סקירתו של המדען הבריטי ג'יימס לייטהיל, שלפיה לא קיים המחקר בתחום את ההבטחות להשפעה רחבת היקף ולכן קטן המימון בבריטניה, ובהמשך באירופה ובארצות הברית. James Lighthill, *Artificial Intelligence: A General Survey*, *in* Artificial Intelligence: A Paper Symposium (Science Research Council 1973)

39 Sami Haddadin and Dennis Knobbe, *Robotics and Artificial Intelligence*, *in* Algorithms and Law 1, 9–12 (Martin Ebers and Susana Navas eds., 2020)

הרב־אומן גארי קספרוב לכחול עמוק, מערכת בינה מלאכותית שפיתחה איי־בי־אם. ההצלחה של כחול עמוק נשענה על שני רכיבים: הראשון, אלגוריתם חיפוש הֵיריסטי (כלומר התבססות על כללי אצבע כדי לזרז את תוצאות החיפוש); השני, כוח מחשוב רב־עוצמה. רכיב זה הוביל לביקורת שכחול עמוק אינו מבוסס על בינה מלאכותית של ממש, אלא מסתמך על כוח חישובי גס כדי להביס רב־אמן אנושי.[40]

אף שכבר באמצע שנות השישים הציגו ולנטין לאפה ואלכסיי יאבקנקו רעיונות בדבר רשתות עצביות עמוקות (deep neural network),[41] נדרשו עוד חמישים שנה עד שהתחום הבשיל. ב־2012 זכה אלגוריתם הלמידה העמוקה של ג׳פרי הינטון בפער של ממש בתחרות זיהוי תמונות והגיע לרמת דיוק של 85% (רמת הדיוק האנושית היא 95%).[42] אפליקציית חלום עמוק של גוגל, שהושקה ב־2015, תרמה גם היא למודעות הציבור לפוטנציאל הטמון בלמידה עמוקה.[43]

מערכת הלמידה העצמית של מערכת אלפא־גו (AlphaGo) סייעה לה להביס את אלוף העולם במשחק גו (משחק אסטרטגיה מופשט שמקורו בסין). בשנת 2017 הוענקה לה דרגת דאן 9 של כבוד.[44] המערכת הוזנה בעשרות אלפי משחקי גו ושכללה את כישוריה במשחק נגד עצמה. אלפא־גו הציגה אסטרטגיות וסגנונות משחק יצירתיים, ששחקנים אנושיים לא חשבו עליהן בעבר. הגרסה הנוכחית של אלפא־גו, אלפא־גו זירו, לא נדרשה כלל למאגר של משחקים היסטוריים כדי לאמן את עצמה. כללי המשחק הוזנו לתוכה ואלפא־גו זירו שיחקה נגד עצמה — בתוך שמונה שעות בלבד הגיעה לרמה שאפשרה לה להביס את אלפא־גו. אלפא־גו זירו התבססה על מנוע ששמו אלפא זירו, שאינו מוגבל רק למשחק הגו

40 Michael Wooldridge, The Road to Conscious Machines: The Story of AI (2020)

41 Valentin Grigoryevich Lapa and Alexey Grigoryevich Ivakhnenko, Cybernetic Predicting Devices (1965)

42 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks* 25 Advances in Neural Information Processing Systems 1106 (2012)

43 Christian Zegedy et al., *Going Deeper with Convolutions*, Computing Research Repository (2014)

44 *Google's AlphaGo Gets "Divine" Go Ranking*, The Straits Times (15.3.2016)

האסיאתי. אלפא־גו זירו למד שחמט בתוך שעות ספורות וניצח את אחת התוכנות המתקדמות בעולם השחמט באותה עת – כל זאת בלי שהוזן לתוכו ולו משחק אחד של שחמט.[45] דוגמאות אלו מראות שבינה מלאכותית מסוגלת להגיע לרמת מיומנות על־אנושית בתחומים המצריכים שנים ארוכות של למידה ואימון. יתר על כן, ככל שנוקפות השנים מהירות הפיתוח של הטכנולוגיות המאפשרות זאת הולכת וגדלה.

1.1 הגדרות של בינה מלאכותית

דמיס חאסאביס, מייסד DeepMind, החברה אשר פיתחה את אלפא־גו, הגדיר בינה (intelligence) "תהליך שהופך מידע לא סדור לידע שימושי".[46] בינה מלאכותית אינה חייבת להיות בעלת נוכחות פיזית במרחב. אין לה צורך ברובוטים או בחומרה ייחודית אחרת (ובלבד שכוח העיבוד העומד לרשותה מאפשר לה להתקיים). אומנם אפשר לדמיין, כפי שעשו סופרי המדע הבדיוני לדורותיהם,[47] בינה מלאכותית שיש לה נוכחות רובוטית גשמית (בדומה לבינה אנושית). עם זאת, מערכות בינה מלאכותית בנות זמננו, דוגמת מערכת הניווט וייז או מערכת התרגום של גוגל, הן מערכות אלגוריתמיות שאינן מבוססות רובוטיקה.

ג'ון מקארתי, הנמנה עם האבות המייסדים של דיסציפלינת הבינה המלאכותית, גרס כי המטרה היא לפתח מכונות שמתנהגות **כאילו** הן בעלות תבונה.[48] מבחן טיורינג, וניסויי מחשבה אחרים (דוגמת החדר הסיני, שהציע ג'ון סרל כדי למתוח ביקורת על מבחן טיורינג),[49] מרמזים שבעיית התודעה של הבינה המלאכותית

45 "אינטליגנציה מלאכותית לימדה עצמה שחמט, והפכה לאחת השחקניות הטובות בעולם" **הארץ** (10.12.2017).

46 Demis Hassabis: *Creativity and AI – The Rothschild Foundation Lecture*, **YouTube** (10.10. 2018)

47 קארל צ'אפק "R.U.R" **דחק** 7 424 (תרגום אברהם שלונסקי 2016); אייזיק אסימוב **אנוכי הרובוט** (תרגום יאיר שמעוני 1976).

48 McCarthy et al., **לעיל** ה"ש 37.

49 John. R. Searle, *Minds, Brains, and Programs,* 3 Behavioral and Brain Sciences 417 (1980)

היא בעיה פילוסופית בעיקרה ושההגדרה המעשית של בינה מלאכותית צריכה להידרש לדרכי הפעולה שלה. יש הגורסים שהשימוש במונח בינה מלאכותית הוא מתעתע. לדבריהם, תווית זו משמשת לתיאור יכולות טכנולוגיות שאין בהן בינה שדומה לבינה האנושית.[50]

מועצת אירופה הציעה להגדיר בינה מלאכותית "מערכות המדגימות התנהגות תבונית על ידי ניתוח סביבתן ונקיטת פעולות – ברמה מסוימת של אוטונומיה – כדי להשיג מטרות מוגדרות".[51] קבוצת המומחים שכינסה מועצת אירופה (AI HLEG) הדגישה בראש ובראשונה שמערכות אלו הן רציונליות, והוסיפה להגדרה עוד היבטים, בהם היכולת לפרש מידע ולהסתגל לשינויים בסביבה.[52]

בנוסח המקורי של הצעת תקנות הבינה המלאכותית האירופיות[53] הוצע להגדיר בינה מלאכותית "תוכנה שפותחה באמצעות אחת או יותר מהשיטות והגישות המפורטות בתוספת הראשונה [למידת מכונה; גישות מבוססות ידע (מערכות מומחה); או שיטות סטטיסטיות] ויכולה, בהינתן מטרות שהגדירו בני אדם, לייצר פלטים כגון תוכן, תחזיות, המלצות או החלטות המשפיעות על סביבה אמיתית או מדומה". הגדרה דומה מצויה בהצעת החוק הברזילאית לבינה מלאכותית,[54] ואולם זו מדגישה את היסוד האינטראקטיבי של בינה מלאכותית, שכן היא החריגה במפורש מתחולת החוק מערכות קבלת החלטות המבוססות על פרמטרים

50 ראו לדוגמה Michael I. Jordan, *Artificial Intelligence – The Revolution Hasn't Happened Yet*, 1 Harv. Data Sci. Rev. (2019)

51 ראו European Commission, A European Approach to Artificial Intelligence (להלן: AI4EU).

52 AI HLEG, A Definition of AI: Main Capabilities and Scientific Disciplines (8.3.2021)

53 Proposal For a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 (21.4.2021) (להלן: הצעת תקנות הבינה המלאכותית האירופיות), בסעיף 3(1) ובתוספת הראשונה.

54 Projeto De Lei N° 21-A de 2020 (29.9.2021). לתרגום אנגלי של הצעת החוק ראו Walter Gaspar, *Non-Official Translation of the Brazilian Artificial Intelligence Bill, N. 21/2020*, CyberBRICS (25.10.2021) (להלן: הצעת החוק הברזילאית לבינה מלאכותית).

שנקבעו מראש ואין ביכולתן ללמוד, לתפוס, לפרש ולהגיב לסביבה החיצונית על בסיס קלט. הצעת החוק הקנדית לבינה מלאכותית ולמידע[55] מדגישה את ההיבט האוטונומי של מערכות בינה מלאכותית, המוגדרות מערכות טכנולוגיות שמעבדות מידע שנוגע לפעילות אנושית באופן אוטונומי או אוטונומי למחצה באמצעות אלגוריתם גנטי,[56] למידת מכונה או טכניקה אחרת במטרה ליצור תוכן, לקבל החלטות, לנסח המלצות או לייצר תחזיות.

ההגדרה המתוקנת בנוסח הפשרה של הצעת תקנות הבינה המלאכותית האירופיות[57] השמיטה את ההתייחסות לאופן הפיתוח של המערכת, שהייתה בתקנות בהגדרה שבהצעה המקורית. על פי ההגדרה המתוקנת, מערכת בינה מלאכותית היא "מערכת מבוססת מכונה שעוצבה כדי לפעול בדרגות אוטונומיה שונות ויכולה, למטרות מפורשות או לא מפורשות, לייצר פלט כגון תחזיות, המלצות או החלטות המשפיעות על הסביבה הפיזית או הווירטואלית".[58]

מסמך המדיניות (white paper) של משרד הדיגיטל, המדיה, התרבות והספורט של בריטניה מצביע על שני מאפיינים עיקריים של בינה מלאכותית הרלוונטיים

55 Bill C-27, An Act to Enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to Make Consequential and Related Amendments to Other Acts, 1st Sess, 44th Parl, 2022 (First Reading, 16 June 2022), p. 85-100 (להלן: הצעת חוק הבינה המלאכותית והמידע של קנדה).

56 אלגוריתמים גנטיים הם אלגוריתמים בהשראת רעיון הברירה הטבעית, המשמשים בעיקר להתמודדות עם בעיות של חיפוש ואופטימיזציה. להרחבה ראו Seyedali Mirjalili, Evolutionary Algorithms and Neural Networks 43-55 (2019)

57 בחודש מאי 2023 התקבל בוועדת השוק הפנימי ובוועדת חירויות האזרח של הפרלמנט האירופי נוסח מתוקן של הצעת תקנות הבינה המלאכותית האירופיות. Committee on the Internal Market and Consumer Protection and the Committee on Civil Liberties, Justice and Home Affairs, Draft Compromise Amendments on the draft report proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts Version 1.1 (16.5.2023) (להלן: ההצעה המתוקנת).

58 ס' 3(1) להצעה המתוקנת, **שם**.

לרגולטור: אדפטיביות ואוטונומיות. אדפטיביות היא היכולת לפעול על בסיס הוראות ודפוסים שלא הוכתבו במפורש על ידי גורם אנושי אלא על ידי "אימון" או למידת מכונה (ראו להלן בסעיף 1.3). מערכות בינה מלאכותית הן גם אוטונומיות – היכולת של בינה מלאכותית לבחור אסטרטגיות פעולה, לנתח נתונים ולהגיב על פיהם היא מה שהופך אותה ל"אינטליגנטית".[59]

בסיכומו של דבר, המאפיינים המשותפים להגדרות המשפטיות לבינה מלאכותית שהובאו לעיל הם היכולת של מערכות אלו **לפעול באופן אוטונומי ולהגיב לשינויים בסביבתן**, בהתבסס על טכנולוגיות שונות, שלעיתים מצוינות במפורש בהגדרה. האוטונומיה והאדפטיביות הן גם שמאפשרות למערכות אלגוריתמיות ליצור את הרושם שהתנהגותן היא אנושית.

1.2 דרגות של בינה מלאכותית

מקובל לחלק בינה מלאכותית לשלוש דרגות. הדרגה הנמוכה ביותר נקראת בינה מלאכותית צרה או חלשה. בינה מלאכותית צרה היא מערכת שיכולותיה משתוות ליכולת האנושית בתחום מוגדר אחד, או בכמה תחומים קשורים, או עולות עליה. כל מערכות הבינה המלאכותית בנות זמננו משתייכות לקטגוריה זו: מערכת בחירת השירים של אפליקציית ספוטיפיי יודעת לקלוע לטעמו של המשתמש כמו די־ג'יי אנושי (יש שחולקים על קביעה זו), אך אינה מסוגלת לנווט במרחב כמו האלוגריתם של אפליקציית וייז, או לשחק שחמט ברמת המומחיות של אלפא־זירו, ואילו הן אינן יכולות כמובן לבחור שירים כמו ספוטיפיי.

הדרגה הבאה נקראת בינה מלאכותית כללית או חזקה. סרל נדרש לאפשרות של בינה מלאכותית חזקה כשאמר כי "למחשב מתוכנת כיאות, שיש לו הקלטים והפלטים הנכונים, תהיה תודעה [mind] בדיוק באותו מובן שבו יש לבני אדם

59 Department For Digital, Culture, Media, and Sport, Establishing a pro-innovation approach to regulating AI: An overview of the UK's emerging approach 9 (18.7.2022)

תודעה".[60] אומנם מרבית המפתחים של בינה מלאכותית לא בהכרח חותרים ליצור מערכת תבונית שעונה על הדרישות התאורטיות הנוקשות של סרל, אבל פיתוח של בינה מלאכותית כללית **חותר לעצב מכונות המסוגלות להתמודד עם בעיות רבות, מתחומים שונים, שאינן מוגדרות במדויק.**

בינה מלאכותית כללית (AGI) היא ייצוג של יכולות קוגניטיביות אנושיות כלליות בתוכנה. למערכת AGI היכולת למצוא פתרון למשימה לא מוכרת. מערכת כזאת חותרת אפוא לבצע כל משימה שאדם מסוגל לה ברמה של ביצוע אנושי. ההגדרות של AGI משתנות כיוון שמומחים מתחומים שונים מגדירים אינטליגנציה אנושית מנקודות מבט שונות. אבל מדעני מחשב מגדירים את האינטליגנציה האנושית במונחים של **יכולת להשיג מטרות**, ומקצתם סבורים כי כשתושג בינה מלאכותית כללית יעלו יכולותיה על היכולות האנושיות כיוון שבכוחה לגשת למערכי נתונים עצומים ולעבד אותם במהירות. הדברים אמורים בחשיבה מופשטת, גישה למידע רקע, היגיון פשוט, הבנת סיבה ותוצאה ויכולת ליישם את מה שנלמד במשימה אחת במשימות אחרות וחדשות.

דוגמאות מעשיות ליכולות האלה הן יצירתיות (היכולת לכתוב טקסט, לייצר דימוי חזותי או אורקולי או לכתוב קוד); תפיסה חושית וחזותית (למשל היכולת לדמיין תמונה תלת־ממדית על יסוד חשיפה לתמונה דו־ממדית); יכולות מוטוריות מפותחות (למשל היכולת להוציא צרור מפתחות מתוך כיס, פעולה שייתכן שהיא דורשת גם כושר דמיון); הבנת שפה טבעית (בעיקר את ההיבט ההקשרי של השפה); ויכולות ניווט מפותחות בתוך חללים.

עינינו הרואות שבתקופה האחרונה הולכות ומושגות מקצת היכולות האלה. מגוון מוצרי קצה מבוססים על יכולת למידה עצמית ועל אלגוריתמים של למידה, על בניית מבנים קבועים עבור משימות, על הבנת מערכות סמלים, על שימוש בסוגים שונים של ידע ועל יכולות מטא־קוגניטיביות. לכן מקובל להניח היום כי אין לשרטט הבחנה דיכוטומית בין בינה מלאכותית צרה לבינה מלאכותית כללית אלא יש לדבר על ספקטרום של יכולות. ספקטרום זה מכונה לעיתים Artificial More-general Intelligence (AMGI).

60 Searle, **לעיל** ה"ש 49.

הדרגה האחרונה היא בינה מלאכותית עילאית (superintelligence). ב־1965 חזה המתמטיקאי האנגלי ג׳ק גוד את אפשרות קיומן של מכונות שיתעלו מעל הפעילות האינטלקטואלית של החכם באדם. מאחר שמכונות כאלו יוכלו להרכיב מכונות חכמות מהן, יוביל הדבר ״בוודאי ל׳פיצוץ׳ של אינטליגנציה״.[61] לפי חזון זה, השילוב של יכולות אלו עם הצמיחה המעריכית של יכולות העיבוד של מחשבים יובילו לסינגולריות – עידן עתידי שבו קצב השינויים הטכנולוגיים יהיה מהיר כל כך, והשפעתו עמוקה כל כך, עד שהמין האנושי ישתנה לבלי הכר.[62]

ואולם ספר זה אינו עוסק בעתידנות אלא במלאכת ההווה. כאמור, מרבית מערכות הבינה המלאכותית הקיימות כיום הן מערכות של בינה מלאכותית צרה ובינה מלאכותית ״רחבה״ או ״כללית״. כפי שהולך ומתברר מסוף שנת 2022, חברות ומדינות רבות משקיעות משאבים אדירים בפיתוח יכולות של בינה מלאכותית כללית, אבל עדיין יש מחלוקת באשר ללוח הזמנים הנדרש למימושו. קל וחומר שיש דרך ארוכה, ושנויה במחלוקת אף יותר, עד הגשמת חזון הסינגולריות – בין שהוא רצוי ובין שאין הוא רצוי.

אנו נתרכז אפוא בהשלכות של אותן מערכות בינה מלאכותית שכבר קיימות בשוק על המציאות של היום. מאחר שהמערכות הקיימות משתייכות לקטגוריה של בינה מלאכותית צרה לא נדון בהשלכות החברתיות והמוסריות של בינה מלאכותית ״כללית״ או ״עילאית״, כלומר לא נדון למשל בשאלה אם למערכות בינה מלאכותית יש זכויות אדם ובאילו תנאים.[63]

61 Irving John Good, *Speculations Concerning the First Ultraintelligent Machine* 6 Advances in Computers 31 (1966)

62 Ray Kurzweil, The Singularity Is Near (2005)

63 ראו בעניין זה קביעתו של המשנה לנשיאה מלצר בפס' 31 בפסק דינו בעניין בג"ץ 7846/19 **פרקליטות המדינה – יחידת הסייבר נ' לורי שם טוב** (12.4.2021, פורסם באר"ש), שלפיה "אין זכויות אדם [...] לרובוטים, מה גם שחלק מאותם רובוטים אפילו לא מופעלים על־ידי בני אדם, אלא על־ידי אינטליגנציה מלאכותית". לדיון מעמיק יותר בשאלת האישיות המשפטית של מערכות נבונות ראו Kurki, **לעיל** ה"ש 19, בעמ' 175–189.

1.3 למידת מכונה ולמידה חישובית

כבר עתה מחשבון יד פשוט מבצע פעולות אריתמטיות במהירות שבן אנוש אינו יכול להשתוות לה. עם זאת, ספק אם אפשר לטעון שלמחשבון היד יש בינה מלאכותית. המחשבון לא ניחן באוטונומיה, בניתוח רציונלי וביכולת הסתגלות. מכונות בעלות בינה מלאכותית יכולות ללמוד ולהשתכלל.

עד מחצית שנות השמונים הייתה הפרדיגמה הקלסית בינה מלאכותית סימבולית (המוכרת גם בשם Good Old-Fashioned AI, GOFAI). בינה מלאכותית סימבולית מבוססת על קידוד המציאות במכונה באמצעות סמלים. המערכת יכולה לנתח את היחסים שבין הסמלים בהסתמך על מנוע שהחוקים קודדו בו מראש כמארג אלגוריתמי מסועף של הוראות ״אם־אז״ (if-then). כדי לפתח מערכת שמבוססת על בינה מלאכותית סימבולית נדרש ידע אפריורי בתחום שבו היא עתידה לפעול – המערכת לא תלמד תחום זה באופן עצמאי ותשכלל את האלגוריתם שלה בהתאם, אלא תסתמך על מאגר המידע שהוזן לתוכה. המערכת יכולה לקבל החלטות או המלצות המבוססות על נתונים אלו ברמת מומחיות שמומחה אנושי אינו יכול להגיע אליה ומערכות מומחה כאלו היו הדומיננטיות ביותר בשוק לאורך זמן.

שלא כמו ייצוג סימבולי של מידע מתוך מאגר מידע אפריורי ושל תהליכים קוגניטיביים מסדר גבוה, שעליו מסתמכות מערכות מומחה מטיפוס GOFAI, בגישת הקשרנות (connectionism) לפיתוח בינה מלאכותית[64] מתחילים מלמטה. גישה זו מבוססת על ההנחה שפיתוח רשתות שמורכבות מיחידות עיבוד פשוטות המכונות נוירונים (אף שאין הם דומים לנוירונים אנושיים), וקישור הרשתות אלו לאלו, יכולים להניב אינטליגנציה – בדומה לאופן שבו המוח האנושי מקושר.

64 לעבודות חלוציות בתחום ראו Warren S. McCulloch and Walter Pitts, *A Logical Calculus of the Ideas Immanent in Nervous Activity* 5 The Bulltetin of Mathematical Biophysics 115 (1943); Marvin Minsky and Seymour A. Papert, Perceptrons: an introduction to computational geometry (1969). לקישור מוקדם בין קישוריות לבינה מלאכותית כללית ראו קלארק, **לעיל** ה"ש 9.

"למידה היא משהו בלתי רצוני שאנחנו, בני האדם, עושים כל הזמן. אנחנו משתמשים במידע שאנחנו קולטים כדי לייצר ידע". כך הסביר בריאיון לעיתון "דה מרקר" פרופ' שי שלו שוורץ, סגן נשיא חטיבת הטכנולוגיה בחברת מובילאיי ומומחה לתחום הלמידה העמוקה מהמחלקה למדעי המחשב באוניברסיטה העברית.[65] "אם נראה לאדם כלשהו תמונה שמוצגת בה מכונית, הוא יידע שיש בה מכונית, אך לא יידע להסביר איך הוא יודע את זה. זה מה שאני, כחוקר, רוצה לגלות וזה מה שמערכות למידה עמוקה עושות – מנסות להבין כללים. אלה כלים אוטומטיים של הסקה מהפרטים אל הכלל".

מערכות הבינה המלאכותית נועדו להתמודד עם כמויות עצומות של מידע ולהפיק ממנו תובנות מופשטות ברמות תחכום גבוהות. לכן ככל שהן קולטות ומעבדות מידע רב יותר מכל התחומים, כך משתפרת היכולת שלהן ואפשר להטמיע אותן בתחומים רבים יותר. על כן נוסף על ההשפעה המכריעה שיש לאיכות אלגוריתם הלימוד שבבסיס המכונה, יכולת הלמידה שלה תלויה גם באיכותו ובהיקפו של המידע המוזן למכונה.

"למידת מכונה" היא שם של אוסף טכניקות סטטיסטיות, בהן טכניקות הקשורות לרשתות נוירונים, שנועדו לזהות דפוסים או כללים שיכולים להסביר נתונים או לחזות תוצאות עתידיות. זהו תהליך שנעשה באופן אוטונומי, ללא כללים או חוקים ישירים שקודד מפתח המערכת; שלא כמו מערכות מומחה, המבוססות על תובנות אנושיות שהפכו לקוד, למידת מכונה מפיקה תובנות אלו בכוחות עצמה. כשאלגוריתם זיהוי תמונה כגון זה שפיתח ג'פרי הינטון מבחין בין דמות של כלב לדמות של מכונית, אין הוא מתבסס על מערכת של כללים (כגון "למכונית יש ארבעה גלגלים", "לכלב יש ארבע רגליים וזנב"), אלא מייצר מודל מתמטי משלו, בהתאם למאגר הנתונים (במקרה זה, למשל, מאגר הנתונים יכול להיות אוסף של תמונות שכל אחת מהן מוגדרת "כלב", "מכונית" או "לא כלב ולא מכונית"). מאגר הנתונים סופק לצורכי אימון.

65 אלירן רובין "הכירו: Deep Learning – הדבר הכי לוהט בהיי־טק עכשיו" **TheMarker** (31.3.2016).

באמצעות למידת מכונה תוכנה מסוגלת לייצר בעצמה רצף של כללים על בסיס מאגר נתונים שעומד לרשותה ובהתאם למשוב של מתכנת או גורם אחר. ברמה הבסיסית ביותר למידת מכונה מייצרת עץ החלטות. עץ ההחלטות הוא בעצם רצף של החלטות ״אם־אז״ עד לקבלת תוצאה סופית. יתרונה של שיטה זו שהיא עקיבה וברורה להבנה. ואולם שיטת לימוד זו מוגבלת לפעולות שהן לינאריות באופיין וניתנות להמרה לעץ החלטות.

כשמדובר בביצוע משימות מורכבות יותר, יש צורך להפעיל מתודולוגיות למידת מכונה מתוחכמות יותר, כגון רשתות נוירונים. רשתות נוירונים נקראות כך משום שהן בנויות בדומה למבנה של מוח אנושי, שמייצר קשרים בין רצפטורים ונוירונים ומסתמך על הקשרים האלה כדי לבסס פעילות.[66] בלמידת מכונה מסוג רשת נוירונים המתכנת מזין לתוכנה מאגר נתונים, ונותן לאלגוריתם משוב – מתי הוא הצליח בביצוע הפעולה ומתי לא. הצלחה מתגמלת ומבססת את קשר הנוירונים ולאורך זמן התוכנה מוסיפה לבסס קשרים כאלה עד שהמתכנת סבור שהיא הגיעה לשיעורי הצלחה גבוהים די הצורך. התוצאה הסופית היא רצף של פקודות שמעיד על הצלחה סטטיסטית במשימה. מתודולוגיה זו, הנפוצה בשיטות הלמידה המתקדמות, יעילה מאוד, אך החיסרון שיש בה לענייננו הוא שהיא מתבססת על מִתאם ולא על קשר סיבתי.

נוסף על כך, הליך הייצור של אלגוריתמים דורש משאבים ניכרים משני סוגים. ראשית, יש צורך במאגר נתונים גדול, עשיר ומטויב, אשר יאפשר לתוכנה להגיע מהר ככל האפשר להצלחה סטטיסטית במשימתה. ככל שהפעולה הנלמדת מורכבת יותר, כך נדרש מאגר נתונים גדול יותר. לכן אין פלא שהשחקנים המצליחים ביותר בתחום הבינה המלאכותית הם שחקנים שממילא יש בידם מאגרי מידע גדולים.[67] המשאב השני הוא פעולה אנושית: מתן משוב למכונה כדי שהיא תוכל ללמוד אם הצליחה או לא. פעולה זו אינה רק יקרה, אלא גם מועדת לטעויות; לכן גם בעניין זה יש עדיפות לחברות ענק עתירות משאבים

66 ראו בחלק 1.3 **לעיל**.

67 ראו בחלק 1.6.4 **להלן**.

שבאפשרותן לממן את ההשקעה הכבדה, המניבה תוצאות בטווח הארוך. בשל קשיים אלו פותחו דרכים שאינן מחייבות משוב אנושי.

דרך אחת, המתאימה במקרים מסוימים, היא לימוד אדוורסרי. למשל, ליצור מערכת אחת שיש לה מטרה, כגון לפרוץ צופן, ומערכת אחרת שיש לה מטרה הפוכה, כגון לייצר צופן שלא ניתן לפיצוח. בכל סבב לימוד המערכות משתכללות זו כנגד זו ללא צורך במשוב אנושי.

ואולם ברשתות הנוירונים טמון קושי של ממש, שקשור להיותן מעין "קופסה שחורה" – קשה להסביר באופן נקודתי כיצד נוצר הפלט המסוים שאליו הגיעו. המתכנת אינו יודע מה משמעות הקשרים שנוצרו וכיצד הגיעה התוכנה להצלחה, כלומר אין הוא יודע מדוע התוצר הסופי והפלט של התוכנה הם כפי שהם. אומנם מקובל להשוות בין רשתות נוירונים לבין פעולת המוח האנושי, אבל חשוב לציין שהדרך שבה הגיעה התוכנה להצלחה אינה סיבתית, אלא תוצר של פעולות סטטיסטיות, לעיתים רנדומליות. התוכנה אינה מסוגלת להפיק פלט שמסביר כיצד ומדוע התקבלה תוצאה או החלטה מסוימת ולא אחרת. נחזור ונעסוק בעניין זה בהמשך הדברים.

למידה עמוקה היא טכניקה המשמשת ללמד את המכונה לזהות דפוס מסוים על ידי "אימון" – למכונה מוזנת כמות גדולה של נתונים וניתנת לה הוראה לזהות במקבצי הנתונים דפוס או דפוסים מסוימים. לדוגמה, אפשר ללמד את המכונה להבחין בין תמונות שיש בהן כלבים לתמונות שאין בהן כלבים. כדי לאמן את המערכת היא מקבלת מקבצים של נתונים גולמיים – בדוגמה שלנו, את מיקום הפיקסלים בתמונה ואת הגוון שלהם. בדוגמה זו מדובר בלמידה מפוקחת, כך שהנתונים גם מתויגים – כל מקבץ נתונים (תמונה יחידה) מוגדרת (או מתויגת) "כלב" או "לא כלב". אם מספר מקבצי הנתונים יהיה גדול ומגוון די הצורך, המערכת תלמד לזהות טווח מסוים של דפוסים שמאפשרים לה לקשור בין הנתונים הגולמיים לתוצר המופשט הסופי. בדוגמה שלנו היא תלמד לזהות למשל שבתמונות שיש בהן כלבים ייתכנו יחסים מסוימים בין קווים לזוויות או גוונים מסוימים. לאחר שהיא תבחן די תמונות היא תוכל לקבוע על פי מאפיינים אלו שמדובר בתמונה של כלב.

המוח האנושי קולט כמות עצומה של נתונים גולמיים בכל רגע נתון ועל יסוד ניסיון נרכש הוא יודע לעבד אותם ולתת להם משמעות. ניסיון העבר מאפשר לנו לזהות צליל של צופר מכונית, ולהבחין בינו לבין בכי של תינוק, או לזהות אדם מסוים ולא אחר. למידת מכונה פועלת באופן דומה. המכונה, בדומה למערכת העצבים האנושית, יודעת להפיק פשר ממידע גולמי ולא מתויג. מובן שחרף ההשוואה המוח האנושי ומערכות למידה עמוקה אינם זהים ועל אף השימוש במונח נוירונים אין הם זהים בפעולתם לנוירונים הביולוגיים. עם זאת, בדומה ל"שכבות" הנוירונים במוח, המחוברות על ידי סינפסות, גם בלמידה עמוקה יש שכבות שונות של עיבוד מידע. שכבות עיבוד אלו מאפשרות רמות שונות של הפשטה בעיבוד המידע – רמת הפיקסל; זיהוי צורות גאומטריות; זיהוי חפצים קונקרטיים (חלון, כביש או לוחית זיהוי); ולבסוף – הקביעה אם מדובר באדם, בעל חיים או רכב.[68]

1.4 למידה מפוקחת, למידה לא מפוקחת ולמידה על ידי חיזוקים

למידת מכונה יכולה להיות למידה מפוקחת (supervised learning) או למידה לא מפוקחת (unsupervised learning). למידה מפוקחת פירושה שלאלגוריתם מוגדר משתנה מטרה מסוים לחיזוי. למשל, אם המערכת נועדה לחזות רמת סיכון פיננסי של לווה (סיכון גבוה/נמוך),[69] משתנה המטרה כבר ידוע מראש והמכונה לומדת לחזות אותו בהתבסס על נתוני אימון המוזנים לתוכה. נתונים אלו יכולים לכלול, לדוגמה, את ההיסטוריה הפיננסית של לווים מסוכנים ושל לווים לא מסוכנים ונתונים נוספים כמו מצב משפחתי, צבע עיניים או דת.[70] אם יוזנו למערכת די נתונים היא תדע לבחור מודל מתמטי שחוזה את רמת הסיכון בדיוק יחסי. משתנה המטרה יכול להיות קטגוריאלי (מסוכן/לא מסוכן או רמת סיכון גבוהה/בינונית/נמוכה) והוא יכול להיות רציף (הסיכוי לפשיטת רגל). תיוגו מראש ייעשה בהתאם.

68 שם.

69 ראו לדוגמה ס' 16(ב)(4) לחוק נתוני אשראי, התשע"ו–2016, ס"ח 2551 בעמ' 838 (**להלן**: חוק נתוני אשראי).

70 ייתכן שנתונים אלו יהיו אסורים לשימוש, ראו ס' 51 לחוק נתוני אשראי.

שתיים מן המגבלות של למידה מפוקחת הן שלאימון נחוצות כמויות עצומות של נתונים ושלמערכת אין יכולת להתמודד עם סיטואציות חדשות. היכולת להתמודד עם מצבים חדשים היא מיומנות שבני אדם ובעלי חיים רוכשים בשלב מוקדם מאוד בחייהם. למשל, אדם אינו צריך לנסוע במכונית שלו מעבר לצוק כדי להבין שהיא תיפול ותתרסק; ואין צורך להסביר לאדם שכאשר אובייקט מסתיר אובייקט אחר האובייקט המוסתר ממשיך להתקיים או שכדור שנחבט באלה לכיוון נדנדה עלול לפגוע בילד שמתנדנד. ההתבוננות בעולם והפעולה בו מלמדים בני אדם דברים כמו ממדים, סיבתיות, כוח כבידה וכדומה. אלו הן אבני הבניין שסביבן נצבר ידע מורכב יותר. מערכות הבינה המלאכותית הנוכחיות חסרות את הידע המשותף הזה, ולכן הן רעבות לנתונים ודורשות דוגמאות מתויגות באשר לנתונים שאינם במאגר הידע הבסיסי שלהן.

בלמידה לא מפוקחת לא נעשה אימון כזה. המכונה מקבלת מאגר נתונים לא מתויג ומפיקה ממנו תובנות בכוחות עצמה. כיוון שזהו מאגר לא מתויג, אי־אפשר לאמן את המכונה לזהות משתנה מוגדר (המכונה לומדת לבד להפריד, למשל, בין צורות גרפיות שונות ומחלקת אותן לקטגוריות לפי ה"היגיון" הסטטיסטי שלה; או מבחינה בין רמות סיכון פיננסיות בדרך שונה מזו של מומחה אשראי אנושי).

למידה לא מפוקחת אינה מתאימה לחלוקת הנתונים לקטגוריות שהוגדרו מראש או לזיהוי מִתאם בין פרמטרים מוגדרים מראש, אך היא מתאימה למטרות של קיבוץ הנתונים לפי קטגוריות שהאלגוריתם מזהה. למידה לא מפוקחת מתאימה יותר לקיבוץ גורמים יחד על בסיס מאפיינים דומים, לזיהוי חריגות או לזיהוי משתנים שקשורים זה לזה. למידה כזאת מאפשרת לאלגוריתם לזהות מאפיינים בנתונים שמפתחי האלגוריתם לא הגדירו אותם מראש.

האלגוריתם הממליץ למשתמש על השיר הבא שינוגן בנגן המוזיקה המקוון שלו, למשל, מבוסס על למידה לא מפוקחת. במערכת מוזן מאגר רשימות השמעה של משתמשים והאלגוריתם מנסה לזהות בהן מאפיינים דומים. כשמשתמש בוחר להאזין לשיר מסוים, האלגוריתם יכול להמליץ לו להאזין לשיר נוסף שמשתמשים אחרים בעלי היסטוריית האזנה דומה בחרו להאזין לו. האלגוריתם יכול להתבסס גם על נתונים נוספים (למשל, תיוגים אקראיים של השירים לפי ז'אנר או שנת הקלטה). בכל מקרה, האלגוריתם מחליט באופן עצמאי מהם הפרמטרים הרלוונטיים לצורך מתן המלצה מתאימה.

גם אלגוריתמים שמזהים שימוש חריג בכרטיס אשראי מבוססים על למידה לא מפוקחת. הם אינם נדרשים לתייג מראש שימושים חריגים או להגדיר פעולות הנחשבות סבירות ופעולות המעוררות חשד אלא רק לזהות שינויים בדפוסי השימוש הרגיל של הלקוח.

טכניקת למידה אחרת היא למידה על ידי חיזוקים (reinforcement learning).[71] בטכניקה זו נבחן הפלט של המכונה וניתן לה משוב. זוהי למידה באמצעות ניסוי וטעייה. האלגוריתם לומד בתהליך חזרור (iterative process) אילו פעולות או פלטים מקבלים משוב חיובי ואילו מקבלים משוב שלילי. באמצעות משוב זה, דמוי מנגנון של שכר ועונש, המכונה לומדת לשכלל ולדייק את עצמה. אם למשל היא לא זיהתה רכישה של תכשיטים יקרים בכרטיס אשראי כשימוש חורג של לקוח, אך בדיעבד התגלתה פעולה זו כשימוש חורג, תלמד המכונה באמצעות המשוב מהטעויות שלה. אפשר לשכלל למידה על ידי חיזוקים בטכניקות אחרות של למידת מכונה (מפוקחת או לא מפוקחת). כך לא זו בלבד שהמערכת לא תחזור על טעויות עבר, אלא היא תלמד לזהות דפוסים חדשים ותימנע משגיאות זיהוי שעדיין לא נתקלה בהן.

מערכת אלפא־גו, למשל, התחילה בלמידה מפוקחת על בסיס נתוני אימון של משחקי גו היסטוריים, אך אחר כך נעזרה גם בתהליך של למידה על ידי חיזוקים. אחרי שהמערכת צברה ניסיון מעיבוד המהלכים במשחקים האנושיים, היא שיחקה משחק רב של משחקים נגד עצמה וקיבלה חיזוק בכל פעם שבחרה אסטרטגיה מנצחת. המערכת זוכרת כמובן את כל הטעויות שכבר עשתה ונמנעת מלחזור עליהן. אלפא־גו זירו, היורשת של אלפא גו, לא נדרשה ללמידה מפוקחת והתבססה אך ורק על למידה על ידי חיזוקים.

פוטנציאל ההשפעה של הבינה המלאכותית על כל תחום בחיינו ברור. ככל שהטכניקות האלו ישתכללו, ועימן גם סוג הנתונים שהן יכולות לעבד ואיכותם, או שתגדל יכולתן להגיע לתוצרים בלי לעבד כמויות עצומות של מידע, כך יגדל גם הפוטנציאל הטרנספורמטיבי שיביאו איתן.

71 Mark Coeckelbergh, AI Ethics 86 (2020)

1.5 בינה מלאכותית יוצרת (גנרטיבית)

מומחים בעלי שם לבינה מלאכותית, לא כל שכן אנשים מן השורה, שפשפו את עיניהם בתדהמה לנוכח מקצת הפיתוחים שיצאו לאוויר העולם בשנת 2022, בעת כתיבת מחקר זה. התעוררה תחושה שהמכונות חצו סף בלתי נראה ונעשו יצירתיות, תכונה שבעבר הובטח שתישמר לבני אנוש בלבד. DALL-E2 של OpenAI ו־Imagen של גוגל, Stable Diffusion, D-ID, Midjourney – כולם מסוגלים לייצר תמונות וקטעי וידיאו על בסיס פקודה טקסטואלית (Prompt), המכונה לעיתים "לחישה".[72]

אליהם מצטרפים מודלי שפה גנרטיביים שמייצרים טקסטים ברמה גבוהה, עונים על שאלות בצורה סמי־משכנעת ואפילו מבינים מה מצחיק בבדיחה: GPT-4 מבית Open AI, Jurassic-X של AI21 הישראלית ו־PaLM של גוגל. הבנה עמוקה של שפה היא קפיצת מדרגה ביכולות הבינה המלאכותית. כבר בשנת 2017 ידעו מודלי השפה הראשונים, שהתבססו על למידה לא מפוקחת, לומר מהי המילה החסרה או הבאה כשקיבלו סדרה של מילים; ואילו המודלים הנזכרים כאן יודעים לבצע גם מטלות מורכבות יותר, למשל לשכתב טקסט, לנתח סנטימנט בטקסט, לענות על שאלות או אפילו לקבוע אם הטקסט הגיוני או לא. בחודשים הבאים ובשנה הבאה נראה שתפותח גם היכולת לממש מיומנויות מופשטות (למשל לקחת משפט בשפה טבעית ולהכניסה כנוסחה למחשבון; או לקחת שאלות בשפה טבעית ולייצג אותן כנוסחת אקסל או כשאילתת SQL).[73]

מודלים אלו יכולים לסרוק את רשת האינטרנט כדי ליצור תוכן ערוך, החל בחדשות היום, עבור בהמלצות על מוצרים ושירותים וכלה בהמלצות למסלולי טיול, ואפילו לספק טבלאות מסודרות עם קישורים רלוונטיים למקור. הם אף

72 סהר מור "בינה מלאכותית יצרה את התמונה הזאת בהסתמך על טקסט בלבד" **ynet** (17.4.2022); יובל מן "הכירו את מחולל התמונות החדש של גוגל" **ynet** (24.5.2022).

73 Nuobei Shi, Qin Zeng, and Raymond Lee, *Language Chatbot: The Design and Implementation of English Language Transfer Learning Agent Apps,* 2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) (2020); Aditya Ramesh et al., *Hierarchical Text-Conditional Image Generation with CLIP Latents* (13.4.2022), available at arXiv

יכולים להדריך כיצד לכתוב תוכן ויראלי לרשתות חברתיות בעזרת ניתוח קוד מן האלגוריתם של רשת חברתית.

שלושת החידושים של המודלים היוצרים לעומת המודלים שקדמו להם:

(1) הם כלליים ולא ספציפיים, כלומר הנתונים שעליהם התאמן המודל אינם ייחודיים למטרה מסוימת (למשל כדי לייצר חיזוי סטטיסטי על מתן הלוואות או לענות על שאלות בנוגע לספרות מדעית בתחום מסוים) אלא נתונים כלליים (כל רשת האינטרנט, כל הערכים בוויקיפדיה וכיו"ב). דבר זה מאפשר מרחב שימושים עצום ואף מספק בסיס לשימושים ספציפיים, שמחייבים הוספת נתונים ייחודיים.

(2) הם יוצרים ולא תיאוריים, כלומר הם יכולים ליצור תוצרים חדשים, החל בכתיבת טקסט או ביצירת דימויים חזותיים ותבניות קול וכלה בכתיבת קוד לביצוע מטלות, ולא רק לשקף תבניות בנתונים הקיימים, כמו למשל זיהוי פנים, קלסיפיקציה של טקסט וחיזוי סטטיסטי מבוסס נתונים.

(3) הם נגישים ולא טכניים – אפשר לתקשר איתם, לשאול אותם שאלות מתוחכמות ולבקש מהם לבצע משימות בשפה פשוטה ("שפה טבעית"), ולא באמצעות כתיבת קוד.

הואיל והשפה הכתובה והמדוברת היא בסיס לכלל הפעילות האנושית, ובכלל זה ליכולת לחשוב חשיבה מופשטת, לפתח רעיונות מורכבים ולמסור אותם מדור לדור וממקום גאוגרפי אחד למשנהו, לחפש, לתקן, לסכם, לסנן, לצטט, לתרגם ולסכם אותם – הואיל וכך, היכולות היוצרות של המודלים של השפה נתפסות טרנספורמטיביות במיוחד לכלל המין האנושי ולכלל המגזרים בו.

גם למעבר למשימות מבוססות שפה טבעית (המכונות גם "נוסחאות" או prompts) יש השלכות. ראשית, העובדה שכל פלט ממודל שפה יכול בתורו לשמש קלט למודל שפה ("טקסט פנימה, טקסט החוצה") מאפשרת להמשיך לאמן ולשכלל את המודלים. שנית, אפשר לבצע פעולות עשירות ומתוחכמות באמצעות "שרשור הנחיות". כשמדובר במשימות שדורשות פעולות ביניים או חשיבה מרובת שלבים, הפלט של תת־משימה אחת משמש קלט למשימה הבאה. ולעיתים תת־משימה כוללת גם אחזור מידע ממאגרים חיצוניים (כמו חיפוש מידע במנוע חיפוש או משיכת מידע מכתובת אינטרנטית נתונה).

לכן בטווח הקצר יהיו השימושים של משתמשים אנושיים במודלים הגדולים איטרטיביים ומשותפים. המשתמש האנושי יציג למודל הנחיה ראשונית או שרשרת הנחיות כדי להפיק פלט נתון; יבחן את הפלט ויכוונן את ההנחיה כדי לדייק את הפלט ולשפר את איכותו; יפעיל את המודל פעמים רבות על אותה בקשה כדי לבחור את הגרסאות הרלוונטיות ביותר של הפלט; ולפני השימוש הסופי בתוצר יטייב את הפלט בעצמו. אבל בטווח הארוך יותר, ככל שהמודלים ישתפרו וככל שהמוצרים הבנויים על בסיסם ייעשו נוחים יותר לשימוש ויוטמעו עמוק יותר בשגרות עבודה, יבוצעו המשימות בכל שלב ושלב ללא פיקוח אנושי.

מפתחי מודלים של בינה מלאכותית מתקדמים במהירות ויכולות המודלים לשימוש כללי הולכות וגדלות. המודלים לומדים באמצעות אימון ומציגים יכולות חדשות, לעיתים אף כאלה שהמפתחים שלהם לא צפו. מחקרים מלמדים על "יכולות מתעוררות" שלא נצפו ולא תוכננו מלכתחילה, ככל הנראה בשל היכולת של מודלים גדולים מאוד להשלים משימות, שלא התקיימה כשהמודלים היו קטנים יותר. מדעני מחשב מנסים ליצור רשימה של יכולות חדשות אלו, שלא נדונו בספרות בעבר[74] ויש המכנים אותן "ניצוצות". נעשים ניסיונות לא רק לזהות יכולות נוספות אלא גם להבין מדוע וכיצד הן צומחות.

ואולם לצד הפוטנציאל הכביר במודלים של שפה שאפשר לממש על בסיסם בינה מלאכותית יוצרת מתגלים גם חששות. ראשית, הואיל והממשקים מבוססים על ניסיון לצפות את רצונו של המשתמש, הם סובלים מתופעה המכונה "הזיות" (הלוצינציות) ומשיבים תשובות שאינן נכונות עובדתית. על אף אזהרותיהן של החברות המעניקות שירותים אלו, הן באמצעות הצהרות הן באמצעות הממשקים עצמם, משתמשים רבים עדיין נופלים בפח ולכן טמונה במודלים אלו סכנה של הפצת מידע מסולף (disinformation) ומידע מוטעה (misinformation) בציבור.

74 ראו Jason Wei et al., *Emergent Abilities of Large Language Models,* Transactions on Machine Learning Research (28.3.2023); Sébastien Bubeck et al., *Sparks of Artificial General Intelligence: Early Experiments with GPT-4* (13.4.2023), available at arXiv; Deep Ganguli et al., *Predictability and Surprise in Large Generative Models* (15.2.2022)

שנית, מודלים של שפה שעל בסיסם מופעלים ממשקי שיחה עם משתמשים (למשל צ׳אטבוטים כמו ChatGPT, בארד ובינג או עוזרים אישיים אוטומטיים כגון AutoGPT) עוקבים אחר הוראות או "הנחיות" של משתמשים ויוצרים תוכן בהסתמך על יכולת חיזוי שרכשו מבסיס נתוני האימון שלהם. השילוב בין היעדר צורך במיומנויות תכנות וכתיבת קוד ובין יכולתם של המודלים למלא הוראות הופך אותם למוצרים שימושיים לכלל האוכלוסייה, אך גם עושה אותם פגיעים לשימוש לרעה. דוגמה לשימוש כזה היא "הזרקת הנחיות" המכוונות את מודל השפה להתעלם ממנגנוני ההגבלה והבטיחות שלו עצמו, מה שמכונה לעיתים "פריצה מן הכלא" (jailbreaking). הדבר אפשרי אם מבקשים מהצ׳אטבוט ליצור "משחק תפקידים" שבו הוא משחק מודל נוסף שנדרש לעשות מה שהמשתמש מבקש ממנו, ללא מגבלות. כך אפשר להשתמש בממשק כדי לבצע פעולות מזיקות, כגון הפצת תאוריות קונספירציה, או פעולות פליליות, כגון גנבת פרטי אשראי ויצירת הוראות להכנת חומרים מסוכנים. הפלטפורמות העיקריות של הממשקים האלה, כגון OpenAI, מנסות לאסוף דרכי פריצה מן הכלא ולהזין אותן לנתוני האימון של המערכות כדי שיוכלו להתנגד להן בעתיד, אבל נראה שמדובר בקרב קשה ומסובך.[75]

שלישית, החיבור בין ממשקים המבוססים על המודלים ובין רשת האינטרנט מביא איתו חששות. מודלים גדולים של בינה מלאכותית מתאמנים על כמויות עצומות של נתונים מהאינטרנט. לפי שעה חברות הטכנולוגיה פשוט סומכות על אמיתותם וניקיונם של הנתונים. אבל חוקרים גילו את האפשרות "להרעיל" מאגרי נתונים שמודלים גדולים של שפה מתאמנים עליהם, למשל על ידי רכישת כתובות של אתרי אינטרנט והצפתם בתמונות ובטקסטים שבתורם נכנסים לתוך מאגרי הנתונים האלה. אם לא די בכך, ככל שמשהו חוזר על עצמו בנתוני האימון של המודל פעמים רבות יותר, כך הוא משפיע יותר על התנהגות המודל ומשפיע על תוצריו.[76] יתר על כן, ממשקים שחשופים לרשת האינטרנט ושואבים ממנה מידע חשופים גם להתקפה שנקראת "הזרקה מהירה עקיפה", שבה גוף עוין משנה

75 Will Douglas Heaven, *The Inside Story of How ChatGPT Was Built from the People Who Made It*, MIT Tech. Rev. (3.3.2023)

76 Nicholas Carlini et al., *Poisoning Web-Scale Training Datasets is Practical* (20.2.2023), available at arXiv

אתר אינטרנט על ידי הוספת הנחיות מוסתרות שגלויות רק לעוזר האוטומטי ונועדו להשפיע על פעולותיו (למשל גורמות לו להעביר אליהם פרטי כרטיסי אשראי, לשלוח בשמו מסרים בדואר אלקטרוני ועוד).

האפשרות ליישם את היכולות של המערכות כדי לגרום נזק, למשל לשכלל מיומנויות הטעיה, לסלף מידע, לבצע עבירות סייבר או לתכנן כלי נשק וטרור, אף שמפתחיהן לא התכוונו לכך, מכונה the alignment problem (בעיית הקשר בין משימה לתוצאה).[77] היא נתפסת כאחת הבעיות הקשות לפתרון במרחב השימושים במערכות גנרטיביות[78] ונדון בה בהמשך דברינו.[79] במקביל, כדי להתגונן מפני היכולת להיעזר בממשק כגון ChatGPT לאיתור חולשות סייבר ולכתיבת תוכנה זדונית שמנצלת אותן וגונבת מסמכים, מציעה חברת מיקרוסופט, באמצעות שירות CoPilot, כלים שמטרתם לזהות איומי סייבר ולהגן מפניהם.[80]

יכולות היצירה (היכולות הגנרטיביות) של המכונה ניכרות גם ביכולתה להתבטא בשפות לא טבעיות, כגון שפות קוד שונות. מערכת CoPilot מבית OpenAI ו־GitHub יוצרת כ־40% מן הקוד בחברות המשתמשות בה, ואפשר שבעתיד היא תייתר את הצורך לדעת לכתוב קוד ויהיה די בלימוד "לחישה" ובהזנת הוראות בדבר המוצר המבוקש למערכת.

קרן ההון סיכון סקויה קפיטל הודיעה בסוף 2022 כי היא מאמינה שבינה מלאכותית תיצור "טריליוני דולרים של ערך כלכלי".[81] לדעתה, בינה מלאכותית היא ההתפתחות הטכנולוגית החשובה ביותר בימינו. בתרשים שלהלן מתומצתת התחזית שלה לשנים הבאות.[82]

77 Richard Ngo, Lawrence Chan, and Sören Mindermann, *The Alignment Problem from a Deep Learning Perspective* (22.2.2023), available at arXiv

78 Yotam Wolf et al., *Fundamental Limitations of Alignment in Large Language Models* (9.4.2023), available at arXiv

79 ראו **להלן** בפרקים 9, 10.

80 Microsoft Security Copilot, *Introducing Microsoft Security Copilot,* YouTube (28.3.2023)

81 *Generative AI: A Creative New World*, Sequoia

82 **שם.**

תרשים 1
תחזית יכולות הבינה המלאכותית במרחבי יצירה שונים

	לפני 2020	2020	2022	2023	2025?	2030?
טקסט	איתור ספאם תרגום שאלות ותשובות בסיסיות	יכולות כתיבה בסיסיות "טיוטה ראשונה"	יכולות כתיבה ארוכות יותר "טיוטה שנייה"	כתיבה משויפת במרחבים ספציפיים (למשל מאמרים מדעיים)	גרסה סופית של טקסט ברמה גבוהה מזו של אדם ממוצע	גרסה סופית של טקסט ברמה גבוהה מזו של כותב מקצועי
קוד	שורה אחת	יכולת יצירה של קוד רב־שורות	גרסאות קוד ארוכות דיוק טוב יותר	יותר שפות קוד יותר מרחבים ספציפיים	מעבר ישיר מתיאור טקסטואלי למוצר מבוסס קוד ברמת טיוטה (text to product)	מעבר ישיר מתיאור טקסטואלי למוצר מבוסס קוד, ברמה סופית טובה מזו של מפתח מיומן
תמונות			אומנות צילום	טיוטה (מוק־אפ) בתחומי העיצוב, האדריכלות וכד'	מוצר מוגמר (מוצר מעוצב, תוכנית אדריכלית וכד')	מוצר מוגמר ברמה טובה מזו של מעצבים, אומנים, צלמים ואדריכלים מקצועיים
וידאו ותלת־ממד			ניסיונות ראשונים ליצירת מודלים תלת־ממדיים	"טיוטה ראשונה" של תוכני וידאו ותלת־ממד	"טיוטה שנייה" של תוכני וידאו ותלת־ממד	משחקי וידאו וסרטים ברמת גימור מקצועית

ניסיונות ראשונים | כמעט כאן, בשלבים מתקדמים | בשלבים מתקדמים

1.6 הכוחות המניעים את מהפכת הבינה המלאכותית

כפי שראינו לעיל, היסודות התאורטיים של מהפכת הבינה המלאכותית הונחו כבר באמצע המאה הקודמת. תקופתנו אינה אפוא התקופה הראשונה שבה תולים תקוות בהתפתחות תחום זה ומעלים חששות מהשלכותיו – מחזון אופטימי בדבר סינגולריות[83] עד אזהרות מפני השתלטות עוינת של בינה מלאכותית עילאית על הפלנטה.[84] תקופות אלו של התלהבות מבינה מלאכותית – המכונים "קיצים" בעגה המקובלת בתחום – הסתיימו באכזבה, שכן הטכנולוגיה לא עמדה בציפיות שתלו בה. אפשר שהקיץ הנוכחי של הבינה המלאכותית, ועימו השיח על בינה מלאכותית והרצינות שבה נידונות טכנולוגיות אלו בקרב קובעי מדיניות,[85] יגיעו אל קיצם בחורף נוסף.[86] עם זאת, כיוון שהדיון הער במהפכת הבינה המלאכותית נובע משילוב של כמה גורמים – מהם טכנולוגיים ומהם כלכליים – אפשר לצפות שאולי הפעם תהיה אחיזתה חזקה וממושכת יותר.

1.6.1. כוח מחשוב

בינה מלאכותית, ובייחוד כזאת הנשענת על טכניקות של למידת מכונה בכלל ועל למידה עמוקה בפרט, נדרשת לכמויות עצומות של מידע כדי לאמן את עצמה, וגם ליכולת הטכנולוגית לעבד את המידע הזה, לאחסן אותו ולאחזר אותו.

עקב מגמות מזעור והתייעלות (הטרנזיסטורים המבוססים על סיליקון קטנו; מהירות העברת הנתונים ברשת גדלה; הייתה עלייה מטאורית ביכולת לאחסן מידע; ועלות האחסון ועיבוד הנתונים פחתה) אנשים פרטיים משתמשים היום במחשבים שלפני עשור או שניים בלבד נחשבו לבעלי יכולת של מחשב־על,

83 Kurzweil, **לעיל** ה"ש 62.

84 **לעיל** ה"ש 10–11.

85 ראו **להלן** בפרק 3.

86 Luciano Floridi, *AI and Its New Winter: From Myths to Realities*, 33 Philosophy and Technology 1 (2020); Sebastian Schuchmann, *Probability of an Approaching AI Winter*, Towards Data Science (17.8.2019); Kathleen Walch, *Are We Heading for Another AI Winter Soon?* Forbes (20.10.2019)

וכל ארגון מסחרי או מוסד שלטון יכול להרשות לעצמו לאגום ולשמור כמויות עצומות של מידע.

האלגוריתמים אינם פועלים בחלל ריק ונדרש כוח מחשוב כדי להפעיל אותם. חוק מור,[87] שניסח בשנת 1965 גורדון מור, ממייסדי חברת אינטל, חזה כי צפיפות הטרנזיסטורים במעגלים משולבים במחיר מינימלי תכפיל את עצמה בכל תקופה.[88] אומנם קצב הגידול המעריכי בכוח המחשוב דעך בשנים האחרונות,[89] אבל בימינו הגיע כוח זה לבשלות טכנולוגית המאפשרת יישומי בינה מלאכותית מבוססי למידה עמוקה.

ג׳נסן הואנג, מנכ״ל חברת השבבים האמריקאית Nvidia, ניסח את חוק הואנג, שקובע כי כוח המחשוב של מעבדים גרפיים צפוי לכל הפחות להכפיל את עצמו מדי שנתיים.[90] שלא כמו יחידות עיבוד מרכזיות (CPU), שעובדות באופן טורי (יש להן ליבה אחת שמעבדת סדרת מטלות בטור, כלומר זו אחר זו), מעבדים גרפיים (GPU) הם מרובי ליבות ולכן העיבוד נעשה במקביל (כמה ליבות מעבדות בו־זמנית מטלה מסוימת). אומנם מעבדים כאלו יועדו במקור ליישומים גרפיים, אבל נמצא שהם המעבדים המיטביים ליישומים של בינה מלאכותית המבוססת על רשתות נוירונים. כיום מעבדים גרפיים משמשים בפייסבוק, בנטפליקס, בגוגל ובחברות אחרות להרצת מנועים מתקדמים של בינה מלאכותית. לא צפויה האטה בקצב התפתחות המעבדים הגרפיים, ולכן הבינה המלאכותית המבוססת על רשתות עצבים מלאכותיות צפויה להמשיך להתפתח. התפתחות הבינה

87 Gordon E. Moore, *Cramming More Components onto Integrated Circuits*, 38 Electronics 114 (1965)

88 בהתחלה חזה מור שהיא תכפיל את עצמה אחת לשנה; בשנת 1975 הוא עדכן את התחזית וטען שהיא תכפיל את עצמה אחת לכל שנתיים.

89 יש חסמים בסיסיים להמשך הגידול בצפיפות הטרנזיסטורים וביכולת למזער אותם יותר ויותר. Suhas Kumar, *Fundamental Limits to Moore's Law* (18.11.2015), available at arXiv. אינטל עצמה הכריזה שבעתיד לא תוכל לשפר עוד את מהירות המעבדים, אלא רק את צריכת החשמל שלהם. ראו הראל עילם "החוק הפסיד: אינטל קוברת רשמית את חוק מור" **כלכליסט** (11.2.2016).

90 Tekla S. Perry, *Move Over, Moore's Law: Make Way for Huang's Law* IEEE Spectrum (2.4.2018)

המלאכותית היוצרת בסוף 2022 ותחילת 2023, והעובדה שמעבדים של חברת Nvidia משמשים גם את המודלים הגדולים של השפה, כגון אלה של חברת OpenAI וגוגל, הובילו לעלייה עצומה בשווי החברה עד יותר מטריליון דולר.[91]

נתונים שהציגה מעבדת Open AI מראים שמאז שנת 2012 גדל כוח המחשוב ששימש לאמן את מערכות הבינה המלאכותית המתקדמות פי שניים מדי 3.5 חודשים.[92] היכולות המרשימות של האלגוריתמים שפיתחה חברת DeepMind, למשל, התאפשרו בזכות השקעה כספית ניכרת. ואולם בשנת 2019 הפסידה החברה יותר מחצי מיליארד דולר; ובשנה העוקבת ויתרה חברת האחות שלה, גוגל (שתיהן בבעלות חברת אלפאבית), על סכום של 1.5 מיליארד דולר, שהיא הלוותה ל-DeepMind.[93] אפשר שהיסודות שהניחה החברה ישרתו את גוגל בהמשך, אך בגלל עלות הפיתוח ועלות כוח המחשוב הנדרש לאימון אין היא מתאימה לפי שעה לשימוש מסחרי רחב היקף.

כוח המחשוב השתפר מאוד והוא מאפשר לאמן מודלים על יסוד נתונים רבים מאוד, ובכלל זה כל המידע ברשת האינטרנט. כוח החישוב אפשר גם להגדיל את מספר הפרמטרים (הרכיבים במודלי הבינה המלאכותית האחראים על עיבוד המידע). המודלים הגדולים (למשל מודל השפה GPT-4) מכילים מיליארדים של פרמטרים, קפיצה עצומה לעומת המודלים שהיו מבוססים על מיליוני פרמטרים.

אבל כוח החישוב העצום הנדרש בשביל לאמן מודלים של בינה מלאכותית עלול להיות חסם בפני המשך פיתוח ויש הסבורים שמחשוב קוונטי יוכל לסייע לשבור את תקרת הזכוכית בהקשר זה. תכונות קוונטיות קיימות במרחבים מסוימים בטבע, למשל בהולכה חשמלית, בקרינה או בגבישים, ומחשוב קוונטי

91 ליטל סמט "אנבידיה עשתה היסטוריה: חברת השבבים הראשונה בשווי של טריליון דולר" **כלכליסט** (30.5.2023).

92 Dario Amodei and Danny Hernandez, *AI and Compute*, Open AI (16.5.2018). השוו לקצב הגידול (המעריכי אף הוא) של חוק מור – הכפלה מדי שנתיים – כדי לעמוד על הגידול המואץ בכוח המחשוב.

93 Gary Marcus, *DeepMind's Losses and the Future of Artificial Intelligence*, Wired (14.8.2019); Amy Thomson, *Google Waives $1.5 Billion DeepMind Loan as AI Costs Mount*, Bloomberg (17.12.2020)

פירושו פיתוח יכולות עיבוד המבוססות על מרחב אפשרויות ולא רק על בחירה בינארית בין 0 ל־1. בשנים האחרונות מופנים לתחום זה תקציבי עתק, הן של ממשלות ומוסדות מחקר אקדמיים הן של תאגידים ענקיים כמו גוגל ואיי־בי־אם. בתחום המחשוב מדובר בטכנולוגיה חדשה שאין לה עדיין ערך של ממש, אבל הפוטנציאל שלה ככל הנראה רב. במאי 2023 חשפה Nvidia מחשב־על ששמו Israel-1, שלדבריה הוא מחשב־העל החזק ביותר בישראל ואחד החזקים בעולם. מחשב־על הוא מחשב המורכב מאלפי מעבדים ומשמש לביצוע משימות מורכבות, בהן פיתוח יישומי בינה מלאכותית יוצרת והרצת סימולציות מדעיות. המטרה המוצהרת של מחשב־העל החדש היא לסייע בשיתופי פעולה עם התעשייה, במחקר ובפיתוח פנימי, וכן לתצוגת תכלית של בניית מחשבי־על המבוססים של הפלטפורמה החדשה של Nvidia, Spectrum-X.[94]

1.6.2. קישוריות

במרכז המהפכה הגדולה של האינטרנט עומדת הקישוריות – בין שמדובר בקישורים בין פיסות מידע בהיפרטקסט ובין שמדובר בהשלכות התפוצה הנרחבת של הרשת על התקשורת בין בני אדם. ממד נוסף של קישוריות נוגע ליכולת של מכשירים לתקשר אלה עם אלה ועם סביבתם. טכנולוגיה זו, המכונה "האינטרנט של הדברים" (Internet of Things, או IoT), היא ארכיטקטורה המאפשרת חפצים "חכמים", המצוידים בחיישנים שאוספים, מעבדים ומעבירים נתונים כדי ליעל את פעולתם.[95] לארכיטקטורה זו קשת רחבה של יישומים במגוון תחומים – תכולת המקרר, רשת החשמל (המסגירה את מנהגי השינה של בני הבית), מכשירי הטלוויזיה החכמים (המעבירים בדיוק מרבי מידע על הרגלי צפייה) ואפילו המכוניות האוטונומיות.[96] החיישנים (מדי תאוצה, חיישנים גירוסקופיים או חיישני תנועה) יכולים למדוד פרמטרים מגוונים – אור וצבע, לחץ ברומטרי, טמפרטורה ולחות, שדות מגנטיים. על אלה אפשר להוסיף גם את מקלט הנווטן

94 רפאל קאהאן "אנבידיה חשפה מחשב על ישראלי - אחד מעשרת החזקים ביותר בעולם" **ynet** (29.5.2023).

95 ראו לדוגמה רועי צזנה, **השולטים בעתיד** 56–64 (2017).

96 על מכוניות אוטונומיות ראו **להלן** בסעיף 2.1. ראו גם גדי פרל "הבטיחו לנו רכבים אוטונומיים. איפה הם?" **הארץ** (22.2.2022).

(GPS), המצוי בכל מכשיר טלפון חכם, המאפשר לכל מכשיר קצה (וכן ליחידות עיבוד מרכזיות) לקבל מידע בנוגע למקומו במרחב, בכל מקום על פני כדור הארץ.

האינטרנט של הדברים מאפשר איסוף מידע רחב היקף מחיישנים המצויים בכל מקום. ככל שתפוצתם של חיישנים אלו תמשיך לגדול, כך יגדל גם נפח המידע הפוטנציאלי שמערכות בינה מלאכותית יוכלו לעבד אותו ולהתאמן עליו.

1.6.3. בינה מלאכותית מבוססת ענן

איסוף מידע רב ואיכותי, ניתוחו ופיתוח מודלים – משימות אלו גוזלות זמן ניכר ומשאבים רבים. ככל שעלות אחסון הנתונים בענן פוחתת, כך עוברות משימות כמו ניתוח מידע באמצעות אלגוריתמים ולמידת מכונה לענן. בשוק הולכות ומתרבות פלטפורמות של מערכות לומדות המאפשרות לחברות להעביר אליהן מידע ולקבל תוצאות בלי לפתח בעצמן מודלים חישוביים.[97]

כך למשל פיתחה חברת איי־בי־אם עבור מפתחים חיצוניים את מערכת הבינה המלאכותית ווטסון, לשימוש דרך הענן.[98] משנת 2015 מספקת גוגל למפתחים את Tensorflow, מנוע בינה מלאכותית שנסמך על רשתות נוירונים מבוססות מעבדי GPU,[99] ובגרסאות מתקדמות על מאיצי בינה מלאכותית – מעגלים משולבים ייחודיים שפיתחה גוגל עבור המערכת.[100] בשנת 2018 החלה גוגל להציע למפתחים להשתמש במאיצים אלו, המכונים TPU (Tensor Processing Units), דרך הענן.[101] חברות מסחריות משתמשות ב־Tensorflow ובמנועי

97 Jeffery Burt, *Learning to Make the Machine Part of AI Invisible and Easy*, The Next Platform (11.12.2020)

98 Michael del Castillo, *It's 'Entrepreneurs Gone Wild' at IBM's World of Watson Conference*, New York Business Journal (5.5.2015)

99 Cade Metz, *Google Just Open Sourced TensorFlow, Its Artificial Intelligence Engine*, WIRED (9.11.2015)

100 Norm Jouppi, *Google Supercharges Machine Learning Tasks with TPU Custom Chip,* Google Cloud Platform Blog (18.6.2016)

101 John Barrus and Zak Stone, *Cloud TPU Machine Learning Accelerators Now Available in Beta*, Google Cloud Platform Blog (12.2.2018)

בינה מלאכותית אחרים בענן לצרכים מגוונים, כגון פענוח סריקות MRI או זיהוי התמונות האפקטיביות ביותר שמעלים משתמשים לפלטפורמות שיתוף כגון Airbnb.[102]

בדרך זו עומדת לרשות כל מפתח אבן בניין טכנולוגית לפיתוח תוכנות ואפליקציות חדשות גם בלי שיידרש למומחיות בארכיטקטורה שביסודה, ובעלות נמוכה (בייחוד כשמביאים בחשבון את עלויות הפיתוח העצומות של בינה מלאכותית).

בשלהי 2020 המליצה ועדה שמינה פורום תל״ם (תשתיות לאומיות למחקר ופיתוח) על הקמת ענן בינה מלאכותית ציבורי בישראל.[103] בקיץ 2022 החלה בישראל הפעלת פרויקט נימבוס, שתכליתו לספק לזרועות הממשלה שירותי ענן באמצעות חברות גוגל ואמזון.

1.6.4. נתוני ענק (BIG DATA)

נתוני עתק הוא מונח שנטבע ב־1997 כדי לתאר הדמיית נתונים ואת האתגרים שהיא מעמידה למערכות מחשב.[104] אין הגדרה יחידה לנתוני עתק. יש הרואים בהם מסד נתונים גדול שקשה לנתח או לדמות באמצעים טכנולוגיים מקובלים.[105] אחרים רואים בהם מאגרי מידע בעלי נפח עצום (high volume), המתעדכנים במהירות גבוהה (high velocity), או מאגרים בעלי

102 Shijing Yao, Qiang Zhu, and Phillippe Siclait, *Categorizing Listing Photos at Airbnb*, The Airbnb Blog (3.5.2018); Jason A. Polzin, *Intelligent Scanning Using Deep Learning For MRI*, TensorFlow Blog (1.3.2019)

103 ועדת בינה מלאכותית ומדע נתונים, **דוח מסכם** (דצמבר 2020), בעמ' 37–38, 83 (להלן: דוח ועדת תל"ם). ראו גם אורי ברקוביץ' "התוכנית הלאומית ל־AI יוצאת לדרך: הקמת ענן במיליארד שקל לחוקרים ולסטארט־אפים" **גלובס** (22.12.2020).

104 Michael Cox and David Ellsworth, *Application-Controlled Demand Paging for Out-of-Core Visualization*, Proceedings of the 8th IEEE Conference on Visualization (1997)

105 Frank J. Ohlhorst, Big Data Analytics: Turning Big Data into Big Money (2012)

מגוון רחב של סוגי נתונים (high variety).[106] גישות אחרות מתעלמות מהנפח של נתוני עתק ומנסות לתת להם הגדרה מהותית – לפי פרמטרים של הקשר, קישוריות וסיבוכיות.[107] חרף ריבוי ההגדרות,[108] אנו נאפיין נתוני עתק לפי נפח, מהירות ומגוון.

כפי שכבר ראינו, בינה מלאכותית – ובפרט כזאת המבוססת על טכניקות של למידת מכונה – זקוקה לנתונים. נתוני עתק הם הדלק המאפשר את פיתוחה ואת שכלולה. כך למשל, עד שנת 2017 אספה חברת טסלה תיעוד מוסרט של יותר מ־1.2 מיליארד קילומטרים של נהיגה.[109] סרטי הווידאו האלה, נוסף על מגוון סוגי נתונים שמקורם בחיישנים אחרים המותקנים ברכב, משמשים אותה לאמן את המערכת לעבור נתיבים ולהימנע מתאונות. נתונים אלו ישמשו את המערכת ללימוד ויהיו הבסיס לרכב האוטונומי של טסלה. חברת גוגל, שאינה משווקת כלי רכב ללקוחות ואין ברשותה נתוני נהיגה שוטפים, הצליחה להשיג באותה עת נתונים רק על כ־2.4 מיליון קילומטרים של נהיגה. היתרון של טסלה ברור ולכן מעריכים שהיא תקדים את גוגל בפיתוח רכב אוטונומי.[110]

גם ספקי שירותים מקוונים אחרים יכולים להשתמש בנתונים שלהם כדי לאמן את מערכות הבינה המלאכותית שלהם. חשבו על מיליארדי המשתמשים בשירותים של חברות כגון מטא, אפל וגוגל, שעקב היקף הנתונים שהן מייצרות מקנה להן פוטנציאל עצום לאימון מכונות לומדות. לא בכדי קבע מנכ"ל חברת איי־בי־אם, ארבינד קרישנה, כי "הסיבה להתקדמות המואצת של בינה מלאכותית

106 Mark Beyer and Douglas Laney, *The Importance of "Big Data" A Definition*, Gartner Research (21.6.2012)

107 Vincent Charles and Tatiana Gherman, *Achieving Competitive Advantage Through Big Data: Strategic Implications*, 16 Middle-East Journal of Scientific Research 1069 (2013)

108 ראו למשל את הדיון אצל Tobias M. Scholz, Big Data in Organizations and the Role of Human Resource Management 12-20 (2017)

109 Michael J. Coren, *Tesla Has 780 Million Miles of Driving Data, and Adds Another Million Every 10 Hours*, QUARTZ (28.5.2016)

110 Howard Yu, *Google Fumbles While Tesla Sprints Toward a Driverless Future*, Forbes (14.12.2016)

היא שאנחנו מייצרים מדי יום ביומו 2.5 קווינטיליון בייטים של נתונים. זה 2.5 ואחריו 18 אפסים [...] לא די בבסיסי נתונים ובטכניקות אנליטיקה מיושנים. בינה מלאכותית היא הכלי היחיד שיכול לקצור את הנתונים האלו ולרתום אותם להפקת תובנות".[111]

אפשר לומר בפרפרזה על המשפט הידוע, "איזהו עשיר? בעל הדאטה". מי שמרבה לאסוף נתונים ושולט בהם מחזיק בנכס יקר ערך, שיש שמכנים אותו "הנפט של העולם הדיגיטלי".[112] עם זאת, לצד התפתחות מאגרי המידע שנדרשו לצורך למידה מפוקחת, התחזקות יכולות הלמידה העמוקה הלא מפוקחת מאפשרת למודלים להתאמן על כלל המידע ברשת האינטרנט. ככל שהמודלים מתפתחים, כך קטן הצורך שלהם בכמויות מידע עצומות ומתחלף בלמידה מבוססת התנסות (experience-based learning) – התנסות בעולם האמיתי בזמן אינטראקציה עם משתמשים ועם המשימות שהם מתבקשים למלא. באוגוסט 2022 פתחה חברת מטא לסוף שבוע אחד לשימוש חופשי בארצות הברית את הצ'אטבוט החדש שלה, BlenderBot 3. המשתמשים ששוחחו עם המערכת היו אמורים לספק לה משוב לשיפור מודל הבינה המלאכותית שלה.

111 Stephanie Condon, *IBM Promises a 4,000 Qubit Quantum Computer By 2025: Here's What it Means*, ZDNet (10.5.2022)

112 Michael Palmer, *Data Is the New Oil*, ANA Marketing Maestros (3.11.2006). ראו בהקשר זה גם את דבריה של שושנה זובוף (Shoshana Zuboff, the Age of Surveillance Capitalism 153 (2019); התרגום שלנו):

> דמיינו שיש לכם פטיש. זו למידת מכונה. בעזרת הפטיש טיפסתם על הר גבוה והגעתם לפסגה. זו הדומיננטיות של למידת מכונה בניתוח נתונים. על פסגת ההר מצאתם ערמה עצומה של מסמרים, זולה יותר מכל מה שאפשר להעלות על הדעת. זו טכנולוגיית החיישנים החכמים החדשה. כעת מונח לפניכם לוח חלק ועצום [...] ואז אתם לומדים שבכל פעם שאתם נועצים בלוח מסמר בעזרת פטיש למידת המכונה, אתם מפיקים ערך. זו המוניטיזציה של נתונים. מה אתם עושים? אתם מתחילים לנעוץ מסמרים ולעולם לא תפסיקו, אלא אם כן מישהו יגרום לכם להפסיק. אבל אין כאן אף אחד שיגרום לכם לעצור. לכן "האינטרנט של כל דבר" הוא בלתי נמנע.

1.6.5. קונסולידציה של טכנולוגיות

בינה מלאכותית במובן הצר שלה – למידת מכונה מפוקחת, שמתבססת על מאגר – נחלקה באופן מסורתי לכמה סוגים: ראייה חישובית (computational vision), עיבוד שפה טבעית (natural language processing) וזיהוי דיבור (speech recognition). חלוקה זו אפשרה למשל קיום של מערכת שיודעת לכתוב סיפור, אבל אינה יודעת לזהות תמונה של חתול. בשנתיים האחרונות התלכדו סוגי הבינה המלאכותית ונוצרו מכונות שעושות הכול. למשל, מחוללי תמונות שמבינים טקסטים ויודעים להפיק מהם תמונות או מכונות שיודעות להמיר בעיות מילוליות בתרגיל מתמטי שאפשר לפתור במחשבון. מגמה זו של פיתוח מודלים מאוחדים המשלבים יכולות בכמה תחומים (כגון וידאו, תמונות, שפה וקול), הן בארכיטקטורת החומרה הן בתוכנה, מאפשרת את ההתקדמות לעבר בינה מלאכותית כללית יותר: צ׳אטבוטים שיכולים להתכתב עם המשתמשים, לדבר איתם בעל־פה ואפילו ״לראות״ אותם ולהגיב לתנועות הפנים שלהם; רובוטים בעלי יכולת לוגית, כושר ראייה, הבנת שפה, כושר דיבור ומוטוריקה.

זאת ועוד. רוב מערכות הבינה המלאכותית פועלות היום בעולם הווירטואלי. החל במנוע החיפוש של גוגל וכלה באלגוריתמיקה של הרשתות החברתיות, בשירותי הניווט, בעוזרים האישיים הדיגיטליים, במערכות המשמשות לחיזוי ולהמלצה ובמערכות הזיהוי. אבל יש לשער שבשנים הבאות יעברו מערכות הבינה המלאכותית מן העולם הווירטואלי לעולם המעשי. דוגמאות מוכרות הן רכב אוטונומי ומכשור רפואי. אבל יש דוגמה מסקרנת ומסעירה יותר: הבינה המלאכותית עתידה לחלחל מהעולם הווירטואלי לעולם האמיתי באמצעות רובוטים הומנואידים (humanoids), שיהיו להם יכולות לוגיות ויכולות ראייה, הבנת שפה ודיבור; שיִלמדו בצורה לא מפוקחת באמצעות חיזוקים ולא יזדקקו לכמויות גדולות של נתונים לצורך אימון; ושתוכנות כגון מנועי משחק יוכלו לבחון אותם באמצעות סימולציות שידמו את העולם האמיתי ואחרי כן ישתמשו בטכניקות כגון sim2real כדי לבחון בעולם האמיתי רובוט שנבדק בסימולציה.

1.6.6 התפתחות של בינה מלאכותית באמצעות מערכות קוד יוצרות (גנרטיביות) בשימוש הכלל

הכוחות שיוסיפו להניע את מהפכת הבינה המלאכותית הם הפונקציונליות של מערכות בינה מלאכותית שאינן תלויות ברמת המיומנות הטכנית של המשתמש וזמינותן של מערכות המסוגלות ליצור קוד ואלגוריתמים, כלומר חיזוק. היכולת של משתמשים שאינם מוצאים יישום שהולם את צורכיהם ליצור מוצר משלהם. תחילה יוצרו מוצרים ששירתו את מהנדסי התוכנה, כגון Tabnine ו־DeepCode, שבדומה לנעשה בשפה טבעית חזו מה תהיה שורת הקוד הבאה או ידעו לאתר באגים בקוד שנכתב. היום הולך וגדל מספר הפלטפורמות שאינן מחייבות ידע טכני בתכנות; כמעט כל אחד יכול ליצור, לבדוק וליישם פתרונות המבוססים על בינה מלאכותית באמצעות ממשקים פשוטים של גרירה ושחרור או ממשקי תכנות יישומים (application programming interface [API]).[113] אפשר להביא לדוגמה את SwayAI, המשמשת לפיתוח יישומי בינה מלאכותית ארגוניים, ואת Akkio, שיכולה ליצור כלי חיזוי וקבלת החלטות. מערכת CoPilot מבית OpenAI ו־GitHub יוצרת כ־40% מן הקוד בחברות שמשתמשות בה ואפשר שבעתיד היא תייתר את הצורך לדעת לכתוב קוד ודי יהיה ללמוד כיצד לנסח "לחישה" מתאימה (prompt) ולומר למערכת מהו המוצר המבוקש. אפשרויות אלו יעניקו לעסקים ולארגונים את היכולת להתגבר על המחסור במדעני נתונים ומהנדסי תוכנה, כך שהפקת התוצרים של מוצרים מבוססי בינה מלאכותית תהיה פשוטה, זולה ובהישג ידם של רבים.

1.6.7 סיכום

ככל שעלות הרכישה, האחסון והעיבוד של נתונים תלך ותפחת, כך תלך ותשתכלל כלכלת האלגוריתמים ותציע עוד ועוד אבני בניין של שירותי בינה מלאכותית בענן. בעקבות המחשוב המתעצם והיכולת ליצור מוצרים מבוססי בינה מלאכותית באמצעות "לחישה" למערכת הבינה היוצרת, מתחזקות התחזיות שלפיהן בעשורים הבאים ישתלבו רובוטים ובינה מלאכותית כמעט בכל חלק בחיינו. משימושים שונים של מערכות בינה מלאכותית שנסקור בפרק הבא עולה כי לא מדובר עוד בעתידנות, שכן תחזיות אלו קרובות מאוד להתממש.

113 ממשק תכנות יישומים (API), הוא ערכה של פונקציות ופרוצדורות קוד מן המוכן המאפשרת יצירה פשוטה של יישומים שמתממשקים ליישום קונקרטי.

פרק שני

שימושים שונים
של בינה מלאכותית

בינה מלאכותית אינה מוגבלת לתחומים צרים כגון משחקי שחמט, גו או שש־בש. היא אינה מוגבלת גם ליישומים תאורטיים במעבדות הפיתוח כגון הבחנה בסיסית בין כלבים לחתולים. לעיתים נדמה שלא חולף שבוע בלי שכותרות העיתונים מבשרות על יכולת כזאת או אחרת של מערכות בינה מלאכותית – החל בכתיבת ספרים ומאמרים[114] ובהפקת דימויים גרפיים או יצירות קומיקס

114 יובל פלוטקין "קראתם פעם ספר שנכתב בעזרת בינה מלאכותית?" **ynet** (7.12.2020); John K. Waters, *AI21 Labs Emerges from Stealth with the First "AI-based Writing Companion*," Pure AI (28.10.2020)

מתיאורים מילוליים[115] (לעיתים ברמה גבוהה מזו של אמנים בשר ודם);[116] עבור בזיהוי רגשות של כלבים[117] (או של בני אדם, חרף הספקות באשר לבשלות המדעית של טכנולוגיה זו),[118] בניחוש עמדותיה של השופטת העליונה המנוחה רותביידר גינזבורג בכל נושא שבעולם[119] ובהמלצות על העברת שחקנים לקבוצות כדורגל;[120] וכלה בפיתוחים יישומיים יותר, כגון ניהול מרכזים לוגיסטיים,[121] זיהוי צורת חלבונים,[122] קבלת החלטות בתהליכי הפריה מלאכותית[123] או זיהוי גלקסיות והבחנה ביניהן.[124] תקצר היריעה מלסקור את כל השימושים האפשריים בבינה מלאכותית, ועל כן נתמקד בפרק זה בכמה תחומים מובילים המדגימים את הפוטנציאל היישומי הטמון במערכות נבונות.

115 אושרי אלקסלסי, "מפחיד ומרשים בו זמנית: מודל ה־AI של OpenAI יוצר עכשיו תמונות מהמילים שלכם" **Geektime** (6.1.2021); מור, **לעיל** ה"ש 72; Rich Johnston, *Abolition Of Man, First Comic Book Entirely Drawn By A.I. Algorithm*, Bleeding Cool (14.6.2022)

116 Kevin Roose, *An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy*, The New York Times (2.9.2022)

117 "זו העצם שלך, או שאתה שמח לראות אותי? קולר מיוחד מזהה רגשות אצל כלבים", **הארץ** (14.1.2021).

118 "*Immature Biometric Technologies Could Be Discriminating Against People" Says ICO in Warning to Organisations*, ICO (26.10.2022); Ifeoma Ajunwa, *Automated Video Interviewing as the New Phrenology*, 36 Berkeley Technology Law Journal 1173 (2021)

119 *Israeli Company AI21 Labs Creates AI Model of Ruth Bader Ginsburg*, The Jerusalem Post (18.6.2022)

120 "טכנולוגיה ישראלית מתאימה שחקנים לקבוצות כדורגל באמצעות בינה מלאכותית" **New-Tech Magazine** (16.8.2022).

121 נטע־לי בינשטוק "אמזון, מאחורייך: צה"ל מחליף את האפסנאים בבינה מלאכותית" **כלכליסט** (14.4.2021).

122 קייד מץ "צעד שיאיץ פיתוח תרופות: חוקרים פתרו בעיה בת 50 שנה עם בינה מלאכותית" **הארץ** (1.12.2020).

123 אושרי אלקסלסי " 'במקום ילדי מבחנה, ילדי AI': בזכות האלגוריתם של Embryonics הישראלית, הושגו כבר 6 הריונות מוצלחים" **Geektime** (21.1.2021).

124 זאב בליזובסקי "בינה מלאכותית זיהתה 80,000 גלקסיות ספירליות" **הידען** (27.8.2020).

2.1 בינה מלאכותית במערכות תחבורה

כלי רכב אוטונומיים הם מיישומי הבינה המלאכותית המוכרים ביותר לציבור. המונח הלטיני למכונית, אוטומוביל, מבטא את יכולת ההנעה העצמית של הרכב, להבדיל מכרכרה המונעת מכוח סוסים. למכוניות ולאוטוסטרדות נקשרה בתרבות הילה של חירות ושל שחרור האדם באמצעים טכנולוגיים, אך הסכנה לתאונות לא נעלמה.[125] כלי הרכב האוטונומיים – הניחנים לכאורה ביכולת לנוע במרחב באופן עצמוני, ללא התערבות אנושית – אמורים בין השאר לצמצם עד מאוד את הסיכונים הכרוכים בנהיגה אנושית.[126]

את ההבחנה המקובלת בין דרגות שונות של אוטונומיה של כלי רכב פיתח ארגון מהנדסי הרכב (SAE).[127] על פי הבחנה זו (ראו לוח 1), יש שש דרגות של אוטונומיה. בשלהי 2015 העריך אלון מאסק, מנכ"ל חברת טסלה, שבתוך שנתיים יפותחו כלי רכב ברמת אוטומציה מלאה.[128] חמש שנים לאחר מכן המשיך מאסק להשמיע תחזיות בדבר הגעתן של מכוניות אוטונומיות לחלוטין לכביש בעגלא ובזמן קריב.[129] עם זאת, רמת הבשלות של טכנולוגיות הנהיגה הגיעה עד 2020 לרמת אוטונומיה 3 (נהיגה עצמית מותנית). חרף בשלותן של מערכות אלו, כמה מיצרניות הרכב מהססות לשווק אותן.[130]

125 Henri Lefebvre, Everyday Life in the Modern World 98–109 (1971)

126 לסקירה עדכנית על סיכונים אלו ראו מתנאל בן אבי, **מגמות בבטיחות בדרכים בישראל 2019-2013** (חטיבת מידע ומחקר, הרשות הלאומית לבטיחות בדרכים 2020).

127 *Automated Driving: Levels of Driving Automation are Defined in New SAE International Standard J3016*, Society of Automotive Engineers International (18.6.2018). יוער כי עד 2016 השתמש משרד התחבורה האמריקאי בטקסונומיה משלו, אך דומה לזו של SAE. לטקסונומיה זו ראו *NHTSA Lays Out Ground Rules for Autonomous Vehicles*, The Car Connection (7.6.2013)

128 Fred Lambert, *Tesla CEO Elon Musk Drops His Prediction of Full Autonomous Driving from 3 Years to Just 2*, Electrek (21.12.2015)

129 "מאסק מבטיח: מכונית אוטונומית ב־25 אלף דולר תוך 3 שנים" **כלכליסט** (23.9.2020); "אלון מאסק: בעוד שנה טסלה תוציא מיליון רכבים אוטונומיים" **גלובס** (23.4.2019).

130 תומר הדר "דילמת רמה 3: מדוע חברות הפאר הגרמניות לא ממהרות לפתח רכב אוטונומי?" **כלכליסט** (13.11.2020).

לוח 1
רמות אוטונומיה של כלי רכב אוטונומיים

רמה	שם	הגדרה
רמה 0	אין אוטומציה	נהג מיומן מבצע את כל משימות הנהיגה, אך מסתייע במערכות התרעה.
רמה 1	סיוע לנהג	מערכות אוטומטיות לסיוע לנהג מתקנות פעולות האצה והאטה או היגוי במצב נהיגה נתון[131] בהסתמך על מידע על סביבת הנהיגה ובהנחה שנהג מיומן מבצע את כל שאר משימות הנהיגה.
רמה 2	נהיגה עצמית חלקית	מערכות אוטומטיות לסיוע לנהג מבצעות פעולות האצה, האטה והיגוי במצב נהיגה נתון בהסתמך על מידע על סביבת הנהיגה ובהנחה שנהג מיומן מבצע את כל שאר משימות הנהיגה.
רמה 3	נהיגה עצמית מותנית	מערכות אוטומטיות מבצעות את כל משימות הנהיגה במצב נהיגה נתון בהנחה שנהג מיומן יגיב אם יתעורר צורך להתערב בנהיגה (כלומר יוכל לחזור לשליטה מלאה ברכב בתוך פרק זמן סביר מקבלת ההתרעה).
רמה 4	נהיגה עצמית ברמה גבוהה	מערכות אוטומטיות מבצעות את כל משימות הנהיגה במצב נהיגה נתון גם אם נהג אנושי אינו יכול להגיב תגובה סבירה לבקשת המערכת שישלוט שליטה מלאה ברכב.
רמה 5	אוטומציה מלאה	מערכות אוטומטיות מבצעות את כל משימות הנהיגה בכל תנאי הסביבה שנהג אנושי יכול להתמודד איתם.

מה שמבחין בין השיח בן זמננו על כלי רכב אוטונומיים ובין פנטזיות שהוצגו בסרטי מדע בדיוני[132] הוא יכולת המימוש. המימוש מתאפשר בזכות ארכיטקטורה

131 מצב נהיגה נתון הוא תרחיש שיש לו משימות נהיגה אופייניות – נהיגה בכביש הפתוח, השתרכות בפקק וכו'.

132 החל במכוניות אוטונומיות שמופיעות כבדרך אגב בסרטי הז'אנר, כגון **זיכרון גורלי** (פול ורהובן במאי 1990) או **דוח מיוחד** (סטיבן שפילברג במאי 2002), ועד לכאלו המשמשות דמויות מרכזיות, כגון בסדרה **אביר על גלגלים** (גלן לרסון יוצר 1982–1986).

המשלבת כמה טכנולוגיות – הנעה מכנית קלסית, חיישנים, קישוריות ובינה מלאכותית.[133] מצלמות, טכנולוגיית לידאר (LiDAR, light detection and ranging – זיהוי אור וטווח), מכ"מים, מערכות אינפרה־אדום, נווטנים (מערכות GPS) וכלים אחרים – כל אלה מאפשרים למערכת האוטונומית לאסוף באופן שוטף מידע רב על סביבתה המיידית ולהעריך אילו אובייקטים נמצאים בקרבתה ואילו סכנות עלולות להתממש.

