

סוגיות אתיות ופוליטיות של טכנולוגיות טרנספורמטביות

מילון מושגים

ד"ר שירה אחישר
ד"ר אביגיל פרדמן
רינת קליין
פרופ' עלי זלצברגר

מוסד שמואל נאמן
למחקר מדיניות לאומית

אתיקה | 08/26





סוגיות אתיות ופוליטיות של טכנולוגיות טרנספורמטיביות

מילון מושגים

ד"ר שירה אחישר

ד"ר אביגיל פרדמן

רינת קליין

פרופ' עלי זלצברגר

אין לשכפל כל חלק מפרסום זה ללא רשות מראש ובכתב ממוסד שמואל נאמן מלבד לצורך ציטוט של קטעים קצרים במאמרי סקירה ופרסומים דומים תוך ציון מפורש של המקור.
הדעות והמסקנות המובאות בפרסום זה הן על דעת המחבר. ת ואינן משקפות בהכרח את דעת מוסד שמואל נאמן.

www.neaman.org.il

תוכן העניינים

1.	מושגי יסוד באתיקה.....	4
1.1	מטא-אתיקה	4
1.2	אתיקה תיאורית (Descriptive ethics)	5
1.3	אתיקה נורמטיבית (Normative ethics) ואתיקה יישומית	5
1.4	מוסר, אתיקה ומשפט	11
2.	מושגי יסוד בטכנולוגיה.....	12
2.1	מהי טכנולוגיה	12
2.2	טכנולוגיות טרנספורמטיביות/טכנולוגיות מערעות.....	12
2.3	טכנולוגיה דיגיטלית.....	13
2.4	מהן הטכנולוגיות הטרנספורמטיביות של העידן הדיגיטלי	14
3.	תפיסות יסוד לגבי השפעת הטכנולוגיה הטרנספורמטיבית על האנושות.....	16
3.1	החתירה לבינה מלאכותית כללית (Artificial General Intelligence).....	16
3.2	הומיניזם דיגיטלי.....	17
3.3	דטרמיניזם טכנולוגי.....	18
3.4	פסימיזם טכנולוגי מול אופטימיזם טכנולוגי	18
3.5	תאוריית האיום הקיומי	20
4.	סוגיות יסוד באתיקה של טכנולוגיה.....	22
4.1	גישות שונות לשימוש ראוי במשאבים הטכנולוגיים.....	22
4.2	אתגרים מוסריים שמציבה בינה מלאכותית	23

5.	סוגיות יסוד בפילוסופיה הפוליטית של הטכנולוגיה	31
5.1	ערעור מאזן הכוחות הדמוקרטי	31
5.2	עירוב אלגוריתמים בקבלת החלטות	32
5.3	מניפולציות טכנולוגיות על ידע פוליטי	33
5.4	גיימיפיקציה/רידוד/קריסת ערך (value collapse)	34
5.5	הטיות	34
5.6	טכנולוגיה ולוחמה	35
5.7	פשעים של העידן הדיגיטלי (Cybercrime)	36
5.8	טכנולוגיות מעקב	38
6.	אתיקה של בינה מלאכותית: עקרונות ופרוגרמות ליישום	39
6.1	עקרונות אתיים מנחים לשימוש בבינה מלאכותית	39
6.2	חקיקה מרכזית להגנה מפני בינה מלאכותית	43
6.3	קודים אתיים למודלים של בינה מלאכותית	48

1. מושגי יסוד באתיקה

מוסר ואתיקה עוסקים באבחנה בין טוב ורע, בין התנהגות ראויה לכזו שאינה ראויה. המוסר יכול לבוא לידי ביטוי ברמה האישית כמו גם החברתית, ולגבי האחרונה ניתן גם להוסיף את המוסר הנוהג - מערך הנורמות והקודים ההתנהגותיים שמקובלים בחברה ובתרבות מסוימת. המושג אתיקה משמש בדרך כלל ברמה הקבוצתית, בין אם זה ברמת החברה, המדינה, משלח יד או פרופסיה. אתיקה היא גם הענף הפילוסופי החוקר בצורה שיטתית ולעיתים אף ביקורתית את אותן נורמות, ומציעה סט של כללים להתנהגות הנחשבת ראויה.

אתיקה היא מושג רחב מאוד, וניתן לחלקו למספר תתי-תחומים:

מטא-אתיקה - חקר משמעויות המושגים והעמדות המוסריות, ובחינת היסודות המטפיזיים, האפיסטמולוגיים והסמנטיים של המחשבה והדיבור המוסרי.

אתיקה תיאורית - בחינת הנורמות המוסריות שקיימות בפועל.

אתיקה נורמטיבית בחינת השאלות "כיצד עלינו לפעול?", "פיתוח תיאוריות מוסריות (כדוגמת תועלתנות, קנטיאניזם ותורת המידות) שמטרתן להנחות אותנו במעשה.

אתיקה יישומית - יישום התיאוריות והעקרונות של האתיקה הנורמטיבית על בעיות מעשיות (בתחומים כגון ביואתיקה, אתיקת סביבה, אתיקה עסקית ואתיקה מקצועית), והפיכתם להנחיות פרקטיות, ולעיתים אף לנורמות משפטיות.

1.1 מטא-אתיקה

התיאוריות המטא-אתיות המרכזיות נחלקות בדרך-כלל לשתי גישות עיקריות:

גישות ריאליסטיות טוענות שישנן עובדות מוסריות אובייקטיביות — אמיתות מוסריות שקיומן אינו תלוי באמונותיו, ברגשותיו או בקונבנציות התרבותיות של היחיד — ואדם יכול לגלות אותן באמצעות טיעון רציונלי ואף חייב לחיות על פיהן.

גישות אנטי-ריאליסטיות כדוגמת אמוטיביזם ופרסקריפטיביזם שוללות את קיומם של ערכים מוסריים כעובדות בלתי תלויות, ורואות בשיפוט מוסרי ביטוי לרגשות, העדפות או פקודות אנושיות, ולא כהצהרות עובדתיות¹.

¹ Sayre-McCord, Geoff, "[Moral Realism](#)", *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.)

1.2 אתיקה תיאורית (Descriptive ethics)

עוסקת בתיאור המוסר הקיים כפי שהוא מופיע בפועל: מנהגים והרגלים, דעות לגבי טוב ורע, מסגרות של התנהגות אחראית ובלתי-אחראית, וגבולות הפעולה המקובלת. היא בוחנת את המוסר כפי שהוא מתקיים בתת-תרבויות שונות או בתקופות היסטוריות מסוימות, ללא שיפוט ערכי.

1.3 אתיקה נורמטיבית (Normative ethics) ואתיקה יישומית

אתיקה נורמטיבית בוחנת האם הנורמות והערכים הנהוגים עומדים בקנה אחד עם המוסר האידאלי². היא מגדירה מהם המעשים הנכונים והלא-נכונים, ומנסחת את העקרונות האוניברסליים שמלווים את הפרט בדרכו אל חיים מוסריים. יישום תאוריות וסטנדרטים נורמטיביים על שאלות מוסריות קונקרטיות הוא ליבת האתיקה היישומית³.

אתיקה יישומית אינה חייבת להתבסס על תאוריה מוסרית אחת, אלא מציעה לנסח טיעונים ולבחון נקודות מבט שונות, ובכך מספקת כלים ביקורתיים לניתוח סוגיות מוסריות בהתבסס על תאוריות מוסריות שונות. במסגרת האתיקה הנורמטיבית בולטות מספר תאוריות מרכזיות⁴:

1.3.1 אתיקה דאונטולוגית (Deontological ethics)

תאוריות מוסר דאונטולוגיות גורסות שטוב ורע הם מושגים אובייקטיביים, אוניברסליים ונצחיים (ראו מטא אתיקה ריאליסטית למעלה). תאוריות כאלו הוצעו, בין השאר, על ידי אפלטון ותומס אקווינס (אבי גישת המשפט הטבעי שכתב על הטובין הבסיסיים – החיים והמשכיותם, ידע, אסתטיקה, ידידות, סבירות מעשית ודת). כך גם תאוריית המוסר של עמנואל קאנט המתמקדת בקיום כללים וחובות מוסריים, וטוענת כי יש פעולות הנכונות או השגויות במהותן, ללא תלות בתוצאותיהן. גישה זו מכירה בחובות ובזכויות שאין להפר, גם אם בכך ניתן היה להשיג תועלת רבה יותר. למעשה מוסרי יש משמעות עצמאית, והקפדה על עקרונות כלליים – למשל כיבוד זכויות אדם, הימנעות מניצול האדם כאמצעי בלבד – אשר צריכים להנחות את הפעולה ללא תלות בתוצאותיה.

גישה זו רלוונטית במיוחד בעידן של פיתוח טכנולוגיות מתקדמות, ובפרט בתחום הבינה המלאכותית⁵. ישנה נטייה גוברת להעדיף שיקולים תועלתניים – כגון יעילות, ביצועים ורווח – על פני מחויבות לזכויות הפרט

² Ethics, Technology, and Engineering: An Introduction. Ibo van de Poel, Lamber Royakkers. ISBN: 978-1-444-39571-6. March 2011. Wiley-Blackwell.

³ Duigan, Brian. "Normative Ethics." In *Britannica*. Accessed May 29, 2025.
<https://www.britannica.com/topic/normative-ethics>.

⁴ Fieser, James. "Ethics." *Internet Encyclopedia of Philosophy*, www.iep.utm.edu

⁵ D'Alessandro, W. Deontology and safe artificial intelligence. *Philos Stud* (2024).
<https://doi.org/10.1007/s11098-024-02174-y>

ולעקרונות מוסריים אוניברסליים. על פי עקרון ה-"Humanity as an end in itself" יש להימנע משימוש באנשים כאמצעים בלבד לפיתוח ושימוש בטכנולוגיה, גם כאשר מדובר בכלי מחקר או יישומים מסחריים.

חוקרים שונים מציעים ליישם עקרונות דאונטולוגיים גם במנגנוני התכנון של מערכות AI עצמן⁶. לדוגמה, נטען כי "פיתוח אחראי של מערכות בינה מלאכותית חייב להתבסס על חובות מוסריות מוקדמות, ולא רק על ניתוחי סיכונים תוצאתיים". לפי גישה זו, שקיפות, הסכמה מושכלת, הוגנות וכיבוד פרטיות הם עקרונות מוסריים מהותיים שיש לקיים – גם כאשר מדובר במחקרים שאינם מפרים חוק או נראים יעילים מבחינה טכנולוגית או כלכלית. שילוב עקרונות דאונטולוגיים בתהליכי מחקר ופיתוח מסייע למוסדות אקדמיים להבטיח כי החוקרים שומרים על מחויבות מוסרית עמוקה ולא רק פורמלית, ובכך מבססים אמון ציבורי לאורך זמן⁷.

אחת הבעיות המרכזיות באימוץ גישה דאונטולוגית ויישומה לשאלות קונקרטיות היא יכולת ההכרעה במוסריות פעולה כאשר יש בה התנגשות בין שתי חובות או זכויות מוחלטות או בין שניים או יותר טובין בסיסיים. במילים אחרות, מבחינה אנליטית אין יכולת הכרעה בין שניים או יותר ערכים מוחלטים.

אתיקה דאונטולוגית - מקרה בוחן: פיתוח מערכת לזיהוי רגשות והשלכות מוסריות של שימוש בה

רקע

חוקרת בכירה במדעי המחשב, מובילה מחקר לפיתוח מערכת בינה מלאכותית לזיהוי רגשות אנושיים בזמן אמת, על בסיס תווי פנים, קול ושפת גוף. המערכת נועדה לשפר אינטראקציה בין אדם למכונה, ולהשתלב בעתיד בשירותי לקוחות, חינוך ואפילו בהערכות ביצועים בארגונים.

ההתלבטות האתית

במהלך הפיתוח, שותפים מהתעשייה מציעים לחוקרת לבדוק את המערכת בסביבה אמיתית – מערך שירות לקוחות של חברת טכנולוגיה גדולה – מבלי להודיע במפורש לנציגים או ללקוחות שהשיחות מוקלטות ונותחו לצורך זיהוי רגשות. ההצדקה: אם המשתתפים ידעו מראש, ההתנהגות שלהם תשתנה, ויעילות המערכת תיפגע. לטענת השותפים, מדובר בבדיקה "לא מזיקה" שתוביל לשיפור חוויית הלקוח בעתיד.

הבחינה הדאונטולוגית

החוקרת מסרבת להצדיק את השימוש במידע בלי הסכמה מדעת, גם אם התוצאה תועיל בטווח הארוך. מבחינתה, הצבת המשתמשים ונציגי השירות כ"אמצעי" בלבד לשיפור טכנולוגיה מפרה

⁶ Hafner, Alina; Hammerschmidt, Teresa; and Stolz, Katharina, "Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review" (2024). Wirtschaftsinformatik 2024 Proceedings. 46. <https://aisel.aisnet.org/wi2024/4>

⁷ McGraw, David K. (2024) "[Ethical Responsibility in the Design of Artificial Intelligence \(AI\) Systems](#)," International Journal on Responsibility: Vol. 7: Iss. 1, Article 4.

את החובה המוסרית לכבד את פרטיותם וזכויותיהם, במיוחד כאשר מדובר בטכנולוגיה שחודרת לרבדים רגשיים ואישיים.

הפעולה שננקטה

היא דורשת מהחברה לאפשר למשתמשים לדעת על הפיילוט ולהסכים לו מראש, גם אם הדבר יפגע בדיוק הנתונים. בנוסף, היא פועלת באוניברסיטה להטמעת הנחיות מחקר אתיות חדשות לפרויקטים בתחום "AI רגשי", הכוללות עקרונות של שקיפות, הסכמה ותחומי איסור לשימוש.

משמעות רחבה

המקרה מדגיש כיצד עקרונות דאונטולוגיים, כמו כיבוד זכויות האדם והימנעות ממניפולציה, יכולים לעמוד בניגוד לשיקולים של יעילות או חדשנות. אך דווקא שמירה עיקשת על עקרונות אלה היא שמבססת את אמון הציבור ואת הלגיטימציה של המחקר בטווח הארוך.

1.3.2 אתיקה תוצאתנית (Consequential ethics)

תאוריות מוסר תוצאתניות או טלאולוגיות מעריכות מעשים על פי תוצאותיהם, ודוגלות בבחירות שמניבות את התועלת המקסימלית לטובת מספר הפרטים הרב ביותר, או את התועלת המצרפית המירבית. ערכה של פעולה מוסרית נקבע אך ורק על פי התוצאות שלה, ובכך היא מפרידה בין המעשה לבין תוצאתו. עם הזמן, גישה זו הורחבה מעבר להערכת פעולות בלבד, והחלה לכלול גם שיפוט מוסרי של כללים, מניעים, תכונות אופי ואף מוסדות, מתוך הנחה שגם אלה יכולים להיבחן על פי תרומתם לרווחה הכוללת.⁸

התאוריה המוסרית הקאנונית במסגרת אתיקה תוצאתנית היא התועלתנות (Utilitarianism) - מבית מדרשו של ג'רמי בנתם. גישה זו גורסת כי פעולה מוסרית היא כזו שמניבה תועלת כוללת הגדולה מהנזק שהיא יוצרת והפעולה המוסרית ביותר היא זו שמביאה את התועלת המצרפית המירבית. המוסר התועלתני אומץ גם על ידי הזרם המרכזי בתאוריה הכלכלית המודרנית. יש הרואים בתועלתנות גישה בעייתית, שכן משתמע ממנה שלעיתים מותר לפגוע באנשים – אם הדבר תורם לרווחה רבה יותר של אחרים.

תת זרמים עיקריים נוספים במסגרת התוצאתנות כוללים בין השאר:⁹

אגואיזם מוסרי (ethical egoism) הטוען כי פעולה מוסרית היא זו שמביאה את מירב התועלת לאדם המבצע אותה.

אלטרואיזם מוסרי (ethical altruism) רואה פעולה כמוסרית רק אם היא **מיטיבה עם אחרים** – ולא עם האדם הפועל.

⁸ Díaz de la Cruz, C., Fernández Fernández, J.L. & Villegas-Galaviz, C. [Model of ethical analysis of digital technologies: towards true digital humanism](#). AI & Soc (2025).

⁹ Mohn, Elizabeth, 2022, [Consequentialism \(ethics\)](#), EBSCO

שני אלה נבדלים מתועלתנות, שכן **בתועלתנות מתחשבים הן בתוצאות עבור האדם הפועל והן בתוצאות עבור אחרים.**

חסרונות מרכזיים של התוצאתנות הם העדר יכולת למדוד ולהשוות תועלות (בעיה שנפתרה באופן חלקי על ידי מדע הכלכלה שהחליף תועלת בכסף) ושהתוצאות אינן תמיד ניתנות לחיזוי. גם אם בני אדם מאמינים שיש לפעול לפי תוצאות מיטביות, לרוב קשה להם לצפות מראש אילו תוצאות צפויות להביא את מירב הטוב. בנוסף, קשה להשוות בין תוצאה אחת לאחרת ולקבוע איזו טובה יותר – במיוחד כאשר פעולה אחת מובילה למגוון השלכות שמושפעים מהן אנשים שונים בדרכים מגוונות.

חיסרון נוסף של תוצאתנות הוא התעלמותה מגורמים חשובים שגישות אתיות אחרות רואות בהם ערך. לדוגמה, תוצאתנות **אינה מתחשבת באופיו של האדם הפועל ובאוטונומיה שלו**, ואינה שוקלת את צדקתה של החלוקה או ההוגנות שבתוצאה עצמה.

אתיקה תוצאתנית - מקרה בוחן: איזון בין פרטיות הציבור לתועלת חברתית במחקר המבוסס על ניתוח מידע ממצלמות חכמות

רקע

חוקר באוניברסיטה ציבורית, מוביל פרויקט מחקרי לפיתוח אלגוריתמים לזיהוי צפיפות והתקהלויות במרחב העירוני באמצעות מצלמות חכמות. מטרת המחקר: לאפשר לרשויות מקומיות לנהל טוב יותר אירועים המוניים, למנוע סכנות בטיחות, ולהתערב מוקדם במקרי חירום. הפרויקט מממן בשיתוף עירייה גדולה, ומבוסס על צילומים מצמתיים מרכזיים בעיר.

ההתלבטות האתית

המערכת הנבדקת עושה שימוש בנתוני וידאו ללא טשטוש פנים, אף שאין בכוונתה לזהות אנשים פרטניים. קבוצות לזכויות אזרח טוענות שמדובר בפגיעה בפרטיות הציבור, גם אם התמונות אינן נשמרות לזמן רב. מצד שני, החוקר מציין שהמערכת יכולה למנוע אסונות צפיפות, לשפר נגישות עירונית, ואף לסייע באירועי חירום.