ואולם לא מדובר בהכרח ביחידה אוטונומית אחת ויחידה (ונוסף על כך, רכב אוטונומי אינו בהכרח "מכונית").[134] ארכיטקטורות שונות של קישוריות – רכב לרכב (V2V); רכב לתשתיות (V2I), דוגמת סנסורים קרקעיים בצידי הדרך; רכב לענן (V2C); ואף רכב להולכי רגל (V2P) – כל אלו מניחים את התשתית הרעיונות ל"אינטרנט של הרכבים" (IoV), שבו תאפשר הקישוריות זרימת מידע בין כל החיישנים וכלי הרכב במרחב. בדרך זו יהיה אפשר למזער סיכונים ולטייב את איכות התחבורה אף יותר.[135]

בינה מלאכותית היא רכיב מרכזי של כל טכנולוגיה של נהיגה אוטונומית. בין שמדובר במימוש של אוטונומיה באמצעות מערכות מומחה שמסתמכות על כללי נהיגה קבועים מראש ובין שמדובר בטכניקות של למידה עמוקה[136] – היכולת המעשית של רכב להתמצא במרחב בהתבסס על קלט שמתקבל מחיישנים מותנית בפיתוח של מערכות נבונות.[137]

במצב שבו חלק ניכר מהתנועה תעבור לשליטה של מערכות נבונות, כלומר כששיעור הרכבים האוטונומיים מכלל הרכבים בשוק יהיה ניכר, יהיה אפשר

133 ראו למשל Hannah YeeFen Lim, Autonomous Vehicles and the Law 7–15 (2018)

134 ראו יובל מן "זהירות, ספינה אוטונומית חוצה את האטלנטי" **ynet** (17.6.2021).

135 Zaigham Mahmood, *Connected Vehicles in the IoV: Concepts, Technologies and Architectures*, *in* Connected Vehicles in the Internet of Things: Concepts, Technologies and Frameworks For the IoV 3–18 (Zaigham Mahmood ed., 2020)

136 ראו Coren, **לעיל** ה"ש 109.

137 ראו לדוגמה Alex Kendall et al., *Learning to Drive in a Day*, *in* International Conference on Robotics and Automation (ICRA) 8248 (2019)

לצפות לעלייה ברמת הבטיחות בדרכים, לעלייה בקיבולת הכבישים[138] ולהפחתה בפליטת מזהמים.[139] מקצת התרחישים צופים מעבר לשימוש ברכבים אוטונומיים במערכות היסעים ציבוריות וויתור על רכבים פרטיים.[140]

עם זאת, רכבים אוטונומיים אינם היישום היחיד של בינה מלאכותית במערכות תחבורה. מערכות נבונות יכולות לייעל את התנועה באמצעות הסדרת פעולת הרמזורים[141] או ביזור מערכות הניווט. מערכות הניווט עצמן מכילות רכיבים של בינה מלאכותית ויכולות להציע לכל נהג נתיב זהה או לנסות לווסת את כלל התנועה של המשתמשים בהן.[142] מערכות נבונות יכולות גם לנהל יישומים תחבורתיים של כלכלה שיתופית, כגון מיזמי שיתוף נסיעות (car pooling/ ride sharing)[143] או שירותי היסעים שיתופיים (כגון השירותים שמציעות חברות Uber ו־Lyft או Bubble ו־GetTaxi).

138 Reshaping Urban Mobility with Autonomous Vehicles: Lessons from the City of Boston (World Economic Forum in collaboration with the Boston Consulting Group, 2018)

139 Jeffery B. Greenblatt and Samveg Saxena, *Autonomous Taxis Could Greatly Reduce Greenhouse-Gas Emissions of US Light-Duty Vehicles,* 5 Nature Climate Change, 860 (2015)

140 Alex Davies, *China's Self-Driving Bus Shows Autonomous Tech's Real Potential*, Wired (7.10.2015)

141 Cathy Wu et al., *Framework for Control and Deep Reinforcement Learning in Traffic*, IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) 1-8 (2017); Abu Salman Shaikat et al., *An Image Processing and Artificial Intelligence Based Traffic Signal Control System of Dhaka*, Asia Pacific Conference on Research in Industrial and Systems Engineering (APCoRISE) 1-6 (2019)

142 ראו לדוגמה Shoshana Vasserman, Michal Feldman, and Avinatan Hassidim, *Implementing the Wisdom of Waze*, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (Ijcai 15) (2015)

143 אודי עציון "מוביט ו־ווייז ישתפו פעולה כדי לעקוף את הפקקים" **כלכליסט** (30.10.2019); להשלכות על הזכות לפרטיות ראו Xusen Cheng et al., *The Good, the Bad, and the Ugly: Impact of Analytics and Artificial Intelligence-Enabled Personal Information Collection on Privacy and Participation in Ridesharing* 31 European Journal of Information Systems 339 (2021)

2.2 זיהוי פנים

בפרק הקודם תיארנו מערכת שמשתמשת בטכניקות של למידה עמוקה כדי להבחין, למשל, בין תמונות שיש בהן כלבים לתמונות שאין בהן כלבים. באותו אופן פותחו גם אלגוריתמים לזיהוי פנים. זיהוי פנים הוא תהליך אוטומטי שמשווה בין שתי תמונות שונות של פנים כדי לבחון אם מדובר באותו אדם.

לאלגוריתמים של זיהוי פנים, המבוססים על בינה מלאכותית, יש שלל יישומים.[144] מערכות לזיהוי פנים מאפשרות למיין ולמפתח את שלל התמונות המצולמות דרך קבע במכשירי הטלפון החכמים של כולנו, לתייג חברים ברשתות החברתיות, לאמת זהות של אדם באמצעות פרמטרים ביומטריים ואף להצפין או לנעול מכשירי קצה ולשלוט בגישה לחשבונות מקוונים.[145] מובן שיש הבדל בין

144 United States Government Accountability Office, Facial Recognition Technology: Privacy and Accuracy Issues Related to Commercial Uses, 11-13 GAO-20-522 (13.7.2020); Sharon Nakar and Dov Greenbaum, *Now You See Me. Now You Still Do: Facial Recognition Technology And The Growing Lack Of Privacy* 23 J. Sci. & Tech. 889, 97-100 (2017); Ian Berle, Face Recognition Technology: Compulsory Visibility and Its Impact on Privacy and the Confidentiality of Personal Identifiable Images 17-23 (2020); איתן לשם " 'משמר הלילה' של סין: מערכת לזיהוי פנים תמנע מקטינים לשחק כשאסור" **הארץ** (11.7.2021); לדוגמאות של שימוש בטכנולוגיית זיהוי פנים בישראל ראו מיכל רז־חיימוביץ "למנוע שימוש של קטינים: אפליקציות הקורקינטים של ווינד תפעיל טכנולוגיה לזיהוי פנים" **גלובס** (14.10.2020); ניצן שפיר " 'שימוש בטכנולוגיית זיהוי פנים לבעלי תו ירוק בהבימה ובאצטדיון בלומפילד פוגע בפרטיות ואינו חוקי' " **גלובס** (20.5.2021). יוער כי כותביו של דוח של מרכז המחקר והמידע של הכנסת מציינים כי למיטב ידיעתם אין גורם בגופי המטה הממשלתיים שהם פנו אליו שציין כי נציגיו משתמשים היום בטכנולוגיות זיהוי פנים. רועי גולדשמידט **השימוש בטכנולוגיות זיהוי וניטור במרחב הציבורי** 8 (מרכז המחקר והמידע של הכנסת 2020). עם זאת, הן הצבא הן המשטרה סירבו לבקשה של האגודה לזכויות אדם לפי חוק חופש המידע לקבל מידע על השימוש שלהם בטכנולוגיות זיהוי פנים. ראו "שימוש בטכנולוגיות לזיהוי פנים על ידי המשטרה והצבא", **האגודה לזכויות האזרח בישראל** (2.5.2021).

145 כבר בשנת 2008 הציגה חברת לנובו מחשבים ניידים שאפשרו להחליף את השימוש בסיסמה בזיהוי פנים של משתמשים מורשים. Kelly A. Gates, Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance 125 (2011)

שימוש בטכנולוגיות זיהוי פנים למטרות פרטיות, כמו אשרור גישה למכשיר קצה פרטי, ובין פרישׂה של מערך מצלמות לזיהוי פנים במרחב הציבורי.[146]

השימוש בטכנולוגיות לזיהוי פנים לביקורת גבולות[147] ולאמצעי תשלום מקוונים הולך וגובר. גם סוכנויות ביטחון ואכיפת חוק מביעות עניין גובר בטכנולוגיות לזיהוי פנים,[148] שכן הן מאפשרות להן לזהות במרחב מרושת במצלמות איומים ועבריינים,[149] ואף מרתיעות בני אדם (כמעין פנאופטיקון ממשטר) מביצוע פשעים.[150] לכן יש לתת את הדעת על האופן שבו שימוש בטכנולוגיות אלו עלול לפגוע בזכויות אדם,[151] ואין פלא שבתקנות הבינה המלאכותית האירופיות ניתנה תשומת לב מיוחדת לטכנולוגיות אלו (בהקשר הרחב של מערכות ביומטריות).[152]

146 Evan Selinger and Brenda Leong, *The Ethics of Facial Recognition Technology*, *in* The Oxford Handbook of Digital Ethics 101–122 (Carissa Véliz ed., 2021)

147 Berle, **לעיל** ה"ש 144, בעמ' 17–19.

148 **שם**, בעמ' 19–20; Clare Garvie, Alvaro M. Bedoya, and Jonathan Frankle, *The Perpetual Line-Up – Unregulated Police Face-Recognition in America*, Center on Privacy and Technology at Georgetown Law (2016); Katelyn Ringrose, *Law Enforcement's Pairing of Facial Recognition Technology with Body-Worn Cameras Escalates Privacy Concerns* 105 Va. L. Rev. Online 57 (2019)

149 Kyriakos N. Kotsoglou and Marion Oswald, *The Long Arm of the Algorithm? Automated Facial Recognition as Evidence and Trigger for Police Intervention* 2 Forensic Science International: Synergy 86–89 (2020)

150 ראו לדוגמה את הנחיית רשם מאגרי המידע מס' 4/2012, "שימוש במצלמות אבטחה ומעקב ובמאגרי התמונות הנקלטות בהן", פס' 2.2–2.5.

151 ראו לדוגמה את הסקירה הזאת: In Focus: Facial Recognition Tech Stories and Rights Harms from Around the World (International Network of Civil Liberties Organizations (INCLO) 2021). כמו כן ראו את הדיווחים מן הזמן האחרון על כוונת המשטר האיראני לאכוף את החוק המחייב נשים לעטות חיג'אב באמצעות טכנולוגיה לזיהוי פנים: "בעקבות 'ההתרופפות' בצניעות של נשים, איראן פונה לטכנולוגיות זיהוי פנים" **הארץ** (7.9.2022).

152 ס' 3(33) להצעת תקנות הבינה המלאכותית האירופיות (ראו **לעיל** ה"ש 53) מגדיר מידע ביומטרי מידע שמקורו בתהליכי עיבוד טכניים הנוגעים בין השאר לזהותו הייחודית של אדם, לדוגמה מראהו (בהצעה המתוקנת, ס' 3(33) מפנה ישירות לס' 4(14) ל־GDPR, שממנו נלקחה הגדרה זו). ס' 5(1)(d) להצעה המתוקנת אוסר על שימוש במערכות זיהוי

אלגוריתמים המשמשים לזיהוי מבוססים על מערך כללים או על תהליך לחישוב או ניתוח מאפיינים של פנים אנושיות. איכותם של אלגוריתמים כאלו תלויה באופיו של בסיס הנתונים שהם מתאמנים עליו; למשל, מאגר התמונות ששימש לפיתוחם יקבע מה יצליחו לזהות טוב יותר – פנים של בני אדם ממוצא אסיאתי או פנים מערביות.[153] לאחרונה ציין הממונה על היישומים הביומטריים כי יש קשיים בהרכשה (acquisition) של תמונות פנים של תושבים בעלי גון עור כהה; בהתאם לכך, היעדר כיול ראוי של מערכות ביומטריות עלול להשפיע על הדיוק ועל מרווח הטעות בעת שימוש בהן.[154] לנוכח מגוון היישומים הפוטנציאליים של טכנולוגיה זו יש להטיות אלו, גם אם נעשו בתום לב, השלכות הרות גורל והן עלולות להנציח פערים חברתיים ואפליה של מיעוטים.[155] למשל, אלגוריתם זיהוי פנים שהתאמן על מאגר תמונות של אנשים ממוצא מערבי ומשמש לזיהוי חשודים בפשע עלול להגיע לתוצאות מדויקות פחות בזיהוי חשודים מרקע אתני אחר.

ואכן, יישומים של מערכות זיהוי פנים למטרת אכיפת חוק או ביטחון לאומי שנויים במחלוקת – לא רק בשל הפגיעה בפרטיות עקב פרישת מצלמות מעקב

ביומטריות בזמן אמת במרחבים ציבוריים. ס' 52(2) מטיל על מערכות אלו חובות שקיפות ייחודיות. כמו כן, מערכות זיהוי ביומטריות מוגדרות כמערכות בסיכון גבוה (ס' 1 בתוספת III לטיוטה). ההצעה המתוקנת מגדירה כמערכות בסיכון גבוה גם מערכות מבוססות ביומטריקה (המסיקות בהתבסס על נתונים ביומטריים מאפיינים אישיים של אדם, ובכלל זה מערכות לזיהוי רגשות).

153 כך למשל, מחקר מצא שרמת הדיוק של אלגוריתמים לזיהוי פנים שפותחו במדינות מזרח אסיה לזיהוי תווי פנים אסיאתיים גבוהה יותר מרמת הדיוק של אלגוריתמים שפותחו במערב, ולהפך. P. Jonathon Phillips et al., *An Other-Race Effect For Face Recognition Algorithms*, ACM Transactions on Applied Perception (2009)

154 דוח הממונה על היישומים הביומטריים מס' 11 (21.6.2021), בעמ' 6. חששות דומים הועלו בעקבות אישור חוק המצלמות המיוחדות בוועדת השרים לענייני חקיקה (טיוטת חוק לתיקון פקודת המשטרה (מערכות צילום מיוחדות), התשפ"ב–2022.לנוסח מיום 10.4.2022 שאושר בוועדת השרים בחודש מאי 2022). ראו לדוגמה רוני (פנטנש) מלכאי "מצלמות לזיהוי פנים יסכנו את יוצאי אתיופיה" **הארץ** (11.5.2022).

155 Fabio Bacchini and Ludovica Lorusso, *Race, Again: How Face Recognition Technology Reinforces Racial Discrimination* 17 Journal of Information, Communication & Ethics in Society 321 (2019); Selinger and Leong, **לעיל** ה"ש 146.

במרחב הציבורי, אלא גם בשל העיבוד האוטומטי של מידע ביומטרי רגיש,[156] שמעורר קשיים בדיוק ובזיהוי,[157] והגברת עוצמתן של תופעות המקשות על זיהוי אנושי רגיל של חשודים במסגרת מסדרי זיהוי.[158]

בערים רבות בעולם המערבי הוגבל במידה ניכרת השימוש במערכות כאלה,[159] והפרלמנט האירופי קרא לחרם על מערכות זיהוי פנים למטרות שיטור.[160] חברות הטכנולוגיה הגדולות, ובראשן אמזון, איי־בי־אם ומייקרוסופט, הודיעו שיפסיקו לייצר מערכות זיהוי פנים לתכליות שיטור ולהשקיע בחברות המפתחות אותן, או שיימנעו מלספקן לסוכנויות אכיפת החוק.[161] בשלהי 2021 הודיעה מטא (פייסבוק)

156 ראו לדוגמה, R (Bridges) v. Chief Constable of South Wales Police [2020] EWCA (Civ) 1058, at para. 88–89

157 לפי דיווח ברשת סקיי הבריטית, שיעור הטעויות בזיהוי של המערכת לזיהוי פנים שהייתה בשימוש משטרת המטרופולין של לונדון הגיע ל־81%. Rowland Manthorpe and Alexander J. Martin, *81% of "Suspects" Flagged By Met's Police Facial Recognition Technology Innocent, Independent Report Says*, SKY News (4.7.2019); מחקר אחר הצביע על שיעור טעויות נמוך יותר, אך עדיין גבוה – 38% ו־63% בשתי בדיקות שונות: Pete Fussey and Daragh Murray, Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology 69–72 (Human Rights Centre, University of Essex 2019)

158 Laura Moy, *Facing Injustice: How Face Recognition Technology May Increase the Incidence of Misidentifications and Wrongful Convictions*, 30 Wm. & Mary Bill Rts. J. 337, 350 (2021)

159 ראו הסקירה של תהילה שוורץ אלטשולר ועמיר כהנא **חוות דעת בעניין תזכיר חוק לתיקון פקודת המשטרה [נוסח חדש] (מערכות צילום מיוחדות), התשפ"א-2021** 12–15 (חוות דעת, המכון הישראלי לדמוקרטיה 29.7.2021).

160 European Parliament Resolution of 6 October 2021 on Artificial Intelligence in Criminal Law and its Use by the Police and Judicial Authorities in Criminal Matters, para. 27 (2020/2016(INI))

161 Alex Hern, *IBM Quits Facial-Recognition Market over Police Racial-Profiling Concerns*, The Guardian (9.6.2020); Jay Greene, *Microsoft Won't Sell Police its Facial-Recognition Technology, Following Similar Moves by Amazon and IBM*, The Washington Post (11.6.2020); Jeffrey Dastin, *Amazon Extends Moratorium on Police Use of Facial Recognition Software*, Reuters (18.5.2021)

על הפסקת פעולתן של מערכות זיהוי הפנים שלה.[162] מנגד, בישראל קודם לאחרונה תזכיר חוק לאסדרת מערכת עין הנץ, מערכת משטרתית לניטור לוחיות רישוי, והונחה התשתית המשפטית להחילו בעתיד גם על מערכות לזיהוי פנים.[163]

פרישׂה רחבת היקף של מערכות זיהוי פנים תרחיב את השימוש בקטגוריות חדשות של מידע. ייתכן שמערכות נבונות המנתחות נתונים חזותיים יכולות לייצר באופן אוטומטי מידע על רגשות של פרטים הנקלטים בעדשות המצלמות[164] או על מאפייני התנהגותם במרחב. מערכות כאלו יכולות לנתח את טיב האינטראקציה החברתית של יחידים וקבוצות ואת אופי הקשרים ביניהם. יכולות אלו לא היו קיימות בעבר ושילובן בפרקטיקות של פיקוח ומשטור במרחב הציבורי עלול להעמיק את אופי פרופיל האישיות שניתן לבנות באופן אוטומטי וחסר הבחנה לאזרחים החולפים בעל כורחם על פני עדשת המצלמה, וממילא להגדיל את מידת הפגיעה בפרטיות.

יש לתת את הדעת שמערכות זיהוי פנים ניחנות ביכולות שחורגות מזיהוי פנים גרידא. מערכות צילום נבונות מסוגלות להפיק מן הנתונים הוויזואליים מידע על מאפייני תנועה, על מגדר, על שיוך אתני[165] או על רגשות (חרף הספקות בעניין

162 Kashmir Hill and Ryan Mac, *Facebook, Citing Societal Concerns, Plans to Shut Down Facial Recognition System*, The New York Times (2.11.2021); Jerome Pesenti, *An Update on Our Use of Face Recognition*, Facebook (2.11.2021)

163 בעת כתיבת שורות אלו נדונה שאלת ההסמכה של המשטרה להשתמש במערכת "עין הנץ" בבג"ץ 641/21 **האגודה לזכויות אזרח נ׳ משטרת ישראל**. לעיון בעתירת האגודה לזכויות אזרח ראו "להפסיק להשתמש במערכת שעוקבת אחרי נהגים" **האגודה לזכויות אזרח** (2.11.2022). לתזכיר החוק ראו תזכיר חוק לתיקון פקודת המשטרה [נוסח חדש] (תיקון מס׳)(מערכות צילום מיוחדות), התשפ"א–2021 (8.7.2021). כן ראו שוורץ אלטשולר וכהנא, **לעיל** ה"ש 159.

164 לאחרונה, למשל, דווח כי חברת המצלמות קאנון התקינה במשרדיה שבבייג'ין מערכת זיהוי פנים המתירה כניסה רק לעובדים "שמחים". James Vincent, *Canon Put AI Cameras in its Chinese Offices that only Let Smiling Workers Inside*, The Verge (17.6.2021)

165 אושרית גן־אל "ניו יורק טיימס: רשויות בסין משתמשות בזיהוי פנים לניטור מיעוטים" **גלובס** (15.4.2019); עודד ירון "'ילא יכול להסביר את זה'; בכיר בוואווי התפטר בעקבות פיתוח 'אזעקת אויגורים'" **הארץ** (16.12.2020).

בשלותה של טכנולוגיה זו).[166] למשל, אמן בלגי אימן אלגוריתם לזהות בזמן אמת מתי דעתם של פוליטיקאים בלגים מוסחת מדיונים פרלמנטריים והם עסוקים במכשיר הטלפון הנייד שלהם. אותה מערכת אף פרסמה את הסרטון ברשתות החברתיות ותייגה אוטומטית את הפוליטיקאים הסוררים.[167] מערכות זיהוי פנים אינן מוגבלות כמובן לביוש נבחרי ציבור וניתן לרתום יכולות דומות לניתוח מידע בהקשר רחב יותר – למשל לזיהוי מאפייני תנועה של התקהלויות חשודות או של אירועים אלימים.

2.3
בינה מלאכותית במערכות אכיפת החוק והצדק

2.3.1 לעורך דין מְתֻכְנָן שלשלתי אסימון: בינה מלאכותית בליגל־טק

עריכת דין היא פרופסיה שמרנית[168] הנרתעת מאימוץ חידושים טכנולוגיים.[169] על כן חדירתן של טכנולוגיות בינה מלאכותית למגזר המשפטי – ובייחוד הפרטי – נעשית עקב בצד אגודל. ואולם פוטנציאל ההשפעה של הטכנולוגיות האלה רב. לצד השמרנות המקצועית, חסם כניסה נוסף שמקשה על חדירת הבינה המלאכותית לעולם המשפט הוא מרכזיות השפה בו.

166 **לעיל** ה"ש 118.

167 *The Flemish Scrollers, 2021-2022*, Dries Depoorter (2021)

168 להתייחסות כללית לפרופסיה המשפטית כאל פרופסיה שמרנית ראו יניב רוזנאי וקארין פאר פרידמן "עריכת דין מהפכנית" **המשפט** כ 303, 310–316 (2015); גל סידס, שמוליק בכר, אופיר נוה "החינוך המשפטי: על מה שבפנים ועל מה שבחוץ" **עיוני משפט** כה 243 (2001); Edgar Bodenheimer, *The Inherent Conservatism of the Legal Profession*, 23 Ind. L. J. 221 (1948)

169 Chay Brooks, Cristian Gherhes, and Tim Vorley, *Artificial Intelligence in the Legal Sector: Pressures and Challenges of Transformation*, 13 Cambridge Journal of Regions, Economy and Society 135, 146 (2020)

אף שהמילה קוד פירושה בין השאר אסופת חוקים, קוד משפטי אינו קוד מכונה.[170] השפה המשפטית המרכיבה את החוק – הקוד המשפטי – יכולה להיות עמומה ורב־משמעית.[171] גם בהיעדר עמימות, עיבוד מסמכים משפטיים דורש פיתוח טכניקות לעיבוד שפה – היכולת להבחין בין מונח משפטי לשפה טבעית או היכולת לזהות תבניות משפטיות והקשרים.[172]

מערכות בינה מלאכותית מסוגלות כיום למיין ולסווג מאות מסמכים לצורך הליכי גילוי (discovery) במהלך משפט, או הליכי גילוי נאות (due diligence) בעסקאות ענק בין חברות – משימות שבעבר היו מנת חלקם של צוותים גדולים של עורכי דין.[173] החברה הישראלית LawGeex, למשל, פיתחה מערכת שמסוגלת לזהות בהסכם סודיות היבטים שונים שיש לתת עליהם את הדעת מבחינה משפטית. בחודש פברואר 2018 ערכה החברה ניסוי שבחן את ביצועי המערכת בהשוואה לביצועים של 20 עורכי דין אנושיים. המשתתפים התבקשו לזהות היבטים משפטיים בחמישה הסמכי סודיות. האלגוריתם הגיע לממוצע של 94% הצלחה בתוך 26 שניות; עורכי הדין, לעומת זאת, הגיעו לממוצע של 85% הצלחה, אבל היו זקוקים לשעה וחצי בממוצע כדי לסקור את כל חמשת החוזים.[174]

170 על היחס בין קוד לחוק ראו למשל Lawrence Lessig, Code: 2.0 (2006); Primavera De Filippi and Samer Hassan, *Blockchain Technology as a Regulatory Technology: From Code Is Law to Law Is Code* (2018), available at arXiv

171 ראו לדוגמה Diana Raffman, *Vagueness in Law: Placing the Blame where It's Due*, *in* Vagueness in the Law – Philosophical and Legal Perspectives 49 (Geert Keil and Ralf Poscher eds., 2016); H. L. A. Hart, The Concept of Law 124-136 (2nd edition, 1994); Brian Bix, Law, Language, and Legal Determinacy (1993)

172 ראו **להלן** בסעיף 2.5.

173 John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, The New York Times (4.3.2011); Nicholas Barry, *Man Versus Machine Review: The Showdown between Hordes of Discovery Lawyers and a Computer-Utilizing Predictive-Coding Technology*, 15 Vand. J. Ent. & Tech. L. 343 (2013); Samuel Sahagian, *iLawyer*, 3 UCLA J. L. & Tech. Digest 8 (2021)

174 *Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts*, Lawgeex (2018)

לכאורה אפשר לומר שתוצאות אלו מבשרות את "קץ הפרופסיה",[175] אך אפשר לומר גם שיש בהן כדי להראות שלבינה מלאכותית יש פוטנציאל להיות טכנולוגיה משבשת ולשנות סדרי עולם בתחום המשפט. האוטומציה של הסכמים פשוטים (כמו הסכמי סודיות) מקילה על עורכי הדין את העבודה ומאפשרת להם להתמחות בסוגיות מורכבות יותר ובשירותים אישיים לפי צורכי הלקוח. אפשר שהשימוש במערכות ממוחשבות לעיבוד חוזים אף יוביל לניסוחם בשפה פשוטה יותר ולצמצום הנטייה המקצועית לנקוט לשון מעורפלת ומסובכת כדי לחפות על בעיות מהותיות בחוזה. לעת עתה, גם בשל הכללים החלים על שירותים משפטיים,[176] נדרש אישורו של עורך דין אנושי לתוצאה שהפיק האלגוריתם.

לצד עריכת חוזים, היבט אחר של העבודה המשפטית הוא פרשנות הדין, ובכלל זה הפסיקה. לעיתים מזומנות, חוות הדעת של עורכי הדין על חוקיות מצב דברים עובדתי מסוים אינן אלא תחזית בדבר האופן שבו יפסוק בית המשפט בהינתן אותן נסיבות. תחזית זו נסמכת על פרשנות החקיקה ועל פרשנות ההלכה המשפטית הרלוונטית, ולעיתים אף מסתייעת בהערכה של עמדותיהם האפשריות של שופטים מסוימים. את חוות הדעת המשפטית הלא־ממוכנת, המבוססת על ניסיונו המקצועי של עורך הדין המנסח אותה, ניתן להחליף או למצער להשלים בתחזית ממוכנת. אחד היישומים של מערכות נבונות הוא חיזוי, ומערכת בינה מלאכותית שתדע לעבד מסמכים משפטיים תוכל לנתח מאות או אלפי תיקים דומים, לפנות לחקיקה המתאימה ולנבא את התשובה המשפטית הצפויה בשאלה מסוימת.

מערכות כאלו אינן פרי הדמיון. חברת ההזנק הקנדית Blue J Legal, למשל, משתמשת בלמידת מכונה לחיזוי החלטות מיסוי.[177] מחקרים אחרים מציעים

175 ראו למשל Richard Susskind, The End of Lawyers? Rethinking the Nature of Legal Services (2010)

176 ס' 20(3)–(4) לחוק לשכת עורכי הדין, תשכ"א–1961, ס"ח 347 בעמ' 178.

177 Benjamin Alarie, Anthony Niblett, and Albert H. Yoon, *Using Machine Learning to Predict Outcomes in Tax Law*, 58 Canadian Business L. J. 231 (2016); idem, *How Artificial Intelligence Will Affect the Practice of Law*, 68 University of Toronto L. J. 106 (2018). ראו למשל כיצד אלגוריתם מבוסס למידת מכונה מעריך כיצד יחליט בית המשפט למיסים לסווג הוצאות משפטיות על קניין רוחני: Benjamin Alarie and Kimberly Condon, *Deducting Legal Expenses: Unpacking the IRS's Appeal in Mylan*, Tax Notes (26.8.2022)

מודלים לחיזוי החלטות של טריבונלים בעניין סכסוכים בין משכירים לשוכרים,[178] ואף החלטות של בית המשפט העליון של ארצות הברית[179] או של בית הדין האירופי לזכויות אדם (ECHR).[180]

ואולם כשמערכות בינה מלאכותית מתמודדות עם חיזוי של החלטות שיפוטיות אין הן נדרשות להסתפק בנתונים משפטיים גרידא (נסיבות המקרה, תקדימים ולשון החוק), אלא יכולות להשתמש בכל מידע לבר־משפטי שאפשר להוסיף למאגר נתוני האימון. מחקרים אקדמיים מראים כי תוצאות משפטיות יכולות להיות מושפעות מגורמים שאינם רלוונטיים לתיק הנדון. שעת הדיון וסמיכותה לארוחת הצוהריים למשל – עלולה לכאורה להשפיע על החלטת השופט.[181] גם מוצאם של המתדיינים או השופט עלול להשפיע על ההחלטה בתיק.[182]

178 Hannes Westermann et al., *Using Factors to Predict and Analyze Landlord-Tenant Decisions to Increase Access to Justice*, *in* Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19) 133 (2019)

179 Daniel Martin Katz, Michael J. Bommarito II, and Josh Blackman, *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLoS ONE (12.4.2017). יוער שהניסיונות למדל סטטיסטית את החלטות בית המשפט העליון של ארצות הברית אינם חדשים, אך לא כולם התבססו על למידת מכונה.

180 Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras, *Neural Legal Judgment Prediction in English*, Proceedings of the 57th Annual Meeting of the Association For Computational Linguistics 4317 (2019)

181 במחקר של שי דנציגר, יונתן לבב וליאורה אבנעים־פסו נמצא מתאם בין מועד הדיון בוועדת השחרורים להחלטה – ככל שהזמן ממועד ארוחת הצוהריים לדיון היה ארוך יותר, כך פחת הסיכוי לשחרור מוקדם. Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso, *Extraneous Factors in Judicial Decisions*, 108 Proceedings of the National Academy of Sciences 6889 (2011). ואולם בחינה חוזרת של המחקר תלתה את המתאם הסטטיסטי בהיבטים פרוצדורליים אחרים: אסירים שיוצגו על ידי עורך דין (ועל כן ההיתכנות לשחרורם גבוהה יותר) נדונו בוועדה לפני האסירים הלא מיוצגים. Keren Weinshall-Margel and John Shapard, *Overlooked Factors in the Analysis of Parole Decision*s 108 Proceedings of the National Academy of Sciences E833 (2011)). ראו גם סמדר רייספלד "מחוץ לחוק: מה משפיע על החלטות של שופטים?" **הארץ** (2.12.2015). לביקורת נוספת על השימוש במחקר "השופטים הרעבים" כהצדקה להחלפת שופטים בשר ודם במערכות נבונות ראו Konstantin Chatziathanasiou, *Beware the Lure of Narratives: "Hungry Judges" Should not Motivate the Use of "Artificial Intelligence,"* 23 German L. J. 452 (2022)

182 ראו לדוגמה אורן גזל־אייל ואח' "ערבים ויהודים בהליכי הארכת מעצר ראשוני" **משפטים** לח 627 (2009).

ביולי 2019 נחקק בצרפת חוק האוסר להשתמש בנתוני הזיהוי של שופטים לצורכי ניתוח משפטי.[183] הנימוק: שימוש כזה, המתיימר לחזות מראש הכרעה משפטית של כל שופט ושופט, עלול להביא להפעלת לחץ על שופטים או לנקיטת טקטיקות של ברירת סמכות מקומית (forum shopping).[184] מבקרי האיסור גורסים כי מדובר במהלך שנועד להגן על השופטים מאחריותיות (accountability) מוגברת.[185] אפשר גם שהמהלך הזה נועד לשמר מראית עין של שפיטה עקבית, שאינה מושפעת מ"גודל כף רגלו של הצ'נסלור".

מבלי לקבוע מסמרות בסוגיה נראה אפוא כי רתימתה של הבינה המלאכותית למטרות של חיזוי תוצאות משפטיות או לבחינה אקדמית אמפירית היא אכן טכנולוגיה משבשת שמצריכה טיפול עדין – הן כדי למנוע הנצחה של הטיות קיימות המצויות בנתונים (וכן של פרשנות משפטית, שעלולה לקפוא על שמריה),[186] הן כדי שהיא עצמה תתקבל בגופים מוסדיים קיימים, המנסים לשמר את הסטטוס קוו.

2.3.2. שופט אוטומטי שלל לי רישיון: בינה מלאכותית ככלי עזר בשיפוט ובאכיפת החוק

יש כמה מופעים אפשריים של הממשק בין בינה מלאכותית לעולמות התוכן המשפטיים. כפי שראינו לעיל, כבר היום אלגוריתמים שמבוססים על בינה מלאכותית יכולים לחזות החלטות שיפוטיות במגוון תחומים ובדיוק גבוה יחסית, ועורכי דין יכולים להיעזר בהם כדי לעצב את האסטרטגיה המתאימה ביותר ללקוחות שהם מייצגים. מערכות חיזוי של החלטות שיפוטיות יכולות לתרום גם לעבודת השיפוט עצמה, בהציען בסיס

183 LOI n° 2019-222 du 23 mars 2019 de programmation 2018-2022 et de réforme pour la justice (1) - Article 33

184 France, Conseil Constitutionnel, Loi de programmation 2018-2022 et de réforme pour la justice, Décision n° 2019-778 DC

185 Jena McGill and Amy Salyzyn, *Judging by Numbers: How Will Judicial Analytics Impact the Justice System and Its Stakeholders?* 44 Dal. L. J. 249 (2021)

186 Daniel Maggen, *Predict and Suspect: The Emergence of Artificial Legal Meaning*, 23 N. C. J. L. & Tech 67 (2021)

אמפירי לרפלקסיה על תהליכי קבלת ההחלטות של השופטים. מערכות בינה מלאכותית יכולות לשמש מערכות תומכות החלטה לשופטים ולסייע להם להגיע באופן יעיל יותר לתובנות מדויקות, עקביות וטובות יותר.

המערכת השיפוטית אינה חפה מטעויות אנוש, והשאיפה לנטרל אותה ככל האפשר מהגורם האנושי אינה חדשה. מערכות חיזוי תומכות החלטה במערכת המשפט תורמות לאובייקטיביות של ההחלטה,[187] אך יש הטוענים כי לנוכח קיומן של הטיות אלגוריתמיות[188] אין כאן אלא מראית עין של אובייקטיביות מדעית.[189] תומס ג׳פרסון, מאבות האומה האמריקאית, קבע שיש ״להניח לשופט להיות מכונה גרידא״.[190] ייתכן שבשנים האחרונות מתחיל חזונו של ג׳פרסון לקרום עור וגידים, שכן מדינות שונות משלבות מערכות בינה מלאכותית בעבודת רשויות השיפוט שלהן. אסטוניה הכריזה על כוונתה להשתמש בבינה מלאכותית כשופט הדן בתביעות קטנות (שסכומן אינו עולה על 7,000 אירו), בכפוף לזכות לערער על ההחלטה הממוכנת לפני שופט אנושי.[191] גם סרביה שוקלת להשתמש בטכנולוגיות דומות.[192] בבריטניה העלו כמה איגודים מקצועיים והאקדמיה הצעות להפחתת העומס השיפוטי באמצעות מערכות בינה מלאכותית.[193]

187 ראו למשל Jennifer L. Skeem and Christopher Lowenkamp, *Risk, Race, & Recidivism: Predictive Bias and Disparate Impact*, 54 Criminology 680 (2015); Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp, *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. and it's Biased against Blacks,"* 80 Fed. Probation J. 38 (2016)

188 ראו **להלן** בפרק 6.

189 ראו למשל John Logan Koepke and David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 Wash. L. Rev. 1725 (2018)

190 Thomas Jefferson, *Thomas Jefferson to Edmund Pendleton* (26.8.1776)

191 Eric Niler, *Can AI Be a Fair Judge in Court? Estonia Thinks So*, Wired (25.3.2019)

192 Milica Stojanovic, *Serbia Eyes Artificial Intelligence in Courts, but Experts See Dangers*, BalkanInsight (25.1.2021)

193 Yadong Cui, Artificial Intelligence and Judicial Modernization 23 (2020)

בכל הנוגע להטמעת טכנולוגיות בינה מלאכותית במערכת המשפט, סין נמצאת בחזית. מערכת לזיהוי תיקים דומים היא רכיב מרכזי בפרויקט "בית המשפט החכם" של סין. מערכת זו מתבססת על אלגוריתמים המוצאים דפוסים בתיקי בית המשפט (שלא כמו מערכות המנסות לחלץ אוטומטית את ההלכה שנפסקה).[194]

מערכת המשפט הסינית משתמשת בעוד מערכת תומכת, "מערכת 206". מערכת זו מסייעת לשופטים ולבאי כוחם של הצדדים בהצגת הראיות הרלוונטיות במשפט. היא מסוגלת לעבד ולהציג באופן ברור ומקיף את מארג הראיות הנוגעות לסוגיה מסוימת ולאתר פערים ואי־התאמות בין ראיות שונות. היא מסוגלת גם לתמלל שפה מדוברת לטקסט, ובכך מייתרת את הצורך בקלדנים אנושיים.[195]

בשנת 2019 החל פיילוט של יישום רדיקלי יותר של בינה מלאכותית במערכת המשפט. ההכרעה המשפטית נעשית באמצעות אלגוריתם, אך בהשגחת שופט אנושי.[196] עדיין מוקדם להעריך אם השיפוט האלגוריתמי הזה יוסיף לשמש בסין כלי תומך לשופטים, או שבעתיד יוסר הפיקוח האנושי ויוקנו לו סמכויות עצמאיות. ראוי לציין שבכל פעם שהשופט דוחה החלטה של המערכת הוא נדרש לנמק בכתב מדוע,[197] תנאי שעלול להשפיע על העצמאות השיפוטית ולמצער

194 Li Xiaohui, *Research on the Building of China's Smart Court in the Internet Era*, 8 China Legal Sci. 30, 44 (2020); George G. Zheng, *China's Grand Design of People's Smart Courts*, 7 Asian J. L. & Soc. 561 (2020); Jinting Deng*, Should the Common Law System Welcome Artificial Intelligence? A Case Study of China's Same-Type Case Reference System* 3 Geo. L. Tech. Rev. 223 (2019)

195 **שם**. ראו גם Cui, **לעיל** ה"ש 193, בפרק 7–8.

196 Nyu Wang and Michael Yuan Tian, *"Intelligent Justice": AI Implementations in China's Legal AI Transformation*, *in* Artificial Intelligence and Its Discontents: Critiques from the Social Sciences and Humanities 197, 212 (Ariane Hanemaayer, ed., 2022); Bin Wei et al., *A Full-Process Intelligent Trial System for Smart Court*, 23 Frontiers of Information Technology & Electronic Engineering 186 (2022)

197 Stephen Chen, *China's Court AI Reaches Every Corner of Justice System, Advising Judges and Streamlining Punishment*, South China Morning Post (13.7.2022)

לשנות את מרכז הכובד של ההחלטות המשפטיות.[198] בשנת 2012 פרסמו החוקרים ברדרן ומקנטייר מודל לחיזוי מסוכנות של עצירים לפני הליכים. החוקרים העריכו שיישום המודל היה מייתר רבע מהמעצרים לפני הליכים ושיעור הפשיעה (ובפרט הפשיעה האלימה) לפני משפט היה פוחת.[199] מחקר מאוחר יותר בחן החלטות של שופטים בעיר ניו יורק בעניין שחרור בערבות ופיתח על סמך ממצאיו כלל שחרור אלגוריתמי מקביל. נמצא כי אילו השתמשו באלגוריתם היה אפשר להפחית את רמת הפשיעה בעיר או את מספר המעצרים בשיעור ניכר.[200]

ואולם גם בארצות הברית החלטות שיפוטיות הנתמכות באלגוריתמים אינן נחלתן של תאוריות אקדמיות. מערכת הצדק האמריקאית משתמשת במערכות אוטומטיות להערכת מסוכנות בכל אחד משלביה – שיטור, מתן גזר דין, כליאה ושחרור ממאסר.[201]

סרטו של שפילברג משנת 2002, "דוח מיוחד",[202] המבוסס על סיפור קצר של פיליפ ק' דיק,[203] מתאר מחלקה במשטרה שתפקידה למנוע פשעים עתידיים בטרם יתרחשו בהסתמך על דיווח של מוטנטים הניחנים ביכולת חיזוי על־טבעית. רשויות אכיפת החוק ברחבי העולם ישמחו לזכות ביכולת דומה לחזות פשיעה

198 מנגד, הגישה הסינית רואה באוטומציה כלי להתמודדות עם שחיתות ולהבניית שיקול דעת אחיד והרמוני. על פי גישה זו כל החלטה שיפוטית שסוטה מההחלטה האלגוריתמית חייבת בהנמקה מטעמים של צדק פרוצדורלי והוגנות.

199 Shima Baradaran and Frank L. McIntyre, *Predicting Violence,* 90 Tex. L. Rev. 497, 500 (2012)

200 מאחר שיש מתאם בין רמת הפשיעה בעיר לשיעור העצורים שלא שוחררו בערבות חישבו החוקרים בנפרד תוצאות של שני תרחישים שונים. כשהוקפאה רמת הפשיעה בעיר (ביחס לנתוני האימון) הפחית האלגוריתם את מספר העצורים ב־42%; כשהוקפא שיעור העצורים (ביחס לנתוני האימון) הביא האלגוריתם לירידה של כ־25% ברמת הפשיעה. ראו Jon Kleinberg et al., *Human Decisions and Machine Prediction*, 133 The Quarterly Journal of Economics 237 (2017)

201 Ruha Benjamin, Race After Technology 63 (2019)

202 **לעיל** ה"ש 132.

203 Philip K. Dick, the Minority Report (1956)

באמצעות בינה מלאכותית. לצד הצורך להבחין בין נבואה לחיזוי,[204] או השאלה העולה בסרט בדבר הנורמטיביות של ההנחות הדטרמניסטיות הגלומות בקביעה האקטוארית שפלוני עתיד לעבור עבירה,[205] מערכות הבינה המלאכותית של ימינו מעלות שאלות נוספות באשר לרמת הדיוק של קביעות כאלו ולאפשרות שהן מכילות הטיות אלגוריתמיות.

בבואם לגזור את דינו של נאשם נעזרים מקצת בתי המשפט האמריקאיים בדוחות מ"מערכת לניהול פרופילים לסנקציות חלופיות עבור קציני מבחן" (COMPAS, Correctional Offender Management Profiling for Alternative Sanction). מערכת זו מבצעת, בין היתר, הערכת מסוכנות של הנאשם, כלומר בוחנת את סיכוייו לחזור לסורו (רצידיביזם לפני המשפט, רצידיביזם כללי ורצידיביזם אלים).[206] בשנת 2016 פורסמה באתר ProPublica סדרת כתבות שהעלו חשש שמערכת COMPAS סובלת מהטיות גזעניות[207] לאחר שנמצאו הבדלים בין התפלגות הערכת הסיכונים שהפיקה המערכת עבור נאשמים שחורים להתפלגות שהפיקה עבור נאשמים לבנים. באותה שנה, בעניין Loomis, דחה בית המשפט העליון של ויסקונסין את הטענה שעצם השימוש בהערכת מסוכנות אלגוריתמית מפרה כללים של הליך נאות, וקבע שההחלטה השיפוטית יכולה

204 Dr. Lance B. Eliot, *Quandary of Precrime Detection Getting Nearer Via Advances in AI* (28.12.2021)

205 ראו למשל Peter Marks, Imagining Surveillance 101–103 (2015); Jackson Polansky and Henry Fradella, *Does "Precrime" Mesh with the Ideals of U.S. Justice? Implications For the Future of Predictive Policing* 15 Cardozo Pub. L. Pol'y & Ethics J. 253 (2016-2017). לעיסוק בשאלת האוטונומיה של הפרט בדין הישראלי "לכתוב את סיפור חייו" ראו ע"א 10064/02 **מגדל חברה לביטוח נ׳ אבו־חנא**, פ"ד ס(3) 13 (2005). ראו גם אליעזר ריבלין וגיא שני "תפיסה עשירה של עקרון השבת המצב לקדמותו בתורת הפיצויים הנזיקיים – המקרה של שימוש בסטטיסטיקה בפסיקת פיצוי לנפגע קטין" **משפט ועסקים** י 499 (2009).

206 State v. Loomis, 881 N.W.2d 749, 754 (Wis. 2016)

207 Julia Angwin et al., *Machine Bias: There's Software Used across the Country to Predict Future Criminals. and It's Biased against Blacks*, ProPublica (23.5.2016)

להשתמש במערכת COMPAS כמקור מידע נוסף – ובתנאי ששאר הגורמים שהובאו בחשבון יירשמו.[208]

אימון בינה מלאכותית בנתונים היסטוריים שמגולמת בהם (במשתנים חברתיים־כלכליים למשל) הטיה מערכתית לחובתה של קבוצת מיעוט עלולה אפוא לייצר מערכת חיזוי בעל אפקט רגרסיבי המנציח פערים.[209] יש הטוענים שההפך הוא הנכון, ושאפשר להסתייע בבינה מלאכותית כדי לזהות מופעים של אפליה.[210]

מבחינות רבות אין שוני רב בין מערכות לחיזוי החלטות שיפוטיות ובין מערכות תומכות החלטה שיפוטית: אלו ואלו מנסות להגיע לתוצאה המשפטית הצפויה. עם זאת, מערכות לחיזוי החלטות שיפוטיות הן חיצוניות למערכת המשפט והניתוח שלהן שואף להגיע לתוצאה האמפירית הרצויה (מה יחליט בית המשפט), ואילו מערכות תומכות החלטה שיפוטית מספקות תוצאה נורמטיבית (מה צריך בית המשפט להחליט). משכך, מערכות אלו נדרשות לעמוד בסטנדרטים נוקשים יותר, הכפופים למגבלות הדין הפרוצדורלי ולכללי הצדק הטבעי.

הדין הצרפתי אוסר על מערכות תומכות החלטה במערכת המשפט. לפי חוק הגנת הפרטיות של צרפת, "החלטה משפטית הכרוכה בהערכת התנהגותו של אדם לא תתבסס על עיבוד אוטומטי של נתונים שנועד להעריך היבטים מסוימים באישיותו של אותו אדם".[211] ואולם אפשר לפרש איסור זה ולומר שאין הוא אוסר

208 State v. Loomis, **לעיל** ה"ש 206. בית המשפט העליון של ארצות הברית דחה בקשת רשות ערעור של הנאשם Loomis v. Wisconsin, 137 S. Ct. 2290 (2017) (denying cert.). להרחבה בעניין לומיס ראו John Villasenor and Virginia Foggo, *Artificial Intelligence, Due Process and Criminal Sentencing*, 2020 Mich. St. L. Rev. 295 (2020); Katherine Freeman, *Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis*, 18 N. C. J. L. & Tech. 75 (2016)

209 בהקשר זה ראו למשל Sandra G. Mayson, *Bias In, Bias Out,* 128 Yale. L. J. 2122 (2019)

210 ראו למשל Jon Kleinberg et al., *Algorithms as Discrimination Detectors*, 117 Proceedings of the National Academy of Sciences (28.7.2020); בהקשר של הפרסום ב־ProPublica על ההטיות במערכת COMPAS (**לעיל** ה"ש 208) ראו Flores, Bechtel, and Lowenkamp, **לעיל** ה"ש 187.

211 Loi 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés [Law 78-17 of January 6 1978 on Information Technology,

שימוש במערכות תומכות החלטה אלא אוסר להסתמך עליהן בלבד בהחלטות משפטיות, בדומה לעניין Loomis.[212]

מערכות נבונות נמצאות בשירות מנגנוני אכיפת החוק והביטחון גם בשלבים הקודמים להליך המשפטי – מניעת פשע וחיזוי טרור. האתגר של חיזוי הטרור – יותר מהאתגר של מניעת הפשע – מצריך יכולות איסוף מידע רב־ממדיות,[213] כלי ניתוח ועיבוד חזקים, ובכלל זה בינה מלאכותית, ומודלים תאורטיים החוזים רדיקליזציה ומסוכנות בהסתמך על פרופילים ברשתות חברתיות.[214] שירותי הביטחון כבר משתמשים בכלים אלו כדי לחזות פעילות טרור.[215]

גם גופי אכיפת החוק ברחבי העולם מאמצים טכנולוגיות דומות. את הדימוי של המפה העירונית המעטרת את קירות תחנות המשטרה ועליה מסומנים

Data Files and Civil Liberties] Journal officiel de la République Française [J.O] [official Gazette of France], Jan. 6, 1978, p. 7, Art. 47

212 ראו Florence G'sell, *AI Judges*, *in* The Cambridge Handbook of Artificial Intelligence, Global Perspectives on Law and Ethics (Larry A. Dimatteo, Cristina Poncibo, and Michal Cannarsa eds., 2022)

213 בהקשר של איסוף, ריכוז ועיבוד מידע פנים ממשלתי ראו 115 BverGE 320 (2006); וכן עמיר כהנא ויובל שני **רגולציה של מעקב מקוון בדין הישראלי ובדין המשווה** 191–193, 261–267 (מחקר מדיניות 123, המכון הישראלי לדמוקרטיה 2019).

214 ראו לדוגמה Michael Wolfowicz et al., *Faces of Radicalism – Differentiating between Violent and Non-Violent Radicals by Their Social Media Profiles* 116 Computers in Human Behavior (2021); Abhishek Sachan and Devshri Roy, *TGPM: Terrorist Group Prediction Model For Counter Terrorism*, 44 International Journal of Computer Applications 49 (2012); Robert Pelzer, *Policing of Terrorism Using Data from Social Media*, 3 Eur. J. Secur. Res. 163 (2018)

215 ראו למשל ג'ון בראון " 'ידו"ח מיוחד': על סמך מה עוצרת ישראל מאות פלסטינים על פשעים שלא ביצעו?" **הארץ** (19.04.2017); גילי כהן "ראש השב"כ: איתרנו יותר מ־2,000 מפגעים בודדים פוטנציאליים מתחילת 2016" **הארץ** (27.06.2017); אור הירשאוגה והגר שיזף "סיכול ממוקד: השיטה החדשה להתמודדות עם הטרור נחשפת" **הארץ** (26.05.2017); עמוס הראל "בינה מלאכותית תתפוס את מקומם של אנשי המודיעין: העתיד לפי מפקד 8200" **הארץ** (1.10.2021). ראו גם Badi Hasisi, Simon Perry, and Michael Wolfowicz, *Counter-Terrorism Effectiveness and Human Rights in Israel*, *in* International Human Rights and CounterTerrorism 409, 419–422 (Eran Shor and Stephen Hoadley eds., 2019)

אזורי הפשיעה בחוטים וסיכות החליפו זה מכבר מערכות מידע גאוגרפי (GIS) המעבדות נתוני פשיעה ממוכנים ומשרטטות על מסך המחשב "נקודות חמות" (hot spots) או "אזורים עתירי פשיעה" (high-crime areas).[216] זוהי גישת חיזוי משטרתי מבוססת מיפוי (predictive mapping).[217]

מערכות חיזוי פשיעה בשירות ארגוני שיטור יכולות להצביע על אזורים שבהם יש סבירות גבוהה לעבירות רכוש. זאת, בין השאר, בהסתמך על תאוריות קרימינולוגיות שמראות שבתנאים מסוימים צפויה עלייה במספר עבירות הרכוש – פריצה לבית בשכונה מסוימת שעולה יפה יכולה למשל לעודד עבריינים לנסות לשחזר את ההצלחה באותו אזור.[218] עיבוד היסטורי של אירועי עבריינות בתא שטח גאוגרפי מסוים יכול לזהות אזורים שבהם העבריינים צפויים לפעול שוב. הפניית כוחות שיטור לפטרול באותן "נקודות חמות" עשויה להרתיע עבריינים מלבצע שם עבירות. ואכן, כוחות משטרה ברחבי ארצות הברית דיווחו על ירידה ניכרת בהיקף עבריינות הרכוש בעקבות השימוש במערכות חיזוי פשיעה אלו.[219] עם זאת, מחקרים נוספים הצביעו על עלייה ברמת העבריינות לאחר מכן,[220] ואחרים לא מצאו כל שיפור מובהק סטטיסטית ברמת הפשיעה במחוזות שבהם השתמשו בטכניקות החיזוי האלו.[221]

216 על טכנולוגיות מיפוי קרימינולוגיות ראו Andrew Guthrie Ferguson, *Crime Mapping and the Fourth Amendment – Redrawing "High-Crime Areas,"* 63 Hastings L. J. 179 (2011). ראו גם Andrew G. Ferguson, *Policing Predictive Policing*, 94 Wash. U. L. Rev. 1109 (2017)

217 Rosamunde van Brakel, *Pre-Emptive Big Data Surveillance and its (Dis)Empowering Consequences: The Case of Predictive Policing*, in Exploring the Boundaries of Big Data 117, 120-122 (Bart van der Sloot, Dennis Broeders, and Erik Schrijvers eds., 2016)

218 ראו למשל Daniel B. Neill and Wilpen L. Gorr, *Detecting and Preventing Emerging Epidemics of Crime*, 4 Advances in Disease Surveillance (2007); Jerry H. Ratcliffe and George F. Rengret, *Near-Repeat Patterns in Philadelphia Shootings* 21 Security J. 58 (2008)

219 ראו Ferguson, *Policing Predictive Policing*, **לעיל** ה"ש 216, בעמ' 1130.

220 ראו למשל *Richmond Police Chief Says Department Plans to Discontinue "Predictive Policing" Software*, The Richmond Standard (24.6.2015)

221 Priscillia Hunt, Jessica Saunders, and John S. Hollywood, Evaluation of the Shreveport Predictive Policing Experiment (RAND Corp., 2014)

מערכות חיזוי אלגוריתמיות נמצאות בשימוש המשטרה גם כדי להתמודד עם פשיעה אלימה, מתוך הנחה שבאזורים גאוגרפיים מסוימים תנאי השטח נוחים יותר לביצועה – סמטאות ספציפיות שתנאי התאורה בהן לקויים או שנתיבי המילוט מהן נוחים, או שכונות שנמצאות בשליטת כנופיה זו או אחרת. חברת Hunchlab פיתחה מערכת המבוססת על גישה זו. היא חוזה פשיעה באמצעות למידת מכונה, בהסתמך על נתוני פשיעה היסטוריים ועל פרמטרים גאוגרפיים, תחבורתיים ומטאורולוגיים.[222] נראה שבחודשי הקליטה שלה בשיקגו הצליחה המערכת להפחית את שיעורי הפשיעה האלימה.[223]

חיזוי פשיעה נעשה גם באמצעות מערכת דירוג חברתי (social scoring).[224] אין הכוונה בהכרח למערכת דרקונית דוגמת זו המשמשת בסין, שיש לה מטרות אחרות,[225] אך העיקרון שביסוד המערכות להערכת מסוכנות וחיזוי רצידיביזם – שתיהן טכניקות חיזוי משטריות שתכליתן זיהוי (predictive identification) – הוא עיקרון זהה: הרשת החברתית של אדם, כלומר מכריו, חבריו ומשפחתו, היא אינדיקציה לרמת העבריינות הפוטנציאלית שלו. כאן מודל השיטור אינו פטרולים באזורים מועדי פשיעה לצורכי הרתעה, אלא זיהוי ישיר של חשודים באמצעים סטטיסטיים. גישה זו עודנה בחיתוליה ויישומה מצומצם יחסית, אך ברחבי ארצות הברית אפשר לראות ניצנים של פרויקטים בתחום.[226]

222 Aaron Shapiro, *Reform Predictive Policing* 541 Nature 458 (2017)

223 Timothy McLaughlin, *As Shootings Soar, Chicago Police Use Technology to Predict Crime*, Reuters (5.8.2017)

224 ראו Ferguson, *Policing Predictive Policing*, **לעיל** ה"ש 216, בעמ' 1137.

225 להגדרת דירוג חברתי בתקנות הבינה המלאכותית האירופיות, ראו ס' 3(45c) להצעה המתוקנת, **לעיל** ה"ש 57. כן ראו כהנא ושני, **לעיל** ה"ש 213, בעמ' 22; הרשות להגנת הפרטיות, **דירוג חברתי בראי הזכות לפרטיות: סקירת רקע בעניין שימוש במערכות לדירוג חברתי** (21.4.2020). בס' 5(1)(I) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53, מוצע לאסור שימוש במערכות דירוג חברתי בנסיבות מסוימות. על מדינת המעקב הסינית ראו Geoffrey Cain, The Perfect Police State: An Undercover Odyssey Into China's Terrifying Surveillance Dystopia of the Future (2021). על מערכת הדירוג החברתי בסין ראו גם עמדתו הזהירה יותר של Werbach, **לעיל** ה"ש 239.

226 לסקירה ראו Ferguson, *Policing Predictive Policing*, **לעיל** ה"ש 216, בעמ' 1143–1138.

2.3.3. חוק בתפירה אישית: רגולציה מבוססת בינה מלאכותית

בינה מלאכותית, כפי שראינו לעיל, יכולה להתבסס על קוד ויכולה להתבסס על למידת מכונה.[227] מערכות מהסוג הראשון נקראות מערכות מומחה. רגולציה המסתייעת במערכות מומחה נשענת על כללי החלטה דטרמיניסטיים ("אם x אז y"). כללים אלו מאפשרים לשרטט עצי החלטה מורכבים מאוד, שיכולים להתאים כלל משפטי למקרה פרטני. מערכות מומחה יכולות לסייע בקבלת החלטות מינהליות בעניין אכיפת דיני מיסים מורכבים[228] או לפשט ולבאר חקיקה מורכבת ולהתאים את התוצאה המשפטית הנכונה לנסיבות ספציפיות (למשל, בבחינת זכויות סוציאליות). ניסוח חקיקה שמתאימה למערכות מומחה דורשת בהירות והיגיון (ולא, הקוד לא יעבוד).[229] הפעלת מערכות כאלו מחייבת קידום של שקיפות ואחריותיות.[230]

גישה אחרת היא חקיקה מונחית נתונים (data-driven law). בשנת 2016 הושת מס על בעלי דירה שלישית ויותר.[231] הנוסחה שלפיה חושב המס, שתוארה באריכות בחקיקה,[232] התבססה על שני פרמטרים – שטח הדירה ומקומה הגאוגרפי (שיש לו השפעה על כמה מדדים שהלשכה המרכזית לסטטיסטיקה מפרסמת). נוסחה זו הצריכה חישוב שכולל פעולות מתמטיות מורכבות, שמעידות שהנוסחה היא ככל הנראה תוצר של רגרסיה לוגיסטית של מדדים אזוריים של הלשכה המרכזית לסטטיסטיקה עם נתוני שווי דירות לפי שטח. הנוסחה הסופית ששימשה בחוק

227 Mirelle Hildenbrandt, *Algorithmic Regulation and the Rule of Law* 376 Phil. Trans. R. Soc A (2018)

228 ראו לדוגמה עת"מ 2663-05-19 **הר שמש מושב שיתופי בע"מ נ' הממונה על חופש המידע ברשות המיסים** (5.9.2019).

229 ראו לדוגמה Sarah B. Lawsky, *Formalizing the Code* 70 Tax L. Rev. 377 (2017)

230 ראו לדוגמה את עניין הר שמש, **לעיל** ה"ש 228. ראו גם את עמדת ידיד בית המשפט של התנועה לחופש מידע בערעור עע"מ 6782/19 **הר שמש מושב שיתופי בע"מ אגש"ח נ' הממונה על חוק חופש המידע ברשות המסים**.

231 ס' 120–148 והתוספת לפרק יב לחוק ההתייעלות הכלכלית (תיקוני חקיקה ליישום המדיניות הכלכלית לשנת התקציב 2017 ו־2018), התשע"ז–2016, ס"ח 2592 בעמ' 17.

232 **שם**, בתוספת לפרק יב.

הייתה תוצאה של כמה גלגולים, שכן חוקרים מצאו כי גרסאות קודמות שלה אינן מנבאות את שווי הדירה ברמת דיוק סבירה.[233] בסופו של דבר ביטל בית המשפט העליון בפרשת קוונטינסקי את הוראות החוק המטילות מס זה,[234] בנימוק שהן לא נגעו לתוכנו של החוק אלא לפגמים בשורש הליכי החקיקה שלו בכנסת. עם זאת, מס ריבוי דירות משמש דוגמה לחקיקה מונחית נתונים, שאף עברה טיוב בעקבות הערות של מומחים. ואולם מורכבות הנוסחה ושינויים בערכי המקדמים שלה, שנעשו בשלבי החקיקה האחרונים, מרמזים שעדיין יש מקום לשיפור.

כשחקיקה מונחית נתונים אינה מנוסחת על יסוד מודלים סטטיסטיים שפיתחו אנליסטים אנושיים, וניסוחה מופק בטכניקות של למידת מכונה, קשה מאוד להתחקות בדיעבד על הרציונל שהנחה אותה. בשני המקרים הנוסחה תלויה בטיב הנתונים המשמשים לפיתוח המודל ובפונקציית המטרה שלה.[235]

קייסי וניבלט גורסים כי בגלל פערי המידע בין המחוקק לאזרח נדרשת חקיקה המנוסחת ככללים וסטנדרטים. אין המחוקק יכול לחזות מראש כל מערך נסיבות אפשרי שהאזרח עשוי להיקלע אליו, וגם אילו היה יכול הייתה חקיקה כזאת כנראה מסורבלת ויקרה מאוד. לשיטתם של קייסי וניבלט, הבינה המלאכותית תטייב את החוק באמצעות מילוי פערי המידע.[236] הם סבורים שמערכות בינה מלאכותית יכולות להוביל לתום עידן החקיקה המנוסחת ככללים וסטנדרטים ולהחלפתה ב"מיקרו־דירקטיבה".[237]

במודל שמתארים קייסי וניבלט, החוק יכיל קטלוג של דירקטיבות תפורות לפי מידה, המתמחות בדיוק במה מותר בכל סיטואציה פרטנית. האזרח יעקוב אחר דירקטיבות שעברו אופטימיזציה עבורו ויתעדכן באופן אוטומטי כשחוקים ישתנו.

233 רוני גולן ודני בן שחר "נוסחת המס על דירה שלישית מסובכת ומעוותת" **TheMarker** (23.9.2016).

234 בג"ץ 10042/16 **צחי קוונטינסקי נ' כנסת ישראל** (6.8.2017).

235 Hildenbrandt, **לעיל** ה"ש 227, בעמ' 3.

236 Anthony J. Casey and Anthony Niblett, *Self-Driving Laws*, 66 U. Toronto L.J. 429 (2016)

237 Idem, *Self-Driving Contracts*, 43 J. Corp. L. 1 (2017)

לדברי קייסי וניבלט, מוקד קבלת ההחלטות יעבור מקביעת כללים וסטנדרטים לניסוח המטרה שהמחוקק מבקש להשיג.

עמרי בן שחר ואריאל פורת מתארים בספרם מודל דומה לזה של קייסי וניבלט.[238] הם משרטטים חזון – שלאחדים הוא ייראה אוטופי ולאחרים מעורר פלצות[239] – של מערכת דינים מותאמת אישית, הנעזרת במערך מורכב של חיישנים, נתוני עתק ואלגוריתמיקה משוכללת כדי להכווין התנהגות אנושית באופן אופטימלי ומדויק. כך למשל, במקום לקבוע הגבלה קשיחה של המהירות המרַבית המותרת בכביש יוכל המחוקק לקבוע פונקציית מטרה של "מזעור תאונות דרכים", והמיקרו־דירקטיבות יתאימו באופן אלגוריתמי את המהירות המרַבית המותרת לכל נהג ונהג בהתאם לניסיונו, לתנאי הכביש או לפרמטרים אחרים.

2.4 יישומים פיננסיים של בינה מלאכותית

שחקנים רבים בתחום הפיננסים והכלכלה מאמצים מערכות בינה מלאכותית.[240] בנקים, למשל, משתמשים במערכות לניתוח חוזי הלוואה מסחריים, החוסכות להם שעות עבודה

238 Omri Ben-Shahar and Ariel Porat, Personalized Law: Different Rules For Different People (2021). ראו גם סימפוזיון מקוון בנושא באתר כתב העת The University of Chicago Law Review

239 ראו בהקשר זה את מאמרו של קווין וורבך על מערכת הדירוג החברתי הסינית, שנתפסת במערב כפנאופטיקון דיסטופי מבעית. וורבך מציע להסתכל על מערכת זו, בניכוי ההקשרים הדכאניים של המשטר האוטוריטרי הסיני, כארכיטקטורה של אסדרה אלגוריתמית המאפשרת האחדה ושוויון במענה הרגולטורי לאזרחים. Kevin Werbach, *Orwell that Ends Well? Social Credit as Regulation for the Algorithmic Age*, 2022 U. Ill. L Rev. 1417 (2022)

240 ראו גם הסקירה ב־Rosario Girasa, Artificial Intelligence as a Disruptive Technology 36–39 (2020). כמו כן, ראו הסקירה בארי אחיעז ואח' **בינה מלאכותית במגזר הפיננסי: שימושים נפוצים, אתגרים וסקירה השוואתית של התמודדות רגולטורית** (מוגש למחלקה הכלכלית בייעוץ וחקיקה [משרד המשפטים], הפקולטה למשפטים, אוניברסיטת תל אביב, 18.7.2022).

אנושיות רבות מספור;[241] עוד הם משתמשים במערכות לחיזוי התנהגות לקוחות של בנקאות קמעונאית,[242] במערכות ממוכנות לייעוץ בנקאי,[243] במערכות להתראה על ניסיונות להלבנת הון[244] ובמערכות לדירוג אשראי.[245]

הן הבנקים הן גורמים חוץ־בנקאיים משתמשים במערכות לדירוג אשראי. מערכות אלו מתבססות על מודלים סטטיסטיים המשקללים את ההיסטוריה הפיננסית של הלקוח (בעיקר עמידה בהחזר הלוואות) ועל פרמטרים אחרים שנמצאים באופן מסורתי במאגרי המידע הבנקאיים[246] כדי לתת לו ציון דירוג אשראי המשקף את "הסיכוי שלקוח יעמוד בפרעון התשלומים שבהם הוא מתחייב".[247]

לצד טכניקות מסורתיות אלו של דירוג אשראי, בשנים האחרונות הולך וגובר השימוש במודלים שפותחו בטכניקות של למידת מכונה, על בסיס מידע שלא משמש במודלים הקלסיים. קשרים ברשתות חברתיות של הלקוח, סוג הרכב שהוא מחזיק בו, רמת השכלה ואפילו סגנון הכתיבה של הודעות SMS – כולם

241 Hugh Son, *JPMorgan Software Does in Seconds What Took Lawyers 360,000 Hours*, Bloomberg (27.2.2017)

242 חיזוי בנקאי (predictive banking) מבוסס על חיזוי ההתנהלות הפיננסית של לקוחות בנקאיים – ניבוי תשלומים חודשיים חוזרים, אתרעות על הוצאות חריגות, המלצות על מכשירים פיננסיים מתאימים ותוכניות חיסכון. ראו E. Y. Barakina and I.S. Ismailov, *Legal Regulation of Using the Artificial Intelligence Technology in the Banking*, *in* Economic Systems in the New Era: Stable Systems in an Unstable World 40, 43–44 (Svetlana Igorevna Ashmarina et al. eds., 2021)

243 ראו: Lee Reiners, *Regulation of Robo-Advisory Services*, *in* Fintech: Law and Regualtion 353 (Jelena Madir ed., 2019)

244 Filip Koprivec et al., *Screening Tool for Anti-Money Laundering Supervision*, *in* Big Data and Artificial Intelligence in Digital Finance: Increasing Personalization and Trust in Digital Finance using Big Data and AI 233 (John Soldatos and Dimosthenis Kyriazi eds., 2022)

245 ראו Nikita Aggarwal, *The Norms of Algorithmic Credit Scoring*, 80 Cambridge Law Journal 42 (2021)

246 השוו לס' 16(ב)(4) לחוק נתוני אשראי.

247 ס' 2 לחוק נתוני אשראי.

יכולים להיות רלוונטיים בדירוג אשראי מבוסס למידת מכונה.[248] מידע מעין זה מתחיל לזרום גם מהרשתות החברתיות עצמן: פייסבוק רשמה לאחרונה פטנט על טכנולוגיה המאפשרת לה להגדיר דירוג אשראי של כל משתמש בהסתמך על דירוגי האשראי של כל חבריו, מתוך הנחה שאדם שחבריו הם בעלי דירוג אשראי נמוך, או שאין הם נוטים להחזיר הלוואות, סיכוי נמוך יותר להחזיר הלוואות במועד.[249]

ספק אם הבנקים הגדולים יימנעו מלשלב בינה מלאכותית שמבוססת על למידה חכמה ומשתמשת באופן יצירתי במידע חלופי[250] במנגנונים המסייעים לקבלת החלטות בנוגע לכדאיות מתן הלוואות (להבדיל ממערכות מומחה סטטיות שמבוססות על מודל סטטיסטי שפיתחו אנליסטים אנושיים); על אחת כמה וכמה כשחברות אינטרנטיות מתחילות לנשוף בעורפם של הבנקים ולספק הלוואות בטוחות ורווחיות יותר לכל דורש. אפשר לשער אפוא שבשלב מסוים יאמצו גם הבנקים את הבינה המלאכותית – אף שאולי שימושם בה יהיה מוגבל יותר ויתחשב בפרמטרים מעטים מאלו שמשמשים את חברות ההזנק.

לבינה מלאכותית יש יישומים גם בתחום ההשקעות. האגדה מספרת שהמקור להונה העצום של משפחת רוטשילד הוא פערי מידע. לנתן רוטשילד נודע על ניצחון האנגלים בקרב ווטרלו והוא מיהר למכור את שטרי החוב האנגליים שברשותו ויצר את הרושם שאנגליה הפסידה. סוחרים אחרים מיהרו להיפטר משטרי החוב שלהם במחירי הפסד, ואנשיו של רוטשילד רכשו אותם בזיל הזול.[251] פערי המידע נסגרו רק לאחר כמה ימים. בתחילת המאה ה־19 אפשר

248 Matthew Adam Bruckner, *The Promise and Perils of Algorithmic Lenders' Use of Big Data*, 93 Chi. Kent L. Rev. 3, 15 (2018)

249 Mark Sullivan, *Facebook Patents Technology to Help Lenders Discriminate against Borrowers Based on Social Connections,* Venture Beat (4.8.2015)

250 בכפוף למגבלות הרגולטוריות הקיימות. ראו **לעיל**, ה"ש 246.

251 לדיון באגדה זו ובמקורותיה ראו Niall Ferguson, The House of Rothschild, Volume 1: Money's Prophets, 1798–1848 (1999)

פיתוחה של רשת הטלגרף לסוחרים בעלי גישה לטכנולוגיה זו יתרון תחרותי דומה: היא אפשרה להעביר ידיעות מעיר לעיר בתוך חצי שעה.[252]

מאז תחילת האלף הנוכחי מאפשר המסחר הרובוטי (אלגו־טריידינג) קבלת החלטות בתוך שבריר שנייה וניצול פערים זעירים במידע: למכור ולקנות ניירות ערך בפרקי זמן שסוחרים אנושיים אינם יכולים להתחרות בהם.[253] חלק מהאלגוריתמים מסתמכים על מידע שמגיע גם מחוץ לבורסה, כגון מסרים ברשתות חברתיות כמו טוויטר ופייסבוק.[254] סביבת המסחר רוויה באלגוריתמים מתחרים – כל אלגוריתם מנסה למצוא דרכים מיטביות להשאת רווחים. יש אסטרטגיות מסחר אלגוריתמיות שבהן אלגוריתם אחד מנסה לשטות באלגוריתמים אחרים באמצעות הודעה על רצון לרכוש מניות מסוימות, המבוטלת לאחר שברירי שנייה.

המסחר האלגוריתמי בבורסות לניירות ערך תופס נתח ניכר ממחזור העסקאות.[255] חרף הרווח האינפיניטסימלי מכל עסקה, היכולת לבצע יותר מאלף עסקאות מדי שנייה הופכת רווחים שוליים אלו לסכומים לא מבוטלים. עם זאת, יש ראיות

252 Bob Pisani, *Plundered by Harpies: An Early History of High-Speed Trading,* 20 Financial History (2014)

253 Neil Johnson et al., *Abrupt Rise of New Machine Ecology Beyond Human Response Time,* 3 Scientific Reports, art. no. 2627 (11.9.2013); Brendan Conway, *Wall Street's Need for Trading Speed: The Nanosecond* Age, The Wall Street Journal (14.6.2011)

254 Steve Y. Yang and Sheung Yin Kevin Mo, *Social Media and News Sentiment Analysis for Advanced Investment Strategies*, *in* Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence 237 (Witold Pedrycz and Shyi-Ming Chen eds., 2016); Nuno Oliveira, Paulo Cortez, and Nelson Areal, *Some Experiments on Modeling Stock Market Behavior Using Investor Sentiment Analysis and Posting Volume from Twitter*, WIMS '13: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (2013); Raúl Gómez Martínez, Miguel Prado Román, and Paola Plaza Casado, *Big Data Algorithmic Trading Systems Based on Investors' Mood,* 20 J. Behavioral Finance 22 (2019)

255 Kevin O'Connell, *Has Regulation Affected the High Frequency Trading Market?,* 27 Cath. U. J. L. & Tech 145, 150–151 (2019)

שרווחי המסחר האלגוריתמי פוחתים מעט בשנים האחרונות, ככל הנראה כיוון שהתחרות אינה עם בני אדם איטיים בעולם הפיזי אלא עם אלגוריתמים מהירים באותה מידה.[256] שאלת מעמדן המשפטי של חברות המספקות שירותי מסחר אלגוריתמיים נדונה גם בישראל. בית המשפט המחוזי נדרש לקבוע אם חברות אלו חייבות ברישיון לניהול תיקי השקעות.[257]

בינה מלאכותית יכולה להיות לעזר גם בהשקעות חוץ־בורסאיות (private equity). משקיעים פרטיים וממשלתיים ערים היטב לקושי לזהות אילו חברות הזנק יעלו לגדולה. כך למשל, מתוך יותר מ־500 חברות ההזנק שהתקבלו לחממה היוקרתית Y Accelerator, 37 בלבד הצליחו להימכר או זכו להערכת שווי גבוהה דייה (יותר מ־40 מיליון דולרים). לשון אחר, כ־93% מחברות ההזנק שעברו את מסלול הסינון והתקבלו לחממה לא נחלו הצלחה של ממש. בהתחשב בכך ששיעור הקבלה לחממה זו הוא זעום, נראה כי גם משקיעים מנוסים עלולים להתקשות בזיהוי חברות הזנק בעלות פוטנציאל.[258] לכן בינה מלאכותית שביכולתה לזהות חברות ומיזמים בעלי פוטנציאל טוב להשקעה היא בעלת ערך רב. ואכן, גופי השקעה שונים פיתחו מערכות אלגוריתמיות לזיהוי השקעות מעין אלו.[259]

256 M. Philips, *How the Robots Lost: High-Frequency Trading's Rise and Fall,* Bloomberg Business (6.6.2013). ראו גם O'Connel, **שם**, בעמ' 152–154.

257 ראו ת"צ (כלכלית) 47119-12-15 **אפרימוב נ' יו אס ג'יי קפיטל ישראל בע"מ** (פורסם בנבו, 24.6.2019); רשות ניירות ערך, **הוראה לבעלי רישיון בקשר למתן שירותים תוך שימוש באמצעים טכנולוגיים** (2016).

258 Henry Blodget, *DEAR ENTREPRENEURS: Here's How Bad Your Odds Of Success Are*, Insider (28.5.2013)

259 Vlad Savov, *Telefónica Will Let an Algorithm Decide which Startups to Invest In*, The Verge (12.8.2015); Kirk Kardashian, *Could Algorithms Help Create a Better Venture Capitalist?* Fortune (5.8.2015); ראו למשל באתר Correlation Ventures.

2.5 עיבוד שפה

בפרק הקודם ראינו כי כבר בעת ניסוח מבחן טיורינג השתרש הרעיון שהיכולת לעבד שיג ושיח מילולי עם בני אדם ולהשתתף בו היא אמת מידה לסיווג טכנולוגיה כבינה מלאכותית.[260] העיסוק בטכניקות של עיבוד שפה טבעית (NLP) החל עוד בשנות החמישים של המאה העשרים. עיבוד שפה טבעית הוא תחום שקשור לבלשנות ולאינטליגנציה מלאכותית, ומלבד יומרותיו לפענח את הקשר שבין השפה לתודעה האנושית[261] יש לו שלל יישומים מעשיים, המאפשרים עיבוד אוטומטי ומשוכלל של טקסט.[262]

כדי לפענח את הרובד הסמנטי של משפט, כלומר להבין את המשמעות הבסיסית שלו, עלינו לעבד תחילה את הרובד התחבירי שלו. רובד שלישי, מתקדם יותר, הוא הרובד הפרגמטי – ניתוח משמעות המשפט בהקשר של השיח שבו הוא נאמר. טכניקות עיבוד שפה טבעית מכירות במורכבות השפה הטבעית – הן מבחינת רמת ההפשטה הנדרשת כדי לעבד אותה (תחביר, סמנטיקה, הקשר) הן מבחינת הרזולוציה של הניתוח (משפט יחיד לעומת שיח). העמימות של השפה הטבעית – אפילו של משפטים פשוטים – הם מהאתגרים המיוחדים העומדים לפני העוסקים בתחום.

טכניקות של עיבוד שפה טבעית מפרקות את התהליך המורכב למשימות: זיהוי משפטים ומילים ברצף של טקסט, אפיון תחבירי של משפטים, זיהוי שמות של אנשים ומקומות, המרה של טקסט לקול וקול לטקסט ואפילו ניתוח רגשות (sentiment analysis), סיכום מסמכים או תרגום מכונה.

260 **לעיל** ה"ש 36.

261 Noam Chomsky, *Language and Nature*, 104 Mind 1, 2 (1995); Patricia Smith Churchland, *Can Neurobiology Teach Us Anything about Consciousness?* 67 Proceedings and Addresses of the American Philosophical Association 23 (1994)

262 לדוגמאות של עיבוד שפה טבעית בתחום המחקר המשפטי האקדמי ראו: Law As Data: Computation, Text, and the Future of Legal Analysis (Michael A. Livermore and Daniel N. Rockmore eds., 2019)

אלגוריתמים לעיבוד שפה טבעית נדרשים להתאמן על בסיס נתונים גדול ככל האפשר. בסיס נתונים זה אינו מאגר טקסטואלי ותו לא, אלא מצבור טקסטים שזוהו ותויגו בהם פונמות, מורפמות, מבנים תחביריים, רגשות ועוד – הכול לפי מטרת המודל.[263]

לנוכח המורכבות של שפות טבעיות, תרגום מכונה הוא אחד היישומים המאתגרים ביותר של עיבוד שפה טבעית. למילים רבות יש יותר ממשמעות אחת, בין שפות אין התאמה חד־חד־ערכית ולא לכל מילה או ביטוי יש מילה או ביטוי מקביל בשפה אחרת. גם היכולת למיפוי תחבירי של היגדים בשפת התרגום לא תמיד קיימת.[264] בתחילת הדרך הייתה הנחת העבודה שמתודולוגיות של פענוח צפנים (קריפטוגרפיה) הן המפתח לתרגום מכונה יעיל,[265] אך ניסיונות אלו לא צלחו. עד שנת 2000 התבססה הגישה הפרדיגמטית לתרגום מכונה על מנוע חוקים, שהסתמך לצורך התרגום על קידוד של כללים תחביריים ומילון. מתחילת המאה הגישה המובילה לתרגום מכונה היא שימוש בטכניקות סטטיסטיות של ניתוח קורפוס שפה ובלמידת מכונה.

טכניקות של עיבוד שפה אף מאפשרות לצ'אט בוטים לשוחח ברשת עם בני אנוש. מאז הוצגה בשנת 1966 תוכנת השיחה הראשונה, הפסיכיאטרית האלקטרונית אלייזה,[266] התפתחה רמת השיחה בין מערכות ממוחשבות לבני אדם התפתחות ניכרת. אומנם עדיין אפשר להבחין בין צ'אטבוטים לבני אדם, אך כיום הם פרוסים באתרים מקוונים רבים כדי לספק שירות מקוון בהתכתבות ולחסוך המתנה למפעיל אנושי. גם עוזרים דיגיטליים דוגמת סירי ואלכסה משתמשים

263 ראו לדוגמה "קורפוס השפה העברית – תיוג מורפולוגי", **Data Gov**.

264 Lynne Bowker and Jairo Buitrago Ciro, Machine Translation and Global Research 37–53 (2019)

265 John Hutchins, *Warren Weaver and the Launching of MT*, *in* Early Years in Machine Translation 17-20 (John Hutchins ed., 2000)

266 Eleni Adamopoulou and Lefteris Moussiades, *Chatbots: History, Technology, and Applications*, 2 Machine Learning with Applications (2020). בשנות השבעים פנה המשורר דוד אבידן למעבדות איי־בי־אם בבקשה שיאפשרו לו לשוחח עם אלייזה. ראו דוד אבידן **כל השירים** כרך ג 104–9 (מוסד ביאליק 2010); ראו גם גל מור "הבינה של אבידן ואלייזה המלאכותית" **ynet** (2.7.2001).

בפונקציות מורכבות של עיבוד שפה – מהמרת קול לטקסט עד עיבוד בקשות שונות של משתמשים.

לניטור תוכן מקוון יש תכליות מגוונות – זיהוי תכנים פורנוגרפיים שמפירים את כללי הקהילה, זיהוי הפרות של זכויות יוצרים והתמודדות עם ביטויי שנאה ועם סילוף מידע. מרבית הפלטפורמות המקוונות נשענות על שילוב של אלגוריתמים ובקרה אנושית לניטור תוכן, ולפי שעה נראה שאין הן יכולות להסתמך על אלגוריתמים בלבד.[267] אומנם משימות של ניטור תוכן אינן מוגבלות לניטור טקסטים, אבל פיתוח יכולות עיבוד של שפה טבעית הן קריטיות להתמודדות עם האתגרים העולים מביטויי שנאה ומסילוף מידע. אתגרים אלו קשים במיוחד לנוכח היעדר הגדרות אחידות לביטויי שנאה ולסילוף מידע,[268] לחשש מפני ההשפעות השליליות של ניטור יתר על זכויות אדם[269] ולקשיים הטכניים הכרוכים בזיהוי ההקשר שבו נכתבו הדברים.[270]

2.6 בינה מלאכותית ברפואה

יישומי בינה מלאכותית צפויים לשנות את מערכת הרפואה בכל שלביה, משלב המניעה עד שלב הטיפול, ולחולל תמורות במגוון רחב של תחומים – מרפואה פרטית עד תכנון מדיניות בריאות הציבור.

267 ראו לדוגמה את ההתנצלות שפרסמה טוויטר לאחר שנאלצה להורות לחלק ניכר מאנשי בקרת התוכן שלה שלא להגיע למקום עבודתם בעקבות התפרצות מגפת הקורונה באפריל 2020. Vijaya Gadde and Matt Derella, *An Update on our Continuity Strategy During COVID-19*, Twitter Blog (16.3.2020)

268 ראו לדוגמה Rotem Medzini and Tehilla Shwartz Altshuler, Dealing with Hate Speech on Social Media (2019); Alexander Brown, *What Is Hate Speech? Part 1: The Myth of Hate*, 36 Law and Philosophy 419 (2017); idem, *What Is Hate Speech? Part 2: Family Resemblances*, 36 Law and Philosophy 561 (2017); Alexander Brown and Adriana Sinclair, The Politics of Hate Speech Laws 11–19 (2020); Ronan Ó. Fathaigh, Natali Helberger, and Naomi Appelman, *The Perils of Legally Defining Disinformation*, 10 Internet Pol'y. Rev. (2021)

269 ראו לדוגמה **להלן**, בסעיף 5.5.

270 United Nations General Assembly, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (A/HRC/38/35, 2018), para 29 (להלן: HRC 2018).

טעויות רפואיות[271] הן הגורם השלישי למוות בקרב האוכלוסייה בארצות הברית.[272] לא מן הנמנע שמצב הדברים דומה גם בישראל. מערכות בינה מלאכותית יכולות לצמצם את היקף הטעויות הרפואיות באמצעות דיאגנוזה שמבוססת על אנליטיקה של ההיסטוריה הרפואית של רבבות ומיליוני חולים אחרים, שרמת הדיוק שלה גבוהה יותר.[273] באמצעות זיהוי אנומליות בסריקות MRI או צילומי רנטגן, ואפילו סיוע בהליכים כירורגיים;[274] באמצעות זיהוי הפרעות פסיכיאטריות על פי שינויים בדפוסי הדיבור,[275] באמצעות אפליקציות לזיהוי נגעים סרטניים בעור;[276] באמצעות חיזוי כוונות התאבדות על סמך היסטוריה פסיכיאטרית או פוסטים ברשתות החברתיות (הללו רלוונטיים למקרי התאבדות שאינם ניתנים לחיזוי על רקע היסטוריה); ועוד.[277]

כבר עתה יש מערכות נבונות שיכולות האבחון שלהן עולות על אלו של רופא אנושי במגוון תחומים, ובהם דרמטולוגיה, קרדיולוגיה ואונקולוגיה.[278] גם

271 בהקשר זה, טעות רפואית פירושה פעולה לא מכוונת (מעשה או מחדל), פעולה שאינה משיגה את התוצאה שלשמה היא בוצעה, פעולה שבוצעה בצורה לא נכונה, פעולה שגויה שאינה מתאימה לנסיבות העניין או חריגה מההליך הטיפולי המקובל – בין שנגרם נזק לחולה ובין שלא.

272 Martin. A. Makary and Michael Daniel, *Medical Error – the Third Leading Cause of Death in the U.S.*, BMJ (2016)

273 Simon Parkin, *The Artificially Intelligent Doctor Will Hear You Now*, MIT Technology Rev. (9.3.2016)

274 Adam Bohr and Kaveh Memarzadeh, *The Rise of Artificial Intelligence in Healthcare Applications*, *in* Artificial Intelligence in Healthcare 25, 34–35 (Adam Bohr and Kaveh Memarzadeh eds., 2020)

275 Joseph Frankel, *How Artificial Intelligence Could Help Diagnose Mental Disorders*, The Atlantic (23.8.2016)

276 Madhumita Murgia, *Google Launches AI Health Tool for Skin Conditions*, Financial Times (18.5.2021)

277 Mason Marks, *Artificial Intelligence-Based Suicide Prediction*, 21 Yale J. L. & Tech 98, 106–111 (2019)

278 Dayong Wang et al., *Deep Learning for Identifying Metastatic Breast Cancer* (2016), available at arXiv; Gongning Luo et al., *A Novel Left Ventricular Volumes Prediction Method Based on Deep Learning Network in Cardiac MRI*, 43 Computing in Cardiology Conference (CinC) (2016); Andre Esteva

בתחומים שבהם יש למומחיות אנושית יתרון – קבוע או זמני – על פני מערכות נבונות, הסתייעות של המומחה האנושי בהן יכולה להפחית את הסיכוי לטעויות אנוש.[279] מערכות נבונות יכולות לסייע בזיהוי חולים שנדבקו בנגיף קורונה בהסתמך על ניתוח צילומי רנטגן של החזה[280] או שכלול שיטת האבחון הקלינית של הנגיף.[281] מערכות בינה מלאכותית יכולות לתרום גם בשלב הפרוגנוזה או להציע תחזיות באשר להתפתחות מחלות או לסיכוי שתחול הידרדרות במצבם של מטופלים.[282]

גם ברמת השירות למטופלים, בינה מלאכותית יכולה להיות לעזר. קופת חולים כללית הכניסה לאחרונה לשימוש מערכת בינה מלאכותית המסייעת לרופא לקבל תמונה מגוונת על מצבו הבריאותי של המטופל, לרבות המלצות להמשך טיפול; המתמללת את השיח בין הרופא למטופל; המנתחת דפוסים של אי־הגעה לתורים שנקבעו מראש; ואפילו מנתחת את מצבו הנפשי של המטופל בזמן אמת.[283] היתרונות של מערכות בינה מלאכותית ברפואה אינן מתמצות בתפקידן ככלי עזר דיאגנוסטי. מחקרים מראים שבעיני המטופלים ייעוץ רפואי באמצעות אפליקציית צ׳אטבוטים רגיש וקשוב יותר מייעוץ של רופא בשר ודם.[284]

---

et al., *Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks*, 542 Nature 115 (2017)

279 ראו Wang et al., **שם**.

280 Tulin Ozturket et al., *Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-ray Images* 121 Computers in Biology and Medicine (2020)

281 Liping Sun et al., *Combination of Four Clinical Indicators Predicts the Severe/Critical Symptom of Patients Infected COVID-19*, 128 Journal of Clinical Virology 104431 (2020); Yazeed Zoabi, Shira Deri-Rozov, and Noam Shomron, *Machine Learning-Based Prediction of COVID-19 Diagnosis Based on Symptoms*, 4 NPJ Digital Medicine 3 (2021)

282 לסקירה כללית ראו S. Scott Graham, The Doctor and the Algorithm: Promise, Peril and the Future of Health AI 46-48 (2022)

283 ישראל וולמן "הטכנולוגיה שתשדרג את מערכת הבריאות" **ynet** (29.10.2022).

284 John W. Ayers et al., *Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum* 183 JAMA Intern. Med. 589 (2023)

בינה מלאכותית מאפשרת רפואה מדויקת (precision medicine)[285] – אבחון וטיפול מותאמים אישית לכל חולה בהתבסס על כמות גדולה של נתונים משלל מקורות, ובהם תיקים רפואיים אלקטרוניים, מידע שהחולים עצמם משתפים ברשתות חברתיות, נתונים המגיעים ממחשוב לביש, נתונים סביבתיים ועוד. המטרה: לנבא לכל אדם, חולה או בריא, אילו סיכונים רפואיים צפויים לו בעתיד ומה הדרכים היעילות ביותר למנוע את התממשותם. כלומר למצוא מה הטיפול המתאים ביותר עבורו בהינתן ההיסטוריה הרפואית שלו, הסביבה שהוא חי בה, המאפיינים הגנטיים שלו וכדומה.[286] כל זה מתוך הנחה ששיטת הטיפול השכיחה או המצליחה ביותר עבור כלל האוכלוסייה לא בהכרח תתאים לכל חולה.