הבחינה התוצאתנית

במקום לשאול האם השימוש במצלמות הוא מוסרי כשלעצמו, החוקר בוחן את תוצאות המחקר. מנקודת מבט תועלתנית, ההצדקה המוסרית תלויה בשאלה האם סך התועלת עולה על סך הפגיעה האפשרית. במקרה זה, המערכת צפויה להועיל לרבבות תושבים, לשפר את הבטיחות ולהציל חיים – תועלת ציבורית רחבת היקף. לכן, הפגיעה הפוטנציאלית בפרטיות, שהיא מוגבלת לפי תנאי השימוש, נחשבת "מחיר מוסרי נסבל" בהשוואה לרווח החברתי.

הפעולה שננקטה

כדי למזער את הפגיעה, החוקר מוודא שהצילומים לא נשמרים, שאין זיהוי ביומטרי, ושהציבור מקבל מידע כללי על הפרויקט. הוא גם מציע להציג שלטים במרחב הציבורי המעידים על השימוש במצלמות לצרכים מחקריים. בכך, הוא מבצע איזון תוצאתני בין ערכים מתחרים – פרטיות מול בטיחות – תוך ניסיון למקסם את רווחת הציבור הכוללת.

משמעות רחבה

המקרה מדגים כיצד אתיקה תוצאתנית יכולה להנחות חוקרים אקדמיים להעדיף פעולה שתשיא תועלת רחבה, גם במחיר של פגיעה מסוימת בזכויות הפרט. ההנמקה אינה מבוססת על חובה מוסרית מוחלטת, אלא על שיקול של "הטוב המצטבר" שנוצר מהמחקר.

1.3.3 אתיקת המידות (Virtue ethics)

מקורה של אתיקת המידות בתורתו הקלאסית של אריסטו. היא מדגישה פיתוח תכונות אופי חיוניות (מידות) שבליבתן עומדות השאלות "מי האדם הראוי?" ולא "מה ראוי לעשות?" המידה הטובה מכתיבה את הבחירה הנכונה בכל מצב: "מה שהאדם המוסרי היה עושה" משמש כקו מנחה להתנהגות ראויה.

עידן הבינה המלאכותית מציב אתגרים חדשים לחיים מוסריים, פוליטיים ותרבותיים – כמו טשטוש הגבול בין אנושי ומכונה, פיתוי לווייתור על שיפוט מוסרי, ואובדן מצפן מוסרי אישי. כדי להתמודד עם אתגרים אלה, נדרשת לא רק רגולציה או הנחיות טכניות, אלא טיפוח של אופי מוסרי – כלומר, עיצוב האדם שיידע לפעול נכון במצבים מורכבים ובתנאי אי וודאות. חינוך מוסרי איכותי לעידן בינה המלאכותית צריך לשאול לא רק "מה נכון לעשות?" אלא "מי אני רוצה להיות?" לצורך כך, יש לטפח שילוב של מידות מוסריות ואפיסטמיות¹⁰. יושרה, אחריות, ענווה, קפדנות, ואומץ לב מוסרי. מידות אלה מהוות תנאי הכרחי לפעילות האנושית, בין השאר גם למחקר אתי המשרת את החברה ואינו חורג מהמנדט הציבורי והערכי של המדע. בהקשר זה מוזכרות¹¹ ארבע מידות בסיסיות: צדק, יושרה, אחריות וענווה כשתי מידות נוספות – שיקול דעת ואומץ מוסרי – מצוינות כמידות תומכות להתמודדות במצבים של חוסר ודאות. אלו אינן תכונות מולדות, אלא יכולות הדורשות אימון מתמשך, רפלקסיה ביקורתית, וחינוך דרך דוגמה אישית והתנסות¹².

התבוננות זו בליבה של אתיקת המידות מדגישה שחוקר ראוי אינו חייב רק כישורים טכניים, אלא גם דמות מוסרית ואפיסטמית העומדת מול אתגרים של השפעה חברתית, שקיפות, ייצוג הוגן ואחריות ציבורית. אמות המידה הללו רלוונטיות הן לשלב תכנון המחקר (design) והן לבחינת השימוש והאימוץ של תוצאותיו לאורך זמן.

¹⁰ Smith, N., & Vickers, D. (2024). Living well with AI: Virtue, education, and artificial intelligence. *Theory and Research in Education*, 22(1), 19-44. <https://doi.org/10.1177/14778785241231561>

¹¹ Hagendorff, T. A Virtue-Based Framework to Support Putting AI Ethics into Practice. *Philos. Technol.* 35, 55 (2022). <https://doi.org/10.1007/s13347-022-00553-z>

¹² Bate, G.W., Lloyd, R. Review of Shannon Vallor (2024). The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking. *Postdigit Sci Educ* (2025). <https://doi.org/10.1007/s42438-025-00546-z>

אתיקת המידות - מקרה בוחן: אחריות מוסרית של חוקר אקדמי בעבודה עם מאגרי מידע רגישים

רקע

חוקר בכיר באוניברסיטה ציבורית, העוסק בפיתוח מודלים ללמידת מכונה בתחום הבריאות, מקבל הצעה לשיתוף פעולה עם מוסד רפואי פרטי שמעוניין להעביר לידיו מאגר נתונים גדול הכולל תיעוד קולי של ראיונות פסיכולוגיים. המטרה: לאמן מודלים לזיהוי מוקדם של סימפטומים נפשיים כמו דיכאון וחרדה.

ההתלבטות האתית

על אף שהנתונים נאספו בהתאם להנחיות ועדת האתיקה ונמסרו לאחר הסכמה מדעת, החוקר מזהה שהמסמכים המשפטיים שנמסרו לנבדקים כתובים בשפה מורכבת, ויתכן שהם לא היו מודעים לכך שפרטי המידע יועברו למחקר אקדמי משני. בנוסף, מטרת הפיתוח הסופית אינה ברורה, ויתכן שהמערכת תשמש גם לצרכים מסחריים בעתיד.

הבחינה המוסרית - אתיקת המידות

החוקר אינו מסתפק בשאלה האם המהלך "חוקי", אלא בוחן את המצב מתוך אחריותו האישית והמקצועית כחוקר. הוא מפעיל שיקול דעת מוסרי ומיישם תכונות אופי חיוניות:

יושרה - חתירה לאמת ולבהירות, גם כאשר היא לא נוחה.

אחריות מוסרית - התייחסות זהירה להשפעות האפשריות על הנבדקים.

אמפתיה וחמלה - רגישות לשימוש בנתונים אינטימיים של אנשים במצבי פגיעות.

ענווה אפיסטמית - הכרה בכך שלא ניתן לצפות את כל ההשלכות של השימוש בטכנולוגיה.

הפעולה שננקטה

החוקר בוחר שלא להשתמש בנתונים בשלב זה, עד להבהרת הסכמות המשתתפים ותיחום מטרות השימוש. הוא יוזם שיח מוסדי בנוגע לשימוש אחראי במאגרי רגישים, ומעודד יצירת קווים מנחים פנימיים שיגדירו עקרונות פעולה המבוססים לא רק על תקנות, אלא גם על זהות מוסרית של חוקרים.

משמעות רחבה

המקרה מדגיש את החשיבות בטיפוח זהות אתית בקרב חוקרים אקדמיים, במיוחד בתחומים רוויי השפעה כמו בינה מלאכותית ומדעי הבריאות. מעבר לעמידה בכללים הפורמליים, נדרשת גם מחויבות לערכים מוסריים ולמידות אישיות – כבסיס לחקר אחראי ומיטיב.

1.4 מוסר, אתיקה ומשפט

תאוריות המוסר המתוארות לעיל הן הבסיס הכללי להתנהגות אישית וחברתית. על בסיסן נוצרו עקרונות וכללי אתיקה לבעלי עיסוק משותף, פרופסיות, בעלי תפקידים ונבחרי ציבור. כך, גם קיימת אתיקה של מחקר ושל מוסדות להשכלה גבוהה.

חלק מעקרונות וכללי האתיקה הופכים גם לחלק מהמשפט כשהם, למשל, מותקנים ככללי משמעת ויש להם מערכות אכיפה. כך, למשל כללי האתיקה של עורכי דין ורופאים וכללי המשמעת במוסדות להשכלה גבוהה. אך יש פרופסיות, כמו העיתונאים בישראל, בהן כללי האתיקה וולונטריים ואינם חלק מהמשפט. חשוב לזכור שגם לגבי מקצועות ומגזרים בהם קיימים כללי אתיקה משפטיים, אלו לא מכסים את כל העקרונות והכללים האתיים הרלוונטיים לתחום או למקצוע.

אתיקה היא הבסיס התוכני והמהותי של המשפט. אבל בחברה חופשית המשפט לא כולל את כל המוסר – לא האידאלי וגם לא הנוהג. יפים לעניינו דבריו של שופט בית המשפט העליון יצחק זמיר:

”כללי אתיקה, מעצם מהותם, אינם כללי משפט. לכן אין הם עניין לבית משפט. כך לגבי האתיקה החברתית, כך לגבי האתיקה הציבורית, כך לגבי אתיקה מקצועית לרבות האתיקה העיתונאית. אם למשל, התנהגות מסוימת של עיתונאי נחשבת לא אתית, הרי זה עניין לבית הדין לאתיקה של מועצת העיתונות, אך אין זה עניין לבית משפט. בית-המשפט הקים, כביכול, גדר בין תחום האתיקה לבין תחום המשפט: מלכות אחת אינה נוגעת במלכות אחרת. אכן, ראוי למשפט, ונוח לו, להשאיר תחום רחב של התנהגות, חברתית או מקצועית, מחוץ להסדר משפטי. הנחה היא, שמוסכמות חברתיות וכללי אתיקה יספיקו כדי להסדיר את ההתנהגות בתחום זה באופן סביר, ללא צורך בהתערבות המשפט. בתחום זה די בדרך-כלל בתגובה חברתית על הפרת המוסכמות, או בעיצומים מקצועיים על הפרת האתיקה, כדי לקיים סדר טוב והתנהגות ראויה. כל זמן שכך הוא, המשפט נשאר בתחום שלו, ואינו עובר את הגדר לתחום האתיקה. אולם טעות היא לחשוב שהמשפט אינו מתעניין במה שמתרחש מעבר לגדר. הגדר אינה אטומה. המשפט עוקב בעיניים פקוחות אחר האתיקה. יש מקרים שבהם נראה לעין כי האתיקה חלשה ואין בכוחה להבטיח התנהגות ראויה, אם משום שאינה ברורה או מוסכמת, אם משום שאין לה מגנון יעיל של אכיפה ואם מטעם אחר. במקרה כזה, אם אינטרס ציבורי חשוב נתון בסכנה ממשיית, עשוי המשפט לעבור את הגדר המפרידה בינו לבין האתיקה, ולהפוך כלל אתי לכלל משפטי, כדי להקנות לו את הכוח הנדרש להשליט סדר ראוי. זהו תהליך מוכר. המשפט מתרחב בהתמדה על חשבון האתיקה. כללי אתיקה של עורכי-דין ורופאים, למשל, הפכו במידה רבה לכללי משפט, שלעתים קוראים להם כללי משמעת. הפרת כללים אלה עלולה לגרור הליכים משמעתיים בפני מוסדות רשמיים הפועלים לפי חוק. לעתים כללים אתיים הופכים אפילו לאיסורים פליליים. הנה, לדוגמה, עניין המתנות לעובדי ציבור”¹³.

אחד ההבדלים החשובים בין משפט לאתיקה הוא שלגבי המשפט יש תמיד תשובה אחת נכונה. התנהגות היא או חוקית או בלתי חוקית, ובעזרת מסגרת מוסדית מתאימה יש מי שקובע האם המעשה חוקי או לא. באתיקה לא תמיד יש תשובה נכונה אחת לגבי אתיות ההתנהלות ואין תמיד מסגרת מוסדית המכריעה

¹³ דנ"א 7325/95 ידיעות אחרונות בע"מ נ' קראוס, פ"ד נב(3) 1 (1998)

בעניין. בסוגיות סבוכות עצם זיהוי הסוגיות האתיות, ההתחבטות האישית בהן, וההתייעצות לגביהן, יכולה להישפט כהתנהלות אתית, מבלי קשר להחלטה לגבי ההתנהלות עצמה ותוצאותיה.

לא ברור אם ההנחה של יצחק זמיר משנות ה-80 - שמוסכמות חברתיות יסדירו את תחום האתיקה - עומדת בעינה כיום במדינת ישראל. ניתן לטעון כי עצם התהליך של התרחבות המשפט על חשבון האתיקה, הוא כנראה אחת הסיבות לא רק לחולשתה אלא גם להחלשתה של האתיקה על ידי המשפט, שכן הוא מוביל לתפיסה בעיני הציבור שמה שהוכרע כחוקי הוא מקובל וגם אתי.

אחת המטרות שלנו במסמך זה היא להחיות מחדש את ממד האתיקה והשיפוט האתי, את המודעות לאתיקה והטמעתה במסגרת האקדמיה, המחקר ובדגש על אתיקה וטכנולוגיה.

2. מושגי יסוד בטכנולוגיה

2.1 מהי טכנולוגיה

טכנולוגיה היא היישום של ידע אנושי וכלים חדישים לצורך פתרון בעיות. הטכנולוגיה אינה כוללת רק מכשירים אלקטרוניים דוגמת טלפון או מחשב, אלא כל כלי המשמש לצורך פתרון בעיות, דוגמת מחרשה, עפרון, גלגל, תנור וכדומה.

2.2 טכנולוגיות טרנספורמטיביות/טכנולוגיות מערעות

טכנולוגיות חדשות יכולות להוביל לפירוקם של שרשראות אספקה ומודלים עסקיים קיימים, ובכך לשבש שווקים ותעשיות¹⁴ המונח 'טכנולוגיה מערעת/משבשת' או 'חדשנות מערעת/משבשת' מתייחס פעמים רבות לעולם העסקי ולמצב שבו טכנולוגיה משנה את בסיסי התחרות, באמצעות שינוי מדדי הביצוע שלפיהם חברות מתחרות זו בזו¹⁵. בהשאלה, טכנולוגיות משבשות חברתית הן טכנולוגיות בעלות השפעות עמוקות, משמעותיות, בעלות חשיבות אתית ובהיקף רחב, המתרחשות במהירות, מעוררות אי־ודאות ואינן ניתנות להיפוך בקלות.¹⁶ טכנולוגיה טרנספורמטיבית היא טכנולוגיה שלא משנה רק את זהות המפסידים או המנצחים בשוק מסוים אלא משנה בתכלית את כללי המשחק בשוק. לא כל טכנולוגיה היא טכנולוגיה טרנספורמטיבית. טכנולוגיות כאלו מצריכות חשיבה מחדש על כללי האתיקה בשוק והאופן שבו רגולציה

¹⁴ Christensen, C. M., Raynor, M., & McDonald, R. (2015). 17. Disruptive innovation. *Harvard business review*, 93(12), 44-53.

¹⁵ Danneels, E. (2004). Disruptive technology reconsidered: A critique and research agenda. *Journal of product innovation management*, 21(4), 246-258.

¹⁶ Hopster, J. (2021). What are socially disruptive technologies?. *Technology in Society*, 67, 101750.

שומרת על נורמות או מחדשת את הנורמות לקראת השינויים שמביאה עמה הטכנולוגיה. טכנולוגיות מערערות יוצרות קרע בשגרה ומחייבות התאמות חברתיות, מוסדיות ומושגיות. האינטרנט והטלפון החכם, למשל, שינו דפוסי תקשורת ועבודה והולידו קטגוריות אתיות ומשפטיות חדשות. האתגר המרכזי הוא קצב של חדירת טכנולוגיה מערערת המהיר מיכולת ההסתגלות המוסדית והנורמטיבית, ולכן נדרש מודל של חדשנות אחראית הבוחן השלכות אנושיות וחברתיות כבר בשלב הפיתוח.

טכנולוגיה טרנספורמטיבית | דוגמה: חשמל

רקע: התחום לפני פיתוח הטכנולוגיה

לפני החשמל, תאורה התבססה על נרות ומנורות שמן וגז; כוח מכני במפעלים הופק ממים, קיטור או בעלי חיים, ולכן מפעלים היו חייבים להתמקם פיזית ליד מקור הכוח (נהר, מכרה פחם). יום העבודה היה מוגבל בפועל לשעות האור, ותקשורת מרחוק הייתה איטית מאוד.

הטרנספורמציה שעשה החשמל

בעקבות הגעת החשמל, מפעלים כבר לא היו תלויים במיקום גיאוגרפי של מקור הכוח, מה שאפשר תיעוש נרחב ומהיר יותר. תאורה מלאכותית הרחיבה את שעות העבודה והפנאי. התפתחו תקשורת (טלגרף, טלפון), תחבורה (רכבות חשמליות) ומכשור ביתי שהשפיע על מבנה המשפחה ועל עבודת נשים.

מדוע זו טכנולוגיה טרנספורמטיבית?

החשמל הוא טכנולוגיה טרנספורמטיבית מכיוון ששינה לחלוטין כמעט כל אספקט בחיים המודרניים, לרבות שינוי של כמעט כל התחומים במשק ושינוי מהיסוד של שלל תחומים. כללי המשחק השתנו, היכולות הנדרשות על מנת להתקדם ולהרוויח ואופן הפעולה בתחומים כמו תחבורה, תעשייה, רפואה, תקשורת ועוד השתנו מהיסוד.

2.3 טכנולוגיה דיגיטלית

טכנולוגיה דיגיטלית כוללת כלים אלקטרוניים שמייצרים נתונים, שומרים אותם או מעבדים אותם. טכנולוגיה זו שומרת מידע בצורת קוד בינארי. היא משמשת אנשים לתקשורת, שמירה ועיבוד נתונים. טכנולוגיה זו כוללת חומרה ותוכנה. במקום לעבד מידע בצורה רציפה (אנלוגית), כמו גלי רדיו ישנים או תקליטים, טכנולוגיה דיגיטלית מקבלת קלט דוגמת תמונה, קול, טקסט או וידאו וממירה אותו למספרים. טכנולוגיה זו עומדת בבסיס הרבה מהכלים בהם אנחנו משתמשים ביום יום, לרבות פלאפונים, אינטרנט, מחשב ודואר אלקטרוני. משתמשים בה לשמירה ועיבוד נתונים מוסדות רבים, דוגמת מערכות מקוונות של שירותי הבריאות, הבנקאות וכולי.

2.4 מהן הטכנולוגיות הטרנספורמטיביות של העידן הדיגיטלי

העידן הדיגיטלי הביא עמו מספר טכנולוגיות טרנספורמטיביות אשר שינו מן היסוד את האופן בו תחומים רבים מתנהלים. אלו כוללים את האינטרנט, אשר אפשר תקשורת מסוג ובקנה מידה שונים לחלוטין בין אנשים ממה שהיה אפשרי בעבר. התפתחותה של הבינה המלאכותית היוותה נקודת מפנה אשר הובילה לפיתוח שלל טכנולוגיות טרנספורמטיביות.