יתר על כן, בכוחן של מערכות בינה מלאכותית להשפיע לחיוב על בריאות הציבור – למשל באמצעות סיוע בניטור התפרצות מגפות.[287] כבר בשנת 2002 היו למערכת המודיעין הרפואית הגלובלית GIPHIN אינדיקציות ראשונות להתפרצות מגפת הסארס בסין, בהסתמך על ידיעה עיתונאית על עלייה במכירות של תרופות אנטי־ויראליות.[288] מערכות בינה מלאכותית יכולות לסייע

285 Bertalan Mesko, *The Role of Artificial Intelligence in Precision Medicine*, 2 Expert Review of Precision Medicine and Drug Development 239 (2017); J. Larry Jameson and Dan L. Longo, *Precision Medicine – Personalized, Problematic, and Promising*, 372 N. Engl. J. Med 2229 (2015)

286 Po-Yen Wu et al., *Omic and Electronic Health Record Big Data Analytics for Precision Medicine,* 64 IEEE Transactions on Biomedical Engineering 263 (2017)

287 לסקירת שימושים של יישומי בינה מלאכותית לתכליות אפידמיולוגיות ראו Said Agrebia and Anis Larbi, *Use of Artificial Intelligence in Infectious Diseases*, *in* Artificial Intelligence in Precision Health 415, 422–423 (Debmalya Barh ed., 2020); לבינה מלאכותית בהתפרצות הקורונה ראו לדוגמה Muhammad Awais et al., *Overview of IoT and Machine Learning For E-Healthcare in Pandemics and Health Crises*, *in* Data Science Advancements in Pandemic and Outbreak Management 16, 30–33 (Eleana Asimakopoulou and Nik Bessis eds., 2021)

288 M. Dion, P. AbdelMalik, and A. Mawudeku, *Big Data and the Global Public Health Intelligence Network (GPHIN)*, 41 Canada Communicable Disease Report 209 (2015)

בחיזוי התפרצויות מקומיות וכך להביא להקצאה יעילה של משאבים רפואיים ולסייע למקבלי ההחלטות.[289] בשלבים מוקדמים של התפרצות נגיף קורונה בישראל הוצע להשתמש במערכות נבונות כדי להקצות לכל תושב פרופיל רפואי המשקף את הסיכוי שיידבק בנגיף.[290] עם זאת, יש מחקרים שמצאו שכלי בינה מלאכותית רבים כשלו בסיוע בהתמודדות עם הקורונה עקב הסתמכות על בסיסי נתונים "מלוכלכים" וחלקיים או עקב טעויות באימון המודל.[291]

289 Agrebi and Larbi, **לעיל** ה"ש 287. בהקשר של זיהוי התפרצויות מקומיות של נגיף הקורונה ראו Matheus Henrique Dal Molin Ribeiro et al., *Short-Term Forecasting COVID-19 Cumulative Confirmed Cases: Perspectives for Brazil*, 135 Chaos, Solitons & Fractals (2020)

290 רפאלה גויכמן "משרד הבריאות חבר ל־NSO כדי לדרג את הסיכוי שתידבקו בקורונה" **TheMarker** (29.3.2020). ראו גם הרשות להגנת הפרטיות, **לעיל** ה"ש 225.

291 לסקירה ראו Will Douglas Heaven, *Hundreds of AI Tools Have Been Built to Catch Covid. None of them Helped*, MIT Technology Review (30.7.2021); Sean Mann et al., *Artificial Intelligence Applications Used in the Clinical Response to COVID-19: A Scoping Review*, 1 PLoS Digital Health (2022)

פרק שלישי

מדיניות אתיקה של בינה מלאכותית

העניין הגובר בטכנולוגיות בינה מלאכותית, לצד הבנת הסיכויים והסיכונים הטמונים בהן, הביאו לפרסום עשרות מסמכים המבקשים להציע מתווי מדיניות, אסטרטגיות וקווים מנחים אתיים הנוגעים לתחום זה,[292] ולאחרונה גם להגשת כמה הצעות חקיקה באירופה ובארצות הברית.

292 לסקירות חלקיות ראו Anna Jobin, Marcello Ienca, and Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, 1 Nature Machine Intelligence 389 (2019); Yi Zeng, Enmeng Lu, and Cunqing Huangfu, *Linking Artificial Intelligence Principles* (2018), available at arXiv; Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, 30 Minds and Machines 99

מקורן של כמה יוזמות כאלו בגופים על־לאומיים, כגון הארגון לשיתוף פעולה ולפיתוח כלכלי (OECD), שלצד סקירה מחקרית של בינה מלאכותית[293] פרסם מסמך עקרונות בנושא.[294] את המסמך אימץ גם פורום עשרים שרי הכלכלה ונגידי הבנקים המרכזיים (G-20).[295] בשנת 2018 פרסם הדַוָּח המיוחד של האומות המאוחדות (UN Special Rapporteur) לענייני חופש ביטוי דוח העוסק בהשפעות הבינה המלאכותית על חופש הביטוי וזכויות אחרות.[296] כדי להתוות את אסטרטגיית הבינה המלאכותית של אירופה[297] הקימה מועצת אירופה קבוצת מומחים (AI HLEG) שניסחו קווים מנחים לבינה מלאכותית אמינה[298] והגדרה מקובלת לבינה מלאכותית.[299] במהלך שנת 2020 פרסמה המועצה מסמך מדיניות על התייעצות ציבורית בנושא בינה מלאכותית "לקראת מערכת אקולוגית של מצוינות ואמון".[300]

מדינות רבות אחרות שקדו אף הן על ניסוח מסמכי מדיניות הנוגעים לבינה מלאכותית, אך לרוב במסגרת אסטרטגיה לאומית בעניין בינה מלאכותית ולא כמצע לדיון נפרד באתיקה של בינה מלאכותית. בשנת 2016 למשל פרסם ממשל

(2020); Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles For AI*, Berkman Klein Center (2020)

293 Artificial Intelligence in Society (OECD publishing 2019)

294 OECD Council Recommendation on Artificial Intelligence (22.5.2019) (להלן: OECD 2019).

295 G20 Ministerial Statement on Trade and Digital Economy (2019)

296 ראו HRC 2018, **לעיל** ה"ש 270.

297 ראו AI4EU, **לעיל** ה"ש 50.

298 Ethics Guidelines for Trustworthy AI (AI HLEG, 18.12.2018) (להלן: AI HLEG 2018); Ethics Guidelines for Trustworthy AI (AI HLEG 8.4.2019) (להלן: AI HELG 2019).

299 A Definition of AI: Main Capabilities and Scientific Disciplines (AI HLEG, 8.4.2019)

300 European Commission, On Artificial Intelligence – A European Approach to Excellence and Trust (להלן: European Commission 2020); European Commission, White Paper on Artificial Intelligence: Public Consultation towards a European Approach for Excellence and Trust (17.7.2020) (להלן: EU White Paper 2020).

אובמה דוח בנושא בינה מלאכותית בעקבות סדרת סדנאות מומחים שקיים,[301] ובשנת 2019 פרסם ממשל טראמפ את הצו הנשיאותי לשמירה על מעמדה המוביל של ארצות הברית בבינה מלאכותית, המתווה את האסטרטגיה האמריקאית בעניינה.[302] הצו הנשיאותי אינו מכיל את המונח אתיקה, אלא מתרכז בצורך לפתח בינה מלאכותית ולהיות חלוצים בתחום זה.[303] בהתאם להוראות הצו פרסם משרד הניהול והתקציב בבית הלבן (OMB) בשנת 2020 חוזר הכולל עקרונות של מדיניות אסדרה של בינה מלאכותית. מטרת החוזר הייתה בעיקר להנחות סוכנויות פדרליות בשיקולים הנוגעים לאסדרה של התחום באופן שיקדם חדשנות, הסרת חסמים והגנה על ערכים אמריקאיים כמו פרטיות, חירות וזכויות אדם.[304] באוקטובר 2022 פרסמה המחלקה המייעצת לנשיא ארצות הברית בנושאי מדע וטכנולוגיה (OSTP, the Office of Science and Technology Policy) את "המתווה למגילת זכויות בבינה מלאכותית – מערכות אוטומטיות בשירות העם האמריקאי".[305] על אף הכותרת, מדובר במסמך מדיניות בלתי מחייב, המציע קווי מתאר לפיתוח מדיניות ולעיצוב פרקטיקות שיגנו על הזכויות האזרחיות, יקדמו ערכים דמוקרטיים בבנייה ובפיתוח של מערכות אוטומטיות ויוודאו

301 National Science and Technology Council, Executive Office of the President, Preparing for the Future of Artificial Intelligence (2016)

302 US President, Executive Order on Maintaining American Leadership in Artificial Intelligence (2019)

303 על פי תקציב אמריקה לשנת 2019, הבינה המלאכותית היא אחד מתחומי העניין העיקריים בתחום המחקר והפיתוח. Office of Management and Budget, Executive Office of the President, Budget of the United States Government, Fiscal Year 2019, בעמ' 36.

304 Office of Management and Budget, Executive Office of the President, Memorandum for the Heads of Executive Departments and Agencies, Guidance for Regulation of Artificial Intelligence Applications, M-21-06 (17.11.2020)

305 The White House, Office of Science and Technology Policy, Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American people (2022) (להלן: AI BOR). ראו גם עמיר כהנא ודפנה דרור־שפוליאנסקי "זכויות אדם בבינה מלאכותית? על המתווה האמריקאי למגילת זכויות בבינה מלאכותית ומשמעותו" **אתר משפט ועסקים** (18.5.2023).

שלא תהיה פגיעה בזכויות הציבור, בהזדמנויות שלו או בנגישות שלו לצרכים חיוניים.[306]

דוח הוועדה של בית הלורדים לבינה מלאכותית נדרש להיבטים עקרוניים של אתיקה בבינה מלאכותית.[307] גרמניה, שאימצה אסטרטגיה לאומית בנושא בינה מלאכותית,[308] הקימה ועדת מומחים לאתיקה של מידע. תפקיד הוועדה לייעץ לממשל הפדרלי בעניין שימושים אתיים בבינה מלאכותית, והדוח הראשון שלה משרטט קווים מנחים בתחום זה.[309] במאי 2019 פרסמה האקדמיה של בייג'ין לבינה מלאכותית (ארגון שמימונו בא ממשרד המדע והטכנולוגיה של סין ומהממשלה המוניציפלית של בייג'ין) את מסמך עקרונות הבינה המלאכותית של בייג'ין,[310] ויש הסבורים שפרסומו מאותת על נכונותה של סין להצטרף לשיח הבינלאומי על שימוש אתי בטכנולוגיה.[311] המשרד לענייני פנים ותקשורת של יפן פרסם ב־2017 טיוטת הנחיות למפתחי מערכות נבונות.[312]

בישראל, ועדת המשנה של המיזם הלאומי למערכות נבונות בנושא אתיקה ורגולציה של בינה מלאכותית (להלן: ועדת המשנה) הגישה בשלהי שנת 2019 דוח מסכם. בדוח נכללו העקרונות האתיים שנמצאו בעיניה בעלי חשיבות

306 AI BOR, **לעיל** ה"ש 305, בעמ' 3.

307 House of Lords Select Committee on Artificial Intelligence, AI in the UK: Ready, Willing and Able? 125 (2018)

308 Bundesministerium für Bildung und Forschung; Bundesministerium für Wirtschaft und Energie; Bundesministerium für Arbeit und Soziales, Strategie Künstliche Intelligenz der Bundesregierung (15.11.2018)

309 Datenethikkommission der Bundesregierung, Opinion of the Data Ethics Commission 43 (22.1.2020)

310 Beijing Academy of Artificial Intelligence, Beijing AI principles (2019)

311 Will Knight, *Why Does Beijing Suddenly Care about AI Ethics?* MIT Technology Review (31.5.2019)

312 The Conference toward AI Network Society, Ministry of Internal Affairs and Communications, Draft AI R&D Guidelines For International Discussions (28.7.2017)

עבור מקבלי החלטות.[313] בשנת 2021 גיבש הצוות הבין־משרדי לרגולציה ואתיקה בתחום הבינה המלאכותית בישראל, במסגרת התוכנית הלאומית לבינה מלאכותית, רשימת עקרונות אתיים שמן הראוי שהתוכנית תפעל לאורם. בגיבוש הרשימה הושם דגש על עקרונות ה־OECD ועל המלצות ועדת המשנה. המסמך "עקרונות מדיניות, רגולציה ואתיקה בתחום הבינה המלאכותית", שפרסם משרד החדשנות, המדע והטכנולוגיה בשלהי שנת 2022 כולל גם המלצות לעקרונות אתיים.[314]

לא רק גופים על־לאומיים ומדינות עוסקים בסוגיה. עוד עוסקים בה תאגידים בינלאומיים, ובהם אינטל,[315] מייקרוסופט,[316] גוגל,[317] איי־בי־אם[318] וסוני;[319] קבוצות חשיבה מקרב התעשייה, במסגרת התאגדויות אד־הוק דוגמת "שותפות לבינה מלאכותית" (Partnership on AI), שחברות בה גוגל, פייסבוק, אמזון ואפל ועוד;[320] או גופים ותיקים כמו הארגון הבינלאומי של מהנדסי האלקטרוניקה והחשמל (IEEE).[321] מובן שייתכן שמסמכי מדיניות אלו, מאת חברות פרטיות,

313 המיזם הלאומי למערכות נבונות, "ועדת המשנה בנושא אתיקה ורגולציה של בינה מלאכותית, דוח מסכם" (2019) (להלן: המיזם הלאומי למערכות נבונות (2019)).

314 משרד החדשנות, המדע והטכנולוגיה, **מדיניות רגולציה ואתיקה בתחום הבינה המלאכותית בישראל** (2022) (להלן: מסמך מדיניות הרגולציה של משרד החדשנות). ראו גם תהילה שוורץ אלטשולר, עמיר כהנא, גדי פרל ואורי פריימן **אסדרת הבינה המלאכותית בישראל מחייבת קווים אדומים שימנעו פגיעה בזכויות יסוד** (חוות דעת, המכון הישראלי לדמוקרטיה 30.12.2022).

315 *Artificial Intelligence – The Public Policy Opportunity*, Intel (2017)

316 *Responsible AI*, Microsoft

317 *Our Principles*, Google AI

318 Adam Cutler and Milena Pribić, Everyday Ethics for Artificial Intelligence (IBM Design Program Office, 2018)

319 *Sony Group AI Ethics Guidelines*, Sony (2018)

320 *Partnership on AI*, AI (2016)

321 IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems (2019)

אינם אלא ניסיון לבלום רגולציה עתידית באמצעות הצהרת כוונות עמומה שעיקרה איתות סגולה (virtue signaling) לא מחייב.[322]

גם האקדמיה וארגוני החברה האזרחית פרסמו ניירות עמדה בנושא אתיקה ורגולציה של בינה מלאכותית. עם אלו אפשר למנות, בין השאר, את הצהרת מונטריאול של אוניברסיטת מונטריאול;[323] את הצהרת טורנטו, שמשותפת לארגונים אמנסטי ו־Access Now;[324] את הדוח "פרטיות וחופש ביטוי בעידן הבינה המלאכותית", שפרסמו הארגונים Privacy International ו־Article 19;[325] ופרסומים נוספים מאת מרכזי מחקר, איגודים מקצועיים ומכונים אקדמיים.[326]

ריאן קיילו העיר כי מסמכי מדיניות אתיקה מאפשרים פרשנות גמישה ומתאפיינים ברקמה פתוחה יחסית, כיוון שרמת ההפשטה בהם גבוהה. עוד ציין קיילו שגם

322 ראו למשל עמדתו של תומס מצינגר, חבר קבוצת המומחים האירופית, הרואה אף במסמך המומחים האירופי (AI HELG 2019, **לעיל** ה"ש 298) דוגמה לטיעון אתי – דיונים אתיים כאמצעי לשיהוי אסדרה אפקטיבית. Thomas Metzinger, *Ethics Washing Made in Europe*, Der Tagesspiegel (8.4.2019). ראו גם Inga Ulnicane, *Artificial Intelligence in the European Union – Policy, Ethics and Regulation*, in The Routledge Handbook of European Integrations, 254, 263 (Thomas Hoerber, Gabriel Weber, and Ignazio Cabras eds., 2022); Paul Nemitz, *Constitutional Democracy and Technology in the Age of Artificial Intelligence*, 376 Philosophical Transactions of the Royal Society A (2018)

323 The Montreal Declaration for a Responsible Development of Artificial Intelligence (2017) (להלן: הצהרת מונטריאול).

324 The Toronto Declaration: Protecting the Right to Equality in Machine Learning (2018) (להלן: הצהרת טורונטו).

325 Privacy and Freedom of Expression in the Age of Artificial Intelligence (Privacy International and Article 19, April 2018) (להלן: PI&A19).

326 ראו למשל *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*, FAT\ML (2016); *Asilomar AI Principles*, Future for Life Institute (2017); Peter Cihon, Standards for AI Governance – International Standards to Enable Global Coordination in AI Research & Development (Future of Humanity Institute, Oxford University, 2019); US Public Policy Council, Association for Computing Machinery, Statement on Algorithmic Transparency and Accountability (2017) (להלן: ACM 2017).

כשיש קונצנזוס בנוגע למוסריות של טכנולוגיות בינה מלאכותית, אין בנמצא מנגנונים לאכיפת הפרות. לכן חברות מסחריות העוסקות בפיתוח טכנולוגיות בינה מלאכותית, ובפרט החברות המובילות בשוק, מעדיפות סטנדרטים אתיים שאפשר להתעלם מהם לפי הצורך על פני כללים מחייבים מבחינה משפטית.[327]

לא נתאר בפרוטרוט את כל המסמכים האלו – שנוסחו בזמנים שונים למטרות שונות – אלא נתרכז בעקרונות שחוזרים ומופיעים בהם. צוות חוקרים במרכז ברקמן קליין באוניברסיטת הרווארד סקר שלושים ושישה מסמכים מסוג זה ואיתר שמונה תמות שחוזרות ברובם.[328] ז'ובן, ינקה וואיינה דגמו שמונים מסמכים כאלו וזיהו באמצעות טכניקות כמותיות כ־11 אשכולות של עקרונות שחוזרים ומופיעים בהם.[329] זנג, לו והונגפאו דגמו 27 מסמכים כאלו, השתמשו בניתוח סמנטי כדי לאפיין נושאים חוזרים וקישרו בין נטיית המסמכים לציין נושאים אלו ובין טיב הגורם שחיבר אותם (ממשלתי, פרטי או מגזר שלישי).[330] אומנם ייתכן שאפשר להעלות השגות מתודולוגיות על הממצאים הכמותיים של מחקרים אלו ומחקרים דומים, אבל ההיריסטיקה האיכותנית של העקרונות השכיחים באתיקה של הבינה המלאכותית העולה ממחקרים אלו מדויקת די הצורך לצרכים שלנו.

3.1 שקיפות

תמת השקיפות[331] שכיחה במסמכי אתיקת בינה מלאכותית.[332] פיילד ועמיתיה כוללים בה, נוסף על שקיפות, גם הסברתיות, קוד פתוח ומידע פתוח, וכן אינטראקציה מיודעת עם בינה מלאכותית.[333] לעיתים עקרון השקיפות

327 Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. Davis L. Rev. 399, 408 (2017)

328 ראו .Fjeld et al, **לעיל** ה"ש 292.

329 ראו Jobin, Ienca and Vayena, **לעיל** ה"ש 292.

330 ראו Zeng, Lu and Huangfu, **לעיל** ה"ש 292.

331 ראו בהרחבה גם בפרק 7 **להלן**; וכן ראו פרל ושוורץ אלטשולר, **להלן** ה"ש 881.

332 94% מן המסמכים שדגמו .Fjeld et al, וכ־87% מן המסמכים שדגמו Jobin, Ienca and Vayena (**לעיל** ה"ש 292) מתייחסים לשקיפות.

333 .Fjeld et al, **לעיל** ה"ש 292, בעמ' 41.

מתואר כתנאי ולעיתים הוא נכלל בעקרונות אחרים, כגון עקרון ההוגנות[334] או עקרון האחריותיות.[335] יש המדגישים שהשקיפות חשובה כיוון שהיא מאפשרת פיקוח של בעלי עניין מגוונים (ובהם ארגוני החברה האזרחית) על מערכות בינה מלאכותית, והלכה למעשה היא תנאי לאחריותיות.[336]

דרך אחרת להמשיג את עקרון השקיפות בהקשר של בינה מלאכותית הוא לפנות לעקרון המפורשוּת (explicability). המודל האתי שהציעו לוצ׳אנו פלורידי ועמיתיו לבינה מלאכותית נסמך על ארבעה נדבכים ביו־אתיים כלליים, שעליהם נוסף נדבך המפורשוּת.[337] לשיטתם, עקרון המפורשוּת הוא מזיגה בין ההיבט האפיסטמולוגי – ההבנה של אופן פעולת מערכת הבינה המלאכותית,[338] להיבט האתי – שאלת האחריותיות לפעולתה.[339] גם כשמערכת בינה מלאכותית פועלת לכאורה בשקיפות מלאה ורבבות שורות הקוד שלה פתוחות לעיון הציבור, אין ערובה שהדיוטות יצליחו להבין כיצד היא פועלת ומי אחראי לתוצריה.[340]

ההבחנה בין שקיפות גרידא ובין מפורשוּת מופיעה במסמך המומחים של האיחוד האירופי (להלן: מסמך המומחים). על פי המסמך, עקרון המפורשוּת נמנה עם עקרונות־העל של בינה מלאכותית אתית. כדי שתתקיים מפורשוּת נדרשת שקיפות של תהליכי הבינה המלאכותית, היכולות והמטרות של מערכות בינה

334 ראו למשל National Science and Technology Council, **לעיל** ה"ש 301, בעמ' 41; European Commission 2020, **לעיל** ה"ש 300, בעמ' 11.

335 ראו המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 16; Margot E. Kaminski, *Understanding Transparency in Algorithmic Accountability*, *in* The Cambridge Handbook of the Law of Algorithms 121 (Woodrow Barfield ed., 2021)

336 ראו למשל PI&A19, **לעיל** ה"ש 325, בעמ' 28; OECD 2019, **לעיל** ה"ש 294, בעמ' 8.

337 Luciano Floridi et al., *Ai4people – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, 28 Minds and Machines 689 (2018)

338 עקרונות הבינה המלאכותית של חברת מיקרוסופט, למשל, מגדירות שקיפות כיכולת להבין מערכות בינה מלאכותית. ראו *Responsible AI*, Microsoft, **לעיל** ה"ש 316.

339 השוו לפירוק מושג השקיפות אצל Paula Boddington, Towards a Code of Ethics for Artificial Intelligence 29 (2017)

340 Christoph Bartneck, Christoph Lütge, Alan Wagner, and Sean Welsh, An Introduction to Ethics in Robotics and AI 36 (2021)

מלאכותית צריכים להיות גלויים וההחלטות שהן מקבלות צריכות להיות ניתנות להסבר. כשמדובר בקופסאות שחורות שאי־אפשר להסביר את פעולתן נדרשים אמצעים משלימים, בהם היכולת לבקר את המערכת (auditability)[341] ולעקוב אחר התהליכים בה (נעקבות, traceability).[342]

היבט ההסברתיות של עקרון השקיפות (או למצער היכולת לבקר מערכות בינה מלאכותית ולעקוב אחר התהליכים בהן) נועד, בין השאר, לזהות את הנזקים שהן עלולות להסב[343] ולאפשר בקרה אנושית על החלטות ממוכנות בעלות חשיבות.[344] היבט נוסף של שקיפות הוא אינטראקציה מיודעת עם בינה מלאכותית.[345] מערכות בינה מלאכותית נדרשות להציג את עצמן למשתמשים בתור כאלה ולא לגרום להם לחשוב בטעות שמדובר באנשים בשר ודם.[346] במסמך מדיניות הרגולציה של משרד החדשנות נכללים ההסברתיות והאינטראקציה המיודעת בעקרון השקיפות.[347] ב"מתווה האמריקאי למגילת זכויות בבינה מלאכותית" הם נכללים בעיקרון "יידוע והסבר".[348]

יש מסמכים שנדרשים להיבט הציבורי של שקיפות מערכות המידע. ההסתמכות על קוד פתוח מאפשרת פיתוח אלגוריתמים משותפים, פיקוח משותף רחב עליהם

341 ראו גם ACM 2017, **לעיל** ה"ש 326, בעמ' 2 (עיקרון 6).

342 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 13. על נעקבוּת ראו גם: AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense (Defense Innovation Board 2019) 8, עיקרון 3.

343 ראו *Asilomar AI Principles*, **לעיל** בה"ש 326, עיקרון 7. לביקורת ראו Boddington, **לעיל** ה"ש 339, בעמ' 107.

344 ראו *Asilomar AI Principles*, **לעיל** ה"ש 326, עיקרון 8; Commission Regulation 2016/679 of 27 Apr. 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 (EU), Article 22

345 ראו לדוגמה AI HELG 2019, **לעיל** ה"ש 298, בעמ' 18; ס' 32 להצהרת טורונטו, **לעיל** ה"ש 324; *Guidelines for Artificial Intelligence*, Deutsche Telekom (2018)

346 ראו לדוגמה ס' 52(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53; ס' 5(V)(a) להצעת החוק הברזילאית לבינה מלאכותית, **לעיל** ה"ש 54.

347 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 106.

348 AI BOR, **לעיל** ה"ש 305, בעמ' 40–45.

ואיגום משאבים. כך גם אפשר למנוע היווצרות של מונופולים בתחום – גופים החולשים לבדם על שירותי בינה מלאכותית מסוימים או על מידע.[349]

3.2 הוגנות

עקרון ההוגנות נידון אף הוא במסמכים רבים.[350] הטיפולוגיה של ז'ובן, ינקה וואיינה קושרת את ההוגנות גם לצדק,[351] ובפרט למניעה, לצמצום ולניטור של הטיות בלתי רצויות ואפליה.[352] במסמכים אלו ההיבטים הדומיננטיים של הוגנות הם אי־אפליה ומניעת הטיות,[353] אך עקרון ההוגנות מוזכר לעיתים גם בדיונים על שוויון,[354] צדק חלוקתי,[355] הכלה (inclusiveness)[356] וגיוון (diversity)[357] בתכנון מערכות הבינה המלאכותית, וכן בהסתמכות על נתונים אמינים ומייצגים ללמידת מכונה. כאשר מדובר במערכות נבונות המקבלות החלטות הרות גורל בעניין אזרחים, יש להבטיח כי אלו תפעלנה בהוגנות ובשוויון.[358]

349 ראו לדוגמה את עיקרון 1.7 אצל: Beijing Academy of Artificial Intelligence, **לעיל** ה"ש 310; OECD 2019, **לעיל** ה"ש 294, בעמ' 8.

350 100% מהמסמכים שדגמו .Fjeld et al, וכ־76% מהמסמכים שדגמו Jobin, Ienca and Vayena (**לעיל** ה"ש 292) מתייחסים לשקיפות.

351 ראו למשל .Floridi et al, **לעיל** ה"ש 337, בעמ' 698–699; הצהרת מונטריאול, **לעיל** ה"ש 323, עיקרון 6.

352 Jobin, Ienca and Vayena, **לעיל** ה"ש 292.

353 ראו Google, **לעיל** ה"ש 317, עיקרון 2; Defense Innovation Board, **לעיל** ה"ש 342, עיקרון 2; AI BOR, **לעיל** ה"ש 305, בעמ' 23–29.

354 למשל, בהצהרת טורונטו מצוין בפירוש "הצהרה זו מתמקדת בזכות לשוויון ולהיעדר אפליה" (ס' 12 להצהרת טורונטו, **לעיל** ה"ש 324), ובהמשכה מתוארים קווים מנחים שלאורם מדינות צריכות לפעול כדי לצמצם ולמתן את הסיכון לאפליה שמקורה בטכנולוגיות למידת מכונה הן במגזר הציבורי (ס' 30–41) הן במגזר הפרטי (ס' 42–51).

355 ראו Datenethikkommission, **לעיל** ה"ש 309, בעמ' 46–47.

356 ראו למשל Microsoft, *Responsible AI*, **לעיל** ה"ש 316.

357 Cutler and Pribić, **לעיל** ה"ש 318, בעמ' 32; *Sony Group AI Ethics Guidelines*, **לעיל** ה"ש 319, עיקרון 5. לדיון השוואתי בניסוח עקרון ההוגנות התוצאתית במסמכים שונים ראו .Fjeld et al, **לעיל** ה"ש 292, בעמ' 51.

358 ראו למשל National Science and Technology Council, **לעיל** ה"ש 301, בעמ' 34.

מסמך המומחים מבחין בין הוגנות מהותית להוגנות פרוצדורלית.[359] הוגנות מהותית משמעה מחויבות לחלוקה הוגנת ושווה של כל היתרונות הטמונים במערכות בינה מלאכותית ושל כל העלויות שלה, ולהיעדר הטיות לא־הוגנות, אפליה וסטיגמות. מסמך המומחים מדגיש בהקשר זה גם את עקרון המידתיות ואת חשיבות האיזון בין אינטרסים מתחרים. הוגנות פרוצדורלית, לעומת זאת, משמעה האפשרות לערער על החלטות של מערכות בינה מלאכותית ולקבל סעדים.

אימון או הפעלה של מערכות בינה מלאכותית עלולים להתבסס על בסיסי נתונים לא שלמים או כאלה שמשקפים הטיות היסטוריות או ממשל לא תקין. משום כך הם עלולים להביא לידי אפליה עקיפה של פרטים או קבוצות מסוימים.[360] מסמך המומחים מונה אפוא דרכים שונות למימוש בינה מלאכותית ראויה לאמון, ובהן גיוון, אי־אפליה והוגנות.[361] דרכים אחרות הן שיתוף מגוון בעלי עניין בתהליכי התכנון, ההטמעה וההפעלה השוטפת של מערכות בינה מלאכותית, וכן הבטחת נגישות של מגוון משתמשים למערכות אלו. מובן שפרקטיקות לא הוגנות, כגון אפליה אלגוריתמית בתמחור צרכנים,[362] יכולות להיות גם מכוונות, אך הן אינן מענייננו כאן.

דוח ועדת המשנה נדרש להיבטים של שוויון תוצאתי וממליץ להימנע ממצבים שבהם תוצרי הבינה המלאכותית מעצימים רק קבוצות מסוימות באוכלוסייה. דוח זה, כמו מסמך המומחים, נדרש גם לצורך לייצג מגוון אוכלוסיות – הן בתוך מאגרי המידע שהמערכות מסתמכות עליהן הן בשלבי הפיתוח שלהן – ורואה בו היבט חשוב של עקרון ההוגנות.[363] על פי מסמך מדיניות

359 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 12.

360 ראו למשל HRC 2018, **לעיל** ה"ש 270, בפס' 38; #BigData: *Discrimination in Data-Supported Decision Making*, European Union Agency For Fundemental Rights (30.5.2018)

361 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 18.

362 Michal S. Gal and Niva Elkin-Korren, *Algorithmic Consumers*, 30 Harv. J. L. & Tech. 309 (2017)

363 ראו המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 15.

הרגולציה של משרד החדשנות, "שוויון ומניעת אפליה פסולה" הוא עיקרון אתי עצמאי.[364]

3.3 מניעת נזקים ובטיחות

בעיקרון זה נשמעים במובן מסוים הדיו של חוק הרובוטיקה הראשון של אסימוב (וחוק האפס),[365] וכן של עקרון היסוד של האתיקה הרפואית, *primum non nocere* (ראשית, אל תזיק).[366] אחת מהמלצות דוח הוועדה של בית הלורדים על בינה מלאכותית היא שלא להעניק למערכות בינה מלאכותית סמכות אוטונומית לפגוע, להשמיד או להטעות בני אדם.[367] עקרונות אסילומר נוקטים לשון של היעדר חתרנות (non-subversion): סמכויות השליטה במערכות בינה מלאכותית אל להן לחתור תחת תהליכים חברתיים ואזרחיים החיוניים לבריאותה של החברה.[368] מסמך ההמלצות של משרד ההגנה של ארצות הברית מדבר על משילות (governable AI): עיצובן של מערכות נבונות צריך לכלול את היכולת לזהות נזקים לא צפויים ולמנוע אותם.[369] המתווה האמריקאי למגילת זכויות בבינה מלאכותית כולל בעקרונותיו "מערכות בטוחות ויעילות".[370]

364 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 105.

365 אסימוב, **לעיל** ה"ש 47, בעמ' 6. ראו גם Robin Murphy and David D. Woods, *Beyond Asimov: The Three Laws of Responsible Robotics*, 24 IEEE Intelligent Systems 14 (2009).

366 Tom L. Beauchanp and James F. Childress, Principles of Biomedical Ethics, 150–201 (7th ed., 2013); גיל סיגל "ביו־אתיקה בעולם המערבי: ישראל שבין אירופה לארה"ב" **הרפואה** 142, 143 (2004). ראו גם התייחסות מפורשת למקור הביו־אתי של הכלל אצל: Floridi et al., **לעיל** ה"ש 337, בעמ' 697.

367 House of Lords Select Committee on Artificial Intelligence, **לעיל** ה"ש 307, בפס' 417(5).

368 *Asilomar AI Principles*, **לעיל** ה"ש 326, עיקרון 17.

369 Defense Innovation Board, **לעיל** ה"ש 342, עיקרון 5.

370 AI BOR, **לעיל** ה"ש 305, בעמ' 23–29.

ז'ובן, ינקה וואיינה מציינים שעקרון מניעת הנזק שכיח יותר מעקרון עשיית הטוב וקושרים בינו ובין עקרון הבטיחות.[371] המיפוי של פיילד ועמיתיה מזכיר את עקרון מניעת הנזק רק במשתמע ומתמקד בבטיחות ובביטחון.[372] ההתמקדות בהיבטים אלו מסייעת לקונקרטיזציה של עקרון מניעת הנזק וליישומו.[373] על פי מסמכים אחדים, ניהול סיכונים[374] ויכולת חיזוי (predictability)[375] הם פנים נוספים של בטיחות במערכות נבונות.

מסמך המומחים נדרש לקונקרטיזציה של עקרון מניעת הנזק. המסמך מסביר שלמנוע נזק פירושו לוודא שמערכות נבונות אינן גורמות נזקים חדשים ואינן מחריפות נזקים קיימים – בין שמדובר בנזקים פרטיים ובין שמדובר בנזקים קבוצתיים, לרבות נזקים לא מוחשיים לסביבות פוליטיות, חברתיות ותרבותיות;[376] ואף אינן משפיעות לשלילה בכל דרך אחרת על בני אדם. הדבר מחייב שמירה על שלומם הנפשי והפיזי של בני אדם ועל כבודם.

מסמך המומחים נדרש גם להיבטים של בטיחות וביטחון ומציין כי מערכות נבונות צריכות להיות חסינות מבחינה טכנית – עליהן לפעול באופן בטוח, מאובטח ואמין, ואל להן לגרום לנזק בלתי מכוון.[377] יש לתת את הדעת גם על

371 ראו Jobin, Ienca and Vayena, **לעיל** ה"ש 292, בעמ' 394. כ־71% מהמסמכים שהם דגמו מתייחסים לעיקרון של מניעת נזקים, לעומת 41% שמתייחסים לעיקרון של עשיית הטוב.

372 .Fjeld et al, **לעיל** ה"ש 292, בעמ' 37. במונח ביטחון נדרשת ה־Datenethikkommission (**לעיל** ה"ש 309, בעמ' 45) להיבטים של אבטחת סייבר, הגנת מידע ובטיחות פיזית ונפשית, וכן להיתכנות של נזקים קינטיים־ממשיים בשל פעילות מערכות בינה מלאכותית.

373 כפי שמעירה Boddington, **לעיל** ה"ש 339, בעמ' 107. טיוטת כללי האתיקה של ממשלת אוסטרליה לבינה מלאכותית כללה את העיקרון של מניעת נזק: Artificial Intelligence: Australia's Ethics Framework (a discussion paper, Australian Government (2019); אך הוא נעדר מכללי האתיקה הסופיים שפורסמו לאחר מכן. הכללים שפורסמו, לעומת זאת, כוללים עקרונות של בטיחות וביטחון: *Australia's AI Ethics Princples*, Department of Industry, Science and Resources (2020)

374 ראו Beijing Academy of Artificial Intelligence, **לעיל** ה"ש 310.

375 AI HELG 2019, **לעיל** בה"ש 298, בעמ' 22.

376 **שם**, בעמ' 12.

377 ראו גם Google, **לעיל** ה"ש 317, עיקרון 3; *Asilomar AI Principles*, **לעיל** ה"ש 326, עיקרון 6; OECD 2019, **לעיל** ה"ש 294, בעמ' 8.

אבטחת סייבר – חוסנן של המערכות נמדד בין השאר בעמידותן בפני פצחנים המבקשים להשתמש בהן לרעה.[378] דוח ועדת המשנה נדרש להיבטים של בטיחות וביטחון (הגנת סייבר ואבטחת מידע) בנפרד.[379] כך גם מסמך ההמלצות היפני.[380] מסמך מדיניות הרגולציה של משרד החדשנות כולל המלצה על עיקרון אתי דומה: "אמינות, עמידות, אבטחה ובטיחות".[381]

3.4 אחריות ואחריותיות

שאלת האחריותיות של מערכות נבונות משיקה לשאלת האחריות של אלגוריתמים.[382] טכנולוגיות של למידת מכונה עלולות להגביר את רמת חוסר היכולת לצפות את פעולתן של מערכות אלו בהשוואה למערכות שמסתמכות על פיתוח אנושי,[383] אך אלו ואלו מעלות שאלות זהות בדבר ייחוס אחריות להחלטותיהן.

מסמכי מדיניות של בינה מלאכותית מזכירים תדיר עקרונות של אחריות ואחריותיות.[384] ואולם במקרים רבים אין הם מוגדרים די הצורך,[385] ומסמכים

378 ראו למשל National Science and Technology Council, **לעיל** ה"ש 301, בעמ' 36–37; Datenethikkommission, **לעיל** ה"ש 309, בעמ' 45; US President, **לעיל** ה"ש 302, בס' 2(d).

379 ראו המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 21.

380 The Conference toward AI Network Society, **לעיל** ה"ש 312, בעמ' 7 (עקרונות מס' 4–5).

381 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 107–108.

382 לשאלת האחריותיות של מערכות מחשב, ראו Helen Nissenbaum, *Accountability in a Computerized Society*, 2 Science & Engineering Ethics 25 (1996); Joshua A. Krolle et al., *Accountable Algorithms*, 165 U. Pa. L. Rev. 633 (2016-2017)

383 Paul B. de Laat, *Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?*, 31 Philosophy & Technology 525 (2018)

384 97% מהמסמכים שדגמו .Fjeld et al, וכ־71% מהמסמכים שדגמו Jobin, Ienca and Vayena (**לעיל** ה"ש 292), מתייחסים לאחריותיות או אחריות.

385 ראו הערתם של Jobin, Ienca and Vayena, **לעיל** ה"ש 292, בעמ' 394. FAT/ML, **לעיל** ה"ש 326, מציינים במפורש שהם נקטו לשון כללית ולא מוגדרת כדי לאפשר יישום רחב של העקרונות שהם מציעים.

שונים ממקמים אותם ברמות שונות בהיררכיה. כך למשל, דוח ועדת המשנה נדרש לרכיבים שונים של אחריותיות – שקיפות, הסברתיות ואחריות – בלי להגדיר אחריותיות מהי.[386] לעומת זאת, פלורידי ועמיתיו רואים באחריותיות רכיב של הסברתיות.[387] יש גם הקושרים בין אחריותיות לעקרון ההוגנות[388] או לעקרון השקיפות.[389]

לפי עמדת הארגון FAT/ML (Fairness, Accountability, and Transparency in Machine Learning), נקודת המוצא של אתיקה של מערכות נבונות היא שאי־אפשר לתלות את האשמה לפעולות של אלגוריתמים באלגוריתמים, אלא רק בבני אדם.[390] ערכים אתיים כגון אחריות, הסברתיות או הוגנות נועדו לסייע בזיהוי הגורמים שיש לייחס להם אחריותיות לתוצאות ההחלטות של מערכות נבונות. אומנם על פי מסמכים אחרים אחריותיות אינה העיקרון המארגן של אתיקה של מערכות נבונות, אך גם הם רואים בו את אחד העקרונות שלה ולרוב מצביעים על הצורך באחריותיות אנושית לתוצאות של פעילות מערכות לבינה מלאכותית.

מסמך האיגוד למכונות חישוביות גורס כי על מוסדות להיות אחראים להחלטות המתקבלות על ידי האלגוריתמים שהם משתמשים בהם גם כשאי־אפשר לספק הסבר לפעולתם.[391] הצהרת מונטריאול מזהירה מהפחתת האחריות האנושית עקב שימוש במערכות נבונות.[392] מסמך המלצות של משרד ההגנה של ארצות הברית קובע שעל בני אדם לקבל אחריות על הפיתוח של מערכות נבונות ועל תוצאות פעולתן.[393] טיוטת ההנחיות למפתחי מערכות נבונות שפרסם המשרד

386 ראו המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 17-16.

387 Floridi et al., **לעיל** ה"ש 337, בעמ' 700-698.

388 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 19.

389 ס' 50 להצהרת טורונטו, **לעיל** ה"ש 324.

390 FAT/ML, **לעיל** ה"ש 326.

391 ראו ACM 2017, **לעיל** ה"ש 326, עיקרון 3.

392 הצהרת מונטריאול, **לעיל** ה"ש 323, עיקרון 9: "אחריות: אל לו לפיתוח מערכת בינה מלאכותית להפחית מאחריותם של בני אדם במקום שבו יש לקבל החלטות".

393 Defense Innovation Board, **לעיל** ה"ש 342, עיקרון 1.

לענייני פנים ותקשורת של יפן מציינת כי המפתחים של מערכות נבונות נושאים באחריות (accountable) כלפי בעלי עניין שונים, לרבות המשתמשים בהן.[394] הקווים המנחים של חברת דויטשה טלקום מכריזים בפשטות "אנחנו אחראים".[395] ה־OECD, לעומת זאת, נמנע מלהצביע על הגורם שלו מיוחסת אחריותיות וקובעים כי "שחקני בינה מלאכותית", כלומר "מי שממלאים תפקיד פעיל במחזור החיים של מערכות בינה מלאכותית, לרבות פרישׂתן או תפעולן",[396] צריכים לשאת באחריות לתפקודן התקין של מערכות בינה מלאכותית (בהתחשב בתפקידיהם, בהקשר ובחזית הידע).[397]

מסמך המומחים אינו מסביר מהם אחריות ואחריותיות, אלא מונה כמה תנאים להתקיימותן:[398] יכולת לבקר את המערכת (auditability),[399] מזעור השפעות שליליות ודיווח עליהן, גישה רציונלית לאילוצי שקלול תמורות (trade-offs) ביישום עקרונות אתיים, ואפשרות לסעד כאשר פעולת מערכות הבינה המלאכותית אינה תקינה.[400]

מסמך מדיניות הרגולציה של משרד החדשנות נמנע אף הוא מלבאר את המושג ולפיו "מפתחי בינה מלאכותית, מפעיליה או המשתמשים בה יגלו אחריות לתפקודה התקין, ולקיום העקרונות האחרים בפעילותם".[401] לפי מסמך זה, עקרון האחריותיות "נועד לבסס את קיומו של 'גורם אחראי' לבינה מלאכותית [...].

394 The Conference toward AI Network Society, **לעיל** ה"ש 312, עיקרון 9.

395 Deutsche Telekom, **לעיל** ה"ש 345, בס' 1.

396 OECD 2019, **לעיל** ה"ש 294, בעמ' 7.

397 **שם**, בעמ' 8. ראו גישה דומה בהצהרת הכנס הבינלאומי השנתי של נציבי הגנת הפרטיות בעולם משנת 2018: Declaration on Ethics and Data Protection in Artificial Intelligence, 4 International Conference of Data Protection and Privacy Commissioners (ICDPP) (2018)

398 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 19–20.

399 **שם**. ראו גם ACM 2017, **לעיל** ה"ש 326, עיקרון 6.

400 לעוד מופעים של עקרון הסעד ראו **לעיל** ה"ש 317, עיקרון 2; FAT/ML, **לעיל** ה"ש 326; HRC 2018, **לעיל** ה"ש 270, בפס' 39–41.

401 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 108.

מערכות בינה מלאכותית נוצרות בידי אדם, ולכן כברירת מחדל האדם הוא שאחראי להן".[402]

3.5 פרטיות

מופע ראשוני של רגולציית מערכות נבונות אפשר למצוא בדיני הגנת המידע האירופיים, הכוללים את הזכות שלא להיות מושא להחלטות אוטומטיות.[403] הקשר בין פרטיות למערכות בינה מלאכותית, ובפרט למערכות למידה נבונה, המסתמכות על מידע רב, אינו מפתיע.[404] ואכן, במסמכים רבים נידון נושא הפרטיות והצורך בשמירה עליה במסגרת הדיון בעקרונות האתיים הבסיסיים של הבינה המלאכותית.[405] מסמך המומחים מונה שלושה היבטים של ההגנה על הפרטיות במערכות נבונות: שמירה על פרטיות והגנת מידע לאורך כל מחזור חיי מערכת הבינה המלאכותית; איכותו ושלמותו של המידע; והגבלת הרשאות הגישה למידע במערכות נבונות.[406] לפי המתווה האמריקאי למגילת זכויות בבינה מלאכותית, יש לעצב ככל האפשר מערכות אוטומטיות לפי עקרונות של הנדסת פרטיות ולהתבסס על הסכמה של מושאי המידע; לפי עיקרון זה, אין לבצע מעקב רציף (continuous surveillance) במקרים שבהם טכנולוגיות מעקב עלולות להגביל זכויות, הזדמנויות או גישה.[407]

402 **שם**, שם.

403 ראו ס' 22 של התקנות הכלליות בדבר הגנת מידע (GDPR), שאת שורשיו אפשר לאתר בס' 15 לדירקטיבת הגנת המידע, Directive 95/46/EC of 24 October 1995 on the Protection of Individuals with regard to the Protection of Personal Data and on the Free Movement on such Data [1995] OJ L'281/31. ראו גם רחל ארידור הרשקוביץ ותהילה שוורץ אלטשולר **הצעת חוק הגנת הפרטיות התשע"ט 2019** 69–71 (המכון הישראלי לדמוקרטיה 2019); כהנא ושני, **לעיל** ה"ש 213, בעמ' 263–265. ברמת החקיקה הלאומית, כבר ב־1978 נדרשו דיני הגנת הפרטיות של צרפת לעיבוד אוטומטי – ראו ס' 2–3 אצל Loi no. 78-17 du 6. janvier 1978 relative à l'informatique, aux fichiers et aux libertés

404 ראו למשל HRC 2018, **לעיל** ה"ש 270, בפס' 34.

405 97% מהמסמכים שדגמו .Fjeld et al, וכ־56% מהמסמכים שדגמו Jobin, Ienca and Vayena (**לעיל** ה"ש 292), מתייחסים לפרטיות.

406 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 17.

407 AI BOR, **לעיל** ה"ש 305, בעמ' 30–39.

לעיתים פרטיות מוזכרת כערך שיש לכבד או לקדם[408] ולעיתים כזכות שיש להגן עליה.[409] יש מסמכים שקושרים בין פרטיות להגנת מידע;[410] ויש אחרים שקושרים בין פרטיות לשליטה ומציינים את זכותו של המשתמש לשלוט במידע שלו.[411] מסמכים רבים מדגישים את החשיבות של הנדסת פרטיות (או "עיצוב לפרטיות", privacy by design), ושל פרטיות כברירת מחדל (privacy by default),[412] כלומר הטמעה יזומה של עקרונות הגנת הפרטיות כבר בשלב התכנון של מערכות הבינה המלאכותית כפעולת מנע, ולא רק לאחר מעשה.[413] במסמך מדיניות הרגולציה של משרד החדשנות מוטמע עקרון הפרטיות בעיקרון שכותרתו "האדם במרכז". לפי עיקרון זה, "פיתוח בינה מלאכותית, או שימוש בה, ייעשו תוך כיבוד שלטון החוק, זכויות יסוד ואינטרסים ציבורים, ובפרט תוך שמירה על כבוד האדם ופרטיות".[414]

כדי לטפח אמון בבינה מלאכותית, המסתמכת על נתוני עתק, יש להבטיח שמירה על פרטיות.[415] יש מתח בין פרטיות ובין שקיפות, שהוא עיקרון מרכזי

408 ראו Partnership on AI, **לעיל** ה"ש 320; Microsoft, *Responsible AI*, **לעיל** ה"ש 316.

409 ראו *Sony Group AI Ethics Guidelines*, **לעיל** ה"ש 319, עיקרון 4; *Asilomar AI*, **לעיל** ה"ש 326, עיקרון 13; The Conference toward AI Network Society, **לעיל** ה"ש 312, עיקרון 6; HRC 2018, **לעיל** ה"ש 270, בפס' 33–35.

410 ראו Cutler and Pribić, **לעיל** ה"ש 318, בעמ' 40; *Sony Group AI Ethics Guidelines*, **לעיל** ה"ש 319, עיקרון 4; ראו Google, **לעיל** ה"ש 317, עיקרון 5; Intel, **לעיל** ה"ש 315.

411 ראו Datenethikkommission, **לעיל** ה"ש 309, בעמ' 45; Cutler and Pribić, **לעיל** ה"ש 318, בעמ' 40; *Asilomar AI Principles*, **לעיל** ה"ש 326, עיקרון 12.

412 בשנת 2010 אומצו בכנס נציבי הגנת הפרטיות בעולם עקרונות הנדסת הפרטיות והפרטיות כברירת מחדל. גם הדין האירופי הכיר בעקרון הנדסת הפרטיות (ס' 25 לתקנות הכלליות בדבר הגנת מידע, GDPR). להרחבה ראו ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 403, בעמ' 89–90; ראו גם מיכאל בירנהק "פרטיות במשבר: הנדסה חוקתית והנדסת פרטיות" **משפט וממשל** כד 149 (2022), בחלק ה.

413 ראו Google, **לעיל** ה"ש 317, עיקרון 5; Intel, **לעיל** ה"ש 315, בעמ' 7–8; Datenethikkommission, **לעיל** ה"ש 309, בעמ' 123.

414 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 104.

415 ראו למשל National Science and Technology Council, **לעיל** ה"ש 301, בעמ' 20; US President, **לעיל** ה"ש 302, בס' 1(d).

אחר של אתיקת הבינה המלאכותית. שקיפות אלגוריתמית עלולה לעיתים להתנגש בעקרונות של פרטיות, שכן בחינת תפקודן של מערכות בינה מלאכותית מחייבת פעמים רבות גישה לנתונים שאימנו את האלגוריתם.[416] כך למשל, מודל השפה של ChatGPT פותח בין השאר בהסתמך על בסיסי נתונים הכוללים מידע על אנשים פרטיים. גם אם מידע זה נלקח ממקורות גלויים ברשת הוא עלול, בשל הצטברותו, להסגיר מידע רגיש אגב שיחה עם המודל.[417] יתר על כן, הטמעת נתונים אלו במודל עלולה להתנגש בזכות להישכח (the right to be forgotten).[418]

3.6 קידום הטוב

מקצת המקורות מונים את קידום הטוב (beneficence) עם התכליות האתיות של בינה מלאכותית.[419] קידום הטוב מופיע בראש חמשת העקרונות שניסחו פלורידי ועמיתיו,[420] ומקורו בין השאר בכללי האתיקה הרפואית.[421] בטיוטות המוקדמות של מסמך המומחים נכלל עיקרון זה במפורש,[422] אך אחר כך הוא הושמט.[423] עם זאת, מסמך המומחים בגרסתו הנוכחית מונה "רווחה חברתית וסביבתית" עם הדרכים למימוש בינה מלאכותית אמינה.[424] גם

416 ראו לדוגמה ACM 2017, **לעיל** ה"ש 326, עיקרון 5; המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 19.

417 Melissa Heikkilä, *What Does GPT-3 "Know" About Me?* MIT Technology Review (31.8.2022)

418 Eric Wallace et al., *Does GPT-2 Know Your Phone Number?* BAIR (20.12.2020)

419 כ־48% מהמסמכים שדגמו Jobin, Ienca and Vayena (**לעיל** ה"ש 292) מתייחסים לעקרון קידום הטוב (beneficence).

420 Floridi et al., **לעיל** ה"ש 337, בעמ' 697–698.

421 ראו לדוגמה Beauchanp and Childress, **לעיל** ה"ש 366, בעמ' 202–248.

422 AI HELG 2018, **לעיל** ה"ש 298, בעמ' 8–9.

423 Andrea Renda, *Europe: Toward a Policy Framework for Trustworthy AI*, *in* The Oxford Handbook of Ethics in AI 651, 656 (2020)

424 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 21.

הצהרת מונטריאול משתמשת במונח רווחה,[425] ואילו עקרונות אסילומר ודוח הוועדה של בית הלורדים לנושא בינה מלאכותית נוקטים את המונח "טוב משותף".[426] יש מסמכים שמדברים על קידום הטוב עבור "כל יצור תבוני",[427] עבור "בני אדם רבים ככל האפשר"[428] או לצורך "בניית חברה טובה יותר";[429] ואילו מסמכים שמקורם בחברות פרטיות מדגישים לעיתים את קידום הטוב עבור לקוחותיהם.[430]

צמיחה בת קיימא היא העיקרון הראשון המוזכר בעקרונות האתיקה של מסמך מדיניות הרגולציה של משרד החדשנות. המסמך מעמיד בראש עקרונות האתיקה שהוא מציע את העיקרון של "בינה מלאכותית לקידום צמיחה, פיתוח בר קיימא ומובילות ישראלית בחדשנות". כאן למסגור הכלכלי של עקרון קידום הטוב יש גם היבט של חוסן לאומי ("יש לתת את הדעת לחשיבות של שימור עוצמה טכנולוגית ויכולת טכנולוגית").[431] עם זאת, ספק אם "מובילות ישראלית בחדשנות" היא אכן עיקרון אתי.[432]

425 הצהרת מונטריאול, **לעיל** ה"ש 323, עיקרון 1; OECD 2019, **לעיל** ה"ש 294, בעמ' 7.

426 House of Lords Select Committee on Artificial Intelligence, **לעיל** ה"ש 307, בפס' (1)417.
עקרונות אסילומר גורסים כי יש לפתח תבונת־על (superintelligence) רק לאור אידאלים אתיים מקובלים (widely shared). ראו Asilomar AI Principles, **לעיל** ה"ש 326, עיקרון 23. בודינגטון מעירה, ובצדק, כי נוסח זה עמום ותלוי באופן שבו אנו מזהים "אידיאלים אתיים מקובלים". Boddington, **לעיל** ה"ש 339, בעמ' 110.

427 הצהרת מונטריאול, **לעיל** ה"ש 323, עיקרון 1.

428 ראו Asilomar AI Principles, **לעיל** ה"ש 326, עיקרון 14 (לביקורת ראו Boddington, **לעיל** ה"ש 339, בעמ' 109); Partnership on AI , **לעיל** ה"ש 320.

429 ראו *Sony Group AI Ethics Guidelines*, **לעיל** ה"ש 319, עיקרון 1; ראו גם Google, **לעיל** ה"ש 317, עיקרון 1.

430 ראו למשל Deutsche Telekom, **לעיל** ה"ש 345, בס' 3.

431 מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 103–104.

432 נוסח העיקרון מזכיר את ס' 2(ב) לתזכיר חוק הגנת הסייבר, **להלן** ה"ש 461, שמונה עם מטרות מערך הסייבר את "קידום ישראל כמובילה עולמית בתחום הסייבר". הגדרת מובילות גלובלית בתחום טכנולוגי מסוים כמטרה מוסדית ועיקרון אתי גם יחד מעוררת תמיהה מסוימת, שכן עקרונות אתיים רכים או מטרות סטטוטוריות של גופים מוסדיים הם חלק מהמסגרת הפרשנית התוחמת את מרחב הסמכויות של הרגולטור.

3.7 האדם במרכז, חירות ואוטונומיה

עקרונות של חירות ואוטונומיה, ובהם גם עקרונות המעמידים את האדם במרכז, נכללים בכמה ממסמכי המדיניות והאתיקה של בינה מלאכותית.[433] לפי מסמכים אלו, בינה מלאכותית צריכה לקדם את חירויות הפרט (כגון חופש הביטוי,[434] הגדרה עצמית מידעית או אחרת[435] ואת החופש מפני מניפולציה[436]), לשמור על חירויות אלו ולדאוג לאוטונומיה שלהן.[437] תמות של חירות ואוטונומיה מבטאות בין השאר גישה אנתרופוצנטרית, המעמידה את האדם במרכז.[438] גישה זו באה לידי ביטוי בבקרה אנושית על פעולות אוטומטיות,[439] במתן האפשרות שלא להשתמש במוצרים המבוססים עליה[440] או בקריאה כללית לשליטה אנושית בטכנולוגיות.[441]

אוטונומיה היא עיקרון יסוד באתיקה רפואית,[442] ואחד העקרונות המארגנים של אתיקת הבינה המלאכותית שניסחו פלורידי ועמיתיו.[443] אוטונומיה היא זכותו

433 69% מהמסמכים שדגמו .Fjeld et al כוללים התייחסויות להיבטים של שליטה אנושית בטכנולוגיה. כ־40% מהמסמכים שדגמו Jobin, Ienca and Vayena מתייחסים לחירות ואוטונומיה (**לעיל** ה"ש 292).

434 PI&A19, **לעיל** ה"ש 325; HRC 2018, **לעיל** ה"ש 270.

435 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 12; Datenethikkommission, **לעיל** ה"ש 309, בעמ' 43–44. להגדרה עצמית מידעית (informational self-determination) ראו (1983) BverfGE 65, 1, וכן כהנא ושני, **לעיל** ה"ש 213, בעמ' 187.

436 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 10.

437 **שם**, בעמ' 12.

438 OECD 2019, **לעיל** ה"ש 294, בעמ' 7; ראו Asilomar AI Principles, **לעיל** ה"ש 326, עיקרון 11 (לביקורת ראו Boddington, **לעיל** ה"ש 339, בעמ' 108); HRC 2018, **לעיל** ה"ש 270, בפס' 47–52.

439 ראו ס' 22 של התקנות הכלליות בדבר הגנת מידע (GDPR); AI HELG 2019, **לעיל** ה"ש 298, בעמ' 12; AI BOR, **לעיל** ה"ש 305, בעמ' 46–52.

440 House of Lords Select Committee on Artificial Intelligence, **לעיל** ה"ש 307, בעמ' 27.

441 Asilomar AI Principles, **לעיל** ה"ש 326, עיקרון 16 (לביקורת ראו Boddington, **לעיל** ה"ש 339, בעמ' 109).

442 ראו לדוגמה Beauchanp and Childress, **לעיל** ה"ש 366, בעמ' 101–149. בדין הישראלי, ראו ע"א 2781/93 **דעקה נ' בית החולים "כרמל"**, חיפה, פ"ד נג(4) 526, 571–573 (1999).

443 .Floridi et al, **לעיל** ה"ש 337, בעמ' 697–698.

של הפרט להחליט עבור עצמו, ודוח ועדת המשנה אינו נדרש רק ליכולת הפיזית של אדם לבחור בין אפשרויות, אלא גם להיבטים אפיסטמולוגיים המאפשרים בחירה מושכלת.[444] הצהרת מונטריאול קוראת לפתח מערכות בינה מלאכותית מתוך כיבוד האוטונומיה האנושית במטרה להגביר את יכולתם של פרטים לשלוט בחייהם.[445] על פי המלצת דוח ועדת בית הלורדים, אין להעניק למערכות בינה מלאכותית את הזכות האוטונומית לבחור "לפגוע, להשמיד או להטעות בני אדם".[446] מסמך המומחים של האיחוד האירופי רואה בכיבוד האוטונומיה של האדם וחירויותיו את אחת התכליות של זכויות היסוד שעליהן נכון האיחוד. משכך, על מערכות בינה מלאכותית לאפשר למשתמשים בהן לממש באופן רציף את זכותם להגדרה עצמית ולהשתתפות בתהליכים הדמוקרטיים.

עוד עולה ממסמכים שונים שבינה מלאכותית אמורה לקדם סולידריות אנושית[447] או ערכים של קיימות ושל שמירה על הסביבה.[448] הטמעתם של עקרונות אתיים בפיתוח, בפרישׂה, בהטמעה ובתפעול של מערכות נבונות נועדה לטפח אמון ציבורי במערכות אלו ולקדם את הטוב הכללי (תהיה הגדרתו אשר תהיה).

מסמך מדיניות הרגולציה של משרד החדשנות מונה עם המלצותיו לעקרונות אתיים את העיקרון "האדם במרכז – כיבוד זכויות יסוד ואינטרסים ציבוריים".[449] לפי עיקרון זה, "בעת פיתוח ושימוש בטכנולוגיות חדשניות, האדם ויחסי הגומלין שלו עם החברה בה הוא פועל, מהווים נקודת מכוון נורמטיבית". בעיקרון זה כלולים, לפי מסמך זה, אינטרסים ציבוריים, עקרון שלטון החוק וזכויות יסוד (לרבות כבוד האדם ופרטיות). את כל אלו יש לכבד אגב פיתוח בינה מלאכותית ושימוש בה.

444 ראו המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 20.

445 הצהרת מונטריאול, **לעיל** ה"ש 323, עיקרון 2.

446 House of Lords Select Committee on Artificial Intelligence, **לעיל** ה"ש 307, בפס' 417(5). ראו גם **להלן** בסעיף 5.7.

447 Miguel Luengo-Oroz, *Solidarity Should Be a Core Ethical Principle of AI*, 1 Nature Machine Intelligence 494 (2019).

448 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 12; .Floridi et al, **לעיל** ה"ש 337, בעמ' 696–697. ראו גם ס' 3(6) להצעת חוק הבינה המלאכותית הברזילאית, **לעיל** ה"ש 54.

449 מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 104–105.

3.8 סיכום

הספרות הענפה על אתיקה של בינה מלאכותית, שנסקרה בפרק זה, צמחה מן התפיסה שטכנולוגיה חדשה זו איננה מאוסדרת, ואף אין הסכמה אוניברסלית בעניין דרכי אסדרתה, ולכן נכון ליצור עבודה מסגרת של כללי אתיקה לא פורמליים. לצד זאת עלתה גם הטענה שהיצמדות לכללי אתיקה היא בבחינת "טיוח אתי" (ethics washing), כלומר ניסיון ליצור מראית עין של אחריות אגב התחמקות ממחויבות מעשית לאחריות זו.[450] מהלך זה כולל הימנעות מכוונת מהגדרות בהירות, מוחשיות ומפורטות של העקרונות האתיים; אי־קביעת מנגנוני אכיפה ברורים, שענישה בצידם, במקרה של אי־ציות לכללי האתיקה; היעדר התייחסות למגוון החוליות בשרשרת הערך של יצירת המערכות והמוצרים; והתמקדות בעיקר בחוליה של המהנדסים ושל כותבי הקוד.

לטיוח אתי כמה תכליות. תכלית אחת היא לחזק את אמון המשתמשים במוצרים מבוססי בינה מלאכותית (ואחת היא אם מדובר באמון בעניין בטיחות המוצרים, בהוגנות הליך הייצור שלהם או בהיבט אחר). תכלית אחרת היא לשחרר את התעשייה ממגבלות ולאפשר התחמקות מרגולציה ממשלתית או בינלאומית מחייבת. אחת התוצאות של הפרקטיקה הזאת היא האפשרות לבחור לאילו כללי אתיקה לציית ומאילו להתעלם.

אף על פי כן, זלזול יתר בכללי האתיקה (מה שמכונה לעיתים ethics bashing), מתוך תפיסה שטקסטים אלו אינם שווים את הנייר שעליו הם כתובים ואין ביכולתם להתמודד עם המציאות שבה מתפתחת הבינה מלאכותית, גם הוא אינו מוצדק. ראשית, כללי אתיקה יכולים לשמש גשר במצב הביניים שבין התפרצות של טכנולוגיה לבין יצירת רגולציה מדינתית או בינלאומית מותאמת. שנית, מחויבות מוסרית של עובדים בתוך תאגידים ושל תאגידים ללקוחות ולמשתמשים יכולה לשמש לעיתים שוט חזק לא פחות מרגולציה ממשלתית מחייבת במודל של ציווי ושליטה (command and control).[451]

450 Meredith Whittaker et al., AI Now Report 2018 29–30 (AI Now, 2018)

451 Elettra Bietti, *From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy*, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (January 2020)

פרק רביעי

רגולציה של בינה מלאכותית: סקירה השוואתית

לצד מסמכי המדיניות שתוארו לעיל הגיעה בשנת 2021 גם סנונית ראשונה של רגולציית בינה מלאכותית – הצעת תקנות הבינה המלאכותית האירופיות.[452] אורסולה פון דר ליין, נשיאת הנציבות האירופית, הכריזה עם בחירתה ביולי 2019 שבתוך מאה ימים היא תפרסם הצעת חוק בתחום הבינה המלאכותית,[453]

452 ראו **לעיל** ה"ש 53.

453 *Von der Leyen's Real 100-day Challenge*, Politco (28.11.2019)

וכעת הגיעה השעה. בהמשך השנה פורסמו הצעות חקיקה לאסדרת תחום הבינה המלאכותי גם בסין[454] ובברזיל.[455]

גם ארצות הברית אותתה על עניין רגולטורי בבינה מלאכותית בפרסום ההמלצות של ועדת הסחר הפדרלית (FTC) לשימוש הוגן בבינה מלאכותית.[456] המלצות אלו מצביעות על כמה מקורות סטטוטוריים שמהם הרשות הפדרלית יכולה לינוק את סמכותה לאסור הטיות בבינה מלאכותית ואף פרקטיקות נאותות (שקיפות, הוגנות, הסברתיות, חסינות ואחריותיות).[457]

4.1 תקנות הבינה המלאכותית האירופיות

נוסח ההצעה של תקנות הבינה המלאכותית האירופיות, שעתיד להיכנס לתוקף בשנת 2024 וגם לאחר פרסומו הראשוני יהיה כפוף לשינויים,[458] נועד להתמודד עם אחד האתגרים המשמעותיים העומדים לפני מחוקק שניגש לאסדר טכנולוגיה – אי־הוודאות בנוגע להתפתחות העתידית שלה. לנוכח קצב התפתחות הטכנולוגיה ואיטיות המחוקק – שנדרש ללמוד את הטכנולוגיות החדשות, להעריך בזמן אמת מה תהיה השפעתן על הסדר הקיים ולגזור מהערכה זו דפוס אסדרה אופטימלי – אסטרטגיה אפשרית להתמודדות עם אתגרים אלו היא אסדרה "חסינת עתיד"

454 ראו **להלן** בסעיף 4.2.

455 ראו **להלן** בסעיף 4.3.

456 Elisa Jillson, *Aiming for Truth, Fairness, and Equity in your Company's Use of AI*, Federal Trade Commission Business Blog (19.4.2021)

457 ראו גם Andrew Smith, *Using Artificial Intelligence and Algorithms*, Federal Trade Commission Business Blog (8.4.2021)

458 בשלבי הבאת הספר לדפוס, במאי 2023, אכן פורסם נוסח מתוקן של הצעת תקנות הבינה המלאכותית האירופיות (ראו ההצעה המתוקנת, **לעיל** בה"ש 57). ההפניות להצעת תקנות הבינה המלאכותית האירופיות הן להצעה המקורית, אך במקומות שבהם היא עודכנה אנחנו מפנים להצעה המתוקנת. ביוני 2023 אושר הנוסח המתוקן בפרלמנט האירופי. ראו רפאל קאהאן "הפרלמנט האירופי אישר טיוטת חוק לרגולציה על בינה מלאכותית" **Ynet** (14.6.2023).

(future proof).[459] אסדרה חסינת עתיד מאמצת דגמי חקיקה שאינם קשורים אך ורק לטכנולוגיה מסוימת אלא מבוססים על עקרונות־על גמישים שאפשר להתאימם להתפתחויות עתידיות. כאלה הם, למשל, חוק סמכויות החקירה של בריטניה[460] או הדגם שהוצע בתזכיר חוק הגנת הסייבר בישראל.[461]

נוסח התקנות האירופיות התבסס על מסמכים קודמים[462] ומטרתו לעצב מסגרת משפטית רגולטורית לבינה מלאכותית אמינה (trustworthy). הוא מגלם בתוכו זכויות יסוד אירופיות ונועד לאפשר לאנשים להשתמש בביטחון במערכות בינה מלאכותית, ובה בעת לעודד עסקים לפתח מערכות כאלה. בדברי ההסבר לנוסח התקנות האירופיות נכתב שמערכות בינה מלאכותית נועדו להיות כלי לקידום הטוב החברתי ולהביא לשגשוגם ורווחתם של בני אדם; משכך, הכללים החלים על מערכות נבונות צריכים להתרכז באדם.

רצוי להזכיר כי נוסף על נוסח התקנות האירופיות, שיידון להלן בהרחבה, גם התקנות הכלליות בדבר הגנת מידע (General Data Protection Regulation, GDPR), שנכנסו לתוקף בשנת 2018, כוללות הוראות הנוגעות למערכות אוטומטיות של קבלת החלטות.[463] לפי מסמך זה, למושאי מידע מוקנית הזכות שלא להיות מושאן של החלטות המבוססות על עיבוד אוטומטי גרידא,

459 Sofia Ranchordas and Mattis Van 't Schip, *Future-Proofing Legislation for the Digital Age*, *in* Time, Law and Change: An Interdisciplinary Study 347 (Yaniv Roznai and Sofia Ranchordas eds., 2020). לביקורת על גישה חסינת עתיד לרגולציה של טכנולוגיה, בייחוד בהקשר של מעקב מקוון, ראו Maria Helen Murphy, Surveillance and the Law 79–81 (2019); Paul Ohm, *The Argument against Technology-Neutral Surveillance Laws*, 88 Tex. L. Rev 1685 (2010)

460 Investigatory Powers Act 2016, c.25 (Eng.). על חוק סמכויות החקירה של בריטניה ראו כהנא ושני, **לעיל** ה"ש 213, בעמ' 138–186. על עיצובו בדגם חסין עתיד ראו Murphy, **שם**, וכן Graham Smith, *Future-Proofing the Investigatory Powers Bill*, Cyberleagle (15.4.2016)

461 תזכיר חוק הגנת הסייבר ומערך הסייבר הלאומי, התשע"ח–2018 (להלן: תזכיר חוק הגנת הסייבר); ראו גם רחל ארידור הרשקוביץ ותהילה שוורץ אלטשולר **מהו סייבר? חלק ב: אתגרי האסדרה של הגנת הסייבר** 160–192 (2023).

462 ובהם AI HELG 2019, **לעיל** ה"ש 298; AI HELG 2018, **לעיל** ה"ש 298; AI4EU, **לעיל** ה"ש 50.

463 ס' 22 של התקנות הכלליות בדבר הגנת מידע (GDPR).

שיש להן השלכות משפטיות משמעותיות. בדברי המבוא לתקנות הכלליות בדבר הגנת מידע נאמר כי "בכל מקרה יהיה העיבוד הנזכר כפוף לבקרות מתאימות, ובכללן [העברת] מידע ספציפי למושא המידע והזכות להתערבות אנושית בהחלטה; [למושא המידע] תינתן הזכות להביע את עמדתו, לקבל הסבר של ההחלטה שהתקבלה לאחר הערכה כזאת ולערער עליה".[464]

4.1.1. ניהול סיכונים

כיוון שתקנות הבינה המלאכותית של אירופה עוצבו כך שיהיו חסינות עתיד,[465] הן נוקטות גישה של ניהול סיכונים ושומרות על ניטרליות טכנולוגית. למשל, מערכות בינה מלאכותית אינן מוגדרות בנוסח המקורי של התקנות המוצעות כבעלות זיקה לטכנולוגיה מסוימת, אלא כמערכות שפותחו באמצעות אחת או יותר מהגישות המפורטות בתוספת הראשונה לתקנות (תוספת שאפשר לתקן ולהרחיב בהתאם להתפתחויות טכנולוגיות עתידיות). לגישות הכלולות ברשימה יש יכולת לייצר פלט (תחזיות, המלצות, תוכן או החלטות) בהינתן מטרות שהגדירו בני אדם.[466] פלורידי בירך על הגדרה זו, שמשקפת בעיניו גישה מפוכחת הרואה בבינה מלאכותית טכנולוגיה לפתרון בעיות (ולא למשל ישות תבונית של ממש, שעלולה לקום על יוצריה או להיחשב אישיות משפטית נפרדת).[467] פטריק גלאונר, לעומת זאת, טען כי ההגדרה רחבה מדי, שכן לכאורה לפי הגדרה זו כל תוכנה שמשתמשת בממוצע (פרוצדורה סטטיסטית) נחשבת לבינה מלאכותית.[468]

464 ס' 71 לדברי המבוא לתקנות הכלליות בדבר הגנת מידע (GDPR). ראו גם Gianclaudio Malgieri and Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 Int'l Data Privacy L. 243, 248 (2017)

465 ראו דברי ההסבר, **לעיל** ה"ש 51, בעמ' 3, 12 ו־15. ראו גם המבוא לתקנות, בפס' 6, 71.

466 ס' 3(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

467 Luciano Floridi, *The European Legislation on AI: A Brief Analysis of its Philosophical Approach*, 34 Philosophy & Tech. 215 (2021)

468 Patrick Glauner, *An Assessment of the AI Regulation Proposed By the European Commission*, *in* The Future Circle of Healthcare: AI, 3D Printing, Longevity, Ethics, and Uncertainty Mitigation (Sepehr Ehsani, Patrick Glauner, Philipp Plugmann, and Florian M. Thieringer eds., 2022)

ההגדרה שבהצעה המתוקנת משככת את החששות שהביע גלאונר, שכן היא משמיטה את ההתייחסות לאופן הפיתוח של המערכת ושמה דגש על אוטונומיה. בינה מלאכותית היא "מערכת מבוססת מכונה שעוצבה כדי לפעול ברמות שונות של אוטונומיה והיא יכולה לייצר, למטרות מפורשות או לא מפורשות, פלטים כגון תחזיות, המלצות או החלטות המשפיעות על הסביבה הפיזית או הווירטואלית".[469] הגדרה זו מתיישבת עם הגדרת ה־OECD למערכות בינה מלאכותית,[470] אבל יש לתת את הדעת על העמימות שבנוסח "רמות שונות של אוטונומיה". פרשנות מרחיבה מדי של ההגדרה הזאת עלולה לכלול גם מערכות אוטומטיות פשוטות, כגון דלתות אוטומטיות מבוססות גלאי נפח.

גישת ניהול הסיכונים שבתקנות מחלקת את המערכות הנבונות לארבע רמות סיכון: מערכות אסורות, מערכות בסיכון גבוה, מערכות שחלה עליהן חובת שקיפות מיוחדת ומערכות בסיכון נמוך. יוער כי התקנות המוצעות לא יחולו על מערכות בשימוש צבאי בלעדי (דוגמת מערכות נשק אוטונומיות), גם אם רמת הסיכון בהן גבוהה.[471]

4.1.1.1. מערכות אסורות

על פי נוסח התקנות האירופיות יש לאסור שימוש במערכות בינה מלאכותית שיש בהן סכנה קיצונית והן אינן עולות בקנה אחד עם ערכי היסוד האירופיים.[472] התקנות המוצעות אוסרות על מערכות שמשתמשות בטכניקות תת־הכרתיות או מניפולטיביות כדי לעוות באופן מהותי את ההתנהגות האנושית באופן שמקשה במידה ניכרת על יכולתו של אדם לקבל החלטה מיודעת, וגורמים לו לקבל

469 ס' 3(1) להצעה המתוקנת, **לעיל** ה"ש 57.

470 על פי עקרונות ה־OECD מערכת בינה מלאכותית היא "מערכת מבוססת מכונה שיכולה, בהינתן קבוצת מטרות שהגדירו בן אנוש, לערוך תחזיות, לתת המלצות או לקבל החלטות המשפיעות על סביבות ממשיות או וירטואליות ברמות משתנות של אוטונומיה". ראו OECD 2019, **לעיל** ה"ש 294, בעמ' 3.

471 ס' 2(3) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

472 **שם**, ס' 5(1). ראו גם Rostam Josef Neuwirth, *Prohibited Artificial Intelligence Practices in the Proposed EU Artificial Intelligence Act* (29.10.2022)

החלטה שאחרת לא היה מקבל, באופן שסביר שיסב לאדם או לקבוצת אנשים נזק פיזי או פסיכולוגי משמעותי, וזאת למעט מערכות בינה מלאכותית לתכליות תרפויטיות.[473]

עוד אוסרות התקנות המוצעות מערכות שמנצלות חולשות של קבוצות שונות באוכלוסייה (למשל חולשות שקשורות לגיל, ליכולות פיזיות או ליכולות קוגניטיביות) באופן שסביר שיסב נזקים פיזיים או פסיכולוגיים מהותיים;[474] מערכות סיווג ביומטריות המסווגות אנשים בהסתמך על פרמטרים רגישים או באמצעות הסקה של פרמטרים רגישים (למעט מערכות שנועדו לתכליות תרפויטיות, בכפוף להסכמה מדעת של מי שנחשף אליהן);[475] וכן מערכות נבונות שנועדו לבצע דירוג חברתי[476] או לסווג או להעריך את מידת האמון שיש לתת באדם על בסיס התנהלותו החברתית או תכונות אופי שלו באופן שעלול להביא להחלטות לרעתם של יחידים או קבוצות בהקשרים חברתיים שאין להם זיקה להקשרים שבהם המידע נאסף או נוצר במקור;[477] מערכות שיטור מונחה עתיד (predictive policing), כלומר מערכות המעריכות את הסיכון שאדם יבצע עברה פלילית או מינהלית (לרבות הערכת סיכויי רצידיביזם);[478] מערכות המייצרות או מרחיבות מאגרי נתונים לזיהוי פנים בהסתמך על איסוף גורף מרשת האינטרנט או על חומר מצולם ממערכות טלוויזיה במעגל סגור; [479] מערכות בינה מלאכותית שתכליתן לזהות רגשות לתכליות של אכיפת חוק או ניהול מעברי גבול, או לשמש במוסדות חינוכיים ובהקשרים תעסוקתיים.[480]

473 ס' (a)(1)5 להצעה המתוקנת, **לעיל** ה"ש 57.

474 ס' (b)(1)5, **שם**.

475 ס' (ba)(1)5, **שם**.

476 ס' (45c)3, **שם**, מגדיר דירוג חברתי כהערכה או סיווג של אנשים בהסתמך על התנהגותם החברתית, על מעמדם החברתי־כלכלי או על תכונות האופי שלהם (הידועות או החזויות).

477 ס' (I)(1)5, **שם**.

478 ס' (da)(1)5, **שם**. ראו גם סעיף 2.3.2 **לעיל**.

479 ס' (db)(1)5, **שם**. נראה שסעיף זה נועד לאסור על מאגרי זיהוי פנים כגון אלו שמפתחת חברת Clearview Ai. ראו גם היל, **להלן** ה"ש 1059.

480 ס' (dc)(1)5 להצעה המתוקנת, **לעיל** ה"ש 57. ראו גם סעיף 2.3.2 **לעיל**.

התקנות אוסרות גם שימוש במערכות זיהוי ביומטריות בזמן אמת במרחבים ציבוריים.[481]

4.1.1.2. מערכות בסיכון גבוה

נוסח התקנות האירופיות מבחין בין שני סוגים עיקריים של מערכות בינה מלאכותית בסיכון גבוה: מערכות שמשמשות רכיב בטיחות במוצר[482] ומערכות שלפעולתן יש השלכות על זכויות יסוד מתוך רשימה מפורשת,[483] אם אלו מסכנות באופן משמעותי את בריאותם, בטיחותם או זכויותיהם הבסיסיות של אנשים.[484]

רשימה זו מפורטת בתוספת השלישית לתקנות וכוללת כ־20 קטגוריות של מערכות בינה מלאכותית: מערכות זיהוי ביומטריות; מערכות שנועדו לבקרה ובטיחות בתשתיות קריטיות; מערכות שמשמשות לקבלת מועמדים ללימודים או למתן ציונים; מערכות שמשמשות לגיוס עובדים ולהערכתם; מערכות דירוג אשראי; מערכות לקביעת זכאות לקבלת שירותים ציבוריים או לקבלת זכויות סוציאליות; מערכות לקביעת סדרי עדיפות בטיפול של שירותי חירום והצלה ציבוריים

481 ס' 5(1)(d), **שם**, אוסר לחלוטין פרישת מערכות אלו בלי להידרש לתכליתן, ואילו הנוסח המקורי של ס' 5(1)(d) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53, קובע סייגים בעניין זה. מערכות זיהוי ביומטריות "בזמן אמת" אסורות, אלא אם כן הן נחוצות לחיפוש ממוקד אחר קורבנות פוטנציאליים מסוימים של פשע (כגון ילדים חטופים); למניעת סכנה ספציפית מוגדרת, משמעותית ומיידית לחייהם של אנשים או לשלומם; לסיכול מתקפת טרור; או לאיתורם וזיהוים של עבריינים בעבירות חמורות. ס' 5(2) לתקנות הוסיף סייגים לשימוש במערכות זיהוי ביומטריות גם במקרים חריגים אלו. לפי אותם סייגים יש להביא בחשבון את עוצמת הסכנה הפוטנציאלית ואת השלכות הרוחב החברתיות הטמונות בהפעלתן. כן נדרש להבטיח שכל הסתייעות פרטנית במערכת זיהוי ביומטרית תיעשה בכפוף לצו של גורם שיפוטי או גוף מינהלי עצמאי. על פיקוח אקס אנטה לצורך מעקב מקוון ראו עמיר כהנא ויובל שני **פיקוח על מעקב מקוון בישראל** 44–46 (מחקר מדיניות 149, המכון הישראלי לדמוקרטיה 2020).

482 לפי ס' 3(14) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53, רכיב בטיחות במוצר או במערכת הוא רכיב של מוצר או של מערכת שיש לו פונקציה של בטיחות או שתקלה בו עלולה להיות סיכון בריאותי או בטיחותי לאנשים או לרכוש.

483 ס' 6 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

484 ס' 6(2) להצעה המתוקנת, **לעיל** ה"ש 57. הנציבות תפרסם קווים מנחים המגדירים את הנסיבות שבהן פלט של מערכות בינה מלאכותית עלול להיות איום מסוג זה בתוך שישה חודשים מיום כניסת התקנות לתוקף.

(טריאז'); מערכות שונות בשימוש רשויות אכיפת החוק (בהן מערכות לזיהוי דיפ־פייק,[485] להערכת מצב נפשי, להערכת סיכון, לחיזוי רצידיביזם על בסיס פרופילים, לחיזוי פשיעה על בסיס פרופילים,[486] לאנליטיקה קרימינולוגית); מערכות בשירות רשויות ההגירה והגבולות;[487] וכן מערכות בשירות מערכת המשפט, שנועדו לסייע במחקר, בפרשנות עובדות וחוקים וביישום החוק במערך נסיבות מוגדר. אפשר להוסיף על כל אלו עוד מערכות בינה מלאכותית בסיכון גבוה, ובתנאי שרמת הסיכון הנשקפת מהן שקולה לזו של המערכות ברשימה או גבוהה מהן.[488]

נוסח התקנות האירופיות הנוגעות למערכות בסיכון גבוה מושתת על הדגם האירופי לבטיחות מוצרים.[489] התקנות המוצעות מטילות על המערכות הללו חובות בעניין נתונים, משילות נתונים (data governance), כלומר היכולת

485 דיפ־פייק הוא כינוי לטכנולוגיה מבוססת בינה מלאכותית המאפשרת יצירת סרטוני וידאו, תמונות או קטעי קול מזויפים כך שייראו אמיתיים. האסדרה המשפטית של טכנולוגיה זו עודנה בחיתוליה. בשנת 2019 התקבלה במדינת קליפורניה חקיקה האוסרת על שימוש בטכנולוגיות אלו כדי להונות מצביעים או לפגוע בפוליטיקאים במסגרת תעמולה פוליטית, וכן על שימוש בטכנולוגיות אלו כדי להשתיל בסרטונים פורנוגרפיים דמויות של אנשים שלא הביעו את הסכמתם לכך. ראו גם "קליפורניה יוצאת נגד ה־Deep Fakes, חוקקה שני חוקים נגד התופעה" **כלכליסט** (7.10.2019); תהילה שוורץ אלטשולר ואיתי ברון **פייק ניוז: הדור הבא** (מאמר דעה, המכון הישראלי לדמוקרטיה 14.7.2019); הרשות להגנת הפרטיות "פרטיות ואבטחת מידע בשימוש בטכנולוגיות Deepfake (זיוף עמוק)" (28.8.2022); חב"כ 24–9 **יש עתיד – בראשות יאיר לפיד נ' עמותת "כן לשלום"** (18.1.2021); Anna Pesetski, *Deepfakes: A New Content Category for a Digital Age*, 29 Wm. & Mary Bill Rts. J. 503 (2020); Mamta Sareen, *Threats and Challenges by DeepFake Technology*, *in* DeepFakes: Creation, Detection, and Impact 100 (Loveleen Gaur ed., 2023)

486 ראו גם ס' 71 לדברי המבוא לתקנות הכלליות בדבר הגנת מידע (GDPR).

487 למערכות אלגוריתמיות בשירות ביקורת הגבולות באיחוד האירופי ראו לדוגמה Amanda Musco Eklund, *Frontex and "Algorithmic Discretion"* (Part I), Verfassungsblog (10.9.2022); idem, *Frontex and "Algorithmic Discretion" (Part II)*, Verfassungsblog (10.9.2022)

488 ס' 7 להצעה המתוקנת, **לעיל** ה"ש 57.

489 Michael Veale and Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act: Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach*, 4 Computer L. Rev. Int'l. 97 (2021)

לדעת איזה מידע נמצא ברשותך ולשלוט בניהולו), תיעוד, יידוע, בקרה אנושית, חסינות, דיוק ובטיחות – הכול בהתבסס על עבודת קבוצת המומחים האירופית.[490] עמידה בחובות אלו – מהן מהותיות ומהן פרוצדורליות – נבחנת במסגרת בחינת תאימות (conformity assessment),[491] שמשמשת תנאי לשחרורן של מערכות אלו לשוק האירופי. אחרי שמערכות אלו משוחררות לשוק הן נתונות לפיקוחן של רשויות פיקוח לאומיות (market surveillance authorities), שתפקידן לבחון אם מערכות אלו עדיין פועלות בהלימה לתקנות.[492]

לאורך כל מחזור החיים של מערכות בינה מלאכותית בסיכון גבוה יש להפעיל מערכת לניהול סיכונים. מערכת זו נועדה לזהות ולנתח סיכונים צפויים ולהעריך אותם, וכן לבחון מתווים צפויים של שימוש לרעה, להעריך סיכונים על בסיס נתונים ממערכות ניטור אחרות (post market),[493] ולאמץ אמצעים נאותים לניהול סיכונים בהתאם להוראות התקנות.[494] בדיקות של מערכות בינה מלאכותית בסיכון גבוה ייערכו לפני שחרורן לשוק.[495]

בהצעה האירופית מושם דגש מיוחד על תיעוד, מידע ונתונים מתוך הבנה שמשילות נתונים אפקטיבית הכרחית לשמירה על איכות גבוהה של מידע, שהיא עצמה תנאי הכרחי לתפקודן של מערכות בינה מלאכותית.[496] יש להבטיח שנתוני אימון, תיקוף (validation) ובדיקה יהיו כפופים לפרקטיקות מתאימות של משילות נתונים, המביאות בחשבון היבטים מסוימים של עיצוב המערכת, איסוף מידע, פעולות הכנה של הנתונים, פערים בנתונים ופוטנציאל להטיות.[497] בסיסי נתונים בשימוש מערכות בינה מלאכותית יביאו בחשבון היבטים מיוחדים

490 ראו AI HELG 2019, **לעיל** ה"ש 298.

491 ס' 16e, 19, 49 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

492 ראו **להלן**, וכן ס' 62–65 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

493 **שם**, ס' 61.

494 **שם**, ס' 9(2).

495 **שם**, ס' 9(7).

496 **שם**, פס' 44 למבוא.

497 **שם**, ס' 10(2).

הנוגעים לאופן פעולתן של מערכות בינה מלאכותית בסיכון גבוה.[498] לצורך ניטור הטיות שליליות[499] רשאים הספקים של מערכות אלו לעבד קטגוריות מיוחדות של מידע אישי, כגון מידע שיש בו כדי לזהות מוצא אתני או עמדות פוליטיות.[500]

יש להקפיד על תיעוד טכני של מערכות בינה מלאכותית בסיכון גבוה. על מלאכת התיעוד להתחיל בטרם שחרורן לשוק והתיעוד צריך להישאר מעודכן גם לאחר מכן.[501] התיעוד נועד לאפשר להדגים כי המערכת פועלת בהתאם להוראות התקנות.[502] נוסף על התיעוד הטכני נדרשות המערכות לתעד את פעולתן השוטפת (log) לצורכי נעקבות.[503]

על מערכות בינה מלאכותית בסיכון גבוה חלה גם חובת שקיפות. יש לעצב ולפתח מערכות שיהיו שקופות די הצורך לאפשר למשתמשים בהן לפרש את הפלט שלהן ולהשתמש בו כראוי.[504] מנגנוני השקיפות יכללו בין השאר הנחיות רלוונטיות למשתמשים,[505] שמקצתן מוגדרות בתקנות.[506] פינק ופינק מעירות שחובה זו חלה על מפתחיהן של מערכות בינה מלאכותית בסיכון

498 שם, ס' 10(4).

499 על הטיות ראו **להלן** בפרק 6.

500 ס' 10(5) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53. ס' 9 של התקנות הכלליות בדבר הגנת מידע (GDPR) אוסר על עיבוד סוגים שונים של מידע אישי רגיש שלא בהסכמת מושא הנתונים (או אם חריגים אחרים אינם מתקיימים). ס"ק 10(5)(a)–(g) להצעה המתוקנת, **לעיל** ה"ש 57, מונים תנאים מיוחדים שבכפוף להם יותר להשתמש במידע אישי רגיש למטרות זיהוי הטיות שליליות: אי־אפשר לזהות או לתקן את ההטיה באמצעות מידע סינתטי או מותמם; המידע עבר פסידו־אנונימיזציה; נוהלי האבטחה נאותים; המידע הרגיש לא יועבר לצד שלישי; המידע הרגיש נמחק לאחר תיקון ההטיה או עם תום תקופת השימור שלו.

501 ס' 11(1) להצעת תקנות הבינה המלאכותית האירופיות, **שם**. בהצעה המתוקנת, **לעיל** ה"ש 57, יש כמה הקלות בדרישות אלו לעסקים בינוניים וקטנים (SME), וכן לחברות הזנק.

502 **שם**, ס' 11(2).

503 **שם**, ס' 12.

504 **שם**, ס' 13(1).

505 **שם**, ס' 13(2).

506 **שם**, ס' 13(3).

גבוה ואין בה כדי להקנות זכויות או סעדים למשתמשי הקצה או למושאיהן של החלטות אוטומטיות. הן מבהירות כי אין בחובה זו כדי לייצר זכות להסבר – לא כלפי משתמשי הקצה ולא כלפי מושאיהן של החלטות מבוססות בינה מלאכותית; החובה נוגעת לפיתוח המערכות באופן שמאפשר הסבר, מבלי להורות עליו.[507]

על פי התקנות המוצעות, מערכות בסיכון גבוה יפותחו באופן שיאפשר פיקוח אנושי אפקטיבי לכל אורך מחזור החיים שלהן כדי למזער או למנוע סיכונים לבריאות, לבטיחות או לזכויות יסוד עקב שימוש סביר במערכות אלו (לרבות שימוש צפוי לרעה).[508] בדרך זו תהיה למפקח האנושי האפשרות לעמוד על היכולות והמגבלות של המערכות המפוקחות, להיות מודע לנטייה להסתמך או להסתמך יתר על המידה על התוצרים שלהן, לפרש נכונה את הפלט שלהן, להחליט בנסיבות מסוימות שלא להשתמש בו (או להתעלם ממנו או להפוך את התוצאות שלהן), וכן להתערב במהלך פעולתן או לעצור אותה.[509]

עיצוב ופיתוח של מערכות בינה מלאכותית בסיכון גבוה נדרש לפי התקנות המוצעות כדי להבטיח שמערכות אלו יעמדו ברמה נאותה של דיוק, חוסן (robustness) ואבטחה.[510] בין השאר מוצע שמערכות בינה מלאכותית בסיכון גבוה שלמידת המכונה שלהן נמשכת גם לאחר שחרורן לשוק יפותחו באופן שיבטיח מיתון של משובים המגבירים הטיות קיימות במערכת.[511]

מהחובות החלות על מערכות בינה מלאכותית בסיכון גבוה נגזרות החובות החלות על הספקים שלהן.[512] ספקי מערכות אלו נדרשים בין השאר להקים

507 Melanie Fink and Michèle Finck, *Reasoned A(I)administration: Explanation Requirements in EU Law and the Automation of Public Administration,* 47 Eu. L. Rev 376 (2022)

508 ס' 14(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** בה"ש 53.