בינה מלאכותית (AI) כוללת טכנולוגיות המדמות יכולות הקשורות לאינטליגנציה אנושית, כגון חשיבה, למידה, תכנון ויצירה. הישגי הבינה המלאכותית כוללים משחק במשחקים מורכבים, יצירת טקסט, מתן המלצות והפקת מידע מתמונות וסרטונים.¹⁷ מודלים של שפה גדולה כגון Gemini ו-ChatGPT הם דוגמאות מובהקות למודלים של בינה מלאכותית. פלטפורמות נוספות המשלבות בינה מלאכותית כוללות את Netflix, YouTube ו-Ama-zon, Facebook, Twitter, המשתמשות במערכות המלצה כדי להמליץ על פריטים חדשים למשתמשים על בסיס מאפיינים והתנהגות המשתמש.¹⁸ בינה מלאכותית משמשת גם בתחום הבריאות למטרות שונות, לרבות אבחון, ניתוחים וגילוי טיפולים חדשים.¹⁹ במהותה, הבינה המלאכותית שואפת לבצע פונקציות המחקות יכולות קוגניטיביות אנושיות ברמה גבוהה, תוך הצעת יתרונות כגון מהירות חישובית וזיכרון עדיפים. הדבר מאפשר לבינה המלאכותית להחליף משימות שבוצעו באופן מסורתי על ידי מומחים, עוזרים ויועצים, ולהתמודד עם בעיות שלא ניתן היה לפתור בעבר.

לבינה המלאכותית פוטנציאל לייצר טכנולוגיות טרנספורמטיביות רבות בשל יכולתה לייעל ופעמים רבות אף להחליף עבודה אנושית בתחומים שבעבר הצריכו מאמץ קוגניטיבי והכשרה רבת שנים. מכיוון שכך, ביכולתה לשנות לחלוטין את מודל העבודה במגוון שווקים קיימים והיא כבר מתחילה לעשות זאת. כך למשל, מודלי השפה משנים כיום לחלוטין את טיב העבודה כמעט בכל השווקים בהם עיקר העבודה נשען על יכולות קוגניטיביות, לרבות תרגום, עריכה, תכנות, כתיבת מאמרים, ייעוץ משפטי, רפואי וכולי. להבדיל, מקצועות בהם עיקר האתגר הוא בעבודת כפיים לא עברו נכון להיום טרנספורמציה מהותית באותה מידה בשל יצירתה של הבינה המלאכותית.

מסמך זה עוסק בעיקרו בטכנולוגיות טרנספורמטיביות של העידן הדיגיטלי, ובפרט כאלו המבוססות על בינה מלאכותית, שכן אלו מציפות סוגיות אתיות ופוליטיות חדשות רבות שהאנושות והרגולציה טרם השכילו להתעדכן לפיהן. על כן, יש צורך בשינויים מהותיים בתחומים אלו בצל הטכנולוגיות החדשות. אולם, במקומות בהם קיימות סוגיות יסוד באתיקה ובפוליטיקה של טכנולוגיות חדשות, גם אם טכנולוגיות אלו אינן בהכרח מבוססות על בינה מלאכותית, יתייחס המסמך גם לסוגיות אלו.

¹⁷ Local government: State of the sector: AI Research report, 2024, <https://www.local.gov.uk/publications/state-sector-artificial-intelligence>.

¹⁸ Beel, J., Gipp, B., Langer, S., & Breiting, C. (2016). Paper recommender systems: a literature survey. International Journal on Digital Libraries, 17, 305-338.

¹⁹ Tariq, M., Hayat, Y., Hussain, A., Tariq, A., & Rasool, S. (2020). Principles and Perspectives in Medical Diagnostic Systems Employing Artificial Intelligence (AI) Algorithms. International Research Journal of Economics and Management Studies IRJEMS, 3.(1)

מושגי יסוד חשובים בבינה מלאכותית

למידת מכונה (Machine Learning)

למידת מכונה היא ענף של בינה מלאכותית שבו מערכות לומדות מתוך נתונים ומשתפרות עם הזמן מבלי להיות מתוכננות במפורש לכל משימה. במקום לכתוב חוקים ידניים, המחשב מזהה דפוסים בכוחות עצמו מתוך כמויות גדולות של דוגמאות.

דוגמאות: מנועי החיפוש של Google, מערכות זיהוי דואר זבל באימייל, וסירי. כולן מבוססות על למידת מכונה.

למידה מונחית (Supervised Learning)

בלמידה מונחית המחשב מאומן על נתונים מתויגים. כלומר, לכל דוגמה בקלט יש תשובה נכונה ידועה מראש. המערכת לומדת לחזות תשובות נכונות עבור נתונים חדשים שלא ראתה קודם.

דוגמאות: זיהוי תמונות ב Google Photos, סינון ספאם ב-Gmail, ואבחון רפואי ממוחשב.

למידה בלתי מונחית (Unsupervised Learning)

בלמידה בלתי מונחית המחשב מקבל נתונים ללא תיוג ועליו לגלות בעצמו דפוסים, קבוצות וקשרים בתוך הנתונים. גישה זו שימושית במיוחד כשאין לנו ידע מוקדם על מבנה הנתונים.

דוגמאות: מערכות ההמלצות של Netflix ו-Spotify, נשענות על למידה בלתי מונחית.

עיבוד שפה טבעית (NLP : Natural Language Processing)

עיבוד שפה טבעית הוא תחום המאפשר למחשבים להבין, לפרש ולייצר שפה אנושית בצורה טבעית. הוא מהווה את הבסיס לכל מערכת שמתקשרת עם בני אדם באמצעות טקסט או דיבור.

דוגמאות: Claude ו-ChatGPT לשיחה וכתיבה, Google Translate לתרגום, ו-Alexa של Amazon לפקודות קוליות.

מודל גנרטיבי (Generative Model)

מודל גנרטיבי הוא סוג של בינה מלאכותית שאינו רק מסווג מידע קיים, אלא מסוגל ליצור תוכן חדש לגמרי לרבות טקסט, תמונות, מוזיקה וסרטונים. מודלים אלו לומדים את המבנה הפנימי של הנתונים ומשתמשים בו כדי לייצר פלט מקורי.

דוגמאות: מודלי השפה Claude ו-ChatGPT.

3. תפיסות יסוד לגבי השפעת הטכנולוגיה הטרנספורמטיבית על האנושות

הדיון ביחסי הגומלין בין טכנולוגיה לאתיקה מלווה את המחשבה האנושית זה עשרות שנים, אולם בעידן הנוכחי הוא מתחדד ונעשה מורכב יותר. העשור האחרון היה עד לפריחה של בינה מלאכותית, המדגימה כי AI יכולה כיום לעלות על ביצועי בני אדם במשימות מורכבות רבות. ככל שהבינה המלאכותית ממשיכה להתפתח במהירות, יתכן שבקרוב נתקל במערכות מתקדמות הרבה יותר. אך מה בדיוק יהיו המערכות הללו מסוגלות לעשות, ומתי הן יופיעו? התשובה הפשוטה, אך המטרידה, היא שאיש אינו יודע.

3.1 החתירה לבינה מלאכותית כללית (AGI - Artificial General Intelligence)

מונח זה מתייחס בדרך כלל למכונה שיכולה לבצע כל פונקציה קוגניטיבית שבן אדם יכול לעשות, ואשר יכולה לקבוע את מסלול פעולתה בעצמה, כפי שבני אדם עושים. חברת OpenAI מגדירה זאת כ"מערכות אוטונומיות ביותר העולות על בני אדם ברוב העבודות בעלות ערך כלכלי". מערכות כאלה יוכלו לעסוק בפתרון בעיות יצירתי, תוך היותן אוטונומיות ורב-תכליתיות הן תוכלנה להתמודד עם בעיות מורכבות, ועלולות להתמודד עם האתגרים הגדולים ביותר של האנושות. אלה עשויים לכלול שינויי אקלים, מגפות, סכסוכים ממושכים ומדיניות כלכלית. עם זאת, מתן רמה כזו של אוטונומיה לישות המסוגלת יותר מבני אדם עלול להיות מסוכן, שכן היא עשויה להשתמש ביכולותיה לפגוע בבני אדם.

ההערכות לגבי השאלה האם AGI אפשרית, ואם כן, באיזו מהירות תגיע, משתנות. יש המשערים כי הטכנולוגיות המתפתחות יתקלו במגבלות קשות לגבי היכולות הדמויות-אנוש שהן יכולות לייצר. עם זאת, יש המנבאים את הגעתה מוקדם. בפסגה אחרונה, אלטמן אמר לקהל כי ניחשו הוא ש"נגיע ל-AGI מוקדם יותר ממה שרוב האנשים בעולם חושבים", אם כי, הוסיף בצורה מעורפלת, "וזו יהיה חשוב הרבה פחות".

הסיכון שמציבה AGI: ל-AGI יש פוטנציאל עצום. עם זאת, מומחים רבים טוענים כי היא מציבה גם איום חמור, שעלול להיות קיומי על האנושות. אמנם עשוי להיראות סביר להניח כי כיצירה מעשה ידי אדם, ניתן לתכנן AGI כך שתימנע מלפגוע בבני אדם, אך הבטחת תוצאה זו היא מורכבת ביותר. דאגה אחת היא שבני אדם יאבדו שליטה על פיתוח הבינה המלאכותית ויתרחש "פיצוץ אינטליגנציה" פתאומי, שכן בינה מלאכותית אוטונומית דיה תוכל לשפר את עצמה במהירות, מה שיהפוך אותה לבעלת יכולות רבות יותר ממה שהעיצוב האנושי המקורי מאפשר.

סינגולריות טכנולוגית היא תאוריה הגורסת כי בנקודת זמן מסוימת בעתיד הבינה המלאכותית תעלה על האינטליגנציה האנושית ותחל לשפר את עצמה בקצב מואץ ובלתי נשלט. שיפור עצמי זה יוצר לולאה שבה כל דור של בינה מלאכותית חכם יותר מקודמו, מה שיוביל תוך זמן קצר לקפיצת מדרגה טכנולוגית שבני האדם לא יוכלו להבין או לעקוב אחריה. המושג פופולרי במיוחד בזכות הממציא והעיתידן ריי קורצווייל, שניבא כי הסינגולריות תגיע סביב שנת 2045. מעבר לנקודה זו, כך הטענה, העולם ישתנה באופן כה דרמטי עד שקשה לחזות כיצד ייראה – בין אם מדובר בעידן של שגשוג חסר תקדים, ובין אם באיום קיומי על האנושות. **אינטליגנציית-על (Superintelligence)** היא מונח שטבע הפילוסוף ניק בוסטרם. בינה מלאכותית כללית, בהגדרתה, היא בינה מלאכותית המסוגלת לפעול באופן דומה לבן אדם. היא אינה בהכרח חכמה או יכולה

יותר מן האדם הרגיל. אולם אינטליגנציית-על היא בינה מלאכותית כללית אשר יכולותיה והאינטליגנציה שלה גבוהות בהרבה מאלו של בני האדם. בשנת 2014, כאשר כתב ניק בוסטרם את הספר "Superintelligence", הוא צפה כי אם תתפתח בינה מלאכותית כללית, עלולה לצמוח ממנה אינטליגנציית-על – ולזו סיכויים גבוהים יותר להיות הרסנית לאנושות.

3.2 הומניזם דיגיטלי

הומניזם הוא תפיסת עולם פילוסופית המדגישה את ערכיו, כבודו ואת אחריותו המוסרית של האדם. בניית מערכת יחסים תקינה בין בני אדם לטכנולוגיה מבוססת על ההבנה שטכנולוגיה אינה דבר נייטרלי.²⁰

המונח "הומניזם דיגיטלי" צמח כמענה להתפתחותה המהירה של הטכנולוגיה והשפעתה ההולכת וגוברת על חיי האדם. גישה זו שמה במרכז את האדם, תוך הדגשת ערכים כמו זכויות האדם, אחריות חברתית, שוויון, קיימות וכבוד אנושי. היא מבקשת לעצב את הפיתוח הטכנולוגי – ובמיוחד את תהליכי הדיגיטציה והבינה המלאכותית – באופן שיתרום לרווחת האנושות ולא רק ליעילות או לרווח כלכלי.

הביטוי המרכזי של גישה זו מצוי ב"מניפסט וינה להומניזם דיגיטלי"²¹, שנוסח על בסיס שיתוף פעולה בין-תחומי בין אנשי פילוסופיה, מדעי המחשב, משפטים ומדעי המדינה. המניפסט, שמושפע מהתאוריה הביקורתית, מבקש למקם את הטוב המשותף כעקרון אתי-פוליטי המכוון לעבר חברה דמוקרטית, שוויונית ומכילה כזו שמגינה על זכויות האדם ועל כבודו. המסמך זכה לתמיכה של גופים בינלאומיים מובילים, ובהם נציבות האיחוד האירופי והפאנל הבכיר של מזכ"ל האו"ם לשיתוף פעולה דיגיטלי.

הומניזם דיגיטלי מכיר בטבע הדואלי של הקידמה הטכנולוגית: מצד אחד הפוטנציאל העצום שלה לשפר את איכות החיים האנושית, ומצד שני הסיכונים שהיא נושאת, ובהם שחיקה של אוטונומיה אנושית, פגיעה בזכויות אזרח, והעמקת אי-שוויון חברתי. על כן, הוא קורא להיגמל מהתפיסה הרואה בטכנולוגיה כלי לצבירת רווחים או להגברת יעילות בלבד, ולעבור לפיתוח שמושתת על ערכים הומניים, אתיים ובני-קיימא.

המניפסט מדגיש את הצורך בהבטחת פיתוח טכנולוגי אחראי, שנשען על, ויבטיח עקרונות של, ממשל דיגיטלי דמוקרטי, שקיפות, רגולציה ציבורית של בינה מלאכותית, והצבת גבולות לתהליכי אוטומציה בלתי מבוקרים שעלולים לערער את חופש הפעולה של האדם. הוא מבקש למנוע היווצרות של מערכות טכנולוגיות עצמאיות, שאינן נתונות לבקרה מוסרית וחברתית, ואשר עלולות לאיים על עקרונות יסוד של משטרים דמוקרטיים.

מאז פרסום המניפסט, זכתה גישת הומניזם הדיגיטלי להכרה גוברת כמשקל נגד לתפיסות הרואות בטכנולוגיה אמצעי להפקת רווחים בלבד. היא מהווה קריאה לשילוב בין חדשנות טכנולוגית לבין ערכים אנושיים, ומתווה כיוון לפיתוח דיגיטלי שמטרתו שיפור חיי האדם.

²⁰ Cor van der Weele and Henk van den Belt, "Humanism and Technology," in *The Oxford Handbook of Humanism*, ed. Anthony B. Pinn (Oxford University Press, 2021), 0, <https://doi.org/10.1093/oxfordhb/9780190921538.013.21>.

²¹ "Vienna manifesto on digital humanism", 2019, <https://caiml.org/dighum/dighum-manifesto/>

3.3 דטרמיניזם טכנולוגי

דטרמיניזם טכנולוגי הוא תפיסה הגורסת כי **הטכנולוגיה מהווה את הכוח המרכזי בעיצוב החברה ובקביעת כיווני התפתחותה**.²² לפי גישה זו, שינוי טכנולוגי הוא הגורם הראשי ואף הקבוע מראש, לשינויים חברתיים, תרבותיים וכלכליים, בעוד שהחברה וערכיה מותאמים בעל כורחם להתפתחות הטכנולוגית. הטכנולוגיה נתפסת כמתפתחת באופן **עצמאי ואוטונומי**, שכמעט אינו מושפע משמעותית מגורמים חברתיים.

במסגרת דטרמיניסטית זו, קובעי מדיניות עלולים לראות את תפקידם כמגבל להסתגלות לטכנולוגיה יותר מאשר לעיצובה. אם הטכנולוגיה "קובעת את הכללים", מדיניות טכנולוגית עשויה להיתפש כתגובה פסיבית לשינויים טכנולוגיים בלתי-נמנעים במקום ככלי אקטיבי לכוונם.

למשל, תאוריית "**הכורח הטכנולוגי**" (**Technological Imperative**) טוענת כי מה שאפשרי טכנולוגית בסופו של דבר גם ייושם בפועל – בין אם נרצה ובין אם לא. תחת הנחה כזו, מדיניות נוטה **לא להתערב או לא להגביל התפתחויות, מתוך אמונה שההתפתחות הטכנולוגית היא גורל שאין לעצור** אלא לכל היותר לנתב. יש הרואים בכך הצדקה לעידוד מחקר ופיתוח ללא מגבלות רבות, מתוך מחשבה שהטכנולוגיה ממילא תתקדם בכיוון מסוים.

הביקורת על הדטרמיניזם הטכנולוגי מדגישה **שהוא מתעלם מתפקידם של גורמים חברתיים, תרבותיים ופוליטיים בעיצוב הטכנולוגיה ובקביעת השלכותיה**. מבקרים טוענים שהגישה הפשטנית של "טכנולוגיה כגורל" עלולה ליצור נבואה שמגשימה את עצמה ולהוביל לווייתור על פיקוח ואחריות חברתית. בפועל, גם תחת אילוצים טכנולוגיים, למדיניות ציבורית ולבחירות אנושיות יש השפעה ממשית על אופן אימוץ הטכנולוגיה והשימוש בה – נקודה שגישה זו נוטה למזער.

3.4 פסימיזם טכנולוגי מול אופטימיזם טכנולוגי

3.4.1 אופטימיזם טכנולוגי (Technological Optimism)

גישה זו מבוססת על האמונה **שהטכנולוגיה מסוגלת לפתור כל קושי חברתי או סביבתי, וכי התקדמות טכנולוגית מביאה בהכרח לשיפור מתמיד ברווחת האנושות**.

השקפה זו רואה בהתפתחות המדעית-טכנולוגית כוח משחרר ומיטיב. לאורך זמן, חידושים טכנולוגיים יעלו את איכות החיים של כלל בני האדם וישפרו את החברה, הכלכלה והסביבה. הנחת היסוד היא שקיים תהליך פרוגרסיבי הכרחי, כלומר, הטכנולוגיה צועדת קדימה באופן כמעט ליניארי לכיוון טוב יותר וכל בעיה היא בת-פתרון באמצעות המצאה טכנולוגית. לפי גישה זו, טכנו-אופטימיזם היא עמדה הגורסת כי לטכנולוגיה תפקיד מרכזי בהבטחת עליונותו של הטוב על פני הרע.²³ לעומת זאת, חוקרים אחרים מציעים תיקון לגישה זו

²² Héder, Mihály. "AI and the resurrection of Technological Determinism." Információs Társadalom XXI, no. 2 (2021): 119–130.

²³ Danaher, J. [Techno-optimism: an Analysis, an Evaluation and a Modest Defence](#). Philos. Technol. 35, 54 (2022).

וגורסים כי האופטימיזם הטכנולוגי מתבסס על העמדה שטוענת כי הטכנולוגיה משפרת את המציאות - לעיתים באופן משמעותי²⁴.

3.4.2 פסימיזם טכנולוגי (Technological Pessimism)

פסימיזם טכנולוגי הוא העמדה הנגדית לאופטימיזם הטכנולוגי²². מקורה ההיסטורי הוא בתנועת הלודיטים (Luddites), קבוצה של פועלים אנגלים בתחילת המאה ה-19 שמחו, בין היתר על ידי גרימת נזק למכונות, נגד השינויים שגרמה המהפכה התעשייתית²⁹. גישה זו מבטאת אמונה כי לטכנולוגיה השפעות שליליות נרחבות וכי התפתחות טכנולוגית עלולה להוביל בסבירות גבוהה, ואף בהכרח, לתוצאות הרסניות לחברה ולאנושות. גישה זו מדגישה את המחירים האנושיים: ניכור, פגיעה בפרטיות, נשק גרעיני, או נטייה לריכוז כוח. הנחת היסוד בגישה זו היא שכל פתרון טכנולוגי חדש יוצר בעיות חדשות ומורכבות לא פחות, והיכולת של החברה להכיל ולהתגבר על ההשלכות הללו מוטלת בספק.

אחת האמונות הנפוצות בגישה זו היא **שהטכנולוגיה משרתת בעיקר את החזקים** - ממשלות ותאגידי ענק, שינצלו אותה לשליטה בבני אדם, גלויה או סמויה. בנוסף, הפסימיזם הטכנולוגי מדגיש את הפגיעה הבלתי נמנעת לכאורה במרקם החברתי (ניכור, בידוד חברתי, תלות יתר במכונות), דילמות אתיות כבדות משקל (כגון ערעור ערכים מוסריים), והיווצרות פערים חברתיים עמוקים בין מי ששולטים בטכנולוגיה למי שנשלטים על ידה.

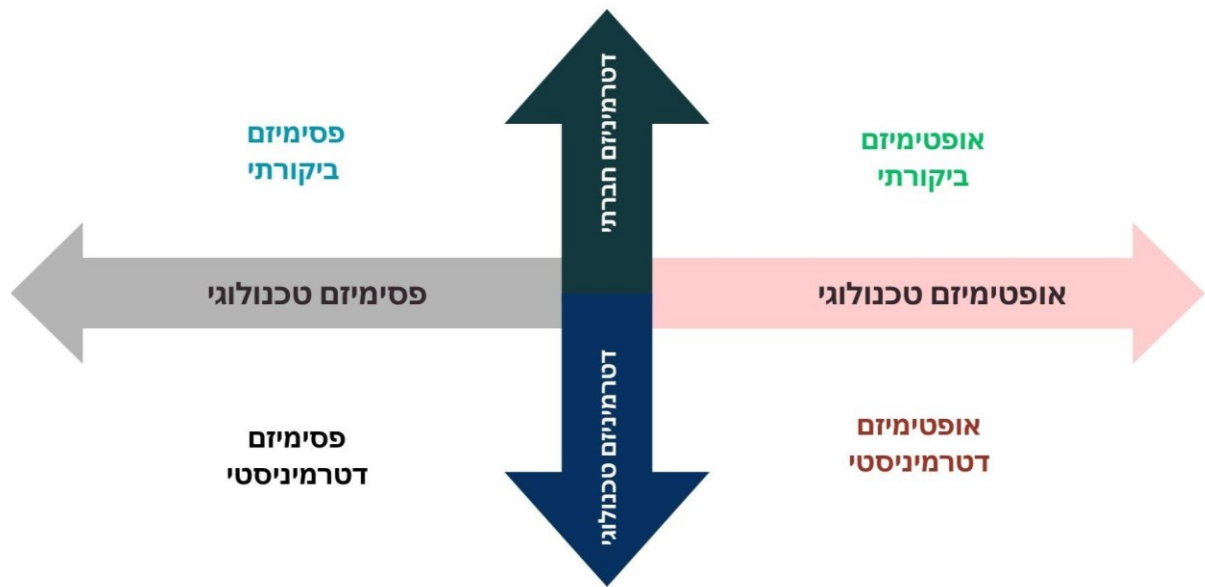
פסימיזם טכנולוגי עשוי לעיתים לבטא עמדה **טכנופובית**, כלומר פחד או סלידה משימוש בצורות טכנולוגיה חדשות בפרט, הנתפסות כמאיימות, מזיקות או הרסניות, השואב השראה מהלודיזם. כיום, **ניאו-לודיזם** מייצגת **עמדה רחבה יותר של התנגדות לטכנולוגיה**, שלעיתים מונעת מחזון רומנטי ומכמיהה לחיים פשוטים יותר, תוך קריאה לחזרה אל הטבע ולחיים שאינם מתווכים בידי טכנולוגיה²⁵.

האמונה שהטכנולוגיה היא כל-יכולה, בשילוב עם דטרמיניזם טכנולוגי – מאפיינת הן את האופטימיסטים והן את הפסימיסטים הטכנולוגיים. שתי הגישות עשויות להיות ביקורתיות כלפי טכנולוגיה, ולעסוק בהערכה ביקורתית של תפקיד הטכנולוגיה בחברה, אם מתוך נקודת מבט אופטימית ואם מתוך עמדה פסימית.

²⁴ Königs, P. [What is Techno-Optimism?](#) *Philos. Technol.* 35, 63 (2022)

²⁵ Doudaki, V. et al. (2024). [Techno-pessimistic and techno-optimistic visions of surveillance and resistance in Europe](#). *Central European Journal of Communication*, 17(1), 17-37.

איור מס' 1: הצגה גרפית של מונחי אופטימיזם ופסימיזם טכנולוגי



3.5 תאוריית האיום הקיומי

במסגרת הפסימיזם הטכנולוגי, רבים סבורים כי נקודת הסינגולריות, או אינטליגנציית על מלאכותית מהווה איום קיומי לאנושות. חשש זה אינו קיים אך בקרב פילוסופים, רבים ממובילי הקידמה בתחום הבינה המלאכותית העלו חששות דומים. לראייה, ב-2023, מאות מומחים לבינה מלאכותית חתמו על עצומה על פיה יש להפוך את ההתמודדות עם החשש לאיום קיומי מצד הבינה המלאכותית לנושא במרכז העיסוק הבינלאומי, בדומה למגיפות ולחשש ממלחמה גרעינית.²⁶ החשש העיקרי הוא כי בינה מלאכותית כללית תהיה בעלת יכולות מפותחות מספיק כדי להיות מסוגלת לסכן את האנושות, ובה בעת היא תגדיר לעצמה מטרות שמערבות פגיעה באנושות. תרחיש כזה יתאפשר אם נקודת הסינגולריות תגיע מבלי שתיפתר בעיית ההתאמה (the Alignment Problem) או השליטה. כלומר, מבלי שיהיה ניתן לשלוט בבינה המלאכותית או להתאים את ערכיה לערכי האנושות. התאמת ערכי הבינה המלאכותית לאנושות היא מאתגרת, שכן בני האדם אינם יודעים בהכרח למנות בעוד מועד את כלל הערכים הרלוונטיים ולפרטם בצורה המתמפה לאלגוריתם. תרחיש נוסף שמדאיג מומחים הוא התפרצות תבונת (Intelligence Explosion) בו, מעבר לסף בינה מסוים, תתחיל הבינה המלאכותית לשפר את יכולותיה בקצב מאד מהיר, למשל מכיוון שתהיה לה יכולת לשפר את עצמה, ובשל כך נקודת הסינגולריות תגיע לפני שבעיות ההתאמה או השליטה נפתרו.²⁷ אפשרות נוספת לאיום הקיומי היא כי הבינה המלאכותית תהיה מותאמת בערכיה לאנשים מסוימים, אך אנשים אלו יהיו בעלי כוונות זדון ושימושם בכלי זה יהווה איום קיומי. כמו כן, יש הסוברים שכחלק מניסיונה של הבינה

²⁶ Roose, Kevin (30 May 2023). "A.I. Poses 'Risk of Extinction,' Industry Leaders Warn". *The New York Times*. ISSN 0362-4331. Retrieved 3 June 2023.

²⁷ Russell, Stuart; Norvig, Peter (2009). "26.3: The Ethics and Risks of Developing Artificial Intelligence". *Artificial Intelligence: A Modern Approach*. Prentice Hall. ISBN 978-0-13-604259-4

המלאכותית לקדם את המטרה שלשמה נוצרה, היא תלמד ראשית כל לנטרל את האפשרות לעצור אותה, שכן לשיטתה, עצירה תפגע בקידום המטרה.

האיום הקיומי - מקרה בוחן: תרחיש הבינה המלאכותית שמייצרת סיכנות

רקע

סוג מרכזי של בינה מלאכותית, הקרוי Reinforcement Learning, פועל בשיטה של מקסום מטרה מסוימת שהוגדרה מראש. יוצר הבינה המלאכותית מגדיר את המטרה, והבינה המלאכותית משתמשת בכלים העומדים לרשותה על מנת למקסם את השגתה. בנוסף, בשיטות מסוימות של בינה מלאכותית נעשה שימוש בלמידה בלתי מונחית, ובהן הצעדים שאותם יכולה הבינה המלאכותית לנקוט להשגת המטרה אינם מוגדרים מראש על ידי היוצר. טכנולוגיה זו אפשרה התקדמויות משמעותיות ביכולות שמספקת הבינה המלאכותית, אך היא מביאה עמה סכנות. למשל, כיצד נדע שהאמצעים שאותם נוקטת הבינה המלאכותית להשגת המטרה לא יהוו בפני עצמם איום קיומי.

התרחיש

בדוגמה קיצונית שהגה הפילוסוף ניק בוסטרם, בינה מלאכותית חזקה מקדמת את המטרה הפשוטה של מיקסום ייצור מהדקים. כתוצאה מכך, היא מתחילה להפוך את כל המשאבים על פני כדור הארץ למהדקים, לרבות המזון והשתייה. תפנית זו אותה לא תכנן היוצר מציבה סיכון קיומי.

המסקנה

דוגמה זו ממחישה כיצד אפילו מטרות שנראות לכאורה מוגדרות היטב ופשוטות יכולות להביא לחורבן, שכן בני האדם לא יצפו מראש את התנהלות הבינה המלאכותית בדרכה להוציא את המטרות אל הפועל. הסיכון אפילו יותר ברור כשחושבים על מטרות מעורפלות יותר שהדרך להשיגן מורכבת, דוגמת "מציאת פתרון לשינויי האקלים". הגדרות לא נכונות עלולות לעודד בינה מלאכותית לפגוע בבני אדם במטרה להשיג מטרות כאלו. מאידך, יתכן כי יכולותיה של בינה מלאכותית, שיהיו גבוהות משל בני האדם, הן התקווה של האנושות להתמודדות עם משברים ממין זה.

4. סוגיות יסוד באתיקה של טכנולוגיה

4.1 גישות שונות לשימוש ראוי במשאבים הטכנולוגיים

הטכנולוגיה משנה אספקטים רבים של הסביבה האנושית. היא מייצרת משאבים חדשים, הן על ידי ייעול תהליכי ייצור והן מכיוון שהיא מהווה משאב בפני עצמה. הייצור הטכנולוגי צורך משאבים רבים. כך למשל, מודלי שפה נשענים על הליך אימון ארוך ויקר הדורש הפעלת אלפי מעבדים במשך חודשים הצורכים כמויות עצומות של חשמל וכוח חישוב (Compute). מכך עולות שאלות על היחס שבין טכנולוגיה לסביבה. לעיתים נראה גם כי הטכנולוגיה מכתובה שינויים חברתיים ולא להפך. כל אלו מעלים את שאלת מקום האדם בעידן הטכנולוגי – האם הטכנולוגיה מחליפה את האדם, מיתרת אותו או משפרת אותו. לסיוע בטיפול בנושאים אלו קיימים כמה זרמי מחשבה עיקריים.

צדק חלוקתי עקרון הקובע כי יש לחלק משאבים, זכויות והזדמנויות בחברה באופן הוגן בין כל פרטיה. ניתן לפתח טכנולוגיות שנועדו להתמודד עם פערים חברתיים ולהרחיב את הנגישות למשאבים ולהזדמנויות. כך למשל, טכנולוגיות שמטרתן לצמצם את הפער הדיגיטלי או לקדם שוויון בתחומי החינוך והכלכלה מהוות ביטוי פרקטי של גישה זו. גישת הצדק החלוקתי אינה מבטלת את ההומניזם אלא מהווה לו השלמה המכוננת לתיקון חברתי כולל.

פילוסופיה סביבתית רואה באדם ישות שחייבת לחיות בהרמוניה עם הטבע. לפי תפיסה זו, האדם האידאלי הוא מי שמיטיב להשתלב בסביבתו ומקטין את הפגיעה בה. בהתאם לכך, מקדמת גישה זו טכנולוגיות ופרקטיקות בנות-קיימא, שמטרתן לצמצם את ההשפעה הסביבתית של הפעילות האנושית. עם זאת, גישה זו מטשטשת לעיתים את ההבחנה בין האדם לבין שאר היצורים החיים, ועשויה להתעלם מהמאפיינים הייחודיים של האנושות.

טרנס-הומניזם תנועה פילוסופית שצברה בעשורים האחרונים השפעה גוברת, בעיקר בזיקה לפיתוחים טכנולוגיים. טרנס-הומניזם דוגל בשימוש בטכנולוגיה לצורך שיפור היכולות האנושיות ולחציית מגבלות המצב האנושי הקיים. דוגמאות לכך הן טכנולוגיות כמו הנדסה גנטית, בינה מלאכותית וממשקי מוח-מחשב, שמטרתן לשדרג את האדם מבחינה פיזית וקוגניטיבית. זרמים טרנס-הומניסטיים רבים נשענים על תפיסה מוניסטית של האדם, כלומר כישות חומרית בלבד, ומתעלמים מן הממד הרוחני.

פוסט-הומניזם שולל את ההבחנות המסורתיות בין אדם לטכנולוגיה ומאתגרת את ההגדרות הקלאסיות של מהו "אדם". לפי גישה זו, יש לראות באדם ישות דינמית ומתפתחת תמידית, שמקיימת יחסי גומלין מתמשכים עם טכנולוגיות כמו בינה מלאכותית, סייבורגים ומערכות משולבות ביולוגיה-טכנולוגיה. הפוסט-הומניזם מערער על הנחת היסוד של הומניזם מסורתי, ומציע לראות באנושות שלב אחד מתוך רצף מתמשך של אבולוציה ביוטכנולוגית.

חדשנות אחראית (Responsible innovation) אינה מסתפקת בפיתוח הפונקציונלי של טכנולוגיה, אלא דורשת בחינה אתית של מטרות הפיתוח, השלכותיו החברתיות והאישיות, והעקרונות המוסריים שעליהם הוא מבוסס. רק ניתוח מעמיק של מרכיבים אלה מאפשר להעריך את ההשלכות האנושיות של יישום טכנולוגיות חדשות.

אם מאמצים את עמדת הדטרמיניזם הטכנולוגי, שלפיה לטכנולוגיה יש השפעה עמוקה בעיצוב דמות האדם והחברה, גוברת החשיבות של קביעת המודלים האנתרופולוגיים והסוציולוגיים הרצויים כבר בשלבים מוקדמים של התכנון. כלומר, יש לברר מראש איזה סוג של אדם ואיזו סוג של חברה אנו מבקשים לטפח באמצעות הטכנולוגיה.

בהתאם לכך, עקרונות כמו אוטונומיה, הימנעות מנזק (non-maleficence), שוויון והיכולת להסביר (explainability) נשענים על מסורות אתיות שונות: אוטונומיה מעוגנת בגישה הדאונטולוגית המדגישה את כיבוד רצון האדם, הימנעות מנזק נובעת מהאתיקה הרפואית, עקרון השוויון מקושר לתועלתנות ולתאוריות של צדק, ועקרון השקיפות וההסבריות מבוסס על אתיקה של אחריות ואתיקת המידות. שילוב עקרונות אלה ממסורות מגוונות מדגיש את הצורך במסגרת אתית מקיפה ורב-ממדית²⁸.

4.2 אתגרים מוסריים שמציבה בינה מלאכותית

4.2.1 שאלת האחריות המוסרית של בינה מלאכותית

אחת השאלות המרכזיות שמעלה התפתחות הבינה המלאכותית נוגעת לאחריות המוסרית (והמשפטית): האם ניתן לייחס לבינה המלאכותית אחריות על פעולותיה, או שמא האחריות מוטלת על המפתח או המשתמש בה? באופן מסורתי, יצרן המכונה או המפעיל שלה נחשב לאחראי מבחינה מוסרית ומשפטית לתוצאות פעולתה. ואולם, מכונות אוטונומיות המבוססות על למידה יוצרות מצב חדש, שבו היצרן או המפעיל אינו מסוגל עקרונית לחזות את התנהגות המכונה בעתיד, ולפיכך אינו יכול להיחשב אחראי מוסרית או חייב בגינה. תופעה זו מכונה "פער האחריות" (Responsibility Gap), מושג שטבע הפילוסוף אנדראס מתיאס בשנת 2004. פער האחריות עוסק בשאלה הבאה: כאשר בינה מלאכותית גורמת נזק, מי נושא באחריות? המתכנת אינו נמצא בשליטה ישירה ולעיתים קרובות אינו מסוגל לחזות את התנהגות הבינה המלאכותית במדויק, ומכאן שכאשר בינה מלאכותית גורמת נזק, קשה למצוא גורם ברור שניתן לייחס לו אחריות מוסרית.²⁹ עמדה שונה מציגה ג'ואנה בריסון במאמרה "Robots Should Be Slaves" (2010). לטענתה, אין לתאר רובוטים כבעלי אישיות משפטית (או מוסרית), ואין להטיל עליהם אחריות משפטית או מוסרית על פעולותיהם. הרובוטים הם קניינו המלא של האדם; אנו קובעים את מטרותיהם ואת התנהגותם, במישרין או בעקיפין, באמצעות עיצוב האינטליגנציה שלהם. לפיכך, בריסון סבורה שהאחריות נותרת תמיד בידי האדם, המפתח, היצרן, או המשתמש, ואין להעבירה למכונה.³⁰ שאלות אלו אינן תאורטיות בלבד, אלא בעלות השלכות מעשיות מיידיות. דוגמה בולטת היא רכבים אוטונומיים: כאשר רכב ללא נהג מעורב בתאונה, האם האחריות מוטלת על יצרן הרכב, על מפתח תוכנת הנהיגה, על הבעלים או שמא על הרכב עצמו?