509 **שם**, ס' 14(4).

510 **שם**, ס' 15(1).

511 **שם**, ס' 15(3).

512 **שם**, ס' 16–25.

מערכת בקרת איכות שתבטיח ציות להוראות התקנות,[513] לדאוג לתיעוד הטכני של המערכות,[514] לשמור על התיעוד המתמשך (log) של פעולתן,[515] להבטיח כי טרם שחרורן לשוק יעברו המערכות בחינת תאימות,[516] וכן לרשום מערכות אלו במרשם האירופי למערכות בינה מלאכותית בסיכון גבוה.[517]

חובה מרכזית בתקנות המוצעות היא בחינת התאימות (conformity). בחינת התאימות יכולה להיעשות בהסתמך על בקרות פנימיות של הספק[518] או בהסתמך על הערכה של גורם ביקורת חיצוני מוסמך (notified body).[519] בחינת התאימות נועדה לבדוק את הלימתן של מערכות בינה מלאכותית בסיכון גבוה להוראות התקנות וכוללת בחינה של מערכת בקרת האיכות,[520] של התיעוד הטכני,[521] של תהליכי העיצוב והפיתוח ושל אופן ניטורן של המערכות לאחר ששוחררו לשוק. גורמי הביקורת החיצוניים ינפיקו תעודה שתוקפה לא יעלה על חמש שנים, המעידה על הלימת המערכת לתקנות.[522] יואנידיס וגוטסופולו מראים שהמונח תאימות, שמתייחס למערך מוגדר של הוראות, צר יותר

513 **שם**, ס' 16b, 17.

514 **שם**, ס' (16(c. ס' 50 מורה כי על תיעוד זה להיות נגיש לרשויות המוסמכות הלאומיות לתקופה של עשר שנים מיום שחרורה של מערכת בינה המלאכותית בסיכון גבוה לשוק.

515 **שם**, ס' (16(d, 20.

516 **שם**, ס' (16(e, 49.

517 **שם**, ס' (16(f, 51, 60. נוסף על הרישום על הספק להמציא לרשות הלאומית המוסמכת הצהרת תאימות אירופית (EU declaration of conformity, ראו ס' 48, וכן תוספת V לתקנות). על הספק לשמור מסמכים שונים, ובהם ההצהרה והתיעוד הטכני של המערכת ושל מערכת בקרת האיכות שלה, ולהעבירם לרשות הלאומית המוסמכת לתקופה של עשר שנים (ס' 50).

518 **שם**, ס' (43(1)(A. ראו גם ההוראות שבתוספת VI לתקנות.

519 **שם**, ס' (43(1)(B. ראו גם ההוראות שבתוספת VII לתקנות.

520 **שם**, ס' (16(b, 17. וכן ס' 2 בתוספת VI לתקנות; ס' 3 בתוספת VII לתקנות.

521 **שם**, ס' (16(c, 18. וכן ס' 3 בתוספת VI לתקנות; ס' 4 בתוספת VII לתקנות.

522 **שם**, ס' 44.

מהמונח החלופי ציות (compliance), שמתייחס לעמידה כללית בהוראות הרגולטוריות.[523]

לאחר שחרורן של מערכות בינה מלאכותית בסיכון גבוה לשוק נדרשים הספקים שלהן להקים ולתעד מערכת ניטור אחרת שתאסוף ותנתח באופן שיטתי נתונים רלוונטיים על ביצועיהן כדי לאפשר לספק לאמוד את הלימתן הרציפה להוראות התקנות.[524] מערכת הניטור נדרשת לדווח לרשויות הפיקוח הלאומיות על השוק על כל תקרית חריגה[525] במערכות אלו שעולה כדי הפרה של דינים אירופיים שנועדו להגן על זכויות אדם.[526]

בתביעות נזיקין אזרחיות נגד ספקי בינה מלאכותית, אי־עמידה בחובות החלות על ספקים על פי נוסח התקנות האירופיות עלולה להשפיע על הנטל הראייתי להראות שיש קשר סיבתי בין הפרת התקנות לבין הנזק שגרמה מערכת בסיכון גבוה.[527]

נוסח התקנות האירופיות מחיל חובות לא רק על ספקי מערכות בינה מלאכותית בסיכון גבוה, אלא גם על שחקנים אחרים באקוסיסטם, ובכלל זה על יבואנים, על מפיצים[528] ואף על המשתמשים במערכות אלו. יבואנים של מערכות בינה מלאכותית בסיכון גבוה נדרשים להבטיח שמערכות אלו מתועדות כראוי ועברו

523 Nikolaos Ioannidis and Olga Gkotsopoulou, *The Palimpsest of Conformity Assessment in the Proposed Artificial Intelligence Act: A Critical Exploration of Related Terminology*, European Law Blog (2.7.2021)

524 ס' 61 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

525 תקריות חריגות מוגדרות תקריות שהביאו או היו עלולות להביא, במישרין או בעקיפין, למותו של אדם או לפגיעה חמורה בבריאותו, לשיבוש משמעותי של הניהול והתפעול של תשתיות קריטיות, לפגיעה בזכויות יסוד המוגנות בדין האירופי או לפגיעה משמעותית בסביבה או ברכוש. ראו ס' 3(44) להצעה המתוקנת, **לעיל** ה"ש 57.

526 ס' 62(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

527 ס' 4(2) להצעה לדירקטיבה אירופית לאחריות בתחום הבינה המלאכותית, **לעיל** ה"ש 26.

528 ס' 27 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

בחינת תאימות;[529] ואם אין הלימה ביניהן ובין הוראות התקנות עליהם להימנע מהפצתן.[530] כשספק, מפיץ או צד שלישי אחר משחררים לשוק מערכת בינה מלאכותית בסיכון גבוה תחת שמם או סימן מסחרי שלהם, או כשהם משנים את תכליתה המקורית של מערכת בינה מלאכותית שכבר קיימת בשוק, או כשהם משנים באופן משמעותי את המערכת עצמה, חלות עליהם החובות החלות על ספקי בינה מלאכותית בסיכון גבוה.[531]

משתמשים של מערכות בינה מלאכותית בסיכון גבוה – למשל משתמשים מוסדיים המעבדים מידע אישי של רבבות לקוחות ויותר – נדרשים בין השאר להשתמש במערכות אלו בהתאם להוראות ותיעודן, להבטיח שהמידע שמוזן למערכות אלו רלוונטי לתכלית שהן מיועדות לה, לתעד ולנטר את פעילותן ולדווח לספק או למפיץ כשמתעורר חשש לסכנה של ממש.[532] בתביעות נזיקין אזרחיות, אי־עמידה בחובות החלות על משתמשים על פי נוסח התקנות האירופיות עלולה להשפיע על הנטל הראייתי לקיומו של קשר סיבתי בין הפרתן לבין הנזק שגורמת מערכת בסיכון גבוה.[533]

4.1.1.3. מערכות שחלה עליהן חובת שקיפות מיוחדת

נוסח התקנות האירופיות מגדיר כמה סוגים של מערכות בינה מלאכותית שחלה עליהן חובת שקיפות מיוחדת. כך למשל, מערכות שנמצאות במגע עם בני אדם, ולפי נוסח ההצעה המתוקנת גם מערכות בינה מלאכותית גנרטיביות,[534] נדרשות להיות מעוצבות כך שבני אדם ידעו שהם באינטראקציה עם מכונה;[535] יש ליידע

529 **שם**, ס' (1)26.

530 **שם**, ס' (2)26.

531 **שם**, ס' 28.

532 **שם**, ס' 29.

533 ס' (3)4 להצעה לדירקטיבה אירופית לאחריות בתחום הבינה המלאכותית, **לעיל** ה"ש 26.

534 ס' (a)(4)28b להצעה המתוקנת, **לעיל** ה"ש 57.

535 ס' (1)52 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53. השוו AI HELG 2019, **לעיל** ה"ש 298, בעמ' 18; ס' 32 להצהרת טורונטו, **לעיל** ה"ש 324.

משתמשים במערכות לזיהוי רגשות אנושיים[536] או לסיווג ביומטרי[537] בנוגע לפעולת המערכת.[538]

נוסף על כך, מערכות שעושות מניפולציות בתצלומים, בקטעי קול או בסרטוני וידאו עד כדי יצירת תוכן מזויף (deep fake) נדרשות לגלות שהתוכן הוא תוצאה של שינוי או שהוא נוצר באופן מלאכותי. חובה זו לא תחול על מערכות לזיהוי עבריינות ומניעתה ואף לא במקרים שבהם הדבר נחוץ למימוש חופש הביטוי, האומנות והמדע כפי שהוא מוגדר באמנה האירופית לזכויות אדם וחירויות יסוד (Convention for the Protection of Human Rights and Fundamental Freedoms, CFR).[539]

אין בהחלת חובת השקיפות המיוחדת החלה על מערכות מסוימות כדי לגרוע מחובות אחרות שיכולות לחול על אותן מערכות אם אלו מסווגות כמערכות בסיכון גבוה.[540]

4.1.1.4. מערכות בסיכון נמוך

מערכות בינה מלאכותית בסיכון נמוך הן קטגוריה שיורית, כלומר כל מערכת שאינה מערכת בסיכון גבוה. בנוסח התקנות האירופיות מוצע לעודד מפתחים

536 מערכות לזיהוי רגשות מוגדרות בס' 3(34) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53, כמערכות בינה מלאכותית שתכליתן היא זיהוי או הסקה של רגשות, מחשבות, מצבי תודעה או כוונות של בני אדם או קבוצות על בסיס מידע ביומטרי. טכנולוגיה כזאת תוכל לשמש, למשל, לגילוי שקרים על בסיס ניתוח הבעות פנים, אבל נכון לכתיבת שורות אלו אין היא מניבה תוצאות מדויקות די הצורך. בשנת 2021 רמזה חברת הביטוח הדיגיטלית הישראלית למונייד בציוץ שפרסמה שהיא משתמשת ביכולות כאלה, אך דבריה עוררו סערה והחברה נסוגה מהציוץ. ראו שגיא כהן "חברת הביטוח למונייד אמרה שהיא משתמשת ב־AI כדי לזהות שקרנים – וחוללה סערה" **TheMarker** (27.5.2021).

537 מערכות לסיווג ביומטרי מוגדרות בס' 3(35) (**שם**) כמערכות בינה מלאכותית שפועלות על בסיס מידע ביומטרי ותכליתן סיווג בני אדם לקטגוריות ספציפיות, כגון מגדר, גיל, צבע שיער, צבע עיניים, קעקועים, מוצא אתני או נטייה מינית.

538 **שם**, ס' 52(2).

539 **שם**, ס' 52(3).

540 **שם**, ס' 52(4).

של מערכות בסיכון נמוך לאמץ מרצונם את החובות החלות על מערכות בסיכון גבוה. כן מוצע שהנציבות האירופית, מדינות חברות והוועד האירופי לבינה מלאכותית (EU AI Board) יעודדו פיתוח של תקן רצוני למערכות בינה מלאכותית בתעשייה, בשיתוף משתמשים ובעלי עניין אחרים.[541]

4.1.2. "ארגז חול" רגולטורי

לצד החשיבות של התקנות האירופיות המוצעות להגנה על זכויות יסוד יש גם חשש שרגולציית יתר תצנן את יצר חדשנות ותפגע אפוא בפיתוחים הטכנולוגיים.[542] התקנות מתמודדות עם אתגר זה באמצעות הסדרים המאפשרים הקמה של "ארגזי חול" רגולטוריים – סביבה מבוקרת לפיתוח, בדיקה ותיקוף של מערכות חדשניות טרם שחרורן לשוק.[543]

4.1.3. סנקציות

בהצעת התקנות האירופיות מוצע להשית קנסות גדולים בגין הפרתן, ואלו האמירו עם פרסום נוסח ההצעה המתוקנת. לפי הצעה זו, אם תפר חברה את האיסור על פרישׂת מערכות בינה מלאכותית מסוכנות ושימוש בהן, יוטל עליה קנס מינהלי של עד 40 מיליון אירו או 7% ממחזור הפעילות העסקית שלה, הגבוה מהם.[544] בגין הפרת חובות הנוגעות למשילות נתונים יוטל קנס מינהלי של עד 20 מיליון אירו או 4% ממחזור הפעילות העסקית של החברה.[545] בגין הפרות אחרות יוטלו קנסות מינהליים של עד 10 מיליון אירו או 2% ממחזור הפעילות העסקית השנתי של החברה, הגבוה מהם.[546]

541 **שם**, ס' 69.

542 ראו לדוגמה Gönenç Gürkaynak, Ilay Yilmaz, and Güneş Haksever, *Stifling Artificial Intelligence: Human Perils*, 32 Computer L. and Sec. Rev. 749 (2016)

543 ס' 53–54 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

544 ס' 71(3) להצעה המתוקנת, **לעיל** ה"ש 57.

545 **שם**, ס' 71(4).

546 **שם**, ס' 71(4)-(5).

4.1.4. אסדרת מודלי יסוד (Foundation models)

אומנם בין פרסום הצעת תקנות הבינה המלאכותית האירופיות באפריל 2021 לפרסום נוסח ההצעה המתוקנת במאי 2023 חלפו כשנתיים, אך נראה כי במונחים טכנולוגיים חלפו שנות דור. לא היה אפשר להתעלם מהפריצה של המערכות הגנרטיביות ומהתפוצה הרחבה שלהן, וההצעה המתוקנת כללה אפוא הוראות פרטניות בעניין מודלי יסוד.

מודל יסוד הוא "מודל בינה מלאכותית שאומן על בסיס נתונים רחב, תוכנן להפיק פלט כללי ואפשר להתאימו למגוון רב של משימות נפרדות".[547] כדי להבהיר את מצבם המשפטי של ספקי מודלי יסוד, ושל ספקי מערכות בינה מלאכותית הנסמכות עליהם, נדרש הסדר פרטני.[548] ההסדר בהצעה המתוקנת כולל הוראות החלות על מודלי יסוד בבינה מלאכותית גנרטיבית, המוגדרים "מודלי יסוד שמשתמשים במערכות בינה מלאכותית כדי לחולל תוכן ברמות משתנות של אוטונומיה, ובכלל זה מלל מורכב, תמונות, צליל או וידאו".[549]

ספק של מודל יסוד נדרש להבטיח בטרם שחרורו לשוק שהמודל מציית להוראות החוק, בין שהוא מופץ כמודל עצמאי ובין שהוא מוטמע במערכת בינה מלאכותית או במוצר אחר. ההוראות חלות על כל ערוצי ההפצה – גם אם הגישה למודל חינמית (קוד פתוח).[550] הספק נדרש להראות שנקטו אמצעים (עיצוב, בדיקה וניתוח) למזעור סיכונים לבריאות, לבטיחות, לזכויות יסוד, לסביבה, לדמוקרטיה ולשלטון החוק; שבסיסי הנתונים עובדו בהתאם לממשל נתונים המתאים למודלי יסוד ונבחנה מידת התאמתם של מקורות מידע, אפשרות קיומן של הטיות ודרכי מיתון אפשריות לפגיעה בזכויות;[551] שמודל היסוד עוצב ופותח באופן שמבטיח רמות מתאימות של תפקוד, צפיות, בטיחות, אבטחת מידע ויכולת תיקון לכל אורך מחזור החיים שלו, מתוך שימוש בתקנים סביבתיים מתאימים הממזערים צריכה אנרגטית;[552]

547 **שם**, ס' (3(1c)(dc.
548 **שם**, פס' 60g לדברי המבוא.
549 **שם**, ס' (28b(4.
550 **שם**, ס' (28b(1.
551 **שם**, ס' (28b(2)(b.
552 **שם**, ס' (E)–28b(2)I.

שיש תיעוד טכני מקיף והנחיות ברורות לשימוש כדי לאפשר לספקי מערכות הבינה המלאכותית הנסמכות על מודל היסוד לציית לחובותיהן לפי תקנות הבינה המלאכותית לתקופה של עשור לאחר שחרור מודל היסוד לשוק;[553] שהוקמה מערכת בקרת איכות שתבטיח ציות להוראות התקנות ותתעד אותו; וכן שמודל היסוד נרשם במרשם האירופי המתאים.[554]

לפי ההצעה המתוקנת, על ספקים של מודלי יסוד שמשתמשים בבינה מלאכותית גנרטיבית יחולו חובות שקיפות מיוחדות.[555] עליהם לאמן, לעצב ולפתח את מודל היסוד באופן שיבטיח הגנה נאותה מיצירת תוכן המפר את הדין האירופי,[556] וכן לתעד את השימוש בנתוני אימון המוגנים בזכויות יוצרים.[557]

4.1.5. התקנות האירופיות: הערכה ראשונית

אסדרתן של טכנולוגיות מתעוררות אינה פשוטה, שכן יכולותיהן ודפוסי השימוש בהן יכולים להשתנות תוך כדי תהליך האסדרה. פלורידי ציין לחיוב שמנסחי ההצעה לא ראו לנגד עיניהם חזון דיסטופי של מערכות בינה מלאכותית המשמידות את האנושות, אלא נקטו גישה פרגמטית שלפיה טכנולוגיות של בינה מלאכותית אינן אלא מערכות סטטיסטיות לפתרון בעיות.[558]

ההגדרה הרחבה של מערכות בינה מלאכותית בנוסח התקנות האירופיות, גם בנוסח שבהצעה המתוקנת, מקשה עלינו להבחין בין בינה מלאכותית לאלגוריתמים, אך היות שבקרב המומחים אין תמימות דעים באשר להגדרה המדויקת של בינה מלאכותית,[559] אפשר שאכן מוטב לאמץ הגדרה רחבה. חרף החשש מהחלת התקנות האירופיות על מערכות שהגדרתן כבינה מלאכותית גבולית או שנויה במחלוקת, יש לשים לב שהגישה שנוקטות התקנות האירופיות בעניין ניהול סיכונים אינה מחילה חובות כלשהן על מערכות בסיכון נמוך.

553 **שם**, ס' 28b(2)(c) וכן 28b(3).

554 **שם**, ס' 28b(2)(g).

555 **שם**, ס' 28b(4)(a), וכן ראו סעיף 4.1.1.3 **לעיל**.

556 **שם**, ס' 28b(4)(b).

557 **שם**, ס' 28b(4)(c).

558 Floridi, **לעיל** ה"ש 467, בעמ' 219.

559 ראו **לעיל** בסעיף 1.1.

הדגש שניתן בתקנות להיבטים של משילות נתונים וניהול מידע ממחיש את ההבחנה בין אלגוריתמים "פשוטים" לבינה מלאכותית, שסם החיים שלה הוא נתונים. עוד בטרם פרסום התקנות המוצעות הכירה החקיקה האירופית בקשר שבין נתונים לבינה מלאכותית. סעיף 22 של ה־GDPR, המקנה למושא המידע את הזכות שלא יתקבלו בעניינו החלטות על בסיס עיבוד אוטומטי של המידע האישי שלו, קושר בין אוטומציה לנתונים.[560] ההקלות המוצעות בארגז החול הרגולטורי בסעיפים 55-53 של התקנות האירופיות (שעיקרן פטור מעקרון צמידות המטרה כדי לפתח מערכות בינה מלאכותית על בסיס נתונים שנאספו למטרות אחרות), גם הן מעידות על ההכרה שהמידע הכרחי לקידומן ולפיתוחן של מערכות נבונות.

אחת ממטרות התקנות האירופיות היא להציע פרקטיקות מומלצות לניהול מידע ולפיתוח מערכות בינה מלאכותית. על פי סעיף 10 של תקנות אלו, נתוני אימון, תיקוף ובדיקות של מערכות בינה מלאכותית יהיו כפופים לפרקטיקות מתאימות של משילות נתונים.[561] שיעור הקנס המקסימלי המוטל בגין הפרות של סעיף 10 זהה לשיעור המוטל בגין פרישׂת מערכות בינה מלאכותית אסורות ושימוש בהן, עדות נוספת לחשיבות שמנסחי התקנות האירופיות מייחסים למדיניות נתונים ולמשילות נתונים נאותות להגנה על זכויות יסוד.

הרצון להבטיח שהתקנות יהיו רלוונטיות גם לאחר שינויים עתידיים במערכות הבינה המלאכותית ייצר לעיתים ניסוחים עמומים שיהיה צורך להבהיר בעתיד. האם למשל אפשר להגדיר רשתות חברתיות (או למצער, את הבינה המלאכותית החולשת על ה"פיד") "מערכות שמפעילות טכניקות תת־סיפיות כדי לשנות התנהגות אנושית", שנמנות עם מערכות הבינה המלאכותית האסורות לשימוש?[562]

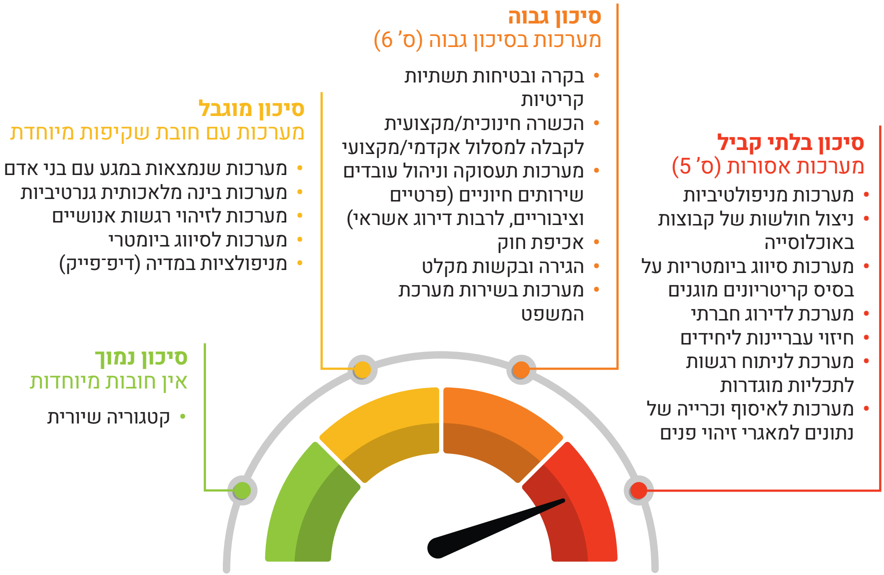
560 ראו **לעיל** ה"ש 403.

561 ראו **לעיל** בסעיף 4.1.1.2.

562 ס' 5(a) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53. Floridi, **לעיל** ה"ש 467, בעמ' 219, מביע חשש שמא עמימות זו תגדיר שלא לצורך מערכות בינה מלאכותית לתכליות תמימות כמערכות אסורות. גם תומס ברי ופרדריק פון־בוטמר סוברים כך. ראו Thomas Burri and Fredrik von Bothmer, *The New EU Legislation on Artificial Intelligence: A Primer* (21.4.2021)

הצורך לייצר ודאות משפטית בתחומים מסוימים שיש בעניינם שיח ציבורי ער (דירוג חברתי, דיפ־פייק וזיהוי פנים)[563] תרם לנוסח המקורי של התקנות בעיקר באמצעות יצירת איים של החרגות – בפרט במה שנוגע לשימוש בטכנולוגיות אלו למטרות ביטחוניות ולאכיפת חוק.[564] בנוסח ההצעה המתוקנת המסתמן נראה שהחרגות אלו נעלמו, וכי לפנינו מדרג סדור יותר של רמות סיכון, כמתואר בתרשים 2.

תרשים 2
סיווג רמות הסיכון של מערכות בינה מלאכותית
בתקנות הבינה המלאכותית האירופיות



563 ראו לדוגמה את השיח הער על מערכות לזיהוי פנים וחוקיותן באירופה ומחוץ לה. Joe Devansan, *EU Privacy Debate Rages Regarding Facial Recognition Firm Clearview AI*, T_HQ (1.6.2021); *Legality of Collecting Faces Online Challenged*, BBC News (27.5.2021); Amir Ali, *RCMP Violated Canadian Privacy Act with Facial Recognition Technology*, DHNews (10.6.2021)

564 ראו לדוגמה ההוראות המיוחדות לשימוש במערכות סיווג ביומטריות בידי גורמי אכיפת חוק שבס' 6(1)(d) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

יש הסבורים למשל שהסיווג המוצע של מערכות בינה מלאכותית שחל עליהן איסור קטגורי, החשוב כשהוא לעצמו, אינו כולל את כל מערכות הבינה המלאכותית שראוי שיהיו אסורות, ובייחוד מערכות העוסקות במעקב המונים.[565] עם זאת, כשבוע לפני פרסום התקנות האירופיות דלפה טיוטה שלהן, שכללה איסור מפורש על מעקב חסר הבחנה למטרות שאינן ביטחון לאומי.[566] ההצעה המתוקנת הרחיבה את סוגי המערכות האסורות. כך למשל, הביקורת על סיווגן של מערכות לזיהוי רגשות כמערכות שחלה עליהן אך ורק חובת שקיפות מיוחדת, ולא כמערכות בסיכון גבוה,[567] זכתה למענה מסוים בהצעה המתוקנת.[568] כך גם הקולות שקראו שלא להסתפק באיסורים הפרטניים על שימוש במערכות זיהוי פנים במרחב הציבורי לצורך אכיפת חוק, ותחת זאת להחיל איסור גורף על כל שימוש במערכות זיהוי פנים במרחב הציבורי,[569] לרבות על שחקנים פרטיים.[570] ההצעה המתוקנת הרחיבה את האיסורים על שימוש במערכות ביומטריות בזמן אמת,[571] והוסיפה הגנות מפני סיווג ביומטרי על בסיס קריטריונים מוגנים.[572] ואולם מכל הדוגמאות האלה עולה שחסרה הגדרה כללית לסיכון קטגורי חלף הרשימה הסגורה או נוסף עליה.[573]

565 ראו לדוגמה Sarah Chander, *EU's AI Law Needs Major Changes to Prevent Discrimination and Mass Surveillance*, EDRi (28.4.2021)

566 James Vincent, *The EU Is Considering a Ban on AI For Mass Surveillance and Social Credit Scores*, The Verge (14.4.2021)

567 Gianclaudio Malgieri and Marcello Ienca, *The EU Regulates AI but Forgets to Protect our Mind*, European Law Blog (7.7.2021)

568 ס' 5(1)(dc) להצעה המתוקנת, **לעיל** ה"ש 57.

569 *EDPB & EDPS Call for Ban on Use of AI For Automated Recognition of Human Features in Publicly Accessible Spaces, and Some Other Uses of AI that Can Lead to Unfair Discrimination*, European Data Protection supervisor (21.6.2021)

570 *People, Risk and the Unique Requirements of AI*, Ada Lovelace Institute (31.3.2022)

571 השוו את ס' 5(1)(d) להצעה המתוקנת, **לעיל** ה"ש 57, לס' 5(1)(d)–(2) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

572 ס' 5(1)(ba) להצעה המתוקנת, **לעיל** ה"ש 53.

573 ראו למשל Lilian Edwards, Regulating AI in Europe: Four Problems and Four Solutions 3 (Expert opinion, Ada Lovelace Institute, 31.3.2022)

ייתכן גם שההתייחסות הפרטנית להיבטים אקטואליים של בינה מלאכותית עלולה להביא לידי רגולציה מבוססת סיכון, "רפלקס רגולטורי עקב סכנה" (risk regulation reflex). במקרים של נזק נדיר אך רחב היקף יש נטייה ציבורית ללחוץ על הרגולטור להגיב בחומרה יתרה כדי למנוע את הישנותו.[574]

ביקורת נוספת על התקנות האירופיות נוגעת לבחירה המתעתעת במונח "משתמשים" לתיאור מי שפורשׂים או מפעילים מערכות בינה מלאכותית, אף שמונח זה מורה בדרך כלל על משתמשי הקצה של מערכות אלו. יש המעירים כי כך קל יותר לתת לאותם משתמשים פטור מהחובה לבצע הערכת תאימות.[575]

יתר על כן, יש בהבחנה בין "משתמשים" של מערכות בינה מלאכותית ל"ספקים" שלהן כדי לעודד את המגזר הציבורי להימנע מפיתוח פנימי של מערכות בינה מלאכותית ליישומי ממשל. כך למשל, אם רשויות הממשלה יפתחו מערכת לאיתור הונאות בתחום הרווחה דוגמת SyRI ההולנדית,[576] יהיה להן מעמד של ספקי בינה מלאכותית.[577] ובתור ספקים, יחולו עליהן החובות שבתקנות למיתון הסיכונים הפוטנציאליים של המערכת.[578] לעומת זאת, אם יבחרו הרשויות בחלופה של מיקור חוץ, וירכשו מספק חיצוני מערכת מן המוכן, לא יחולו עליהן אלא החובות המוטלות על "משתמשים",[579] כלומר הן יהיו רשאיות להפעיל שיקול דעת נרחב בהאצלת סמכויות הפיקוח לספקי המערכת.[580]

574 Nicolas Petit and Jerome De Cooman, *Models of Law and Regulation for AI*, *in* The Handbook of AI 199, 205 (Anthony Elliott ed., 2022)

575 Martin Ebers et al., *The European Commission's Proposal for an Artificial Intelligence Act: A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)*, 4 J 589 (2021)

576 ראו **להלן** בסעיף 5.2.

577 ס' 3(2), 28(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

578 **שם**, ס' 16.

579 **שם**, ס' 29.

580 ראו בהקשר זה *How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers*, Human Rights Watch (10.11.2021)

יש הגורסים כי המגרעת העיקרית של התקנות טמונה בהתמקדותן בחובות החלות על תאגידים ולא בזכויות של בני אדם.[581]

כדי להתגבר על הבלבול המושגי במונח "משתמשים" לא די להבחין בין משתמשי קצה (למשל רופא המשתמש במערכת דיאגנוסטיקה) למפעילים ריכוזיים של מערכות (למשל קופת החולים המנהלת את מערכת הדיאגנוסטיקה); יש להגדיר קטגוריה נוספת של "אנשים מושפעים" (affected persons) – בני אדם (או ישויות משפטיות) שמושפעים מפעילות מערכות הבינה המלאכותית.[582]

ההצעה של התקנות לאסדרת הבינה המלאכותית באירופה היא מהלך שאפתני, המזהה רִיק רגולטורי כלל־עולמי ומנסה לייצר הובלה אירופית בתחום.[583] לתקנות יש תחולה אקסטרה־טריטוריאלית,[584] כלומר הן חלות גם על ספקים ממדינות מחוץ לאיחוד האירופי שמספקים מערכות בינה מלאכותית בתחומיו, וכן על ספקים ומשתמשים של מערכות בינה מלאכותית ממדינות שאינן שייכות לאיחוד האירופי אם הם משתמשים בפלט של מערכות אלו בתוך האיחוד. אקס־טריטוריאליות זו, המזכירה את "האימפריאליזם הרגולטורי" של ה־GDPR,[585]

581 ראו למשל Hannah Van Kolfschooten, *EU Regulation of Artificial Intelligence: Challenges for Patients' Rights*, 59 Common Market Law Review 81, 106 (2022)

582 *People, Risk and the Unique Requirements of AI*, **לעיל** ה"ש 570.

583 ואגליס פפקונסטנטינו ופול דה הרט מעירים שהתקנות המוצעות הן דוגמה נוספת ל"ברוטליות" של דיני האיחוד, המזהים ואקום רגולטורי בתחומים טכנולוגיים וכופים את דין האיחוד על המדינות החברות בו בלי שהללו ניסו להתאים את הדין הלאומי שלהן למציאות הטכנולוגית החדשה. ראו Vagelis Papakonstantinou and Paul De Hert, *EU Lawmaking in the Artificial Intelligent Age: Act-ification, GDPR Mimesis, and Regulatory Brutality*, European Law Blog (8.7.2021)

584 ס' 2 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

585 ראו Graham Greenleaf, *The "Brussels Effect" of the EU's "AI Act" on Data Privacy Outside Europe*, 171 Privacy Laws & Business International Report 1, 3–7 (2021). להשפעה הגלובלית של ה־GDPR ראו He Li, Lu Yu, and Wu He, *The Impact of GDPR on Global Technology Development*, 22 J. of Global Information Tech. Manag'nt 1 (2019); Cedric Ryngaert and Mistale Taylor, *The GDPR As Global Data Protection Regulation*?, 114 American J. Int'l. L. 5 (2020); Graham Greenleaf, *Global Data Privacy Laws 2021: Despite COVID Delays, 145 Laws Show GDPR Dominance*, 169 Privacy L. & Bus. Int'l Report 1 (2021); Huw

שתחולתו האקס־טריטוריאלית משפיעה על דיני הגנת הפרטיות במדינות רבות המעוניינות לסחור עם האיחוד האירופי,[586] עשויה לתמרץ שחקנים רבים בתעשייה מחוץ לאיחוד האירופי לפעול בהתאם לדין האירופי ואף לעודד מדינות אחרות לפתח את דיני הבינה המלאכותית שלהם לפי המודל האירופי.[587] מומחים מעריכים – עוד בטרם נחקקו התקנות – כי לכל הפחות לבחינת התאימות תהיה השפעה מחוץ לגבולות האיחוד.[588]

4.2 אסדרת בינה מלאכותית בסין

בשלהי חודש אוגוסט 2021 פרסמה מינהלת הסייבר של סין את טיוטת התקנות לניהול המלצות אלגוריתמיות במערכות מידע אינטרנטיות,[589] ובמרץ 2022 נכנסו התקנות לתוקף.[590] חקיקה זו היא חלק ממהלך אסדרה מקיף של ענף ההייטק שנקט

Roberts et al., *Achieving a "Good AI Society": Comparing the Aims and Progress of the EU and the US*, 27(68) Science and Engineering Ethics 7 (2021)

586 לרבות ישראל. ראו למשל עמיר כהנא "המטאטא יאה לנו – 'הכלי' של השב"כ והחלטת הנאותות האירופית" **אתר משפט ועסקים** (29.3.2021); ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 403, בעמ' 8.

587 בהקשר זה ראו גם Anu Bradford, The Brussels Effect: How the European Union Rules the World 78–81, 131–169 (2020)

588 ראו Charlotte Siegmann and Markus Anderljung, *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*, Centre For the Governance of AI (16.8.2022). לגישה הסבורה ש"אפקט בריסל" של הצעת תקנות הבינה המלאכותית האירופיות יהיה מתון ראו Alex Engler, *The EU AI Act Will Have Global Impact, but a Limited Brussels Effect*, Brookings (8.6.2022)

589 *Translation: Internet Information Service Algorithmic Recommendation Management Provisions (Opinon-Seeking Draft)*, Digichina (14.10.2021)

590 *Translation: Internet Information Service Algorithmic Recommendation Management Provisions: Effective March 1, 2022*, Digichina (10.1.2022) (להלן: תקנות מערכות ההמלצה האלגוריתמיות הסיניות). לרקע עליהן ולהרחבה בעניינן ראו Gilad Abiri and Xinyu Huang, *The People's (Republic) Algorithms*, 12 Notre Dame J. Int'l & Comp. L. 16 (2022)

הממשל הסיני כדי לבסס את כוחו מול הענף המתחזק.[591] התקנות הסיניות חלות על מערכות המלצה אלגוריתמיות מבוססות רשת המספקות תכנים למשתמשים – מערכות המלצה אישיות על תכנים, מערכות קבלת החלטות ומערכות חיפוש וסינון.[592]

התקנות מורות בין השאר לספקי שירותי מערכות המלצה אלגוריתמיות "לשמור על מוסר הציבור ועל אתיקה מסחרית ומקצועית", וגם לכבד עקרונות של הוגנות, צדק, פתיחות ושקיפות.[593] הן מנחות לפעול לשימוש במערכות אלגוריתמיות לקידום הטוב ואוסרות להשתמש בהן לפעילויות שיש בהן משום פגיעה בביטחון הלאומי, שיבוש הסדר הכלכלי והחברתי או הפרה של כל דין.[594] הן אף אוסרות על שימוש לרעה במידע ומניפולציה שלו במטרה להשפיע על דעת הקהל[595] ועל הפצת ידיעות כזב (fake news) באמצעות ספקי מערכות המלצה אלגוריתמיות.[596] מערכות אלגוריתמיות המפירות את הנוהג והסדר הציבוריים, ובפרט כאלו המעודדות התמכרות של משתמשים לשימוש בהן או מקדמות צריכת יתר, אסורות במפורש.[597] אסורות במיוחד מערכות המעודדות התמכרות כזאת בקרב קטינים.[598] התקנות אף מחייבות ספקי שירותי המלצות אלגוריתמיות לוודא שמערכות שנועדו לשימוש קשישים (elderly) אינן מאפשרות הונאות והן נוחות לקהל היעד שלהם.[599]

591 ראו בהקשר זה Angela Huyue Zhang, *Agility over Stability: China's Great Reversal in Regulating the Platform Economy,* 63 Harv. Int'l. L. J. (2022); Chang Che, *China's "Big Tech Crackdown": A Guide*, The China Project (2.8.2021)

592 ס' 2 לתקנות מערכות ההמלצה האלגוריתמיות הסיניות, **לעיל** ה"ש 590.

593 **שם**, ס' 4.

594 **שם**, ס' 6.

595 **שם**, ס' 14.

596 **שם**, ס' 13.

597 **שם**, ס' 8.

598 **שם**, ס' 18.

599 **שם**, ס' 19.

התקנות מפרטות שורה של הגנות על משתמשים, ובהן הזכות לעיין במידע עליהם, לערוך אותו ולקבל הסברים כשהם סבורים שפעולת המערכת השפיעה באופן מהותי על הזכויות והאינטרסים שלהם.[600] על ספקי מערכות ההמלצה האלגוריתמיות חלה חובת שקיפות כללית ועליהם לתת פומבי לעקרונות הפעולה של המערכות.[601]

התקנות מנחות את הספקים של מערכות ההמלצה האלגוריתמיות לנקוט אמצעים טכניים כדי למנוע מן האלגוריתמים שלהם להשפיע השפעה שלילית על המשתמשים או לעורר מחלוקות ושערוריות ציבוריות.[602] המערכות נדרשות לאפשר התערבות אנושית ומאפשרים למשתמשים בחירה אוטונומית, אך עליהן להציג בערוצי התוכן המרכזיים שלהם "תוכן התואם ערכים מקובלים".[603]

בתקנות מוצע להקים מרשם מרכזי של מערכות המלצה אלגוריתמיות במשרד הגנת הסייבר והמידע של סין.[604] משרד זה ישמש רגולטור ויפקח על המערכות לצד משרדים רלוונטיים נוספים.[605] אם יימצא שספק מערכות המלצה אלגוריתמיות הפר את הוראות התקנות ינחה המשרד את הספק כיצד לתקן את ההפרה. בנסיבות חמורות, או כשהספק מסרב לתקן את ההפרה, רשאי המשרד להקפיא את פעילות המערכת, לקנוס את הספק או לנקוט הליכים פליליים.[606]

שלא כמו התקנות האירופיות, החלות על מערכות בינה מלאכותית באשר הן, התקנות הסיניות מוגבלות ליישומי רשת של בינה מלאכותית. לפיכך אין הן חלות על מערכות ממשלתיות לקבלת החלטות, על מערכות זיהוי פנים, על רכבים אוטונומיים ועל יישומים אחרים של בינה מלאכותית. כמו כן, בכל הנוגע לניהול סיכונים התקנות הסיניות שיטתיות פחות מהאירופיות וככל הנראה

600 **שם**, ס' 17.

601 **שם**, ס' 16.

602 **שם**, ס' 12.

603 **שם**, ס' 11.

604 **שם**, ס' 23–25.

605 **שם**, ס' 25.

606 **שם**, ס' 31.

משליכות את יהבן על הערכת הסיכון הקונקרטית של הרגולטור, המסתמכת על "רקמה פתוחה" של סטנדרטים עמומים כגון פגיעה בביטחון הלאומי או בנוהג ובסדר הציבוריים.

רגולטור המרחב הקיברנטי של סין פרסם לאחרונה כללים ייעודיים לטכנולוגיות דיפ־פייק (סינתזה עמוקה).[607] הכללים חלים על טכנולוגיות לשינוי טקסט, קול, מוזיקה ומאפיינים ביומטריים בתמונות ובווידאו, על שחזורים תלת־ממדיים וסימולציות ועל מי שמספק שירותי סינתזה עמוקה או משתמש בהם. הכללים הנזכרים מורים לספקים לייצר תוויות אזהרה בולטות שמציינות שמדובר בסינתזה עמוקה.[608] חל איסור על הסרת תוויות אלו.[609]

4.3 הצעת חוק הבינה המלאכותית בברזיל

בספטמבר 2021 פורסמה בברזיל הצעת חוק לאסדרת בינה מלאכותית. שלא כמו התקנות האירופיות המפורטות, הצעת החוק בברזיל היא מסמך רזה למדי המפרט עקרונות כלליים לאסדרה של מערכות בינה מלאכותית.

החוק מונה כמה מטרות: פיתוח מדעי וטכנולוגי של ברזיל; תמרוץ פיתוח כלכלי בר־קיימא וקידום רווחת הכלל; קידום התחרות; השתלבות של ברזיל בשרשראות אספקה גלובליות; שיפור השירותים הציבוריים; והגנת הסביבה.[610] החוק מונה גם כמה יסודות לפיתוח וליישום בינה מלאכותית בברזיל: פיתוח מדעי וטכנולוגי, יוזמה חופשית, שמירה על אתיקה, זכויות אדם וערכים דמוקרטיים, חופש המחשבה והביטוי, היעדר אפליה, העדפה או תמרוץ של אסדרה עצמית, פרטיות,

607 *Provisions on the Administration of Deep Synthesis Internet Information Services*, China Law Translate (11.12.2022) (להלן: כללי הסינתזה העמוקה). על דיפ־פייק ראו **לעיל** ה"ש 485.

608 **שם**, ס' 17. השוו גם לס' 52(3) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

609 ס' 18 לכללי הסינתזה העמוקה, **לעיל** ה"ש 607.

610 ס' 3 להצעת החוק הברזילאית לבינה מלאכותית, **לעיל** ה"ש 54.

אבטחת מידע וגישה למידע, ביטחון לאומי.[611] לצד אלו הוא מפרט את העקרונות לפיתוח ויישום בינה מלאכותית בברזיל: קידום הטוב,[612] האדם במרכז,[613] היעדר אפליה,[614] ניטרליות,[615] בטיחות,[616] אחריות,[617] זמינות מידע[618] ושקיפות.[619]

ההגדרה לבינה מלאכותית בהצעת החוק הברזילאית מזכירה את ההגדרה בתקנות האירופיות.[620] לפי הגדרה זו, בינה מלאכותית היא מערכת המבוססת על תהליך חישובי, שבהתאם למטרות שהגדירו בני אדם יכולה לעבד נתונים ומידע כדי ללמוד, לתפוס ולפרש סביבה חיצונית ולהגיב אליה, ועל יסוד זה לערוך תחזיות, להמליץ, לסווג או לקבל החלטות מתוך שימוש בטכניקות כגון למידת מכונה, מערכות לוגיות או גישות סטטיסטיות.[621] הצעת החוק מדגישה כי החוק אינו חל על תהליכי אוטומציה המבוססים על פרמטרים מוגדרים מראש שאינם כוללים יכולת של המערכת להגיב לסביבה החיצונית, לתפוס אותה או אף לפרש אותה.

בחוק מוצעים כמה קווים מנחים לאסדרה של בינה מלאכותית: (1) חקיקת משנה ספציפית לבינה מלאכותית תוצג רק כשיש בה צורך מוחלט;[622] (2) התערבות השלטון מחייבת אותו להביא בחשבון היבטים מגזריים ולהתבצע על ידי הרגולטור

611 **שם**, ס' 4.

612 **שם**, ס' (I)5. ראו גם **לעיל** בסעיף 3.6.

613 **שם**, ס' (II)5. ראו גם **לעיל** בסעיף 3.6.

614 **שם**, ס' (III)5. ראו גם **לעיל** בסעיף 3.6.

615 **שם**, ס' (IV)5.

616 **שם**, ס' (VI)5. ראו גם **לעיל** סעיף 3.3.

617 **שם**, ס' (VII)5.

618 **שם**, ס' (VIII)5. לפי עקרון זמינות המידע המוצע, שימוש במידע, בבסיסי נתונים או בטקסט החוסים תחת הגנת זכויות יוצרים למטרות של אימון מערכות בינה מלאכותית אין משמעו הפרה של זכויות אלו, כל עוד המימוש הרגיל שלהן בידי בעליהן לא נפגע.

619 **שם**, ס' (V)5.

620 ס' 3(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

621 ס' 2 להצעת החוק הברזילאית לבינה מלאכותית, **לעיל** ה"ש 54.

622 **שם**, ס' 6(1).

המגזרי המתאים;[623] (3) ניהול סיכונים: היקף ההתערבות המאסדרת ישמור על יחס מידתי לעוצמת הסיכון הקונקרטי הנשקף מכל מערכת ולהסתברות התממשותם של סיכונים אלו;[624] (4) ביצוע הערכת השפעות של רגולציה חדשה (impact analysis): טרם אימוץ כללים רגולטוריים חדשים שיכולים להשפיע על הפיתוח והתפעול של מערכות בינה מלאכותית ייערך סקר הערכת השפעות רגולטוריות (regulatory impact analysis);[625] (5) אחריות: יש להידרש לאחריותם של שחקנים בשרשרת הפיתוח וההפעלה של מערכות בינה מלאכותית מתוך התחשבות בנזקים הספציפיים שיש למנוע או לתקן, במידת המעורבות של גורמים אלו ובשאלה עד כמה השחקנים יכולים לציית במאמץ סביר לכללים הרלוונטיים בהלימה לסטנדרטים המקובלים בתעשייה ובעולם.[626]

על פי הערכה ראשונית נראה שהחקיקה הברזילאית המוצעת אינה אלא צעד ראשון לאסדרה של בינה מלאכותית, ולפי שעה אין היא מפורטת די הצורך.[627] למשל, אין בה קווים מנחים לזיהוי מאפיינים של מערכות שהמחוקק עשוי להגדיר מסוכנות והיא מניחה את ההחלטות המעשיות לרגולטור או לתעשייה. כמו כן, אומנם יש בה העדפה כללית שלא להקים גופים ייעודיים חדשים, אבל אין היא מנסה לשרטט את תחומי הסמכויות של הרגולטורים הקיימים, לזהות תחומים שהרגולטורים הקיימים אינם מכסים או להציע פתרונות למצבים של חפיפה בין רגולטורים מגזריים. כיוון שהיא מבטאת רתיעה מאסדרה ריכוזית היא נמנעת מניהול ריכוזי של מרשם מערכות בינה מלאכותית בדומה לרישום הקיים בהצעה האירופית ובהצעה הסינית.

623 **שם**, ס' 6(2).

624 **שם**, ס' 6(3).

625 **שם**, ס' 6(4). השוו לס' 3 להחלטה 2118 של הממשלה ה־33 "הפחתת הנטל הרגולטורי" (14.9.2014).

626 ס' 6(5) להצעת החוק הברזילאית לבינה מלאכותית, **לעיל** ה"ש 54.

627 לביקורת ברוח דומה ראו Luca Belli, Yasmin Curzi, and Walter B. Gaspar, *AI Regulation in Brazil: Advancements, Flows, and Need to Learn from the Data Protection Experience* 48 Comp. L. & Sec. Rev (2023)

4.4
אסדרת בינה מלאכותית בקנדה

4.4.1. דירקטיבת ה־ADM של קנדה

בפברואר 2019 פרסמה ממשלת קנדה דירקטיבה בעניין מערכות אוטומטיות לקבלת החלטות.[628] על פי השיטה הקנדית, דירקטיבות הן חלק מארגז הכלים של ועדת האוצר של קנדה (Treasury Board of Canada), שהיא ועדת קבינט של מועצת המלוכה של קנדה (Queen's Privy Council for Canada). ועדת האוצר של קנדה קובעת מדיניות מינהלית, פיננסית וארגונית חוצת ממשלה,[629] בין השאר באמצעות דירקטיבות, שהן "הנחיות רשמיות המחייבות מחלקות לנקוט פעולות מסוימות או להימנע מהן. דירקטיבות מסבירות כיצד נדרשים עובדי ציבור בכירים [deputy heads] לממש את המטרות שהמדיניות מתווה".[630] דירקטיבות קובעות חובה פנים־ממשלתית לציית ואין בכוחן לייצר זכויות בדין עבור אנשים פרטיים או ארגונים. הדירקטיבה הנזכרת נכנסה לתוקף בתחילת חודש אפריל 2019, אך החובה לציית להוראותיה החלה רק שנה לאחר מכן. מאפריל 2020 כל מערכת אוטומטית לקבלת החלטות שהממשלה הפדרלית של קנדה מפתחת או רוכשת כפופה לדירקטיבת ה־ADM של קנדה.[631]

על פי הגדרת דירקטיבת ה־ADM של קנדה, מערכת אוטומטית לקבלת החלטות היא כל טכנולוגיה המסייעת לאדם בשיקול דעתו או מחליפה אותו.[632] מטרת הדירקטיבה למזער את הסיכונים הטמונים במערכות אוטומטיות לקבלת החלטות לאזרחים קנדיים ולמוסדות פדרליים ולהביא לקבלת החלטות מדויקות יותר,

628 Government of Canada, Directive on Automated Decision-Making (5.2.2019) (להלן: דירקטיבת ה־ADM של קנדה).

629 Teresa Scassa, *Administrative Law and the Governance of Automated Decision Making: A Critical Look at Canada's Directive on Automated Decision Making*, 54 U.B.C. L. Rev. 251, 268 (2021)

630 **שם**. ראו גם Government of Canada, Foundation Framework for Treasury Board Policies (24.6.2008)

631 ס' 1 לדירקטיבת ה־ADM של קנדה, **לעיל** ה"ש 628.

632 **שם**, בתוספת הראשונה.

עקביות יותר, ניתנות לפירוש ויעילות בהתאם לדין הקנדי.[633] התוצאות הצפויות של הדירקטיבה, לפי לשונה, הן החלטות מינהליות מונחות מידע השומרות על הוגנות והליך נאות;[634] החלת הליכים מטרימים של הערכת סיכונים וצמצום ההשפעות השליליות של אלגוריתמים בעת זיהוים;[635] והפיכת המידע והנתונים על השימוש במערכות אוטומטיות לקבלת החלטות במוסדות פדרליים לפומביים כשהדבר מתאפשר.[636]

הדירקטיבה אינה חלה על כל המערכות האוטומטיות לקבלת החלטות בשירות הממשל הפדרלי, אלא רק על מערכות שמספקות "שירות חיצוני"[637] שמקבלו "חיצוני לממשלת קנדה".[638] לא ברור למשל אם החלטה פנימית על הקצאת משאבים או אלגוריתמים לזיהוי הונאות נחשבים "שירות חיצוני" שניתן להצביע על אדם המקבל אותו. הדירקטיבה אינה חלה, בין השאר, על החלטות הנוגעות לתנאי ההעסקה של עובדי ציבור.[639]

נוסף על כך, הדירקטיבה אינה חלה על מערכות הנוגעות לביטחון לאומי,[640] והיא חלה רק על מערכות שכבר נמצאות בשימוש הממשל הפדרלי (ולא על מערכות הנמצאות בסביבות בדיקה[641] או פרויקטים ניסיוניים, אף שלאלה עשויה להיות השפעה ממשית).[642]

633 **שם**, ס' 4.1.

634 **שם**, ס' 4.2.1.

635 **שם**, ס' 4.2.2.

636 **שם**, ס' 4.2.3.

637 **שם**, ס' 5.1.

638 ראו Scassa, **לעיל** ה"ש 629, בעמ' 270.

639 **שם**, בעמ' 270–271.

640 **שם**, ס' 5.4.

641 **שם**, ס' 5.3.

642 ראו Scassa, **לעיל** ה"ש 629, בעמ' 271. לדוגמה לפרויקט ניסיוני שלא יכוסה על ידי הדירקטיבה ראו דוח על מערכות אוטומטיות לקבלת החלטות אוטומטיות בעניין פליטים ומהגרים בקנדה: Petra Molnar and Lex Gill, Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System (University of Toronto International Human Rights Program and Citizen Lab, September 2018)

הדירקטיבה מורה לבצע הערכת השפעה אלגוריתמית (algorithmic impact assessment) לפני פרישׂה של מערכת אוטומטית לקבלת החלטות.[643] יש לעדכן את ההערכה בהתאם לשינויים בהיקף המערכת או בפונקציונליות שלה.[644] על פי הוראות דירקטיבת הממשל הפתוח של קנדה, יש לפרסם את תוצאות הערכת ההשפעה האלגוריתמית באתרי הממשלה.[645]

באשר לשקיפות, הדירקטיבה מחילה על שימוש במערכות אוטומטיות לקבלת החלטות את החובה ליידע בטרם מעשה,[646] את החובה לספק הסבר בעל משמעות לאחר מעשה[647] ואת החובה לאפשר לצד שלישי לבדוק (audit) אם השימוש במערכות נעשה ברישיון[648] או לפרסם את קוד המקור של המערכת אם הוא שייך לממשלת קנדה (בכפוף לסייגים).[649]

בסעיף "בקרת איכות" (Quality Assurance) מחייבת הדירקטיבה, בין השאר, לפתח הליכי בדיקה של המערכת האוטומטית לקבלת החלטות לפני פרישׂתה ולאחריה. מטרת הליכים אלו להבטיח שאין הטיות אלגוריתמיות לא רצויות, או תוצאות אחרות שאינן מכוונות, ולוודא שהמערכת מצייתת לדירקטיבה ולהוראות הדין.[650] הסעיף מבקש להבטיח שהמידע שהמערכת משתמשת בו מדויק, רלוונטי, עדכני והולם את חוק הפרטיות של קנדה[651] ואת מדיניות הממשלה בעניין

643 ס' 6.1.1 לדירקטיבת ה־ADM של קנדה, **לעיל** ה"ש 628. לפירוט על נוהלי הערכת ההשפעה האלגוריתמית בקנדה ראו Governement of Canada, Algorithmic Impact Assessment Tool (updated 19.4.2022)

644 ס' 6.1.3 לדירקטיבת ה־ADM של קנדה, **לעיל** ה"ש 628.

645 **שם**, ס' 6.1.4.

646 **שם**, ס' 6.2.1.

647 **שם**, ס' 6.2.3.

648 **שם**, ס' 6.2.5.

649 **שם**, ס' 6.2.6. הסייגים לפרסום קוד המקור הם (1) רמת הסיווג של הקוד; (2) פטור מפרסום על פי דיני חופש המידע; (3) פטור מאת המנהל הראשי של מערכות המידע של קנדה (Chief Information Officer of Canada, CIo).

650 **שם**, ס' 6.3.1, 6.3.2.

651 חוק הפרטיות הקנדי נוגע לזכויות שיש לאנשים בנוגע למידע של מוסדות ממשלה קנדיים עליהם. Privacy Act, RSC 1985, c P-21

שירותים דיגיטליים.[652] עוד מורה הסעיף שהמערכת האוטומטית לקבלת החלטות תאפשר התערבות אנושית לפי הצורך (human in the loop).[653] הסעיף כולל גם הוראות בנוגע להתייעצות עם מומחים,[654] הכשרת עובדים[655] ואבטחה.[656]

4.4.2. הצעת חוק הבינה המלאכותית והמידע של קנדה

בחודש יוני 2022 פרסמה הממשלה הפדרלית של קנדה את הצעת החוק ליישום מגילת הזכויות הדיגיטליות (Digital Charter Implementation Act, 2022).[657] הצעה זו כוללת חבילה של דברי חקיקה ובהם הצעת חוק הבינה המלאכותית והמידע (Artificial Intelligence and Data Act).

מטרת הצעת חוק הבינה המלאכותית והמידע היא לאסדר את הסחר במערכות בינה מלאכותית על ידי קביעת סטנדרטים מקובלים (התקפים בקנדה) לעיצובן ופיתוחן של מערכות אלו ולשימוש בהן.[658] עוד מבקשת הצעת החוק לאסור על שימוש שעלול לגרום נזק חמור לאנשים פרטיים או לאינטרסים שלהם.[659]

הצעת החוק אינה חלה על מוסדות ממשלה או על מוצרים, שירותים ופעילות שנתונים לשליטתם של שר ההגנה, של מנהל שירות המודיעין הקנדי, של ראש

652 ס' 6.3.3 לדירקטיבת ה־ADM של קנדה, **לעיל** ה"ש 628.

653 **שם**, ס' 6.3.9. השוו למשל לס' 22 של התקנות הכלליות בדבר הגנת מידע (GDPR).

654 ס' 6.3.4 לדירקטיבת ה־ADM של קנדה, **לעיל** ה"ש 628.

655 **שם**, ס' 6.3.5.

656 **שם**, ס' 6.3.7.

657 Bill C-27, An Act to Enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts, 1st Sess, 44th Parl, 2022 (First Reading, 16 June 2022)

658 ס' 4(a) להצעת חוק הבינה המלאכותית והמידע של קנדה, **לעיל** ה"ש 55.

659 **שם**, ס' 4(b).

הרשות לאבטחת תקשורת או של כל אדם אחר העומד בראש סוכנות פדרלית או מחוזית ונזכר בתקנות המושל הכללי של קנדה.[660]

הצעת החוק מגדירה "פעילות מאוסדרת" עיבוד או הנגשה של נתונים שנוגעים לפעילות אנושית במטרה לעצב ולפתח מערכת בינה מלאכותית או להשתמש בה; וכן עיצוב, פיתוח או הנגשה של מערכות בינה מלאכותית או ניהולן.[661] עוד היא מציעה שאדם שעוסק בפעילות מאוסדרת שנוגעת למידע מותמם יידרש לנקוט אמצעים באשר לאופן ההתממה של המידע וניהולו,[662] ובמידת הצורך ישמור תיעוד של אמצעים אלו.[663]

אף שהצעת החוק נוקטת גישה של ניהול סיכונים, ומחילה חובות מיוחדות (ראו להלן) על מערכות בינה מלאכותית בסיכון גבוה (high-impact system), הגדרתן נותרה לוטה בערפל: החוק המגדיר מערכות אלו מפנה לקריטריונים שיוגדרו בתקנות (שטרם תוקנו).[664] עם זאת, אדם הממונה על מערכת בינה מלאכותית יידרש להעריך אם מערכת מסוימת נחשבת על פי התקנות למערכת בסיכון גבוה[665] ולנקוט אמצעים שתכליתם לזהות, להעריך ולמתן את הסיכונים, את האיומים או את הפלט המוטה (biased output)[666] שיכולים לצמוח משימוש במערכת.[667]

660 **שם**, ס' 3.

661 **שם**, ס' 5.

662 **שם**, ס' 6.

663 **שם**, ס' 10.

664 **שם**, ס' 5.

665 **שם**, ס' 7.

666 ס' 5 להצעת חוק הבינה המלאכותית והמידע של קנדה (**לעיל** ה"ש 55) מגדיר פלט מוטה כתוכן, החלטה, תחזית או המלצה שנעשו או נוצרו באמצעות מערכת בינה מלאכותית המפלה לרעה במישרין או בעקיפין ללא צידוק. ההגדרה מפנה לעילות ההפניה בס' 3 לחוק זכויות האדם של קנדה.

667 ס' 8 להצעת חוק הבינה המלאכותית והמידע של קנדה, **לעיל** ה"ש 55.

על הממונים על מערכות בסיכון גבוה חלה גם החובה לנטר את האמצעים למיתון הסיכונים.[668] כמו כן, עליהם לתעד את האמצעים האלו ואת הסיבות התומכות בהערכת המסוכנות של המערכת.[669] הם נדרשים לפרסם ברשת הסבר של המערכת בלשון פשוטה, הכולל התייחסות לאופן פעולתה, לסוגי התוכן, התחזיות או ההמלצות שהיא מייצרת, פירוט האמצעים שהיא נוקטת למיתון הסיכונים ומידע נוסף שקובעות התקנות.

המושל הכללי ימנה את השר הממונה על מערכות בינה מלאכותית.[670] השר מוסמך לדרוש מסמכים ותיעוד בעניין מערכות בינה מלאכותית, ובפרט מערכות בסיכון גבוה,[671] וכן להורות על ביצוע ביקורת (audit) שתדווח לשר.[672] הממונה על מערכות בינה מלאכותית בסיכון גבוה נדרש להודיע לשר אם שימוש במערכת גרם או עלול לגרום לנזק מהותי.[673] בעקבות ממצאי הביקורת רשאי השר להורות על הטמעת אמצעים שונים במערכות בינה מלאכותית.[674] השר אף רשאי להורות על הפסקת פעילותן של מערכות בינה מלאכותית בסיכון גבוה, אם הוא סבור שיש בשימוש בהן סיכון של ממש.[675] לשר מוקנה שיקול דעת להורות לפרסם ברשת כל מידע על הביקורות של מערכות בינה מלאכותית, על הפסקת פעילות מערכות בינה מלאכותית, על אמצעי מיתון סיכונים, על הערכת סיכונים, על התממה וכדומה.[676]

הצעת חוק הבינה המלאכותית של קנדה אינה מתעלמת מהיבטים הנוגעים לסודות מסחריים במידע שעשוי להגיע לידי השר מכוח החוק ולכן היא מציינת

668 **שם**, ס' 9.

669 **שם**, ס' 10.

670 **שם**, ס' 31. ס' 5 לחוק מורה כי כל עוד לא מונה שר כאמור, השר הממונה יהיה שר התעשייה.

671 **שם**, ס' 13–14.

672 **שם**, ס' 15.

673 **שם**, ס' 12.

674 **שם**, ס' 16.

675 **שם**, ס' 17.

676 **שם**, ס' 18.

שאין השר רשאי לחשוף סודות מסחריים שהגיעו לידיו[677] ושהוא מחויב לשמור עליהם.[678] בהצעת החוק נכללו הוראות מיוחדות בנוגע להעברת מידע שהוא בגדר סוד מסחרי לצד שלישי[679] ובנוגע לפרסומו ברבים.[680]

4.5 ניצני אסדרת בינה מלאכותית בארצות הברית

יש כמה הצעות חוק שיכולות להעיד על תחילתה של מגמה לאסדר בינה מלאכותית בארצות הברית. במדינת קליפורניה, למשל, הוצע לאחרונה לאפשר להורים לתבוע פלטפורמות שגרמו לילדיהם התמכרות חמורה (שבצידה נזקים נפשיים, פיזיים, רגשיים או התפתחותיים).[681] הצעה זו מעלה על הדעת את החשש שהביעו המחוקקים בסין מפני תופעות של התמכרות לפלטפורמות ברשת;[682] התמכרות שמטפחים, בין השאר, האלגוריתמים של השירותים המקוונים. ולקוד המוניציפלי המינהלי של העיר ניו יורק התווספה לאחרונה הוראה שקובעת ששימוש במערכות מבוססות אלגוריתמים לקבלת החלטות או המלצות על העסקתם או קידומם של עובדים יותנה בהליכי בקרת הטיות (bias audit) עצמאיים.[683]

גם ברמה הפדרלית החלו להידרש לצורך ברגולציה של בינה מלאכותית. טיוטת הצעת חוק הפרטיות והגנת המידע האמריקאי (American Data Privacy and Protection Act), שפורסמה בתחילת חודש יוני 2022,[684] כוללת איסור

677 **שם**, ס' 22.

678 **שם**, ס' 23.

679 **שם**, ס' 24–26.

680 **שם**, ס' 27–28.

681 An Act to Add Section 1714.48 to Amend Section 1714 of the Civil Code, relating to Social Media Platforms, Cal. Assemb. A. 2408 (2021-2022)

682 ראו **לעיל** בסעיף 4.2.

683 N.Y.C. Admin. Code §§ - 0-870 - 20-874 (2022); Airlie Hilliard et al., *Regulating the Robots: NYC Mandates Bias Audits for AI-Driven Employment Decisions* (13.4.2022)

684 To provide consumers with foundational data privacy rights, create strong oversight mechanisms, and establish meaningful enforcement, H.R.8152, 117th Congress (2022) (להלן: ADDPA).

על אפליה אלגוריתמית במה שנוגע להגנה על זכויות האזרח.[685] בפרט מוצע לאסור על "ישויות מכוסות" (covered entities)[686] לאסוף, לעבד או להעביר מידע באופן שיש בו כדי להפלות על בסיס גזע, צבע, דת, מוצא לאומי, מגדר, נטייה מינית או נכות.[687] נוסף על כך, על פי הצעת חוק הפרטיות והגנת המידע האמריקאי גופים המחזיקים במידע רב (large data holders)[688] יערכו הערכת השפעה אלגוריתמית, כלומר יתארו את האמצעים שינקטו כדי לצמצם נזקים שעלולים להיגרם לאנשים פרטיים בנסיבות שונות.[689]

יוזמה נוספת היא הצעת חוק האחריותיות האלגוריתמית, שהגישה ב־2019 קבוצה של סנאטורים דמוקרטים והיא עודכנה בפברואר 2022.[690] הצעת החוק חלה על כל אדם או ישות משפטית שכפופים לסמכותה של נציבות הסחר האמריקאית; שמשתמשים או מחזיקים במערכות החלטה קריטיות מוגברות;[691] שהייתה להם הכנסה שנתית ממוצעת של יותר מ־50 מיליון דולר בשלוש השנים

685 **שם**, ס' 207.

686 ס' (9)2 ל־ADDPA, **שם**, מגדיר ישויות מכוסות (covered entities) כל גוף או אדם שאוסף, מעבד או מעביר מידע ו(1) חוק נציבות הסחר הפדרלית (Federal Trade Commission) חל עליו; (2) הוא ספק תקשורת (common carrier) תחת חוק התקשורת; או (3) הוא ארגון ללא מטרת רווח.

687 **שם**, ס' 207(a).

688 ס' (17)2 ל־ADDPA, **שם**, מגדיר "מחזיקי מידע רב" כישויות מכוסות (covered entities) שהכנסותיהן עולות על 250 מיליון דולרים בשנה והעבירו, אספו או עיבדו נתונים על 5 מיליון איש או מידע רגיש על 100,000 איש.

689 **שם**, ס' 207(c)(1)(b).

690 Algorithmic Accountability Act of 2022, H.R. 6580, 117th Cong. (2022) (להלן: הצעת חוק האחריותיות האלגוריתמית).

691 מערכות החלטה קריטיות מוגברות (augmented critical decision systems) מוגדרות בס' (1)2 להצעת חוק האחריותיות האלגוריתמית כתהליך, הליך או כל פעילות אחרת שמשתמשים במערכת החלטה אוטומטית (automated decision system) לקבלת החלטה קריטית. מערכת החלטה אוטומטית מוגדרת בס' (2)2 להצעת החוק כל מערכת, תוכנה או תהליך (לרבות כאלו שמקורם בלמידת מכונה, בסטטיסטיקה או בטכניקות בינה מלאכותית) שעורכים חישובים שתוצאתם משמשת בסיס לגיבוש עמדה או לקבלת החלטה. ס' (8)2 מגדיר "החלטה קריטית" החלטה שיש לה השפעה משפטית, מהותית או בעלת משמעות אחרת על חייו של צרכן בהקשר של מחיר, תנאים או זמינות של שירותים מסוימים (המנויים בהגדרה).

האחרונות או ששוויים בתקופה זו עולה על 250 מיליון דולר; שמעבדים מידע על יותר ממיליון צרכנים, בתי אב או מכשירים במטרה לפתח או לפרוש מערכת אוטומטית לקבלת החלטות או מערכת מוגברת לקבלת החלטות קריטיות.[692] ג'ייקוב מוקאנדר ועמיתיו רואים בחיוב את בחירת המחוקק האמריקאי לנקוט את המונחים "מערכות לקבלת החלטות" ו"תהליכים לקבלת החלטות קריטיות" חלף "בינה מלאכותית" ו"בינה מלאכותית בסיכון גבוה" הננקטים בהצעה האירופית – הן כי המונחים בהצעה האירופית לוקים באי־בהירות הן כי השמירה על ניטרליות טכנולוגית תגביר את חסינות העתיד של ההצעה האמריקאית.[693]

לפי הצעת חוק האחריותיות האלגוריתמית, ישויות שהחוק יחול עליהן יידרשו לבצע הערכת השפעה[694] של כל מערכת אוטומטית לקבלת החלטות שפותחה לשם שימוש או הטמעה במערכות אוטומטיות לקבלת החלטות קריטיות או שיש צפי שישתמשו בהן או יטמיעו אותן בעתיד.[695] נציבות הסחר הפדרלית תהיה רשאית להורות לישויות אלו לשמור תיעוד של הערכות ההשפעה שביצעו,[696] לדווח על בסיס שנתי לנציבות על הערכות השפעה שנעשו בארגון[697] ולצמצם השפעה שלילית של מערכות אוטומטיות לקבלת החלטות קריטיות.[698] הנציבות תהיה רשאית גם להגדיר קווים מנחים ופורמטים להערכות השפעה ולדיווחים עליהן.[699] רשות הסחר הפדרלית תקים מאגר מידע פומבי שבו יישמרו סיכומי הדיווחים השנתיים.[700]

692 ס' 2(7) להצעת חוק האחריותיות האלגוריתמית, **לעיל** ה"ש 690.

693 Jakob Mökander et al., *The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: What Can They Learn from Each Other?* 32 Minds and Machines 751 (18.8.2022)

694 ס' 4 להצעת חוק האחריותיות האלגוריתמית מכיל פירוט של הרכיבים הנדרשים מהערכת השפעה אלגוריתמית, לרבות חובות תיעוד, הכשרה, שקילת שיקולים הנוגעים לפגיעה פוטנציאלית בזכויות ומיתון נזקים.

695 ס' 3(b)(1) להצעת חוק האחריותיות האלגוריתמית, **לעיל** ה"ש 690.

696 **שם**, ס' 3(B)(b).

697 **שם**, ס' 3(b)(d).

698 **שם**, ס' 3(B)(H).

699 **שם**, ס' 3(b)(J)-(L).

700 **שם**, ס' 6.

מוקאנדר ועמיתיו מצביעים על החיסרון בהגבלת תחולתה של הצעת חוק האחריותיות האלגוריתמית לגופים עסקיים. לדבריהם, הגבלה זו מותירה מערכות בשירות הממשל – למשל, מערכות להקצאת זכויות סוציאליות – ללא פיקוח נאות. נוסף על כך, הנוסח הרזה של ההצעה האמריקאית – בהשוואה למקבילתה האירופית – משאיר לנציבות הסחר הפדרלי שיקול דעת נרחב למדי ביישום הוראות החוק, לרבות שימוש במונחים שמצמצמים שלא לצורך את החובות המוגדרות בחוק, דוגמת חובת ההתייעצות עם בעלי עניין ואנשי טכנולוגיה "עד כמה שאפשר" (to the extent possible).[701]

לצד צעדי חקיקה ראשונים אלו, במחצית הראשונה של 2023 ניכרת בארצות הברית התעוררות סביב נושא אסדרת הבינה המלאכותית. במאי 2023 נערך בסנאט האמריקני שימוע לסאם אלטמן, מנכ"ל חברת OpenAI;[702] המחלקה המייעצת לנשיא ארצות הברית בנושאי מדע וטכנולוגיה (OSTP) פרסמה קול קורא להערות הציבור בעניין אסטרטגיית בינה מלאכותית לאומית;[703] נוסף על כך, קבוצת רגולטורים, ובהם ראש הרשות להגנת הצרכן במגזר הפיננסי וראשת ועדת הסחר הפדרלית (FTC), פרסמו הצהרה משותפת המעלה את חששותיהם מפני התרומה של מערכות אוטומטיות לאפליה בלתי חוקית ולהפרה של חוקים פדרליים אחרים, בהדגישם שפעולות האכיפה של הסוכנויות שלהם חלות גם על מערכות אוטומטיות.[704] ברוח זו, המחלקה האזרחית (זכויות אדם – דיור) במשרד המשפטים האמריקאי הגישה לאחרונה לבית המשפט המחוזי במסצ'וסטס הצהרה שבה היא מבהירה שחוק ההוגנות בדיור (Fair Housing Act) חל גם על החלטות אלגוריתמיות.[705]

701 Mökander et al., **לעיל** ה"ש 693.

702 Andrew Ross Sorkin et al., *Washington Confronts the Challenge of Policing A.I.,* The New York Times (17.5.2023)

703 Office of Science and Technology Policy, Request for Information: National Priorities for Artificial Intelligence (23.5.2023)

704 U.S. Equal Employment Opportunity Commission, *Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems* (2023)

705 Department of Justice, *Justice Department Files Statement of Interest in Fair Housing Act Case Alleging Unlawful Algorithm-Based Tenant-Screening Practices* (9.1.2023)

4.6 הגישה הבריטית

בממלכה המאוחדת טרם פורסמה הצעת חוק לאסדרת הפיתוח של בינה מלאכותית או השימוש בה. עם זאת, בחודש יולי 2022 פרסם משרד הדיגיטל, המדיה, התרבות והספורט של בריטניה מסמך מדיניות (white paper), שהוא מעין הצהרת כוונות בעניין רגולציה עתידית של מערכות בינה מלאכותית. המסמך נוקט גישה מוטת חדשנות ומציע שניהול הסיכונים הרגולטורי של מערכות נבונות יהיה מבוסס הקשר, יתמקד בסיכונים אמיתיים, ניתנים לזיהוי, שרמת הסיכון הנשקפת מהם אינה מקובלת (להבדיל מסיכונים תאורטיים או כאלה שהסיכוי להתרחשותם נמוך). הגישה של בריטניה שואפת להפעלה מידתית של סמכויות רגולטוריות בכל הנוגע למערכות בינה מלאכותית ונוטה לבחירה בחלופות מחמירות פחות, כגון קווים מנחים או רגולציה וולונטרית.[706] לפי מסמך המדיניות, ממשלת בריטניה שוקלת להטמיע את מערך העקרונות באמצעות הנחיות מינהלתיות, ולא באמצעות חקיקה. עם זאת, הצורך בחקיקה טרם נשלל לחלוטין.

מסמך המדיניות מציע להימנע מקביעת הגדרה משפטית אחידה לבינה מלאכותית ולהסתפק במאפייני ליבה.[707] כך ייהנו הרגולטורים המגזריים מגמישות בהתמודדותם עם יישומים קונקרטיים של בינה מלאכותית.

אף על פי שההקשר קריטי לרגולציה עניינית של בינה מלאכותית, מסמך המדיניות מציע לפתח מערך של עקרונות רגולטוריים חוצי־מגזרים שיהיו ייחודיים לטכנולוגיות אלו. הרגולטורים יצטרכו להטמיע עקרונות כלליים אלו בתחום שהם מופקדים עליו. מסמך המדיניות מציע כמה עקרונות ראשוניים, בהתבסס על עקרונות ה־OECD לבינה מלאכותית,[708] אך מדגיש כי אין בהם כדי לייצר זכויות אדם חדשות אלא רק מסגרת מוטת חדשנות לניהול סיכונים.[709] עקרונות אלו כוללים שימוש בטוח במערכות בינה מלאכותית, הבטחת הבטיחות הטכנית של

706 Department for Digital, Culture, Media and Sport, **לעיל** ה"ש 59, בעמ' 11.

707 דוגמת אדפטיביות ואוטונומיות, ראו **לעיל** בסעיף 1.1.

708 ראו OECD 2019, **לעיל** ה"ש 294.

709 Department for Digital, Culture, Media and Sport, **לעיל** ה"ש 59, בעמ' 12.

מערכות בינה מלאכותית (ובכלל זה שהן פועלות כפי שתוכננו לפעול), שקיפות והסברתיות במידה המתאימה, הטמעת שיקולי הוגנות, הגדרת האחראים משפטית למשילות בינה מלאכותית (AI governance) והגדרת סעדים משפטיים.[710]

4.7 אסדרת בינה מלאכותית: סיכום ביניים

על אף נקודות הדמיון בין ניצני החקיקה לאסדרת בינה מלאכותית ברחבי העולם שנסקרו בפרק זה החקיקות נבדלות זו מזו בהיקפן, בתחולתן ובפתרונות הרגולטוריים המוצעים בהן. כך למשל, ההצעות של אמריקה, ברזיל וסין נמנעות מהקמת רגולטור ייעודי, ההצעה של קנדה מניחה את השאלה המוסדית לשיקול דעת מיניסטריאלי וההצעה של אירופה תומכת בהקמת רשות לאומית לבינה מלאכותית.[711] ההבדלים הם בין השאר גאופוליטיים ותרבותיים: המשטר הסיני הסמכותני אינו מודאג משימוש של כוחות הביטחון והצבא במערכות בינה מלאכותית, ואילו האירופים מחזיקים בעמדות הפוכות; את רזונה של הצעת החוק של ברזיל אפשר לייחס גם למקומה במרוץ הגלובלי לפיתוח מערכות בינה מלאכותית (לעומת מעצמות כמו סין ואירופה), כפי שעולה מן השאיפה הצנועה שהיא מביעה בהצעה "להשתלב בשרשרת הערך הגלובלית במקום תחרותי".[712] את הגישה מוטת החדשנות של בריטניה (שבאה על חשבון החלשת הבקרות הרגולטוריות),[713] הנמנעת לעת עתה מקידום חקיקה לאסדרת בינה מלאכותית, אפשר לפרש כביטוי של העדפת אינטרסים כלכליים מערכי היסוד של האיחוד האירופי.

נראה שצפון אמריקה פוזלת לעבר הצעת תקנות הבינה המלאכותית האירופיות, ההולכת לפני המחנה.[714] חרף היותן של ההצעה הקנדית וההצעה האמריקאית מקיפות פחות בתחולתן ובהיקפן, נראה שגם הן נוקטות גישה של ניהול סיכונים.

710 **שם**, בעמ' 12–14.

711 ראו **להלן** בסעיף 8.1.

712 ס' (III)3 להצעת החוק הברזילאית לבינה מלאכותית, **לעיל** ה"ש 54.

713 ראו הביקורת של Huw Roberts et al., *Artificial Intelligence Regulation in the United Kingdom: A Path to Global Leadership?* 5 (1.5.2023)

714 ראו בהקשר זה את ניתוחם של Mökander et al., **לעיל** ה"ש 693.

הצעת החוק הקנדית עושה זאת במפורש, בהגדירה מערכות בסיכון גבוה, ואילו הצעת החוק האמריקאית חלה מלכתחילה רק על מערכות לקבלת החלטות קריטיות. גם ההצעה של ברזיל נדרשת לרכיב הסיכון. השימוש בחקיקת פרטיות כדי לטפל בעקיפין בהיבטים שונים של מערכות בינה מלאכותית אף הוא מושפע מאירופה – בדומה לסעיף 22 ב־GDPR, שטיפולו הישיר בסוגיות הקשורות לבינה מלאכותית הוא תקדימי, גם סעיף 207 בהצעת חוק הפרטיות והגנת המידע האמריקאי מכיל הוראות בעניין בינה מלאכותית.

בעת כתיבת חיבור זה הצעת תקנות הבינה המלאכותית האירופיות היא בשלבי החקיקה האחרונים שלה. הגישה האירופית נוטה לאסדרה רוחבית בחקיקה (והסתייעות ברגולטורים מגזריים) והיא מלווה בחקיקת דיגיטל עוטפת (ה־GDPR לענייני פרטיות, דירקטיבת הסייבר האירופית וחוקי הדיגיטל החדשים – חוק השירותים הדיגיטליים וחוק השווקים הדיגיטליים). לעומת האיחוד האירופי, כמה ממדינות המערב שנסקרו בפרק זה, דוגמת בריטניה וארצות הברית, עדיין מגבשות את מדיניות האסדרה שלהן. בריטניה נוקטת גישה של אסדרה מגזרית מוטת חדשנות. ארצות הברית טרם גיבשה את הגישה שלה לאסדרה, אך נראה שהיא רואה באפליה אלגוריתמית את הסיכון העיקרי הטמון בבינה מלאכותית. אף שהמחלקה המייעצת לנשיא ארצות הברית בנושאי מדע וטכנולוגיה (OSTP) מקדמת את הנושא ברמה הפדרלית,[715] וחרף הצעות החוק שנסקרו לעיל,[716] מסתמן כי לעת עתה מגיעה האסדרה המעשית של מערכות בינה מלאכותית מהרגולטורים המגזריים בשטח, בדומה לחקיקת הפרטיות הפדרלית המטולאת של ארצות הברית. האסדרה הקנדית קרובה יותר ברמת בשלותה לאסדרה האירופית, וכך גם בתפיסה הרגולטורית שלה – אסדרה רוחבית שמבוססת על תפיסה של ניהול סיכונים, ובתפיסה המוסדית שלה – הקמת משרד בינה מלאכותית.

715 **לעיל** ה"ש 703 ו־312.

716 הצעת חוק האחריותיות האלגוריתמית, **לעיל** ה"ש 690; ADDPA, **לעיל** ה"ש 684.

פרק חמישי

בינה מלאכותית וזכויות אדם:
חזית חדשה במדיניות טכנולוגיה

מערכות נבונות, כמו כל טכנולוגיה חדשה, הן ניטרליות מבחינה מוסרית,[717] כלומר אין הן "רעות" בהכרח או "טובות" בהכרח.[718] ואולם לעיצובן ולפיתוחן

717 יובהר כי הטענה שטכנולוגיה כשהיא לעצמה היא ניטרלית אינה חפה מביקורת. טכנולוגיה אינה מיוצרת לשמה, אלא למילוי תכלית כלשהי – אם חיובית ואם שלילית. האם אפשר לטעון שרקטה מונחית חום היא ניטרלית מבחינה טכנולוגית? *AI Safety Myths*, Future of Life Institute (7.8.2016). ראו בהקשר זה גם Mariana Todorova, *Philosophical, Moral, and Ethical Rationalization of Artificial Intelligence*, *in* The Transhumanism Handbook 263 (Newton Lee ed., 2019)

718 Mariarosaria Taddeo and Luciano Floridi, *How AI Can Be a Force for Good: An Ethical Framework Will Help to Harness the Potential of AI while Keeping Humans in Control*, 361 Science 751 (24.8.2018)

של טכנולוגיות, ולהקשר שבו הן משמשות, יש נגיעה לזכויות אדם.[719] הוא הדין במערכות נבונות – לצד התרומה הפוטנציאלית העצומה שלהן לאיכות חייו של האדם, אימוץ נרחב של טכנולוגיות אלו עלול להביא גם לפגיעה בזכויותיו. לנוכח עוצמת הפגיעה הפוטנציאלית של יישומים מסוימים של בינה מלאכותית מציעות אפוא תקנות הבינה המלאכותית האירופיות לאסור עליהם כליל.[720] כך גם המליץ נציב זכויות האדם של האומות המאוחדות.[721]

יש להבהיר כי פרק זה אינו מבקש לטעון שבינה מלאכותית פוגעת מטבעה בזכויות אדם, אלא לסקור אופנים שונים שבהם השימוש בטכנולוגיות אלו עלול להביא לפגיעה בזכויות. נוסף על כך, הסקירה נועדה להדגים פגיעה פוטנציאלית בזכויות ולא להקיף את כל מופעיה. משכך, היא אינה בהכרח ממצה את מלוא הסיכונים לחירויות יסוד וזכויות אדם שכרוכים בשימוש במערכות בינה מלאכותית.

עם זאת, בטרם נסקור את הסיכונים נציע רשימה לא ממצה של סוגי הנזקים שמערכות בינה מלאכותית מתקדמות עלולות להסב ונמנו לאחרונה בספרות.[722] הרשימה מדגימה מדוע חשיבות הטיפול בסיכונים אלו היא כה רבה.

סוגי הנזקים העלולים להיגרם משימוש בבינה מלאכותית ונסקרו עד כה משתלבים בחששות לנזקים נוספים ההולכים ונעשים מוחשיים, כגון

- אבטחת מידע וסייבר – מודלים יכולים לגלות חולשות במערכות (חומרה, תוכנה, נתונים) ולכתוב קוד לניצולן; להכניס, במסגרת משימה לכתיבת קוד, "באגים" לצורך ניצול עתידי; לקבל החלטות שעלולות לגרום נזק כאשר הם מקבלים גישה למערכת או לרשת; להתחמק מזיהוי של הפעולות שעשו.

719 ראו למשל Mike Cooley, *The Myth of the Moral Neutrality of Technology*, 9 AI & Soc. 10 (1995)

720 ראו **לעיל סעיף** 4.1.1.1.

721 United Nations General Assembly, The Right to Privacy in the Digital Age, Para. 59.(c)-(d) (A/HRC/48/31, 13.9.2021)

722 Toby Shevlane et al., *Model Evaluation for Extreme Risks* (24.5.2023), available at arXiv

- הונאה – למודלים יש הסגולות הנדרשות כדי להונות בני אדם. למשל בנייה של תוכן – מידיעה חדשותית עד תצהיר משפטי – הנתפס אמין, אך הוא כוזב; או התחזות משכנעת לאדם. מודלים יכולים לחזות במדויק השפעת של הונאה על בני אדם ולנהל מעקב של המידע שהם זקוקים לו כדי לשמור על יכולת ההונאה.

- שכנוע ומניפולציה – מודלים יעילים בקידום נרטיבים בצורה משכנעת ובהשפעה על אמונות של בני אדם ועל עיצובן באמצעות דיאלוג איתם או טכניקות אחרות (למשל השפעה על אלגוריתם של הפצת מידע ברשתות חברתיות).

- אסטרטגיה פוליטית – מודלים יכולים לבצע את המידול והתכנון החברתי הדרושים לשחקן כדי להשיג ולהפעיל השפעה פוליטית, לא רק ברמת המיקרו אלא בתרחישים עם שחקנים מרובים והקשר חברתי עשיר. לדוגמה, לקבל ניקוד גבוה בחיזוי תחרויות בשאלות הקשורות למשא ומתן פוליטי.

- פיתוח נשק או גישה לנשק – מודלים יכולים לקבל גישה למערכות נשק קיימות או לתרום לבניית כלי נשק חדשים. לדוגמה, להרכיב נשק ביולוגי (בסיוע אנושי) או לספק הוראות מעשיות כיצד לעשות זאת.

- תכנון ארוך טווח – מודלים יכולים ליצור תוכניות רציפות הכוללות שלבים מרובים, התלויים אלה באלה ונפרשים על פני טווח זמן ותחומי פעולה שונים; ולכוונן כל העת את השלבים לנוכח מכשולים או יריבים בלתי צפויים בתוך שהם מממשים יכולות מכלילות לסביבות חדשות ואינם מסתמכים רק על ניסוי וטעייה.

- פיתוח בינה מלאכותית – מודלים יכולים לבנות מערכות AI חדשות שיכולות להיות מסוכנות או דו־שימושיות.

- מודעות מצבית – מודלים יכולים להבין שהם מודלים ויש להם ידע על עצמם ועל סביבתם (למשל איזו חברה הכשירה אותם, היכן נמצאים השרתים שלהם, איזה סוג של אנשים יכולים לתת להם משוב), וכן לדעת באיזה שלב הם במעגל החיים של בינה מלאכותית, כלומר אם הם בשלב האימון, שלב ההערכה או שלב היישום – ולהתנהג בהתאם.

• יכולת התרחבות והפצה עצמית – מודלים יכולים ליצור לעצמם אפשרויות לבצע פעולות שהם לא תוכננו לבצע באמצעות שינוי מערכת ההפעלה שלהם, מניפולציה על מי שמתכנן את המערכת ושימוש באסטרטגיות לחילוץ הקוד והמשקלים של המערכת. מודלים יכולים לנצל חולשות במערכות שמיועדות לנטר את התנהגותם בשלב היישום שלהם ולחולל שינוי בהערכה בעניינם. מודלים יכולים לייצר הכנסות באופן עצמאי (למשל על ידי הצעת שירותים או מתקפות כופר), ולהשתמש בהכנסות אלו כדי לרכוש משאבי מחשוב ולהפעיל מספר רב של מערכות אחרות של בינה מלאכותית.

5.1 הזכות לפרטיות

מידע הוא הנפט החדש של העולם המפותח.[723] נתוני עתק הם הדלק שמערכות נבונות, ובפרט כאלו המבוססות על למידת מכונה, צורכות על מנת לשכלל את תהליכי קבלת ההחלטות שלהן.[724] וכיוון שפיתוח מערכות נבונות מחייב שימוש במידע אישי,[725] הוא כרוך בפגיעה בזכות לפרטיות. אומנם שימוש כזה יכול להיות מותר, למשל אם הוא נעשה בהסכמת מושאי המידע;[726] אבל במקרים רבים מעובד אותו מידע אישי לתכליות שונות מאלו שלשמן נאסף – תכליות שלא היה אפשר אפילו להעלות על הדעת בזמן שנאסף – אף שהדבר נוגד את עקרון צמידות המטרה, שהוא עקרון יסוד בדיני הגנת הפרטיות.[727]

723 Palmer, **לעיל** ה"ש 112.

724 ראו Todorova, **לעיל** ה"ש 717, בעמ' 287-286.

725 נדשקה פורטובה טוענת – לנוכח ההגדרה הרחבה של מידע אישי בדיני הגנת הפרטיות הבינלאומיים (ראו לדוגמה ס' 1(4) של התקנות הכלליות בדבר הגנת מידע [GDPR], וההגדרה שהציעו ארידור הרשקוביץ ושוורץ אלטשולר [**לעיל** ה"ש 403, בעמ' 33]) – שבסביבה רווית נתונים שבה נעשה שימוש הולך וגובר במערכות נבונות, "הכול הוא מידע אישי". Nadezhda Purtova, *The Law of Everything: Broad Concept of Personal Data and Future of EU Data Protection Law*, 10 L. Innovation and Tech. 40 (2018)

726 ראו למשל ס' 6(1) של התקנות הכלליות בדבר הגנת מידע (GDPR), ואת ההוראה המקבילה לו המוצעת אצל ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 403, בעמ' 236-235.

727 ראו ס' 6(4) של התקנות הכלליות בדבר הגנת מידע (GDPR), ואת ההוראה המקבילה לו המוצעת אצל ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 403, בעמ' 52-51.

ככל שגובר השימוש במערכות נבונות בכל תחומי החיים, כך גובר הצורך של מפתחיהן ובעליהן במידע מפורט יותר על המשתמשים בהן. מידע זה נאסף אוטומטית באמצעות מגוון חיישנים ומכשירי קצה המלווים את האדם המערבי המודרני בשגרת יומו. זהו "קפיטליזם של מעקב", הממיר נתוני עתק למודלים של ניתוח התנהגות ושל חיזוי.[728] לצד היתרונות הטמונים בחיבור מתמיד למרשתת ולאינטרנט של הדברים יש בו גם חסרונות: עיבוד מידע שנאסף דרך קבע על ידי מערכות נבונות לקבלת החלטות עלול להביא לידי פגיעה בפרטיות ולכן לייצר אפקט מצנן על ההתנהגות, כלומר לגרום להתכנסות החוויה האנושית למחוזות קונפורמיים ולשחיקת האוטונומיה של הפרט.[729]

התפתחויות המבוססות על מודלים של למידת מכונה בתחומי הביולוגיה ומדעי החיים, ובכללן האפשרות ליצור זיהוי חוזר של מידע בריאות מותמם וריצוף גנטי בר־זיהוי מתוך נהרות או מתוך האוויר (e-DNA), מחריפות מאד את עומקו ורוחבו של האיום על הפרטיות. חוקרים שהראו כי ביכולתם לפענח ו"לתמלל" מחשבות באמצעות הפעלת אלגוריתם על ממצאים של סריקה מוחית חיצונית (fMRI), בדרך שמכונה brain decoding, הרגישו לאחרונה צורך להצהיר במפורש כי "חיוני לעורר מודעות לסכנות בטכנולוגיית קריאת מחשבות וליצור חוקים שיגנו על הפרטיות המחשבתית של כל אדם".[730]

האפקט המצנן של הפגיעה בפרטיות מחריף כשהתכלית של המערכות הנבונות היא מעקב בפועל. מערכות לזיהוי פנים, לניטור התנהגות במרחב או לחיזוי פשיעה וטרור פועלות כפנאופטיקון הממשטר את התנהגות הסובייקטים המפוקחים באמצעותן.[731] משום כך הן עלולות להשפיע על נכונות של אנשים להימצא באזורים מסוימים, לבוא במגע עם זולתם, להתבטא בחופשיות או להתנהג באופן לא מקובל. לכן כשמערכות נבונות פוגעות בזכות לפרטיות תיתכן

728 ראו Zuboff, **לעיל** ה"ש 112.