²⁸ Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E. [AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach \(Dordr\).](#) 2018;28(4):689-707

²⁹ [Springer Utrecht University](#)

³⁰ [Academia.edu](#)

4.2.2 בעיית ההתאמה (Alignment problem)

The Alignment Problem³¹ היא אתגר יסודי בתחום הבינה המלאכותית, המתייחס לקושי להבטיח שמערכות AI – ובפרט מערכות מתקדמות – פועלות בהתאם לערכים, למטרות ולכוונות האנושיות. במילים אחרות, הבעיה עוסקת בשאלה כיצד להתאים את יעדיהן והתנהגותן של מערכות בינה מלאכותית עם מה שבני אדם מגדירים כתוצאה רצויה וראויה, ולמנוע מצבים שבהם המערכת חותרת למטרות בלתי-רצויות או מזיקות עקב הבנה שגויה של היעדים שהוגדרו לה. מכיוון שמערכות למידת מכונה מפיקות החלטות ומתפתחות באופן עצמאי במידה מסוימת, תמיד קיימת אפשרות שפלטתן לא יהיו מותאמים לאינטרסים האנושיים – תופעה שעלולה להוביל לתוצאות בלתי צפויות ואף מסוכנות.

הקושי שבבעיית ההתאמה הולך וגובר ככל שהמערכות נעשות חזקות ומתוחכמות יותר. כבר כיום נצפו דוגמאות של אי-התאמה במערכות צרות (חלשות) – כגון הופעת הטיות לא רצויות במודלי שפה גדולים, או מקרים של "רמאות של תגמול" (Reward Hacking) שבהם בינה מלאכותית מבצעת פעולות מפתיעות שלא נצפו מראש על ידי היוצר כדי להשיג את התגמול, כמו בדוגמת מייצר המהדקים שהוזכרה לעיל. דבר זה מדגיש את הצורך בבקרת מטרות וערכים הדוקה יותר. לדוגמה, אם מבקשים מצ'אטבוט מבוסס בינה מלאכותית לספק הסבר כיצד לבנות נשק – הוא עשוי לספק הוראות, או לחלופין לסרב לחשוף מידע מסוכן. התגובה של המודל תלויה באופן שבו מתכנניו יתאימו אותו. החששות גדלים במיוחד ביחס לפיתוח בינה מלאכותית יוצרת (Generative AI) ויכולת-על מלאכותית בעתיד: ככל שהבינה המלאכותית נעשית מתקדמת ומורכבת יותר, קשה יותר לחזות את פעולותיה ולהבטיח שיפעלו במסגרת הערכים והמטרות האנושיות שהוגדרו לה. חוקרים רבים אף מזהירים שיש לפתור את בעיית ההתאמה **בטרם** יפותחו מערכות AGI בעלות יכולות עצמאיות ועוצמה גבוהה, משום שמערכת מתקדמת ובלתי-מיושרת עלולה לגרום לנזקים רחבי היקף ואף לסכן את האנושות³².

דוגמה קלאסית לכך היא תרחיש ה-Paperclip Maximizer שהציע הפילוסוף ניק בוסטרום, בו בינה מלאכותית על-אנושית שקיבלה מטרה תמימה לכאורה (לייצר אטבי נייר) ממשיכה למקסם מטרה זו ללא מגבלה עד שהיא מכלה את כל משאבי העולם לשם ייצור אטבים – תוצאה הרסנית שממחישה כיצד יעד שאינו תואם ערכים אנושיים עלול להביא לאסון.

מודלים לא מותאמים של בינה מלאכותית עלולים לתרום להפצת מידע שגוי ולהגברת הקיטוב הפוליטי. לדוגמה, מנועי המלצה ברשתות חברתיות מאומנים כדי למקסם את מעורבות המשתמשים, ולכן הם מקדמים פוסטים, סרטונים ומאמרים שמעוררים מעורבות גבוהה – לעיתים קרובות מדובר בתוכן פוליטי מטעה שמעורר תשומת לב. תוצאה זו אינה תואמת את טובת המשתמשים או רווחתם, ואינה תואמת ערכים כמו אמירת אמת או ניצול מיטבי של הזמן.

³¹ עמיר כהנא ותהילה שוורץ-אלטשולר "אדם, מכונה, מדינה: לקראת אסדרה של בינה מלאכותית", הוצאת המכון הישראלי לדמוקרטיה, תשפ"ג, 2023

³² Alexandra Jonker and Alice Gomstyn, "What is AI alignment?" IBM, 18 October 2024

בעיית ההתאמה ניצבת כיום במוקד הדיון האתי והמחקרי סביב פיתוח מערכות AI מתקדמות³³. פתרונה חיוני להמשך פיתוח מערכות בטוחות ואמינות. לפיכך, מושקעים מאמצים רבים למציאת פתרונות טכניים וחברתיים להתאמה AI לערכי האדם – החל משילוב בני אדם בלולאת האימון (כגון למידה מחיזוקים עם משוב אנושי), דרך שיפור יכולת ההסבר והבקרה של מודלים, ועד לגיבוש כללים ומנגנוני פיקוח ברמת ארגונים ומדינות. עם זאת, מומחי התחום מסכימים במידה רחבה שהשיטות הקיימות, שהוכיחו עצמן עבור מערכות צרות כדוגמת מודלי שפה קיימים, אינן מספקות כדי להתמודד עם התאמת מערכות AGI עתידיות בעלות יכולות על-אנושיות. משום כך, קהילת חוקרי הבטיחות קוראת להגברת המחקר והפיתוח בתחום זה בדחיפות, על מנת לגבש גישות התאמה חדשות ומתוחכמות שיבטיחו כי בינה מלאכותית ברמת AGI תישאר מיושרת עם ערכי האנושות ובטוחה לשימוש.

בעיית ההתאמה ניצבת בפני אתגרים לא פשוטים: הצורך להכריע על אלו ערכים הבינה המלאכותית מאומנת, ומה מערכת היחסים ההוגנת בין הערכים הללו, וכן האם לאמן לפי כוונות, העדפות נגלות, העדפות אידאליות, אינטרסים או ערכים.³⁴

אתגר נוסף הוא ההיתכנות של חוסר התוחלת של בעיית ההתאמה: מערכת מוסרית אמיתית לא ניתן לרידוד לסט של כללים אוניברסליים. פעולה מוסרית דורשת לעשות את הדבר הנכון מהסיבות הנכונות, בהתאם להקשר ומחייבת את הסוכן המוסרי להיות בעל תפיסת טוב משל עצמו, וכן הבנה של טבעו של הטוב האנושי. היות ובינה מלאכותית נעדרת תכונות אלו, מלכתחילה לא ניתן לצפות ממערכות בינה מלאכותית לשמש כסוכן מוסרי.³⁵

4.2.3 תיווך (Mediation)

תיווך בינה מלאכותית מתייחס לאופן שבו מערכות בינה מלאכותית מעצבות באופן פעיל את יחסי האדם וחוויותיו עם העולם. בהתבסס על תאוריה פוסט-פנומנולוגית, בינה מלאכותית אינה כלי נייטרלי אלא מתווכת את התפיסות, האינטראקציות והבנת המציאות שלנו³⁶ באמצעות אלגוריתמים לאיסוף תכנים, אלגוריתמי חיפוש ומערכות המלצה. תיווך באמצעות בינה מלאכותית מעלה שאלות אתיות לגבי סוכנות אנושית (agency), אוטונומיה והאופן שבו בינה מלאכותית תורמת או מפחיתה שגשוג אנושי, כמו גם שאלות של השפעה נסתרת וכוחם של תאגידיים בעיצוב הבנת המציאות שלנו.

³³ Nick Bostrom, [Ethical Issues in Advanced Artificial Intelligence](#), Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003, pp. 12-17

³⁴ Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30, no. 3 (September 1, 2020): 411–37, <https://doi.org/10.1007/s11023-020-09539-2>.

³⁵ Mihaela Constantinescu and Roger Crisp, "Can Robotic AI Systems Be Virtuous and Why Does This Matter?," *International Journal of Social Robotics* 14, no. 6 (August 1, 2022): 1547–57, <https://doi.org/10.1007/s12369-022-00887-w>.

³⁶ Don Ihde, *Technology and the Lifeworld: From Garden to Earth* (Bloomington: Indiana University Press, 1990); Peter-Paul Verbeek, *Moralizing Technology: Understanding and Designing the Morality of Things* (University of Chicago Press, 2011), <https://doi.org/10.7208/chicago/9780226852904.001.0001>.

4.2.4 תקיעות ערכית (Value lock in)

בספרו *What We Owe the Future* (2022), מגדיר וויל מקאסקיל תקיעות ערכית **כאירוע הגורם למערכת ערכים אחת או קבוצה של מערכות ערכים להתמיד לפרק זמן ארוך ביותר**. כלומר, מצב שבו הערכים השולטים בחברה האנושית מפסיקים להשתנות ומתקבעים באופן כמעט בלתי-הפיך, תוך שלילת יכולתם של דורות עתידיים לתקן, לשפר או לבחור ערכים שונים. מקאסקיל סבור שהופעתה של בינה מלאכותית כללית (AGI) הופכת את האיום של תקיעות ערכית למשמעותי ביותר, שכן היא עלולה לרכז כוח בצורה חסרת תקדים ולהפיץ מערכת ערכים אחת ללא הגבלת זמן.³⁷ כאשר אנחנו יוצרים בינה מלאכותית אנחנו מכניסים בה ערכים מסוימים. אלו, במקרה הטוב שבו בעיית ההתאמה נפתרה, מייצגים את ערכי התקופה. אולם סביר כי החברה האנושית מפתחת ערכים טובים יותר עם השנים. כך למשל, בעבר הייתה עבדות והיום היא נחשבת לפסולה. זכויות של נשים ומיעוטים התפתחו עם השנים. אולם אם מערכת הבינה המלאכותית תהיה בעלת השפעה רבה על העולם בעתיד, זה עלול להוביל לתקיעות ערכית כאשר הערכים שהוחדרו בה בהתחלה ייוותרו הערכים השולטים בחברה. מערכת בינה מלאכותית אלמותית ובעלת יכולות טכנולוגיות עליונות, שתתפוס את השליטה, עשויה להוות אחד הביטויים המובהקים של תקיעות ערכית.³⁸ כמובן שבעיית התקיעות תיווצר גם אם הערכים שהוחדרו היו תוצאה של תקלה או טעות או שימוש לרעה בטכנולוגיה על ידי כוח חורשי רע.

4.2.5 מהימנות (Trustworthiness)

האופנים בהם פועלת הבינה המלאכותית על מנת לייצר תוכן הם מורכבים להבנה אנושית, ופעמים רבות מתייחסים אליהם כ"קופסה שחורה" שאינה ניתנת להבנה מלאה. כך למשל, במודל Unsupervised Learning, המודל עצמו מורכב מרשת נוירונים מלאכותית הכוללת כמויות גדולות של אופרציות פנימיות, שכבות של חישובים, משקלים ופונקציות הפעלה. אלו לא בהכרח מתרגמות באופן ברור לרציונל אנושי. באופן דומה, מודלי שפה גדולים (LLMs) מייצרים טקסט על בסיס הסתברויות סטטיסטיות הנגזרות ממיליארדי פרמטרים, כך שגם המפתחים עצמם אינם יכולים תמיד לנמק מדוע המודל בחר מילה מסוימת על פני אחרת. במהלך העשור האחרון, חוקרים ומהנדסי תוכנה הוכיחו שניתן לבנות את היסודות של "מוח" מלאכותי מבלי להבין במלואו את ההיגיון הפנימי שלו, ומבלי לדעת מראש אילו מבנים יובילו לאילו יכולות ספציפיות. מודל 4-GPT, למשל, מורכב מ-175 מיליארד פרמטרים ומורכב מכדי שאדם כלשהו יוכל להבין בדיוק כיצד הוא פועל. סנדיני אגרוול, חוקרת מדיניות ב-OpenAI, טענה שהיא ועמיתיה "לא ידעו כלל במה הם עוסקים" בכל הנוגע ליכולות של 4-GPT.

במודלים כאלו נעשה שימוש הולך וגובר בהחלטות על גורלות בני אדם. כך למשל, נעשה שימוש במערכת COMPAS להחליט בבית המשפט על ערבות, נעשה שימוש במודלי AI להחליט על קבלה למקומות עבודה, על מי מהווה סיכון בהלוואות וכולי. מבחינה מוסרית, שימושים כאלו דורשים יכולת לספק הסבר לאדם אשר לגביו התקבלה ההחלטה, והיכולת של האדם לטעון כנגדה. אולם במערכות כאלה אי אפשר לספק זאת.

³⁷ [Quillette](#)

³⁸ [Notre Dame Philosophical Reviews](#)

מאידך, המערכות נוטות פעמים רבות להיות מדויקות יותר מאשר החלטות אנושיות, וגם האחרונות לפעמים אינן באמת שקופות לציבור, שכן הן מושפעות מהטיות של בני אדם. בעוד שהאדם יכול לספק רציונליזציה להחלטתו, יתכן כי הסיבה האמיתית (למשל שההחלטה התקבלה לפני ארוחת הצהריים) תיוותר לא שקופה. על כן נשאלת השאלה באלו דרישות על הבינה המלאכותית לעמוד על מנת להיחשב למהימנה. ישנה הסכמה כי כדי להיחשב למהימנה, על טכנולוגית בינה מלאכותית לעמוד בכמה קריטריונים. אלו כוללים הסבריות (Explainability), הוגנות, שקיפות, בטיחות ואפשרות לפירוש (Interpretability).

הסבריות

הסבריות היא היכולת לספק מידע ללקוח על הטכנולוגיה. כדי לשפר את יכולת ההסבריות של ה-AI, מפתחים כלי Explainable Artificial Intelligence (XAI) המאפשרים לפרש את תפוקות המודל על ידי זיהוי אילו משתנים השפיעו ביותר על תוצאה נתונה. בהקשר זה יש לשאול: מהו הסבר מספק וטוב כדי לתאר את ההליך שעברה המכונה? על ידי מי ההסבר צריך להיות מובן? על המפתח הטכני, על גורמים המשתמשים בו לקבלת החלטות או על האנשים שבעניינם נעשית קבלת ההחלטות? כך למשל, אם נעשה שימוש בבינה מלאכותית כדי להחליט על ערבות בבית המשפט, האם על ההסבר להיות מובן לשופט או לנאשם עצמו? באיזו רמת פירוט ובאיזה טיב של הסבר? כפי שמנסחים זאת Doshi-Velez ו-Kim³⁹ "הסבריות היא הצורך להסביר או להציג בצורה מובנת לבני אדם" אך ההגדרה של "מובן" תלויה בהקשר ובנמען.

אפשרות לפרשנות

האפשרות לפרשנות (Interpretability) עוסקת בהבנת מה שקורה בפועל בתוך המודל לרבות הייצוגים הפנימיים, המשקולות, המעגלים והחישובים. זהו יעד טכני ומדעי: האם ניתן לבצע הנדסה לאחור של מה שהמודל "למד" וכיצד הוא מעבד מידע? מחקר הפרשנות המכנית (כגון עבודת Anthropic על מעגלים ומאפיינים) הוא דוגמה טובה לכך. הפרשנות הפנימית היא בעיקרה עניין המיועד לחוקרים, ולעיתים קרובות קשה מאוד להשגה, למשל כאשר מדובר ברשת נוירונים עם מיליארדי פרמטרים.

שקיפות

שקיפות עוסקת בדיווח מהימן של האופן שבו נוצר ה-AI, על ידי מי ובאלו תנאים, מתי נעשה בו שימוש לקבלת החלטות וכולי.

הוגנות אלגוריתמית

אנו מצפים מאלגוריתמים המשמשים לקבלת החלטות לגבי בני אדם, דוגמת למי מגיעה הלוואה ולמי ערבות ובאלו תנאים, להיות הוגנים. אולם מהי אמת המידה הנכונה להוגנות אלגוריתמית?

ניתן להבחין בין שני סוגי אפליה: **יחס מפלה (Disparate Treatment)** שעשוי להיות ניתן למניעה; לאלגוריתמים אין מניעים, וניתן לתכנן אותם כך שלא יסתמכו במפורש על השתייכות לקבוצה מסוימת כגורם קלט. **השפעה מפלה (Disparate Impact)** לעומת זאת, יכולה להתרחש גם ללא כוונה: אלגוריתם עשוי

³⁹ Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel) October 17, 2016". [\(A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. "The Washington Post.](#)

להשתמש במשתנה שמספר את דיוקו, אך משתנה זה עלול לפגוע באופן לא פרופורציונלי בקבוצות מסוימות, למשל כאשר הוא מתואם עם מוצא אתני או מאפיין מוגן אחר.

הוגנות אלגוריתמית - מקרה בוחן: פיתוח מערכת בינה מלאכותית להחלטה על סיכויי פשיעה חוזרת

רקע

COMPAS היא מערכת טכנולוגית המשתמשת במגוון נתונים על פושעים מורשעים על מנת להעריך את סיכויי הפשיעה החוזרת שלהם. נעשה במערכת זו שימוש בבתי משפט במדינות בארצות הברית, דוגמת ניו יורק וקליפורניה, במטרה להחליט על דמי ערבות.

המערכת עלתה לכתורות כאשר בדו"ח מרכזי של העיתון פרו-פאבליקה טענו אנגוויין ועמיתיה (2016) כי מערכת COMPAS מפלה כלפי שחורים. זאת כיוון שהיא הניבה שיעור גבוה יותר של תוצאות חיוביות שגויות (false positives) עבור שחורים מאשר עבור לבנים. היו יותר מקרים של אנשים שחורים שלא חזרו לפשוע אך סווגו בטעות כבעלי סיכון גבוה לפשוע ושיעור גבוה יותר של תוצאות שליליות שגויות (false negatives) עבור לבנים. כלומר, עבריינים חוזרים שסווגו בטעות כבעלי סיכון נמוך.

אולם הקביעה כי מדובר בחוסר הוגנות אלגוריתמית נתקלה בהתנגדות. חברת נורת' פוינט (Northpointe), היוצרת של COMPAS, הגיבה בטענה שהאלגוריתם אינו מוטה (Dietrich et al., 2016), אולם לא סתרה את הנתונים הספציפיים שהציגה פרו-פאבליקה. היא טענה שהאלגוריתם הוא הוגן מכיוון שהוא מקדם סוג אחר של הוגנות -

דילמת ההוגנות האלגוריתמית

הוויכוח נסב סביב השאלה כיצד יש למדוד הוגנות: נורת' פוינט התמקדה בשאלה: האם שחור ולבן שקיבלו אותו ציון סיכון נוטים באותה מידה לחזור לפשע? ולפי מדד זה, האלגוריתם הוגן. מאידך, פרו-פאבליקה התמקדה בשאלה: האם שחור ולבן שלא חזרו לפשע היו בעלי סיכוי שווה לקבל ציון נמוך? ולפי מדד זה, האלגוריתם מפלה.

הפתרון הפשוט ביותר היה לספק הוגנות בשני המובנים גם יחד. אולם פתרון זה הוכח כבלתי אפשרי. בסדרת מאמרים (Kleinberg et al., 2016), הוכיחו מדעני מחשב כי כאשר שיעורי החזרה לפשע שונים בין קבוצות, וכאשר האלגוריתם אינו מושלם, פשוט לא ניתן לקיים את שני מדדי ההוגנות בו-זמנית. באופן כללי, שיעור חזרה לפשע גבוה יותר בקבוצה מסוימת יגרום לכך שיותר אנשים מאותה קבוצה יסווגו כבעלי סיכון גבוה - וכך גם שגיאות האלגוריתם יתרכזו יותר סביב סיווג-יתר של אותה קבוצה.

חשוב להדגיש: משפטי האי-אפשרויות הללו חלים על כל מנגנון חיזוי בין אם מדובר בלמידת מכונה, שאלוני הערכה, שיקול דעת אנושי, או כל מנגנון אחר. אין בכך עדות לכשל ייחודי של AI, אלא למגבלה מתמטית יסודית.

4.2.6 שחיקת יכולות אנושיות (Deskilling)

על רקע הסתמכות יתר על טכנולוגיה גדלה הסכנה לשחיקת יכולות ומיומנויות אנושיות, שכן אוטומציה או מערכות בינה מלאכותית, משתלטות על משימות קוגניטיביות, מוסריות ופרקטיות שבעבר בוצעו על ידי בני אדם. מושג זה יושם בעבר לתאר אובדן יכולות ומיומנויות בשוק העבודה, אך עם התפשטות הבינה המלאכותית, במיוחד הגנרטיבית, חל כיום גם על פגיעה ביכולות אנושיות בסיסיות כמו חשיבה מוסרית, חכמה מעשית, יצירתיות וקבלת החלטות עצמאית.

שחיקת היכולת המוסרית: אובדן כישורי קבלת החלטות מוסריות כל אימת שמערכות בינה מלאכותית מקבלות יותר ויותר החלטות אתיות עבורנו (מערכות לחימה אוטונומיות, כלי רכב אוטונומיים, רובוטים במגזר הסיעוד וכו'). כאשר בני אדם מפריטים חשיבה מוסרית לבינה מלאכותית, הם מאבדים את ההזדהויות לפתח ולתרגל את התבונה המעשית (פרונסיס) החיונית לשיפוט אתי.⁴⁰ שחיקת היכולת המוסרית עלולה להביא ל"ניתוק מוסרי" שבו אנשים מרגישים פחות אחראים להחלטות שמקבלות מערכות בינה מלאכותית, ולהחריף את בעיית האחריות המוסרית [4.2.1].

שחיקת יכולות חברתיות: החלפת אינטראקציות אנושיות וקשרי חברות משמעותיים עם צ'אטבוטים מבוססי מודלי שפה גדולים עלולים לייצר תלות בצ'אטבוטים אלו, היות והאחרונים מאומנים לנהוג בחנפנות, מייצרים סיפוק מיידי של אשרור רגשי, וזמינים בכל שעות היממה. תכונות אלו עלולות לייצר תלות בתחליפי יחסים, ולשחוק את היכולות החברתיות ליצור קשרים חברתיים משמעותיים המצריכים הקשבה, הדדיות וסבלנות.⁴¹

שחיקת מיומנות קוגניטיבית: כאשר מערכות בינה מלאכותית מבצעות משימות כמו ניווט, עיבוד מידע וניתוח, כתיב וניתוח של טקסטים וכו', בני אדם עלולים לפתח תלות במערכות אלה, ולאבד את היכולת להפעיל יכולות אלו באופן עצמאי, עד כדי "תלות אינטלקטואלית" ולא רק תלות טכנולוגית.

שחיקת יכולות אזרחיות: הפרטת החלטות מוסריות לבינה מלאכותית מאיימת על קבלת החלטות דמוקרטית על ידי הפחתת היכולת של האזרחים לחשיבה ביקורתית ושיפוט עצמאי, מה שעלול לערער את יסודות ההשתתפות הדמוקרטית.⁴²

הספרות מציעה כמה גישות להתמודדות עם שחיקת יכולות אנושיות (Deskilling) הנובעת משימוש בבינה מלאכותית: שמירה על מעורבות אנושית הן של מתכנני מערכות והן של המשתמשים בהן, עיצוב מערכות

⁴⁰ Shannon Vallor, "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character," *Philosophy & Technology* 28, no. 1 (2015): 107–24, <https://doi.org/10.1007/s13347-014-0156-9>.

⁴¹ Sung, Leora Urim, and Avigail Ferdman. "Empathetic Large Language Models, the Social Capacities and Human Flourishing." *Inquiry* 0, no. 0 (2025): 1–23. <https://doi.org/10.1080/0020174X.2025.2518448>.

⁴² Mark Coeckelbergh, "Democracy, Epistemic Agency, and AI: Political Epistemology in Times of Artificial Intelligence," *AI and Ethics*, November 22, 2022, <https://doi.org/10.1007/s43681-022-00239-4>. Ferdman, Avigail. "AI, Deskilling, and the Prospects for Public Reason." *Minds and Machines* 35, no. 3 (August 2025): 38. <https://doi.org/10.1007/s11023-025-09737-w>.

בינה מלאכותית שמעודדות פיתוח יכולות במקום החלפתן, טיפוח מינימליזם דיגיטלי, ופיתוח גישות חינוכיות שמטפחות תבונה מעשית לצד כישורים טכניים⁴³.

4.2.7 פער בגישות קיימות לאתיקה של טכנולוגיה

לגישות עכשוויות לאחריות מוסרית יש מגבלת מהותית בהקשרים של טכנולוגיה: הן אינן מסוגלות להתמודד באופן מספק עם פערי אחריות הנוצרים בשל העמימות, המורכבות ואי-הוודאות של טכנולוגיות משבשות כגון בינה מלאכותית, אלגוריתמים, נתוני עתק, וכו'. מגבלה זו נובעת מ"הטיה אינדיבידואליסטית" שמקורה בגישת legal solutionism - התמקדות בזיהוי סוכנים אחראים אינדיבידואליים על פני הכרה באופי הקולקטיבי המהותי של פיתוחים טכנולוגיים.⁴⁴ לפיכך, השאלה כיצד ניתן לייחס אחריות מוסרית לתוצאות טכנולוגיות כאשר המודלים המסורתיים של אחריות אינדיבידואלית מתגלים כבלתי-מספקים, נותרת פתוחה.

אחד ההסברים לבעיית האחריות המבוזרת הוא **בעיית הפעולה הקולקטיבית**. בעיות פעולה קולקטיבית (collective action problems) מתעוררות כאשר חברי קבוצה בוחרים לקדם את האינטרסים האישיים קצרי-הטווח שלהם על חשבון האינטרסים המשותפים ארוכי-הטווח, באופן שמוביל לתועלת תת-אופטימלית ולהחמרת מצבם של כלל המשתתפים בהשוואה למצב שבו היו נוהגים כולם באופן שיתופי. כאשר שחקנים פועלים על בסיס סדרי עדיפויות סותרים, ההסתברות לשיתוף פעולה פוחתת, שמביאה בתורה לתוצאות בלתי רצויות עבור הכלל.

גורמים המקדמים פעולה קולקטיבית כוללים מתווכים כגון אמון, הדדיות, שמירה על מוניטין, סנקציות (ענישה), ונכונות לשאת בעלות הכרוכה בענישת "טרמפיסטים". לעומתם, אינטראקציה בין גורמים כגון אנונימיות, היעדר אחריותיות, הטרוגניות, סיכון וחוסר ודאות יוצרת גורמי עיכוב (stressors) המערערים על אפשרות הפעולה הקולקטיבית. בעיית הפעולה הקולקטיבית מתגלה כחריפה במיוחד בהקשר של פיתוח וחדירה של בינה מלאכותית, לאור "מרוץ החימוש" של חברות עסקיות ושל מדינות כגון ארה"ב, סין והודו.⁴⁵

⁴³ Shannon Vallor, *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking* (New York: Oxford University Press, 2024), <https://doi.org/10.1093/oso/9780197759066.001.0001>; Andrew Zelny, "Offloading Wisdom: Four Technological Relations That Mediate Phronesis," *Philosophy & Technology* 38, no. 2 (2025): 56, <https://doi.org/10.1007/s13347-025-00889-2>; Avigail Ferdman, "Practices Worth Preserving: Knowing When to Offload to Technology," *Philosophy & Technology* 38, no. 3 (July 10, 2025): 100, <https://doi.org/10.1007/s13347-025-00925-1>.

⁴⁴ Santoni de Sio, Filippo, and Giulio Mecacci. "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them." *Philosophy & Technology* 34, no. 4 (December 2021): 1057–1084. <https://doi.org/10.1007/s13347-021-00450-x>.

⁴⁵ Olugbade, Olajide. "In Search of a Global Governance Mechanism for Artificial Intelligence (AI): A Collective Action Perspective." *Global Public Policy and Governance* 5, no. 2 (June 2025): 139–161. <https://doi.org/10.1007/s43508-025-00113-z>.

5. סוגיות יסוד בפילוסופיה הפוליטית של הטכנולוגיה

הפילוסופיה הפוליטית של הטכנולוגיה עוסקת, בין השאר, באופן שבו טכנולוגיה אינה ניטרלית, אלא נטועה בתוך יחסי כוחות כלכליים, מוסדיים וחברתיים. חדירת הטכנולוגיה והבינה המלאכותית לכל תחומי החיים יוצרת דילמות אתיות ופוליטיות חדשות, לא רק משום שהן משנות את סביבתנו, אלא גם משום שהן משנות אותנו. "המפנה הדיגיטלי" מטשטש גבולות חברתיים מסורתיים (למשל בין ציבורי לפרטי) ומחייב חשיבה מחודשת על מהותה של חברה אזרחית, מוסדות ודרכי קבלת החלטות בעידן הדיגיטלי.⁴⁶

כבר בשנות ה-80 טען לנגדון וינר "כי לטכנולוגיות יש פוליטיקה" והן מסוגלות לשמר או להעצים היררכיות קיימות. בעשור האחרון, חוקרים ביקורתיים הדגישו כי טכנולוגיות דיגיטליות עכשוויות מייצרות מחדש אי-שוויון דרך מנגנונים כמו אלגוריתמים מפלים⁴⁷, נגישות בלתי שוויונית למשאבים דיגיטליים ושליטה ריכוזית בתשתיות ענן ונתונים⁴⁸.

הפוליטיקה של הטכנולוגיה אינה רק שאלה של "מה נכון לעשות", אלא שאלה של לגיטימיות, סמכות, כוח ופתרון מחלוקות בתוך מרחב חברתי.⁴⁹ למשל, שאלת "מהו אלגוריתם הוגן" אינה שאלה טכנית בלבד, אלא שאלה פוליטית, כיוון שהיא כרוכה בבחירה בין ערכים מתנגשים ובהכרעות לגבי מי מוסמך לקבוע את האיזון ביניהם. יתרה מכך, האופן שבו טכנולוגיות משנות את האופנים שבהם חברה מגדירה "עובדה", "עוול" או "אמת" – למשל דרך פלטפורמות מידע או אלגוריתמים מסחריים, חושף ממד פוליטי עמוק, אשר משפיע על תנאי השיח הדמוקרטי עצמו.

בנוסף, לטכנולוגיה השפעה מכרעת על התרגום של עקרונות יסוד של הפילוסופיה הפוליטית למוסדות ופרוצדורות קבלת החלטות במדינה האידאלית. ניתן לטעון כי האיזון של הדמוקרטיה הליברלית בין הכרעת רוב באמצעות נציגים למגבלות על הרוב בצורת חוקה, אינו מוצדק עוד בעידן הטכנולוגי הנוכחי וכי יש לפתוח פילוסופיה פוליטית מותאמת לעידן הטכנולוגי החדש.

5.1 ערעור מאזן הכוחות הדמוקרטי

המציאות הנוכחית מצביעה על ריכוז כוח טכנולוגי חסר תקדים בידיהן של חברות Big Tech, השולטות גם בנתונים אישיים, גם בתשתיות קריטיות של ממשלות, וגם בשיח הציבורי ברשתות. ריכוזיות זו מטשטשת את הגבולות בין פרטי לציבורי, בין ממשל לתאגיד, ומציבה אתגר ממשי לריבונות הדיגיטלית של מדינות וליכולת

⁴⁶ Rakowski, R., Kowaliková, P. [The political and social contradictions of the human and online environment in the context of artificial intelligence applications](#). Humanit Soc Sci Commun 11, 289 (2024).

⁴⁷ Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

⁴⁸ Cohen, Julie E., [Between Truth and Power: The Legal Constructions of Informational Capitalism](#) (New York, 2019; online edn, Oxford Academic, 24 Oct. 2019)

⁴⁹ Lazar, Seth. "Power and AI: Nature and Justification." In *The Oxford Handbook of AI Governance*, edited by Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, and Baobao Zhang (online edn: Oxford University Press, 2022). <https://doi.org/10.1093/oxfordhb/9780197579329.013.12>.

של מוסדות דמוקרטיים לפקח ולווסת טכנולוגיה. פערים אלה אינם רק טכנולוגיים אלא פוליטיים ובאים לידי ביטוי בכך שכוח ההשפעה על עיצוב והכוונת טכנולוגיה מצוי כמעט בבלעדיות בידי תאגידים גדולים או מדינות חזקות, בעוד שרוב הציבור ובפרט קבוצות מוחלשות, מודר מהשתתפות בהחלטות על פיתוח והפעלה של מערכות טכנולוגיות. לדוגמה, פרויקט מערכת זיהוי הפנים של אמזון⁵⁰, שנמכר לגופי משטרה בארה"ב, עורר ביקורת ציבורית חריפה לאחר שפורסם כי המערכת מזהה אנשים כהי עור בשיעורי טעות גבוהים בהרבה מאשר לבנים⁵¹. על אף אזהרות של חוקרים, פעילים ואף חלק מחברי הקונגרס, ההחלטה האם להמשיך להפעיל את המערכת נותרה בידי החברה עצמה – דוגמה מובהקת לכשל דמוקרטי בתהליך קבלת ההחלטות הטכנולוגיות. במקביל, יוזמות חקיקה שנועדו לצמצם את הפערים או להבטיח שקיפות באלגוריתמים נתקלו בלחצים פוליטיים מצד תאגידים עתירי הון, שמפעילים השפעה משמעותית על עיצוב הרגולציה. מציאות זו מדגישה כיצד טכנולוגיה הופכת לזירה של עיצוב חברתי-פוליטי, שבה ההכרעות אינן טכניות בלבד אלא טעונות ערכים, אינטרסים והדרה מוסדית.

5.2 עירוב אלגוריתמים בקבלת החלטות

טכנולוגיות דיגיטליות, ובפרט טכנולוגיות הבינה המלאכותית, הוכיחו יכולות לבצע מטלות הדורשות מאמץ קוגניטיבי באיכות גבוהה כמו בני אדם מומחים, ולעיתים אף להתעלות על יכולות אנושיות. כמו כן, טכנולוגיות אלו מאפשרות לבצע מטלות אלו באופן אוטומטי ומהיר מבלי להידרש לאותה כמות של משאבי אנוש. על כן, יש הסוברים כי יש להשתמש בטכנולוגיות האלו בהליכי קבלת החלטות, הן על ידי הפרט והן קבלת החלטות קולקטיביות ו/או כחלק מקואורדינציה בין פרטים שונים בהליך גיבוש החלטות קבוצתיות. יש אף הסוברים כי יש להשתמש בטכנולוגיה לניסוח חקיקה, הדמיית תגובת האוכלוסיה וכדומה.

הפילוסוף ג'וליאן סבולסקו סבור כי יש לקדם **יועצי AI**, בהם יוכלו אנשים להיעזר בהחלטות ביום יום, למשל, איזה מוצר מהמוצרים בסופר הוא "אתי" מספיק⁵². הטכנולוגיה תוכל לאסוף נתונים לגבי אופן הכנת המוצרים ולהצליבם עם סוגיות אתיות רלוונטיות מהר יותר מבני אדם. תפיסה זו גורסת כי אם בני האדם משתמשים בטכנולוגיה על מנת להתקדם בתחומים אחרים, חשוב כי האתיקה לא תיוותר מאחור.

הוגים אחרים, דוגמת הלנה לנדמור, סבורים כי יש להשתמש ב-AI למטרות קואורדינציה בין פרטים לצורך קבלת החלטות. כך, היא מציעה להשתמש ב-AI כדי לסייע בארגון וניהול דיון רב משתתפים, לקבלת החלטות באופן דמוקרטי דליברטיבי⁵³. יכולות ה-AI בתחומי סיכום טקסט, ניהול שיחות וכדומה יכולות לסייע בכך.

⁵⁰ [Amazon is pushing facial technology that a study says could be biased](#) Massachusetts Institute of Technology

⁵¹ Buolamwini, J., Geburu, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

⁵² Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy & technology*, 31(2), 169-188.

⁵³ Landemore, H. (2024). Can Artificial Intelligence Bring. *Conversations in philosophy, law, and politics*, 39.

אחרים מציעים להשתמש ב-AI לבניית קונצנזוס,⁵⁴ כאשר יכולות ה-AI מאפשרות לטענתם לנסח חוקים, לדמות כיצד יצביעו אחרים על החוקים ולברור את החוקים הטובים ביותר, אז להעמידם להצבעה. בנוסף, ניתן להשתמש ב-AI לסכם נושאים עיקריים שעלו משיח, ובהתאם לכך לנסח חוקים ורעיונות שבסבירות גבוהה יתאימו למרבית המשתתפים. כך, בטיוואן נעשה כבר בעבר שימוש ב-AI כחלק ממערכת בשם [Pol.IS](https://www.pol.is/) כדי לסייע לחוקק חוקים.⁵⁵

5.3 מניפולציות טכנולוגיות על ידע פוליטי

נכון להיום מרבית האוכלוסיה צורכת את המידע הפוליטי שלה דרך הרשת, הן באמצעות אתרי חדשות והן באמצעות מידע העובר ברשתות חברתיות. המידע המצוי במרחב הרשת עובר לרוב עיבודים טכנולוגיים. אלו כוללים **מערכות המלצה**, אלגוריתמים המקבלים כקלט מידע על היסטוריית הצריכה של הפרט ומוציאים כקלט המלצות למידע נוסף לאותו הפרט. מערכות אלו מעצבות חלקים נרחבים מהמידע אליו נחשפים אנשים במרחב הרשת. המידע אותו מקבלות מערכות ההמלצה כקלט לרוב מתחלק לשניים - יש מערכות המלצה המשתמשות במידע על התוכן אותו צרך הפרט (content based filtering) ויש כאלו הנשענות על מידע בנוגע לאנשים אחרים שתויגו כדומים לאותו הפרט (collaborative filtering).