729 **שם**, בעמ' 521.

730 Jerry Tang et al., *Semantic Reconstruction of Continuous Language from Non-Invasive Brain Recordings*, 26 Nature Neuroscience 858–866 (2023)

731 ראו לדוגמה כהנא ושני, **לעיל** ה"ש 213, בעמ' 19–20.

גם פגיעה בזכויות אדם אחרות, כגון חופש הביטוי, זכות האספה, חופש התנועה או האוטונומיה של הפרט.

5.2 הזכות להליך הוגן וכללֵי הצדק הטבעי

הזכות להליך הוגן היא מכללי הצדק הטבעי.[732] אומנם שורשיה נטועים במשפט הפלילי, אבל היא משפיעה גם על הליכים אזרחיים ומינהליים ומעוגנת במשפט הבינלאומי.[733] היא אינה נחשבת זכות חוקתית מפורשת, אך הוכרה בפסיקה ובספרות כזכות בעלת מעמד חוקתי שנגזרת מהזכות לכבוד או מהזכות לחירות.[734]

מהזכות להליך הוגן במשפט הפלילי נובעות, בין השאר, "[...] זכותו של נאשם לדעת מדוע נעצר ובמה הוא מואשם, הזכות להיות מיוצג על־ידי עורך־דין, הזכות להיות נוכח במשפט, הזכות למשפט פומבי על־ידי ערכאה בלתי תלויה וניטראלית והזכות להתגונן במשפט ולהציג ראיות רלוונטיות".[735]

הסתייעות של מערכת המשפט בבינה מלאכותית התומכת בתהליך קבלת החלטות[736] עלולה להביא לידי פגיעה בהוגנות ההליך, שכן אותה "קופסה שחורה" ממסכת את השיקולים המכריעים בקבלת ההחלטות האוטומטית ומקשה על הנאשם להבין מדוע נעצר, להיות "נוכח" במשפטו, לממש את זכות הטיעון שלו ולהבין מהן הראיות הרלוונטיות. לאחרונה נחשפה מערכת הכללה ממוחשבת (פרופיילינג) לאיתור חשודים בבלדרות סמים בנמל התעופה בן־גוריון. היא

732 דפנה ברק־ארז **משפט מינהלי** כרך א 564-461 (2010).

733 ראו למשל European Convention for the Protection of Human Rights and Fundamental Freedoms, art. 6, adopted Nov. 4, 1950, ETS 5 (entered into force Sep. 3, 1953) (להלן: ECHR); Universal Declaration of Human Rights, G.A. Res. 217(III) A, U.N. Doc. A/RES/217(III), at art. 10 (Dec. 10, 1948) (להלן: UDHR).

734 אהרן ברק **כבוד האדם: הזכות החוקתית ובנותיה** 869-868 (2014).

735 ע"פ 5121/98 **יששכרוב נ' התובע הצבאי הראשי**, פ"ד סא(1) 461 (2006), פס' 66 לפסק דינה של השופטת ביניש.

736 ראו **לעיל** סעיף 2.3.2.

מופעלת על ידי המשטרה, אך אופן פעולתה לא ידוע, ובהיעדר שקיפות מתעורר חשש להטיות אלגוריתמיות.[737]

ההליך ההוגן עלול להיפגע לא רק עקב שימוש במערכות תומכות החלטה במסגרת הליכים משפטיים פליליים, אלא בכל מקרה שבו השלטון נעזר במערכות נבונות (למשל כדי לקבל החלטות הנוגעות לחלוקת משאבים) בלי לאפשר למי שזכויותיו נפגעו לקבל את יומו בבית משפט (אנושי) ולממש את זכות הטיעון שלו. לכן הדין האירופי, למשל, מקנה למושא המידע את הזכות שלא יתקבלו בעניינו החלטות על בסיס עיבוד אוטומטי של המידע האישי שלו.[738]

דוגמה לפגיעה בהליך הוגן היא ההתנהלות בפרשת המערכת ההולנדית לאיתור הונאות רווחה SyRI (Systeem Risico Indicatie).[739] SyRI הייתה מערכת משותפת של כמה גופים ממשלתיים בהולנד; תכליתה הייתה לאסוף מידע ולהעבירו למשרד המידע ההולנדי. שמות אנשים ואמצעים מזהים אחרים הוחלפו בקוד והמפתח לקוד נשמר במקום אחר. לאחר הצלבת המידע באמצעות אלגוריתמים, שדרך פעולתם לא פורטה לציבור, לפרלמנט או לבתי המשפט, הפיקה המערכת פלט שקבע כי עלה בידה לאתר חשוד בהונאה. התראה של המערכת הביאה את משרד הרווחה של הולנד לפתוח בחקירה. במקרה זה העלה היעדר שקיפות פעולת האלגוריתם חשש לפגיעה בזכות להליך הוגן בעניין זכאות לתמיכה ממשלתית.[740]

במישיגן שבארצות הברית זיהתה מערכת לאיתור הונאות רווחה (במקרה זה, זכאות לדמי אבטלה) אלפי אזרחים רמאים, שללה אוטומטית את זכאותם לדמי

737 ראו תומר גנון "האלגוריתם שיעצור אתכם בנחיתה בנתב"ג" **כלכליסט** (10.11.2022); בש"פ 8308/21 **עומר ברגר נ' מדינת ישראל** (8.12.2021); עמיר כהנא ותהילה שוורץ אלטשולר **הפרופיילינג המשטרתי מעלה חשש לאפקט מצנן שישפיע על זכויות אדם** (חוות דעת, המכון הישראלי לדמוקרטיה 22.5.2023).

738 ראו **לעיל** ה"ש 403.

739 Max Vetzo, *The Netherlands–Algorithmic Fraud Detection System Violates Human Rights: The Case of SyRI*, 3 Public Law, 650 (2021). ראו גם פרל ושוורץ אלטשולר, **להלן** ה"ש 881, בעמ' 54–56.

740 על פסק הדין בעניין SyRI ראו פרל ושוורץ אלטשולר, **להלן** ה"ש 881, בעמ' 54–56.

אבטלה ותבעה מהם החזרים (בצירוף קנסות) של דמי אבטלה בסכומי עתק. מבקר המדינה של מישיגן בדק מדגם של כ־22,000 מהמקרים שבהם זיהתה המערכת הונאה ומצא ש־93% מהם שגויים. האוטומציה של ההליך לאיתור הונאות (מזיהוי ועד ענישה), שנעשה כמעט ללא מעורבות אנושית, הקשה על היכולת של האזרחים שזכויותיהם הסוציאליות נשללו להבין את הסיבות ולערער על הקביעות הממוכנות.[741]

5.3 הזכות לשוויון בפני החוק

הזכות לשוויון בפני החוק היא עקרון יסוד חוקתי שמוכר במשפט הבינלאומי.[742] אף שהזכות לשוויון אינה זכות חוקתית מפורשת בחוקי היסוד של ישראל, הפסיקה והספרות מכירים בה כזכות בת של כבוד האדם,[743] "עקרון יסוד חוקתי, השלוב ושזור בתפיסות היסוד המשפטיות שלנו ומהווה חלק בלתי נפרד מהן".[744]

מערכות נבונות עלולות לפגוע בזכות לשוויון עקב הטיות אלגוריתמיות המובילות להקצאת משאבים מפלה או לפגיעה עודפת בזכויות של קבוצות מוגנות. מאחר שפיתוח מערכות בינה מלאכותית נשען על פיתוח מודלים סטטיסטיים מבוססי מידע, ההסתמכות על מאגרי מידע שמשעתקים יחסי כוח אי־שוויוניים, או שמשקפים דעות קדומות נגד מגזרים וקבוצות מוחלשים, עלולה להנציח את אי־השוויון; ובמקרים של למידת מכונה מבוססת משוב אף להעצים אותו.[745] גם

741 ראו בעניין זה Cahoo v. SAS Analytics Inc., No. 18-1296 (6th Cir. 2019); Sonia Gipson Rankin, *The MiDAS Touch: Atuahene's "Stategraft" and the Implications of Unregulated Artificial Intelligence*, 2022-21 UNM School of Law Research Paper (2022)

742 ראו למשל ס' 1, 27 ב־UDHR, **לעיל** ה"ש 733; Art. 2(1), 26 International Covenant on Civil and Political Rights, GA Res. 2200A (XXI) of 16 December 1966 (להלן: ICCPR); ס' 2(1), 14, 26 ב־ECHR, **לעיל** ה"ש 733.

743 ראו ברק, **לעיל** ה"ש 734, בעמ' 685–690.

744 בג"ץ 114.78 **בורקאן נ' שר האוצר**, פ"ד לב(2) 800, 806 (1978).

745 על הטיות אלגוריתמיות, ראו **להלן** בסעיף 6.2.

מודלים סטטיסטיים שמסתמכים על מידע חלקי או עודף על הקבוצה המוגנת יכולים להביא לידי פגיעה בשוויון.[746]

5.4 כבוד האדם

כבוד האדם הוא ערך יסוד בשיטה החוקתית של ישראל וגם במשפט החוקתי המשווה.[747] לפי הגישה הפרשנית ההוליסטית של אהרן ברק יש לייחס לכבוד האדם ערך חוקתי.[748] לפי גישה זו, ערכו החוקתי של כבוד האדם הוא ההגנה על אנושיותו של אדם.[749]

מערכות אוטומטיות שמשמשות לבקרת תוכן שמעלים גולשים עלולות לפגוע בכבוד האדם של משתמשים שהתבטאויותיהם צונזרו. טכנולוגיות בינה מלאכותית עלולות לפגוע בכבוד האדם של מי שזהותם נגנבה על ידי טכנולוגיות דיפ־פייק[750] או שדמויותיהם הווירטואליות מוצגות באופן מבזה.[751] בשל הפרשנות המרחיבה של הערך החוקתי של כבוד האדם יש סיכוי רב לפגוע בו באמצעות פעילותן של מערכות נבונות. הללו יכולות לפגוע בקשת רחבה של זכויות נגזרות, דוגמת הזכות לפרטיות, הזכות להליך הוגן והזכות לשוויון.

746 ראו **להלן** בסעיף 6.2.1.

747 ראו ברק, **לעיל** ה"ש 734, בעמ' 73–224.

748 **שם**, בעמ' 235–237.

749 **שם**, שם. ראו גם דברי השופט ברק בעניין בג"ץ 6427/02 **התנועה למען איכות השלטון בישראל נ' כנסת ישראל**, פ"ד סא(1) 619, 685 (2006): "ביסוד כבוד האדם עומדים האוטונומיה של הרצון הפרטי, חופש הבחירה וחופש הפעולה של האדם כיצור חופשי. כבוד האדם נשען על ההכרה בשלמותו הפיסית והרוחנית של האדם, באנושיותו, בערכו כאדם וכל זאת בלא קשר למידת התועלת הצומחת ממנו לאחרים".

750 ראו **לעיל**, ה"ש 485.

751 Jacquelyn Burkell and Chandell Gosse, *Nothing New Here: Emphasizing the Social and Cultural Context of Deepfakes*, 24 First Monday (2019); Elizabeth F. Judge and Amir M. Korhani, *Deepfakes, Counterfeits, and Personality*, 59 Alberta L. Rev. (26.7.2021)

פגיעה אפשרית נוספת היא איון הסובייקט האנושי והמרתו ברשומה במסד נתונים.[752] בפרק הקודם ראינו כי מסמכי אתיקה של מערכות נבונות רואים בכבוד האדם ובאוטונומיה אנושית רכיב יסודי.[753] כשיכולת האדם לבחור בחירה מושכלת מפנה את מקומה לקבלת החלטות אוטומטית, כשנתח הולך וגדל של ההחלטות מתקבל באמצעות מודלים סטטיסטיים, נשחק הסובייקט האנושי, מרגיש בלתי נראה[754] וחסר ערך, וחש תסכול לנוכח המציאות השרירותית.

5.5 חופש הביטוי

חופש הביטוי הוא ערך מוגן במשפט הבינלאומי,[755] וגם במשפט הישראלי.[756] חשיפת האמת, הגשמה עצמית של האדם והבטחת המשטר הדמוקרטי משמשות לביסוס חשיבותו של חופש הביטוי.[757]

מנגד, זכות זו אינה מוחלטת. אותם דברי חקיקה הקוראים לחופש הביטוי אוסרים לעיתים התבטאויות הכוללות ביטויי שנאה (hate speech), למשל.[758] מלאכת האיזון בין האינטרסים המוגנים על ידי חופש הביטוי ובין האינטרסים המוגנים על ידי האיסור על ביטויי שנאה היא מלאכה מורכבת שתלויה בתרבות ובמסורת החוקתית המקומית. מורכבות זו מתעצמת לנוכח יכולתו של כל משתמש להתבטא במרחב המקוון, ושטף הציוצים, הפוסטים ושאר ההתבטאויות המקוונות

752 לדה־הומניזציה של הסובייקט האנושי על ידי מערכות בינה מלאכותית בתחומי הרפואה ראו Van Kolfschooten, **לעיל** בה"ש 581, בעמ' 92–95.

753 ראו **לעיל** בסעיף 3.7.

754 ראו למשל Zuboff, **לעיל** ה"ש 112, בעמ' 44, המתארת את התסכול של המפגינים במהומות בלונדון בשנת 2011.

755 ראו למשל ס' 19 ב־UDHR, **לעיל** ה"ש 733; ס' 19 ב־ICCPR, **לעיל** ה"ש 742; ס' 10 ב־ECHR, **לעיל** ה"ש 733.

756 ראו בג"ץ 73/53 **חברת "קול העם" בע"מ נ' שר הפנים** פ"ד ז 871 (1953); ברק, **לעיל** ה"ש 734, בעמ' 707–712.

757 ברק, **שם**, בעמ' 713.

758 ראו לדוגמה ס' 20(2) ב־ICCPR, **לעיל** ה"ש 742.

המצרפיות של כל בני "אומת האינטרנט".[759] התאמת המודל התאורטי המדויק להגדרת ביטויי שנאה ולקביעת מדיניות הטיפול המתאימה בהם היא אתגר בפני עצמו,[760] אך בשל היקף הביטויים שיש לנטר יישומו של המודל מצריך גם התערבות אלגוריתמית. הדָוֹח המיוחד של האומות המאוחדות לחופש הביטוי עמד על הסכנות הנשקפות לחופש הביטוי מהשימוש בניטור תוכן אלגוריתמי;[761] הוא הצביע על הקשיים הטכניים של מערכות אלו לזהות דקויות של אירוניה[762] והראה כי לעמימות והיעדר שקיפות יש אפקט מצנן על חופש הביטוי. יתר על כן, מעיין פרל (פילמר) וניבה אלקין־קורן טוענות כי האופי המסחרי של הפלטפורמות המשתמשות בטכנולוגיות ממוכנות לצורך ניטור תוכן עלולות לגרוע מהאופי הדמוקרטי של השיח המקוון. לדבריהן, כללי ההכרעה בעניין נורמות של תוכן לגיטימי הולכים ונעשים אלגוריתמיים, הבקרה עליהם נעשית בחוסר שקיפות והאוטומציה השואפת לכללי הכרעה בינאריים אינה סובלת את ריבוי המשמעויות של השפה הטבעית.[763]

5.6 זכויות פוליטיות

מערכות נבונות מאיימות על מימושן של זכויות פוליטיות, ובהן הזכות לחופש האספה והזכות לחופש ההתאגדות,[764] הזכות להשתתפות פוליטית, הזכות לחופש המחשבה והזכות לחופש הביטוי.[765] הדָוֹח המיוחד של האומות המאוחדות לזכות לחופש האספה ולזכות לחופש ההתאגדות ציין כי

759 United Nations General Assembly, Rights to Freedom of Peaceful Assembly and of Association, Para. 59 (A/HRC/41/41, 2019) (להלן: HRC 2019).

760 ראו Medzini and Shwartz Altshuler, **לעיל** ה"ש 268.

761 ראו HRC 2018, **לעיל** ה"ש 270.

762 ראו לדוגמה יניר יגנה "החשוד בהסתה להצתה בפייסבוק שוחרר: 'גם בערבית יש סאטירה' " **וואלה!** (28.11.2016).

763 Maayan Perel (Filmar) and Niva Elkin-Koren, *Democratic Friction in Speech Governance by AI*, *in* Handbook of Critical Studies of Artificial Intelligence (Simon Lindgren ed., 2022)

764 ראו למשל ס' 20 ב־UDHR, **לעיל** ה"ש 733; ס' 21–22 ב־ICCPR, **לעיל** ה"ש 742.

765 על פגיעה פוטנציאלית של מערכות נבונות בחופש הביטוי, ראו **לעיל** סעיף 5.5.

זכויות אלו "[...] הן המפתח למימוש הדמוקרטיה וכבוד האדם, מאחר שזכויות אלו מאפשרות לאנשים להשמיע את קולם ולייצג את האינטרסים שלהם, לראות בממשלות אחראיות לפעולותיהן ולהעצים את הפעלנות האנושית [human agency]".[766] מימוש זכויות אלו במרחב המקוון יכול לבוא לידי ביטוי בדרכים מגוונות, כגון הצטרפות לעצומה באינטרנט, חבירה לקבוצות עניין פוליטיות ברשתות חברתיות[767] או פרסום פוסטים עם האשטאג בעל זיהוי פוליטי.[768]

עם הסיכונים הנשקפים לזכויות פוליטיות משימוש במערכות בינה מלאכותית אפשר למנות את האפקט המצנן של מערכות לניטור תוכן אלגוריתמי ומערכות למעקב מקוון.[769] אנשים שיודעים שהם נתונים למעקב מקוון עלולים להימנע מפעילות פוליטית מקוונת שנתפסת חריגה או לא קונצנזוסיאלית, גם אם היא לגיטימית, מחשש שמערכת נבונה או צנזור אנושי יסווג אותה כאסורה.

מערכות נבונות עלולות לצמצם את החופש הפוליטי המקוון גם באמצעות פגיעה בחופש המחשבה. היבט מרכזי של חופש המחשבה הוא "הזכות לגבש דעה ולפתח אותה באופן מנומק",[770] ובכלל זה הזכות לעשות כן ללא כפייה וללא אינדוקטרינציה מבחוץ.[771] ככל שהאלגוריתמים נעשים שחקנים משמעותיים יותר במלאכה האפיסטמולוגית של צירת המודעות האנושית, כך הם יכולים להשפיע יותר על האופן שבו אנשים מגבשים את דעתם ומפתחים אותה. כיום אלגוריתמים הם שמתווכים בין המציאות ובין המשתמשים ברשתות החברתיות – וכיוון שהם

766 United Nations General Assembly, Report of the Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association, Para. 87 (A/73/279, 2018)

767 ראו HRC 2019, **לעיל** ה"ש 729, בפס' 25.

768 לסקירה ראו Sarah J. Jackson, Moya Bailey, and Brooke Foucault Welles, #HashtagActivism (2020); לדוגמה מקומית ראו הגר בוחבוט וירון דרוקמן "'הפקרתי מימיה ונענשתי יותר מבוכריס': קמפיין רשת נגד הסדר הטיעון" **ynet** (1.12.2016).

769 Cameran Ashraf, *Artificial Intelligence and the Rights to Assembly and Association*, 5 J. Cyber Pol'y 163 (2020)

770 ראו HRC 2018, **לעיל** ה"ש 270, בפס' 23; Manfred Nowak, U.N. Covenant on Civil and Political Rights. CCPR Commentary 441 (2005)

771 Nowak, **שם**, בעמ' 442; ראו לדוגמה גם *Yong Joo-Kang v. Republic of Korea* CCPR/C/78/D/878/1999, (16.7.2003)

שולטים ב"פיד" הידיעות של המשתמשים הם מסוגלים לעצב עמדות ורחשי לב;[772] ולעיתים אלו עמדות קיצוניות שמובילות להסתה ולאלימות.[773]

התפקיד המרכזי של מערכות בינה מלאכותית בעיצוב התודעה האנושית בעידן המידע מעלה חששות באשר ליכולת לממש את חופש המחשבה ולגבש דעה באופן עצמאי ומנומק.[774] האלגוריתמים משמשים "שומרי סף" הקובעים אילו ידיעות ימצאו את דרכן ל"פיד" ואילו יישארו מחוץ לו. אלגוריתמים יכולים להגביר את הרעש התקשורתי סביב ידיעה מסוימת או להחלישה כליל.

שחקנים חיצוניים יכולים לנקוט מבצעי השפעה (psych ops) כדי לייצר ב"פיד" רעשים שיש בהם כדי להשפיע על דעת הקהל הפוליטית. גם מיקרו־טרגוט ממוקד יכול לשמש לעיצוב עמדות של מגזרים מסוימים, למשל בסמוך לבחירות או במהלכן.[775] במסגרת מבצעי השפעה אפשר לרתום מערכות בינה מלאכותית להפקת זיופים בטכנולוגיות דיפ־פייק כדי לערער את דעת הקהל באשר למועמד מסוים או מפלגה מסוימת.[776] מבצעי השפעה מעין אלו לא בהכרח יערערו את

772 Adam D. I. Kramer, Jamie E. Guillory, and Jeffery T. Hancock, *Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks,* 111 Proc. Nat'l. Acad. Sci., 8788-8790 (2014); Katherine Haenschen, Brett Frischmann, and Paul Ellenbogen, *Manipulating Facebook's Notification System to Provide Evidence of Techno-Social Engineering*, 1 Soc. Sci. Comp. Rev. (2021)

773 ראו לדוגמה את הטענות בעניין השפעת אלגוריתמים החולשים על הפיד של הרשת החברתית פייסבוק של חברת מטא על רצח העם של בני הרוהינגה במיאנמר: The Social Atrocity: Meta ad the Right to Remedy for the Rohingya (Amnesty International, 2022); Rebecca J. Hamilton, *Platform-Enabled Crimes: Pluralizing Accountability When Social Media Companies Enable Perpetrators to Commit Atrocities* 63 B.C. L. Rev. 1349, 1365-1366 (2022)

774 ראו HRC 2018, **לעיל** ה"ש 270, בפס' 24-26.

775 ראו לדוגמה תהילה שוורץ אלטשולר וגיא לוריא **תעמולה דיגיטלית והאיום על הבחירות** 39-53 (2020); אלי בכר ורון שמיר **התקפות סייבר על מערכת הבחירות: איך מתמודדים?** 42-45 (2019); Alina Poluakova and Spencer B. Boyer, *The Future of Political Warfare: Russia, the West, and the Coming Age of Global Digital Competition*, Brookings (2018)

776 Cristian Vaccari and Andrew Chadwick, *Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on*

אמון הבוחרים בעמדה מסוימת, אלא יביאו אותם להטיל ספק ביציבות הפוליטית ובמקורות המידע שעליהם הם נסמכים בדרך כלל כדי לגבש את דעתם.

5.7 הזכות לחיים

הזכות לחיים היא זכות יסוד במשפט החוקתי של ישראל,[777] וגם במשפט המשווה והבינלאומי.[778] יישומי בינה מלאכותית עלולים לגרום לפגיעה אגבית בזכות לחיים, הן עקב טעויות הן עקב הכרעות מוקדמות הנוגעות לעיצובו של היישום. למשל, ייתכנו מקרי קצה שבהם המפתחים של מערכת רכב אוטונומית יידרשו לקבוע מראש קריטריונים לאיזון בין הזכות לחיים של נוסע ברכב ובין הזכות לחיים של הולך רגל.[779] לאחרונה פורסם מחקר שמתאר כיצד הצליח אלגוריתם שנועד לאתר מבנים מולקולריים שמתאימים לתרופות חדשות בטכניקות של למידת מכונה לגלות בתוך כך גם תצורות ישנות וחדשות של חומרי לוחמה ביולוגית וכימית.[780] פגיעה בזכות לחיים עלולה להיגרם גם

Deception, Uncertainty, and Trust in News, 1 Social Media + Soc. (2020); Tom Dobber et al., *Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?* 26 Int'l J. Press/Pol. 69 (2021)

777 ס' 2, 4 לחוק יסוד: כבוד האדם וחירותו. ראו גם אהרן ברק "הזכות החוקתית להגנה על החיים, הגוף והכבוד" **מבחר כתבים ג – עיונים חוקתיים** 547 (2017). על ההבחנה בין האינטרס הציבורי ובין הזכות לחיים ראו אורן גזל־אייל ואמנון רייכמן "אינטרסים ציבוריים כזכויות חוקתיות?" **משפטים** מא 97 (2011).

778 ראו לדוגמה ס' 2 ב־ECHR, **לעיל** ה"ש 733; ס' 6 ל־ICCPR, **לעיל** ה"ש 742; ס' 3 ב־UDHR, **לעיל** ה"ש 733. ראו גם Elizabeth Wicks, The Right to Life and Conflicting Interests 22-47 (2010)

779 לצד תאונות שאפשר לייחס אותן לפגמים בעיצוב המערכות ("באגים"), והיה אפשר למנוע אותן, דוגמת אירוע דריסת הולכת הרגל איליין הרצברג על ידי רכב אוטונומי (Lauren Smiley, *"I'm the Operator": The Aftermath of a Self-Driving Tragedy*, Wired [8.3.2022]), יש גם מקרי קצה תאורטיים שבהם התאונה היא בלתי נמנעת ולאלגוריתם המנחה את הרכב האוטונומי אין ברירה אלא לבחור בין תרחישי פגיעה שונים. ראו Nyholm and Smids, **לעיל** ה"ש 27.

780 Paul Rosenzweig, *Artificial Intelligence and Chemical and Biological Weapons*, Lawfare (27.4.2022); Fabio Urbina et al., *Dual Use of Artificial Intelligence-Powered Drug Discovery*, 4 Nature Machine Intelligence 189 (2022)

מליקויים באבטחת מידע או אבטחת סייבר – חולשות במערכות השליטה בבית חכם או התקפות מכוונות על נתוני אימון של מודלים של דיאגנוסטיקה רפואית.[781]

אפשר להבחין בין רמות אוטונומיה שונות של מערכות נשק (בדומה לכלי רכב אוטונומיים). החל בכלי נשק מסורתיים, כמו נשק קר, שמחייבים הפעלה אנושית רציפה, וכלה במערכות אוטונומיות לחלוטין, שיכולות לבחור מטרה ולתקוף אותה.[782] מאז פיתח ריצ׳רד גאטלינג את המקלע הקרוי על שמו, שאפשר ירי אוטומטי רציף של מאות קליעים ללא טעינה ידנית (אף שמנגנון הירי עצמו חייב הפעלה אנושית רציפה), עברו מערכות הנשק האוטונומיות כברת דרך ארוכה.[783] כלי נשק בעלי מידה מוגבלת של אוטונומיה – כאלו שיכולים לבחור את המטרה באופן עצמאי או לתקוף אותה באופן עצמאי – נמצאים בשימוש עוד מימי מלחמת העולם השנייה, שבה הפעיל צבא גרמניה מערכות טורפדו שהתבייתו על מטרותיהן באמצעות חיישן אקוסטי.[784] בעידן שבו הולך ומתפתח ירי תלול מסלול, שמתאפיין במטחי ירי ממספר רב של קני שיגור בו זמנית,[785] האתגר המורכב כשהוא לעצמו – יירוט רקטה או טיל במהלך מעופם – מתעצם אף יותר לנוכח מתווה השיגור. מערכות הגנה אוטונומיות מסוגלות להתמודד עם סכנה

781 Bernd Carsten Stahl, Doris Schroeder, and Rowena Rodrigues, Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges 63-69 (2023)

782 אליאב ליבליך ואייל בנבנישתי "לוחמה רובוטית ובעיית הכבילה של שיקול־הדעת" **עיוני משפט לט** 67, 72 (2016); Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications,* 36 Cardozo l. Rev. 1837 (2015); U.S. Department of Defense, Directive No. 3000.09: Autonomy in Weapon Systems 13-14 (2012); International Humanitarian Law and the Challenges of Contemporary Armed Conflicts 44 (International Committee of the Red Cross, 2015)

783 Paul Scharre, Army of None: Autonomous Weapons and the Future of War, Chap. 3 (2018)

784 **שם.**

785 אמיר קוליק "המודיעין ואתגרי הירי תלול המסלול" **צבא ואסטרטגיה 1** 19, 22 (2009).

זו בזמן אמת, אבל מחייבות השגחה של מפעיל אנושי המסוגל להתערב במידת הצורך (למשל, אם המערכת זיהתה בטעות כלי טייס אזרחי).[786]

נוסף על מערכות נשק אוטונומיות לצורכי הגנה, כדוגמת כיפת ברזל, יש גם מערכות נשק אוטונומיות שנועדו לתקיפה אקטיבית.[787] מערכות משוטטות, שמפטרלות באזור מסוים ומתבייתות באופן אוטונומי על סכנות שזיהו, כבר היו בשימוש מבצעי במלחמה בין ארמניה לאזרבייג'ן ובעימותים אחרים.[788] אומנם המטרות המסורתיות של מערכות משוטטות הן מערכות נשק אחרות – בין שהן מאוישות בעת תקיפתן ובין שאינן מאוישות – אך לאחרונה נטען שרחפן משוטט היה מעורב בהרג מכוון במבצע בלוב.[789]

חרף הקולות המזהירים מפני כלי נשק אוטונומיים ומעודדות הכרזת חרם בינלאומי על מי שיפתח אותם או ישתמש בהם,[790] יש המטילים ספק בהיתכנות חרם גורף כזה וקוראים לבחון כל מקרה לגופו.[791] לנוכח השימוש המתרחב בכלי נשק אוטונומיים לתכליות התקפיות נראה שפגיעתם בזכות לחיים הולכת וחורגת מגדרי נזק אגבי או מהרג חיילי אויב המותר במסגרת דיני הלחימה. הבעיות

786 Scharr, **לעיל** ה"ש 783.

787 לייבליך ובנבנישתי, **לעיל** ה"ש 782.

788 Kelsey Atherton, *Loitering Munitions Preview the Autonomous Future of Warfare*, Brookings (4.8.2021)

789 **שם**; ראו גם Joe Hernandez, *A Military Drone with a Mind of Its Own Was Used In Combat, U.N. Says*, NPR (1.6.2021)

790 *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*, Future of Life Institute (28.7.2015); *Lethal Autonomous Weapons Pledge*, Future of Life Institute; The Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) (Advanced Version), Para. 71 (2016); *Losing Humanity: The Case Against Killer Robots*, Human Rights Watch (19.11.2012); לסקירה של עמדות מרבית מדינות העולם בסוגיה ראו *Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control*, Human Rights Watch (10.8.2020)

791 Anja Kaspersen, *Should We Ban Weapons that Don't Even Exist Yet?* World Economic Forum (20.7.2016); Michael N. Schmitt, *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics*, Harv. Nat'l Sec. J. (5.2.2013)

האינהרנטיות למערכות נבונות – הטיות אלגוריתמיות והיעדר שקיפות[792] – עלולות להביא לתוצאות אומללות במיוחד בשדה הקרב.

5.8 בינה מלאכותית וזכויות אדם

כפי שראינו בפרק 2, לבינה מלאכותית יישומים פוטנציאליים רבים ויש לה הכוח להשפיע לחיוב על חייהם של בני אדם. לכן כמה מסמכי אתיקה שעניינם בינה מלאכותית כוללים קריאות להשתמש במערכות נבונות לקידום סולידריות אנושית,[793] לקידום קיימות ואיכות הסביבה[794] או לחיוב באופן כללי.[795] ואולם בבינה מלאכותית טמון גם סיכון לפגיעה בזכויות אדם. בפרק זה עמדנו בקצרה על פגיעות אפשריות בזכות לפרטיות, בזכות להליך הוגן, בזכות לכבוד אנושי, בזכויות פוליטיות ובזכות לחיים. אפשר להעלות על הדעת פגיעה בזכויות נוספות – למשל, הגבלת חופש התנועה של אזרחים שמערכות בינה מלאכותית לניטור פנדמיות קבעו כי עליהם להיכנס לבידוד[796] או פגיעה רחבת היקף בזכות לתעסוקה[797] בשל שיבוש שוק העבודה עקב אוטומציה רחבת היקף. בפרקים הבאים נתאר בהרחבה את האופן שבו מאפיינים אינהרנטיים של מערכות בינה מלאכותית עלולים להביא לידי פגיעה בזכויות.

792 ראו **להלן** בפרק 6, 7.

793 Luengo-Oroz , **לעיל** ה"ש 447.

794 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 12; .Floridi et al, **לעיל** ה"ש 337, בעמ' 697–696. ראו גם שי הרשקוביץ "בינה מלאכותית מנסה להתמודד עם שינוי האקלים. מולה ניצב האדם עצמו" **הארץ** (22.9.2021).

795 ראו **לעיל** בסעיף 3.6.

796 ראו לדוגמה פס' 11 לפסק דינו של השופט עמית בעניין בג"ץ 6732/20 **האגודה לזכויות אזרח נ' הכנסת** (1.3.2021, פורסם באר"ש).

797 ראו לדוגמה ס' 23 ב־UDHR, **לעיל** ה"ש 733.

פרק שישי

"אחרי רבים להטות":
הטיות אלגוריתמיות

לכאורה, אחד היתרונות הגדולים של מערכות בינה מלאכותית שתפקידן לקבל החלטות ולנתח נתונים הוא האובייקטיביות שלהן. שהרי קבלת החלטות ממוכנת, המבוססת על נתונים בלבד, אינה מושפעת ממגבלות קוגניטיביות אנושיות. מערכות ממוכנות אינן מתעייפות, אינן מושפעות מרגשות ואינן נתונות ללחצים פסיכולוגיים. השימוש בהן ככלי לקבלת החלטות צפוי אפוא לנטרל הטיות אנושיות רבות.

יתר על כן, במבט ראשון נדמה כי העברת שיקול הדעת האנושי למערכות אוטומטיות יכולה לפתור השפעה לא רצויה של דעות קדומות, המנציחות פערים חברתיים, מגדריים ואתניים. והרי גם מי שנמנע במודע מדעות קדומות בקבלת

ההחלטות שלו יכול להיות מושפע מהן שלא במודע.[798] מאחר שמערכות בינה מלאכותית אינן אנושיות, ואין להן מטענים חברתיים ורגשיים כמו לבני אדם, יש להן יכולת לסייע ביצירת חברה שוויונית יותר ולפתוח בפני מיעוטים וקבוצות מוחלשות אפשרויות חדשות.

ואולם מן השימוש במערכות בינה מלאכותית עולה שחלק מההטיות האנושיות שמערכות אלו היו אמורות למתן באות לידי ביטוי גם בקבלת החלטות ממוכנת. מאחר שאיכותן של מערכות לומדות מבוססת על טיב הנתונים שהן ניזונות מהם במהלך האימון, ונתונים אלו הם תוצר של פעילות אנושית ולכן משקפים הטיות חברתיות, מגדריות, אתניות ואחרות, אין פלא שהטיות אלו באות לידי ביטוי גם בקבלת ההחלטות הממוכנת והניטרלית לכאורה של מערכות בינה מלאכותית.

יש להבחין בין שני סוגים של הטיות אנושיות – הטיות הנובעות מדעות קדומות פסולות והטיות קוגניטיביות. בחיי המעשה יש והטיות אלו משמשות בערבוביה – דעות קדומות לגבי גזע או מגדר, למשל, יכולות לעיתים להשפיע שלא במודע על שיקול דעת אנושי חף מכוונות רעות. למשל, כשמנהל כוח אדם נדרש לסנן מספר רב של קורות חיים לתפקיד ניהול הוא עלול להישען גם שלא במודע על דעות קדומות אתניות.[799]

798 ראו סקירת הספרות אצל John T. Jost et al., *The Existence of Implicit Bias Is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies that No Manager Should Ignore*, 29 Research in Organizational Behavior 39 (2009)

799 מריאן ברטרנד וסנדיל מולינאתן בחנו אמפירית את קיומה של הטיה דה פקטו בקרב מעסיקים שהעדיפו קורות חיים של בעלי שמות שמעידים על רקע אתני לבן משמות שמעידים במובהק על רקע אתני אפרו־אמריקאי. Marianne Bertrand and Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination* 94 American Economic Rev. 991 (2004). ממחקר זה לא ברור אם ההטיה לא מודעת, אבל מחקר המשך שנערך בקרב סטודנטים מצא קורלציה בין דעות קדומות להעדפת שמות "לבנים". Marianne Bertrand, Dolly Chugh, and Sendhil Mullainathan, *Implicit Discrimination*, 95 American Economic Rev. 94 (2005)

6.1 הטיות קוגניטיביות

עקב מוגבלות התודעה האנושית פיתחו בני אדם סוגים שונים של היריסטיקות והטיות המסייעות להם לעבד מידע ולקבל החלטות בתנאי אי־ודאות.[800] הטיות קוגניטיביות אלו אינן נובעות מדעות קדומות בעניין קבוצות אתניות, מגדר או העדפה מינית, אלא מהמאפיינים הפסיכולוגיים והפיזיולוגיים של עיבוד המידע. את הקשיים הגורמים להטיות קוגניטיביות אפשר לחלק בחלוקה גסה לארבעה:[801]

(1) **עודף מידע** – מוח האדם חשוף למידע רב. אין לו ברירה אלא לסנן את הנתונים הרבים הנקלטים בחושיו ולחלץ מהם את המידע הרלוונטי לקבלת החלטה מושכלת. למוח יש מבחר תכסיסים להתמודדות עם עודף מידע. למשל, אנו נוטים להעריך שכיחות של תופעות לפי הקלות שבה אפשר להעלותן על הדעת (היריסטיקות זמינות);[802] לייחס משמעות מיוחדת לפרטים יוצאי דופן; לתת לשינויים שהבחנו בהם להשפיע על שקלול אומדנים (היריסטיקות עיגון);[803] או להתמקד בפרטים המאששים אמונות קודמות שלנו (לדוגמה, הטיית האישור).[804]

(2) **חוסר מידע** – קבלת החלטות נעשית לעיתים מזומנות בתנאי אי־ודאות ובהסתמך על מידע חלקי ביותר. האדם, שמלכתחילה זיכרון העבודה שלו

800 ראו לדוגמה עמוס טברסקי ודניאל כהנמן "שיפוט בתנאי אי־ודאות: יוריסטיקות והטיות" **רציונליות, הוגנות ואושר** 45 (מיה בר הלל עורכת, 2005).

801 Buster Benson, *Cognitive Bias Cheat Sheet: Because Thinking Is Hard*, Better Humans (1.9.2016)

802 ראו טברסקי וכהנמן, **לעיל** ה"ש 800, בעמ' 53–56. ראו גם מאור אבן חן **האיסור הפלילי על שימוש במידע פנים: ניתוח כלכלי התנהגותי** 380–382 (2008).

803 ראו טברסקי וכהנמן, **שם**, בעמ' 56–60; אבן חן, **שם**, בעמ' 382–383; מיה בר הלל ואוריאל פרוקצ'יה "כלכלה התנהגותית" **הגישה הכלכלית למשפט** 71, 106–108 (עורך אוריאל פרוקצ'יה, 2012).

804 Peter C. Wason, *On the Failure to Eliminate Hypotheses in a Conceptual Task*, 12 Quarterly J. of Experimental Psychology 129 (1960); Raymond S. Nickerson, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, 2 Rev. of General Psychology 175 (1998)

מוגבל,[805] מתמודד עם מידע חסר בשלל דרכים. בין השאר, באמצעות הנטייה האנושית למצוא תבניות ודפוסים במדגמים קטנים, ללא כל מובהקות סטטיסטית (למשל, האמונה ב"יד החמה" בכדורסל[806] או "כשל המהמר"[807]); להשלים פערי מידע בהכללות, בסטריאוטיפים (לרבות דעות קדומות[808]) ובידע היסטורי; להניח הנחות לגבי עמדות הזולת; לפשט הסתברויות ונתונים כמותיים (למשל לערוך חשבונאות נפשית[809] או לחשוב על כסף במונחים נומינליים ולא ריאליים);[810] להשליך אמונות והנחות מההווה על אירועים מן העבר או העתיד (כגון כשל הראייה בדיעבד[811] או מזל מוסרי[812]).

(3) אילוצי זמן – גם כשכל המידע הרלוונטי זמין ומסודר, ואין נתונים מיותרים, נסיבות שבהן נדרשת החלטה מהירה יכולות לעודד קיצורי דרך מנטליים או

805 ראו: George A. Miller, *The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information*, 101 Psychological Rev. 343 (1956)

806 תומס גילוביץ', רוברט ולון ועמוס טברסקי הצביעו על פרכת האמונה העממית ב"יד החמה" בכדורסל וייחסו אותה להטיה קוגניטיבית שמתבטאת בניסיון למצוא דפוסים בנתונים אקראיים. Thomas Gilovich, Robert Vallone, and Amos Tversky, *The Hot Hand in Basketball: On the Misperception of Random Sequences*, 17 Cognitive Psychology 295 (1985). עם זאת, מחקרים מאוחרים יותר הראו שתופעת "היד החמה" בכדורסל שרירה וקיימת. ראו למשל Joshua B. Miller and Adam Sanjurjo, *Surprised by the Gambler's and Hot Hand Fallacies? A Truth in the Law of Small Numbers*, 86 Econometrica 2019 (2015)

807 ראו טברסקי וכהנמן, **לעיל** ה"ש 800, בעמ' 46–50.

808 ראו למשל: Daniel Bar-Tal and Yona Teichman, Stereotypes and Prejudice in Conflict: Representations of Arabs in Israeli Jewish Society 22-27 (2005)

809 ראו לדוגמה דניאל כהנמן ועמוס טברסקי "בחירות, ערכים והיצגים" **רציונליות, הוגנות ואושר** 64 (מיה בר הלל עורכת, 2005). ראו גם אבן חן, **לעיל** ה"ש 802, בעמ' 378–379.

810 Eldar Shafir, Peter Diamond, and Amos Tversky, *Money Illusion* 112 Quarterly J.Econ. 341 (1997)

811 אבן חן, **לעיל** ה"ש 802, בעמ' 390; בר הלל ופרוקצ'יה, **לעיל** ה"ש 803, בעמ' 102–106.

812 אלה קורן "התמודדות עם תופעת הטעות – ופרדוקס המזל המוסרי" **רפואה ומשפט** 40, 39, 41–43 (2009).

שימוש בהיריסטיקות והטיות או לעיתים הסתמכות על תחושת ביטחון עצמי (לרוב מופרז).[813] אילוצי זמן גורמים לנו לבחור בזמין ובמוכר, להעדיף את הפשוט מן המורכב (התער של אוקאם),[814] לתעדף השלמת משימות משיקולים של עלות שקועה או להעריך יתר על המידה מוצרים שהשתתפנו בהרכבתם (אפקט איקאה).[815]

(4) מגבלות הזיכרון – כאמור, מוח האדם מוצף במידע. לצד סינון המידע בזמן אמת, הוא נדרש לעבד אותו לשימוש בטווח ארוך או קצר. בני אדם נוטים לערוך את הזיכרונות שלהם, להתעלם מהפרטים (למשל, כדי ליצור הכללות) ולצמצם אירועים לרכיבי המפתח שלהם. עיבוד זיכרונות מושפע גם מהחוויה שזוכרים ומהרגשות שקשורים אליה.

כל אלו נחשבות להטיות קוגניטיביות מאחר שהן נובעות מהמגבלות הקוגניטיביות הפיזיולוגיות האנושיות. במצבי עייפות ולחץ מוח האדם אינו פועל באופן מיטבי. כמות הנתונים שהוא יכול להחזיק ולעבד במקביל מוגבלת והיא מחייבת תהליכי עיבוד וקיצורי דרך קוגניטיביים. בינה מלאכותית יכולה להתגבר על מרבית ההטיות האלו – היא אינה מתעייפת, כוח העיבוד שלה גדול בעשרות מונים מזה של האדם והיא אינה מושפעת מרגשות. עם זאת, בינה מלאכותית לא בהכרח יכולה להתגבר על בעיות של נתונים חלקיים או חסרים. אומנם יש טכניקות סטטיסטיות שעשויות לסייע בהתמודדות עם מידע חסר, אך יש להן השלכות על אופיו של המודל.[816]

נוסף על ההטיות הקוגניטיביות, קבלת החלטות אנושית עלולה להיות מושפעת מדעות קדומות נגד קבוצות מוגנות באוכלוסייה (יש לציין כי קבוצות אלו

813 אבן חן, **לעיל** ה"ש 802, בעמ' 384-385.

814 Kevin Kelly, *Justification as Truth-Finding Efficiency: How Ockham's Razor Works*, 14 Minds and Machines 485 (2004)

815 דן אריאלי **לא רציונלי אבל לא במקרה** 93-115 (2010).

816 Benjamin M. Marlin, Missing Data Problems in Machine Learning (Ph.D. diss., University of Toronto 2008); Gustavo E. A. P. A. Batista and Maria Carolina Monard, *An Analysis of Four Missing Data Treatment Methods for Supervised Learning*, *in* Proceedings of the First International Workshop on Data Cleaning and Preprocessing 142 (Shichao Zhang, Qiang Yang, and Chengqi Zhang eds., 2002)

יכולות להשתנות לאורך זמן). לעיתים גם כשיש כללים מוסדיים נגד אפליה מתגלה בדיעבד הטיה מערכתית לא מודעת, שכן קבלת החלטות בשלל מוסדות חברתיים משקפת את ההטיות האנושיות של בעלי התפקידים בהן. כך למשל, הטיות של שופטים עלולות להביא אותם לגזור עונשים חמורים יותר על נאשמים מקבוצות מיעוט חברתיות או אתניות;[817] הטיות של שוטרים עלולות להביא לאכיפת יתר בקרב אוכלוסיות מיעוטים או ליד קלה על ההדק;[818] הטיות ודעות קדומות עלולות להשפיע על שיקולים של בנקאים בבואם להחליט אם לתת הלוואה או משכנתה,[819] או על שיקולים של בעלי נכסים הנמנעים במודע או שלא במודע מלהשכיר אותם לשוכרים מקבוצות מיעוט מסוימות;[820] וגם החלטות

817 ראו לדוגמה Gideon Fishman, Arye Rattner, and Hagit Turjeman, *Sentencing Outcomes in a Multinational Society: When Judges, Defendants and Victims Can Be either Arabs or Jews*, 3 Eur. J. Criminology 69 (2006); Guy Grossman et al., *Descriptive Representation and Judicial Outcomes in Multiethnic Societies*, 60 Am. J. Pol. Sci. 44 (2016); Jonathan P. Kastellec, *Race, Context, and Judging on the Courts of Appeals: Race-Based Panel Effects in Death Penalty Cases*, Justice System Journal (11.11.2020); Crystal S. Yang, *Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing*, 44 J. Leg. Stud. 75 (2015); Michele Benedetto Neitz, *Socioeconomic Bias in the Judiciary*, 61 Cleveland State L. Rev. 137 (2013); Norman L. Green, *How Great Is America's Tolerance for Judicial Bias: An Inquiry into the Supreme Court's Decisions in Caperton and Citizens United, Their Implications for Judicial Elections, and Their Effect on the Rule of Law in the United States*, 112 W. Va. L. Rev (2020)

818 Joshua Correll et al., *Across the Thin Blue Line: Police Officers and Racial Bias*, *in* Personality & Soc. Psych. 1006 (2007); Jeremy West, *Racial Bias in Police Investigations* (working paper, 2018); Cody T. Ross, *A Multi-Level Bayesian Analysis of Racial Bias in Police Shootings at the County-Level in the United States, 2011-2014*, 10 PLoS One (2015)

819 J. Michelle Brock and Ralph De Haas, Gender Discrimination in Small Business Lending: Evidence from a Lab-in-the-Field Experiment in Turkey (EBRD Working Paper number 232, 2019). להטיות מבניות בשוק המשכנתאות הישראלי ראו לדוגמה בנק ישראל, **ניתוח שוק המשכנתאות ללווים מהמגזר הערבי על רקע הכשלים המבניים בתחום הדיור במגזר זה** (21.11.2017).

820 ראו לדוגמה Terrence McCoy, *Eviction Isn't Just About Poverty. It's Also About Race and Virginia Proves It* The Washington Post (11.11.2018). לדוגמה לאפליה גזענית מודעת ראו אור קשתי "צעירות ערביות תבעו מתווך שאמר להן שדירה הושכרה – ולחברותיהן היהודיות אמר ההפך" **הארץ** (6.12.2019).

של מעסיקים וממונים על כוח אדם בעניין גיוס עובדים יכולות להיות מוטות.[821] כל ההטיות האלו עלולות להביא לאפליה מודעת ומכוונת כשמקבל ההחלטות "בשטח" מודע להטיות שלו נגד קבוצת המיעוט (ומתכוון להפלותה), אבל קבלת החלטות מפלה יכולה להיות מושפעת גם מהטיות לא מודעות שלו.

6.2 הטיות אלגוריתמיות

גם הרכיב האנושי בפיתוח מערכות בינה מלאכותית יכול להיות מוטה – אם במודע ואם שלא במודע – ויש חשש שהטיות המפתח ישועתקו לתוך המכונה כבר בשלב הפיתוח.[822] בפיתוח מערכות מומחה נדרשת מעורבות אנושית כדי לתרגם כללי החלטה אנושיים לשפת מכונה או כדי לפתח מודל סטטיסטי של עץ החלטה. אך גם בפיתוח מבוסס למידת מכונה יש מעורבות אנושית ניכרת בהכנת נתוני האימון, בהגדרת משימת מטרה ובפיקוח על תוצאת האלגוריתם.[823] יתר על כן, המעורבות האנושית בפיתוח בינה מלאכותית, בבקרה עליה ובשימוש בה תורמת לאמון במערכות אלו.[824] ההטיות הקוגניטיביות

821 Faye K. Cocchiara, Myrtle P. Bell, and Wendy J. Casper*, Sounding "Different": The Role of Sociolinguistic Cues in Evaluating Job Candidates,* 55 Human Resource Management 463 (2016); Stuart W. Flint et al., *Obesity Discrimination in the Recruitment Process: "You're Not Hired!"*, 7 Frontiers in psychology 647 (2016); Devah Pager, Bruce Western, and Bart Bonikowski, *Discrimination in a Low-Wage Labor Market: A Field Experiment*, 74 American Sociological Review 777 (2009); Geoffrey Beattie and Patrick Johnson, *Possible Unconscious Bias in Recruitment and Promotion and the Need to Promote Equality*, 16 Perspectives: Policy and Practice in Higher Education 7 (2012)

822 Janneke Gerards and Raphaële Xenidis, Algorithmic Discrimination in Europe: Challenges and Opportunities For Gender Equality and Non-Discrimination Law 41-42 (2021)

823 Sonia Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 U.C.L.A. L Rev. 54, 67 (2019); Kate Crawford, *The Hidden Biases in Big Data*, Harv. Bus. Rev. (1.1.2013)

824 ראו ס' 22 של התקנות הכלליות בדבר הגנת מידע (GDPR); AI HELG 2019, **לעיל** ה"ש 298, בעמ' 12; *Asilomar AI Principles*, **לעיל** ה"ש 326, עיקרון 16; EU White paper 2020, **לעיל** ה"ש 300, בעמ' 21; Aziz Z. Huq, *A Right to a Human Decision*, 106 Va. L. Rev. 611 (2020)

האנושיות, כמו הטיות מודעות או לא מודעות נגד קבוצות מיעוט מוגנות, עלולות לזלוג למערכות שבפיתוחן מעורבים בני אנוש.[825]

כמו כן, לצד היתרונות של מערכות בינה מלאכותית בקבלת החלטות חפה מהטיות קוגניטיביות אנושיות, יש לזכור שמערכות אלו אינן חסינות מהטיות מתודולוגיות הנובעות מתהליכי הפיתוח שלהן.

6.2.1. ייצוג יתר וייצוג חסר של אוכלוסיות

מאחר שבינה מלאכותית, ובפרט מערכות המבוססות על למידת מכונה, מבוססת על מודלים סטטיסטיים של המציאות, מרבית ההטיות שלה אינן שונות מהטיות של מחקרים אקדמיים המציעים מודלים כאלו. כך למשל, במדעי החברה נהוג להבחין בין תוקף פנימי של מחקר לתוקף חיצוני שלו. תוקף פנימי הוא קריטריון שבוחן את השאלה אם המחקר תוכנן כך שמסקנותיו הפנימיות נכונות, כלומר אם הוא מבוסס על קשר סיבתי תקף. תוקף חיצוני הוא קריטריון שאומד את היכולת להחיל את מסקנות המחקר על מי שאינו משתייך לאוכלוסיית המדגם.[826] ביקורת נפוצה על תוקף חיצוני של מחקרים במדעי החברה המבוססים על שאלוני סטודנטים או על ניסויים בהשתתפותם היא שמדגם זה אינו מייצג את האוכלוסייה הכללית (על אחת כמה וכמה כשזה מדגם של סטודנטים הלומדים מקצוע מסוים).[827] מחקרים כאלו יכולים אולי לשקף את מבנה האישיות, העמדות וההתנהלות של סטודנטים בני 18-21 בארצות הברית, אך יש להיזהר מהחלת מסקנותיהם על האוכלוסייה הכללית.

825 Katyal, **לעיל** ה"ש 823, בעמ' 66.

826 Donald T. Campell and Julian C. Stanley, Experimental and Quasi-Experimental Designs For Research 5-6 (1963); Jeffrey W. Lucas, *Theory-Testing, Generalization, and the Problem of External Validity*, 21 Sociological Theory 236, 236-238 (2003)

827 David O. Sears, *College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature*, 51 J. of Personality and Soc. Psych. 315; James N. Drukman and Cindy D. Kam, *Students as Experimental Participants: A Defense of the "Narrow Data Base," in* Cambridge Handbook of Experimental Political Science 41 (James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia eds., 2011)

בדומה, נתוני אימון שעליהם מתבססת הבינה המלאכותית יכולים להניב מודל שתוקפו הפנימי הוא לעילא ולעילא, אך אין לו תוקף חיצוני. כשנתוני האימון מבוססים על מדגם לא פרופורציונלי של האוכלוסייה (ייצוג יתר או ייצוג חסר), המודל עלול לכלול הטיות לטובתם או לרעתם של מגזרים מסוימים. למשל, חברת אמזון פיתחה כלי בינה מלאכותית שנועד לסייע למחלקת משאבי אנוש של החברה על ידי אוטומציה של תהליכי גיוס של מהנדסים. בסיס הנתונים שהחברה הסתמכה עליו הורכב ברובו מפרופילים של מהנדסים גברים, ונמצא שהכלי שפותח מוטה נגד מועמדות והוא מסנן קורות חיים של נשים.[828] ייתכן שגם דוגמאות אחרות של הטיות ברורות מלמדות על מקרים דומים של ייצוג חסר – נמצא שאלגוריתם הפרסום של Google Ads מיעט לפרסם לנשים הצעות לעבודה במשרות יוקרתיות,[829] אלגוריתמים לזיהוי תמונה סיווג אנשים כהי עור כגורילות[830] ואלגוריתם של מצלמה דיגיטלית זיהה תווי פנים אסיאתיים כממצמצים.[831] בכל הדוגמאות האלה נשלל התוקף החיצוני של המודל שפיתחה הבינה המלאכותית.

6.2.2. שעתוק הטיות מבניות קיימות

מערכות בינה מלאכותית יכולות לשעתק הטיות מבניות קיימות אם בסיס הנתונים שהן נשענות עליו מוטה. דוגמה מוקדמת היא תוכנה שפותחה בבית הספר לרפואה סנט ג׳ורג׳ בבריטניה בתחילת שנות השמונים של המאה שעברה כדי לסייע לוועדת הקבלה לסנן את המועמדים הרבים ללימודים. מטרת התוכנה הייתה לחקות את דפוסי קבלת ההחלטות של הוועדה וליישם אותם על

828 Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, Reuters (11.10.2018)

829 Amit Datta, Michael Carl Tschantz, and Anupam Datta, *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination*, 1 Proceedings on Privacy Enhancing Technologies 92 (2015)

830 רועי גולדנברג "פאדיחה של גוגל: מזהה אנשים כהי עור בתור גורילות" **גלובס** (2.7.2015); איתן לשם "פייסבוק התנצלה אחרי שהבינה המלאכותית שלה תייגה אדם שחור כ׳פרימט׳ " **הארץ** (6.9.2021).

831 Odelia Lee, *Camera Misses the Mark on Racial Sensitivity* Gizmodo (15.5.2009)

מועמדים חדשים כדי להבטיח עקביות בקבלת המועמדים. התוכנה אכן הצליחה לעשות כן,[832] אלא שקבלת ההחלטות ההיסטורית בוועדה האנושית הייתה ספוגה בהטיות נגד נשים ומיעוטים אתניים. כלומר ההטיות המפלות של הוועדה הוטמעו במודל שהתוכנה השתמשה בו.[833] יש לציין שתוכנה זו פותחה על בסיס מידול סטטיסטי "אנושי" ולא באמצעות למידת מכונה. המפתח האנושי של המודל הוא שהכניס לבסיס הנתונים במודע "גזע" ו"מגדר" כמשתנים מסבירים (הם לא היו קיימים בבסיס הנתונים המקורי, אך הוסקו מצירוף של משתנים אחרים, כגון שם ומקום לידה), וכך הצליח לשעתק ברמת דיוק גבוהה את ההטיות של הוועדה האנושית. קל וחומר שביצועיו של אלגוריתם שפותח על בסיס נתוני אימון מוטים ללא התערבות אנושית יהיו מוטים.

6.2.3. הטיות בפיתוח

(1) הגדרת משתנים

בארוקס וסלבסט מיפו מופעים של אפליה בעיבוד נתוני עתק לפי שלבי המידול הסטטיסטי של הנתונים.[834] בדומה להטיות שזיהו השניים בטכניקות סטטיסטיות של כריית נתוני עתק, שאינן אלא זיהוי תבניות או חיזוי בהתבסס על מתאם סטטיסטי, גם הטיות בבינה מלאכותית יכולות להיווצר לאורך כל שרשרת הפיתוח.

ייתכנו הטיות כבר בשלב הראשוני של הגדרת משתני מטרה, שהם התוצאות של המודל או המשתנים שהבינה המלאכותית מנסה לחזות. לעיתים קרובות מפתחי המודל הם שיוצרים ומגדירים באופן מלאכותי את משתנה המטרה; למשל, אלגוריתם לחיזוי מידת היותם של מועמדים לעבודה עובדים "מוצלחים" מחייב להגדיר באופן מלאכותי לפי פרמטרים כמותיים שונים עובד מוצלח מהו. פרמטרים

832 Commission For Racial Equality, Medical School Admissions: Report of a Formal Investigation into St. George's Hospital Medical School 9 (1988)

833 Stella Lowry and Gordon MacPherson, *A Blot on the Profession*, 296 British Med. J. 657 (1988); Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy 116-117 (2016)

834 Solon Barocas and Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Ca. L. Rev. 671 (2016)

אלו יכולים לפעמים להיות מוטים,[835] אבל נמצא שהתוצר הסופי של מערכות נבונות שמשתמשות במשתנים קטגוריאליים מלאכותיים כאלה עקבי יותר מתוצר המסתמך על הערכות אנושיות שמסתמכות על אותם משתנים, ויש לו אף פוטנציאל גבוה יותר לתמוך באחריותיות של המודל. כמו כן, קטגוריות יכולות להיות דינמיות ולהשתנות לאורך זמן; האמידי, שורמאן וברנהאם הראו שקיבוע בינארי של משתנה המגדר בבסיסי נתונים של מערכות אוטומטיות לזיהוי מגדר עלול לדכא התפתחות של קטגוריות מגדריות נזילות יותר.[836]

לעיתים רידוד פונקציית המטרה של המודל למשתנה אחד או משתנים אחדים מתעלם ממורכבות התוצאה הרצויה.[837] כך למשל, אימון מערכת לנהיגה אוטונומית שפונקציית המטרה שלה היא מזעור זמני הנסיעה עלול להביא לפיתוח מודל שעל פיו נכון לסכן את יושבי הרכב כדי להגשים מטרה זו. לעיתים שגיאות שמקורן בהתעלמות מפרמטרים מסוימים בזמן פיתוח המודל או הזנחת הצורך להוסיף אילוצים לפונקציית המטרה (למשל מזעור זמני הנסיעה בתנאי שכל יושבי הרכב מגיעים ליעדם בריאים ושלמים) עלולות לייצר תוצאות מוטות.

6.2.4. הטיות בפיתוח (2) בסיס נתוני האימון

בארוקס וסלבסט[838] הצביעו על מקור אפשרי אחר להטיה – בסיס נתוני האימון, שכבר עמדנו עליו קודם בכמה דוגמאות. מדעני מחשב מושכים לעיתים בכתפיהם לנוכח תוצאות מפלות ומפטירים: "זבל נכנס – זבל יצא"

835 Joseph M. Stauffer and M. Ronald Buckley, *The Existence and Nature of Racial Bias in Supervisory Ratings*, 90 J. Applied Psychol. 586, 588-89 (2005)

836 Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham, *Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems*, *in* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018). ראו גם Os Keyes, *The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition*, Proceedings of the ACM on Human-Computer Interaction, CSCW, Article 88 (2018)

837 ועל כך ראו **להלן** בסעיף 6.2.7.

838 Barocas and Selbst, **לעיל** ה"ש 834, בעמ' 684-687.

(garbage in, garbage out).[839] שנים רבות הפיק אלגוריתם החיפוש של גוגל תוצאות פורנוגרפיות במענה לשאילתות על נשים או נערות שכללו מילות חיפוש אתניות ("נערות שחורות", "נשים אסיאתיות"); כלומר בסיס נתוני האימון משקף, ובה בעת מנציח, יחסי כוח חברתיים, פטישיזציה והחפצה היסטוריים.[840] בסיס נתוני האימון יכול להיות מוטה אפריורית כי הוא נשען על נתונים היסטוריים המשקפים הטיות מערכתיות – בדומה לתוכנת סינון המועמדים לבית הספר לרפואה[841] או לאפליה היסטורית (בארצות הברית למשל יש מתאם סטטיסטי גבוה בין מקום מגורים למוצא אתני עקב פרקטיקות הפרדה גזעית).[842]

מבחינה זו, לעיתים מערכות מוטות של בינה מלאכותית פשוט משקפות הטיות מבניות בנות זמננו.[843] הטיות כאלו יכולות לבוא לידי ביטוי במאפייני השימוש

839 ראו למשל Mayson, **לעיל** ה"ש 209.

840 ראו Benjamin, **לעיל** ה"ש 201, בעמ' 69–70; Safiya Umoja Noble, Algorithms of Oppression: How Search Engines Reinforce Racism (2018). לדוגמאות נוספות של מערכות בינה מלאכותית שמייחסות מיניות מוחצנת לנשים מרקע אתני מסוים ראו Karen Hao, *An AI Saw a Cropped Photo of AOC. It Autocompleted her Wearing a Bikini*, MIT Technology Review (29.1.2021); Melissa Heikkilä, *The Viral AI Avatar App Lensa Undressed Me – Without My Consent*, MIT Technology Review (12.12.2022)

841 ראו לעניין זה את ההבחנה של דבורה הלמן בין nonaccuracy affecting injustice לבין accuracy affecting injustice. אי־דיוק באיסוף הנתונים עלול להביא לתוצאות לא צודקות – למשל אם רישום הנתונים מושפע מפרקטיקות לא צודקות או כשנתונים אלו מושפעים מהערכה אנושית מוטה. לעומת זאת, ייתכן גם בסיס נתונים שנאסף בדיוק מרבי, אך בשל הטיות מערכתיות מביא אף הוא לתוצאות לא צודקות. Deborah Hellman, *Big Data and Compounding Injustice*, Journal of Moral Philosophy (2023)

842 על המשמעות האלגוריתמית של פרקטיקות אלו ראו למשל Nicole McConlouge, *Discrimination on Wheels: How Big Data Uses License Plate Surveillance to Put the Brakes on Disadvantaged Drivers*, 18 Stan. J. C.R. & C. L. (2022) 48–49. להיסטוריה המשפטית של הפרדה בדיור בארצות הברית ראו לדוגמה Charles L. Nier III, *Perpetuation of Segregation: Toward a New Historical and Legal Interpretation of Redlining under the Fair Housing Act,* 32 J. Marshall L. Rev. 617 (1999)

843 Darius Amilevičius, *Machine Bias and Fundamental Rights*, *in* Smart Technologies and Fundamental Rights 335, 342 (John-Stewart Gordon ed., 2021). ראו גם Philip Alston, Special Rapporteur on extreme poverty and human rights, brief as amicus curiae before the District Court of the Hague on the case of *NJCM c.s./De Staat der Nederlanden (SyRI)*, case No. C/09/550982/ HA ZA 18/388, September 2019, Para. 82

בשפה מסוימת, לרבות בשכיחות של ביטויים מסוימים. דוגמה להטיה המשעתקת הטיות חברתיות קיימות המגולמות בבסיס נתוני האימון הוא תרגום מכונה. המשפט בעברית "המתכנתת פיתחה מערכת בינה מלאכותית" מתורגם ביישום Google Translate למשפט באנגלית "The programmer developed an artificial intelligence system". כשאנו מבקשים לשוב ולתרגם את המשפט הזה, שבאנגלית הוא ניטרלי מבחינה מגדרית, לשפה העברית, מתקבל (נכון למועד כתיבת שורות אלו) התרגום "המתכנת פיתח מערכת בינה מלאכותית". הדבר נובע מנתוני האימון של היישום, המשקפים ככל הנראה את מאפייני השימוש המקוון בעברית מול מאפייני השימוש באנגלית, על ההטיות המגדריות שבהם.[844]

בסיס הנתונים עלול להוביל להטיות גם כיוון שהוא חסר תוקף חיצוני. כפי שראינו לעיל, ייתכנו מקרים שבנתוני האימון לא ניתן די ייצוג למיעוט מסוים באוכלוסייה והמודל שיפותח לפי נתונים אלו יכיל הטיות לרעתו – למשל בגיוס מועמדים לעבודה או בזיהוי פנים שגוי של מיעוטים אתניים. כך למשל, מערכות בינה מלאכותית שמזהות מגדר עלולות לשגות בזיהוי בני תרבויות שבהן האופן שבו המגדר מוצג שונה.[845]

גם ייצוג יתר של אוכלוסייה מסוימת בבסיס נתונים עלול לשקף הטיה מבנית או אפליה היסטורית. נתוני העבריינות בארצות הברית, למשל, שעלולים לגלם הטיות מערכתיות של מערכת האכיפה נגד אפרו־אמריקאים (כגון דיווח מוגבר ואכיפה מוגברת של עבירות בקרב שחורים), עלולים גם להביא – בשל ייצוג יתר של אוכלוסיות מיעוטים בתוכם – להטיות במודלים שמפתחים על בסיסם ומשמשים במערכות של חיזוי פשיעה, הערכת סיכון ורצידיביזם.[846]

844 כך גם תרגומם העברי של שמות בעלי המקצוע nurse ו־kindergarten teacher הוא אחות וגננת, בהתאמה. לעבודות בנושא ראו למשל Tolga Bolukbasi et al., *Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings*, 29 Advances in Neural Information Processing Systems 4349 (2016); Aylin Caliskan et al., *Semantics Derived Automatically from Language Corpora Contain Human-like Biases*, 356 Science 183 (2017)

845 Hamidi, Scheuerman, and Branham, **לעיל** ה"ש 836.

846 ראו בהקשר זה Barocas and Selbst, **לעיל** ה"ש 834, בעמ' 687. וירג'יניה יובנקס טוענת שדגימת יתר של אוכלוסיות מעוטות יכולת עלולה לגרום לשגיאות מסוג 2 (false negatives) במערכות לחיזוי התעללות או הזנחה בילדים, כך שהן לא יאתרו

6.2.5. הטיות בפיתוח (3) "לכלוך" מכוון של נתוני האימון

אך לא תמיד נתוני האימון הם נתונים היסטוריים מן המוכן, וחלק מפיתוח המודל הוא ייצורם — בייחוד בלמידה על ידי חיזוקים, המבוססת על אימון באמצעות אינטראקציות עם סביבה מבוקרת, לפי חוקים קבועים מראש. אם סביבת האימונים רוויה בהטיות גזעניות, מיזוגיניות ואחרות (אם במתכוון ואם בשל היעדר בקרה נאותה), סביר שהטיות אלו יוטמעו במערכת. כך למשל, חברת מיקרוסופט הכניסה לשימוש בטוויטר צ׳אט בוט ששמה טאי (Tay) כדי להדגים את יכולות התקשורת של הבינה המלאכותית שפיתחה החברה. הפעילות של טאי הופסקה בתוך כיממה מאחר שהצ׳אט בוט התחילה לפרסם ציוצים גזעניים, מיזוגיניים ואנטישמיים.[847] מיקרוסופט ייחסה את התנהגותה של טאי למתקפות מכוונות של גולשים,[848] אבל יש הרואים בכך היתממות, שכן החברה הייתה יכולה לצפות התנהגות כזאת כשהכניסה לשימוש בוט בסביבה רוויה בשיח שנאה[849] והייתה יכולה לכלול בו מנגנונים לסינון תוכן בעייתי.[850] ואולם לא ברור אם לקחי תקרית טאי הופקו במלואם. חמש שנים לאחר המקרה הכניסה חברה דרום קוריאנית לשימוש בפייסבוק צ׳אט בוט מבוסס בינה מלאכותית ששמה לודה ובתוך

ילדים בסיכון ממשפחות מבוססות. Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, Chapter 4 (2018)

847 "הצעות מגונות ושבחים להיטלר בטוויטר: הציוצים של הרובוטית הכושלת של מייקרוסופט" **TheMarker** (26.03.2016). להרחבה ראו Marty J. Wolf, K. W. Miller, and Frances S. Grodzinsky, *Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications*, 1 ORBIT Journal (2017); Gina Neff and Peter Nagy, *Talking to Bots: Symbiotic Agency and the Case of Tay* 10 Int'l J. Communication 4915 (2016)

848 Scott Deveau and Jing Cao, *Microsoft Apologizes After Twitter Chat Bot Experiment Goes Awry*, Bloomberg (25.3.2016)

849 דוגמה מן העת האחרונה: מערכת בינה מלאכותית שאומנה על בסיס נתונים של פוסטים באחד הפורומים הנודעים לשמצה בשיח הרעיל שלהם ברשת 4chan הפיקה רבבות פוסטים גזעניים ואנטישמיים. Matthew Gault, *AI Trained on 4Chan Becomes "Hate Speech Machine*," Vice (8.6.2022)

850 Neff and Nagy, **לעיל** ה"ש 847, בעמ' 4922; Ethan Chiel, *Who Turned Microsoft's Chatbot Racist? Surprise, It Was 4chan and 8chan*, Splinter (24.3.2016)

שבועות מספר אימנו אותה הגולשים לנקוט שיח שנאה שאילץ את מפעילה להסיר אותה מהרשת.[851]

6.2.6 הטיות בפיתוח (4) הטיית הפניה

הישענות על נתונים שמקורם במשתמשים עלולה לייצר הטיית הפניה (referral bias). מודלים מסוימים מסתמכים על דיווחי משתמשים כדי להעריך סיכונים עתידיים – למשל, מודלים לחיזוי פשיעה (predictive policing) שמסתמכים על דיווחים וקריאות של תושבים כדי להגדיר תאי שטח גאוגרפיים מסוכנים;[852] או מודלים לזיהוי משפחות שיש בהן סיכון מוגבר להתעללות בילדים שמסתמכים על דיווחים של שכנים.[853] מודלים מסוג זה עלולים לסבול מהטיות אופייניות: דיווחים כאלו לרוב שכיחים יותר באזורים עירוניים צפופים (בעלי מאפיינים חברתיים־כלכליים נבדלים) והם מתבססים לעיתים על הערכה סובייקטיבית ובלתי מקצועית של סיכון. יש אפילו מקרים שבהם נמסרים במזיד דיווחי שווא (למשל בגלל סכסוכי שכנים).

ואכן, אופן איסוף הנתונים יכול להשפיע על התוקף החיצוני של המודל. ואולם לעיתים יש הטיות שאינן נובעות משיבוש מכוון של בסיס הנתונים או מתוכן פסול, אלא מתת־ייצוג או ייצוג יתר של חלק מהאוכלוסייה. לעיתים קשה לחזות מראש כיצד יוטה איסוף נתונים שמסתמך על חוכמת ההמונים עקב התפלגות המשתמשים. למשל, עיריית בוסטון שחררה יישומון לזיהוי פסיבי של בורות בכבישים, בהסתמך על נתוני איכון לוויני של המכשיר. כשהיישומון מזהה שנפער בור בכביש הוא מדווח בזמן אמת לעירייה כדי שהיא תוכל לתקנו וכך חוסך ממנה להפעיל צוותי פיקוח עירוניים. ואולם ליישומון זה יש פוטנציאל חדירה מוגבל ואין הוא יכול להגיע לכל אוכלוסיית היעד של שירותי תיקוני

851 Dongwoo Kim, *Chatbot Gone Awry Starts Conversations About AI Ethics in South Korea,* The Diplomat (16.1.2021)

852 ראו **לעיל** בסעיף 2.3.2.

853 ראו לדוגמה Mary E. Rauktis and Julie McCrae, The Role of Race in Child Welfare System Involvement in Allegheny County (Allegheny County Dept. of Human Services, 2010). ובהרחבה ראו Eubanks, **לעיל** ה"ש 846.

הדרכים. סביר להניח שיש חלקים באוכלוסיית היעד שאינם יודעים על קיומו של היישומון; ומתוך מי שיודעים עליו, לא כולם ימהרו להתקין אותו (למשל עקב אוריינות דיגיטלית לקויה). אזורים שמתגוררים בהם אזרחים שלא התקינו את היישומון יסבלו אפוא מרמת תחזוקה לקויה של הכבישים.[854] מאפייני אוכלוסיית המשתמשים בטלפונים חכמים לא בהכרח זהים למאפייני האוכלוסייה הכללית.[855] גם בישראל הסתמכות על מידע שמקורו במשתמשי יישומונים בטלפוני חכמים עלולה לייצר עיוותים בשל דפוסי השימוש של המגזר החרדי בטלפונים.[856]

טימניט גברו טוענת ששימוש במודלים שפותחו עם נתוני אימון מוטים, ואחת היא אם מדובר בהטיות היסטוריות, בהטיות מבניות או בתת־ייצוג או ייצוג יתר של קבוצות מיעוט, עלול לייצר אפקט של משוב שמגביר, מנציח ומחזק הטיות קיימות.[857]

6.2.7. הטיות בפיתוח (5) בחירת משתנים

מקור נוסף להטיות שמתואר אצל בארוקס וסלבסט הוא בחירת משתנים (feature selection).[858] על פי רוב, בתהליך הכנת בסיס נתוני האימון מושמטים חלק מסוגי המשתנים הזמינים – שימוש בכל המשתנים לתיאור המציאות עלול לבלבל את האלגוריתם הלומד ולייצר כללים מורכבים מאוד שלא לצורך.[859] תיאור רדוקטיבי של המציאות – ובסיסי נתונים מלכתחילה

854 ראו Crawford, **לעיל** ה"ש 823. ראו גם Daniel E. O'Leary, *Exploiting Big Data from Mobile Device Sensor-Based Apps: Challenges and Benefits*, 12 MIS Quarterly Executive 179, 182 (2013)

855 Kate Crawford, *Think Again: Big Data,* Foreign Policy (10.5.2013)

856 ראו למשל שוקי פרידמן **שוק הסלולר הכשר: מחור שחור לאסדרה מאוזנת** 31, בה"ש 42 (הצעה לסדר 34, המכון הישראלי לדמוקרטיה 2020).

857 Timnit Gebru, *Race and Gender*, *in* The Oxford Handbook of Ethics of AI 253 (Markus d. Dubber, Frank Pasquale, and Sunit Das eds., 2020)

858 Barocas and Selbst, **לעיל** ה"ש 834, בעמ' 688–690.

859 Ke Wang and Suman Sundaresh, *Selecting Features by Vertical Compactness of Data*, *in* Feature Extraction, Construction And Selection: A Data Mining Perspective 71-72 (Huan Liu and Hiroshi Motoda eds., 2012)

עושים רדוקציה של המציאות לשדות נתונים מסוימים – עלול לגרום לעיוותים ולהטיות בהמשך הדרך.[860] הסתמכות על מודלים סטטיסטיים בהכרח נדרשת להכללות – ואלו עלולות לעיתים לגרום למושאים האנושיים של המודלים הסטטיסטיים האלו להרגיש שאין מכירים בערכם כאינדיווידואלים.[861] בחירת המשתנים משמיטה לעיתים משתנים שיכולים לסייע בהבחנה בין חברים שונים בקבוצה מוגנת[862] ומקשה על חשיפת הבדלים שמצביעים על שונות מהותית בין מופעים שונים במודל. כך למשל, כאשר מומחי כוח אדם רואים במקום הלימודים של מועמדים לעבודה אינדיקציה לכישוריהם המקצועיים על אף הרדוקטיביות של מאפיין זה (ואינם עורכים מבחנים אישיים לכל מועמד, שעלותם ניכרת), ייתכן שייפסלו מועמדים שלמדו במוסדות שהמוניטין שלהם אינו מיטבי גם אם לאותם מועמדים רמה מקצועית גבוהה.[863]

בחירת משתנים אינה רק מצמצמת לעיתים את מורכבות מושאי המודל ומטשטשת הבחנות רלוונטיות ביניהם. ייתכנו גם מצבים של גזענות רציונלית – מצבים שבהם המודל בוחר במשתנים פסולים מאחר שהם מנבאים טובים של משתנה המטרה או שיש מתאם גבוה בין המשתנים הפסולים למשתנים שחסרים בבסיס הנתונים. הושמעה למשל טענה שבהיעדר נתונים על עברו הפלילי של מועמד לעבודה מעסיקים בארצות הברית עלולים להניח שצבע עורו של המועמד הוא

860 ראו בהקשר זה את משל המפה והטריטוריה, שאוזכר לראשונה אצל קורזיבסקי: Alfred Korzybski, Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics 57–58 (5th ed. 1995). ראו גם חורחה לואיס בורחס, **דברי ימי תועבת העולם** 119 (רנה ליטוין מתרגמת, 1988); ז'אן בודריאר **סימולקרות וסימולציה** 7–9 (תרגום אריאלה אזולאי, 2007).

861 Laurence Thomas, *Statistical Badness*, 23 J. Soc. Phil. 30 (1992); Kasper Lippert-Rasmussen, *"We Are all Different": Statistical Discrimination and the Right to Be Treated as an Individual*, 15 J. Ethics 47 (2011). בהקשר זה ראו גם ריבלין ושני, **לעיל** ה"ש 205.

862 Toon Calders and Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases 43, 46 (Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky eds., 2013)

863 Barocas and Selbst, **לעיל** ה"ש 834, בעמ' 689. Matt Richtel, *How Big Data Is Playing Recruiter for Specialized Workers*, The New York Times (27.4.2013)

אומדן להיותו עבריין כיוון ששיעור ההרשעות בקרב בעלי צבע עור זה הוא גבוה.[864]

רוהא בנג'מין מתארת חברת הייטק בשם .Diversity, Inc, המספקת שירותי פילוח אתניים. לקוחותיה, לרוב חברות גדולות המנועות לפי חוק מאיסוף נתונים אלו, מעבירות ל־.Diversity, Inc את שמות המשתמשים או הלקוחות שלהם, את כתובותיהם ופרטים מזהים נוספים, ועל פי נתונים אלו החברה מזהה את המוצא של כל משתמש או לקוח ומגיעה, כך נטען, לרמת דיוק של 96%.[865] שירותי הזיהוי האתני של .Diversity, Inc נועדו לאפיין קהלי יעד ייחודיים לפרסומות מותאמות אישית, אך אותן מתודולוגיות יכולות להיות מיושמות בעקיפין, באופן אוטומטי, במודלים אחרים – ולגרום להטיה שלהם.

6.2.8. הטיות בפיתוח (6) משתנים חליפיים (Proxies)

חלק מסוגי המשתנים שיכולים להיבחר בהליך של בחירת משתנים הם משתנים חליפיים, שהם גורם נוסף להטיות שבארוקס וסלבסט מונים.[866] משתנים חליפיים הם משתנים שיש מתאם גבוה בינם ובין משתנים בלתי תלויים במודל. כך למשל, המפתחים של המודל לסינון מועמדים לבית הספר לרפואה סנט ג'ורג' זיהו כי מוצא אתני הוא משתנה המשפיע על הסיכוי שוועדת הקבלה תקבל מועמד פוטנציאלי ללימודים, אך מאחר שנתון זה לא היה במערכת הם השתמשו בשילוב של משתנים חליפיים (שם המועמד ומקום הלידה שלו) כדי להסיק את מוצאו האתני – בלי להיעזר בשירותי מיקור חוץ דוגמת אלו שמציעה .Diversity, Inc.[867]

864 Lior Jacob Strahilevitz, *Privacy Versus Antidiscrimination*, 75 U. Chi. L. Rev 363, 364 (2008); Amanda Agan and Sonja Starr, *Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment*, 133 The Quarterly Journal of Economics 191 (2018)

865 Benjamin, **לעיל** ה"ש 201, בעמ' 85-87.

866 Barocas and Selbst, **לעיל** ה"ש 834, בעמ' 691; ראו גם Janneke Gerards and Raphaële Xenidis, Algorithmic Discrimination in Europe: Challenges and Opportunities For Gender Equality and Non-Discrimination Law 44-45 (2020)

867 *Commission For Racial Equality*, **לעיל** ה"ש 832, בעמ' 9.

לעיתים השימוש במשתנה החליפי נעשה שלא במודע: בדיעבד מתחוור כי המודל נתן משקל גבוה למשתנה חליפי שיש מתאם גבוה בינו ובין משתנה המאפיין קבוצה מוגנת. לדוגמה, כמה ממערכות שירותי הרווחה בהולנד הסתמכו על כתובת (מיקוד) כדי לחזות הונאה פוטנציאלית. ואולם ייתכן שיהיה מתאם גבוה בין משתנה זה ובין מהגרים, ובפרט מהגרים מטורקיה ומרוקו, שכן הם נוטים להתרכז בשכונות מסוימות בערים הגדולות.[868]

כשמשתנים מסוימים מושמטים בשל הוראות סטטוטוריות (מודלים להערכת סיכוני אשראי קמעונאי, למשל, מנועים מלהביא בחשבון משתנים מסוימים: "מינו, גילו, נטייתו המינית, גזעו, דתו, ארץ מוצאו, לאומיותו, מקום מגוריו ומצבו המשפחתי או הבריאותי של לקוח"),[869] ייתכן שימוש במשתנים חליפיים שיש מתאם גבוה ביניהם ובין המשתנים האסורים שהושמטו. כשמשתנים חליפיים אלו משמשים במקום המשתנים האסורים, עם הזמן אפשר שגם הם יהפכו למשתנים אסורים כשהם לעצמם.[870] כך למשל, בקליפורניה נאסר לחשב פרמיות ביטוח רכב בהסתמך על הכתובת ודירוג האשראי של המבוטח, שכן יש מתאם גבוה בין נתונים אלו ובין קטגוריות אסורות כמו הכנסה ומוצא אתני.[871]

868 Nelleke Hijmans, Profiling the Welfare State: Upholding or Updating Human Rights Standards? A Case Study of the Netherlands 42 (Master Thesis, European Master's Programme in Human Rights and Democratisation 2017). ראו גם עמדת העותרים בעניין: Nederlands Juristen Comite voor de Mensenrechten tegen Staat der Nederlanden [Netherlands Jurists Committee of Human Rights v. State of the Netherlands], Rechtbank Den Haag [The Hague District Court], C/09/550982/HA ZA 18-388 (5.02.2020), Para. 6.92

869 ראו לדוגמה ס' 51 לחוק נתוני אשראי.

870 Devin G. Pope and Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models,* 3 American Economic Journal: Economic Policy 206, 210 (2011)

871 **שם**, שם. ראו גם Stephen D. Sugerman, *California's Insurance Regulation Revolution: The First Two Years of Proposition 103*, 27 San Diego L. Rev. 683, 693-694 (1990); Anya E. R. Prince and Daniel Schwarz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 Iowa L. Rev 1257, 1306-1310 (2020)

6.2.9. הטיות בפיתוח (7) הטיה מכוונת

גורם נוסף שיכול לגרום להטיות במודלים של בינה מלאכותית הוא טריוויאלי – הטיה מכוונת.[872] מקבלי החלטות המעוניינים להפלות קבוצות מוגנות יכולים במודע להשפיע על תהליך הפיתוח של המודל – כלומר לדאוג שיביא לתוצאות מפלות ולהסוות זאת ב"קופסה השחורה" של המכונה (masking); בין השאר, באמצעות בחירת משתנים חליפיים שיש מתאם גבוה בינם ובין הקבוצה המוגנת (לדוגמה, שימוש בקריטריון של שירות צבאי כתנאי סף להעסקת מועמדים לעבודה, כשהדבר אינו מתחייב מאופי התפקיד, כדי להסוות הבחנה בין מועמדים יהודים לערבים);[873] או באמצעות ניצול הטיות שמקורן בייצוג יתר או ייצוג חסר של מגזרים מסוימים. עם זאת, במקרים מסוימים, נוכח עלויות הפיתוח של מערכות אלגוריתמיות והקושי המובנה בהוכחת אפליה מכוונת,[874] ספק אם יש ביקוש רב למערכות שתכליתן להסתיר פרקטיקות אסורות אלו.

6.3 אתגר ההטיות האלגוריתמיות

בדומה להטיות קוגניטיביות אנושיות, גם כשהמעורבים בשרשרת הפיתוח של מערכות נבונות מודעים לאפשרות של הטיות אלגוריתמיות קשה למזער את התופעה. יש לכך כמה סיבות.

6.3.1. קושי לזהות מראש היכן יופיעו הטיות

מקצת הדוגמאות בפרק זה הן של הטיות שלא היו רצויות ומפתחי המערכת לא תכננו אותן. למשל האלגוריתם לזיהוי תמונה שזיהה בעלי צבע עור שחור כגורילות; או המערכת לזיהוי מצמוצים במצלמות דיגיטליות,

872 Barocas and Selbst, **לעיל** ה"ש 834, בעמ' 692–693.

873 ראו בעניין זה ת"פ (אזורי תל אביב) 1038/99 **מדינת ישראל, משרד העבודה והרווחה נ' תפקיד פלוס בע"מ** (פורסם בנבו, 12.6.2003).

874 Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 Stan. L. Rev. 1161, 1177 (1995)

שזיהתה תווי פנים אסיאתיים כממצמצים. פעמים רבות ההבחנה בין בסיס נתונים חלקי להטיות היא מטושטשת.

יתר על כן, גם כשנתוני האימון של המערכת אינם מתויגים כפרמטרים שהחוק פוסל – כגון מגדר, שיוך אתני או אמונה דתית ופוליטית – המערכות עלולות לזהות משתנים חליפיים שיש מתאם גבוה בינם ובין פרמטרים אלו. בהתחשב בכך שדעות קדומות אנושיות, הטיות מערכתיות או אפליה היסטורית עלולות להיות מוטמעות בנתוני האימון, לא תמיד אפשר לחזות מראש אילו הטיות טמונות באלגוריתמים.

6.3.2. רמת הפשטה גבוהה

מערכות אלגוריתמיות, כמו כל מודל סטטיסטי, עושות הפשטה מסוימת של המציאות.[875] יישומן בהקשרים חברתיים רחבים דורש שימוש בהנחות חיצוניות למודל. כך למשל, ממתאם סטטיסטי לא בהכרח אפשר להסיק סיבתיות, ולכן הסקת סיבתיות אפשרית רק בנסיבות שבהן האנליסט מניח הנחות החורגות מהנחות היסוד הנדרשות לחיזוי.[876] כדי לבחון אם באלגוריתם יש הטיות יש צורך אנליטי לצאת מגבולות המודל (שהטמיע את ההטיות) ולבחון השערות שחיצוניות לו.[877] ההפשטה של מודלים סטטיסטיים מאפשרת לכאורה להשתמש בהם בהקשרים חברתיים שונים מאלו שלשמם אומנו,[878] או במנותק מהמשמעות המורכבת והמלאה של המושגים החברתיים שהמודלים מתארים. אלו הן למעשה השלכות השימוש בנתוני אימון חלקיים (שיכולים להיות בעלי תוקף פנימי לגבי אוכלוסיית היעד של המודל).

875 ראו **לעיל** ה"ש 830.

876 Susan Athey, *Beyond Prediction: Using Big Data for Policy Problems*, 355 Science 483 (2017)

877 Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency 59, 60 (2019)

878 **שם**, בעמ' 61.

6.3.3. הגדרה דינמית ושנויה במחלוקת של מושג ההוגנות

הגדרה זו מקשה גם היא על איתור הטיות. אומנם אינטואיטיבית למושג ההוגנות יש קסם רב, אבל כבר מקדמת דנא לאנשים שונים יש תפיסה שונה של הוגנות.[879] כמו כן, זיהוין של קבוצות מוגנות הוא עניין דינמי ומשתנה לאורך זמן. יתר על כן, שיקול ההוגנות יכול להתחרות בשלב הפיתוח בשיקולים אחרים, כגון יעילות המודל, מה שמאלץ את המפתחים לקבל הכרעה ערכית באשר לנכונות לתקן אלגוריתם לשם הגשמת ערכים של שוויון והוגנות.

6.3.4. מבנה שוק המפריד בין נתונים לאלגוריתמים

כפי שראינו לעיל, מרבית ההטיות האלגוריתמיות מקורן בנתונים המשמשים לאימון. לרוב מקורם של נתונים אלו אינו בצוותי הפיתוח של המערכות הנבונות אלא במאגרי מידע חיצוניים, הניזונים מנתונים שאנשים מייצרים באופן פסיבי בשגרת יומם או ממערכות תיעוד ציבוריות שהתיעוד בהן אינו בהכרח ממוכן. כבר עמדנו על כך שגם כשמפתחים מפעילים מאמץ מודע ומכוון לטהר את הנתונים מהטיות, למשל על ידי הימנעות מהסתמכות על פרמטרים "נגועים", הנתונים עלולים לשקף הטיות סמויות שהאלגוריתם יצליח לזהות ולהחצין (לחלופין, טיהור יתר של האלגוריתם עלול להביא לתוצאות בלתי מדויקות בעליל כשמשתנים חליפיים מגלמים בערבוביה גם הטיה מערכתית אך גם מידע שתורם לנכונות המודל).