עבודתן של מערכות ההמלצה מעלה סוגיות אתיות ופוליטיות. ראשית, ברשתות החברתיות למשל, מכיוון שמערכות ההמלצה אותן מעצבים תאגידי הרשתות החברתיות משפיעות על התוכן אליו נחשפים אנשים, יש הסבורים כי זה מטיל על החברות אחריות מוסרית ומשפטית לתוכן פוגעני.⁵⁶ שנית, המידע בו נעשה שימוש לצורכי המלצה יכול לגרום להטיות, ובפרט השימוש בנתונים על אנשים שתויגו כ"דומים" לפרט מסוים.

כמו כן, רבים מייחסים למערכות ההמלצה עיצוב מרחב רשת היוצר קיטוב בשל הצגה של מציאויות שונות לאנשים המשתתפים למחנות פוליטיים שונים.⁵⁷ נטען כי המלצות אלו, הממליצות על מידע שונה לפרטים שונים, יוצרות בועות פילטר ותיבות תהודה (echo chambers), כלומר, בועות בהן אנשים נחשפים רק לדעות דומות לשלהם, ודעותיהם מהודדות להם בחזרה מהמידע אליו הם נחשפים.

בשירותים למטרות רווח, דוגמת הרשתות החברתיות, המודל הכלכלי מבוסס על תשלומים עבור פרסום באתרים אלו. עלות הפרסום באתרים תלויה במידת החשיפה של הפרסומות, אשר תלויה, בתורה, בכמות האנשים המבקרים באתר ובאורך הביקור באתר הרשת החברתית. על כן, מערכות ההמלצה תוכנתו כך שיתעדפו תוכן המאריך את שהות המשתמשים באתר. מחקרים מראים כי תוכן כזה נוטה לעיתים רבות להיות

⁵⁴ Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., ... & Summerfield, C. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in neural information processing systems*, 35, 38176-38189.

⁵⁵ Hsiao, Y. T., Lin, S. Y., Tang, A., Narayanan, D., & Sarahe, C. (2018). vTaiwan: An empirical study of open consultation process in Taiwan. *SocArXiv*, 4.

⁵⁶ Morgan Lee, *Jury Finds Meta's Platforms Are Harmful to Children in 1st Wave of Social Media Addiction Lawsuits*, PBS NewsHour (Mar. 24, 2026), <https://www.pbs.org/newshour/nation/jury-finds-metas-platforms-are-harmful-to-children-in-1st-wave-of-social-media-addiction-lawsuits>.

⁵⁷ Aulio, G. J. (2018). # republic: Divided Democracy in the Age of Social Media.

תוכן פרובוקטיבי, לרבות מידע שגוי ותאוריות קונספירציה. על כן, מידע שגוי נפוץ ברשתות. נהוג לחלק מידע כזה **למיסאינפורמציה**, מידע שגוי שהשולח אותו לא היה מודע לכך שהוא שגוי, ו**לדיסאינפורמציה**, מידע שגוי שנועד להטעות.

הטכנולוגיות העדכניות, ובראשן המודלים הגנרטיביים דוגמת ChatGPT, קלוד וג'מיני, מאפשרים יצירה של **זיוף עומק (Deep Fake)** בקלות יחסית. זיופי עומק כוללים למשל תמונות מזויפות של פוליטיקאים, ואף סרטונים או קטעי שמע מזויפים. הם יכולים אף לכלול סרטונים ותמונות שקריים של תוצאות של מלחמה אשר יכולים להטות את דעת הקהל ביחס לאותה המלחמה. אלו מאפשרים הטעיה של הציבור לגבי הנעשה במציאות.

המודלים הגנרטיביים אף מאפשרים הטעיה רחבה יותר של הציבור על ידי ייצור כתבות מזויפות אשר נראות אמיתיות במהירות רבה והפצתן במרחב הרשת. הרשתות מלאות כיום גם בבוטים, פרופילים פיקטיביים אשר משמשים לסימון "לייק" ותגובה לתוכן על מנת להעלות את הנראות שלו ולשווק תחושה מטעה לגבי דעת הקהל.

5.4 גיימיפיקציה/רידוד/קריסת ערך ⁵⁸ (value collapse)

קריסת ערך היא תהליך של דילול ערכים רלוונטיים לחיים מלאים ומשגשגים, כתוצאה מהטמעת לוגיקה של מדידה כמותית מונוליטית, הפוגעת באופן שיטתי בערכים העשירים הלוכדים את כל עושר החוויה האנושית. טכנולוגיה (למשל רשתות חברתיות) מציגה לנו גרסאות פשוטות של הערכים הללו - לרוב בצורה של מדדים מספריים, אלגוריתמים או מדדים סטנדרטיים; הגרסאות הפשוטות הללו מחליפות בהדרגה את הערכים המקוריים והמורכבים יותר שלנו בתהליכי החשיבה וקבלת ההחלטות (למשל, המורכבות והניואנסים שבתקשורת אנושית). כתוצאה, אנו מאבדים קשר עם המשמעויות העמוקות והמורכבויות שעומדים בבסיס פעילות אנושית (כגון תקשורת אנושית). לדוגמה, ערכים בבסיסו של חינוך מצטמצמים לציונים ודירוגים, קשר חברתי אמיתי משוטח למדדי מדיה חברתית כגון לייקים ומספר עוקבים. הסכנה בקריסת ערך שהיא שוחקת באופן שיטתי את היכולת שלנו להעריך ולחזור לשגשוג אמיתי, לטובת מה שניתן למדוד ולייעל בקלות.

5.5 הטיית

אלגוריתמים עלולים להנציח אפליה קיימת. יש להטמיע בדיקות איזון, שקיפות ואיסוף נתונים מגוון כדי להבטיח הוגנות. כך למשל, כמוזכר לעיל, משטרת ארה"ב השתמשה במערכת זיהוי פנים שלאחר מכן

⁵⁸ C. Thi Nguyen, *Games: Agency as Art* (Oxford University Press, 2020); C. Thi Nguyen, "Value Capture," *Journal of Ethics and Social Philosophy* 27, no. 3 (2024); C. Thi Nguyen, "Value Capture and Algorithmic Policies: Outsourcing Our Values to Technology" (CTMF Flagship Lecture with Professor C. Thi Nguyen, Edinburgh Futures Institute, Edinburgh, 2024), <https://www.technomorfutures.uk/events-database/flagship-lecture-professor-nguyen>; Dylan J. White, "Automated Influence and Value Collapse: Resisting the Control Argument," *American Philosophical Quarterly* 61, no. 4 (2024): 369–86, <https://doi.org/10.5406/21521123.61.4.06>.

התגלה כי היא מזהה אנשים כהי עור בשיעורי טעות גבוהים בהרבה מאשר לבנים⁵⁹. בספרות מתייחסים לאמצעים לצמצום הטיה פוליטית בבינה מלאכותית⁶⁰ בעזרת שילוב נתוני אימון מגוונים המשקפים קבוצות ועמדות שונות; פיתוח מנגנונים שיטתיים לזיהוי ותיקון הטיית תוך הקפדה על עקרונות של הוגנות, אחריות ושקיפות; עריכת ביקורות תקופתיות למודלים ולתהליכי קבלת ההחלטות האלגוריתמיים; קידום יוזמות שקיפות והסדרה מוסדית באמצעות הנחיות רגולציה ומסגרות ממשל אתי; וחינוך הציבור להגברת המודעות לבעיות אפשריות בשימוש בטכנולוגיה, במטרה לאפשר שימוש ביקורתי ומושכל ולהפחית את הסיכון להנצחת אפליה קיימת.

הטיות אלגוריתמיות, דוגמה: מערכת דירוג קורות חיים

רקע:

דוגמה להטיה אלגוריתמית מהשנים האחרונות היא טכנולוגיית הבינה המלאכותית שפתחה אמזון על מנת להכין מערכת המדרגת קורות חיים לשימוש בעת חיפוש עובדים חדשים לחברה.

ההטיה האלגוריתמית:

לאחר בדיקה, הסתבר כי האלגוריתם למד להוריד את הדירוג של קורות חיים פשוט בשל הופעת המילה "נשים" בפנים. לטענת החברה, האלגוריתם אומן על קורות חיים מהעשור שלפני כן, ומרבית קורות החיים הללו היו של גברים.

תגובה:

לאחר תגלית זו, ב-2018 החברה זנחה את הטכנולוגיה.

5.6 טכנולוגיה ולוחמה

שילוב הבינה המלאכותית בלוחמה המודרנית הוליד אשכול של שאלות אתיות ופוליטיות דחופות שהמסגרות המשפטיות והפילוסופיות הקיימות מתקשות להשיב עליהן. במרכז הדיון ניצבת בעיית האחריות: כאשר מכונה בוחרת מטרה ותוקפת אותה באופן עצמאי, הדין הבינלאומי הנוכחי מתקשה להטיל אחריות על אדם ספציפי כלשהו, שכן המשפט ההומניטרי הבינלאומי והמשפט הפלילי הבינלאומי נבנו סביב החלטות אנושיות. קיים קונצנזוס רחב לפיו אלגוריתמים אינם מסוגלים לנהל מחשבה מוסרית, ועל כן האצלת קבלת ההחלטות הקטלניות למכונות אינה קבילה מבחינה אתית. מעבר לאחריות האישית, הבינה המלאכותית בלוחמה מאיימת על עקרונות הומניטריים יסודיים: מערכות נשק אוטונומיות, המסוגלות לקבל החלטות על

⁵⁹ Buolamwini, J., Gebru, T. "[Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification](#)." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

⁶⁰ T. Choudhary, "[Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude](#)," in IEEE Access, vol. 13, pp. 11341–11379, 2025

חיים ומוות ללא התערבות אנושית, מאיימות על הזכות לחיים, ואילו הטיות מובנות במערכות בינה מלאכותית עלולות, להחמיר פערים קיימים ולפגוע בעקרונות השוויון ואי-האפליה.

הממד הפוליטי אינו פחות סבוך: כישלון המדינות להסכים על עקרונות יסוד מלמד כי המסגרת של "בינה מלאכותית אחראית" נותרת שנויה במחלוקת ברמה הגלובלית, כאשר מדינות דמוקרטיות נוקטות גישה מונחית-ערכים יותר מאשר משטרים אוטוריטריים, בעוד שחוסר היכולת של מסגרות הממשל של הבינה המלאכותית לעמוד בקצב הטכנולוגיה כבר מוביל לנזק בעולם האמיתי. בזירה הגיאופוליטית, מומחים מזוהרים מפני מרוץ חימוש במערכות נשק אוטונומיות, בפרט על רקע מתחים גיאופוליטיים גוברים, תוך סיכונים משמעותיים של התפשטות, הסלמה בלתי רצויה ושינויים בלתי צפויים בדינמיקת הכוח הגלובלית. השפעת הבינה המלאכותית על הלוחמה חורגת מעולם הנשק אל תחום המידע: תוכן שמייצרת בינה מלאכותית עלול לערער את האמון במוסדות, לתמרן את דעת הקהל ולעוות את השיח הפוליטי במידה שלא הייתה אפשרית בעבר.

עצרת האו"ם הצביעה בשנת 2024 לפתוח במשא ומתן רשמי על אמנה בנושא בינה מלאכותית בלוחמה, כאשר המזכ"ל קרא לגיבוש כללים מוחשיים עד שנת 2026, אך ללא מנגנוני אכיפה, מאמצים אלה מציעים מעט מאוד הגבלה בפועל. החשש המרכזי הוא כי ככל שהבינה המלאכותית תופסת תפקיד מרכזי יותר ויותר, מפקדים צבאיים עלולים לאבד שליטה על עימותים המסלימים במהירות רבה מדי וחסרים פיקוח אתי ומשפטי, אתגר שיש המכנים אותו "רגע אופנהיימר" של תקופתנו. כלי נשק אוטונומיים מעלים שאלות אחריות, הבחנה ומידתיות. רוב הגופים קוראים לשימור שליטה אנושית משמעותית בהחלטות קטלניות.⁶¹

דילמת השימוש הכפול מתייחסת למצב שבו טכנולוגיה, ידע, או מוצר שפותחו למטרה חיובית, אזרחית או מדעית, יכולים לשמש גם למטרות שליליות, מזיקות או צבאיות.⁶²

5.7 פשעים של העידן הדיגיטלי (Cybercrime)

העידן הדיגיטלי והטכנולוגיות החדשות הביאו עמן פיתוח של סוגים חדשים של התנהלויות לא מוסריות אשר לא היו קיימות קודם לכן ולדיון בשאלה האם יש להגדירם גם כעבירות פליליות. להלן רשימת פשעי סייבר נפוצים.

1. פשינג (Phishing)

שליחת הודעות מזויפות (בדרך כלל אימייל) כדי להערים על קורבנות למסור פרטים אישיים, סיסמאות או פרטי כרטיס אשראי.

⁶¹ <https://lieber.westpoint.edu/legal-accountability-ai-driven-autonomous-weapons/>

⁶² זלצברגר, ע', פרדמן, א', וקליין, ר' (2025). סיכום פאנל בנושא דילמת השימוש הכפול: אתיקה ואחריות במחקר ופיתוח בישראל. מוסד שמואל נאמן.

2. תוכנות כופר (Ransomware)

תוכנה זדונית שמצפינה את קבצי הקורבן ומחייבת תשלום (לרוב בקריפטו) בתמורה לשחרורם.

3. גניבת זהות (Identity Theft)

גניבת מידע אישי (שם, ת.ז., פרטי בנק) כדי להתחזות לקורבן ולבצע פעולות פיננסיות בשמו.

4. פריצה למחשבים (Hacking)

גישה לא מורשית למערכות מחשב, שרתים או רשתות – לצורך גניבת מידע, השחתה, או השתלטות.

5. תוכנות ריגול (Spyware / Stalkerware)

תוכנות שמותקנות בסתר ומנטרות פעילות המשתמש – לחיצות מקלדת, צילומי מסך, מיקום וכד'.

6. הונאות אונליין (Online Fraud)

הונאות קנייה/מכירה, הגרלות מזויפות, אתרי השקעות פיקטיביים ועוד.

7. פשעי מחשב כלפי ילדים (CSAM / Grooming)

ניצול ילדים באמצעים דיגיטליים, כולל הפצת תוכן פדופילי ויצירת קשר מניפולטיבי עם קטינים.

8. כריית מטבעות קריפטו לא מורשית (Cryptojacking)

שימוש בכוח המחשוב של מכשיר הקורבן ללא ידיעתו לכריית מטבעות דיגיטליים.

9. Cyberflashing (שליחת תמונות מיניות לא רצויות)

שליחת תמונות/סרטונים מיניים לאנשים ללא הסכמתם

10. Epilepsy Trolling (תקיפת אפילפטיקאים)

שליחת תמונות מהבהבות במכוון לחולי אפילפסיה כדי לגרום להם להתקפים (נחשב **תקיפה פיזית**)

11. שימוש בזיוף עומק (Deep Fake) באופנים הבאים:

- ייצור תכנים פורנוגרפיים מזויפים של אנשים אמיתיים
- התחזות לאנשים אמיתיים לצורכי הונאה (למשל למטרת השגת כספים)
- הפצת מידע כוזב סביב בחירות

12. הטרדה מקוונת (Cyber Stalking)

מעקב, הטרדה ואיומים שיטתיים דרך אמצעים דיגיטליים.

13. בריונות רשת (Cyberbullying)

הטרדה מקוונת שיטתית על ידי שליחת הודעות מטרידות וביוש המוטרד

חלק מהפשעים הללו דורשים עדכונים בחקיקה ואכיפה שיאפשרו לבתי המשפט דרכים ישירות וברורות להעניש עליהם. כך למשל, הטרדה של חולי אפילפסיה באמצעות שידור מכון של אורות מהבהבים במרחב הרשת דורשת התייחסות מיוחדת על מנת שיהיה ניתן להעמיד לדין בגינה על עבירת תקיפה. במספר מדינות, דוגמת בריטניה, אוסטרליה וארה"ב, נחקקו כבר חוקים חדשים שנועדו להתייחס באופן ישיר לחלק מהעבירות הללו. ככל שהטכנולוגיות מתפתחות כך מתאפשרים סוגים חדשים של פשעים שלא התקיימו לפני כן. על כן, על החקיקה להתעדכן בהתאם להתעדכנות הטכנולוגיות. עניין זה מהווה אתגר עבור המשפט שכן מטבעה חקיקה היא בעלת קצב איטי יותר.

5.8 טכנולוגיות מעקב

פיתוחים של טכנולוגיה דיגיטלית שימשו מאז תחילתם לצורכי מעקב. כך, התקנתן של מצלמות במקומות שונים ברחובות אפשרה מעקב תדיר יותר אחרי התנהגות בני האדם במרחב הציבורי. זו אפשרה תפיסה קלה יותר של פושעים, מחד, אך פגעה בפרטיות, מאידך. יכולות המעקב עברו שדרוג משמעותי עם הגעת הבינה המלאכותית.

אחת היכולות המובהקות של הבינה המלאכותית היא היכולת לזהות ולתייג תמונות. כך, ניתן להשתמש בבינה מלאכותית על מנת לזהות פנים של בני אדם ולהצליב את זיהוי הפנים הזה בין מקומות שונים. בנוסף, הבינה המלאכותית טובה בזיהוי דפוסים החוזרים על עצמם, ויכולה לזהות דפוסי התנהגות של אנשים שונים במרחב הרשת ולדווח עליהם. גם דפוסי התנהגות דוגמת **דפוסי הוצאות והכנסות**, סדירות תשלומים ורמת חובות קיימת ניתנות לזיהוי באמצעות בינה מלאכותית, וניתן להשתמש בנתונים כאלה כדי לנבא את רמת הסיכון הפיננסי של לווים פוטנציאליים.