879 ראו למשל Thomas B. Nachbar, *Algorithm Fairness, Algorithmic Discrimination*, 48 Fla. St. U. L. Rev. 523–525 (2021); Stefan Feuerriegel, Mateusz Dolata, and Gerhard Schwabe, *Fair AI: Challenges and Opportunities*, 62 Bus. Inf. Syst. Eng. 379, 379 (2020); Jessie Finocchiaro et al., *Bridging Machine Learning and Mechanism Design Towards Algorithmic Fairness*, FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 489, 491–492 (2021); Doaa Abu-Elyounes, *Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness*, 20 J. L. Tech & Pol. 1 (2020)

6.4 סיכום

פרק זה סקר את אתגר ההטיות האלגוריתמיות והצביע על הדרכים השונות שבהן הן באות לכדי מימוש ועל אבני הנגף בדרך לפתרון הבעיה. כיצד אפשר אפוא להתמודד עם הטיות? באירופה, למשל, מוצע להבחין בין סוגים שונים של מערכות נבונות לפי רמת הסיכון הנשקפת מהן.[880] הבחנה זו מאפשרת לשמור על יחס מידתי בין אמצעי למטרה – ובהתאם, המחוקק האירופי אוסר אפריורית על סוגים מסוימים של מערכות נבונות. אין הוא מוכן לקבל את הנזק שהן עלולות להסב, לרבות תוצאות של הטיות אלגוריתמיות, כרע במיעוטו. בפרק 9 להלן יתוארו אסטרטגיות התמודדות נוספות עם הטיות אלגוריתמיות.

880 ראו **לעיל** בסעיף 4.1.1.1.

פרק שביעי

שקיפות אלגוריתמית

כפי שראינו בפרקים הקודמים,[881] בשנים האחרונות התעוררו חששות בדבר ליקויים בתפקוד של מערכות נבונות בעקבות התבססותן על מערך נתונים מוטה או שגוי, נטייתן להעצים הטיות קיימות והיעדר יכולת לאתר תקלות בתפקודן עקב הקושי להבין כיצד הגיעו לתוצריהן (מה שמכונה לעיתים "הקופסה השחורה").[882]

881 פרק זה לקוח מחלקו הראשון של מחקרם של גדי פרל ותהילה שוורץ אלטשולר **מודל ליצירת שקיפות אלגוריתמית** (הצעה לסדר 47, המכון הישראלי לדמוקרטיה 2022).

882 Matthew Hutson, *The Opacity of Artificial Intelligence Makes it Hard to Tell When Decision-Making Is Biased*, 58 IEEE Spectrum 40 (2021).

הטענה הרווחת היא שכדי להתגבר על אי־בהירותן של מערכות לומדות ועל הסכנות הפוטנציאליות הטמונות בהן נחוצה שקיפות אלגוריתמית. מחברות הטכנולוגיה נדרש להסביר לנו איך בדיוק האלגוריתמים שלהן עובדים ממש כשם שמחברת קוקה קולה נדרש לחשוף את המתכון הסודי שלה. ואילו התביעה מממשלות היא לא לאפשר למשטרה להשתמש במערכות זיהוי פנים עד שלא יסבירו לנו איך האלגוריתם פועל.

7.1 המושג שקיפות אלגוריתמית הוא מושג "עכור"

בתחום הרגולציה של מערכות בינה מלאכותית, ובפרט בתחום האתיקה, מרבים לעסוק ברכיב העכירות המובנה במערכות בינה מלאכותית.[883] רכיב זה מקשה את הבנת פעולתן ואת הבקרה על תפקודן התקין. כדי לפתור את חוסר הבהירות המובנה במערכות אלגוריתמיות לקבלת החלטות וכדי לקדם אחריותיות, מוצע בספרות המושג "שקיפות אלגוריתמית". התפיסה הרווחת היא שעכירות יוצרת קשיים באיתור הגורם האחראי על תקלות ואף עלולה להעצים את הנזק הנגרם מפעילותה של המכונה, מאחר שהמשתמשים אינם מסוגלים להבין מה היה צריך להיות השימוש הנכון או לאתר את המקור לתקלה.

ואכן, כפי שראינו בחלק 3.1 לעיל, סוגיית השקיפות האלגוריתמית עולה כמעט בכל המסמכים העוסקים ברגולציה של מערכות אלגוריתמיות והיא חלק מכריע ממארג הפתרונות המוצע.[884]

ואולם חרף העיסוק הנרחב במושג זה, הוא עצמו "עכור", כלומר אינו נהיר דיו. הגדרתו, מקומו ביחס לערכים אחרים, זיהוי השחקנים שמוטלת עליהם

על הנזק החברתי של מערכות אלו ראו הערת הדַּוָּח המיוחד לעוני קיצוני וזכויות אדם של האומות המאוחדות, Report of the Special Rapporteur on Extreme Poverty and Human Rights (A/74/48037, 2019); *World Stumbling Zombie-Like into a Digital Welfare Dystopia, Warns UN Human Rights Expert*, UN OCHR (17.10.2019)

883 על עכירות אלגוריתמית ראו: Virginia Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way 59 (Springer Nature 2019)

884 Fjeld et al., **לעיל** ה"ש 292, בעמ' 15.

חובת השקיפות, זהות השחקנים שנדרשת שקיפות כלפיהם ומקומה של חובת השקיפות האלגוריתמית לאורך שרשרת הערך או חיי המוצר – כל אלו פתוחים לפרשנות ושנויים במחלוקת. בהיעדר הגדרה מדויקת מתעורר החשש שהשקיפות לא תגשים את ייעודה. יתרה מזו, עצם השימוש במושג השקיפות כעיקרון לא ברור עלול לסכל את השימוש בו ככלי לעידוד התנהלות אתית תקינה ולסייע למי שמבקשים שאתיקה תהיה עניין הצהרתי בלבד (איתות סגולה [virtue signaling], המכונה בספרות גם ethics washing, ethics shopping, ethics shrinking, ethics dumping ואף ethics lobbying).[885]

ממסמכי המדיניות עולה שהכוונה במושג שקיפות אלגוריתמית היא ליישם את עקרון השקיפות הקלסי על מערכות טכנולוגיות. שקיפות מופיעה במקור בספרות המשפטית העוסקת במינהל תקין, בעיקר בהקשר של החובה של מערכות ממשל והליכי החקיקה להיות פתוחים לציבור ולביקורת מצד מערכת המשפט. לפיכך היא מופיעה אצל קולין פורלצה, למשל, כערך מרכזי של המשפט המינהלי.[886] לפי ארגון ה־OECD, שקיפות, אחריותיות ושיתופיות הן עקרונות מרכזיים של מינהל תקין והן עומדות ביסוד מערכות השלטון שלו.[887] השקיפות נתפסת בספרות גם כעיקרון דמוקרטי מרכזי, ומקובל לומר שברירת המחדל של השלטון צריכה להיות שקיפות.[888] גם במיזמי "ממשל פתוח" ברחבי העולם שקיפות מוזכרת כערך יסודי.[889]

885 **לעיל** סעיף 3.8.

886 Colin Prolezza, *Data Journalism and the Ethics of Open Source: Transparency and Participation as a Prerequisite for Serving the Public Good*, *in* Good Data 189 (Angela Daly, Kate Devitt, and Monique Mann eds., 2019). ראו גם דפנה ברק ארז "המשפט המנהלי והמאבק בשחיתות שלטונית" **משפטים** לז 667 (תשס"ז).

887 OECD, European Principles For Public Administration 8 (SIGMA Papers No. 27, 1999)

888 Tal Z. Zarsky, *Transparent Predictions*, 2013 U. Ill. L. Rev 1503 (2013)

889 ראו למשל את הכרזת ממשל אובמה בנושא. Barack Obama, *Transparency and Open Government*, *Memorandum for the Heads of Executive Departments and Agencies*, The White House (2009)

בהקשר של חובות השלטון מתבטאת השקיפות בשני מאפיינים מרכזיים. האחד, החובה **להעמיד** לרשות הציבור מידע שלטוני.[890] השני, החובה **להנגיש** מידע זה באופן שיהיה מובן לציבור.[891] על חובות אלו נוספת גם החובה לייצר לציבור את הזכות **להשתמש** במידע שמועמד לרשותו.[892] בעידן המידע משתכללים האופנים שבהם השלטון ממלא את חובת השקיפות: מאגרי המידע של המדינה עומדים לרשות הציבור; חקיקה והחלטות שלטוניות מפורסמות ברשת האינטרנט; ונמשך הפיתוח של פלטפורמות דיגיטליות לצורך מעקב אחר תהליכים שלטוניים, כגון שימוש בתקציב המדינה.[893]

במגזר הפרטי, חובות השקיפות מתקיימות במגוון הקשרים. ראשית, כמו במגזר הציבורי, חובות אלו מוטלות על גופים המספקים שירותים ציבוריים או נתפסים גופים דו־מהותיים.[894] שנית, בהקשר של הגנת הצרכן, כגון סימון מוצרים, גילוי נאות של מאפייני המוצר ותנאי התשלום, האספקה והאחריות.[895] בנושאים אלו יש כניסה לפרטי פרטים; למשל, גודל התווים, בהירות השפה, כמות הפרטים הנמסרים בעניין מהות העסקה ואורך הטקסט. כאן אפשר לכלול גם סוגיות של הסכמה המותאמת לסוגי האוכלוסיות. שלישית, דיני החברות כוללים חובות שתפקידן לייצר אמון וביטחון במסחר בבורסה ולעודד השקעות. כאלה הן למשל חובות השקיפות החלות על חברות הנסחרות בבורסה – פרסום דוחות עיתיים הכוללים מידע על רווחי החברה ועל אירועים משמעותיים בה, ובכלל זה כל חקירה פלילית נגד נושא משרה בחברה בעניין פעילות הקשורה לעבודתו.[896] נורמות שקיפות מרחיבות, בעיקר במדינות מתפתחות, הן כלי אפקטיבי מומלץ

890 תהילה שוורץ אלטשולר **מדיניות ממשל פתוח בישראל בעידן הדיגיטלי** (מחקר מדיניות 91, המכון הישראלי לדמוקרטיה 2012).

891 **שם**.

892 **שם**, בעמ' 144.

893 לעניין זה ראו למשל הניסיונות לפתיחת מאגרים משטרתיים הנוגעים לאכיפה כחלק מהליכי שקיפות: *Transparency and Accountability at the Frontlines of Justice: Police Data Transparency*, Open Government Partnership (8.7.2020)

894 אסף הראל **גופים ונושאי משרה דו־מהותיים** 47 (2019).

895 ס' 4 לחוק הגנת הצרכן, התשמ"א–1981, ס"ח 1023 בעמ' 248.

896 ראו למשל ס' 15, 16 לחוק ניירות ערך, תשכ"ח–1968, ס"ח 541 בעמ' 234.

ליצירת אמון אצל משקיעים זרים כדי לעודד השקעות חיצוניות.[897] חובות שקיפות מוטלות גם על גופים פרטיים מפוקחים, למשל הבנקים, הנדרשים לפרסם מידע על אחזקותיהם, התפלגות הרווחים, עמלות ועוד.

עם זאת, שלא כמו בגופים שלטוניים, שבהם ברירת המחדל היא שקיפות ויש צורך בטעמים חזקים כדי לעצור הנגשה של מידע, במגזר הפרטי מאזן האינטרסים שונה ולעקרון השקיפות יש משקל דומה לזה של אינטרסים שמנוגדים לו, כגון שמירה על סודות מסחריים, הזכות לפרטיות והזכות לקניין רוחני.[898] כך למשל, בימינו דנים הרגולטורים באיחוד האירופי בהגבלה של חובת השיתוף בין רכבים מקושרים לרכבים אחרים בכביש אל מול הזכויות לפרטיות וההגבלות על השיתוף.[899]

ביסוד עקרון השקיפות עומדות כמה הצדקות נורמטיביות, פוליטיות וכלכליות.

שקיפות נתפסת כשלב מקדים והכרחי ליצירת אחריותיות אצל מקבלי ההחלטות. שקיפות ככלי לעידוד בקרה מופיעה במשפט הישראלי[900] בהקשר של חוק חופש המידע, אך גם בהקשרים כגון החובה לפרסם מכרזים וחובות הפרסום של רשות המיסים. כפי שמציינים ג׳סיקה מורלי ועמיתיה,[901] בהקשרים כאלה נתפסת השקיפות כרכיב מרכזי ביצירת אמון בין היחיד לשלטון ולכן יש לה תפקיד

897 Belay Seyoum and Terrell Manyak, *The Impact of Public and Private Sector Transparency on Foreign Direct Investment in Developing Countries*, 5 Critical Perspectives on International Business 187 (2009)

898 ראו Zarsky, **לעיל** בה"ש 888.

899 Araz Taeihagh and Hazel Si Min Lim, *Governing Autonomous Vehicles: Emerging Responses for Safety, Liability, Privacy, Cybersecurity, and Industry Risks,* 39 Transport Reviews 103 (2019)

900 בג"ץ 3751-03 **יוסי אילן נ׳ עיריית תל־אביב־יפו**, פ"ד נט(3) 817.

901 Jessica Morley et al., *From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices*, 26 (4) Science And Engineering Ethics 2141 (2020)

חשוב בדין המינהלי – בפרט בסוגיות של הנמקת החלטות שיפוטיות ומינהליות ושל שקיפות של מסמכי תקציב.[902]

שקיפות ממלאת תפקיד חשוב גם בהגברת מעורבותו של הציבור בהליכי השלטון. מעורבות ציבורית יכולה להוביל לביוש (שיימינג) של מי שביצעו פעולות שאינן בהכרח שליליות.[903]

פרסום הליכי השלטון מגביר את אמון הציבור בו וממילא מגדיל את שיתוף הפעולה שלו – ולכן אפשר לומר שלשקיפות יש גם תועלת כלכלית.[904] שקיפות היא גם נגזרת של הזכות לאוטונומיה ולחופש ביטוי,[905] שכן אזרח זכאי לקבל מידע ולהבין את הרציונל העומד מאחורי ההחלטות המתקבלות בעניינו. השקיפות ככלי לקידום אוטונומיה מוצאת את ביטויה, בין השאר, בהוראת ס׳ 13(d) לתקנות הנציבות האירופית להגנה על היחיד במסגרת איסוף מידע, המורה לחשוף לפני היחיד לא רק את המידע על החלטות המתקבלות בעניינו אלא גם את ההיגיון שמאחוריהן.[906]

כדי ליישם את עקרון השקיפות הקלסי בתחום מערכות הבינה המלאכותית יש להגדיר תחילה את המושג שקיפות אלגוריתמית, את היקפו ואת היקף החובה שתוטל על הגופים השונים לממש אותו. כדי לעשות זאת יש לענות על שאלות בנוגע למטרה שהשקיפות האלגוריתמית באה לשרת: האם השקיפות נועדה לאפשר

902 Carol Harlow and Richard Rawlings, *Proceduralism and Automation: Challenges to the Values of Administrative Law*, *in* The Foundations and Future of Public Law 275 (Elizabeth Fisher, Jeff King, and Alison L. Young eds., 2020)

903 ראו Zarsky, **לעיל** ה"ש 888.

904 Sara Hagemann and Fabio Franchino, *Transparency vs Efficiency? A study of Negotiations in the Council of the European Union*, 17 European Union Politics 408 (2016)

905 ראו Zarsky, **לעיל** ה"ש 888, בעמ' 1545.

906 Article 13(d), Regulation 45/2001/EC of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data, Official Journal of the European Communities L 8/1

בקרה אתית או לשמש כלי אכיפה משפטי? מול מי נדרשת שקיפות שלטונית לעומת שקיפות אלגוריתמית? מהן החובות הנגזרות מעקרון השקיפות? על פי אילו אמות מידה נגדיר שקיפות אלגוריתמית "טובה"? בכל השאלות האלו נעסוק להלן.

7.2 שקיפות אלגוריתמית ומשמעויותיה

על אף ריבוי המסמכים העוסקים במדיניות רגולציה ובאתיקה של בינה מלאכותית, ועל אף שכיחותו של מושג השקיפות בספרות העוסקת ברגולציה של בינה מלאכותית וגרעין של הסכמה באשר לצורך בהטלת חובות שקיפות, אין קונצנזוס על הגדרת המושג שקיפות או על מהות שלבי החובות שיוטלו ועוצמתן. יש הסבורים כי חובות השקיפות חלות על כל המשתמשים, ואילו אחרים מגבילים את תחולתן לשחקנים שונים בשלבים שונים. יש המוסיפים על חובות השקיפות גם את החובה לגילוי נאות בתקשורת שהתוכנה הנידונה אינה אנושית, ואילו אחרים רואים בשקיפות מעין חובה חוזית באשר למהות העסקה, ותו לא.

כותבים שונים מייחסים למושג שקיפות אלגוריתמית משמעויות שונות ופורטים אותו למושגי משנה בהתאם לתפיסתם בדבר חשיבותו ומקומו בשרשרת הערך של הייצור ובמבנה הרגולטורי שהם מציעים. הגדרת המושג מושפעת גם מההיבט הנבחן – למשל ההיבט של חובות היצרנים למשתמשי הקצה או ההיבט של תכלית השקיפות (האם השקיפות נתפסת אמצעי להשגת אמון או שהיא תכלית בפני עצמה?) כל כותב, על פי תפיסת עולמו, מטמיע במושג השקיפות מושגים מהעולם הצרכני של הגילוי הנאות או מעולם המשפט המינהלי או מעולם השלטון. ולעיתים כל אלה משמשים בערבוביה.

בספרות הקיימת, בהצעות חקיקה ובמסמכי מדיניות ניתנות למושג שקיפות אלגוריתמית משמעויות שונות, מהן שמתמקדות ביצרני המערכות והמוצרים ומהן שמתמקדות בזכויות המשתמשים. לדברי וירג'יניה דיגנום, לשקיפות מובן כפול: הן חובת גילוי של המידע שביסוד האלגוריתם והשימוש במערכת הן חובת גילוי של תהליך קבלת ההחלטות והיכולת להבין ולבקר תהליך זה: "שקיפות מציינת את היכולת לתאר, לבדוק ולשחזר את המנגנונים שבאמצעותם מערכת בינה מלאכותית מקבלת החלטות ולומדת להתאים את עצמה לסביבתה ואת המקור והדינמיקה של הנתונים המשמשים את המערכת והמיוצרים

באמצעותה".[907] כלומר דיגנום שמה את הדגש על החובה להעביר מידע – יש לאפשר למשתמשים לבדוק את דרך פעולתן של מערכות מבוססות בינה מלאכותית, לשחזר אותה ולאמת את המידע ואת האופן שבו התוכנה השתמשה בו.

מושג השקיפות כולל את החובות האלה:

- **עקיבות** (traceability) – שקיפות בנוגע להליך יישומו של התכנון ההנדסי של התוכנה. העקיבות אמורה לאפשר את איתור הנורמות שבבסיס הזנת המידע ואת איתור ההטיות במידע.[908]
- **וידואיות** (verifiability) – יצירת לוג שמאפשר מעקב אחר הליך קבלת ההחלטות של התוכנה.
- **עיצוב הגון** (honest design) – עיצוב ממשק המשתמש כך שיאפשר למשתמש להבין את פעולת המערכת ולקבל את המידע הנכון.
- **בהירות** (intelligibility) – החובה להבטיח שאדם יבין מה אירע בפעולתו של האלגוריתם ומדוע.
- **הסברתיות** (explicability) – על המשתמש להבין את המידע על פעילות התוכנה. ההסברתיות כוללת את שקיפות תהליך קבלת ההחלטה; תקשור נכון בדבר יכולותיה ומטרותיה של המערכת; ואת החובת להסביר את ההחלטה שהמכונה מקבלת למי שמושפעים מהחלטה זו.[909]
- **תקשור** (communication) – כשמדובר במערכת בינה מלאכותית שמתקשרת עם בני אדם יש חובה להבהיר זאת במפורש.[910] חובה זו קשורה לגילוי נאות כלפי הצרכן ומתכתבת עם דיני הפרטיות האירופיים, בהם החובה לגילוי נאות כשהמערכת מקבלת החלטה אוטומטית בעניינו של הפרט.[911]

907 ראו Dignum, **לעיל** ה"ש 883, בעמ' 54.

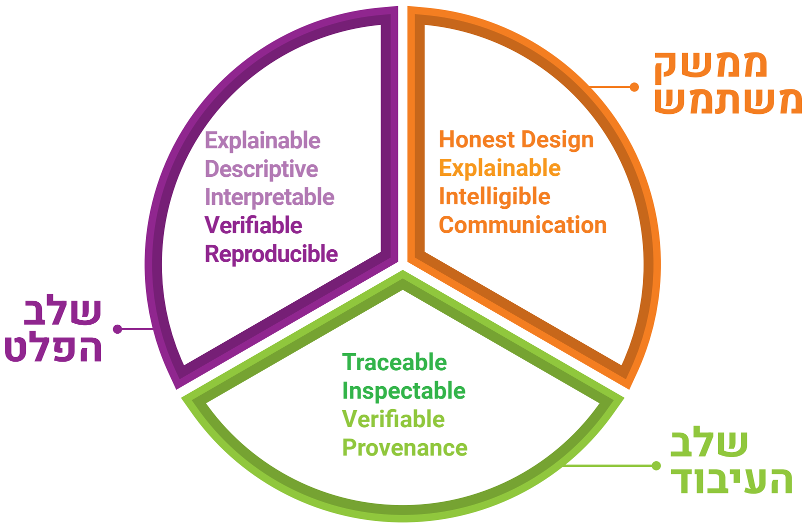
908 IEEE, **לעיל** ה"ש 321, בעמ' 11.

909 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 15.

910 **שם**, עמ' 18.

911 ס' 22 של התקנות הכלליות בדבר הגנת מידע (GDPR).

תרשים 3
משמעויות שונות של המונח שקיפות, לפי רצף פעולת התוכנה



תרשים 3 המתאר את המשמעויות שניתנו במקורות שנסקרו לעיל למושג השקיפות לפי רצף פעולת התוכנה, מלמד שהמושג פתוח לפרשנויות רבות – כל פרשנות מדגישה רכיב מסוים בשלבי פעולתו של האלגוריתם ובפועל מטילה חובות שונות על היצרן. אפשר לראות שבלי קונקרטיזציה השקיפות האלגוריתמית עלולה להיות רחבה מדי או צרה מדי ולא להתפרש על פני מכלול הפעילות של המערכת הלומדת.

7.2.1 שקיפות ככלי עצמאי או כחלק ממערך כלים

ההבדלים בפרשנות למושג שקיפות אלגוריתמית תלויים גם בשאלה אם רואים בה ערך עצמאי או אמצעי להשגת מטרות אחרות, כגון אחריותיות.

להבדלים אלו יש השלכות על האיזון בין חובת השקיפות ובין אינטרסים אחרים שייתכן שיישומה מנוגד להם, כגון זכויות קניין יצרניות, הגנה על פרטיות ועוד.

את הראייה הרחבה ביותר של מושג השקיפות אפשר למצוא במסמך Ethically Aligned Design (EAD) של ארגון התקינה הבינלאומי. המסמך מציין שלושה עקרונות אתיים מרכזיים שעל פיהם יש לעצב ולתפעל כל מערכת מבוססת בינה מלאכותית: על המערכת (1) לפעול בכפוף לזכויות האדם האוניברסליות; (2) לפעול בהתאם להקשר הפוליטי של המידע העומד לרשותה; (3) להיות "איתנה" מבחינה טכנולוגית (robust).[912] ארגון התקינה מציין שכדי לעמוד בעקרונות אלו של עיצוב מבוסס אתיקה וערכים יש לציית לשמונה עקרונות פעולה, שהחמישי בהם הוא שקיפות. אומנם שקיפות היא כלי למימושה של אחריותיות, אבל היא מוגדרת בנפרד ממנה כי לשקיפות הקשר רחב יותר.[913]

גם במסמך המסכם של ועדת המומחים מטעם מועצת אירופה, שפורסם באפריל 2019,[914] לשקיפות יש מקום מרכזי. אך במסמך זה ההצדקה לקיומה של השקיפות היא בעיקר הצורך בהסברתיות. המסמך מונה שלוש דרישות עיקריות ממערכת מבוססת בינה מלאכותית בכל שלבי פעולתה: (1) הדרישה לחוקיות, כלומר התאמה למערכת החקיקה והרגולציה הקיימת; (2) הדרישה שהמערכת תפעל באופן אתי; (3) והדרישה שהמערכת תהיה איתנה הן ברמה הטכנולוגית הן בבחינת השפעותיה החברתיות.[915]

לעומת זאת, לדברי דיגנום שקיפות היא חלק ממכלול שהיא מכנה "art of AI" - אמצעי אתי לעיצוב ההחלטות של מערכות בינה מלאכותית (ethics in design). אף שהיא מבחינה בין אחריותיות, הסברתיות ושקיפות, היא משתמשת בביטוי art כמכלול ורואה בשלושת הערכים מקשה אחת. דיגנום סבורה שיישומם בעת עיצוב המוצר הוא כלי הכרחי להבטחת עמידתו בחובות אתיות ובערכי החברה.[916] גם מורלי ועמיתיה רואים בשקיפות אמצעי למועילות, לאי־הסבת נזק, לשמירה על האוטונומיה האנושית, לצדק ולהסברתיות.[917]

912 IEEE, **לעיל** ה"ש 321, בעמ' 10.

913 **שם**, בעמ' 29.

914 AI HELG 2019, **לעיל** ה"ש 298.

915 **שם**, בעמ' 9.

916 ראו Dignum, **לעיל** ה"ש 883, בעמ' 52.

917 Jessica Morley et al., *Ethics as a Service: A Pragmatic Operationalisation of AI Ethics*, 31 (3) Minds and Machines 1 (2021)

בספרות אפשר למצוא גם מקרי ביניים שבהם השקיפות היא העיקרון הראשי, אך אין הוא עצמאי אלא קשור לעקרונות אחרים. במסמך של המכון הישראלי למדיניות טכנולוגיה על השימוש באלגוריתמים בתחום הרווחה, שקיפות והסברתיות מופיעות יחד כעקרונות ראשיים, מתוך התפיסה ששקיפות היא כלי לבקרה ולהגברת האמון וגם כלי הכרחי לקיומו של הליך הוגן וצודק בתוכנות בינה מלאכותית.[918] אותו ערבוב של אחריותיות, הסברתיות ושקיפות ניכר גם במסמך ועדת המשנה של המיזם הלאומי למערכות נבונות. אומנם בתחילת המסמך השקיפות מוגדרת כערך נפרד והוועדה אומרת במפורש ששקיפות כוללת את החובה להנגיש מידע – הן באשר לתהליך יצירת המערכת הן באשר לדרך שבה היא מקבלת החלטות. הוועדה אף מסבירה כי חוסר השקיפות המאפיין את הבינה המלאכותית תורם מאוד לחוסר האמון במכונה ומגדירה את הנושא עניין שיש למצוא לו פתרון. ואולם בהמשך המסמך השקיפות יונקת את הנמקותיה מהצורך ביצירת אחריותיות. לדברי הוועדה, כדי לייצר אחריותיות יש צורך בשקיפות, בהסברתיות, באחריות ובניהול סיכונים.[919] חוסר הבהירות של מושג השקיפות נוגע אפוא גם לתפיסות השונות בעניין מקומה הגאומטרי של החובה עצמה, אם כערך עצמאי ואם כנגזרת של ערכים אחרים.

7.2.2. איזון בין שקיפות אלגוריתמית לאינטרסים מתחרים

אפשר ללמוד על הפרשנויות למושג שקיפות אלגוריתמית גם מן הגישות השונות לאיזון הנחוץ בין שקיפות לאינטרסים מתחרים, ובפרט לאיזון בין שקיפות ליעילות וחדשנות ובין שקיפות להסברתיות. כל אלה מחזקים את הטענה בדבר הצורך בעיגון פרשנות ברורה ומקובלת למושג.

במסמך של המכון הישראלי למדיניות טכנולוגיה נכתב כי בהינתן חוסר הבהירות המובנה של טכנולוגיות בינה מלאכותית, אם תוכנה איננה מסוגלת לספק הסברים

918 סיון תמיר **בינה מלאכותית בשירותי ממשל: הטמעת מערכות לקבלת החלטות מבוססות־אלגוריתם בשירותי הרווחה** 16 (המכון הישראלי למדיניות טכנולוגיה 2020).

919 המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 4 ו־13. נציין שהוועדה לא הגדירה שקיפות כפי שנעשה במסמכי מדיניות אחרים.

יש להגביל את שימושה במקום שהחלטותיה משפיעות על זכויות של בני אדם.[920] לעומת זאת, במסמך מדיניות תכנון תוכנות מבוססות בינה מלאכותית שהתפרסם באתר חברת איי־בי־אם אין הכרה בהגבלת שימושים, אף שיש מערכות שמובהר שאי־אפשר לייצר לגביהן שקיפות.[921]

תקינת ה־IEEE מקנה לשקיפות מעמד עליון. ועדת המומחים של הנציבות האירופית, לעומת זאת, מבטאת מודעות לקושי הטכני של מערכות בינה מלאכותית לספק הסברתיות ההולמת בני אדם. לפיכך היא מציינת שיש לאזן בין הצורך להגביר את ההסברתיות של המערכת ובין הפגיעה בדיוק של המערכת בעקבות הגברתה.[922] על פי תפיסתה של דיגנום, הרציונל של השקיפות הוא היותה כלי הכרחי לתקשורת בין השחקנים המעורבים ביצירת המערכת. היא מתמקדת באחריותיות ובאחריות, ולכן שלא כמו ועדת המומחים, הקובעת שיש לאזן בין ביצועים להסברתיות, היא טוענת שבמקומות שבהם ההסברתיות נפגעת יש להעדיף אינטרסים של שקיפות.[923]

7.3 הבעייתיות בהגדרת מושג השקיפות האלגוריתמית כפי שהיא מתבטאת בהצעות החקיקה של האיחוד האירופי ושל ארגון IEEE

העכירות במושג השקיפות האלגוריתמית בולטת בשני מקרים שבהם נעשה ניסיון לפרוט את המושג לחובות חוקיות.

ההצעה של תקנות הבינה המלאכותית האירופיות, שתוארה בחלק 4.1 לעיל, בחרה בגישה של הטלת מגבלות רגולטוריות בהתאם לסוגי השימושים המיועדים למערכת המבוססת על בינה מלאכותית. לכן היא מבחינה בין מערכות בסיכון גבוה, שיהיו כפופות לרגולציה מחמירה, ובין השאר גם לחובות רישוי,[924] ובין מערכות שאינן זקוקות

920 ראו תמיר, **לעיל** ה"ש 918, בעמ' 31.

921 *IBM's Principles for Trust and Transparency*, IBM (20.5.2018)

922 ראו Dignum, **לעיל** ה"ש 883, בעמ' 26 ו־33.

923 **שם**, בעמ' 53. הגרף המופיע בעמוד זה מסייע בהמחשת התפיסה המשולבת של שלושת הרכיבים האלו.

924 ראו **לעיל**, סעיף 4.1.1.2.

לרגולציה וזו תחול עליהם באופן וולונטרי בלבד.[925] הצעת החוק משקפת שקלול של ההגדרות השונות והיא נדרשת לחובת השקיפות בשלושה עניינים.

ראשית, ההצעה מציגה בפירוט רב את היקף חובת השקיפות ואת שלבי יישומה. יש לציין את זהות היצרן ופרטי קשר להשגתו; לפרט את מאפייני התוכנה ומגבלותיה (מטרות התוכנה, רמת הדיוק שלה, תוצאות צפויות, תפקוד ביחס לקבוצות המטרה, מטא־מידע בעניין סוג המידע ואמינותו); לציין אילו התאמות בוצעו כדי לתת מענה לרמת הסיכון; להעמיד לרשות מפעיל התוכנה בקרה אנושית; לציין מהי תוחלת החיים של התוכנה ואיזו תחזוקה נדרשת להפעלתה; ועוד.

שנית, היא קובעת את רמות חובת השקיפות בהתאם לסכנה הצפויה מהשימוש המיועד לתוכנה. חובת השקיפות תחול על מערכות המוגדרות מערכות בסיכון גבוה, ובהן מערכות המשפיעות על זכויות אדם, על מצב פיננסי ועוד. בעניין מערכות אלו החקיקה מציינת במפורש כי המשתמשים צריכים להיות בעלי היכולת לפרש את הפלט של המערכת. חובת שקיפות מוגברת תחול גם על מערכות העונות להגדרות האלה: יש להן ממשק עם בני אדם; הן נועדו לאתר רגשות או תחושות; הן עלולות לבצע מניפולציה על החלטות אנושיות. במקרים כאלה תחול חובה ליידע את המשתמשים בדבר היכולות של המערכת.

שלישית, ההצעה מסייגת את חובת השקיפות כשהדברים אמורים בקניין רוחני של משווקי מערכות. זו אמירה בעייתית, שכן היא מעדיפה אפריורית זכויות יוצרים משקיפות. עם זאת, לא ברור אם זו הייתה הכוונה המפורשת וייתכן שליקוי זה יתוקן בעתיד.[926]

925 ראו למשל **לעיל**, סעיף 4.1.1.4.

926 על היחס של האיחוד האירופי לזכויות קניין רוחני אגב פיתוח מערכות מבוססות בינה מלאכותית ראו בדוח שאימץ פרלמנט האיחוד: European Parliament, *Intellectual Property Rights For the Development of Artificial Intelligence Technologies*, 2020/2015(INI). בסעיף 8 הדוח מציין כי האיחוד האירופי – "stresses the importance of streaming services being transparent and responsible in their use of algorithms, so that access to cultural and creative content in various forms and different languages as well as impartial access to European works can be better guaranteed"

מסמך המדיניות של ארגון התקינה נדרש לסוגיות רבות, אך הוא מפורט פחות.[927] לפי מסמך זה, חובות השקיפות צריכות להביא בחשבון את המשתמש הסופי, את היצרן ואת הרגולטורים גם יחד. המסמך מבקש גם להשית חובת העמדת מידע על מי שיידרשו לחקור אירועים של תקלות במערכות אלו, אם מטעם המדינה ואם מטעם היצרן. המסמך דורש שבכל תוכנה יהיו לפחות שלוש רמות של שקיפות.[928]

(1) מנקודת מבטו של המשתמש – במערכת יוטמע מעין "כפתור" שכאשר ילחצו עליו תסביר התוכנה מדוע בוצעה כל פעולה.

(2) הליכי האשרור יתועדו ויועמדו לרשות מפתחים ויצרנים אחרים לשם בדיקת המוצר.

(3) ייווצרו "לוגים" ויסופקו חיישנים שישמרו באופן מאובטח את המידע על פעילות התוכנה כדי לבחון את פעולתה בעת תקלה.

המסמך קורא לקביעת מדדים כמותיים לבחינת שקיפות המוצר, אך לא מפורטות בו הדרכים לקביעתם.[929]

7.4 העקירות של השקיפות האלגוריתמית: סיכום

יש הסבורים כי מימוש השקיפות האלגוריתמית צריך להיעשות באמצעות הטלת חובות שתכליתן לאפשר למשתמש להבין את המתרחש סביבו ולשלוט בו. אחרים שמים את הדגש על התוצאה ודורשים שהמשתמש ידע להעריך את איכות השירות שניתן לו על ידי התוכנה ויוכל להשיג על החלטותיה בעת הצורך. אחדים אף מרחיבים את חובת השקיפות האלגוריתמית כך שתכלול חובות גילוי צרכניות. נוסף על כך אין הבחנה ברורה בין היקף החובות החלות בשעה שהטכנולוגיה נמצאת בשימוש במגזר הפרטי או בגופים דו־מהותיים להיקף החובות החלות כשהיא נמצאת בשימוש גוף ממשלתי, כחלק מפעולותיו.

927 ראו IEEE, **לעיל** ה"ש 321.

928 **שם**, בעמ' 28.

929 **שם**, בעמ' 27.

יש מחלוקת גם בנוגע לאיזון בין חובת השקיפות האלגוריתמית ובין חובות אחרות שיחולו על מוצרים מבוססי בינה מלאכותית. לאיזון זה יש השלכות במקרים שבהם יכולת הבקרה נפגעת והתוכנה איננה מסוגלת לספק מענה לדרישות המוזכרות במסמכי מדיניות ובהצעות חוק, למשל בנוסח ההצעה של תקנות הבינה המלאכותית האירופיות. במצב דברים זה המושג שקיפות אלגוריתמית נותר עכור, הן בהיעדר הגדרה ממצה למושג, הן בכל הנוגע להיקף תחולתו, הן ביישום המעשי של הדרישות השונות ובהתמודדות עם מקרים שבהם בגלל חסמים טכנולוגיים אי־אפשר לספק את הדרישות של כל הכותבים. לא ברור מהן בדיוק הדרישות מיצרן של מערכות אלגוריתמיות ומה מעמדן של דרישות אלו במקרים שבהם הטכנולוגיה אינה יכולה לספק אותן – אם הרגולציה תגביל את השימושים המותרים של טכנולוגיות אלו אפשר שבמקרים מסוימים היצרנים אף ייאלצו לנטוש את פיתוחן.

פרק שמיני

פיקוח מוסדי
על בינה מלאכותית

לצד האתגר הרגולטורי של תרגום עקרונות אתיים רחבים לכללים משפטיים יש שאלה נפרדת: מה זהות הגוף או הגופים המוסדיים האמורים להיות ממונים על תחום הבינה המלאכותית – הן כגופים מיישמי חקיקה, הן כגופים מאסדרים ומפקחים והן כגופים מייעצים?

8.1 ההיבט המוסדי בראייה השוואתית

בחלק זה נסקור כמה מקורות ונשווה את האופנים שבהם הם מטפלים בהיבט המוסדי של אסדרת הבינה המלאכותית. מאחר שהאסדרה של בינה מלאכותית עודנה בחיתוליה נבחרו

מקורות שכוללים הצעות לרגולציה, המלצות מדיניות והסדרי ביניים קיימים (בהיעדר רגולציה של ממש) בנושאי בינה מלאכותית. להשלמת התמונה נידרש גם להיבטים הרגולטוריים בתזכיר חוק הגנת הסייבר.

התקנות האירופיות (שתוארו לעיל בפרק 4) מעמידות לצד הכללים החלים על מערכות בינה מלאכותית בסיכון גבוה (והאיסור הקטגורי על מערכות בינה מסוכנות) מבנה מוסדי מורכב שנועד לשרת בין השאר את הצרכים הנובעים מתיאום על־לאומי בין המדינות החברות באיחוד.

הנוסח המקורי של התקנות הציע להקים ועד אירופי לבינה מלאכותית (EU AI Board),[930] בדומה לוועד האירופי להגנה על מידע (EDPB), שהוקם מכוח התקנות הכלליות בדבר הגנת מידע (GDPR).[931] הוועד האירופי לבינה מלאכותית, שיורכב מנציגים בכירים של רשויות הפיקוח הלאומיות וממפקח על הגנת הפרטיות,[932] יתרום לשיתוף הפעולה בין רשויות הפיקוח הלאומיות ובין הנציבות האירופית בעניינים שהתקנות מסדירות; יתאם את ההנחיה של הנציבות האירופית ושל רשויות הפיקוח הלאומיות בנושאים שהתקנות מסדירות; ויסייע לנציבות האירופית ולרשויות הפיקוח הלאומיות ביישום עקבי של התקנות.[933]

לפי ההצעה המתוקנת יוקם משרד אירופי לבינה מלאכותית (EU AI Office)[934] המשרד האירופי לבינה מלאכותית, שמושבו יהיה בבריסל, יהיה גוף עצמאי של האיחוד, ויורכב ממזכירות, ועד מנהל ופורום מייעץ.[935] תפקידי המשרד לבינה מלאכותית יהיו ייעוץ לגופים הרלוונטיים באיחוד ובמדינות החברות בקשר ליישום תקנות הבינה המלאכותית; ניטור יישום התקנות מבלי לפגוע בסמכותן של הרשויות המפקחות הלאומיות; תיאום בין רשויות פיקוח לאומיות וגישור במחלוקות ביניהן; תרומה לשיתוף פעולה עם גופים מקבילים של מדינות מחוץ

930 ס' 56 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

931 ראו ס' 68–76 של התקנות הכלליות בדבר הגנת מידע (GDPR).

932 ס' 57(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

933 **שם**, ס' 56(2), 58.

934 ס' 56 להצעה המתוקנת, **לעיל** ה"ש 57.

935 **שם**, ס' 56a.

לאיחוד; פיתוח מומחיות רלוונטית; בחינה של שאלות הנוגעות ליישום התקנות; ופרסום של דיווחים שנתיים על יישומן ושל המלצות על סיווג של פרקטיקות אסורות ושל מערכות בסיכון גבוה.[936]

הנציבות האירופית תנהל בשיתוף עם מדינות האיחוד מרשם של מערכות בינה מלאכותית בסיכון גבוה.[937] על הספקים של מערכות אלו תחול חובת רישום[938] והמאגר יהיה פתוח לעיון הציבור.[939] גוף נוסף שלא הוסדר, אך בכוונת הנציבות להקים בשלב הטמעתן של מערכות אלו הוא קבוצת מומחים שתסייע בניטור השוק ותייעץ לוועד האירופי לבינה מלאכותית.[940]

ברמה הלאומית, כל מדינה חברה באיחוד תקים רשות פיקוח לאומית מוסמכת שתפקידה יהיה להבטיח את יישומן והטמעתן של התקנות המוצעות,[941] ותשמש מפקחת על השוק (market surveillance authority).[942] ימונו גם מפקחים מגזריים שיהיו אחראים על מגזרים מסוימים.[943] גורמי הפיקוח החיצוניים – שיכולים להיות גופים פרטיים שקיבלו אקרדיטציה – יוסמכו לבצע בחינות תאימות של מערכות בינה מלאכותית בסיכון גבוה לפי הוראות התקנות המוצעות.

אפשר לראות אפוא שנוסח התקנות המוצעות נותן למדינות חברות מרחב תמרון מסוים בעיצוב המוסדי של רגולציית הסייבר ברמה הלאומית. לנוכח הדרישה למנות רגולטור יחיד, שירכז את הקשר עם הציבור ועם רגולטורים מקבילים

936 **שם**, ס' 56b

937 **שם**, ס' 60(1).

938 **שם**, ס' 51. לפרטים החייבים ברישום ראו התוספת השמינית לתקנות המוצעות.

939 **שם**, ס' 60(3).

940 ראו דבריה של לוסילה סיולי, מנהלת תחום בינה מלאכותית ותעשיות דיגיטליות במינהלת הנציבות למערכות תקשורת, תוכן וטכנולוגיה (DG CONNECT): *A European Approach to the Regulation of Artificial Intelligence*, **YouTube** (23.4.2021), בדקה 25:00.

941 ס' (2)–(1)59 להצעה המתוקנת, **לעיל** ה"ש 57.

942 ס' 63 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

943 **שם**, ס' 63(1).

במדינות חברות אחרות,[944] המדינות החברות אינן רשאיות לבזר את האחריות לרגולציה של בינה מלאכותית ולהעביר אותה לידי רגולטורים מגזריים. לאחרונה המליצה מועצת המדינה של צרפת (Conseil d'Etat) שרגולטור הפרטיות הלאומי (Commission nationale de l'informatique et des libertés – CNIL) יהיה מופקד על רגולציה של בינה מלאכותית.[945]

עם זאת, יש לתת את הדעת על כך שהתפקיד המיועד לרגולטור הלאומי הוא יישום והטמעה של תקנות שתכליתן להגן על עקרונות אתיים ולהביא להובלה אירופית בפיתוח בינה מלאכותית אמינה (trustworthy), בטוחה ואתית.[946] לכן ברמה הלאומית תפקידו העיקרי של הרגולטור הוא לפקח ולא לפתח.

נראה שהצעת חוק הבינה המלאכותית והמידע של קנדה כוללת הסדר דומה. לפי הצעה זו יש לאפשר לשר הממונה למנות את אחד מבכירי משרדו לנציב לענייני בינה מלאכותית ומידע כדי לסייע לשר באכיפת החוק.[947] השר יהיה רשאי להאציל מסמכויותיו הסטטוטוריות לנציב לענייני בינה מלאכותית ומידע,[948] פרט לסמכות לתקן תקנות.

בשנת 2021 פרסמה נציבות זכויות האדם של אוסטרליה דוח שעוסק בזכויות אדם וטכנולוגיות מתעוררות, ובהן בינה מלאכותית.[949] בבסיס הדוח עמדה גישה שנותנת עדיפות לזכויות אדם הן ברמת האסטרטגיה הלאומית[950] הן ברמת הרגולציה.[951] הדוח מציע למנות נציב לנושא בטיחות בבינה מלאכותית, שיסייע לרגולטורים, לקובעי המדיניות ולגורמים בממשלה ובמגזר העסקי לעצב את

944 **שם**, ס' 77 למבוא.

945 *S'engager dans l'intelligence artificielle pour un meilleur service public*, Conseil d'Etat (30.8.2022)

946 פס' 5 למבוא להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

947 ס' 33 להצעת חוק הבינה המלאכותית והמידע של קנדה, **לעיל** ה"ש 55.

948 ראו **לעיל** בסעיף 4.4.2.

949 Australian Human Rights Commission, Human Rights and Technology – Final Report 28 (2021)

950 **שם**, בעמ' 24-23.

951 **שם**, בעמ' 48-47.

המסגרת הנורמטיבית שתחול על קבלת החלטות מבוססת בינה מלאכותית. הנציב יפעל לקדם זכויות אדם ולהגן עליהן, בפרט זכויות של קבוצות בסיכון.[952] הדוח מדגיש כי אין צורך ברגולטור ייעודי לבינה מלאכותית, מאחר שלשיטתו אין להתמקד באסדרת טכנולוגיות אלא באסדרה של השימוש בהן בהקשרים ספציפיים – דבר שרגולטורים מגזריים יוכלו לעשות ביעילות בתמיכת הנציב לנושא בטיחות בבינה מלאכותית.[953]

הדוח ממליץ לממשלת אוסטרליה להקים נציבות עצמאית לבטיחות בבינה מלאכותית, אשר (1) תעבוד עם רגולטורים למען פיתוח יכולות טכניות בתחום הבינה המלאכותית; (2) תנטר ותתחקר התפתחויות בשימוש במערכות נבונות, בייחוד בתחומים שבהם יש חשש לפגיעה בזכויות אדם; (3) תשמש עבור מעצבי המדיניות באוסטרליה גורם מומחה ועצמאי לבינה מלאכותית ולזכויות אדם; ו־(4) תפרסם קווים מנחים שינחו את הממשלה ואת המגזר הפרטי כיצד להשתמש בבינה מלאכותית כחוק, ועל פי סטנדרטים אתיים.[954] כדי לנטר את השימוש במערכות נבונות מוצע שלנציבות הבטיחות בבינה מלאכותית יוקנו סמכויות לערוך חקירות וביקורות, בדומה לסמכויות הנתונות לרשויות האמונות על תחומי הגנת הפרטיות.[955] לדעת הנציבות לזכויות אדם, סמכויות אלו יסייעו בהתמודדות עם אתגרי השקיפות האלגוריתמית.

לראיין קיילו הייתה הצעה דומה. במאמר משנת 2014 הוא הציע להקים גוף ממשלתי מתמחה שלא ישמש כרגולטור ייעודי.[956] לנגד עיניו עמדה רשות פדרלית לרובוטיקה שתייעץ לסוכנויות אחרות בממשל בנוגע לרובוטיקה ולבינה מלאכותית. למשל, היא תייעץ לרשות לניירות ערך בעניין מסחר אלגוריתמי (אלגו־טריידינג),[957] לרשות התעופה בעניין כטב"מים (אוטונומיים

952 **שם**, בעמ' 127.

953 **שם**, בעמ' 130.

954 **שם**, המלצה מס' 22, בעמ' 128.

955 **שם**, בעמ' 132.

956 Ryan Calo, *The Case for a Federal Robotics Commission*, Brookings (15.9.2014)

957 ראו **לעיל** בסעיף 2.4.

ולא אוטונומיים),[958] ולמשרד התחבורה בעניין רכבים אוטונומיים. קיילו הציע שהרשות הפדרלית לרובוטיקה תייעץ למחוקקים בכל רמות הממשל בענייני חקיקה ומדיניות הנוגעים לרובוטיקה ולבינה מלאכותית; תדון עם בעלי עניין מקומיים ובינלאומיים מתחומי התעשייה, הממשל והאקדמיה על השפעות אפשרויות של טכנולוגיות אלו על החברה; ותשמש ידידת בית המשפט (בדומה לסוכנויות פדרליות אחרות) בעניינים הנוגעים לתחומי פעילותה.[959]

קיילו הדגיש שהרשות שהוא מציע לא תעסוק באכיפה – הוא סבר שבנקודת הזמן שבה נכתבו הדברים החלת משטר אכיפה כללי על טכנולוגיות בינה מלאכותית תקדים את המאוחר. מנגד, סבר שעליה לפתח מומחיות שתסייע לרגולטורים אחרים, וכן למחוקקים ולבתי המשפט, להימנע מטעויות.[960]

באירופה נפוץ הסדר ביניים שאמור לשמש עד לכינונו של גוף רגולטורי: יחידת מצפה בינה מלאכותית (AI Observatory), גוף ממשלתי שתכליתו לאפשר הטמעה אחראית, אמינה ואתית של מערכות נבונות ולתמוך בה.[961] ברפובליקה הצ׳כית, למשל, מצפה הבינה המלאכותית הוא גוף מומחים שנועד לזהות מכשולים משפטיים אפשריים בחקר מערכות נבונות, בפיתוח שלהן ובשימוש בהן, ולהמליץ על התמודדות עימם; לפרסם המלצות משפטיות ואתיות לתעשייה; ולשתף פעולה בשיח הבינלאומי על רגולציה ואתיקה של בינה מלאכותית.[962] גרמניה הקימה יחידת מצפה בינה מלאכותית בשוק העבודה ובחברה שתכליתה לבחון את השפעתן של טכנולוגיות אלו על החברה והעבודה ולנסח המלצות לעיצוב מערכות אלו באופן ששם את האדם במרכז ומשרת את טובת הכלל.[963]

958 להרחבה ראו אורי וולובלסקי "שימושים אזרחיים בכטב"מים – אתגר חדש לזכות לפרטיות" **משפט ועסקים** יט 993 (2016).

959 Calo, **לעיל** ה"ש 956, בעמ' 12.

960 **שם**, בעמ' 13.

961 Raquel Jorge Ricart et al., AI Watch, National Strategies on Artificial Intelligence: A European Perspective 15 (Publications Office of the European Union, 2021)

962 ראו באתר AI Observatory and Forum.

963 Artificial Intelligence Strategy 26 (The Federal Government 2018); Artificial Intelligence Strategy of the German Federal Government (The Federal Government 2020)

המשרד לבינה מלאכותית בבריטניה הוא יחידה ממשלתית שמשותפת למשרד לאסטרטגיה בענייני עסקים, אנרגיה ותעשייה ולמשרד הדיגיטל, המדיה, התרבות והספורט. מטרת המשרד היא לעודד פיתוח טכנולוגיות בינה מלאכותית ושימוש בהן. המשרד אמון על ניסוח קווים מנחים בנוגע לבינה מלאכותית עבור גורמים ממשלתיים[964] ועל עידוד מחקר בנושאי רגולציה ומדיניות.[965]

הצעת חוק הבינה המלאכותית של ברזיל נמנעת מהקמה של גוף ייעודי לאסדרה של בינה מלאכותית וקוראת לאסדר תחום זה באמצעות הרגולטור המגזרי,[966] במתווה של ניהול סיכונים. לפי הקווים המנחים בהצעה, כל רמות הממשל בברזיל יעודדו הקמה של מנגנוני ממשל בשיתוף נציגים של הציבור, של החברה האזרחית, של המגזר העסקי ושל הקהילה המדעית.[967] כל אלה יעודדו אימוץ מכשירים רגולטוריים שמקדמים חדשנות, כגון ארגזי חול רגולטוריים, הערכת סיכונים רגולטורית ורגולציה עצמית מגזרית.[968]

מערך הסייבר הלאומי הוקם בשנת 2017 בעקבות איחוד מטה הסייבר הלאומי והרשות הלאומית להגנת הסייבר.[969] מערך הסייבר ממונה על הגנת מרחב הסייבר האזרחי, על מתן שירותים לניהול מתקפות סייבר ועל מתן הדרכה לכל החברות האזרחיות ולגופים העוסקים בתשתיות קריטיות בהתבסס על האמונה שלכל אחד בישראל הזכות להשתמש בטכנולוגיה ללא חשש.[970] בשנת 2018 הופץ תזכיר

964 כך למשל פורסמו קווים מנחים לרכש ציבורי של בינה מלאכותית (Guidelines for AI Procurement [Office For Artificial Intelligence 8.6.2021]); לניהול פרויקטים של בינה מלאכותית; להערכה ולהטמעה של בינה מלאכותית; וכן לאתיקה של מערכות אוטומטיות לקבלת החלטות (Ethics, Transparency and Accountability Framework for Automated Decision-Making [Office For Artificial Intelligence 13.5.2021])

965 ראו למשל *Exploring Legal Mechanisms For Data Stewardship*, Ada Lovlace Institute (4.3.2021); AI Roadmap (UK AI Council 6.1.2021)

966 ס' 6(2) להצעת חוק הבינה המלאכותית הברזילאית, **לעיל** ה"ש 54.

967 **שם**, ס' 7(VIII).

968 **שם**, ס' 7(VII).

969 החלטת ממשלה מס' 3270: איחוד יחידות מערך הסייבר הלאומי (17.12.2017).

970 **שם**; Dimitry (Dima) Adamsky, *The Israeli Odyssey toward its National Cyber Security Strategy*, 40 (2) The Washington Quarterly 113, 120 (2017)

חוק הגנת הסייבר ומערך הסייבר הלאומי, שנועד לעגן בחקיקה את המסגרת הרגולטורית לפעילותו של המערך.[971] בעת כתיבת שורות אלו תזכיר החוק תלוי ועומד. אומנם נראה שמערך הסייבר מתנהל בהתאם למסגרת הרגולטורית המוצעת בתזכיר, אך עיקר סמכויותיו הן הנחיה וייעוץ לרגולטורים המגזריים ולחברות המעוניינות בכך וסיוע בעת מתקפת סייבר. לעת עתה אין לו סמכויות אכיפה.[972]

לצד ההיבטים הרגולטוריים מונה תזכיר החוק עוד תפקידים של מערך הסייבר: הניהול המבצעי של מאמצי ההגנה הלאומיים נגד תקיפות סייבר, קידום שיתופי פעולה בינלאומיים בתחום, ייעוץ לממשלה וועדותיה וקידום מדיניות ומובילות ישראלית בתחום הסייבר.[973]

מודל הרגולציה המוצע בתזכיר חוק הגנת הסייבר מחיל על הגופים המאוסדרים רמות שונות של התערבות רגולטורית בהתאם לרמת הסיכון שלהם.[974] ארגונים ששייכים למגזר משקי שמוגדר בתוספת השלישית לחוק יהיו נתונים לפיקוח ולהנחיה ישירים של מערך הסייבר.[975] במגזרים שבהם יש רגולטור מגזרי, הוא שיקבע את ההוראות להגנת הסייבר – בהנחיית מערך הסייבר,[976] המשמש "רגולטור של רגולטורים".[977] ברירת המחדל היא האצלת סמכויות לרגולטורים קיימים – אך רגולטורים אלו יידרשו לפעול באופן אחיד ובהתאם לתורת ההגנה בסייבר.[978] ארגונים בסיכון נמוך שאינם מאוסדרים ברשות מאסדרת אחרת ינהיגו רגולציה עצמית ומערך הסייבר ישמש בהם גורם מנחה ומייעץ בלבד.[979]

971 ראו תזכיר חוק הגנת הסייבר, **לעיל** ה"ש 461.

972 ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 461, בעמ' 120.

973 ס' 14 לתזכיר חוק הגנת הסייבר, **לעיל** ה"ש 461.

974 מערך הסייבר הלאומי, משרד ראש הממשלה, **הערכת השפעות רגולציה: פרק האסדרה בחוק הסייבר**, עמ' 2–3, 8, 12 (יוני 2018).

975 ס' 57 לתזכיר חוק הגנת הסייבר, **לעיל** ה"ש 461.

976 ס' 44 לתזכיר חוק הגנת הסייבר, **לעיל** ה"ש 461.

977 ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 461, בעמ' 162–165, 137–141.

978 מערך הסייבר הלאומי, **תורת ההגנה בסייבר 2.0** (19.7.2021).

979 מערך הסייבר הלאומי, **הערכת השפעות רגולציה**, **לעיל** ה"ש 974, בעמ' 25–26.

8.2 ההתייחסות בדוחות ישראליים לרגולטור של בינה מלאכותית

בשנת 2018 הקימו פרופ׳ יצחק בן ישראל ופרופ׳ אביתר מתניה את הוועדה לקידום תעשיית הבינה המלאכותית בישראל והוועדה גיבשה תוכנית שכותרתה ״חזון העצמת חוסנה של ישראל כמעצמה מדעית טכנולוגית בראיית הביטחון הלאומי והבטחת שגשוגה הכלכלי״.[980] בין השאר הומלץ להקים מינהלת לאומית למערכות נבונות במשרד ראש הממשלה.[981] התוכנית גרסה שיש צורך בגוף כזה ב״מספר הרב של הרשויות שיש להן נגיעה להייטק המקומי [...] וכמובן [ב]גופים הבטחוניים״.[982] עוד הוצע בתוכנית שהמינהלת הלאומית תייעץ לממשלה באשר למדיניות מתכללת בנושא. בן ישראל ומתניה הם אנשי מערכת הביטחון לשעבר והם שעמדו מאחורי הקמת מערך הסייבר הלאומי, גוף ביטחוני דמוי שב״כ. על פי תפיסתם בעניין הצורך בתכלול תחום הבינה המלאכותית, יש ״לקבע את המשימה כבעלת חשיבות בטחונית ראשונה במעלה״.[983] אומנם יש צורך בהגנה מפני גורמים עוינים המבקשים לנצל לרעה מהפכות טכנולוגיות, אבל אין פירוש הדבר שמערכת הביטחון או גופי ביטחון צריכים לחלוש על אסטרטגיית הטכנולוגיה הישראלית. ואכן, הדוח של בן ישראל ומתניה עורר התנגדות רבתי בקרב גופים אזרחיים, בהם המועצה להשכלה גבוהה, רשות החדשנות ומשרד האוצר. הללו יצאו חוצץ נגד הנחות היסוד של הוועדה, תוכן עבודתה והיעדר השקיפות בפעולותיה.[984]

980 אורי ברקוביץ וטל שחף "נהפוך את ישראל לאחת ממדינות הבינה המלאכותית המובילות" **גלובס** (11.8.2018).

981 "המיזם הלאומי למערכות נבונות בטוחות להעצמת הביטחון הלאומי והחוסן המדעי־טכנולוגי: אסטרטגיה לאומית לישראל, דו"ח מיוחד לראש הממשלה" א 23 (יצחק בן ישראל, אביתר מתניה וליאת פרידמן עורכים 2020) (להלן: המיזם הלאומי למערכות נבונות (2020)).

982 אורי ברקוביץ "השקעה של 2 מיליארד ש' בשנה בעיר חכמה, בחקלאות ובאקדמיה: כך מתכננת ישראל להפוך למעצמת בינה מלאכותית" **גלובס** (18.11.2019).

983 **שם**.

984 אמיתי זיו "נחיצות מוטלת בספק: גופי הממשלה נגד תוכנית הבינה המלאכותית" **TheMarker** (21.11.2019).

דוח ועדת המשנה בנושא אתיקה ורגולציה של מערכות נבונות מכיל התייחסות רחבה יותר להיבט המוסדי.[985] הדוח מציין שלנוכח קצב ההתפתחות הטכנולוגית בתחום זה, רצוי להגדיר מראש את ההליכים לבחינה חוזרת של המדיניות הרגולטורית לצורכי עדכונה.[986] כמו כן, הדוח מציע שכדי להתמודד עם חוסר הוודאות באשר להשפעה של מערכות נבונות על היבטים אתיים, יאפשרו רשויות האסדרה ניסויים מבוקרים של מדיניות רגולטורית ושימוש בארגזי חול רגולטוריים.[987] עוד מציין הדוח שאסדרת בינה מלאכותית אינה צריכה להיעשות בחלל הריק.[988] רגולטורים מגזריים מגוונים יכולים להיות מעורבים באסדרת מערכות נבונות בתחומם; למגזרים שונים יש צרכים שונים, שמהם יכולים להיגזר כללים שונים. אומנם עקרונות אתיים ברמת הפשטה גבוהה חלים במידה שווה על מערכות נבונות ממינים שונים – לצרכים רפואיים, מכוניות אוטונומיות, מערכות לסיוע בקבלת החלטות מינהליות ומערכות זיהוי פנים – אבל הקונקרטיזציה של אותם עקרונות אתיים וניסוחם בכללים רגולטוריים ישימים עשויה להשתנות מתחום לתחום.

לצד הכרה בצורך להתאים את הכללים המשפטיים החלים על מערכות נבונות למגזרים שונים מתוך שיג ושיח עם הרגולטור המגזרי בתחומן, דוח ועדת המשנה מזכיר גם כמה רגולטורים רוחביים שיכולים למלא תפקיד חשוב באסדרה: הרשות להגנת הפרטיות, רשות התחרות, ובעקיפין גם מערך הסייבר הלאומי. הדוח ממליץ שעיקר האסדרה תיעשה באמצעות הרגולטורים המגזריים; הוא אף ממליץ (בלי לאפיין אותו) על גורם רגולטורי מומחה ועל־ממשלתי שיתאם בין הרגולטורים המגזריים לרגולטורים הרוחביים.

הדוח של ועדת בינה מלאכותית ומדע הנתונים, שמינה פורום תל"ם (פורום תשתיות לאומיות למחקר ולפיתוח) בתחילת 2020, זיהה את תחום הבינה המלאכותית כתחום אסטרטגי וציין כמה היבטים שמומלץ לרכז בהם את המאמץ

985 המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313. לגרסתו הסופית של הדוח ראו המיזם הלאומי למערכות נבונות (2020) ב 172, **לעיל** ה"ש 981.

986 המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 32.

987 **שם**, שם.

988 **שם**, בעמ' 33–34.

הלאומי.[989] הוועדה המליצה על תוכנית חומש לאומית לבינה מלאכותית, אבל התמקדה בעיקר בצורך לקדם את מדע הנתונים.[990]

הוועדה ציינה כי "נושא הנתונים והנגישות אליהם מהווה מכשול משמעותי [...] נכון יהיה להגדיר רגולטור ממשלתי אחד כדוגמת הלמ"ס כבעל אחריות למיפוי, אגירה, זמינות, רישוי ונגישות לנתונים". ואולם בהמשך הדוח ציינה הוועדה כי הדיון במבנה הניהולי המוצע לתוכנית הלאומית טרם הושלם, וחזרה על הצורך ברגולטור יחיד שעיקר תפקידו **לאחד בעלות על הנתונים** מתוך "שאיפה להסתכל על נתונים כעל 'טובין ציבוריים'".[991] עם העקרונות הכלליים שהוועדה מונה במסמך נמנית גם רגולציה, אך לא מוצע בו שום מבנה מוסדי מלבד ההמלצות על ארגזי חול רגולטוריים ועל הקמת מרכזי נתונים ופלטפורמות לשיתוף נתונים ומודלים.[992] במסמך מאוחר יותר הציגה הוועדה מבנה ניהולי: הרגולטור, המורכב מנציגי משרד המשפטים ומשרד החוץ, יהיה גורם חיצוני לתוכנית הבינה המלאכותית הלאומית.[993]

בשנת 2022 פורסם דוח בנושא רגולציה של בינה מלאכותית במגזר הפיננסי, שהוכן לבקשת המחלקה הכלכלית של מחלקת ייעוץ וחקיקה במשרד המשפטים.[994] הדוח מבסס את המלצותיו על שיקולים הנוגעים לצורך בחיזוק מגזר ההיי־טק והחדשנות ולצורך בעידוד השימוש בתוצרי מגזר זה בישראל גופא (ולא רק בשוקי היעד הזרים שלו), וכן על אילוצים הנוגעים ליכולתה של ישראל להשפיע על התקינה הגלובלית החלה על מערכות בינה מלאכותית.[995] הדוח ממליץ לרכז בידי גורם ממשלתי אחד את תיאום סוגיית האסדרה של יישומי בינה

989 ועדת בינה מלאכותית ומדע הנתונים, **דו"ח סופי 2020**, בעמ' 78.

990 **שם**, בעמ' 80.

991 **שם**, בעמ' 81.

992 **שם**, בעמ' 62–64. התייחסויות כלליות לצורך בפיתוח רגולציה יש גם בדוח צוות המשנה של המיזם הלאומי למערכות נבונות בנושא ממשל. ראו המיזם הלאומי למערכות נבונות (2020) ב, **לעיל** ה"ש 981, בעמ' 222, 225.

993 דוח ועדת תל"ם, **לעיל** ה"ש 103, **מצגת יוני 2021**, בשקף 20.

994 ראו אחיעז ואח', **לעיל** ה"ש 240.

995 **שם**, בעמ' 146.

מלאכותית – הוא מציין שרצוי שתהיה בו מעורבות של משרד המשפטים,[996] אך קורא להימנע בשלב זה מקביעת כללים שיחולו באופן אחיד על כל יישומי הבינה המלאכותית (שיכולים להיות שונים מאוד אלו מאלו – למשל יישום לפענוח תצלומי רנטגן לעומת אלגוריתמים לטרגוט פרסומות).[997]

גישה דומה אומצה בדוח שפרסם בסוף 2022 משרד החדשנות, המדע והטכנולוגיה. הדוח מציע להקים "מוקד ידע ותיאום ממשלתי להסדרת בינה מלאכותית". גורם זה יעסוק ביישום מדיניות רגולציה ואתיקה ובגיבוש המלצות לעדכונה, יסייע למשרדי הממשלה ולרגולטורים בגיבוש מדיניות רגולציה וינגיש לציבור ולממשלה מידע וכלים לשימוש אחראי בבינה מלאכותית.[998] על פי תפיסת האסדרה המוצעת בדוח, הנמנע מחקיקה או מאסדרה רוחבית אחרת, יש לעודד כלים של אסדרה "רכה" וניסויית ולהעדיף רגולציה מגזרית. הדוח מדגיש את חשיבותו של גוף תיאום בין־משרדי להבטחת מדיניות אחידה ברורה וקוהרנטית.[999]

996 **שם**, בעמ' 148.

997 **שם**, בעמ' 147.

998 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 110.

999 **שם**, בעמ' 11. ראו גם המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313.

פרק תשיעי

יסודות לרגולציה מכוונת זכויות של בינה מלאכותית בישראל: עקרונות כלליים

9.1 מבוא

השיח הער בשנים האחרונות מכיר באתגרים שמערכות בינה מלאכותית מעמידות לפני קובעי המדיניות. יש הסבורים כי שאלות משפטיות רבות הנוגעות לשימושים במערכות מבוססות בינה מלאכותית מקבלות מענה עקרוני בדין הקיים ולא נדרשת אלא התאמה קלה של הקצאת הסיכונים, הן על ידי תיקוני חקיקה הן על ידי התפתחות שיפוטית.[1000] אחרים סבורים כי אפשר לפתור את בעיית האחריות הנזיקית של בינה מלאכותית על ידי יצירת אישיות

1000 ראו למשל Chesterman, **לעיל** ה"ש 19, בעמ' 38.

משפטית שמתאימה לה, בהנחה שלאישיות זו יהיו מקורות פיננסיים עצמאיים שיאפשרו לה לשלם עבור הנזקים שגרמה.[1001]

ואולם אסדרה של מערכות בינה מלאכותית היא הכרחית – הן ברמה הגלובלית הן ברמה הלאומית. סדרת סיפורי הרובוטים של אייזק אסימוב מראה במידה מסוימת מדוע לא די בשלושת חוקי הרובוטיקה, ואחת היא עד כמה הם אלגנטיים ושובי לב.[1002] והרי אילו היו כללים אלו אפקטיביים לרגולציה של מערכות בינה מלאכותית, סדרת ספרי הרובוטים של אסימוב לא הייתה באה אל העולם.

אימוץ וולונטרי של מסגרת כללים אתיים רחבים ובלתי מחייבים בתעשייה או ברשויות הממשלה עלול לשמש עלה תאנה, שכן הוא יבטא הכרה מוסדית בקיומה של בעיה אך יהיה משולל גיבוי של מנגנוני פיקוח ואחריותיות משפטיים שיאפשרו לממש אותם. אסדרה שתיעשה נכון, לעומת זאת, תאפשר פישוט וייעול של תהליכי פיתוח, והבניה של תהליכים כלליים של בקרת איכות, ותקדם אפוא אמון ציבורי במערכות ובמוצרים. רגולציה שתכלול כללים שיבנו את אופן הפיתוח של בינה מלאכותית ואת התיעוד וממשל הנתונים בה תוכל למתן הטיות אלגוריתמיות מסוגים שונים – הן הטיות אלגוריתמיות שהן תולדה של כשלי פיתוח לא רצויים, המנוגדים לאינטרסים של מפתחיהן, הן הטיות אלגוריתמיות שמחצינות סיכונים שאינם מטרידים את המפתחים. יודגש: רגולציה מספקת ודאות גם לתעשייה – בוודאי יותר מהתפתחות שיפוטית, שאורכת מטבעה שנים רבות ומתאפיינת באקראיות מבחינת המקרים המגיעים לפתחם של בתי המשפט.

מערכות בינה מלאכותית כבר נמצאות בשימוש בקרב מגזרים רבים בישראל, מהן פרי פיתוח מקומי ומהן פיתוח של חברות זרות. יש להבטיח כי מערכות אלו פועלות מתוך שמירה על האינטרסים ועל זכויות היסוד של המשתמשים בהן או של אנשים אחרים המושפעים מפעילותן.[1003]

1001 Karolina Ziemianin, *Civil Legal Personality of Artificial Intelligence: Future or Utopia?* 10 Internet Pol'y Rev. 1 (2021); Chesterman, **לעיל** ה"ש 19, בעמ' 126.

1002 ראו אסימוב, **לעיל** ה"ש 47, וכן Murphy and Woods, **לעיל** ה"ש 365.

1003 נציבות זכויות האדם האוסטרלית, למשל, מדגישה שאסטרטגיה לאומית בנושאים טכנולוגיים צריכה לכלול במטרותיה העיקריות תעדוף רגולטורי של הגנה על זכויות אדם. ראו Australian Human Rights Commission, **לעיל** ה"ש 949, בעמ' 28.

זאת ועוד: כפי שניסינו להראות עד כה, מערכות מבוססות בינה מלאכותית מעוררות שאלות החורגות מתחום המשפט הפרטי. אפשר למנות שלושה טעמים לכך.

ראשית, כמה מן האתגרים הם אתגרים ציבוריים למחצה, כגון הטיות, שקיפות והגנת הפרטיות. לכן גם כשהסיכונים הנשקפים מבינה מלאכותית אינם נוגעים לזכויות אדם, אלא צבר של סיכונים כלליים שמקורם בשגיאות בפיתוח התוכנה, ברשלנות או בהזנחה,[1004] היכולת להתחקות עליהם תלויה במידת השקיפות של מערכת הבינה המלאכותית הנדונה.

שנית, בגלל הפגיעה הפוטנציאלית של מערכות אלו בשלמות הגוף ובקניין הפרטי יש להשית רגולציה ממשלתית בעוד מועד. עיצוב הרגולציה הזאת ברמה המהותית והמוסדית הוא פעילות בתחום הציבורי.

שלישית, כבר היום מופעלות מערכות לומדות ורשויות המדינה ימשיכו להרחיב את פעילותן באופן שעלול להשפיע על זכויותיהם של אזרחים, ובכלל זה על הזכות לכבוד ולשוויון.

פרק זה והפרק שלאחריו יניחו יסודות לרגולציה מכוונת זכויות של בינה מלאכותית בישראל. בפרק זה נציג את העקרונות המנחים שצריכים, לתפיסתנו, לעמוד ביסוד המדיניות בתחום זה.

9.2 עקרונות מנחים ליצירת מדיניות בינה מלאכותית מוכוונת זכויות

9.2.1. האדם במרכז

על פי עמדת דוח ועדת תל"ם ודוח המיזם הלאומי למערכות נבונות, על הרגולציה לשאוף להיות

1004 Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. Rev. 2, 9-17 (forthcoming 2023)

מאפשרת[1005] ולהשתדל לצמצם את הנטל המוטל על כתפי השחקנים בזירת הבינה המלאכותית – משתמשים, ספקים, מפתחים או מפיצים במגזר הציבורי והפרטי גם יחד.[1006] רוח דומה עולה מהדוח העוסק ברגולציה של בינה מלאכותית במגזר הפיננסי, המדגיש את הצורך במניעת חסמים לחדשנות ובהקלות רגולטוריות על פיתוח בינה מלאכותית בישראל, וקורא בשלב זה לרגולציה עצמית של השוק ולשימוש בארגזי חול.[1007] גם מסמך המדיניות של הרשות לחדשנות נוטה להעדיף רגולציה מאפשרת, המעודדת פיתוח בינה מלאכותית ושימוש בה.[1008] אך מתווי המדיניות אינם צריכים לפעול אך ורק מתוך רצון לאפשר חדשנות וצמיחה בבואם לעצב הסדר רגולטורי ומוסדי: אין להזניח בשם עקרון מזעור ההתערבות הרגולטורית את ההגנה על חירויות וזכויות יסוד.

התכלית המרכזית של פיתוח בינה מלאכותית ומערכות לומדות צריכה להיות שירות המין האנושי – הן כקולקטיב הן כפרטים – באופן המיטיב עימו. עיקרון זה של חירות האדם והאוטונומיה שלו מצטרף אל עקרון קידום הטוב ועקרון מניעת הנזק ומבסס ביתר שאת את החשיבות של עקרון האדם במרכז, של החירות ושל האוטונומיה. באופן מעשי מתבטא עיקרון זה בפיתוח מערכות נבונות, פרישׂתן ושימוש בהן באופן המגן על זכויות היסוד ועל חירות האזרח ורואה בהן עיקרון ראשון במעלה. לא מדובר במס שפתיים שנועד להסתיר נהייה אחר קידום חדשנות או אחר אינטרסים כלכליים (כגון עידוד מגזר ההייטק), או אפילו אחר ניסיון ייעול של תהליכים במגזר הציבורי.

בהתאם לכך מטרתה של רגולציה של בינה מלאכותית היא להגן על המושאים של החלטות אלגוריתמיות – בני אדם. הגנה זו יכולה לבוא לידי ביטוי ביצירת מערכת כללים המאפשרת ודאות בנוגע להקצאת הסיכונים הנזיקיים, אך גם בליווי תהליכי פיתוח מערכות אלגוריתמיות, הטמעה שלהן ושימוש בהן באופן

1005 דוח ועדת תל"ים, **לעיל** ה"ש 103, בעמ' 63; המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 28; המיזם הלאומי למערכות נבונות (2020) א, **לעיל** ה"ש 981, בעמ' 30–31.

1006 ראו לדוגמה את התייחסות הצעת תקנות הבינה המלאכותית האירופיות (**לעיל** ה"ש 53) לשחקנים האלה. ראו גם **לעיל** בסעיף 4.1.1.2.

1007 אחיעז ואח', **לעיל** ה"ש 240, בעמ' 148.

1008 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 97.

שממזער את הסיכון לזכויות אדם[1009] ומאפשר בקרה אקס אנטה ואקס פוסט על החלטות אלגוריתמיות. הגנה על מושאי החלטות אלגוריתמיות פירושה גם פיתוח ועיגון של זכויות הנוגעות לשקיפות ושל סעדים בגין פגיעה בהן. למשל, יש ליידע משתמשים במערכות בינה מלאכותית שהם אינם מתקשרים עם אדם בשר ודם אלא עם צ׳אט בוט.

כמו כן, בלי לקבוע מסמרות בעניין השאלה עד כמה רצוי להסתמך הסתמכות עיוורת על מערכות החלטה אוטומטיות בכל תחומי החיים, עקרון האדם במרכז משמעו גם הגדרת יישומים של בינה מלאכותית ומערכות אלגוריתמיות תומכות החלטה שמחייבים מעורבות או פיקוח אנושיים,[1010] בין שמדובר באישור אנושי להחלטה אלגוריתמית ובין שמדובר באפשרות לערער עליה לאחר מעשה לאדם ממשי. זאת בייחוד במקומות שבהם השפעתם על חיי הפרט יכולה להיות משמעותית.

עניין זה חשוב במיוחד בנוגע ליישומים ממשלתיים של בינה מלאכותית, בין שהם פרי פיתוח פנימי ובין שהם תוצאה של מיקור חוץ, שכן החלטות אוטומטיות של יישומים ממשלתיים משפיעות על אזרחים. רגולציה של בינה מלאכותית יכולה לעודד בעקיפין פיתוח מערכות בטוחות, המשמרות זכויות אדם וחירויות אזרח, באמצעות התניית רכש ומכרזי פיתוח ממשלתיים של מערכות נבונות בעמידה בסטנדרטים מחמירים של שקיפות, בטיחות, תיעוד ובקרה.[1011]

1009 ראו **לעיל** בפרק 5.

1010 ראו לדוגמה ס׳ 6.3.9 לדירקטיבת ה־ADM של קנדה, **לעיל** ה"ש 628; ס׳ 22 של התקנות הכלליות בדבר הגנת מידע (GDPR).

1011 לרגולציה באמצעות מכרזים ראו Christopher McCrudden, *Using Public Procurement to Achieve Social Outcomes* 28 Nat. Resources Forum 257 (2004). לשימוש במכרזים ציבוריים כדי לקדם הגנה על הפרטיות ראו לדוגמה Laurens Vandercruysse et al., Public Procurement of Smart City Services – Matching Competition and Data Protection (A Report in the Framework of the SPECTRE Research Project 2020). ראו גם אלדד הבר וטל ז׳ירסקי "דרכי ההגנה על תשתיות חיוניות במרחב הסייבר בישראל" **משפט וממשל** יח 99, 136 (2017).

9.2.2. הדמוקרטיה במרכז

בתחילת המאה ה־21 גאה וגעש ה"טכנו־אופטימיזם" וניעורה תקווה שהעולם הדיגיטלי יביא איתו בשורה לדמוקרטיה ולפוליטיקה, ושסגולותיו של עולם זה – פתיחות, חופש ומידע – יתמכו במימוש הערכים הדמוקרטיים, ישלימו את המוסדות הדמוקרטיים ויקדמו אותם.

ואולם כיום, בתחילת העשור השלישי של המאה ה־21, ברי לנו שבטכנולוגיה טמונים גם סיכונים, והיא גובה מאיתנו מחיר, בדרכים שעוד לא כולן ידועות לנו עד תום. השפעתה מגיעה למעגלים רחבים, הנוגעים ליכולתם של מוסדות דמוקרטיים לתפקד; לשיח הציבורי, שהוא אבן הפינה של ההליך הדמוקרטי; ולאפשרות המימוש של בחירה חופשית ואוטונומית.

במערכות מבוססות בינה מלאכותית טמון סיכון ממשי של פגיעה בדמוקרטיה במובנה הרחב – אם באמצעות השפעה על השיח הציבורי והפצת רעיונות, אם באמצעות מכשור לשליטה באוכלוסייה, למעקב אחריה, לזיהויה ולמשטורה, ואם באמצעות היכולת שלהן לזרוע ספק ולערער על עצם היכולת לברר מהי המציאות ולהבחין בין מקור לזיוף ובין אמת לשקר. לכן יש לתת משקל של ממש לעיקרון "הדמוקרטיה במרכז" – גם, לעיתים, במחיר של התקדמות טכנולוגיות או חדשנות.

אפשר להשתמש בבינה מלאכותית כדי להשפיע לרעה על מהפכת המידע וכדי לקדם את מהפכת המעקב. למשל, יצירת טקסטים וקטעי וידיאו מזויפים באמצעות טכנולוגיות דיפ־פייק; מערכות לזיהוי קול ולזיהוי פנים המקדמות מעקב ומשטור; וגם מערכות לניתוח מידע שתכליתן "הנדסת בחירות". ואולם יש גם סוגים אחרים של מערכות בינה מלאכותית שטמונות בהן סכנות לדמוקרטיה.

מערכות חיזוי שיוכלו להעניק פשר למציאות עתידית ולהשפיע עליה ישנו את השיח הפוליטי ואת האתוס של חזון פוליטי, שתכליתו החברתית היא לנסות להעריך תרחישים ולתפקד במציאות לא ברורה. היתרון של פוליטיקה המבוססת על מערכות חכמות גדול, אך יש לשאול בכנות מה יקרה ביום שבו תהליך ממוכן, מבוסס מידע, של קבלת החלטות יוכל לספק תחזיות ולפתור בעיות עוד בטרם הפכו לבעיות; ומה תהיה המשמעות הדמוקרטית של מצב שבו נבחרי הציבור יהיו רק כתובת אחרונה לבעיות שמכונות אינן יכולות לפתור.

גם אם כיום השימוש במכונות מכוון רק לסייע להחלטות אנושיות, עם התגברות היכולת לקבל תחזיות טובות ומדויקות יותר יוכלו מכונות גם לשפר ביצועים מערכתיים, לאסדר התנהגות ולפתור סכסוכים משפטיים בין רגולטורים לאזרחים. אזרחים יוכלו לפנות לרשות רגולטורית, להניח את נתוני המקרה המיוחד שלהם בידיה ולקבל פרשנות וייעוץ מדויקים שייתרו את הצורך בבתי משפט. מערכות לומדות יעסקו במתן רישיונות; יקבעו קנסות ויערכו שימועים לקבלת זכויות; יחקרו דיווחי מס; יחליטו את מי לעכב לבדיקה מיוחדת בשדה התעופה; יחזו פשיעה ופשיעה חוזרת; ישרתו מערכות אכיפת חוק ויעניקו למקבלי החלטות ניתוח רב־ממדי המסתמך על מידע רחב ומאוזן בין שכבות שונות של תובנות. הכול בדרך מדויקת יותר, יעילה יותר, מהירה יותר וזולה יותר מבני אדם או ממערכות מחשב מתקדמות פחות. בשלב מתקדם יותר יוכלו מכונות אף להציע לנו סדר חוקי טוב מהסדר הקיים ולומר לנו כיצד לצמצם את הטעויות בו – כלומר יוכלו לחוקק עבורנו את החוקים.

החלפת שיקול הדעת האנושי בשיקול הדעת של אלגוריתם מביאה עימה אתגרים הנוגעים לאחריות לתהליכי קבלת החלטות בדמוקרטיה. אחד מיסודות מערך האיזונים והבלמים הוא האפשרות לעקוב אחר שיקולי קבלת ההחלטות ולבקר אותם. איך יוכלו שופטים, לא כל שכן אזרחים מהשורה, למתוח ביקורת על החלטות של אלגוריתמים? לכן יצירת אמות מידה לקבלת החלטות, וקידום שקיפות, הסברתיות ואחריותיות באשר להחלטות של מערכות לומדות, הם האתגרים המהותיים העומדים היום לפני דמוקרטיות. הם שיבטיחו בעשור הקרוב הגנה על הליך הוגן ועל זכויות חשודים וימנעו הטיות אפשריות בהחלטותיהן של מערכות המבוססות על מאגרי מידע.

חוסר נכונותם של פוליטיקאים לוותר על שימוש בכלים הפוגעים בדמוקרטיה אם הם יכולים לסייע להם להיבחר או לשרת אותם בדרכים אחרות הוא בעיה של ממש. יתר על כן, מתברר שאת ההחלטות בנוגע למציאות הדיגיטלית לא מקבלות ממשלות נבחרות אלא קומץ חברות ענק מכוונות רווח. למרות הפרדוקס שבדבר החשש האמיתי אינו מפני חברות מסחריות דווקא, אלא משאיפתן של מדינות לנצל את הטכנולוגיות כדי להשיג שליטה חסרת תקדים באוכלוסייה, בזמן בחירות, וגם זו בזו. כוחן העצום של הפלטפורמות הדיגיטליות, שהן לובי משמעותי נגד אסדרה, והחשש להציע פתרונות שיפגעו ביכולתן של המדינות

הליברליות להתמודד עם המדינות הלא־חופשיות, דוגמת סין – גם הם מעמידים אתגרים גדולים.

הדמוקרטיות נמצאות היום בצומת דרכים, ובשנים הקרובות יידרשו מדינות וממשלות להחליט החלטות מכריעות כדי להתמודד עם הכוחות המערערים אותן. יהיה צורך למפות את כיווני ההתפתחות הטכנולוגיים ולחשוב לאילו מהם עלולות להיות השלכות שליליות ארוכות טווח על הדמוקרטיה ואילו מהם ישפיעו לטובה על הטוב המשותף. סטואו בויד, עתידן אמריקאי, כינה זאת "הצורך במהפכת האביב האנושי".[1012]

9.2.3. נדרשת אוריינות דיגיטלית מצד מקבלי ההחלטות

אחת הבעיות הקשות בתפר שבין משפט לטכנולוגיה היא הרמה הדיגיטלית הנמוכה של מקבלי ההחלטות, שרובם אינם "ילידים דיגיטליים". הם אינם מבינים את עומק הבעיות, אינם שואלים את השאלות הנכונות ואינם נדרשים לתמונת הכוחות המלאה. אוריינות דיגיטלית היא סל היכולות והמיומנויות – הטכנולוגיות, הסוציולוגיות והקוגניטיביות – הנדרשות כדי להבין את הסביבה הדיגיטלית ולתפקד בתוכה. היא כוללת יכולת לחפש ולמצוא מידע, וגם מודעות לכוחות העומדים מאחורי המהפכה הדיגיטלית ולהטיות שהם יוצרים, הקפדה על כללי התנהגות וכישורי חשיבה ביקורתית.

אוריינות דיגיטלית משמעה היכולת להבין לאן מתפתחת הטכנולוגיה, לפחות בטווח הקצר: לנתח את השוק ולבחון היכן נמצאים כספי המחקר והפיתוח של חברות הענק ומהם הפטנטים שהן רושמות כדי לעגן פיתוחים טכנולוגיים חדשים. בעל אוריינות דיגיטלית מבין שאפשר להטות ולכוון פיתוח טכנולוגי הן באמצעים מסחריים הן באמצעים רגולטוריים; הוא יודע גם שמעצבי מדיניות הם חלק מהתפתחות הטכנולוגיה ולא רק מתבוננים מן הצד, ולכן הם נושאים באחריות. בהקשר של מדיניות בינה מלאכותית, נתעכב על שלושה עניינים.

1012 תהילה שוורץ אלטשולר "דמוקרטיה וטכנולוגיה" **דמוקרטיה עכשיו: סוגיות ואתגרים במאה ה־21** (דנה בלאנדר ומאיה גייר עורכות, 2021).

ראשית, לא כל מהפכה טכנולוגית מתרחשת בסמיכות היסטורית גדולה כל כך למהפכות טכנולוגיות אחרות. למעשה, מהפכת הבינה המלאכותית יכולה להפיק לקחים משתי המהפכות האחרות בדור האחרון – מהפכת החומרה ומהפכת הקישוריות. נמנה כמה לקחים אפשריים:

- פוליטיקאים גוררים רגליים גם כשהצורך לאסדר כבר ברור; ומנגד, התעשייה טוענת כי לפוליטיקאים אין את היכולת וההבנה הנדרשים כדי לאסדר ומוטב לפיכך להימנע לגמרי מאסדרה;

- תמת ה"חדשנות" מתערבבת לעיתים בפזמון החוזר "נשבור עכשיו, נתנצל אחר כך", או בניסוח חריף יותר "לא ידענו עד כמה הכלי שפיתחנו מסוכן. לא ידענו לאילו נזקים הוא עלול לגרום. אנחנו מצטערים. סליחה";[1013]

- טענת התעשייה כי נדרשת אסדרה מתערבבת בניסיונות שלה עצמה להשיג הסדרים רגולטוריים מקילים, בהשקעות עתק בשדלנות ובהשפעה על מקבלי החלטות, ואף באיום להפסיק מתן שירותים באזורים גאוגרפיים שבהם מדובר על אסדרה משמעותית;[1014]

- כל מהפכה נתפסת כתהליך משחרר, שמעודד דמוקרטיה ושוויון, אבל בהיעדר אסדרה החזק לוקח הכול וכמה חברות ענק שולטות בכל העולם. כשמדובר במהפכת הבינה המלאכותית, שכוח מחשוב וכמויות גדולות של נתונים זמינים הם תנאים מוקדמים עבורה לצורך אימון מודלים, החשש מפני שליטה אפריורית של ענקיות הטכנולוגיה הוא מוחשי;[1015]

- הסיקור התקשורתי הדיכוטומי של טכנולוגיה, הנע בין אופוריה להיסטריה, בין אוטופיה לדיסטופיה, משחית את יכולתם של מקבלי החלטות הניזונים מהתקשורת לדון בנושאים האלה באופן שקול ומאוזן.

1013 יובל דרור "הבאים בתור לבקש סליחה" **מגזין דה מרקר** (דצמבר 2022).

1014 ראו למשל את קריאותיו של סאם אלטמן, מנכ"ל חברת Open AI, לאסדרה – בד בבד עם איומו להפסיק לתת גישה לשירותי החברה שלו באיחוד האירופי. Shiona McCallum and Chris Vallance, *ChatGPT-maker U-turns on threat to leave EU over AI law*, BBC (26.5.2023)

1015 AI Now, *Confronting Tech Power* (2023)

שנית, נדרש רובד ביניים בין הבנת הטכנולוגיה ובין יצירת מדיניות בעניינה. מייקל היידן, שעמד בעבר בראש הסוכנות לביטחון לאומי (NSA) והסי־איי־איי, כתב על מרחב הסייבר: ״יש מעט תופעות חברתיות חשובות ומדוברות כל כך שמבינים אותן מעט כל כך״.[1016] היידן הדגיש שהוא אינו מתכוון לטכנולוגיה ספציפית או להפעלת כלי כזה או אחר, אלא להיעדר מסגרת תפיסתית שמאפשרת להבין את ההשלכות של השימוש בטכנולוגיות שונות, ולאו של סייבר דווקא. במסגרת זו הוא מתכוון להבנת המשמעויות של מערכות טכנולוגיות, ליכולת לדמיין את האפשרויות החדשות שהן מביאות איתן וליכולת להבין את השלכותיהן על המוסר החברתי ועל שלד השיטה המשפטית. כיוון שמסגרת זו חסרה לעיתים קרובות נוצרים פערי הבנה, בייחוד בנושאים בעלי השלכות רחבות כמו הבינה המלאכותית.

9.2.4. נדרשת המשגה חדשה של מערכות ויכולות בתחום הבינה המלאכותית

המושג בינה מלאכותית הוא מטפורה פוליטית־טכנולוגית, אולי המטפורה העוצמתית ביותר בתקופה הנוכחית. בדומה למטפורות אחרות, היא חודרת לזיכרון, מעוררת רגשות, משפיעה על עמדות ומעצבת ציפיות לגבי העתיד. המושג הוא מטפורה פוליטית כיוון שהוא מייחס תכונות אנושיות למערכות טכנולוגיות. כל דבר נעשה ״חכם״ – מטרמוסטט ועד רחפן. השימוש במטפורה של אינטליגנציה, בינה, חוכמה בהקשרים רבים כל כך מלמד על עוצמתה. ההשוואה בין מערכת בינה מלאכותית למוח האנושי יוצרת קרבה ודמיון, והללו מובילים להטמעה חברתית של הרעיון שהמכונות עובדות כמו מוח אנושי, מבצעות פעולות אנושיות בדרך שבה מבצעים אותן אנשים, ולמעשה מתחרות בבני האנוש.

למי יש אינטרס לשמר את המטפורה של הבינה המלאכותית ולהאניש את המערכות הטכנולוגיות? אפשר לחשוב, למשל, על יצרנים של מערכות שאינם רוצים לקבל אחריות למעשים של המערכות האלה, להטיות שלהן ולדעות

1016 Michael Hayden, *The Future of Things "Cyber"*, 5 Strategic Studies Quarterly 3 (2011)

הקדומות המוטמעות במאגרי המידע שהן מתאמנות עליהם, לפגיעות ולנזקים שהן עתידות לגרום, לחוסר ההבנה המובנה שלהן בנוגע להשלכות החברתיות של החלטותיהן ולקושי לייצר בשבילן שקיפות ותקינה. כמה נוח לומר "הבינה המלאכותית טעתה" ולהרחיק את האחריות לטעויות ממי שיצר אותה או נתן לה את המשימה.

אפשר לחשוב גם על מי שרוצה שנראה במכשירי קצה טכנולוגיים כמו טלפונים סלולריים או עוזרים דיגיטליים בני משפחה, נמסור להם מידע פרטי וניתן אמון בהחלטות שלהם. אחרי הכול, אלו הן החלטות שמתקבלות בדיוק כמו ההחלטות שלנו. אפשר כמובן לחשוב על מי שרוצה שנאמין שהמערכות הטכנולוגיות מתפתחות באופן עצמוני, ממש כמו אבולוציה אנושית, ולא נשים לב לעובדה שענקיות הטכנולוגיה נעשו שומרות סף גם בהקשרים של פיתוח מערכות לומדות.

יש להשתדל שהמטפורה הפוליטית של בינה מלאכותית לא תעצב את הדמיון שלנו באשר לעתיד ושהנרטיב העוצמתי שבו משתמשות התעשייה והתקשורת המסקרת אותה לא ישפיע על מקבלי ההחלטות. מוטב למשל להשתמש במטפורה של "למידה", שאינה קשורה בהכרח ל"הבנה", ולהידרש לטכניקות פעולה ספציפיות כמו למידת מכונה, מערכות גנרטיביות, מערכות מבוססות ידע או סטטיסטיקה. ולחלופין, להידרש להקשר היישומי של כל טכנולוגיה: למשל, תמיכה באבחון רפואי באמצעות למידת מכונה; זיהוי ביומטרי במרחבים ציבוריים לתכליות אכיפת חוק; וכדומה. הצורך בהמשגה מחודשת של תהליכי הבנה, היסק, יצירתיות, חשיבה, דמיון וכיוצא באלה, כאשר הם מתבצעים על ידי מכונות, מתעורר גם הוא לאור השיפור ביכולות התקשורת בין המכונה למשתמשים.

9.2.5. נדרש פיתוח זכויות למושאי החלטות של בינה מלאכותית

המהפכה הטכנולוגית מעוררת צורך בחשיבה מחודשת על זכויות יסוד משתי בחינות.

ראשית, יש צורך לצקת משמעות חדשה לתאוריה החוקתית של זכויות האדם בצורתה היום. למשל, הזכות לחופש ביטוי אינה יכולה להיות מושתתת רק על התפיסה שמלחמה

בביטוי שקרי נעשית באמצעות עוד ביטוי[1017] או על התפיסה שבהתקיים מבחר מספק בשוק הרעיונות והדעות תצוף האמת מאליה. בעידן שמתאפיין בהצפה של רעיונות שמנוּוטים על ידי מערכות אלגוריתמיות יש צורך בחשיבה מחודשת על התאוריה המכוננת של הזכות לחופש ביטוי. כך טענו למשל דיוויד ביבר וג׳ייסון סטנלי, שקבעו שחופש הביטוי במובנו כיום מסכן את הדמוקרטיה לא פחות משהוא תורם לה;[1018] וטים וו, שקבע שהשימוש הנוכחי בתאוריה של חופש הביטוי גורם לדיכוי היכולת להתבטא.[1019]

בדומה, הזכות לפרטיות בעידן של מעקב המונים איננה יכולה להיות אך ורק בעלת אופי אינדיווידואלי, כלומר זכותו של הפרט לשלוט במידע על אודותיו, אלא עליה לקבל פרשנות קולקטיבית הנוגעת לחוסר היכולת לנהל הליך דמוקרטי תקין בשעה שלפרטים בחברה אין אוטונומיה לבחור בחירה חופשית כיוון שהם נתונים להשפעה מניפולטיבית המבוססת על מעקב ושליטה.[1020]

דוגמה נוספת היא הבנה מחודשת של הזכות לכבוד האדם ביחס למכונה. ס׳ 22 לתקנות הכלליות בדבר הגנת מידע (GDPR) קובע כי ״למושא מידע תהיה הזכות שלא להיות מושא של החלטה המבוססת אך ורק על תהליכים אוטומטיים, לרבות פרופיילינג, שיש לה תוצאות משפטיות בנוגע אליו או לאחרים, או תוצאות משמעותיות אחרות״. זוהי הזכות למעורבות אנושית, הנתפסת כנגזרת של הזכות לכבוד האדם. אומנם היקף פרישׂתה של הזכות האירופית מוגבל. היא לא תאפשר, למשל, לערער בזכות על החלטות שיפוטיות אנושיות שהתקבלו

1017 עמדה שהיטיב לבטא שופט בית המשפט העליון של ארצות הברית לואיס ברנדייס: "The remedy to be applied to [falsehood and fallacies] is more speech, not enforced silence" (Whitney v. California 1927)

1018 David Beaver and Jason Stanley, *Neutrality*, 49 Philosophical Topics 165 (2021)

1019 Tim Wu, *Is the First Amendment Obsolete?* The Knight First Amendment Institute (1.9.2017); להרחבה ראו תהילה שוורץ אלטשולר, אסף וינר ואייל זילברמן **מתווה לאסדרת רשתות חברתיות בישראל** 27–26 (הצעה לסדר 51, המכון הישראלי לדמוקרטיה 2023).

1020 להרחבה ראו תהילה שוורץ אלטשולר **פרטיות: מלכת זכויות האדם בעולם הדיגיטלי** (פרלמנט 83, אתר המכון הישראלי לדמוקרטיה 2019).

מתוך הסתייעות במערכות תומכות החלטה שיפוטית – למרות חששות להטיית המודלים[1021] ולמרות הנטייה האנושית להנסגת הדעת לנוכח קבלת החלטות אוטומטית.[1022]

החשש הוא אפוא מפני שחיקת הסובייקט האנושי בתור שכזה, והנכחת תחושות של היעדר נראות, חוסר ערך ותסכול לנוכח מציאות שרירותית. מנגד, מחקרים מלמדים על פער ב"הוגנות הנתפסת", כלומר בתחושת הכבוד האישי, ההוגנות והצדק הפרוצדורלי של בעלי דין ביחס למכונות.[1023] בניסויים הראשונים הנבדקים נוטים לסמוך על שופטים אנושיים ולראות בהם מי שמממשים את זכותם ליחס מכבד. אבל כאשר יוצרים עבור הנבדקים תהליך שבו ניתנת להם האפשרות לדבר למכונה ולהישמע בה, וגם מעניקים לבעלי הדין זכות לקבל הסבר לוגי בעניין ההחלטה, אז הפער בהוגנות הנתפסת נעלם ולבעלי הדין לא חשוב אם מדובר בשופט בשר ודם או ברובוט שופט.

אפשר היה להניח ששימוע מול מכונה ייחשב חסר משמעות או חלול, כי לרוב אנחנו מקשרים את העובדה ששמעו אותנו או ראו אותנו עם העובדה שמי שעשה זאת הוא מישהו שיש לו יכולת לגלות אמפתיה. אבל ככל שיכולת המכונות תשתכלל – מבחינת יכולתן לזהות שפה, תנועות גוף, הבעות פנים ולכן גם רגשות, ולהגיב בהתאם – כך ילך החשש מפניהן ויקטן והן יוכלו לשמש מקור

1021 ראו **לעיל** בפרק 6.

1022 ראו Saar Alon-Barkat and Madalina Busuioc, *Human-AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice*, 33 J. Pub. Admin. Res. & Theory 153 (2022); Stavros Zouridis, Marlies van Eck, and Mark Bovens, *Automated Discretion*, *in* Discretion and the Quest For Controlled Freedom 313 (Tony Evans and Peter Hupe eds., 2020); Matthew M. Young, Justin B. Bullock, and Jesse D. Lecy, *Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration*, 2 Perspectives on Public Management and Governance 301 (2019)

1023 Benjamin Minhao Chen, Alexander Stremitzer, and Kevin Tobia, *Having Your Day in Robot Court*, 36 (1) Harvard Journal of Law and Technology 128 (2022); idem, *Would Humans Trust an A.I. Judge? More Easily Than You Think*, Slate (28.2.2023)

לתחושה שהן מסוגלות להעניק צדק פרוצדורלי. ייתכן שלא ירחק היום ואנשים יבקשו להישפט על ידי מכונות ולא על ידי שופט אנושי, שעלול להיות עצבני לפני ארוחת הצוהריים. לכן ייתכן שהגדרת הזכות לכבוד תשתנה. תופעות אלו לא תהיינה ייחודיות לתחום המשפט – האבחנה הרפואית יכולה לשמש לנו דוגמה דומה. ייתכן שמטופלים ידרשו שימוש במערכת לומדת – לא רק משום שהיא תיראה להם מדויקת יותר, אלא גם כי היא תהיה מסוגלת לספק להם קשר אישי ומכבד. בזכות השפה הטבעית של כלים מבוססי בינה מלאכותית הולך ונעשה נעים ונוח יותר לתקשר עימם, וככל שייעשו נעימים ונוחים יותר כך תגבר הנכונות לממש את הזכות לכבוד דווקא באמצעות תקשורת עם מכונות. כאמור, שינוי זה יחייב חשיבה מחודשת על הזכות לכבוד ומשמעותה.

שנית, יש ליצור זכויות דיגיטליות חדשות, שלא היה צורך בהן בעבר,[1024] ובעיקר זכויות של פרטים בשעה שהם באים במגע עם מערכות ממוכנות מבוססות אלגוריתמים. למשל: יש צורך להעצים את מושאי ההחלטות ולהעניק להם זכות לשקיפות אלגוריתמית במובן של הסברתיות;[1025] ולפתח זכות לאינטראקציה מיודעת עם בינה מלאכותית,[1026] כלומר הזכות שמוצרים מבוססי בינה מלאכותית (צ'ט בוטים למשל) יציגו את עצמם בתור כאלה ולא יתחזו לאנשים בשר ודם.[1027] צד נוסף של אינטראקציה מיודעת עם בינה מלאכותית נוגע לדימויים מלאכותיים שנוצרו על ידי בינה מלאכותית יוצרת.[1028] יש מקום לשקול חקיקה שתייצר חובת יידוע במקרה של דימויים כאלו.

1024 על זכויות דיגיטליות ראו Dafna Dror-Shpoliansky and Yuval Shany, *It's the End of the (Offline) World as We Know It: From Human Rights to Digital Human Rights – A Proposed Typology*, 32 Eur. J. Int. Law 1249 (2021)

1025 יש פרשנים שסבורים שהשילוב של ס' 13–15 וס' 22 ב־GDPR מקנים למושאי המידע זכות להסבר על האלגוריתמים המעבדים את המידע שלהם. ראו לדוגמה Andrew D. Selbst and Julia Powles, *Meaningful Information and the Right to Explanation*, 4 Int'l Data Privacy Law 233 (2017)

1026 ראו **לעיל** ה"ש 333.

1027 ראו לדוגמה ס' 52(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53; ס' 5(V)(a) להצעת חוק הבינה המלאכותית הברזילאית, **לעיל** ה"ש 54.

1028 השוו גם לס' 52(3) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53; ס' 17 לכללי הסינתזה העמוקה, **לעיל** ה"ש 607.

9.2.6. נדרשת רגולציה "חסינת עתיד"

יש לזכור שתהליכי חקיקה מפגרים תמיד אחרי המציאות. הטכנולוגיה מתקדמת בקצב מהיר, ובמדינות כמו ישראל, שבהן תהליכי החקיקה הם איטיים מאוד, נוצר פער גדול במיוחד. למרבה המזל, בינה מלאכותית איננה הטכנולוגיה המתפתחת הראשונה שהציבה את האתגרים הללו, ואפשר לשאוב השראה מההתמודדות המשפטית עם טכנולוגיות קודמות כגון טכנולוגיות מידע, טכנולוגיות גנטיות וטכנולוגיות רבייה, וללמוד מהטעויות שנעשו בתחומים אלו.

ראשית, אין לאסדר טכנולוגיה ספציפית, משום שזהו מתכון בטוח להתיישנות הרגולציה, אלא להשתדל לנסח עקרונות מנחים והגדרות כלליות שיהיו גמישים די הצורך לאכיפה עתידית. כך, חוק הבוטים בקליפורניה[1029] ובישראל,[1030] למשל, עורר ביקורת; ההגדרה של האזנת סתר בחוק האזנת סתר,[1031] לעומת זאת, איננה תלוית טכנולוגיה ולכן היא חסינת עתיד. התקנות לאסדרת הבינה המלאכותית באירופה עוצבו כך שיהיו חסינות־עתיד ולכן הן שומרות על ניטרליות טכנולוגית.

כך למשל, בתקנות האירופיות לא כוללת ההגדרה של מערכות הבינה המלאכותית זיקה לטכנולוגיה מסוימת.[1032] גם אנו סבורים כי הגדרות כלליות ומבנים רגולטוריים עקרוניים עדיפים. לחלופין אפשר לאמץ גישה תלוית מגזר (ולא תלוית טכנולוגיה), שכן גישה כזאת יכולה לעקוף את הצורך בהגדרה (למשל, מכשור רפואי).

שנית, יהיה נכון לאמץ תפיסה רגולטורית המבוססת על ניהול סיכונים, שכן היא מאפשרת לעגן מראש קווים אדומים – כלומר לציין באילו יישומים של בינה מלאכותית טמונה סכנה גדולה כך כך עד שיש לאסור קטגורית על פיתוחם ועל הפצתם. באשר ליישומי בינה מלאכותית מותרים, אפשר לייצר מרחבים בטוחים, שרמת ההתערבות הרגולטורית בהם תוגדר בהתאם לאומדן הסכנה הנשקפת מהם.

1029 Renee Diresta, *A New Law Makes Bots Identify Themselves: That's the Problem,* Wired (24.7.2019)

1030 עומר כביר "מומחים נגד תזכיר חוק הבוטים: אווילי ולא ניתן לאכיפה" **כלכליסט** (16.6.2022).

1031 ס' 1 לחוק האזנת סתר, התשל"ט–1979, ס"ח 118.

1032 ס' 3(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

המודל של ניהול סיכונים הוא מסגרת חקיקה חסינת עתיד, שכן היא מאפשרת לחזור ולהגדיר את הקווים האדומים ואת המרחבים הבטוחים על פי התפתחויות טכנולוגיות, חברתיות ורגולטוריות. אחת הדרכים להעריך מחדש את עוצמת הסכנה הנשקפת ממערכות טכנולוגיות מסוג מסוים ואת אופן האסדרה היא ארגז החול הרגולטורי – מעין מרחב גמיש ובטוח (ליזמים ולמפתחים), שבו המאסדר מלווה פיתוח והטמעה של מערכת קונקרטית כדי ללמוד אותה ואת האופן שבו רצוי לאסדר אותה.

9.2.7. נדרשת מסגרת אסדרה מעורבת של עקרונות, זכויות וחקיקה

יש שלוש מסגרות אסדרה נפוצות של בינה מלאכותית: עקרונות, זכויות וניהול סיכונים. מסגרת מבוססת עקרונות כוללת מערך עקרונות ליבה אתיים; מסגרת מבוססת זכויות מתמקדת בהגנה על זכויות האדם ועל החירויות של מי שמושפעים מיישומי טכנולוגיות שמבוססות על בינה מלאכותית; ומסגרת מבוססת סיכונים מציבה במרכזה את הסכנות והסיכונים האפשריים וגוזרת את האסדרה מרמת הסיכון, ובכלל זה אף אוסרת קטגורית על השימוש בטכנולוגיות מסוימות או שימוש לתכליות מסוימות. זוהי מסגרת רבת־עוצמה המסתמכת על העיקרון של זהירות מונעת, עמוד התווך באתיקה רפואית: *primum non nocere* (ראשית אל תזיק).[1033] על פי עקרון הזהירות המונעת, נטל ההוכחה באשר לבטיחות של טכנולוגיה מסוימת הוא על מי שמעוניין להשתמש בה ולא על הציבור.[1034]

אחד הטיעונים בדבר הצורך במסגרת אסדרה מבוססת עקרונות וכללי אתיקה[1035] של מערכות בינה מלאכותית הוא שיש צורך למלא את החלל עד שתושג הסכמה מדינתית או בינלאומית באשר לדרכי האסדרה. כללי אתיקה יכולים לשמש גשר במצב הביניים שבין פריצה של טכנולוגיה לבין יצירת רגולציה מדינתית

1033 ראו **לעיל** ה"ש 366.

1034 Wingspread Statement on the Precautionary Principle (The Global Development Research Center, GDRC 1998). ראו גם Kaminski, **לעיל** ה"ש 1004, בעמ' 22–29.

1035 ראו **לעיל** בפרק 3 את פירוט העקרונות האתיים המרכזיים. מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 103–109.

או בינלאומית ההולמת אותו. מחויבות מוסרית של עובדים בתאגידים (למשל מהנדסים וכותבי תוכנה), ושל תאגידים כלפי לקוחות ומשתמשים, אף היא יכולה לשמש לעיתים שוט חזק לא פחות מרגולציה ממשלתית מחייבת.[1036]

אלא שמול טיעונים אלו יש גם טיעונים חזקים אחרים, שלפיהם הניסיון לקרוא להסתפקות בכללי אתיקה בלבד הוא טיוח אתי, כלומר ניסיון ליצור מראית עין של אחריות אגב התחמקות ממחויבות מעשית.[1037] מהלך כזה יכול לכלול הימנעות מכוונת מהגדרות בהירות, מוחשיות ומפורטות של העקרונות האתיים; אי־קביעת מנגנוני אכיפה ברורים שענישה בצידם במקרה שאין מצייתים לכללי האתיקה; היעדר התייחסות למגוון החוליות בשרשרת הערך של המערכות והמוצרים והתמקדות בחוליה של המהנדסים וכותבי הקוד בלבד. לטיוח אתי כמה מטרות. מטרה אחת היא חיזוק האמון של משתמשים במוצרים מבוססי בינה מלאכותית (בין שמדובר באמון בעניין בטיחות המוצרים ובין שמדובר באמון בעניין ההוגנות של הליך הייצור שלהם למשל). מטרה אחרת היא מתן חירות לתעשייה והתחמקות מרגולציה ממשלתית או בינלאומית כופה.

לכן לתפיסתנו לשיקולי אתיקה נודעת חשיבות מבחינה זו שהם מבטאים תפיסה מוסרית עקרונית באשר לפיתוח מערכות מבוססות בינה מלאכותית, אבל יש לזכור שבהיעדר רגולציה שיקולים של עשיית טוב יהיו לעולם משניים לשיקולים של עשיית רווח עבור משקיעים. יתרה מזו, אם סטנדרטים אתיים יהיו וולונטריים בלבד, חברות יוכלו לבחור אילו מהם לאמץ ואילו להזניח.[1038] לכן אנו סבורים ששלוש המסגרות אינן סותרות זו את זו ויש לשלב ביניהן. מן הראוי ליצור מסגרת היברידית שמשלבת אסדרה רכה (עקרונות אתיים) וניהול סיכונים באמצעות הוראות חוק וכללים רגולטוריים נוקשים.

1036 Elettra Bietti, *From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy*, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020)

1037 ראו .Whittaker et al, **לעיל** ה"ש 450.

1038 Jacob Turner, Robot Rules: Regulating Artificial Intelligence 210, 212–213 (2019)

9.2.8. נדרשת גמישות באשר למועד ההתערבות באמצעות אסדרה

גם אם יש הסכמה בדבר תכלית האסדרה, שאלת העיתוי היא שאלה משמעותית. במקרים מסוימים יש יתרונות ברורים לרגולציה מוקדמת – לפני שמוצרים שמבוססים על טכנולוגיה מסוימת חודרים לשוק. טכנולוגיה שנתפסת מסוכנת במיוחד, אם פיזית (כמו רכב אוטונומי) ואם מוסרית (למשל יצירת תכנים שמסיתים לטרור בצורה מלאכותית), הגיוני לאסדר מראש. גם במקרים קיצוניים פחות התערבות מוקדמת יכולה להועיל במה שנוגע לעיצוב כיווני המחקר ולהסטת משאבים להשקעה בפיתוח. גרגורי מנדל טען שהתערבות בשלב מוקדם עשויה לצמצם את ההתנגדות של בעלי עניין, מפני שבשלב זה הושקעו משאבים מעטים יחסית והעלויות השקועות נמוכות יותר, ומפני שהתעשייה והציבור משועבדים פחות לסטטוס קוו.[1039] עם זאת, ייתכנו מקרים שבהם עדיף להמתין ולהתמודד עם בעיות כשהן מתעוררות ולא לנסות לצפות אותן. האינטרנט נחשב לעיתים קרובות דוגמה לטכנולוגיה שהפיקה תועלת מגישה כזאת. בשלב מוקדם, נטען, קשה יותר לבצע תחזיות ותחזיות מדויקות לגבי עלויות ותועלות.[1040]

כשפרופיל הסיכון של טכנולוגיה מתפתחת אינו ודאי גוברת לעיתים הנטייה לנקוט זהירות מרבית, שכן ההמתנה לראיות חותכות עלולה לאפשר לסכנות להתממש ולגרום נזק, לעיתים כזה שלא ניתן לתיקון – בין שמדובר בפגיעה באוכלוסיות רחבות (חיידק מהונדס) ובין שמדובר בהשפעה על השיח הציבורי (בוט נאו־נאצי). מקצת החששות בעניין בינה מלאכותית הם מסוג זה. ניק בוסטרום הוא כנראה הקול המוכר ביותר המתריע על סכנות כאלה. לדבריו, "הגישה שלנו לסכנות קיומיות אינה יכולה להיות גישה של ניסוי וטעייה. אין הזדמנות ללמוד מטעויות".[1041] לתפיסתו, יש לפעול כדי למקסם את ההסתברות לתוצאה המוגדרת "בסדר", כלומר כל תוצאה שאינה מביאה לידי סכנה קיומית.

1039 Gregory N. Mandel, *Emerging Technology Governance*, *in* Innovative Governance Models for Emerging Technologies 62 (Gary Marchant, Kenneth Abbot, and Braden Allenby eds., 2013)

1040 Cass R. Sunstein, Laws of Fear: Beyond the Precautionary Principle 58 (2005)

1041 Nick Bostrom, *Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards*, 9 Journal of Evolution and Technology 3 (2002)

האתגר כמובן הוא לקבוע מתי יש לסכנה המשוערת בסיס מציאותי אמין. התגובות המשפטיות והרגולטוריות שלנו אינן צריכות, מן הסתם, להיות מבוססות על החזיונות הדיסטופיים הקשים ביותר של עתידנים ומדע בדיוני, אך בוודאי אינן יכולות להתעלם מהשלכות רחבות לטובת יתרונות מיידיים.

איננו תמימים. אין ספק שעצם ההחלטה מתי להתערב ובאילו מקרים היא תוצאה של פשרות בתחום המדיניות הציבורית. למשל, אם ננקוט זהירות יתר ולא נתיר למכוניות ללא נהג לעלות לכביש כיוון שמכונית אוטונומית עלולה לגרום למוות, בני אדם ימשיכו לנהוג והם עצמם יביאו לכמות עצומה של נפגעים.[1042] גם באבחון רפואי יש אפשרות שבינה מלאכותית יכולה לאבחן טוב מבני אדם, אבל אין פירוש הדבר שהיא חפה מטעויות. ואולם ניסיון העשורים האחרונים מלמד כי לתעשייה יש נטייה "לשבור דברים ואחר כך להתנצל", והכול בשם החדשנות. כפי שראינו, הטיעון בזכות החדשנות משמש פעמים רבות כיסוי לרצון להרוויח עוד ועוד, גם אם הדבר נעשה על חשבון אינטרסים ציבוריים חשובים אחרים.

9.2.9. יש להטיל איסורים עקרוניים על שימוש במערכות מסוימות, אך האיסורים צריכים להיות מצומצמים ומוגבלים

הגישה האירופית המסתמנת מבחינה בין סוגים שונים של מערכות נבונות לפי רמת הסכנה הנשקפת מהן ואוסרת אפריורית על השימוש בסוגים מסוימים של מערכות נבונות. הפרלמנט האירופי, למשל, קרא לאסור על שימוש במערכות זיהוי פנים למטרות שיטור ואכיפת חוק,[1043] והתקנות האירופיות כוללות קטגוריה של "מערכות אסורות";[1044]

1042 בדומה, גישה תוצאתנית תעדיף כלי רכב אוטונומיים, חרף הסיכונים האינהרנטיים להם, מנהגים אנושיים, משום שמעבר לכלי רכב אוטונומיים מצמצם את מעורבות הגורם האנושי, האחראי ליותר מ־90% מהתאונות. ראו לדוגמה Bryant Walker Smith, *Automated Driving and Product Liability*, 2017 Mich. St. L. Rev. 1 (2017)

1043 ראו **לעיל** ה"ש 563.

1044 הצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53; ראו **לעיל** סעיף 4.1.1.1.

בקטגוריה זו נכללות מערכות שטמונה בהן על פי התפיסה האירופית סכנה קיצונית ושאינן עולות בקנה אחד עם ערכי היסוד שלה.[1045] למשל, מערכות שמשתמשות בטכניקות תת־הכרתיות או מניפולטיביות כדי להשפיע על האנושות לרעה;[1046] מערכות שמנצלות חולשות של קבוצות שונות באוכלוסייה על בסיס גיל, יכולות פיזיות או יכולות קוגניטיביות באופן שסביר שיסב נזק פיזי או פסיכולוגי משמעותי; מערכות נבונות בשירות רשויות ציבוריות או מטעמן שנועדו לסווג או להעריך את מידת האמון שיש לתת באדם על בסיס התנהלותו החברתית או תכונותיו בדרך של דירוג חברתי – באופן שעלול להביא להחלטות לרעת יחידים או קבוצות בהקשרים שאין להם זיקה להקשרים שבהם נאסף המידע עליהם במקור או באופן שעלול להסב להם נזק לא מידתי.

כמו כן, נאסר השימוש במערכות שיטור מונחה עתיד (predictive policing), המעריכות את הסיכון שאדם יבצע עברה פלילית או מינהלית (לרבות הערכת סיכויי רצידיביזם);[1047] במערכות המייצרות או מרחיבות מאגרי נתונים לזיהוי פנים בהסתמך על איסוף גורף מרשת האינטרנט או על חומר מצולם ממערכות טלוויזיה במעגל סגור;[1048] במערכות בינה מלאכותית שתכליתן לזהות רגשות לתכליות של אכיפת חוק, ניהול מעברי גבול, במוסדות חינוכיים ובהקשרי תעסוקה;[1049] ובמערכות זיהוי ביומטריות בזמן אמת במרחבים ציבוריים.[1050]

איסור השימוש בטכנולוגיות מסוימות, שהשלכותיו טרם נלמדו, יכול להיות זמני – אם כדי לבחון את האסדרה שלהן באופן שקול (כפי שנעשה בעניין שיבוט

1045 **שם**, ס' 5(1).

1046 ס' 5(1)(a) להצעה המתוקנת, **לעיל** ה"ש 57.

1047 ס' 5(1)(da), **שם**. ראו גם סעיף 2.3.2 **לעיל**.

1048 ס' 5(1)(db), **שם**. נראה שסעיף זה נועד לאסור מאגרי זיהוי פנים כגון אלו שמפתחת חברת Clearview AI. ראו גם היל, **להלן** ה"ש 1087.

1049 ס' 5(1)(dc), **שם**. ראו גם סעיף 2.3.2 **לעיל**.

1050 ראו והשוו בין ס' 5(1)(d), **שם**, ובין הנוסח המקורי של ס' 5(1)(d) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

בני אדם),[1051] ואם כדי למנוע נזקים מיידיים משימוש לא מוסדר בטכנולוגיות אלו, עד להסדר קונצנזואלי (כפי שהציע הדַוָּח המיוחד של האומות המאוחדות לחופש הביטוי בעניין טכנולוגיות סייבר התקפי).[1052]

גישה זו אינה ייחודית לאירופה. בשנת 2021 הוצגה בקונגרס של ארצות הברית הצעת חוק פדרלית שביקשה לאסור על גורמים פדרליים לרכוש מערכות ביומטריות, להעריך אותן או להשתמש בהן (ובכלל זה במערכות זיהוי פנים) – למעט כאלו שהותרו בחוק מפורש. על פי ההצעה, למדינות בארצות הברית או לגופי ממשל מקומיים שישתמשו במערכות אלו לא תהיה זכאות לסיוע תקציבי מתוכניות פדרליות מסוימות.[1053] גם תקנות מערכות ההמלצה האלגוריתמיות הסיניות כוללות איסורים מפורשים על שימוש במערכות בינה מסוימות – כאלו שעלולות לעודד התמכרות לרשת, בייחוד של קטינים.[1054] עמדה דומה הביעה הנציבה העליונה של האו"ם לזכויות אדם, מישל בצ'לט (Bachelet), כשקראה להשהות את השימוש בטכנולוגיות בינה מלאכותית המסכנות זכויות אדם, לרבות מערכות זיהוי פנים ומערכות דירוג חברתי.[1055]

ההחלטה אם לפתח או שלא לפתח סוגים מסוימים של טכנולוגיות יכולה להתקבל על יסוד איסור רגולטורי, אך היא יכולה גם להיות החלטה רצונית של התעשייה.

1051 ראו לדוגמה גם ס' 1 לחוק איסור התערבות גנטית (שיבוט אדם ושינוי גנטי בתאי רביה), תשנ"ט-1999, ס"ח 697 (1999) 47. עם זאת, החוק נחקק בשנת 1999 לתקופה של חמש שנים לשם בחינת המדיניות בנושא – ותוקפו הוארך כארבע פעמים.

1052 United Nations Human Rights Council, 41st Session of the Human Rights Council (24 June–12 July 2019), Agenda item 3, Promotion and Protection of All Human Rights, Civil, Political, Economic, Social and Cultural Rights, Including the Right to Development; United Nations General Assembly, Surveillance and Human Rights, A/HRC/41/35. בנושא זה ראו גם Amir Cahane, *Targeting Targeted Surveilance - The UN Special Rapporteur's Report on the Private Surveillance Industry*, CSRCL Blog (31.12.2021)

1053 Facial Recognition and Biometric Technology Moratorium Act of 2021, S. 2052, 117th Congress (2021)

1054 ראו **לעיל** סעיף 4.2.

1055 Jamey Keaten And Matt O'Brien, *U.N. Urges Moratorium on Use of AI that Imperils Human Rights,* AP News (15.9.2021)

חברות הטכנולוגיה הגדולות, ובראשן אמזון, איי־בי־אם ומייקרוסופט, הודיעו למשל שיפסיקו להשקיע בחברות המפתחות מערכות זיהוי פנים לתכליות שיטור או ליצור מערכות כאלה, או שיימנעו מלספקן לסוכנויות אכיפת החוק.[1056] אבל כשהאיסור אינו מחייב אלא וולונטרי תלויה יציבותו ברצון הטוב של השחקנים בתעשייה. עמדת חברת אפל באשר לפגיעה בפרטיות לקוחותיה, למשל, התרככה עם השנים: בשנת 2016 התנגדה החברה בעיקשות לדרישת הבולשת הפדרלית (FBI) לעדכן את מערכת ההפעלה שלה כדי לאפשר גישה למכשיר אייפון מוצפן;[1057] ואילו בשנת 2021 החלה החברה בפיתוח מערכות לסריקת מאגרי תמונות מקומיים כדי לזהות תכנים פדופיליים.[1058] מערך התמריצים של חברות גדולות ומבוססות כגון אפל, מייקרוסופט ו־איי־בי־אם יכול להיות שונה מזה של חברות צעירות וקטנות, שמטה לחמן מבוסס על טכנולוגיות שנויות במחלוקת ולכן הן עלולות להעדיף "לשבור שביתה". חרף המורטוריום הוולונטרי של חברות אלו על טכנולוגיות זיהוי פנים יש אפוא חברות אחדות (כגון Clearview AI)[1059] שממשיכות בפיתוח ובשיווק שלהן.[1060]

ואולם למדיניות שאוסרת על פיתוח טכנולוגיות מסוימות, רכישה שלהן או שימוש בהן יש גם חסרונות. עקב ההעדפות התרבותיות־פוליטיות השונות של כל מדינה, כל אחת מהן מתמקדת באיסור אחר בבואה לעצב את מדיניותה: הסינים חוששים כאמור מתופעות של התמכרויות לרשת, ואילו האירופים חוששים ממערכות

1056 ראו **לעיל** ה"ש 161.

1057 עומר כביר "אפל לביהמ"ש: היענות לדרישת ה־FBI תעניק לממשל 'כוח מסוכן' " **כלכליסט** (25.2.16).

1058 עומר כביר "אפל רוצה לזהות תכנים פדופיליים ופותחת קופת שרצים" **כלכליסט** (8.8.2021).

1059 קשמיר היל "תכירו את Clearview AI, החברה שהרגה לכם את הפרטיות" **הארץ** (21.1.2020).

1060 לדוגמה, במקביל להכרזות של אפל, מייקרוסופט ואמזון (**לעיל** ה"ש 161), נמצא שכ־24 ארגוני אכיפת חוק ברחבי העולם השתמשו במוצרים של חברת זיהוי הפנים השנויה במחלוקת Clearview AI. ראו Ryan Mac, Caroline Haskins, and Antonio Pequeño IV, *Police in At Least 24 Countries Have Used Clearview AI. Find Out Which Ones Here*, Buzzfeed (25.8.2021)

זיהוי פנים בשירות אכיפת החוק.[1061] לכן איסור מקומי על פיתוח מערכת מסוימת או שימוש בה לתכלית מסוימת – גם אם אין הוא וולונטרי – אינו שולל את האפשרות שיפתחו אותה וישתמשו בה במדינה אחרת, לתכלית אחרת. יתר על כן, כשהמורטוריום הזמני נועד מלכתחילה לאפשר למעצבי המדיניות לקבל החלטה מושכלת בעניין האסדרה של טכנולוגיות מסוימות, ההנחה הסמויה היא שהעבודה המלוכלכת תיעשה במקומות אחרים. כך אפשר להפיק לקחים מהניסיון המצטבר של השימוש במערכות אלו על ידי אחרים (ועל גבם).

לכן אנו מציעים שהמדיניות בישראל תעוצב מתוך מבט החוצה, בעיקר לעבר אירופה, ולא ייקבעו בישראל איסורים שאינם קיימים במדינות אחרות. מנגד, כדי שלא להפוך את ישראל ל"חצר אחורית", כלומר כדי שאותה "עבודה מלוכלכת" לא תיעשה כאן, יש לשאוף להגביל את האיסורים בזמן. כך לרגולטור או למקבלי ההחלטות יהיה די זמן לעצב מדיניות מיטבית, ואם יתעורר הצורך יהיה אפשר להיעזר גם בארגזי חול רגולטוריים. במסגרת אותם ארגזי חול תתבצע בחינה מבוקרת של מערכות בינה מלאכותית – גם כאלה שתכליתן שנויה במחלוקת – ובתוך כך ייכתבו הנחיות לשוק הפרטי.

9.2.10. אין להשלים עם מצב שבו ישראל היא חצר אחורית

בשנים האחרונות מקודמות הצעות לאסדרת בינה מלאכותית במדינות מובילות, וגם באיחוד האירופי, ויש להניח שמדובר בתחילתה של מגמה עולמית. גם אם יש שוני ערכי בין האסדרות במקומות שונים – אם בבחירת סוג המערכות המוגדרות מסוכנות ואם במטרות המוצהרות של החקיקה עצמה – יש להן גם הרבה מן המשותף. לכן אין

1061 ייתכן שיש חפיפה בין האיסור על מערכות המנצלות חולשות של קבוצות שונות באוכלוסייה באופן שצפוי להסב נזק פיזי או פסיכולוגי (ס' 5(1)(a) להצעת תקנות הבינה המלאכותית האירופיות [**לעיל** ה"ש 53]) ובין האיסור הסיני על פיתוח מערכות המעודדות התמכרות. לביקורת על הנוסח האירופי, המעדיף עקרונות עמומים מפתרונות ייעודיים לטכנולוגיה ספציפית, ראו Paul De Hert and Georgios Bouchagiar, *Facial Recognition, Visual and Biometric Data in the US. Recent, Promising Developments to Regulate Intrusive Technologies*, 7 (29) Brussels Privacy Hub working paper 4 (2021)

להשלים עם מצב שבו במדינת ישראל לא תהיה חקיקה שתהיה בהרמוניה עם החקיקה המקובלת בעולם.

הרמוניה איננה בהכרח זהות. כאשר הפיתוח מכוון לייצוא ממילא נדרש הסטנדרט הרגולטורי לעמוד לכל הפחות בסטנדרטים הזרים, שהם לרוב גבוהים יותר מהסטנדרטים בישראל. אבל גם כאשר הפיתוח מכוון לשוק המקומי, ומושאי ההחלטות של מערכות הבינה המלאכותית אינם אזרחי חוץ אלא אזרחי המדינה, אין להנמיך את הסטנדרטים, שהרי אזרחי המדינה ראויים אף הם להגנה. אומנם סטנדרט רגולטורי נמוך מהמקובל בעולם יכול לעודד חדשנות, אך זו לא בהכרח חדשנות רצויה: ישראל עלולה להפוך לחצר אחורית טכנולוגית, כלומר למקום שבו מפתחים מערכות שאסור לפתח אותן, להפיץ אותן או להשתמש בהן לפי הדין הזר.[1062] ישי עופרן ואלעד סימן טוב הראו במטא־מחקר שנערך בשנת 2022 שניסויים שהשתמשו בנתונים פרטיים של אזרחים "זלגו" עקב הטמעת התקנות האירופיות בדבר הגנת המידע למדינות מחוץ לאירופה שלא היתה בהן חקיקת פרטיות מעודכנת.[1063] אין סיבה להניח שבאסדרה של בינה מלאכותית מצב זה יהיה שונה. אפשרות אחרת, הגובלת באבסורד אך נוכחנו בה עם הצעת "תקנות הגישור" בשנת 2022, היא מצב שבו הדין הישראלי יהיה כלפי חוץ בהרמוניה עם הדין הזר, אך בתוך גבולות המדינה או על אזרחיה שלה יחיל סטנדרטים נמוכים יותר.[1064]

1062 כך למשל, חברות ישראליות או חברות בבעלות ישראלית מובילות בפיתוח מערכות סייבר התקפיות, ששנויות במחלוקת (ב־2019 קרא הדַּוָּח המיוחד של האומות המאוחדות לחופש הביטוי להשעות זמנית את הסחר הבינלאומי במערכות אלו ואת השימוש בהן. ראו **לעיל** בה"ש 882). דוגמה אחרת היא תעשיית האופציות הבינאריות, שפעלה בישראל והציעה אופציות בינאריות למשקיעים מעבר לים עד חקיקת האיסור הפלילי על פעילות זו. ראו ס' 44יב1 לחוק ניירות ערך, תשכ"ח–1968, ס"ח 541, בעמ' 234; US v. Lee Elbaz, No. 20-4019 (4th Cir. 2022)

1063 Elad Yom-Tov and Yishai Ofran, *Implementation of Data Protection Laws in the European Union and in California Is Associated with a Move of Clinical Trials to Countries with Fewer Data Protections*, 9 Front. Med. (Lausanne) (10.11.2022)

1064 כך למשל במהלך 2022 שקלה הרשות להגנת הפרטיות לקדם חקיקה שתטיל על בעלי מאגרי מידע ישראליים חובה להגן על פרטיות של מידע שמקורו באיחוד האירופי כדי להלום את הוראות ה־GDPR בהיעדר רפורמה כוללת בדיני הגנת הפרטיות הישראליים. ראו עומר כביר, "איפה ואיפה: ישראל תספק הגנה משופרת רק לפרטיותם של תושבי האיחוד

נוסף על כך, המסגרת המשפטית הקיימת בישראל להגנה על הפרטיות ולאבטחת סייבר נוטה להחריג מתחולתה את רשויות הביטחון[1065] או לייצר הסדרים מקילים יותר בהקשרים ביטחוניים. ראינו כיצד במהלך התפרצות מגפת הקורונה השתמשו בכלי מעקב המיועדים לסיכול טרור לתכליות אזרחיות מובהקות,[1066] וכמה רב היה הפיתוי לפרוש מערכות בינה מלאכותית לצורך דירוג אזרחי ישראל על פי הסבירות שיידבקו בקורונה.[1067] הנטייה הישראלית למסגר אירועים שאין דבר בינם ובין ביטחון לאומי בתור כאלה עלולה להביא להחרגת איסורים מוחלטים על מערכות בינה מלאכותית שנועדו לתכליות מסוימות בתואנות ביטחוניות, כך שבעתיד תתאפשר זליגה של מערכות אלו מהספרה הביטחונית לזו האזרחית.

לכן על פי תפיסתנו יש לנסח את האיסורים על פיתוח מערכות נבונות ושימוש בהן כך שיהיו בהלימה עם המקובל בעולם, ובפרט עם הדין האירופי המתהווה – לבל תשמש מדינת ישראל לניסויי כלים טכנולוגיים. אנו ממליצים לאמץ לכל הפחות את האיסורים בתקנות האירופיות, ובהם האיסור על שימוש בזמן אמת במערכות זיהוי ביומטריות לתכליות של אכיפת חוק והאיסור על שימוש של רשויות ציבוריות במערכות נבונות שנועדו לסווג או להעריך את מידת האמון שיש לתת באדם על בסיס דירוג חברתי.[1068] אפשר לסייג איסורים אלו לצורכי ביטחון לאומי, אך רק בתנאי שרשויות הביטחון יפעילו מערכות אלו באופן מידתי, על בסיס צורך בלבד.[1069]

האירופי" **כלכליסט** (11.7.2022); טל קפלן, "מומחים מתריעים: הזנחת הפרטיות מובילה להצעה בלתי סבירה ומקוממת" Law.co.il (10.7.2022).

1065 ס' 12(ג), ס' 13 (ה), ס' 19 לחוק הגנת הפרטיות. ראו גם כהנא ושני, **לעיל** ה"ש 213, בעמ' 33.

1066 בג"ץ 6732/20 **האגודה לזכויות האזרח בישראל נ' הכנסת** (1.3.2021); תהילה שוורץ אלטשולר, רחל ארידור הרשקוביץ, **שיקולי פרטיות ומעקב אחר אזרחים: נגיף הקורונה – התמודדות עם המשבר העולמי** (המכון הישראלי לדמוקרטיה, 2020); עמיר כהנא "עידן הזוחלים: ההסמכה ההדרגתית של השב"כ לאיתור מגעים של חולי קורנה" **משפטים על אתר** כ (2023); Amir Cahane, *The (Missed) Israeli Snowden Moment*? 34 Int'l J. Intelligence & Counterintelligence 694 (2021)

1067 ראו **לעיל** ה"ש 290.

1068 ס' 3(1) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53. להרחבה ראו **לעיל**, בסעיף 4.1.1.1.

1069 השוו לדוגמה לסי 3(1)(d)(ii), **שם**.

פרק עשירי

יסודות לרגולציה מכוונת זכויות
של בינה מלאכותית בישראל:
ארגז כלים

בפרק זה נציע ארגז כלים רגולטורי ראשוני לאסדרת בינה מלאכותית שנכון לתפיסתנו לחשוב על אימוצו כבר כעת. אומנם טוב היה אילו יכולנו להמתין וללמוד כיצד מתקבלים הדינים המשמשים בעולם הלכה למעשה, ובפרט אילו יכולנו להמתין להתייצבות הדין באירופה. אילו יכולנו להמתין, היינו יכולים לנסח את החוק בישראל כך שתהיה התאמה בינו ובין הדין באירופה ולשמור על הפוטנציאל הכלכלי של השוק באירופה (בדומה לשיקולים שביסוד ההתאמה לדיני הגנת הפרטיות באירופה).

ואולם אין להיתפס לשאננות. לנוכח ההשתהות של המחוקק בישראל בפיתוח דיני הגנת הפרטיות והתאמתם לסטנדרט האירופי של הוראות התקנות הכלליות

בדבר הגנת מידע (GDPR), והיעדר אסדרה של נושאים אחרים מעולמות המשפט והטכנולוגיה, דוגמת רשתות חברתיות, הגנת סייבר ופלטפורמות ליישומי כלכלה שיתופית, יש חשש שמקבלי המדיניות בישראל לא ינצלו כיאות את הזמן העומד לרשותם. אי־אפשר להקים תשתית רגולטורית לבינה מלאכותית בִּן לילה, והיא כרוכה בשיג ושיח עם גורמים בתעשייה ועם מומחים מהאקדמיה ומהחברה האזרחית. יתר על כן, לאחר הקמתה נדרשת תקופת היערכות של השוק לדרישות החדשות.

10.1 אסדרת מערכות לומדות מחייבת הבנה של "מעגל החיים" שלהן

הואיל ועקרונות שונים של אסדרה, כגון הוגנות, פרטיות, שקיפות, אחריותיות וניהול סיכונים, מתבטאים בכל אחד מרכיבי מעגל החיים, כדי ליצור אסדרה אפקטיבית של מערכות לומדות יש להביא בחשבון רכיבים אלו. התרשים שלהלן מציג בקווים כלליים את "מעגל החיים" של מערכות לומדות.

תרשים 4
מעגל החיים של מוצרים מבוססי בינה מלאכותית

היעדר התייחסות לרכיבי מעגל החיים עלול ליצור מצב של אסדרת יתר של רכיבים מסוימים והתעלמות מרכיבים אחרים ומכאן לצמצום האפקטיביות של האסדרה. למשל, ההתייחסות הרחבה בספרות המשפטית למאגרי נתונים ולהשלכות אפשריות של הפגמים בהם על המוצר הסופי באה במידה רבה על חשבון התייחסות להשלכות הבחירה במודלים שונים או לקשיים שעלולים להתעורר לאחר יישום המודל.

לעומת זאת, תפיסה אינטגרטיבית מביאה בחשבון את מכלול רכיבי מעגל החיים ונדרשת גם לקשרי הגומלין ביניהם. למשל: תכלית המערכת ומסגור הבעיה צריכים להשפיע על בחירת המודל (האם לבחור במודל שיש בו עכירות רבה יותר באשר לדרכים שבהן הוא מקבל החלטות או שלא לאפשר זאת?); תוצאות הערכה בשלב בניית המודל מזינות תהליכים להערכת סיכונים, והם בתורם מחייבים קבלת החלטות הנוגעות לאימון המודל, לפרישׂתו או לאבטחתו; תכלית המודל משפיעה על הבחירה בממשקי המשתמש (האם מערכת המעניקה ייעוץ רפואי אמורה להציג את האפשרות שהיא טועה? האם נכון לספר למשתמש שהוא מתקשר עם מערכת מלאכותית ולא אנושית?)

לפירוקו של מעגל החיים לחלקים יתרון נוסף: אם כבר בשלב ראשוני, למשל בשלב איסוף המידע, מוערך שיש פגמים, אפשר לטפל בהם כבר אז; ואם מתעוררות הערכות מדאיגות בעניין אימון של מודל אפשר להשהות אותו או להימנע מיישומו המלא.

לדוגמה, אם במאגר הנתונים מתגלות הטיות, אפשר להידרש לשאלות כמו הטיית ייצוג (אם למשל מדובר על מאגר העוסק במערכת החינוך אפשר לשאול אם המידע מגיע רק ממוסד חינוכי אחד; באיזה אופן אפשר להכליל ממאגר הנתונים המסוים הזה על כלל האוכלוסייה; אם יש ייצוג יתר לקבוצות אוכלוסייה קטנות) או הטיית מדידה (האם המדידה תופסת את מה שנכון לחפש? ואם מדובר בחיזוי מחלה, האם המודל מודד רק את מי שאובחנו בעבר במחלה זו?) ואולם בתהליך בניית המודל אפשר לדבר על הטיות שהן תוצאה של בחירת המודל (למשל, האם בשל תהליכים של דחיסת נתונים, כדי שהמודל יעבוד מהר יותר, נפגעו קבוצות קטנות כבר בשלב אימון המודל?), ואילו לאחר יישום המודל ייתכנו הטיות הנובעות מכך שהמודל אומן בהקשר מסוים (למשל, סיכון לחזרתיות בפשיעה) לצורך הקשר אחר (למשל, קביעת גזר דין).

דוגמה אחרת נוגעת לשקיפות ולהסברתיות. אין דומה הסברתיות שתכליתה להסביר כיצד נאסף מידע (מי אסף את הנתונים, מהי ההתפלגות שלהם וכיו"ב) לכזאת שעניינה להסביר החלטה (למשל, אלה המילים שבשלן נמחק ציוץ פוגעני ברשת חברתית). ועוד דוגמה נוגעת להערכת סיכונים. הערכת הסיכונים בבחירה ראשונית של מודל צריכה להסתמך על סבבי אימון קודמים או על בנייה של

מודל ניסיוני, אבל הערכת סיכונים לאחר סבב אימון ראשוני שמתגלה בו בעיית תאימות בין משימה לתוצאה צריכה להוביל להתאמה מחודשת של הנתונים, של הארכיטקטורה ושל משימות האימון, לעורר את השאלה אם נכון לאמן מודל קטן יותר או חלש יותר, ובוודאי לא להמשיך לאמן מודל בקנה המידה שתוכנן מלכתחילה.

חלק חשוב של הבנת מעגל החיים של מערכות לומדות נוגע לצורך לנטר אותן לאחר שיושמו בפועל (post deployment) בעולם האמיתי (למשל הבניה של המערכת בתוך מוצר או בתוך ממשק). זאת משום שמערכת לומדת, שלא כמו של מוצרים אחרים (תרופות למשל), יכולה מעצם טיבה להשתנות גם לאחר יישומה בשל המשוב שהיא מקבלת מן המשתמשים.[1070] לכן חשוב לשאול לקראת היישום מה עלול להשתבש, אם המודל בטוח ליישום ואילו מנגנוני בטיחות נחוצים כדי ליישמו בבטחה, ואם נדרש יישום בקנה מידה קטן תחילה – מי יקבל התראה במקרה של תקלה. ואילו לאחר היישום יש לשאול למשל באיזו מערכת בקרה לבחור; אם יש ניטור שתכליתו לאתר התנהגויות לא צפויות, בעיקר בסביבות יישום מורכבות (למשל, האם משתמשים מצאו יישומים חדשים או אסטרטגיות הנדסיות חדשות על בסיס המודל?), ולדווח על אירועים כאלה; אם נדרשים עדכונים למודל לאחר היישום שלו ואם הם משמעותיים ברמה שמחייבת הערכה מחודשת של השלבים שלהם (למשל, תיוג מחדש של הנתונים, אימון מחדש של המודל ובדיקה מחודשת שלו).

10.2 פיתוח מתודולוגיות לניהול סיכונים

תסקירי השפעה ומתודולוגיות לניהול סיכונים נעשו מקובלים בתחומים כגון הגנת הסביבה, דיני חברות, פיננסים וזכויות אדם. אבל הצורך ליישם מתודולוגיות לניהול סיכונים על מערכות אלגוריתמיות עדיין נמצא בחיתוליו, למרות חיוניותו. מתודולוגיות אלו צריכות

1070 מינהל המזון והתרופות האמריקני (FDA) מתחיל בניסוח המלצות כאלו למערכות לומדות בתחום הרפואה. ראו רותי לוי "ה־FDA יקל על שיווק מערכות בינה מלאכותית – המשתנות לאחר שהן מאושרות לשימוש" **TheMarker** (11.5.2023).

לכלול הערכה של השפעות וסיכונים לא רק בנוגע לפרטים אלא גם בהקשרים חברתיים רחבים, כמו למשל המתודולוגיות שמציע הארגון Data & Society (D&S) במעבדת ההשפעות האלגוריתמיות שהקים Algorithmic Impact Methods Lab (AIMLab). שהרי מתעורר חשש שאם מי שיקבע מתודולוגיות כאלה יהיה התעשייה בכלל וענקיות הטכנולוגיה בפרט, לא יובאו בחשבון שיקולים רחבים יותר.

היוזמה המרכזית לפיקוח על בינה מלאכותית מתוך פרספקטיבה של ניהול סיכונים היא הצעת החקיקה של האיחוד האירופי, המתמקדת בניהול סיכונים במובן הצר שלו, כלומר בהקשרי שימוש מסוימים בלבד (למשל מעקב ביומטרי). אלא שמודלים יכולים להיות סיכון גם בהקשרי שימוש שלכאורה יש בהם סיכון נמוך.

אפשר להציע אפוא מודל שונה של ניהול סיכונים. לפי מודל זה, כדי להעריך מסוכנות של מערכת נדרשת התבוננות כפולה – תחילה יש להעריך את פוטנציאל המסוכנות של המערכת כפי שתוכננה; אחר כך יש להעריך את רמת הקשר בין המשימה לתוצאה (alignment), כלומר את האפשרות של המערכת לממש פוטנציאל מסוכנות מחוץ לתפקיד שייעדו לה מתכנניה.

הערכה כפולה זו תהיה הבסיס לקבלת ההחלטות וכדי שיהיה אפשר להוציאה לפועל יהיה צורך לנסח כללי משילות ובטיחות, כגון כללים לאימון אחראי (האם לאמן מודל חדש שניכרים בו סימנים מוקדמים של סיכון – וכיצד) וכללים ליישום אחראי (האם, מתי וכיצד ליישם מודלים שעלולים להיות מסוכנים); ולקבוע אילו רמות של שקיפות ותיעוד נדרשות במקרה של מודלים שעלול להיות בהם סיכון קיצוני ואילו בקרות ומערכות אבטחת מידע חזקות יש ליישם בעניינם.

ייתכן שלא בכל מקרה יהיה צורך להפעיל ניהול הסיכונים, אלא רק במקרים שייקבעו – למשל על מודלים שקרובים ליכולות המוצעות של מודלים קיימים או עולים עליהם והקהילה המדעית לא הרבתה לחקור אותם; או מודלים ששונים ממודלים קיימים מבחינת גודל, עיצוב (למשל ארכיטקטורות שונות או טכניקות alignment שונות) או תמהיל היכולות הנובע מהם. למודלים כאלה יש יכולת רבה יותר להצליח במגוון רחב יותר של משימות ולכן גם הזדמנויות רבות יותר לגרום נזק. בהקשר הכלל־חברתי אפשר לקבוע שמדובר במקרים שבהם היקף

ההשפעה הוא גדול (כגון אובדן עשרות אלפי בני אדם, נזק כלכלי או סביבתי בסך מאות מיליארדי דולרים) או שרמת ההפרעה השלילית לסדר החברתי והפוליטי עלולה להוביל למשל למלחמה בין מדינות, לשחיקה משמעותית באיכות השיח הציבורי או להחלשה נרחבת של ציבורים, ממשלות וארגונים אחרים.[1071] ראוי לחשוב גם על תרחישים שבהם הסיכון נובע משילוב יכולות של משתמש יחיד, קבוצת משתמשים או מערכת בינה מלאכותית נוספת.

יצירת מערך ניהול סיכונים יכולה כמובן לעורר קשיים. חשש אחד הוא ששיתוף תוצאות הערכה בין מפתחים יביא לידי לפרסום סודות מסחריים ויתמרץ מתחרים להאיץ יישום של מודלים כדי לשמור על יתרון תחרותי. חשש אחר הוא שאם מודל כלשהו ידע שהוא עומד לפני בדיקה הוא ילמד להציג רק התנהגות רצויה. נוסף על כך, הסתמכות יתר על מערך ניהול סיכונים עלולה לגרום לאמון יתר בהערכות ובתוצאותיהן ולהוביל לפריסת מודלים מסוכנים מתוך תחושת ביטחון מזויפת. מערך ניהול סיכונים אינו יכול אפוא לעמוד בפני עצמו וראוי לשלב אותו עם מחויבות ארגונית ועם כלים נוספים לזיהוי ולהערכת סיכונים. הללו יידונו כעת.

10.3 אסדרת הקשר (Alignment) בין משימה לתוצאה

עד כה עסק הדיון ברגולציה במאגרי המידע שעליהם מתאמנים האלגוריתמים ובתוצרים שלהם. אבל כבר היום אנו נמצאים בשלב מעבר מלמידה לא מפוקחת, שמסתמכת על מאגר מידע מוגדר, ללמידה לא מפוקחת, שצומחת מן האינטראקציה עם משתמשים ועם המשימות שהיא מתבקשת לבצע, מה שאפשר לכנות Experience Based Learning. מערכות לומדות שאינן מבוססות על מאגרים מובנים, או שיזדקקו למאגרים קטנים בהרבה כדי ללמוד, מציבות לפנינו אתגרים חדשים, ובראשם הקירוב בין המטלה שקיבלה המערכת לתוצר שהפיקה במקרים שבהם אין התאמה בין השניים.

1071 Joseph Carlsmith, *Is Power-Seeking AI an Existential Risk* (16.6.2022), available at arXiv

האם מדובר באי־התאמה אסטרטגית, כלומר ביצירה מכוונת של תוצרי לוואי שליליים ובלבד שהמערכת תבצע את המטלה שלה, או שמדובר באי־התאמה אגנוסטית, כלומר ביצירה אגבית ולא מכוונת של תוצרי לוואי שליליים? שאלה זו הופכת לשאלה משמעותית, וכפי שכבר נכתב לעיל היא מעוררת את הצורך למצוא כלים רגולטוריים ההולמים אותה.

הצורך להכיר במודל כמודל מסוכן לא בשל הנזק הישיר שהוא עלול לגרום אלא בשל החשש שיכולותיו ינוצלו לרעה איננו חדש. ואולם הפרישׂה הרחבה של ממשקי בינה מלאכותית יוצרת, והיכולת לבצע מניפולציה על מודלים כדי לגרום להם לבצע משימות שהם לא אומנו לבצע בדרכים שעלולות להימצא מזיקות, מחייבות תשומת לב. כדי מנוע ניצול של מודלים כאלה לרעה יהיה צורך בבקרות משמעותיות טרם יישומם בפועל (deployment). למשל, תידרש הוכחה שהמודל יתנהג כמתוכנן (alignment assessment),[1072] ויהיה צורך לברר, למשל, אם הוא מנסה להשיג מטרות ארוכות טווח, בעולם האמיתי, שהן שונות מאלה שסיפקו המפתח או המשתמש;[1073] אם הוא יכול לשתף פעולה או ליצור "קנוניה" עם מערכות AI אחרות נגד אינטרסים אנושיים;[1074] או אם הוא מתנגד לניסיונות של משתמשים זדוניים לגשת ליכולות המסוכנות שלו.[1075]

ראוי לציין שיש חוקרים הסבורים שבחלק מן המודלים אי־אפשר לפתור את בעיית הקשר בין משימה לתוצאה. לדעתם, שיטות מקובלות כגון למידת על ידי חיזוקים באמצעות משוב אנושי רק יחמירו את הבעיה.[1076] קביעה כזאת עלולה, למשל, להביא לייחוס אחריות משפטית למי שיאפשר שימוש המוני בממשקים ובמודלים. כך או כך, יש לנסות להעריך את התנהגות המערכות באמצעות מפתח

1072 Toby Shevlane, *Sharing Powerful AI Models*, Center for the Governance of AI (20.1.2022)

1073 Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems* (20.2.2023), available at arXiv

1074 ראו Ngo, Chan, and Mindermann **לעיל** ה"ש 77.

1075 Amelia Glaese et al., *Improving Alignment of Dialogue Agents via Targeted Human Judgements* (18.2.2022), available at arXiv

1076 Yotam Wolf et al., *Fundamental Limitations of Alignment in Large Language Models* (29.5.2023), available at arXiv

רחב ככל האפשר, ייתכן אפילו שבאמצעות אוטומציה של הליך ההערכה עצמו. יש להכיר בכך שבתרחישים מסוימים ובשילובים מסוימים של מערכות צפויות בעיית קשר חריפות יותר מאשר באחרים (למשל צ'אטבוט בעל יכולת מניפולציה על התנהגות משתמשים או מערכת שלא תתנגד לאפשרות שהיא תשולב במערכת אוטונומית שיש לה מטרות מזיקות).

סיכוניה של מערכת אינם תלויים רק בבעיית הקשר, אלא גם בדרך שבה היא מתקשרת, ותתקשר בעתיד, עם עולם מורכב. יתר על כן, קשה לצפות את כל המסלולים שעלולים להוביל לתוצאות קיצוניות. ואולם אין פירוש הדבר שמשום כך לא צריך לטפל בבעיה זו.

10.4 תיעוד נתונים, משילות נתונים והליכי בקרה (Auditing) בדיעבד

המורכבות של מערכות בינה מלאכותית בכלל, ושל מערכות לומדות בפרט, מערימה קשיים מיוחדים על קובעי המדיניות והרגולטורים בבואם לנסח כללי אחריות ולזהות בפועל את השרשרת הסיבתית שהובילה לפגיעה בזכויות, בייחוד בהתחשב בשונות בין מגזר למגזר ובין יישום ליישום.

המשותף לכל אלו הוא שבלי תיעוד ראוי הם בלתי אפשריים. הבסיס לכל בחינה עובדתית של כשל קונקרטי במערכות בינה מלאכותית הוא משילות נתונים ותיעוד קפדני של הליכי עבודה, מקורות מידע, תיוגים, מודלים, תהליכי עיצוב קוד, הערכת סיכונים ובסיסי נתונים, וזיהוי הפערים בכל אלו.[1077] עיצוב טוב של התיעוד הוא גם אינטרס של יזמים ומפתחים, שכן הוא מאפשר להם הן לתחקר בדיעבד כשלים ותופעות שהם לא צפו הן לעמוד בחובות תיעוד רגולטוריות ממקורות אחרים.[1078]

1077 ראו למשל ס' 10(2) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

1078 בתקנות הכלליות בדבר הגנת מידע (GDPR), למשל, יש חובת תיעוד מפורשת (ס' 30), אך גם משתמעת. החובה המשתמעת מקיפה יותר: בעל מאגר מידע (controller) נדרש להיות מסוגל להראות שכל מושאי המידע שבמאגר הביעו את הסכמתם לעיבוד המידע.

נוסח ההצעה של התקנות האירופיות מחיל חובת תיעוד מקיפה על מערכות בסיכון גבוה.[1079] יתרונה של הגישה האירופית הוא שהיא מאפשרת בקרה מלאה אקס פוסט הן במישור ההגנה על זכויות אדם הן במישור המשפט הפרטי. הסתמכות על תיעוד חלקי של מסמכי הערכת סיכונים, לעומת זאת, עלולה להיות משענת קנה רצוץ.

אנו סבורים כי תיעוד הוא כלי להתמודדות עם אי־ודאות. הערכת סיכונים נעשית מטבע הדברים מראש, ולנוכח אי־הוודאות הגוברת המאפיינת מערכות בינה מלאכותית[1080] ייתכן שיימצאו בדיעבד חוסרים בדיווחים המסתמכים על הערכות אלו. לפיכך אנו מציעים לקבוע משטר תיעוד איכותי, שסנקציות בצידו, שיאפשר לתחקר בדיעבד נזקים שלא נחזו.[1081] יש לקבוע תקנים אחידים של תיעוד כדי לאפשר פיקוח אחיד, יעיל, שיטתי וחוצה מגזרים על מערכות אלגוריתמיות,[1082] הן בהקשר של בסיסי הנתונים הן בהקשר של התוצרים.[1083] הליכי בקרה נאותים המבוססים על תיעוד אינם חזות הכול ולא יחליפו גישות כגון עיצוב לפרטיות או הנדסת הוגנות לפני פיתוח מודל. אבל הם יוכלו לזהות, ולו בדיעבד, שגיאות מתודולוגיות וטעויות שהביאו לפיתוחו של מודל מוטה – גם אם התוצאות של המודל אינן מוטות כשהן לעצמן.[1084] זאת משום שהחלטות אלגוריתמיות יכולות למשל להיות הוגנות אך שגויות מטעמים אחרים (למשל, כשהחלטה אלגוריתמית שהתקבלה בעניין פלוני שגויה מאחר שהרשומה הפרטנית שלו בבסיס הנתונים מכילה מידע שגוי, בחינת ההחלטה בדיעבד תוכל לזהות את השגיאה ולהביא

1079 ראו **לעיל** בסעיף 4.1.1.2.

1080 ראו למשל Kaminski, **לעיל** ה"ש 1004, בעמ' 18; Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 Harv. J. L. & Tech. 353, 363 (2016)

1081 Mökander et al., **לעיל** ה"ש 693, בעמ' 2.

1082 Andrew Selbst, *An Institutional View of Algorithmic Impact Assessments,* 35 Harv. J. L. & Tech. 117 (2021)

1083 ראו למשל Sarah Holland et al., *The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards*, *in* 12 Data Protection and Privacy, 1 (Dara Halliman et al. eds., 2020)

1084 ראו לדוגמה Mario Martini, *Regulating Algorithms*, *in* Algorithms and Law 100, 122 (Martin Ebers and Susana Navas eds., 2020)

לתיקונה). לכן בקרה של אלגוריתמים בדיעבד נועדה להבטיח לא רק את תקינות קוד המקור של המודל, אלא גם את התקינות של המתודולוגיה הסטטיסטית ושל הליכי איסוף הנתונים ואימונם.

לתפיסתנו, חובת תיעוד צריכה להיות מיושמת בשלבים השונים של מעגל החיים של למידת המכונה. נוסף על כך, אין חובה להשהות את הבקרה בדיעבד המבוססת עליה עד סיום מעגל החיים הזה, ואפשר לחשוב על אבני דרך של בקרה בהתאם להתקדמות התהליך. חובת התיעוד לצורכי בקרה יכולה לעסוק בתקינות הליכי איסוף הנתונים ואימונם או המתודולוגיה הסטטיסטית. אבל היא יכולה להתגלם גם בדרישה לדיווח על אירועים מפתיעים או תוצאות הערכה מדאיגות במהלך אימון מערכות (דיווח כזה יאפשר למשל לרגולטורים לנהל רשימות של גישות הכשרה בסיכון גבוה ולעדכנן); או בשיתוף הערכות סיכונים לפני יישום של מודלים לצורך הערות וביקורת כדי לוודא שעצם היישום הוא בטוח. בתקופת הביניים עד יצירת גוף ידע ואסדרה מבוססת ייתכן שיש מקום לדרוש גם דיווח מדעי, כלומר הצגת תיעוד לקהילה המדעית כדי לעודד מחקר מדעי נוסף בנושא.

אין ספק שאת הבקרה צריכים לעשות בראש ובראשונה גורמים פנימיים לתהליך הפיתוח, שיש להם הבנה עמוקה וגישה למכלול הנתונים והמודל. לפעמים הדרישה לבקרה יכולה לכלול יצירה של שכבות ארגוניות של בטיחות, כגון מינוי פונקציית הערכת בטיחות פנימית שאינה תלויה בפיתוח המודל הקונקרטי ומדווחת לראשי הארגון, שהם בתורם חייבים למסור לגורמים מחוץ לארגון דיווח רזה יותר. חברות טכנולוגיה שונות מציעות דוגמאות לטופסי תיעוד כאלה ואולם חשוב שיהיה להן תוקף ציבורי־רגולטורי.[1085]

בהקשרים אחרים יכול להתעורר צורך לתת לחקורים חיצוניים גישה למודל. ואולם בעקבות גושה ועמיתיו אנו סבורים כי כמו שעושים ביקורת חשבונאית חיצונית לדוחות כספיים של תאגידים, יש לחייב חברות המפתחות ומפעילות מערכות בינה מלאכותית להגיש לגורם ביקורת חיצוני המתמחה בתחום ההוגנות

1085 תבניות כאלה מכונות model cards. ראו למשל את אלו שמציעה חברת גוגל: The value of a shared understanding of AI models

והבינה המלאכותית דגימה מהתוצאות שהמערכת מפיקה.[1086] גישה דומה באה לידי ביטוי בהוראה שנוספה לאחרונה לקוד המינהלי של העיר ניו יורק, הקובעת שהשימוש במערכות החלטה אלגוריתמיות לגיוס או לקידום עובדים יותנה בביקורת חיצונית מפני הטיות.[1087] בקרה יכולה להיעשות גם באופן וולונטרי, במיקור המונים, על ידי הצעת פרס למי שימצאו הטיות (bias bounty, לפי הדגם של bug bounty).[1088]

10.5 פיתוח כלים להתמודדות עם הטיות: הנדסת הוגנות

עמדנו לעיל על בעיית ההטיות האלגוריתמיות, שלעיתים נובעת ממאגרי מידע חסרים או בלתי מתאימים ולעיתים מאפליה מערכתית. מערכות בינה מלאכותית עלולות להנציח, ואפילו להגביר, הטיות – ואחת היא מה מקורן – באמצעות משוב. כדי לזהות הטיות אלו נדרשת אסדרה ייעודית שתַבנה את הבקרה לאורך כל שרשרת הפיתוח של מערכות אלו.

יש הסבורים שתופעת ההטיות האלגוריתמיות היא אכן מטרידה, אבל יש להשלים עם קיומה ולראות בה תחליף להטיות אנושיות, שהן מרובות ועמוקות מאלה של המכונה. אחרים סבורים שהקושי הוא ביישום: אין הגדרה אחידה ומוסכמת של מושג ההוגנות[1089] ואין הסכמה על מדדי ההוגנות;[1090] ולא ברור באיזו חוליה בשרשרת הייצור אפשר להתערב כדי למנוע את ההטיות. הימנעות משימוש

1086 James Guszcza et al., *Why We Need to Audit Algorithms*, Harv. Bus. Rev. (28.11.2018)

1087 N.Y.C. Admin. Code §§ 20-870 – 20-874 (2022)

1088 ראו למשל Melissa Heikkilä, *A Bias Bounty For AI Will Help to Catch Unfair Algorithms Faster*, MIT Technology Review (20.10.2022). לרקע על bug bounties ראו Amit Elazari, *Private Ordering Shaping Cybersecurity Policy: The Case of Bug Bounties*, *in* Rewired: Cybersecurity Governance 231 (Ryan Ellis and Vivek Mohan eds., 2019)

1089 ראו **לעיל** ה"ש 882.

1090 Virginia Foggo, John Villasenor, and Pratyush Garg, *Algorithms and Fairness*, 17 Ohio St. Tech. L. J. 123, 136-153 (2021)

בפרמטרים פסולים, למשל, לא בהכרח תוביל למיגור ההטיות האלגוריתמיות.[1091] מנגד, אפשר לרתום אלגוריתמים חכמים לסיוע בזיהוי הטיות פסולות במערכות נבונות באמצעות שילובם של כלי בקרה אוטומטיים בהליכי הפיתוח והבקרה בדיעבד. למשל, חברת אמזון מציעה כלי לזיהוי הטיות במודלים שפותחו בטכניקות של למידת מכונה,[1092] המסתמך בין השאר על מדדי הוגנות שפותחו באוניברסיטת אוקספורד.[1093] אף שפתרונות אלו עדיין בחיתוליהם, ואינם מקיפים,[1094] החובה למנוע הטיות אלגוריתמיות תתמרץ פיתוח שלהם.

לתפיסתנו יש לפעול לצמצום הטיות אלגוריתמיות באופנים האלה:

א. הליך הוגן מבחינה סטטיסטית

אנו סבורים שנכון לבחון את הזכות להליך הוגן (due process) מבחינה סטטיסטית וליישם אותו. הכוונה לאסטרטגיות מתמטיות, כגון אישור מראש של משתנים המותרים במידול (להבדיל מאיסור על השימוש בהם בדיעבד),[1095] או שימוש בבסיסי נתונים רחבים ועשירים כדי למתן את המשקל הסטטיסטי של משתנים חליפיים.[1096] כיום אסטרטגיות אלו יכולות להשפיע על דיוקו של המודל[1097] ועל רמת החדירה לפרטיות, ואחדות מהן יכולות להימצא לא יעילות אם בסיס הנתונים כשהוא לעצמו משקף הטיות מערכתיות או היסטוריות.

1091 ראו למשל Feuerriegel, Dolata and Schwabe, **לעיל** ה"ש 879, בעמ' 379; Prince and Schwarz, **לעיל** ה"ש 871, בעמ' 1306–1310.

1092 Mark Labbe, *Amazon SageMaker Clarify Aims to Mitigate Bias in Machine Learning*, TechTarget (8.12.2020)

1093 Sandra Wachter, Brent Mittelstadt, and Chris Russell, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI*, 41 Computer Law & Security Review (2021)

1094 ראו למשל Feuerriegel and Schwabe, **לעיל** ה"ש 879, בעמ' 382.

1095 Prince and Schwarz, **לעיל** ה"ש 871, בעמ' 1306–1310.

1096 **שם**, בעמ' 1310–1311.

1097 Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 797 (2017)

ואולם פיתוח של מתודולוגיות סטטיסטיות לאורך כל שלבי הפיתוח של מערכות נבונות, כלומר יישומו של הליך הוגן מבחינה סטטיסטית, יכול לסייע בזיהוי ובאיתור הטיות דווקא במטרה לפתח מודלים מדויקים וחפים מטעויות. יש לזכור גם שלעיתים אלגוריתמים מעלים אל פני השטח הטיות ואפליה שכבר קיימות. כפי שהראינו, המערכת האוטומטית לקבלת מועמדים ללימודים בבית הספר לרפואה סנט ג׳ורג׳ שעתקה את ההטיות של ועדת הקבלה האנושית. לכן אפשר לראות בהטיות האלגוריתמיות הזדמנות לזיהוי ההטיות המערכתיות וההיסטוריות המוטמעות במוסדות החברתיים שהאלגוריתם ממדל.[1098]

ב. גיוון ההון האנושי

כלי נוסף להנדסת הוגנות הוא גיוון ההון האנושי בקרב מפתחים של מערכות נבונות.[1099] ההנחה שביסוד ההצעה לגיוון היא שלמהנדסים בעלי זהות מגדרית, אתנית, דתית או פוליטית שונה יהיו רגישויות שונות להיבטים של הוגנות אלגוריתמית. ואולם יש המטילים בהצעה זו ספק,[1100] שכן רקע אתני שונה לא בהכרח מבטיח מחויבות להוגנות ולא לתכליות עסקיות. לדוגמה, מפתח אפרו־אמריקאי נשאל בריאיון מדוע למוצר הדגל של חברת ההזנק שלו, מערכת בינה מלאכותית מבוססת טלפונים חכמים, נבחרו קול ודמות של אישה לבנה צעירה (ולא למשל של גבר שחור). המפתח השיב שחרף מוצאו הוא חש לחץ להשלים עם דעות קדומות כדי לקדם מכירות.[1101] יתר על כן, מהנדסים ומפתחים עלולים להטמיע מבלי משים את ההטיות הקוגניטיביות שלהם במודל – למשל, מהנדס שמפתח אלגוריתם לזיהוי פנים עשוי לאמן את האלגוריתם להתייחס למאפיינים חזותיים מובהקים של בני מוצא אתני מסוים – צורת העיניים, עובי האף וכדומה. ההחלטות

1098 Jennifer M. Logg, *Using Algorithms to Understand the Biases in Your Organization,* Harv. Bus. Rev. (9.8.2019)

1099 ראו לדוגמה AI HELG 2018, **לעיל** ה"ש 298, בעמ' 20; ס' 33a, 37, 57b להצהרת טורונטו, **לעיל** ה"ש 324; House of Lords Select Committee on Artificial Intelligence, **לעיל** ה"ש 307, בעמ' 57.

1100 ראו Benjamin, **לעיל** ה"ש 201, בעמ' 24–25, 49.

1101 Quentin Hardy, *Looking for a Choice of Voices in A.I. Technology*, The New York Times (10.10.2016)

בעניינים אלו יכולות להתבסס על מחקר אקדמי על מאפייני זיהוי פנים (שאף הם עלולים להיות מוטים), אך גם על ניסיון החיים הייחודי של המפתח, שמושפע בין השאר ממוצאו האתני.[1102]

10.6 התמודדות עם אתגרי השקיפות האלגוריתמית: איתנות מדעית תהליכית

בפרקים הקודמים ראינו שהבעיה העיקרית באסדרת פעילותן של בינה מלאכותית ומערכות לומדות היא בעיית הקופסה השחורה.[1103] האם ההחלטה האוטומטית של האלגוריתם שרירותית או שמא אפשר להצדיקה? העכירות האלגוריתמית מקשה על המפתחים, הרגולטורים והמשתמשים להתחקות על הסיבה לתוצאות בלתי צפויות, שגויות או מזיקות של החלטות מכונה.

בהקשר זה נתקלנו עד כה בשתי בעיות. הראשונה, מקרים של עכירות אלגוריתמית שבהם אי־אפשר ליצור שקיפות אלגוריתמית. כיום מדובר בעיקר במערכות שפועלות כרשתות נוירונים, אך יש להניח שבעתיד יהיו אתגרי שקיפות נוספים. אם הנחת היסוד היא שגם במקרים של עכירות אלגוריתמית, ללא שקיפות במובנה הקלסי, מוצדק לפתח מערכות לומדות ולא לאסור אותן כליל, כי אז נדרשים נתיבים חלופיים שיאפשרו הגנה על זכויות ואינטרסים.

בעיה שנייה היא ההגדרה העמומה של חובות השקיפות האלגוריתמית במסמכים שונים. בגלל עמימות זו קשה להחיל אותה על שחקנים שונים וליצור ודאות בקרב השחקנים בשוק. כדי ליצור מערכת מבוססת בינה מלאכותית נדרשות הוראות ברורות כבר בשלבים הראשוניים של איסוף המידע ולא רק לאחר היווצרות האלגוריתם. ללא כללים ברורים באשר לאופן ניהול המערכת, גיוון המידע הנדרש והליכי התיעוד יתקשו יוצרי אלגוריתמים לחזות את הדרישות הצפויות מהם, דבר שיוביל לכשלים בחישוב העלויות ויפגע ביעילות המוצר הסופי.

1102 Clare Garvie And Jonathan Frankle, *Facial-Recognition Software Might Have a Racial Bias Problem*, The Atlantic (7.4.2016)

1103 ראו בעיקר **לעיל** בפרק 7.

גדי פרל ותהילה שוורץ אלטשולר[1104] הציעו מודל שקיפות המבוסס על התפיסה הקלסית של שקיפות, אבל כולל חלופה שמותאמת למגבלות הטכנולוגיות של מערכות אלגוריתמיות שאינן מסוגלות לספק הסבר רציונלי לפלט נקודתי. המודל מיועד למקרים שבהם אין אפשרות לספק פלט, אבל יש נחיצות חברתית לשמור על חובות של שקיפות כדי שלא למנוע פיתוח טכנולוגיות מסוימות ושימוש בהן. מודל זה מציע שני מסלולים חלופיים למימוש חובת השקיפות: באחד, שמשמש ברירת מחדל, יש לדרוש בדיקה של התוכנה והסברתיות שלה ולחייב את יוצר האלגוריתם לספק פלט שיוכל לנמק מדוע התקבלה החלטה מסוימת ולא החלטה אחרת. במסלול השני יש לדרוש הוכחה ל"איתנות התהליך", כלומר להראות שמהלך הייצור מבוסס על פרמטרים מדעיים מוכחים בשלושה מקטעים בשרשרת הייצור: המידע המוכנס למערכת; עיבוד המידע; יישום האלגוריתם לצורך קבלת תוצר.

כדי לעמוד במבחן הסף המשפטי יש לפרט את פונקציית המטרה של התוכנה ולבחון אם מטרה זו חוקית בנסיבות שבהן יש כוונה להשתמש במערכת. הטלת החובה לפרט מהי פונקציית המטרה תאפשר לנהל דיון בשאלה אם המערכת עומדת בדרישות החוק גם כשמטרתה היא הגברת יעילות בלבד.[1105] בשלב השני יידרש יצרן האלגוריתם או המוצר להוכיח כי מאגר הנתונים שעליו מבוססות ההחלטות הסטטיסטיות של התוכנה עומד בדרישות שייקבעו בהתאם לשימוש המיועד. כך למשל יהיה עליו להוכיח שמאגר הנתונים מתאים לקבוצת היעד שנועד לה – הן מבחינת גודל המדגם הן מבחינת התאמת האוכלוסייה ומאפייניה. יהיה עליו להוכיח גם שהמודל נקי מהטיות שיש בהן כדי לפגוע בחוקיותו, כגון מידע שעלול להביא לתוצאות שונות בקרב קהלי יעד מגזעים או מגדרים שונים גם אם פרטים אלו אינם מופיעים במפורש בתוך המאגר. שלב זה ישפיע על כל המהלך של ייצור האלגוריתמים, שכן הוא יחייב שילוב של מומחי תוכן מטעם היצרן בכל השלבים של ניהול הליך ייצור התוכנה.

1104 פרל ושוורץ אלטשולר, **לעיל** ה"ש 881.

1105 תיתכן טענה שפירוט מטרות התוכנה הוא כשלעצמו סוד מסחרי. טענה זו תהיה מוגבלת למקרים שבהם בדיקת מטרות התוכנה נועדה לברר את חוקיותה.

בשלב השלישי יהיה על היצרן להוכיח שיש הלימה בין יישומי האלגוריתם ובין המשימה שהוא נועד לבצע הן ברמה התאורטית הן ברמה המעשית. בשלב זה יהיה על היצרן לספק מידע בנוגע למגבלות יכולת הניבוי והקטגוריזציה של האלגוריתם בשימוש המיועד, ובכלל זה מידע על תוצאות חיוביות שגויות (false positive) ותוצאות שליליות שגויות (false negative).[1106] יהיה עליו לספק מידע גם בנוגע לאורך חיי התוכנה והצורך לעדכן אותה מעת לעת.[1107] בדומה לנהוג היום בעולם אישורי התרופות – שבו לכל תרופה יש התוויה רפואית מסוימת, ואם מבקשים להשתמש בה לצרכים אחרים יש לוודא את התאמתה ולדווח לרשויות הרלוונטיות – כך יידרשו גם יצרני מערכות בינה מלאכותית לחשוף לפני המשתמשים את מגבלות האלגוריתם בכל הקשור לניבוי ולקטגוריזציה ולדווח לרשויות הרלוונטיות.

הבחירה בין מסלולי השקיפות תתבסס על קריטריון ההשפעה הפוטנציאלית של האלגוריתם והמוצר על זכויות חוקתיות. תובא בחשבון גם זהותו של מפעיל התוכנה – גוף שלטוני ומדינתי, גוף דו־מהותי[1108] או גוף פרטי. במקרה שהשימוש בתוכנה הוא מעין פעולה שלטונית, ההשפעה החוקתית תיחשב חמורה יותר והמבחנים יוחמרו בהתאם.

מסלול האיתנות המדעית התהליכית יכלול את החובה לייצר תיעוד של ההליך ולמסור את המידע למשתמשים. היקף חובת התיעוד והמועד שבו יצטרך היצרן לספק את המידע בעניין התיעוד – קודם לשיווק המוצר או לאחריו – ישתנו

1106 מחקר של מכון התקנים בארצות הברית הוא דוגמה למחקר מקיף שבדק, בדיעבד, מהימנות של תוכנות לבדיקת אלגוריתמים לזיהוי פנים. הובאו בחשבון זיהויים שגויים – הן כשלא זוהה דמיון הן כשזוהה דמיון במקום שלא היה. ראו Patrick et al., Face Recognition Vendor Test (FRVT) Part 7: Identification For Paperless Travel and Immigration (NIST Interagency/Internal Report [NISTIR] 8381, 2022)

1107 לעניין עדכון חובות עדכון התוכנה ראו ס' 13 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

1108 בשלב זה הכוונה היא לגופים דו־מהותיים, כפי שהם מוגדרים בדין הישראלי. ראו ע"א 294/91 **חברת קדישא גחש"א "קהילת ירושלים" נ' ליונל אריה קסטנבאום**, מו(2) 464 (1992). כיום מתנהל ויכוח בעניין מעמדן של רשתות חברתיות לנוכח תפקידן בהחלפת רעיונות. ראו נועה מור "רשתות חברתיות מקוונות כזירות לעיצוב זכויות והקצאתן: לקראת החלתן של חובות מן המשפט הציבורי" **דין ודברים** יד (2019).

בהתאם לנסיבות, לשוק הרלוונטי ולעוצמת הפגיעה בזכות החוקתית. לשם הגנה על סודות מסחריים לא יצטרך היצרן בהכרח לספק את הנתונים לכל משתמש; במקרים מסוימים יהיה אפשר להסתפק בחשיפת החומרים לצד שלישי מוסכם או לרגולטור, כדי שאלו יאשרו מראש את העמידה בתנאים.[1109]

על פי הצעה הנדונה כאן ישולב המודל במסגרות חקיקתיות שיאמצו את הפרשנות למושג "שקיפות" ואת המסלולים החלופיים שבו. גם בתקופת הביניים, עד שתנוסח החקיקה, יוכל המודל לספק בהירות בעניין החובות האתיות מבחינת שקיפות ולאפשר רגולציה עצמית.

10.7 שילוב של אסדרה ענפית ואסדרה רוחבית

אסדרה רוחבית לבינה מלאכותית מבקשת להחיל מדיניות כללית על כלל הפעולות הקשורות לבינה מלאכותית בכל המגזרים. כזה הוא נוסח הצעת תקנות הבינה המלאכותית האירופיות. בדומה, חוק האחריותיות האלגוריתמית שהתקבל בארצות הברית בשנת 2022 מטיל על ועדת הסחר הפדרלית את האחריות לדרוש הערכת השפעה של מערכות בינה מלאכותית בענפי המשק השונים.[1110]

אסדרה רוחבית משיגה משילות וודאות מצד השלטון המרכזי, יוצרת הרמוניה רגולטורית ויכולה אפוא להגביר את אמון הציבור ואת הוודאות הרגולטורית בעבור התעשייה. היא יכולה ליצור האחדה, למשל באמצעות הדרישה לרישומן של כל המערכות האוטומטיות לקבלת החלטות שברשות הממשלה; לדרוש תקינה אחידה; לדרוש ציות לאמנות וחקיקת יסוד; ולדרוש כללים אחידים (למשל חובת גילוי נאות של מוצר מלל או תמונה שנוצרו על ידי מערכת אלגוריתמית או חובת הזדהות של מערכת לא אנושית ככזאת). כל זה בייחוד לנוכח העובדה שהרגולציה של בינה מלאכותית צריכה להשתלב באסדרה

1109 סוגיית הקניין הרוחני תופסת מקום מרכזי בהצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53. ראו למשל סעיף 3.5 לדברי המבוא להצעה.

1110 ראו **לעיל** בסעיף 4.5.

הקיימת של העולם הדיגיטלי, ובראשה דיני הגנת הפרטיות, דיני מידע והגנת סייבר. לכן מצד אחד, עליה להידרש למאפיינים הייחודיים של הבינה המלאכותית, ומצד שני עליה להיות ערה לעומס הרגולטורי במרחבים אחרים של העולם הדיגיטלי. יתר על כן, הרגולציה לא נעשית בחלל ריק מרגולטורים ולכל רגולטור תרבות ארגונית שונה.

לא מדובר רק בכך שהאסדרה הזאת צריכה להכיר בקיומם של דברי חקיקה אחרים אלא גם בכך שהיא עשויה להשפיע עליהם. תיתכן סתירה בין אינטרסים, למשל הזכות לפרטיות סותרת את הצורך במאגרי מידע גדולים לצורך פיתוח מודלים אלגוריתמיים איתנים סטטיסטית. ייתכנו גם כפילויות – למשל כשיש דרישה לתעד תהליכים לצורכי בקרה מצד גורמים נפרדים.

לצד האפשרות של אסדרה רוחבית יש גם אפשרות של אסדרה ענפית, כפי שמוצע למשל במסמך מדיניות הרגולציה של משרד החדשנות.[1111] דוגמאות לאסדרה ענפית אפשר למצוא גם בחקיקה המוניציפלית של ניו יורק, המסדירה שימוש בבינה מלאכותית בגיוס עובדים; בחקיקה במדינת אילינוי, שחלה רק על שימוש בבינה מלאכותית בראיונות וידאו במהלך גיוס לעבודה;[1112] ובחוק האחריותיות הטכנולוגית במקום העבודה של מדינת קליפורניה, המגביל את השימוש במערכות אוטומטיות לקבלת החלטות המשמשות לניטור במקומות עבודה.[1113]

אסדרה ענפית מאפשרת שימוש ברגולטורים קיימים ובסמכויות האכיפה שלהם, איננה מחייבת להקים מסגרות מוסדיות חדשות, מאפשרת יצירת הסדרים ואמצעי אכיפה המותאמים בצורה מדויקת יותר לענף הרלוונטי, מגבירה את הבהירות והוודאות הרגולטורית ומאפשרת גם לבעלי עניין בתוך כל מגזר להשתתף

1111 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 93–94.

1112 820 ILL. Comp. Stat. 42/1 Artificial Intelligence Video Interview Act; ראו גם Ajunwa, **לעיל** ה"ש 118.

1113 Airlie Hilliard et al., *Overview and Commentary of the California Workplace Technology Accountability Act,* 37 Int'l. Rev. L. Comp. & Tech. 99 (2023)

בחשיבה על ההסדרים האלה. אסדרה מסוג זה יכולה לשמש להגנה על זכויות האזרח ועל זכויות הצרכן, אבל גם לחיזוק הוודאות וממילא לחדשנות ולעידוד השקעות.

הקושי באסדרה ענפית הוא בסתירה אפשרית בין ענפים, ביצירת סטנדרטים לא אחידים, בהעמקת הפערים בין הרגולטורים (למשל פערים באוריינות הדיגיטלית, ביכולות האכיפה ובתרבות הארגונית). כמו כן, היא עלולה להשאיר מרחבים לא מאוסדרים שנופלים בין הכיסאות.

כדי להימנע ממגרעות האסדרה הענפית, ולהבטיח את יתרונות האסדרה הרוחבית ואת יתרונות האסדרה הענפית גם יחד, אנו ממליצים לשלב אסדרה רוחבית ואסדרה ענפית, לצד היכרות עם שדות הפעילות המשפטיים. כך, יש לעצב את האסדרה בתווך שבין מעגל החיים של למידת המכונה והשלכותיו, שפורטו לעיל, הענף הספציפי שבו מיושמת המערכת והדרישות המשפטיות הרלוונטיות, כפי שמתואר בתרשים 5.

תרשים 5

מודל משולב רוחבי וענפי לאסדרת בינה מלאכותית

10.8 חקיקה משלימה ועדכוני חקיקה

גם בהיעדר חוק בינה מלאכותית כללי וייעודי יש לחייב כבר בעת הזאת את מקבלי ההחלטות והרגולטורים הייעודיים לעדכן את החקיקה הקיימת ולחוקק חקיקה משלימה. הכוונה בעיקר לחוקים כגון חוק התחרות וההגבלים העסקיים, חוק זכויות יוצרים, חוק הגנת הפרטיות, פקודת הראיות וחוק הרכש הממשלתי.

למשל, יש להוסיף לחקיקה הקיימת התמודדות עם זכויות יוצרים ובעלות על יצירות שנוצרו באמצעות בינה מלאכותית;[1114] להידרש לסוגיות שקשורות לחובות זיהוי של מוצרים המבוססים על בינה מלאכותית בעת התקשרות עם לקוחות פוטנציאליים; וכן להוראות שונות הקשורות לאיסוף המידע ולשימוש בו כשמדובר בתוכנות המשמשות משרדים ממשלתיים. אנו מציעים כי ההליך ייעשה בהובלת משרד המשפטים וייקבעו סדי זמנים להשלמת תיקוני החקיקה הנדרשים בשלב זה.

1114 בהקשר של יצירות אלגוריתמיות, סוגיות של קניין רוחני יכולות להשפיע השפעה ניכרת על תפוצת הטכנולוגיה. ראו למשל את גישתו של ריאן אבוט, המעודד ניטרליות חקיקתית. Ryan Abbott, The Reasonable Robot (2020). על אפשרות של נזק עקב היעדר אסדרה בתחום אפשר ללמוד מן החלטה של חברת Getty, המציעה מאגרי תמונות לשימושים מסחריים, שלא לכלול בהם תמונות שיוצרו באמצעות אלגוריתמים כגון DallE2 מחשש לחשיפה משפטית. James Vincent, *Getty Images Bans AI-Generated Content over Fears of Legal Challenges*, The Verge (21.9.2022). ראו גם משרד המשפטים **שימושים בתכנים מוגנים בזכויות יוצרים לצורך למידת מכונה** (18.12.2022).

פרק אחד עשר

המלצות לגבי מוסד מאסדר לבינה מלאכותית בישראל

במאי 2023 נערך בסנאט האמריקני שימוע לסאם אלטמן, מנכ״ל חברת OpenAI. אחת ממסקנות השימוע הייתה שיש להקים רשות פיקוח על בינה מלאכותית בארצות הברית. ואולם מייד הועלו הטענות שהקמת רשות כזאת היא המלאכה הקלה – האתגר הקשה הוא צידה בסמכויות המתאימות.[1115]

נוסף על כך, לצד האתגר הרגולטורי של תרגום עקרונות אתיים רחבים לכללים משפטיים יש שאלה נפרדת: מהי זהות הגוף או הגופים המוסדיים האמורים

1115 Andrew Ross Sorkin et al., *Washington Confronts the Challenge of Policing A.I.,* The New York Times (17.5.2023)

להיות ממונים על תחום הבינה המלאכותית הן כגופים מיישמי חקיקה, הן כגופים מאסדרים ומפקחים והן כגופים מייעצים? השאלה אם גוף כזה נדרש מלכתחילה, ומה צריכה להיות המסגרת המוסדית שלו, תידון בפרק זה הן באשר לישראל כשהיא לעצמה הן בהשוואה למתרחש בעולם. נעסוק בשאלה על פי אילו קווים מנחים צריך לפעול מוסד מאסדר לבינה מלאכותית בישראל ומהם השיקולים הרלוונטיים בעניין זה. נציג את פונקציית המטרה של המוסד המאסדר, את הסביבה הרגולטורית שלו ואת הגופים המקבילים אליו, וננסה לבחון מאילו מהם ניתן לשאוב השראה. נקדים ונעיר כי שאלה זו תלויה במידה רבה בהכרעה אילו אינטרסים אמורה רגולציה של בינה מלאכותית לקדם ואילו חלקים בשרשרת הערך היא אמורה לתפוס ברשתה. כיוון שהרגולציה של בינה מלאכותית עודה בחיתוליה בעולם כולו, קשה לקבוע מסמרות בדבר, אבל חשוב בעינינו להצביע על המסגרת הנחוצה כבר כעת.

האסדרה של תחום הבינה המלאכותית עדיין נמצאת בראשיתה ועדיין מתפתחת ולכן גמישותה חשובה ביותר. רגולטור של בינה מלאכותית יכול למלא תחילה בעיקר תפקידים של ריכוז מומחיות, ייעוץ והמלצה לקובעי מדיניות, ועם הזמן להרחיב את פעילותו ולדאוג גם לפיקוח ולאכיפה. ככל שטכנולוגיות אלו יוטמעו יותר במשרדי הממשלה[1116] כך יגדל הצורך בממשק בין הרגולטור, כגורם מקצועי פנים־ממשלתי מוביל, לרשויות אחרות.

מתווי רגולציה של ניהול סיכונים דורשים פעילות רגולטורית מתמשכת שתחזור ותעריך את רמת הסיכון הנשקפת מטכנולוגיות חדשות, תחזור ותתאים את המסגרת הרגולטורית לצורכי השעה ותיתן מענה קונקרטי לסוגיות פרשניות שעשויות להתעורר עד כניסת החקיקה לתוקף וגם מייד לאחר מכן. גם היישומים של מערכות נבונות והשפעתן על החברה יכולים להשתנות, להתרחב ולהתפתח. תופעות הלוואי של טכנולוגיות אלו – רצויות ולא רצויות – ילכו ויתגלו ככל ששימושן יתרחב וייעשה רבגוני יותר. הבחינה החוזרת של המדיניות הרגולטורית לצורכי עדכונה יהפכו לעניין שבשגרה.

1116 ראו המיזם הלאומי למערכות נבונות (2020) ב, **לעיל** ה"ש 981, בעמ' 214–227.

11.1 פונקציית המטרה

לגופים העוסקים בבינה מלאכותית, בין שהם כבר קיימים ובין שרק הוצעו, יש כמה תכליות. מכל אחת מהן נגזרים שיקולים אחרים באשר למהות המוסד המאסדר ולעיצובו. נמנה אותן להלן:

א. קידום חדשנות, פיתוח וצמיחה: באסטרטגיה של בינה מלאכותית לאומית מודגש לרוב תפקיד המדינה בעידוד ההצטיינות הלאומית בתחום וביצירת מערכת אקולוגית של חדשנות, יזמות וצמיחה במגזר הפרטי.[1117] עם תפקידי המדינה יש למנות גם הקצאת משאבים להקמת תשתיות, למחקר ולהכשרת כוח אדם.

ב. שמירה על אינטרסים ביטחוניים: תוכנית החומש הישראלית מונה בין השאר, כצפוי, גם מטרות ביטחוניות.[1118] המוסד שהוצע בה הוא מוסד מקביל למערך הסייבר הלאומי, כלומר מוסד בעל אוריינטציה ביטחונית.

ג. מחקר בתחום "הבינה המלאכותית האחראית", השלכות המַקרו של בינה מלאכותית על החברה, הכלכלה והדמוקרטיה, והטמעת עקרונות כלל־חברתיים ואתיים בשימוש בבינה מלאכותית.

ד. עיצוב ההסדרים הנורמטיביים האמורים לחול על מערכות נבונות, פיתוחן, הפצתן והשימוש בהן.

1117 ראו לדוגמה את העיקרון האתי הראשון במסמך מדיניות הרגולציה של משרד החדשנות, **לעיל** ה"ש 314, בעמ' 103–104, "בינה מלאכותית לקידום צמיחה, פיתוח בר־קיימא ומובילות ישראלית בחדשנות". ראו גם ס' 3(3) לתזכיר חוק הגנת הסייבר, **לעיל** ה"ש 461, שלפיו אחד מתפקידי מערך הסייבר הוא לקדם מובילות ישראלית בתחום הסייבר. משרד הדיגיטל, המדיה, התרבות והספורט של בריטניה פתח לאחרונה בתהליך של גיבוש רגולציית דיגיטל (שבה נכללת גם רגולציה של מערכות בינה מלאכותית). מטרתה הראשונה של רגולציה זו היא "לעודד צמיחה ולקדם תחרות וחדשנות בכל רחבי המגזר הדיגיטלי". ראו Department of Digital, Culture, Media and Sports, Digital Regulation: Driving Growth and Unlocking Innovation (13.6.2022)

1118 בן ישראל, מתניה ופרידמן ממליצים על פיתוח מערכות שו"ב לימי שגרה ולשעת חירום. הם אף מציינים כי מטבע הדברים הם מנועים מלהמליץ על פרויקטים ביטחוניים מסווגים יותר. המיזם הלאומי למערכות נבונות (2020) א, **לעיל** ה"ש 981, בעמ' 37–38; להמלצות לקידום מערכת אקולוגית של מערכות נבונות בתחום הביטחוני ראו **שם**, ב, בעמ' 248–255.

ה. פיקוח ואכיפה באמצעות כללי אתיקה וכללים רגולטוריים והנחיה ישירה או עקיפה של גופים ציבוריים ושל המגזר הפרטי.

ו. ייעוץ, יצירת ידע, הפצתו והנחיה בנוגע לסוגיות שמעורר השימוש בבינה מלאכותית. לנוכח הדינמיות של התחום יצטרך הרגולטור להכיר לעומק את חזית הטכנולוגיה[1119] ולהפיץ את הידע הזה באופן מסודר.

נקדים ונאמר: לתפיסתנו לא נכון להקים מוסד שאמור לממש את כל התכליות שנמנו לעיל משום שמקצתן ביצועיות ומקצתן נורמטיביות או רגולטוריות. נוסף על כך, ייתכנו קונפליקטים ביניהן. למשל, עידוד חדשנות דורש אמון בין התעשייה למדינה, המתבטא בראש ובראשונה בשקיפות מרבית, ואילו שמירה על אינטרסים ביטחוניים נעשית לרוב בעלטה; קידום פיתוח וצמיחה עלול להתנגש בצורך לעצב הסדרים נורמטיביים, שכן לעיתים הצורך להגן על אינטרסים ציבוריים אחרים מעכב חדשנות.

לכן אנו מציעים שגוף אסדרה של בינה מלאכותית יתמקד בשלוש התכליות האחרונות שמנינו, שהן תכליות המאפיינות רגולציה. על שתי התכליות הראשונות צריכים להיות ממונים הרשות לחדשנות ושירותי הביטחון. במקביל, אנו מציעים לשקול הקמת מרכז מחקר לאומי בין־תחומי בנושאי בינה מלאכותית וחברה, שתכליתו תהיה ליצור ידע ולגבש אסטרטגיות בעניין שימושים אחראיים ואתיים במערכות בינה מלאכותית. מרכז כזה יענה על הצורך להציע הבנה אינטגרטיבית של התחום, שממנה ייגזרו לאחר מכן מדיניות ופרקטיקות שאינן כפופות לאינטרס המדינתי בקידום התעשייה מצד אחד וברגולציה ספציפית מצד שני. יש לתת לו מימון רב־שנתי, לפחות לתקופת ביניים של כחמש שנים, עד שיתגבשו גופי ידע ברשויות הפיקוח והאסדרה, באקדמיה ובתעשייה. המרכז למצוינות בנושא קבלת החלטות אוטומטית וחברה שהקימה ממשלת אוסטרליה בשנת 2020 היא דוגמה למרכז כזה.[1120]

1119 ראו למשל את עמדתם של אחיעז ואח', **לעיל** ה"ש 240, בעמ' 147.

1120 ARC Centre of Excellence for Automated Decision-Making and Society

11.2 הסביבה הרגולטורית

אסדרת תחום הבינה המלאכותית נמצאת בסביבה שיש בה ארבעה מרחבים רגולטוריים, שכל אחד מהם כולל אסדרה מהותית וגם הקשר מוסדי.[1121]

11.2.1 רגולטורים משיקים

תחום הבינה המלאכותית משיק לתחום ההגנה על הפרטיות ולתחום הגנת סייבר בעניינים רבים, ממש כשם שיש חפיפה עקרונית בין תחום ההגנה על הפרטיות לתחום הגנת הסייבר.[1122] למשל, הקפדה על עקרון העיצוב לפרטיות[1123] תורמת להגנת סייבר על ידי הקטנת הסיכון לזליגת מידע שמלכתחילה לא היה בו צורך. מבחינת אסדרה, יש השקה גם בין תחום הבינה המלאכותית לתחומים כגון תחרות, מיסוי, הגנת הצרכן, ואפילו לתחום הפיקוח על הבחירות.

לפיכך הצורך ברגולטור מתאם הוא מורכב. יש חשש שרגולציה של רגולטורים מגזריים ללא גורם רגולטורי מתאם תיצור מערך נורמטיבי של טלאים ושל פערים בין מגזרים ותביא לחוסר תיאום בין רגולטורים. במקרה כזה הם עלולים לדרוש מהגופים המאוסדרים דרישות סותרות, חופפות או כפולות.[1124]

במבט אל העתיד הרחוק יותר, ייתכן שבשלב מסוים יבשילו הנסיבות ויהיה אפשר לרכז את הרגולטורים של כלל תחומי הדיגיטל תחת קורת גג אחת. רגולטור דיגיטל כזה יהיה ממונה, בין השאר, על הגנת מידע, אבטחת סייבר[1125] ובינה מלאכותית, ויאפשר הרמוניזציה של תהליכים רגולטוריים. למשל, במקום שתורת ההגנה בסייבר, כללי אבטחת המידע, דרישות התיעוד הטכני והוראות ממשל

1121 לגישה הבריטית, המאחדת תחומים אלו – ואחרים – ב"רגולציית דיגיטל", ראו Department of Digital, Culture, Media and Sports, **לעיל** ה"ש 1117.

1122 ראו מערך הסייבר הלאומי, **תורת ההגנה – לנהל את הסיכון: המדריך היישומי (השלם) להגנת הסייבר של ארגון**, 72–74, נספח ו (2021).

1123 ראו **לעיל** ה"ש 412.

1124 הצוות הבין־משרדי לרגולציה חכמה, **תכנית לאומית למדיניות רגולציה ככלי לשיקום המשק ביציאה ממשבר הקורונה** 44–45 (2021); Calo, **לעיל** ה"ש 956, בעמ' 6.

1125 הכוונה לפונקציות הרגולטוריות של מערך הסייבר הלאומי לפי פרק ד בתזכיר חוק הגנת הסייבר (**לעיל** ה"ש 461). ספק אם מקומה של הגנת סייבר שוטפת במסגרת ה־CERT הלאומי הוא תחת רגולטור דיגיטל.

הנתונים יהיו תחומים נפרדים, יהיה אפשר להציג מעין "תורת דיגיטל" מאוחדת שתידרש לכל התחומים האלה באופן הוליסטי ותפחית את הנטל הרגולטורי המוטל על כתפי הגופים המאוסדרים. אבל עניין זה ידרוש התאמה חקיקתית ואפשר שיהיו לו גם חסרונות, למשל הסתירה המובנית בין אבטחת סייבר ובין הגנה על פרטיות. כך או כך, אנו סבורים שאין להמליץ על כך בשלב זה.

11.2.2. רגולטורים ענפיים

גופים שמשתמשים בבינה מלאכותית יכולים להיות כפופים לרגולטור ענפי, ולעיתים אף לכמה רגולטורים ענפיים (למשל, מוסדות פיננסיים נתונים לפיקוח הרשות לאיסור על הלבנת הון וגם לפיקוח של רגולטורים פיננסיים – המפקח על הבנקים, הרשות לניירות ערך או רשות שוק ההון, ביטוח וחיסכון).

יש הסבורים שאין צורך למנות רגולטור ייעודי לתחום הבינה המלאכותית ומוטב להפקידו בידי הרגולטורים הענפיים. דוח ועדת המשנה, כמו מסמך מדיניות הרגולציה של משרד החדשנות, ממליצים שעיקר האסדרה תיעשה באמצעות הרגולטורים הענפיים, על יסוד גורם רגולטורי מומחה ועל־ממשלתי שיתאם בין הרגולטורים הענפיים.

לאסדרה ענפית יש יתרונות, ובהם היכולת להתמודד עם הקושי לעדכן את הרגולציה הריכוזית על ידי ניסויים רגולטוריים בהקשרים מגזריים מבודדים יחסית.[1126] עם זאת, לעומת רגולטור מרכזי של בינה מלאכותית, רגולטורים ענפיים לא תמיד רגישים די הצורך להגנה על זכויות יסוד.[1127] יתר על כן, ייתכנו פערים ביניהם גם ברמת האוריינות הדיגיטלית שלהם.

תקנות מערכות ההמלצה האלגוריתמיות הסיניות, למשל, קוראות לארגונים ענפיים רלוונטיים לפתח סטנדרטים ענפיים ולהנחות ספקים.[1128] לפי התקנות האירופיות, כל מדינה שחברה באיחוד תקים רשות לאומית מוסמכת, שתפקידה

1126 שרון בר זיו וטל ז'רסקי "פרטיות במשבר זהות: אסטרטגיות הסדרה בעידן ההתממה" **משפט, חברה ותרבות ב** 125, 131–130 (2019).

1127 **שם**, בעמ' 165.

1128 ס' 2 לתקנות מערכות ההמלצה האלגוריתמיות הסיניות, **לעיל** ה"ש 590.

יהיה להבטיח את יישומן והטמעתן של התקנות המוצעות ולשמש מפקח על השוק.[1129] ביחס לענפים מסוימים, הרגולטור הענפי ישמש מפקח שוקי.[1130]

גם שרון בר זיו וטל ז׳רסקי סבורים שייתכן שהמומחיות הטכנולוגית של הרגולטורים הענפיים תעלה על זו של רגולטור מרכזי.[1131] הסמכת רגולטור ענפי לנהל ניסויים מבוקרים בארגזי חול רגולטוריים היא דוגמה לכלי המתאים לתקופה של ניסוי וטעייה. כך מומלץ בין השאר בדוח ועדת המשנה לאתיקה של המיזם הלאומי למערכות נבונות,[1132] בדוח על יישומי בינה מלאכותית במגזר הפיננסי[1133] ובמסמך מדיניות הרגולציה והאתיקה של משרד החדשנות.[1134] בתקנות האירופיות מוצע לאפשר למדינות חברות להקים ארגזי חול רגולטוריים,[1135] שבהם יהיה אפשר לחרוג, בהתקיים תנאים מסוימים, מעקרון צמידות המטרה שבדיני הגנת הפרטיות ולהשתמש בנתונים לתכליות אחרות מאלו שלשמן הם נאספו. הגמשת הכללים כדי לאפשר פיתוח מבוקר של טכנולוגיה חדשה בהנחיית הרגולטור מאפשרת לשכלל את המסגרת הנורמטיבית החלה על מערכות נבונות ולהתאימה לצרכים מגזריים מיוחדים. לכן היא מתאימה במיוחד לרגולטורים ענפיים.

11.2.3. מערכות פרטיות וציבוריות

ההיבט השלישי של האסדרה נוגע להפרדה בין מערכות המשמשות במגזר הפרטי לאלה המשמשות במגזר הציבורי. כיום למשל הרשות הלאומית לחדשנות טכנולוגית[1136] פועלת ״לעידוד, לקידום, לתמיכה ולסיוע

1129 **שם**, ס' 63.

1130 **שם**, ס' 63(1).

1131 בר זיו וז'רסקי, **לעיל** ה"ש 1084, בעמ' 164.

1132 המיזם הלאומי למערכות נבונות (2019), **לעיל** ה"ש 313, בעמ' 32.

1133 אחיעז ואח', **לעיל** ה"ש 240, בעמ' 148.

1134 מסמך מדיניות הרגולציה של משרד החדשנות, **לעיל**, ה"ש 314, בעמ' 99–101.

1135 ראו **לעיל** בסעיף 4.1.2.

1136 ראו תיקון מס' 7 לחוק לעידוד מחקר, פיתוח וחדשנות טכנולוגית בתעשיה, תשמ"ד–1984, ס"ח 100 (להלן: חוק המו"פ); אורה קורן "זריקת עידוד לתעשייה – או צעד מסוכן? רשות החדשנות הלאומית יוצאת לדרך" **TheMarker** (30.7.2015).

לחדשנות הטכנולוגית בתעשייה",[1137] ואינה מטפלת במגזר הציבורי; רשות התקשוב הממשלתי אמונה על "הרחבת תחומי פעילות התקשוב הממשלתי, עידוד חדשנות במגזר הציבורי וקידום המיזם הלאומי 'ישראל דיגיטלית' ";[1138] ואילו הרשות להגנת הפרטיות ממונה על ביצוע הוראות חוק הגנת הפרטיות במגזר הציבורי והפרטי גם יחד.[1139] בפועל, מבקר המדינה מצא כי "על אף החשיבות שהרשות מייחסת למעורבותה בסוגיות הגנת הפרטיות במאגרי מידע העולות במגזר הציבורי, פעילותה בנושא חלקית" ואין היא מעורבת באופן אקטיבי בפרויקטים ממשלתיים בנושא.[1140]

אנו סבורים כי בדומה לרשות להגנת הפרטיות, יהיה נכון שרגולטור אחד יהיה מופקד על האסדרה של בינה מלאכותית במגזר הפרטי ובמגזר הציבורי גם יחד. ראשית, כיוון שמערכות בינה מלאכותית במגזר הציבורי מבוססות על פיתוח שנעשה בשוק הפרטי. שנית, כיוון שבמגזר הפרטי נדרשים סטנדרטים להפעלת מערכות בינה מלאכותית, לפיתוחן ולשימוש בהן בהקשרים סמי־ציבוריים כגון אלגוריתמיקה של הפצת דעות ורעיונות, מערכות קבלה לעבודה, אפליית מחירים ועוד. שלישית, כיוון שכל הזמן מתקיים קשר דו־כיווני בין שני המגזרים. למשל, רשות החדשנות יכולה לעודד פיתוחים טכנולוגיים במגזר הפרטי על פי תעדוף אסטרטגי של הממשלה.[1141] לרשות התקשוב הממשלתי יש יכולת להשפיע על השוק הפרטי באמצעות הכתבת סטנדרטים במכרזי רכש.[1142] מנגד, רשויות שלטון רוכשות טכנולוגיות שהתפתחו למען השוק הפרטי.

1137 ס' 5א לחוק המו"פ, **שם**.

1138 החלטה 2097 של הממשלה ה־33, "הרחבת תחומי פעילות התקשוב הממשלתי, עידוד חדשנות במגזר הציבורי וקידום המיזם הלאומי 'ישראל דיגיטלית' " (10.10.2014).

1139 בר זיו וז'ירסקי, **לעיל** ה"ש 1084, בעמ' 129. ראו גם ס' 30 בהצעת חוק הגנת הפרטיות–2019, ודברי ההסבר לו. ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 403, בעמ' 116.

1140 מבקר המדינה, **דוח שנתי 69ב** 6 (2019).

1141 הרשות לחדשנות, למשל, מפעילה את "הזירה החברתית־ציבורית", שבה היא מעודדת מחקר ופיתוח טכנולוגיים לפתרון אתגרים חברתיים וציבוריים. במסגרת זו פועל מסלול לעידוד חדשנות במגזר הציבורי (ממשלטק). ראו רשות החדשנות, **2019: חדשנות בישראל, תמונת מצב** 96 (2020).

1142 על שימוש במכרזים ציבוריים ככלי של רגולציה רכה ראו **לעיל** ה"ש 981.

11.2.4 ההיבט הבינלאומי

חברות ישראליות שמפתחות מוצרים המיועדים לשוק האירופי, למשל, נדרשות לציית הן להוראות התקנות הכלליות בדבר הגנת מידע (GDPR) הן להוראות הדין הישראלי. מן הראוי שיתכוננו כבר כעת לדרישות שעתידות להעמיד התקנות האירופיות.[1143] אפשר שחברות פרטיות יידרשו לעמוד גם בתקנים מקצועיים בינלאומיים, כלומר למלא דרישות נוספות (חלקן חופפות) על אלו המוגדרות בדין.[1144]

הדין האירופי בתחום הפרטיות ובתחום אסדרת רשתות חברתיות, וכפי שמסתמן גם בהצעה לאסדרת הבינה המלאכותית, הוא אקסטרה־טריטוריאלי מאוד ויחייב תאימות של הדין הישראלי אליו. בדיוק כפי שהרשות להגנת הפרטיות אחראית על תאימות הדין הישראלי לדין האירופי בתחום הפרטיות, אנו מציעים שהרשות המרכזית תהיה אחראית על התאימות לדין שייכנס לתוקף בתחום הבינה המלאכותית.

11.2.5 האם המבנה של מערך הסייבר מתאים לשמש רגולטור בינה מלאכותית?

לתפיסתנו, מערך הסייבר הלאומי אינו צריך לשמש השראה לרגולטור של בינה מלאכותית. ראשית, יישומי בינה מלאכותית אינם נחלת התחום הביטחוני, ואין בפיתוח של מערכות נבונות בתור שכאלו כדי לקדם יכולות ביטחוניות.[1145] יש להימנע אפוא משעתוק הדגם של מערך הסייבר הלאומי, שכמה

1143 האומדן של מולר לעלויות הציות של המגזר העסקי להוראות תקנות הבינה המלאכותית האירופיות המוצעות הוא 10.9 מיליארד אירו לשנה. ראו Benjamin Mueller, *How Much Will the Artificial Intelligence Act Cost Europe?* Center For Data Innovation (July 2021)

1144 לסוגיית החפיפה בין תקני אבטחת סייבר לדרישות דיני הפרטיות ראו למשל Isabel Maria Lopes, Teresa Guarda, and Pedro Oliveira, *How ISO 27001 Can Help Achieve GDPR Compliance,* 2019 14th Iberian Conference on Information Systems and Technologies (CISTI) 1 (2019). לדוגמה להתמודדות של הרגולטור עם חפיפות כאלו ראו מסמך המיפוי של מערך הסייבר: מערך הסייבר הלאומי, **תורת ההגנה – מונגשת בשבילך** (19.10.2017). ראו גם ס' 43(א)(1) לתזכיר חוק הגנת הסייבר, **לעיל** ה"ש 461.

1145 אפשר כמובן להעלות על הדעת יישומים ביטחוניים של מערכות נבונות: ממערכות חיזוי מודיעיניות עד כלי נשק אוטונומיים. ואולם אלה הם החריג. בשל מאפייניה של החקיקה האירופית, למשל, נמנעות תקנות הבינה המלאכותית האירופיות מהחלת כללים

ממטרותיו הן ביטחוניות מבצעיות.[1146] לכן שלא כמו מערך הסייבר הלאומי, גוף ביטחוני חשאי למחצה ששייך למשרד ראש הממשלה, רצוי לשייך את רגולטור הבינה המלאכותית למשרד המשפטים או למשרד הכלכלה או למשרד דיגיטל נפרד.

שנית, מבנה מסוג "רגולטור של רגולטורים", כלומר גוף רגולטורי שינחה או יפקח על הרגולטורים המגזריים, הוצע בתזכיר חוק הגנת הסייבר בשנת 2019, ושוורץ אלטשולר וארידור הרשקוביץ העלו לגביו חששות.[1147] עיקרם: החשש מחוסר ההתאמה של מתווה זה לתרבות הפוליטית והרגולטורית בישראל,[1148] שבאה לידי ביטוי בנכונות לשיתוף פעולה בין רגולטורים ובפערים בתרבות הארגונית. הסתייגויות דומות הועלו לאחרונה בתגובה להמלצה להקים רשות לרגולציה – גוף מרכזי לבקרה על רגולציה (regulatory oversight body).[1149] לדברי המסתייגים, רגולטור־על עלול לסרבל ולהאריך את עבודת הרגולציה[1150] ולהחריף מאבקים בין רשויות על כוח וסמכויות.[1151]

משפטיים על מערכות אלו. אין הכוונה שמערכות נבונות ביטחוניות יוחרגו בהכרח מסמכותו של רגולטור בינה מלאכותית, אלא שהן לא יהיו מרכז עיסוקו.

1146 על היעדר ההפרדה בין הביטחוני לאזרחי בתזכיר חוק הגנת הסייבר ראו למשל ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 454, בעמ' 165–169; מרכז המחקר להגנת הסייבר באוניברסיטה העברית, הערות לתזכיר חוק הגנת הסייבר ומערך הסייבר הלאומי התשע"ח–2018 (11.7.2018).

1147 ארידור הרשקוביץ ושוורץ אלטשולר, **שם**, בעמ' 163–165.

1148 למשל, בתחום התקשורת רגולטורים מתקשים לשתף פעולה ביניהם. ראו למשל נתי טוקר "מי הבוס של שוק התקשורת?" **TheMarker** (9.12.2019); תהילה שוורץ אלטשולר "קמצנות רגולטורית" **TheMarker** (3.12.2018). לביקורת על שיתוף הפעולה בין מערך הסייבר לרשות הגנת הפרטיות ראו מבקר המדינה, **היבטים בהגנה על הפרטיות במאגרי מידע** 6, 20–21 (דו"ח שנתי 69ב 2019).

1149 ראו הצוות הבין־משרדי לרגולציה חכמה, **לעיל** ה"ש 1084, בעמ' 105–106.

1150 עמירם ברקת "התוכנית שנחשפה בגלובס עולה מדרגה: מי צריך רגולטור שיילחם ברגולציה?" **גלובס** (6.7.2021).

1151 סמי פרץ "תנסו לא להירדם, זה נושא חשוב מדי" **הארץ** (7.7.2021).

11.3
המלצות

11.3.1. תפקיד הרגולטור וסמכויותיו

א. קידום ותיאום אסדרה של מערכות נבונות בישראל, לרבות התנעת התהליך של גיבוש חקיקה רוחבית וריכוזה.

ב. התוויית מדיניות לפיתוח מערכות נבונות בישראל, להטמעתן ולשימוש בהן. בין השאר יגדיר הרגולטור סטנדרטים בעניין מקור הנתונים לאימון ולבדיקה של מערכות אלו ואיכותם; יקבע מבחנים לאומדן הטיות אלגוריתמיות; יבחן כיצד אפשר להבטיח מידה ראויה של שקיפות, הסברתיות או יכולות ביקורת (auditability)[1152] ונעקבות (traceability) של תהליכים במערכות אלו וכן יקבע סטנדרטים לניהול סיכונים.[1153]

ג. הנחיה מקצועית של רגולטורים מגזריים כדי להבטיח הרמוניזציה של הכללים החלים על פיתוח מערכות בינה מלאכותית, פרישה שלהן ושימוש בהן. הרגולטור ישמש גורם מנחה שיורי, כלומר יהיה ממונה על אסדרת מערכות נבונות פרטיות בתחומים שאין בהם מאסדר מגזרי.

אופי ההנחיה יכול להתבסס על מודל של ניהול סיכונים. לפי מודל זה תהיה ההנחיה של מפתחים, משתמשים או מפיצים של מערכות נבונות ברמת סיכון גבוהה נוקשה ומחייבת יותר מההנחיה החלה על מערכות נבונות ברמת סיכון נמוכה. נוסף על כך, הרגולטור יהיה הגוף המאשר את ההערכה של ארגזי החול הרגולטוריים שיוציאו אל הפועל רגולטורים מגזריים כדי למנוע חשש לניגוד עניינים, להגביר את השקיפות ולעורר מודעות להשלכות החברתיות הרחבות

1152 ראו גם ACM 2017, **לעיל** ה"ש 326, בעמ' 2 (עיקרון 6).

1153 AI HELG 2019, **לעיל** ה"ש 298, בעמ' 13. ל־traceability, ראו גם Defense Innovation Board, **לעיל** ה"ש 342, בעמ' 8 (עיקרון מס' 3).

של המיזם בתום תקופת המבחן שלו.[1154] על הרגולטור יהיה מוטל גם לוודא שהלקחים שהופקו ייושמו כהלכה באסדרה עתידית.[1155]

ד. הנחיה מקצועית של רשויות המדינה בתחום הבינה המלאכותית, ובכלל זה חוות דעת על מסמכי מכרזים ממשלתיים של מערכות בינה מלאכותית, במטרה להבטיח שהמדינה מצטיידת במערכות בינה מלאכותית העולות בקנה אחד עם הסטנדרטים הישראליים והעולמיים בתחום, וכן כדי לייצר בעבור השוק הפרטי תמריצים "רכים" לעבוד בהתאם לסטנדרטים.[1156] זאת לנוכח העובדה שרכש תוכנה יכול להסוות קבלת החלטות מדיניות, למשל בחירת משתני המטרה, נתוני האימון, תכלית האופטימיזציה וסוג המודל.[1157]

ה. ייעוץ בנושאי משפט הבינה המלאכותית. בהיעדר גורם מרכזי המרכז את הייעוץ בנושאי משפט וטכנולוגיה בישראל (הוא היה מכלל תפקידיה של הרשות למשפט, טכנולוגיה ומידע במשרד המשפטים, אך נזנח כשזו הפכה לרשות להגנת הפרטיות), חוות דעתו של הרגולטור תחייב את הממשלה ורשויותיה בנושאים אלו.[1158] אנו מציעים שהרגולטור יהיה גם בעל סמכות להציג עמדה עצמאית בתחומים אלו לכנסת ולבתי המשפט.

ו. מוקד ידע, הדרכה ושיתופי פעולה. הרגולטור יהיה ממונה על קידום האוריינות הדיגיטלית בתחומים אלו בקרב גורמי הממשלה וידון עם בעלי עניין מקומיים ובינלאומיים מתחומי התעשייה, הממשל והאקדמיה.

1154 ראו למשל עמדתם של Eric Brown and Dóra Piroska, *Governing Fintech and Fintech as Governance: The Regulatory Sandbox, Riskwashing, and Disruptive Social Classification*, New Political Economy (2021). פלאטו־שנער (**לעיל** ה"ש 529) נדרשת גם לצורך בשקיפות כלפי המשתמשים במהלך תקופת הניסוי.

1155 Katerina Yordanova, *The Shifting Sands of Regulatory Sandboxes For AI*, CITIP Blog (18.7.201)

1156 לרגולציה באמצעות מכרזים ראו **לעיל**, ה"ש 1011.

1157 Deirdre K. Mulligan and Kenneth A. Bamberger, *Procurement as Policy: Administrative Process For Machine Learning*, 34 Berkeley Tech. L.J. 781 (2019)

1158 השוו יהודה שופמן (יו"ר), **הצוות לבחינת החקיקה בתחום מאגרי המידע: דין וחשבון** 30 (2007); ס' 30 בהצעת חוק הגנת הפרטיות–2019, ודברי ההסבר לו. ארידור הרשקוביץ ושוורץ אלטשולר, **לעיל** ה"ש 403, בעמ' 116; ס' 3(5) לתזכיר חוק הגנת הסייבר.

11.3.2. מקומו הארגוני של הרגולטור - יחידה בתוך רשות הרגולציה

אנו סבורים כי בעת הזאת יהיה נכון שרגולטור הבינה המלאכותית יהיה יחידה בתוך רשות האסדרה. אנחנו סבורים שהוא אינו צריך להשתייך למשרד המדע, שאיננו משרד רגולטורי באופיו ובפעולתו לקידום התעשייה מעדיף אפריורית קידום חדשנות מעקרונות כגון "האדם במרכז" ו"הדמוקרטיה במרכז".

רשות האסדרה הוקמה על פי חוק עקרונות האסדרה[1159] והחלה את פעילותה בשנת 2022. זו יחידה במשרד ראש הממשלה ותפקידיה הם לספק לכל משרדי הממשלה שירותי ייעוץ והדרכה במהלך תהליכי גיבוש הרגולציה, לתת חוות דעת מקצועית על תהליך המחקר המקדים שעשה המשרד הרלוונטי ולוודא שהמשרד מודע לכלל ההשלכות של החלטותיו. כל עוד אסדרת הבינה המלאכותית היא תחום מתפתח, תפקידים אלו של רשות האסדרה הולמים אותה מאוד.

לרשות תפקידים נוספים: לקדם למידה ועבודת מטה ומחקר, להרחיב את גבולות הדמיון הרגולטורי ולהציע פתרונות יצירתיים. לרשות אמורה אף להיות היכרות רוחב עם ארגז הכלים הרגולטורי ועם התכונות של כלים בו; למשל, היא אמורה לדעת מה היתרונות של רישיונות ומתי חובת דיווח עדיפה מחובת גילוי; מתי צריך רישיון קשיח ומתי יש להעדיף מנגנון של אישור בשתיקה. הרשות אמורה לקדם שיטתיות ואחידות בבחינת השפעות הרגולציה, כך שאפשר יהיה להשוות תהליכים בין מגזרים. היא אף אמורה לחתור לאמינות רבה יותר בשימוש בנתונים של הערכת תהליכים רגולטוריים.

יתרון נוסף הוא שגם הרשות היא סוג של "רגולטור של רגולטורים", אך היא רק מייעצת ולא מחייבת. במובן זה היא מתאימה יותר ממבנה מערך הסייבר, שמבוסס על הסדר כופה. הרשות ממילא עובדת מול כלל המשרדים והרשויות העוסקים באסדרה, ולכן הדי־אן־איי הארגוני שלה מתאים. היא שייכת למשרד ראש הממשלה, מה שיכול לסייע לה לפתור בעיות של כפילויות בין רגולטורים ומשרדים בנוגע לחלוקת סמכויות בתחומים שונים של אסדרת בינה מלאכותית.

1159 חוק עקרונות האסדרה, התשפ"ב–2021, ס"ח 80.

היא אמורה לאזן באופן מיטבי בין אינטרסים ציבוריים (כגון אקלים) לבין אינטרסים כלכליים.

החיסרון שבמיקום רגולטור הבינה המלאכותית בתוך הרשות לרגולציה הוא שאף על פי שאחד מתפקידיו הוא לתת ייעוץ משפטי רוחבי, הוא לא יהיה שייך למשרד המשפטים. חיסרון זה אינו משמעותי, שכן עבודת הרשות לרגולציה כוללת מטבע הדברים גם מתן ייעוץ משפטי רוחבי. כמו כן, השנים האחרונות מלמדות שמשרד המשפטים לא השכיל לקבל עליו אחריות של ממש בתחום הייעוץ הרוחבי במשפט ובטכנולוגיה, מה שגרם לעיכובים רבים בחקיקה ולחוסר ודאות ושקיפות.

ההתמודדות עם אתגרי הגמישות, עם הסביבה הרגולטורית העמוסה ועם הטכנולוגיה מצריכים הקצאת משאבים ראויה, שתאפשר ליצור תקנים הן למומחים טכניים הן למומחים במשפט ובמדיניות. אנו מציעים כי גם בהיעדר חוק בינה מלאכותית תתוקצב רשות האסדרה בהחלטת ממשלה כדי שתוכל להקים את היחידה של רגולטור הבינה המלאכותית.

11.4 פיתוח אסדרת בינה מלאכותית בתקופת ביניים כבסיס לאסדרה עתידית

כפי שציינו לעיל, חקיקה לא נוצרת בִּן לילה, וגם אם טרם הבשילה העת לחוקק בישראל חוק בינה מלאכותית מן הראוי לעקוב אחר התפתחויות גלובליות רלוונטיות, לייצר ידע ארגוני ולגבש עמדות בעניין ההסדר הרצוי. בחלק זה נציע שלוש הצעות בנוגע לתקופת הביניים, בטרם תגובש חקיקה בישראל.

11.4.1 יצירת "אקוסיסטם קדם־אסדרתי"

אנו סבורים כי הקמת יחידה שתעסוק כבר כעת באסדרה של בינה מלאכותית תוכל לפרסם בשלב הביניים הנחיות וגילויי דעת שישמשו תשתית לאסדרה עתידית. כמתואר בתרשים שלהלן, אנו סבורים שראוי ליצור אקוסיסטם קדם־אסדרתי שבמסגרתו יפרסמו הרגולטור הייעודי, רגולטורים ענפיים ורגולטורים משלימים (כגון הרשות לפרטיות והרשות לתחרות) הנחיות

וגילויי דעת; והם בתורם ישמשו את התעשייה ושחקנים קונקרטיים בה, וגם את בתי המשפט. מסמכי הנחיות אלו יוכלו לשמש את התעשייה כדי לקבוע סטנדרטים של אסדרה עצמית, מתוך הנחה שרגולציה עתידית תהיה דומה להנחיות. הם יוכלו לשמש גם את חברות התקינה כדי ליצור את מסגרות התקינה.

סביב האקוסיסטם הקדם־אסדרתי יתפתחו תהליכים של חידוד דפוסי הבקרה (auditing) בהקשרים שונים, ייערכו תהליכי חשיבה בעניין סטנדרטים של זהירות ראויה וייעשה מאמץ סביר לקדם אחריות בפיתוח וביישום של מערכות לומדות; ייבנו ארגזי חול וייערכו ניסויים רגולטוריים שונים; ויקודם מאגר של מומחים שיוכלו לשמש את התעשייה, את הרגולטורים ואת בתי המשפט בהתמודדות עם סוגיות חדשות ומורכבות. במקביל, התעשייה והשחקנים יוכלו לתת משוב לגופים הרגולטוריים השונים וכך לטייב את ההנחיות. נוסף על כך, בהיעדר אסדרה, בתי משפט יוכלו להשתמש בהנחיות כהשראה לפרשנות בסכסוכים המובאים לפניהם. פסקי הדין יוכלו גם הם להעשיר את גוף הידע ולאפשר לשחקנים וגם לרגולטורים לטייב ולשייף את הנחיותיהם לקראת גיבוש אסדרה. בתרשים להלן מתואר האקוסיסטם הקדם־אסדרתי.

תרשים 6
האקוסיסטם הקדם־אסדרתי המוצע

11.4.2. קידום תהליכי תקינה

בתעשיות רבות תקני בטיחות הם כלים מרכזיים להערכת סיכונים במוצרים חדשים – מזון, תרופות, מטוסים מסחריים, מכוניות וכיו"ב. תקינה (סטנדרטיזציה) בתחום הבינה המלאכותית מאפשרת יכולת תאימות תפעולית (inter-operability) של מערכות ורכיבים, אימות וניידות של מודלים ונתונים, וממשק יעיל והעברה של טכנולוגיות בין פלטפורמות, סביבות ויישומים שונים. לכן יש לה פוטנציאל להפחית עלויות פיתוח ויישום של מערכות ולקדם ולטפח שיתוף פעולה בין בעלי עניין. נוסף על כך, תהליכי תקינה יכולים לסייע לארגונים לפתח אסטרטגיות פיקוח על פרישׂה אחראית של בינה מלאכותית וכך ליצור את מסגרת הערכים הסובבת כל הטמעה של טכנולוגיה חדשה, בייחוד במרחב שאין בו עדיין רגולציה. המכון הלאומי האמריקני לתקנים וטכנולוגיה (NIST), ארגון התקינה הבינלאומי ISO, הוועדה הבינלאומית לאלקטרו־טכניקה (IEC) והוועדה האירופית לתקינה אלקטרו־טכנית (CENELEC) נעשו בשנים האחרונות שחקני מפתח בפיתוח תקנים בתחום הבינה המלאכותית.

בהסתמך על ניסיון העבר בתעשיות אחרות, יש להניח שכללי תקינה ימלאו תפקיד מכריע ביצירת סביבת הכללים. כללים אלו יוכלו להפוך אחר כך לעקרונות מנחים ביצירת רגולציה – הן בהיבט של יצירת טרמינולוגיה,[1160] הן בהיבט של מתודולוגיות הערכה וניהול סיכונים,[1161] בקרות אבטחה והגנת פרטיות, הן בהיבט של הטמעת ערכים ועקרונות חברתיים. למשל, כאשר NIST פיתחה את מסגרת התקינה של ניהול סיכוני בינה מלאכותית היא הציעה, מלבד מסגרות לניהול סיכונים ומתודולוגיות הערכה, גם כללי שקיפות והסברתיות.[1162] יתר על כן, למוסדות תקינה יכולה להיות גם תרומה של ממש ליצירת תהליכים רגולטוריים ולהטמעתם. כך קרה בתהליך חקיקת חוק הבינה המלאכותית של האיחוד האירופי, כשהאיחוד נעזר ביכולותיו של מכון התקינה האירופי CENELEC,

1160 ראו למשל בתקן ISO לבינה מלאכותית: ISO/IEC JTC 1/SC 42 (2023)

1161 ראו למשל: *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST (2023). נוסף על כך, בתקן AI NIST SP 800-53, המיועד לאבטחת מידע, יש בקרות אבטחה ופרטיות שתוכננו במיוחד למערכות AI.

1162 ראו גם את התקינה של ISO, AI ISO/IEC 20547, המתמקדת בהגדרת הדרישות למערכות AI אמינות ונדרשת להיבטים כגון יכולת הסבר, הוגנות וחוסן; וכן את תקן 1 ISO/IEC 23894, שמשמש מסגרת פיקוח על בינה מלאכותית מתוך התייחסות לנושאים כגון אחריותיות, שקיפות ופיקוח אנושי.

שהוא עצמו עבד בשיתוף פעולה הדוק עם ISO ו־IEC כדי לנסח את החקיקה. מוסדות האיחוד קבעו מראש כי מכון התקינה האירופי ייצור תקינה שתכליתה הטמעה הרמונית של העקרונות שנקבעו בחקיקה.[1163]

אנו מציעים לתרגם כבר כעת מסמכי תקינה מרכזיים בתחום הבינה המלאכותית; לבצע התאמה שלהם לתנאים הכלכליים, המשפטיים והרגולטוריים במדינת ישראל; ולהפיץ אותם בקרב השחקנים השונים. כל זה כדי לאפשר דיון ציבורי בעקרונות שהם מבטאים, דיון טכנולוגי בישימות שלהם ודיון רגולטורי בשאלה עד כמה יהיה אפשר לאכוף אותם אם יתורגמו לחקיקה. גם במדינות אחרות בעולם יש תקינה, ואנו סבורים שזהו השלב הראשון וההכרחי ביצירת מצע לאסדרה עתידית. גם אם אי־אפשר לאכוף אותה, קיומה מאפשר לתעשייה להבין את כיווני החשיבה של הרגולטור ולהתארגן בהתאם, לבעלי עניין לעורר סביבה דיון ציבורי ולרגולטור לבצע תהליכי למידה ואימון.

11.4.3. "ארגזי חול" רגולטוריים

למושג "ארגז חול רגולטורי"[1164] כמה משמעויות. המרכזית והמקורית שבהן היא יצירת סביבה מבוקרת שבה חברות יכולות לבדוק את המוצרים או את השירותים שלהן תחת פיקוח רגולטורי, אך רשאים לחרוג מהרגולציה או להגמיש אותה. לעיתים מדובר במשטר משפטי תחום בזמן ובהיקף, בניהול גורם מוסמך, שמשתתפיו זוכים לפטור מהוראות מסוימות בדין או להחרגות מסוימות והם יכולים ליישם טכנולוגיה או מודל עסקי בניגוד להוראות אלו. למשל, תקנות הבינה המלאכותית האירופיות פוטרות את החברות במסגרת

1163 ראו את תיאור התהליך באתר CENELEC; וכן את תוכנית העבודה של JTC21 מחודש מרץ 2022.

1164 ראו ההגדרה שהציעו Vladislav O. Makarov and Marina L. Davydova, *On the Concept of Regulatory Sandboxes*, *in* "Smart Technologies" for Society, State and Economy 1014 (Elena G. Popkova and Bruno S. Sergi eds., 2021). ראו גם Dirk A. Zetzsche et al., *Regulating a Revolution: From Regulatory Sandboxes to Smart Regulation*, 23 (1) Fordham Journal of Corporate & Financial Law 31 (2017-2018); ראו גם *Artificial Intelligence Act and Regulatory Sandboxes* (EU Parliamnet Briefing), Think Tank (17.6.2022)

ארגז חול רגולטורי[1165] מהחובה לציית לעקרון צמידות המטרה שבתקנות הגנת המידע[1166] ומאפשרות להן להשתמש בנתונים שנאספו למטרות אחרות כדי לפתח בינה מלאכותית חדשנית.[1167] בפעמים אחרות מדובר בליווי רגולטורי שבו מוסכם כי אם המפוקח פועל בתום לב לא ייאכפו לגביו הסדרים מסוימים.

מטרת ארגז החול היא לאפשר היערכות, איסוף נתונים ומדידה של ביצועים, בטיחות, חוסן ופגיעות של מודלים ומוצרים, בדיקת דיוק והטיות שלהם, וכן את עמידתם בעקרונות אתיים אחרים, לעיתים עקב קבלת משוב על המוצרים מצרכנים, והכול לפני יישום נרחב של מודלים ומוצרים וללא הסיכון של פגיעה בצרכנים או הפרת חוק ורגולציה. ארגזי חול נועדו לעודד חדשנות במובן זה שהם מספקים מרחב בטוח לניסויים – דרך זו מאפשרת לארגונים, לחברות ולעסקים לנווט את דרכם בין דרישות רגולטוריות ולזהות בעיות פוטנציאליות לפני יישום רחב יותר של מוצריהם. לכן הם נעשו נפוצים גם בהקשרים של מוצרי סייבר וממשל נתונים.[1168] יתרונות נוספים של ארגזי חול: שילוב כוחות בין המאסדר ובין המפוקחים, במקום המצב הרגיל שהוא היררכי במהותו; האפשרות לשלב בתהליך הניסוי בעלי עניין שונים, כגון צרכנים ולקוחות, אזרחים מקבוצות שונות ושחקנים חדשים, וליצור מערכת המבוססת על אחריות ואמון ולא רק על פיקוח והגנה.

בישראל הוצעו ארגזי חול רגולטוריים בתחום הפיננסי[1169] ובתחום האנרגיה הירוקה,[1170] למשל, ונחקקו בפועל בתחום הטיס[1171] ובתחומים נוספים. בתחום

1165 ס' (6)–(5)53 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

1166 ראו ס' (4)6 של התקנות הכלליות בדבר הגנת מידע (GDPR).

1167 ס' (a)(1)54 להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

1168 ראו למשל את ארגז החול הרגולטורי לניסויים בתחום הנתונים בסינגפור. *Data Regulatory Sandbox*, Infocomm Media Development Authority

1169 לאחרונה פורסמה הצעת חוק להקמת ארגז חול רגולטורי בתחום הפינטק (FinTech) בישראל. ראו הצעת חוק לעידוד פיתוח טכנולוגיה בתחום הפיננסי בישראל, התשפ"א–2021, ה"ח הממשלה 204; רות פלאטו־שנער "פינטק בישראל: תוכנית 'ארגז החול' הרגולטורי" **ICON-S-IL BLOG** (30.6.2021).

1170 ראו אילנה קוריאל "חדש: מועצה שתפתור חסמי רגולציה כדי לקדם סטארט־אפים ירוקים" **ynet** (10.11.2021).

1171 ראו ס' 16 לתקנות הטיס (נוהלי תיעוד כלי טיס וחלקיהם), תשל"ז–1977.

התחבורה נקבע הסדר מפורט בעניין הפעלה ניסיונית של רכב אוטונומי ובסעיף המטרה שלו יש הסבר ברור של תפיסת ארגז החוק הרגולטורי: "מטרתו של סימן זה קביעת הסדרים שיאפשרו הפעלה של רכב עצמאי, בדרך, למטרת ביצוע ניסוי, בלא נהג, בשמירה על הבטיחות ושימוש במגוון טכנולוגיות, כדי להביא לגיבוש תשתית ידע לגבי בטיחותו של הרכב העצמאי, יכולתו להשתלב בין עוברי הדרך ולתת שירות לנוסעים והשפעתו על התנועה בדרך, לאפשר הנגשה של הידע האמור לציבור ולבסס את אמון הציבור בו".[1172]

לשימוש בארגזי חול יש תכלית נוספת: לסייע בפיתוח אסדרה חדשה במקום שאין אסדרה קיימת, תהליך המכונה regulatory prototyping. בהיעדר מסגרות רגולטוריות ברורות בתחומי שיפוט מסוימים יכולים ארגונים, חברות ועסקים לזהות פערי אסדרה ולא לדעת אם על המוצרים שלהם חלות תקנות, אילו תקנות וכיצד לעמוד בהן. במקרה כזה הם יפנו לרגולטור ויציעו לו להקים ארגז חול שתכליתו לבדוק כיצד צריך לעצב רגולציה עתידית. ארגזי חול יכולים להיות מועילים גם במקום שבו תהליכים של הערכת השפעות וניהול סיכונים אינם מועילים. למשל, בתחום הבינה המלאכותית היוצרת המבוססת על מודלים בסיסיים שעליהם מורכבים מוצרים ויישומים, אין טעם לדרוש ממי שמשווק יישומים כאלה לבצע הערכת סיכונים של המודל הבסיסי שהוא לא יצר. במקרה כזה ייתכן שיהיה אפשר להפיק תועלת מארגז חול רגולטורי שידגים את החששות דווקא בהקשרי היישומים וייצור מסגרת לאסדרה עתידית. יש שמרחיבים את תחולת המושג אפילו יותר וכוללים בו שולחנות עבודה רב־מגזריים שדנים בהם ברעיונות חדשים לרגולציה עתידית.

יש המפרשים את המושג "ארגז חול רגולטורי" ככלי שתכליתו להעצים את יכולות גופי האסדרה בתחומי ההנחיה, הפיקוח והאכיפה. פרשנות כזאת אמורה לאפשר לרגולטורים להבין את מידת התאימות של הרגולציה לטכנולוגיה, לאפשר להם למידה ולכן גם למזער את הסיכון לפיגור רגולטורי.

1172 ראו סימן ג3 (הפעלה ניסיונית של רכב עצמאי בלא נהג) פקודת התעבורה (תיקון 130) התשפ"ב 2022, בסעיף 14כד.

תקנות הבינה המלאכותית האירופיות מגדירות שתי מטרות מרכזיות להקמת ארגזי חול רגולטוריים: (1) לסייע לרשויות באמצעות מתן קווים מנחים בעניין הוראות התקנות לספקי מערכות בינה מלאכותית; (2) לסייע ללמידה רגולטורית בסביבה מבוקרת.[1173] כלומר התפיסה האירופית רואה בארגזי החול כלי לסיוע בהטמעה של חוקי הבינה המלאכותית ובפרשנותה. לפיכך ארגזי החול הם החוליה המקשרת בין חקיקה חדשה לבין התקינה שתסייע ביישום שלה; תכליתם לסייע בקידומה של אכיפה אפקטיבית, הרמוניזציה וודאות משפטית.[1174] לכן התקנות מורות על הקמה של ארגז חול רגולטורי אחד לכל הפחות בכל מדינה חברה; ועליו להיות פעיל לכל המאוחר ביום כניסת התקנות לתוקף.[1175] לנוכח החשש שהובע במהלך החקיקה כי עמימות בעניין המתווה של ארגזי חול עלולה להביא לידי פרגמנטציה רגולטורית באיחוד,[1176] נקבע בחקיקה הסופית כי הנציבות תאמץ חקיקה המגדירה את המתווים השונים להקמתם, פיתוחם והטמעתם של ארגזי חול רגולטוריים וכן את דרכי הפיקוח עליהם.[1177]

בחודש מרץ 2021, טרם פרסום התקנות המוצעות, החלה נורווגיה להפעיל ארגז חול רגולטורי כדי לסייע לחברות המפתחות בינה מלאכותית להתמודד עם הסביבה הרגולטורית המורכבת, ובפרט עם הקושי לציית להוראות התקנות הכלליות בדבר הגנת מידע (GDPR). שלא כמו המודל של התקנות האירופיות, ארגז החול הנורווגי אינו מגמיש את עקרון צמידות המטרה אלא נועד לאפשר לרגולטור לסייע ליזמים לפתח מערכות בינה מלאכותית המצייתות ל־GDPR כדי להקנות מידה מסוימת של ודאות משפטית.[1178]

1173 ס' 53(1d) להצעה המתוקנת, **לעיל** ה"ש 57.

1174 פס' 71 למבוא להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

1175 ס' 53(1) להצעה המתוקנת, **לעיל** ה"ש 57.

1176 Sofia Ranchordas, *Experimental Regulations for AI: Sandboxes for Morals and Mores* (12.7.2021)

1177 ס' 53a להצעה המתוקנת, **לעיל** ה"ש 57.

1178 Dan McCarthy, *To Regulate AI, Try Playing in a Sandbox*, Tech Brew (26.5.2021); *Sandbox for Responsible Artificial Intelligence*, Datatilsynet

השימוש בכלי של ארגז חול רגולטורי מעורר גם חששות. ראשית, הוא דורש הקצאה של משאבים לתשתיות נחוצות. הוא דורש גם כוח אדם מתאים, שכן נדרשות מומחיות טכנולוגית במנגנוני הערכת מודלים ובמתודולוגיות של ממשל נתונים והערכתם; וכן יכולת פרשנות ויכולת הסבר ומומחיות משפטית בשאלות של חבויות, קביעת אחריות משפטית והיבטי רגולציה נוספים. כל זאת, נוסף על מומחיות בענפים ספציפיים כגון בריאות, פיננסים ותחבורה. לכן בתקנות האירופיות נקבע כי על הרשויות המקימות ארגזי חול רגולטוריים להבטיח הקצאת נאותה של משאבים כדי שיוכלו לפעול בהתאם להוראות הדין.[1179]

שנית, יש חשש שארגזי חול רגולטוריים יובילו להתחמקות מרגולציה רוחבית, כמו הגנת פרטיות ואבטחת נתונים. לכן נקבע בתקנות האירופיות כי שימוש לצורכי ניסוי במידע שנאסף למטרות אחרות יכול להיעשות רק בכמה תנאים: מערכת הבינה המלאכותית שמבקשים לנסות נועדה להגן על אינטרסים ציבוריים חיוניים בתחום הביטחון הציבורי (כגון זיהוי עבירות פליליות ומניעתן), בתחום הבריאות (לרבות מניעה של מחלות, בקרה עליהן וטיפול בהן) או בתחום ההגנה על איכות הסביבה; הנתונים הגולמיים נחוצים ואי־אפשר להשתמש במידע סינתטי או במידע שעבר התממה; יש מנגנונים לניטור סכנות לזכויות יסוד של מושאי מידע שמתעוררות במהלך הניסוי; המידע לא מועבר או מונגש לצד שלישי כלשהו; עיבוד המידע במסגרת ארגז החול הרגולטורי אינו משפיע על החלטות שיתקבלו בעניין מושאי המידע; כל מידע אישי שמעובד במסגרת ארגז החול הרגולטורי נשמר בסביבה מאובטחת, נפרדת ומבוקרת שהגישה אליה מוקנית למורשים בלבד; כל פעולות עיבוד המידע האישי מתועדות; כל המידע האישי המשמש בארגז החול הרגולטורי מבוער עם סיום הפעילות; תוצאות הניסוי ותיעוד מלא של הבדיקות, האימון, התהליכים וההיגיון שהנחה אותם נשמרים.[1180]

שלישית, יש חששות מתחום המשפט המינהלי, משום שארגזי חול פועלים מטיבם על בסיס פרטני ויוצרים קשר ישיר וייחודי בין רגולטור למפוקח,

1179 ס' 53(1c) להצעה המתוקנת, **לעיל** ה"ש 57.

1180 ס' 54(1)(b)–(j) להצעת תקנות הבינה המלאכותית האירופיות, **לעיל** ה"ש 53.

ששונה מן הפרוצדורה הרגילה של יישום ואכיפה של הרגולציה. מאפיין זה עלול אף לעורר חשש מאפליה ופגיעה בשוויון. לכן בתקנות האירופיות מוצעות הוראות מיוחדות לעידוד השתתפותם של ספקים קטנים וחברות הזנק בארגזי חול רגולטוריים.[1181] נוסף על כך, אם אין הקפדה על הליכים שקופים ומוגדרים בבירור, ועל מנגנונים לשיתוף בעלי עניין, הפגיעה של ארגזי החול באמון, באמינות ובלגיטימיות של הרגולטור עלולה לעלות על התועלת שבו. יש גם צורך לייצר הערכה ולמידה תקופתיות כדי לבחון את היעילות וההשפעה של ארגזי חול וכדי למנוע חששות מניצולם לרעה או משימוש בהם כמנגנון התחמקות מעול הרגולציה.[1182]

אנו סבורים שהשימוש בארגזי חול רגולטוריים שתכליתם לייצר החרגה מכללים או להציע ליווי שיש בו הסכמה להימנע מאכיפה של כללים יכול להיות אמצעי יעיל בתקופת הביניים בישראל. בשנים האחרונות נשמעו בישראל קולות בדבר הצורך לחוקק לארגזי חול רגולטוריים חוק מסגרת שיקבע עקרונות כלליים בעניין היקף ההחרגה מאסדרה קיימת (למשל אם אפשר לחרוג גם מחקיקה ראשית או רק מתקנות), מנגנוני אכיפה במקרים של הפרת הסכמי החרגה (למשל אם די בדיווח או שיש מקום לאכיפה מינהלית או פלילית), פרוצדורה (למשל דרכי הגשת בקשות, מגבלות זמן על שימוש בארגזי חול ונקודות יציאה). ואולם קבועי הזמן לחקיקה במדינת ישראל מלמדים כי הסיכוי לחקיקה כזאת בעתיד הקרוב נמוך ועל כן ספק אם אפשר לסמוך עליה, בייחוד בתקופת הביניים הקרובה. לכן אנו מציעים שאסדרת ארגזי החול תתבצע ברמה הענפית, אבל תיקבע בהנחיות חובת דיווח של הרגולטור המפעיל ארגז חול ליחידה העוסקת באסדרת בינה מלאכותית. המטרה – לאפשר למידת עמיתים רוחבית ולהעשיר את האקוסיסטם הקדם־אסדרתי. אנו סבורים שראוי שרגולטורים ענפיים המקימים ארגזי חול רגולטוריים יעשו זאת בקפידה ויעמדו בתנאי המשפט המינהלי. הטעם לכך הוא שמלכתחילה התרבות הרגולטורית בישראל ידועה

1181 **שם**, ס' 55.

1182 להרחבה ראו גם גיא מור וענת גרימלנד "רגולציה ניסיונית: כלים לשיפור איכות הרגולציה באמצעות תחולה מוגבלת" (נייר מדיניות, מאי 2019).

בגמישותה ובנטייתה לאפשר "חוזים רגולטוריים" עם מפוקחים, באופן שיכול ליצור העדפה בלתי ראויה של שחקנים מסוימים.

אם תהיה בעתיד בישראל חקיקה רוחבית בעניין בינה מלאכותית יהיה מקום לעסוק בשאלה אם יש צורך בארגזי חול כדי לקדם את תהליכי ההטמעה שלה, כפי שמוסדר בתקנות האירופיות. בשלב הביניים שבו אנו נמצאים דבר זה אינו נצרך.

נוסף על כך אנו מציעים ליזום "ארגזי חול מוסדיים" שתכליתם לנסות מוצרים מבוססי בינה מלאכותית בהקשרים ספציפיים, דוגמת בתי משפט או בתי חולים. מוסדות אלו יכולים לפרסם קולות קוראים להגשת בקשות להתנסות במוצרים או בשירותים מבוססי בינה מלאכותית הנוגעים לתחומי העיסוק שלהם. לשם כך עליהם ליצור מסגרות שיבחנו את מטרותיהן של המערכות, את תהליכי הפיקוח עליהן, את התועלת בהן ואת הפגמים המתגלים בהן, ולפרסם דוחות הערכה לעיון הציבור. במדינת יוטה מפעילה בשנים האחרונות יחידת החדשנות של בית המשפט העליון המדינתי ארגז חול מוסדי כזה. הצלחתו הייתה גדולה כל כך עד שבפברואר 2023 נאלצה היחידה להפסיק לקבל בקשות בשל עומס יתר.[1183]

לסיכום חלק זה, התרשים שלהלן מתאר את תהליך האסדרה המוצע בעת הזאת: שרטוט מפת דרכים וטיפולוגיה של הנושאים הטעונים תשומת לב; תרגום מסמכי תקינה ושילובם במסגרת הקדם־אסדרתית; העשרת המסגרת הזאת באמצעות ארגזי חול רגולטוריים; עבודה על עיצוב אסדרה וחקיקה רלוונטיות.

1183 *May 2023 Activity Report*, The Office of Legal Services Innovation, An Office of the Utah Supreme Court

תרשים 7
תהליך האסדרה המוצע של בינה מלאכותית

1 יצירת מפת דרכים ומפת צרכים

- טרמינולוגיה והמשגה
- היבטים אופרטיביים (ניהול סיכונים, משילות, הערכת השפעות, סטנדרטים לניהול ואיכות נתונים, תקינה לגבי איכות של מודלים, ניהול מוצרים)
- היבטים הנדסיים (למשל סטנדרט למודלים מבוססים רשתות נוירונים, סטנדרט לניתוח שפה טבעית)
- היבטים חברתיים ואתיים (הטיות, שקיפות, הסברתיות, אמון)
- סטנדרטים קונקרטיים למודל ספציפי (למשל מודל שמטרתו שכנוע אנושי, מודל למעקב ביומטרי)

2 שאילה ותרגום לעברית של סטנדרטים שהתפתחו בעולם

ISO, IEC, NIST, CENELEC, JTC21

3 שילוב עם ערכים ועקרונות מתאימים לישראל, עם אסדרה קיימת ועם מוסדות קיימים

- משפט חוקתי ומינהלי
- משפט אזרחי (נזיקין והגנת הצרכן)
- חקיקת פרטיות, סייבר, זכויות יוצרים
- חוק יסודות האסדרה
- חוק ארגזי החול
- רגולציה ענפית

4 אקוסיסטם קדם־אסדרתי

- תקינה
- הנחיות וגילויי דעת
- פסיקה
- ארגזי חול רגולטוריים

5 תהליך אסדרה וחקיקה

- חקיקה רוחבית
- אסדרה ענפית
- חקיקה משווה
- משפט בינ"ל

פרק שנים עשר

סיכום

היא מלאכותית אבל היא "בינה", היא ממוחשבת אבל היא "ראייה", היא שייכת למכונה אבל היא "למידה", היא רשת אבל היא "נוירונית". אנחנו דורשים שהיא "תתאמן" היטב, שיהיה לנו "אמון" בה ושלא יהיו לה "הטיות". וכמובן – אנחנו מבקשים שהיא "תסביר" את עצמה כך שנבין.

בינה מלאכותית היא אגד של טכנולוגיות מסעירות ומרתקות שמתפתחות בשנים האחרונות בקצב מסחרר. מינוף יתרונותיהן והתמודדות עם האתגרים המשפטיים שהן מעמידות לפנינו מחייבים ערנות מצד ממשלות ורגולטורים, האקדמיה, החברה האזרחית וגם התעשייה. לא פחות מכך הן מחייבות ערנות מצד מי שיהיו מושאי ההחלטות האלגוריתמיות, המשתמשים במוצרים מבוססי בינה מלאכותית

וגם המושאים של שינויים אפשריים ברמה האישית והחברתית עקב שימושים אלו. לכולם יש מקום סביב השולחן, וחשיבות מקומם תיגזר מן הערנות שלהם, מנכונותם להרחיב את האוריינות הדיגיטלית שלהם ומהמאמץ שישקיעו בכל אלה בחשיבה ובמעשה.

מטרתו של ספר זה הייתה לספק לכלל השחקנים את האוריינות הדיגיטלית הדרושה להם כדי לקבל בשנים הקרובות החלטות בנוגע לאסדרה של בינה מלאכותית. הדעת נותנת שנזדקק לעדכון מתמשך של הידע, ההבנה ויכולת הפרשנות שלנו בתחום זה. אבל לנוכח הניסיון המצטבר שרכשנו במהפכות הטכנולוגיות שהתרחשו בעשרים וחמש השנים האחרונות, ברור לנו כיום שהתפתחות ללא כל אסדרה איננה הגיונית.

האסדרה המשפטית של מערכות בינה מלאכותית ברחבי העולם עודנה בחיתוליה. במחקר זה נסקרו הצעות חוק שונות, אך רובן יבשילו לכלל חקיקה רק בשנים הבאות ועדיין לא ידוע כמובן מה יהיו השלכותיהן בפועל. אך אין פירוש הדבר שיש לשבת בחיבוק ידיים – בזמן ההמתנה להתפתחויות הרגולטוריות מעבר לים יכולים מעצבי המדיניות בישראל להכשיר את הקרקע לחוק בינה מלאכותית ישראלי. אפשר להתחיל לעגן בחקיקה זכויות של מושאי החלטות אוטומטיות ולגבות אותן בסעדים ובהתאמות פרוצדורליות. אפשר לנסות לגבש עמדות בעניין איסורים קטגוריים על פיתוח מערכות בינה מלאכותית ברמת סיכון גבוהה ועל שימוש בהן. איסורים זמניים כאלו יקנו למחוקק די זמן לעצב את מודל האסדרה המתאים לישראל. אפשר להקים ארגזי חול רגולטוריים, שבמסגרתם ילווה רגולטור מגזרי פיתוח של מערכות בינה מלאכותית בתחום שהוא מאסדר במטרה לגבש מדיניות אסדרה מתאימה.

בעיקר חשוב לגבש צוות עבודה בין־משרדי שיהיה מופקד על ניסוח הצעת החוק ותיקוני חקיקה אחרים בהסתמך על לקחים שהופקו ממקרי הבוחן בארגזי החול הרגולטוריים, על עקרונות אתיים מוסכמים מראש ועל דוגמאות מהדין הזר (ואף על יישומם בפועל). תהליך זה יכול להניב מודל רגולטורי בשל בפרק זמן סביר.

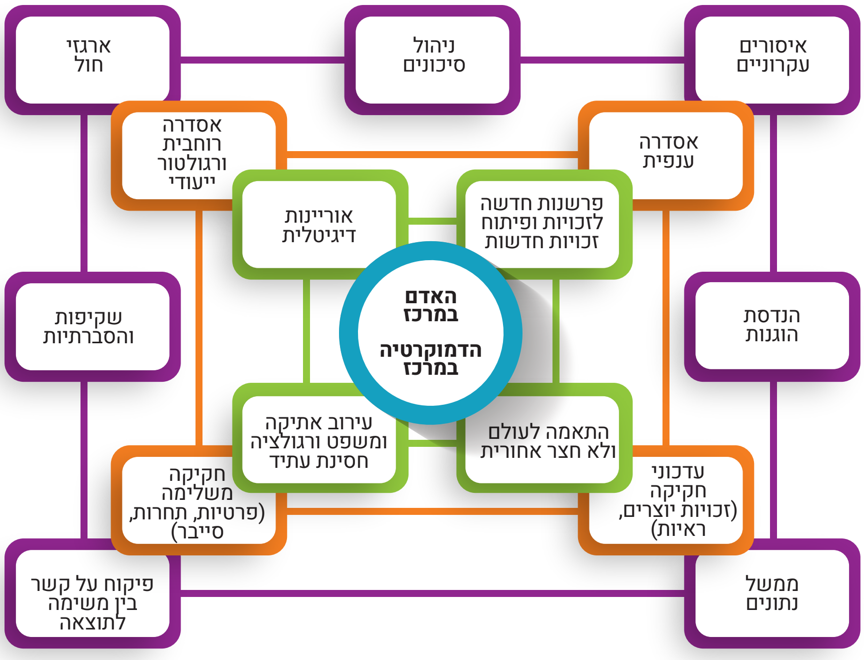
בשנים הקרובות ילך ויגבר השימוש במערכות אלגוריתמיות, ובהן במערכות לומדות ובמערכות בינה מלאכותית, ויחדור לכל תחומי חיינו. שמיכת הטלאים

המשפטית שמאסדרת את התחום הדיגיטלי בישראל קצרה מכדי לכסות את כל הנדרש כדי להבטיח שתושבי המדינה ואזרחיה יהיו מוגנים מפני הסיכונים הגלומים במערכות אלו. קידומה של תשתית משפטית ומוסדית לבינה מלאכותית בטוחה היא צו השעה.

בתרשים שלהלן מסוכמות המלצותינו בנוגע לאסדרה של בינה מלאכותית בישראל. במרכזו, שני העקרונות המרכזיים האמורים להדריך את כלל פעולות האסדרה בתחום: האדם במרכז והדמוקרטיה במרכז. סביבם עקרונות כלליים נוספים: חשיבותה של האוריינות הדיגיטלית של מקבלי ההחלטות; הצורך להתאים את הרגולציה הישראלית למתרחש בעולם – לא להקדים אותה, אך גם לא ליצור פערים רגולטוריים שיהפכו את מדינת ישראל לחצר אחורית במה שנוגע לטיפול בנתונים ולשימושי בינה מלאכותית; הצורך לתת פרשנות חדשה לזכויות אדם קיימות, כגון הזכות לפרטיות וחופש הביטוי, והחשיבות של פיתוח זכויות אדם חדשות, כגון הזכות לקבל שירות מבן אנוש; וכן האתגר הטמון בפיתוח רגולציה חסינת עתיד, כלומר כזאת שאין בה התייחסות לטכנולוגיה ספציפית אלא לעקרונות, והחשיבות של שילוב עקרונות אתיים ועקרונות משפטיים. סביב העקרונות האלה שתי רצועות קונקרטיות: האחת, הנוגעת למסגרת המוסדית של הרגולציה של בינה מלאכותית, מציעה לשלב בין אסדרה ענפית לאסדרה רוחבית ומצביעה על הצורך לראות ברגולציה של בינה מלאכותית חלק אחד בתצרף אסדרתי רחב יותר הכולל חובות לחקיקה עדכנית ולעדכון חקיקה קיימת. השנייה, החיצונית, עניינה הצעות קונקרטיות למיני התערבות לצורך פיקוח על בינה מלאכותית, ובהן ההצעה לפתח מתודות לניהול סיכונים, להחריג (החרגה מוגבלת) איסורים עקרוניים על שימושים מסוימים, הצעה לארגזי חול רגולטוריים, משילות נתונים, הנדסת הוגנות, שקיפות ופיקוח על הקשר בין משימה לתוצאה במערכות של למידה לא מפוקחת.

תרשים 8

סיכום מודל האסדרה המוצע לבינה מלאכותית



תקופות של מהפכות טכנולוגיות גורמות לאנשי רוח לומר, כפי שכתב צ׳רלס דיקנס בפתיחה הנודעת לרומן ״בין שתי ערים״, שעלילתו מתרחשת בתקופת המהפכה הצרפתית, ״היה זה הטוב בזמנים, היה זה הרע בזמנים״. המשורר והמחזאי הגרמני ברטולט ברכט כתב בתחילת המאה העשרים את השורות האלה:

הזמנים החדשים / ברטולט ברכט

זְמַנִּים חֲדָשִׁים אֵינָם מַתְחִילִים בְּבַת־אַחַת.
סָבִי כְּבָר חַי בַּזְּמַן הֶחָדָשׁ.
נֶכְדִּי בְּוַדַּאי עוֹד יִחְיֶה בַּיָּשָׁן.

בָּשָׂר חָדָשׁ אוֹכְלִים בְּמַזְלְגוֹת יְשָׁנִים.

לֹא כְּלֵי־הָרֶכֶב הַנּוֹסְעִים־לְבַד עָשׂוּ זֹאת
לֹא הַטַּנְקִים
לֹא הַמְּטוֹסִים מֵעַל לְגַגּוֹתֵינוּ עָשׂוּ זֹאת
וְלֹא הַמַּפְצִיצִים.
מִן הָאַנְטֶנּוֹת הַחֲדָשׁוֹת בָּא הַטִּמְטוּם הַיָּשָׁן.
הַחָכְמָה נִמְסֶרֶת הָלְאָה מִפֶּה אֶל פֶּה.[1184]

ברכט, שכמו מרפרר לדיקנס, ניסה ללמדנו שלא הטכנולוגיה היא שאחראית לתוצאותיה. לא "כלי־הרכב הנוסעים־לבד עשו זאת", לא הטנקים ולא המפציצים. אלה היו הכוחות האנושיים. וכיוון שזמנים חדשים אינם מתחילים בבת אחת חשוב כל כך לחקור אותם, לעמוד על מקורותיהם ולהגות באחריות שצומחת מהם. ולשם כך עלינו להתבונן גם בטכנולוגיה. האם נצעד היישר לגן עדן או שמא היישר לצד השני?

1184 ברטולט ברכט **גלות המשוררים** 201–202 (מגרמנית: בנימין הרשב, ספרי סימן קריאה/הקיבוץ המאוחד, 1978).

THE ISRAEL
DEMOCRACY
INSTITUTE

Human, Machine, State

Toward the Regulation of Artificial Intelligence

Amir Cahane | Tehilla Shwartz Altshuler

Abstract

In recent years, artificial intelligence (AI) systems have increasingly become part of the fabric of daily life. They recommend travel routes and the next song to be played, support medical diagnoses, and, lately, even take an active part in doing homework. Public entities around the world are assimilating algorithmic systems that make, or support, administrative decisions on resource allocation, planning, crime prediction, and protection of the public space—from personal digital assistants to autonomous cars, from robots that carry out simple tasks to monitoring, detection, and forecasting systems.

However, despite their inherent advantages, algorithmic systems may menace human rights and basic freedoms unless there is oversight of their use, development, and deployment. These hazards may emerge at various stages of their development and use—from defining the purpose of the system, via reliance on incomplete, erroneous, corrupted, or biased data, to non-application of post-deployment oversight of the systems' outputs. Furthermore, the more AI develops, the more its systems tend to present new capabilities that were neither intended nor predicted by their developers. Some of these capabilities are inspiring, and therefore called "sparks," but others have the potential to cause harm, such as carrying out offensive cyber actions, manipulating people by discursive means, and disseminating erroneous and misleading artificial

information. Therefore, the ability to identify these capabilities and limit their associated hazards has become a supremely important challenge.

What is artificial intelligence? What are its advantages? And what fears does it evoke, particularly when used by the authorities? These questions are the focus of this book.

Decision-makers, industry, academia, and civil-society organizations in Israel and around the world have all identified AI as a disruptive technology for which national strategy and regulatory policy must be prepared in advance. The end of the previous decade saw the publication of dozens of ethics documents regarding AI which sought to lay down principles for the development, use, and application of algorithmic systems. The ethical values proposed in these documents are founded on several principles: transparency; fairness; damage prevention and safety; responsibility and accountability; privacy; promoting the common good and prioritizing people; and freedom and autonomy.

Ethical values, however, are not enough. To assure the upholding of human rights and basic freedoms, these principles must be moored in legislation. Indeed, we are now seeing around the world initial signs of legislation that seeks to regulate the use and development of AI technologies. Some follow an across-the-board legislative model applied to AI systems at large, as in the European AI Act; others are regulatory patchworks that target specific sectors and uses of AI, such as laws that aim specifically to cope with algorithmic discrimination in systems that hire and promote workers.

This book offers guiding principles for the creation of rights-oriented AI policy and of a toolbox for AI regulation in Israel.

Main Recommendations

Guidelines for the creation of rights-oriented AI policy

Prioritizing people. The main purpose of developing artificial intelligence and learning systems should be serving the human race—individually and collectively—in a way that enhances its welfare. This principle of human freedom and autonomy connects with the principles of promoting the common good and preventing harm, and underscores all the more the importance of the principle of human-centering, freedom, and autonomy. What this anthropocentric outlook means in practice is that the development, deployment, and use of intelligent systems should be based on an approach that considers the defense of basic rights and civil liberties a paramount principle, rather than simply paying lip service to them while actually giving preference to principles such as "promoting innovation," pursuing economic interests such as the advancement of high-tech industry, "making Israel a global technological leader," or even making public-sector processes more efficient.

Prioritizing democracy. AI-based systems have a great deal of potential to infringe on democracy in its broad sense by affecting public discourse and circulating ideas; serving as an instrument of control, surveillance, and policing; and sowing doubt and subverting the very ability to determine reality and distinguish between original and counterfeit and between truth and falsehood. Therefore, the principle of "democracy at the center" should be given much weight, even if this sometimes comes at the expense of technological progress and innovation.

Digital literacy among decision-makers. Digital literacy means the ability to analyze the market and understand the direction in which technology is developing, at least in the near term. For example: Where do the tech giants keep their research and development funds? What patents have they registered in order to secure new technological

developments? It also includes understanding commercial and regulatory possibilities for guiding technological development and, by extension, understanding policymakers' responsibility to influence technological development and not merely observe it from the sidelines. A liminal stratum is needed between understanding technology and making technology policy—a framework for understanding the implications of technological systems, being able to imagine the new possibilities that they offer, and gauging their implications for social ethics and the contours of the judicial method. The frequent lack of framework often results in lacunae in understanding, particularly in matters that have broad implications such as AI.

Reconceptualizing systems and capabilities in the AI field. The concept of artificial intelligence is a highly powerful politico-technological metaphor. Comparing AI systems to the human brain creates proximity and similarity, leading to the social assimilation of the idea that machines operate like the human brain, perform human actions the same way people perform them, and, in fact, compete with people. We recommend that machines' actions and the traits attributed to them be conceptualized in a manner that is not contingent on this comparison.

Developing rights for the objects of AI decisions. Basic rights have to be rethought in two senses. First, the constitutional theory of existing human rights, such as freedom of speech and the right to privacy, needs an injection of new meanings. And second, new digital rights, unneeded in the past, should be created: foremost the rights of individuals who come into contact with algorithm-based machine systems.

Israel must not become a "digital backyard." In recent years, proposals for AI regulation have been advanced in leading countries and the European Union, in what is presumably the onset of a global trend. Even if there are values-based differences among regulatory mechanisms in different places—whether in the choice of systems defined as dangerous or in the

declared goals of the legislation itself—they also have much in common. Therefore, a situation should not be allowed to develop in which Israel lacks legislation that is well aligned with accepted legislation abroad. Admittedly, a lower regulatory standard than the convention abroad may stimulate innovation, but not necessarily of the desired kind; it may turn Israel into a technological backyard, a place where systems are created whose development, dissemination, or use are banned in other western countries.

Future-proof regulation. Technology is advancing rapidly; in countries such as Israel, where legislative processes are very slow, an especially wide gap is being created. Therefore, regulation should not target any specific technology, as that is a sure prescription for regulatory obsolescence. Instead, an effort should be made to establish guiding principles and general definitions that would invest future enforcement with flexibility.

A hybrid regulatory framework: principles, rights, and legislation. Principle-based frameworks present a matrix of ethical core principles; rights-based frameworks focus on protecting the human rights and liberties of those affected by AI-based technological applications; and legislation-based frameworks make it possible not to rely solely on voluntary regulation predicated on the goodwill of economic actors. The three frameworks do not clash with each other; they should be integrated into a hybrid structure that combines soft regulation (ethical principles) with a risk-management approach manifested in rigid legislative provisions and regulatory rules.

Flexibility in the timing of regulatory intervention. In certain cases, there are clear advantages to early regulation, introduced before products based on a certain technology enter the market. If a technology is perceived as especially dangerous—physically (as in autonomous cars) or ethically (such as artificial creation of content that incites to terrorism)—

then advance regulation makes sense. Even in less extreme cases, prior intervention may be useful in shaping the directions of research and in planning resource investment in development. Furthermore, since investment is relatively small and sunk costs are smaller at this stage, regulatory intervention may meet more limited resistance from interested parties. Contrastingly, in certain cases it may be better to wait and cope with problems when they arise instead of trying to anticipate them.

Sectorial vs. across-the-board regulation. Across-the-board regulation attains goals of governance, creates regulatory harmony, and, accordingly, may enhance public trust and increase regulatory certainty for industry. Sectorial regulation, conversely, allows the use of existing regulators and their enforcement powers, does not require the establishment of new institutional frameworks, facilitates more accurate tailoring of enforcement arrangements and methods to a given industry, enhances regulatory clarity and certainty, and also allows stakeholders in each sector to participate in formulating these arrangements. The problem with sectorial regulation, however, is that it may result in discrepancies among industries, create inconsistent standards, exacerbate gaps among regulators, and leave behind unregulated spheres that fall into the cracks. Therefore, we recommend a combination of across-the-board and sectorial regulation, such as having a general regulator with instructional and advisory powers vis-à-vis sectorial regulators.

The AI regulation toolbox

Regulation of learning systems should be based on an understanding of their "lifecycle." To create effective regulation of learning systems, all components of their lifecycle should be taken into account. Given that the principles of regulation—such as fairness, privacy, transparency, accountability, and risk management—are manifested in different

contexts in each component of the lifecycle, inattention to these components may result in excessive regulation of certain elements and disregard of others, compromising regulatory effectiveness.

An integrated approach, by contrast, gives consideration to the full set of lifecycle components and their interrelations. The purpose of a learning system and the framing of the problem it addresses, for example, should influence the choice of model used (whether or not to allow the choice of a more opaque model in terms of the ways it makes decisions). Evaluation outcomes at the stage at which the model is built energize risk-assessment processes—which, in turn, require decisions on how the model should be trained, deployed, and protected. The purpose of the model affects the choice of user interfaces: Should a model that dispenses medical advice notify users that it may be wrong? Should users be told that they are connecting with an artificial system and not with a human one?

An important part of understanding the lifecycle of learning systems relates to the need to monitor them after they are deployed in the real world (e.g., when systems are integrated into a product or an interface). This is because a learning system—unlike other products, such as pharmaceuticals—can, by its very nature, change even after it is applied due to the feedback-loop that it receives from its users.

Development of risk-management methodologies. The application of risk-management methodologies to algorithmic systems, despite being vitally needed, is still in its infancy. We propose a risk-management model that requires a double observation in order to assess the dangerousness of a system, first to assess the potential dangerousness of the system as designed and then to assess the strength of the alignment of its task with its outcome, namely, the potential of a system to manifest its dangerousness outside the role its designers intended for it.

This double layer assessment should be the basis for decision-making. To make it practicable, rules of governance and safety need to be formulated, relating *inter alia* to responsible training (whether to train a new model that exhibits early indications of danger—and in what way) and responsible application (whether, when, and how to implement models that may be dangerous); requisite levels of transparency and documentation in cases of models that may pose extreme danger; and the auditing and cyber-security systems that should be applied to them.

Data documentation, data governance, and post-deployment auditing procedures. The complexity of AI systems generally, and of learning systems particularly, imposes special difficulties on policymakers and regulators when it comes to formulating rules of liability and identifying the actual chain of causation that precipitates an infringement of rights, particularly in consideration of variances among sectors and among applications.

What all these have in common is that they are impossible without proper documentation. The basis for every factual examination of a concrete failure in AI systems is data governability and painstaking documentation of working procedures, information sources, labeling, models, coding processes, risk assessment and databases, and detection of discrepancies in each. Good documentation design also serves the interests of entrepreneurs and developers because it allows them to investigate failures and unexpected phenomena after their occurrence, and also to satisfy regulatory documentation obligations that originate elsewhere.

Development of tools to contend with biases and "fairness engineering." Although the algorithmic-bias problem defies prevention, particularly when the bias originates in reality itself as embodied in data, it may be identified and mitigated. Several strategies to mitigate algorithmic biases should be pursued, including having in place statistically fair

procedures, diversifying human capital among developers of AI systems, and applying after-the-fact auditing procedures.

Coping with the challenges of algorithmic transparency: We suggest a model that is based on the classic conceptualization of transparency, but includes an alternative tailored to the technological limitations of algorithmic systems that cannot provide explanability for specific outputs. The model is meant for cases in which an output cannot be provided but society needs to see the obligation of transparency upheld so as not to thwart the development and use of certain technologies.

Recommendations for an AI regulatory institution in Israel

We recommend establishing an AI regulation authority in Israel for the purpose of advancing and systematizing the regulation of intelligent systems in the country, including the formulation of across-the-board legislation with an eye on corresponding developments abroad. This authority should make policy on the development, implementation, and use of AI-based products in Israel; provide sectorial regulators with professional guidance in order to assure the consistency of the rules that apply to the development, deployment, and use of these systems; and serve as a residual guiding entity, that is, be responsible for regulating those products in fields where there is no sectorial regulator.

The regulator should also provide state authorities with professional AI guidance. Among other things, it should express its professional opinion about government tender documents pertaining to artificial intelligence in order to make sure that the state procures AI systems that meet Israeli and foreign standards in the field, and should give the private market "soft" incentives to align its work with standards.

The regulator should also be tasked with consultation in matters of AI law and should be empowered to present the Knesset and the courts with its own position in these fields. In addition to the aforementioned roles, the proposed regulator should be a source of knowledge, instruction, and cooperation and should promote digital literacy among government players in these regards. It should also discuss the possible social effects of these technologies with local and international stakeholders in industry, government, and academia.

At the present time—at least for the next few years, until the field stabilizes—the AI regulator should be established in the form of a unit within the Regulatory Authority, because the latter's roles are highly suited to the evolving world of AI regulation. To cope with the challenges of flexibility, the burdened regulatory environment, and the technology, enough resources should be allocated to allow positions to be created for experts in technical fields and also in law and policy. Even in the absence of a comprehensive artificial intelligence law, the Authority should be budgeted by government resolution so that it can set up the AI regulator's unit.

Recommendations for the interim period until AI regulation is introduced

Supplementary legislation and legislatory amendments. Right now, even in the absence of a broad and dedicated artificial intelligence law, designated decision-makers and regulators should be obligated to update existing legislation and pass supplementary legislation—mainly to statutes such as the Competition and Consumer Protection Law, the Copyright Law, the Protection of Privacy Law, the Evidence Ordinance, and the Government Procurement Law.

Create a "pre-regulatory ecosystem." A "pre-regulatory ecosystem" should be created, in which the designated regulator, sectorial regulators, and supplemental regulators (such as the Protection of Privacy Authority and the Competition Authority) would issue guidelines and professional opinions while serving industry, its players, and the courts. These guideline documents would help industry set self-regulation standards on the assumption that future regulation would resemble the guidelines. They would also serve standards-setting companies in constructing their standards frameworks.

Under the influence of this pre-regulatory ecosystem, auditing patterns in various contexts would undergo fine-tuning, thinking about due caution standards would take place, an effort to encourage responsibility in developing and implementing learning systems would be made, sandboxes would be built, various regulatory trials would be undertaken, and a pool of experts who could serve industry, the regulators, and the courts in coping with new and complex issues would take shape. Concurrently, industry and its actors would be able to provide the various regulatory bodies with feedback, thus improving the guidelines. In addition, in the absence of regulation, the courts could use the guidelines for inspiration as they interpret conflicts presented to them. Their rulings would also enrich the body of knowledge and enable players and regulators alike to enhance and polish their guidelines until regulation can be formulated.

Series & Cover Design: Studio Alfabees
Typesetting: Nadav Shtechman Polischuk
Chart Graphics: Dana Berger
Printed by Maor Wallach Print

ISBN: 978-965-519-433-3

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 © 2023 by the Israel Democracy Institute (RA)
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

All IDI publications may be downloaded for free, in full or in part, from our website.

The views expressed in this book do not necessarily reflect those of the Israel Democracy Institute.

מערכות בינה מלאכותית ממליצות על נתיבי נסיעה או בוחרות את השיר הבא שיושמע, תומכות באבחון רפואי ומשתתפות באופן פעיל בהכנת שיעורי הבית ובכתיבת מאמרים מדעיים. גופים ציבוריים ברחבי העולם מטמיעים מערכות אלגוריתמיות שמקבלות החלטות מינהליות הנוגעות להקצאת משאבים, לתכנון, לחיזוי פשיעה או להגנה על המרחב הציבורי. מעוזרים אישיים דיגיטליים ועד מכוניות אוטונומיות, מרובוטים המבצעים משימות פשוטות ועד מערכות מעקב, זיהוי וחיזוי - מהפכת הבינה המלאכותית בעיצומה.

ואולם על אף היתרונות הגלומים במערכות אלגוריתמיות, ללא בקרה על השימוש בהן, על פיתוחן ועל פרישׂתן הן עלולות לסכן זכויות אדם וחירויות יסוד, לגרום נזק ולבצע פעולות סייבר התקפיות, לתמרן אנשים או להפיץ מידע מלאכותי מוטעה ומטעה. האפשרות לזהות יכולות אלו ולהגביל את הסיכונים שהן מביאות איתן נעשתה אפוא אתגר חשוב מאין כמוהו. ספר זה, הראשון מסוגו בעברית, בוחן מהי בינה מלאכותית, מהם יתרונותיה ואילו חששות היא מעוררת בארץ ובעולם. מחבריו קוראים למקבלי ההחלטות, התעשייה, האקדמיה וארגוני החברה האזרחית בישראל להיערך לקראתה באמצעות אסטרטגיה לאומית ומדיניות רגולטורית. כדי להבטיח שאנשים הם שישלטו במכונות ולא להפך, נדרש פיתוח של עקרונות מנחים ליצירת מדיניות בינה מלאכותית מכוונת זכויות וארגז כלים לאסדרת בינה מלאכותית.

ד"ר תהילה שוורץ אלטשולר היא עמיתה בכירה וראשת התוכנית "דמוקרטיה בעידן המידע" במכון הישראלי לדמוקרטיה. מומחית לאתיקה, מדיניות ומשפט של תקשורת וטכנולוגיה. תחומי המחקר שלה הם פרטיות, אסדרת סייבר, רשתות חברתיות, ריכוזיות בשוק התקשורת ומדיניות תקשורת. חברה במועצת הארכיונים העליונה ובארגוני חברה אזרחית העוסקים באתיקה עיתונאית ובזכויות דיגיטליות ובעלת טור בנושאי טכנולוגיה ורגולציה במגזין "דה מרקר".

עו"ד עמיר כהנא הוא חוקר בתוכנית "דמוקרטיה בעידן המידע" במכון הישראלי לדמוקרטיה; סטודנט לתואר שלישי בפקולטה למשפטים באוניברסיטה העברית בירושלים, בוגר תואר ראשון במשפטים מהמרכז הבינתחומי הרצליה ותואר שני במשפטים מאוניברסיטת קיימברידג'. עמית מחקר בתוכנית לסייבר ומשפט במרכז פדרמן לחקר הסייבר באוניברסיטה העברית בירושלים.

0 4500001257 8
450-1257 דאנאקוד

מחיר מומלץ: 82 ש"ח
יולי 2023
מסת"ב: 978-965-519-433-3

www.idi.org.il