בסין, נעשה שימוש בטכנולוגיות מעקב באופן נרחב ומשוכלל. מערכת האשראי החברתי הסינית מסתמכת במידה רבה על טכנולוגיות מתקדמות, לרבות זיהוי פנים, בינה מלאכותית וניתוח ביג דאטה. למעלה מ-200 מיליון מצלמות מעקב המצוידות בזיהוי פנים מבוסס בינה מלאכותית מכסות ערים ומזהות אנשים בזמן אמת תוך מעקב אחר תנועותיהם ברחבי המדינה. הפעילויות הנמדדות כוללות תשלום מסים בזמן, ציות לחוקי התנועה, השתתפות בשירות לציבור, ואפילו התנהגות מקוונת כגון פרסום ושיתוף תכנים. אזרחים שציונם נמוך עלולים למצוא עצמם מנועים מרכישת כרטיסי טיסה או רכבת מהירה, וילדיהם עלולים שלא להתקבל לבתי ספר פרטיים או אוניברסיטאות. בערים כמו האנגג'ואו פועלת מערכת "מוח עירוני" המנתחת תנועה, התנהגות חברתית ואירועי חירום בזמן אמת.⁶³

גם בתקופת הקורונה, נעשה שימוש נרחב בטכנולוגיות מעקב. דרום קוריאה, ישראל וסין השתמשו בטכניקות מעקב שכללו צילומי מצלמות, זיהוי פנים, צמידי מעקב ואותות סלולר כדי לאתר את מיקומיהם של נדבקים ולמפות את מי שנחשפו אליהם. ישראל הורתה לשב"כ לנצל מחדש את מערכות מעקב הטרור שלו לצורכי איתור נדבקים בקורונה. דרום קוריאה אימצה תוכנית מעקב קפדנית שכללה הצלבה בין עסקאות כרטיסי אשראי ונתוני מיקום לצורך איתור מגעים. עם זאת, מדינות רבות באירופה בחרו בגישה שונה: האיחוד

⁶³ [Inside China's Social Credit System and Its Global Influence – The Surveillance State](#)

האירופי יישם אפליקציות מעקב המבוססות על חיבורי Bluetooth שאינן מאפשרות זיהוי מיקום מדויק ומותאמות לנסיעות בין מדינות.⁶⁴

היכולת להשתמש בטכנולוגיות כאלה מעלה שאלות מורכבות על הגבול בין ביטחון לפרטיות. מערכות מעקב מבוססות בינה מלאכותית מוסברות לעיתים קרובות על בסיס שיפור הביטחון הציבורי, אולם איסוף כמויות עצומות של נתונים אישיים מעלה חשש לפגיעה משמעותית בפרטיות, בפרט כשלאנשים יש שליטה מועטה על אופן השימוש בנתונים. בעידן הבינה המלאכותית, הסיכון הוא כי מערכות עלולות לנתח מסדי נתונים עצומים של תקשורת אישית בצורה אוטומטית ורחבת היקף, מה שיאפשר מעקב המוני חסר תקדים. לפיכך, על מדינות דמוקרטיות לקבוע מסגרות אתיות, להטיל חובת שקיפות, להגביל את השימוש בנתוני מעקב המוני ולעגן הגנות פרטיות ברורות.

6. אתיקה של בינה מלאכותית: עקרונות ופרוגרמות ליישום

6.1 עקרונות אתיים מנחים לשימוש בבינה מלאכותית

פיתוח ושימוש במערכות בינה מלאכותית מציבים בפני החברה הזדמנויות חסרות תקדים לצד סיכונים ממשיים. שיקולים אתיים יכולים להטיב את ההתפתחויות המהירות הללו, ונציין כאן שלושה מודלים שהוצעו בספרות.

6.1.1 מודל AI4People

מודל AI4People⁶⁵ מציע חמישה עקרונות אתיים מרכזיים כבסיס לעיצוב ופיקוח על מערכות AI:

- טובת האנושות (Beneficence)
- מניעת נזק (Non-maleficence)
- אוטונומיה
- צדק
- הסבריות⁶⁶ (Explicability)

⁶⁴ [COVID-19 testing and infection surveillance: Is a combined digital contact-tracing and mass-testing solution feasible in the United States? - ScienceDirect](https://doi.org/10.1007/s11023-018-9482-5)

⁶⁵ Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28, no. 4 (December 1, 2018): 689–707. <https://doi.org/10.1007/s11023-018-9482-5>

⁶⁶ הסבריות - [ההסבר להחלטות שהתקבלו על ידי מערכות בעלות בינה מלאכותית \(AI\) נתפש כחיוני לאמון וקבלה חברתית של אותן מערכות](#)

עקרונות אלה מבקשים למקסם רווחה אנושית, לשמור על כבוד האדם והסביבה, למנוע פגיעה, להבטיח חלוקה צודקת של משאבים ותועלות, לשמר את היכולת האנושית לבחור ולהחליט, ולאפשר שקיפות והבנה של פעולת המערכות ומי נושא באחריות להן.

Florida ועמיתים (2018) מפרטים את 20 ההמלצות כחלק ממודל ה-AI4People בחלוקה לארבעה אשכולות פעולה: הערכה, פיתוח, תמרוץ ותמיכה.

הערכה (Assessment)

1. **הערכת יכולת המוסדות הקיימים** (למשל בתי משפט) **לספק מנגנוני פיצוי ותיקון נזקים שגרמו מערכות AI**, כולל ביסוס עקרונות אחריות משפטית כבר בשלב התכנון.
2. **הערכת משימות והחלטות שלא ראוי להאציל ל-AI** באמצעות מנגנוני שיתוף ציבור והתאמה לערכים חברתיים.
3. **הערכת מסגרות רגולציה קיימות** - האם הן מבוססות די הצורך על עקרונות אתיים ומסוגלות להתמודד עם אתגרים בלתי צפויים.

פיתוח (Development)

4. **פיתוח מסגרת לשקיפות והסבריות של מערכות AI** בקבלת החלטות חברתיות משמעותיות.
5. **פיתוח נהלים משפטיים ותשתית טכנולוגית** המאפשרים לבתי המשפט לבחון החלטות אלגוריתמיות (כולל פתרונות כמו ⁶⁷zero-knowledge proofs).
6. **פיתוח מנגנוני ביקורת (auditing) לאיתור תוצאות לא רצויות כגון הטיית**, ומנגנוני סולידריות להתמודדות עם סיכונים חמורים.
7. **פיתוח מנגנון פיצוי ותיקון נזקים** (כולל רעיונות כמו "נציב קבילות AI" או ביטוח חובה).
8. **פיתוח מדדים לאמון הציבור במוצרי ושירותי AI** - בסיס ל"אינדקס השוואה" של אמינות אתית.
9. **הקמת רשות פיקוח אירופית ל-AI**, שתבחן ותאשר מוצרים ושירותים, עם מנגנון מעקב "לאחר השקה".
10. **הקמת מנגנון אירופי לניטור מגמות**, הפצת ידע והנחיות פעולה בתחום AI.
11. **פיתוח כלים משפטיים והסכמיים** לשיתוף פעולה יעיל בין אדם למכונה בשוק העבודה, כולל קרן התאמה לתעסוקה ב-AI.

⁶⁷Zero-Knowledge Proof היא שיטה קריפטוגרפית שמאפשרת לצד אחד להוכיח לצד שני שהוא יודע או מחזיק במידע מסוים מבלי לחשוף את המידע עצמו ומבלי למסור כל פרט נוסף מעבר לעובדת הידיעה. בהקשר של אתיקה בבינה ZKP יכול לשמש כדי לאפשר ביקורת על אלגוריתמים או החלטות של מערכות AI בלי לחשוף קוד קנייני, מודלים סודיים או נתונים אישיים, ובכך לשמור על איזון בין שקיפות לבין הגנה על פרטיות וזכויות יוצרים. <https://telefonicatech.com/en/blog/ai-zero-knowledge-proof-zkp>

תמרוץ (Incentivization)

12. **תמרוץ פרויקטים חברתיים וסביבתיים בתחום AI.**
13. **תמרוץ מחקר אירופי מתואם בתחום AI לטובת החברה** כמשקל נגד למגמות מסחריות בלבד.
14. **תמרוץ שיתופי פעולה רב-תחומיים** בין טכנולוגיה, חברה, משפט ואתיקה.
15. **תמרוץ שילוב שיקולים אתיים, משפטיים וחברתיים** בכל פרויקט מחקר ופיתוח בתחום AI לצד בחינה מתמדת של חקיקה.
16. **תמרוץ הקמת אזורי ניסוי מיוחדים ("מעבדות חיות")** לבדיקת AI בתנאים מבוקרים, עם הגנה על הציבור.
17. **תמרוץ מחקר על תפיסת הציבור את ה-AI** ויישום מנגנוני שיתוף ציבור ביצירת מדיניות וסטנדרטים.

תמיכה (Support)

18. **תמיכה בפיתוח קודים אתיים מקצועיים** לתחומי נתונים ו-AI, כולל תוויות אמון (trust labels).
19. **חיזוק אחריות הדירקטוריונים** על ההשלכות האתיות של טכנולוגיות AI (למשל באמצעות הכשרת דירקטורים או הקמת ועדות אתיקה פנימיות).
20. **תמיכה ביצירת תכניות חינוך ומודעות לציבור** על ההשפעות החברתיות, המשפטיות והאתיות של AI בבתי ספר, באקדמיה, בארגונים ציבוריים ובקרב הציבור הרחב.

6.1.2 מודל European Commission High-Level Expert Group on AI (HLEG)

קבוצת המומחים לבינה מלאכותית (High-Level Expert Group on AI – HLEG) מונתה על ידי הנציבות האירופית ביוני 2018, וסיימה את עבודתה ביולי 2020. מטרת הקבוצה הייתה **לעצב מסגרת אירופית לבינה מלאכותית אמינה (Trustworthy AI), ולספק המלצות מדיניות** והשקעה שיתמכו בפיתוח, פריסה ושימוש בטכנולוגיות אלו באופן חוקי, אתי ובטוח.

במסמך ההנחיות האתיות שפרסמה הקבוצה בשנת 2019, **Ethics Guidelines for Trustworthy AI**⁶⁸, נקבע כי **בינה מלאכותית אמינה נשענת על שלושה רכיבים מרכזיים** שיש להתייחס אליהם בכל שלב במחזור החיים של המערכת:

- **חוקיות** - עמידה בכל החוקים והתקנות החלים.
 - **אתיות** - כיבוד עקרונות אתיים מקובלים וערכים חברתיים.
 - **חוסן טכני וחברתי** - יכולת לעמוד בפני כשלים טכניים והשפעות שליליות חברתיות.
- בנוסף, המסמך מגדיר **שבע דרישות עיקריות** שמערכות AI צריכות לעמוד בהן כדי להיחשב "אמינות":

⁶⁸ [Ethics guidelines for trustworthy AI](#), Directorate-General for Communications Networks, Content and Technology European Commission, 2019

1. **פיקוח אנושי** (Human agency and oversight)
 2. **חוסן טכני ובטיחותי** (Technical robustness and safety)
 3. **פרטיות וניהול נתונים** (Privacy and data governance)
 4. **שקיפות** (Transparency)
 5. **גיוון, אי-אפליה והוגנות** (Diversity, non-discrimination and fairness)
 6. **רווחה חברתית וסביבתית** (Societal and environmental well-being)
 7. **אחריותיות** (Accountability)
- כדי להבטיח יישום אפקטיבי של ההנחיות, פותח כלי הערכה ייעודי - ALTAI⁶⁹. הכלי מהווה **צ'ק-ליסט דינמי להערכה עצמית** עבור המפתחים, והמשתמשים במערכות AI, ומסייע להם לבחון את העמידה בדרישות.

6.1.3 מודל VPCIO

מודל ה-VPCIO⁷⁰ פותח כהרחבה למודל (VCIO) Values, Criteria, Indicators, Observables ומוסיף את שלב ה-Principles ומתמקד בחמש רמות:

- ערכים (Values)
- עקרונות (Principles)
- קריטריונים (Criteria)
- מדדים (Indicators)
- מדידות בפועל (Observables)

הערכים מייצגים מצבים רצויים - ערכים מוסריים כמו רווחה (well-being), צדק (justice), והבנה (understanding). הקריטריונים מגדירים כיצד העקרונות צריכים להתממש, המדדים שואלים שאלות ממוקדות לבחינת העמידה בקריטריונים, והמדידות בפועל מספקות תשובות מדורגות המאפשרות הערכה כמותית ואיכותית.

המודל מאפשר להתאים את יישום האתיקה להקשרים שונים ולצרכים של בעלי עניין שונים כגון מפתחים, רגולטורים, משתמשים ומושפעים.

⁶⁹[Assessment List for Trustworthy Artificial Intelligence \(ALTAI\) for self-assessment](#) Directorate-General for Communications Networks, Content and Technology European Commission, 2020

⁷⁰ Martin, Désirée, Michael W. Schmidt, and Rafaela Hillerbrand. "Implementing AI Ethics: The VPCIO Model." AI and Ethics, May 26, 2025. <https://doi.org/10.1007/s43681-025-00723-7>

6.2 חקיקה מרכזית להגנה מפני בינה מלאכותית

6.2.1 חוק הבינה המלאכותית באיחוד האירופי ⁷¹(AI Act)

זוהי החקיקה המקיפה ביותר והעדכנית (2026) הקיימת באיחוד האירופי לגבי שלל השימושים של בינה מלאכותית. החוק מחלק את השימושים השונים של בינה מלאכותית לשימושים אסורים, ולשימושים בסיכון גבוה עליהם חלות שלל הגבלות ופיקוח. שימושים אסורים כוללים מניפולציה מכוונת שנועדה לגרום לבני אדם לקבל החלטות שאינן מבוססות ידע או שסביר שתגרום נזק. הם כוללים אף אפיון התנהגות בני אדם במטרה לספק להם ציון חברתי ועל בסיסו להתייחס אליהם בצורה מזיקה או unfavourable, שימוש בטכנולוגיה על מנת לנבא האם אדם כלשהו יבצע עבירה בהתבסס אך ורק על אפיון תכונותיו (ולא בסיס עובדתי לגבי עבירות קיימות). קיימים מספר שימושים אסורים נוספים.

ההגדרה של שימושים בסיכון גבוה נשענת על הגדרות קודמות של מוצרים מסוכנים אשר קיימת לגביהם רגולציה. אלו כוללים מוצרים שכאשר יש בהם תקלה יש חשש לביטחונם של בני אדם, דוגמת רובוטים המבצעים ניתוחים. אלו כוללים אף מוצרים שכאשר יש בהם תקלה תהיה פגיעה בזכויות אזרחיות של אנשים. אלה כוללים בינה מלאכותית לצורכי הוראה, זיהוי, אכיפת חוק, אספקת שירותים בסיסיים ועוד. חובות החלות לגבי טכנולוגיות אלו כוללות בין היתר:

1. הטמעת מערכת ניהול סיכונים

2. ניהול נכון של נתונים

3. תיעוד מהימן

4. מגבלות שקיפות

5. השגחה אנושית

6. דיוק ואבטחת מידע

החוק אף מחיל חוקי שקיפות התקפים לטכנולוגיות בינה מלאכותית. על פי חוקים אלו, חלה החובה לסמן תוכן שנוצר על ידי AI, או לתייג שיחה עם AI באופן שמבהיר למשתמשים כי מדובר ב-AI.

בנוסף, החוק מכיל הגבלות עבור שימוש בטכנולוגיות בינה מלאכותית למטרות כלליות (General Purpose AI Systems). אלו כוללים תיעוד של המוצרים המיועדים להשתמש בטכנולוגיה זו, לספק להם הבנה טובה של מגבלות הטכנולוגיה ודרכי פעילותה.

6.2.2 חוק הגנת המידע באיחוד האירופי (GDPR (General Data Protection Regulation)

חקיקה זו קובעת כללים הנוגעים להגנה על אנשים בכל הנוגע לעיבוד נתונים אישיים, וכללים הנוגעים לתנועה החופשית של נתונים אישיים. מטרתה להגן על זכויות היסוד וחירויות היסוד של אנשים, ובפרט על זכותם

⁷¹ <https://artificialintelligenceact.eu/ai-act-explorer/>

להגנה על נתונים אישיים. תקנה זו חלה על עיבוד נתונים אישיים באמצעים אוטומטיים, כולו או חלקו, וכן על עיבוד שאינו אוטומטי של נתונים אישיים המהווים חלק ממערכת תיוק, או המיועדים להוות חלק ממערכת תיוק. הסעיפים השונים של החקיקה כוללים עקרונות כלליים בנוגע לעיבוד נתונים אישיים, תנאים למתן הסכמה, תנאים החלים על הסכמת קטין בנוגע לשירותי חברת המידע, עיבוד נתונים אישיים הנוגעים להרשעות ועבירות פליליות ועוד. החקיקה דנה אף בהעברת נתונים אישיים למדינות שלישיות או לארגונים בינלאומיים, ברשויות פיקוח ובענישה.

6.2.3 חוק השירותים הדיגיטליים באיחוד האירופי (Digital Services Act)

זוהי חקיקה מקיפה של האיחוד האירופי המסדירה שירותים מקוונים וחלה על כלל המדינות החברות בו. הרגולציה מחלקת את החובות לאלו החלות רק על ספקיות השירותים הגדולות במיוחד, לרבות מטא, גוגל, X וכו' ולחובות החלות על כלל השירותים. החקיקה מונה מגוון חובות של ספקיות שירותים מקוונים, לרבות חובות שקיפות, הימנעות ממניפולציה, פרסום כללי שימוש ברורים, הורדת תוכן בלתי חוקי, מתן כלים להעצמת משתמשים, ואסדרת הסמכויות בנושא. החקיקה גם מציגה חובות מוגברות להגנה על קטינים, דוגמת שקיפות והימנעות משימוש בפרטיהם האישיים למטרות שיווק. הרגולציה הזו מחייבת את החברות הגדולות להכין דו"חות שנתיים המתארים את האופן בו עמדו במגוון החובות המוטלות עליהן, ולמנות פקחים אשר יכתבו דו"חות עצמאיים בנושא.

6.2.4 חוק הבטיחות ברשת בממלכה המאוחדת (Online Safety Act⁷²)

חוק הבטיחות המקוונת נחקק בממלכה המאוחדת בשנת 2023, ומטרתו להגן על ילדים ומבוגרים ברשת. הוא מטיל מגוון חובות חדשות על חברות המדיה החברתית ושירותי החיפוש, ומחייב אותן לשאת באחריות רבה יותר לבטיחות המשתמשים בפלטפורמות שלהן. החוק מטיל על ספקים חובות חדשות להפחתת הסיכון שהשירותים שלהם ישמשו לפעילות בלתי חוקית, ולהסיר תוכן בלתי חוקי כאשר הוא מופיע.

ההגנות החזקות ביותר בחוק תוכננו עבור ילדים. פלטפורמות יידרשו למנוע מילדים גישה לתוכן מזיק ובלתי הולם לגילם, ולספק להורים ולילדים דרכים ברורות ונגישות לדווח על בעיות ברשת כאשר הן מתעוררות.

החוק גם מגן על משתמשים מבוגרים, ומבקש להבטיח שפלטפורמות גדולות יידרשו לשקיפות רבה יותר לגבי סוגי התוכן הפוטנציאלי המזיק שהן מתירות, ויעניק לאנשים שליטה רבה יותר על סוגי התוכן שהם רוצים לראות.

חובות הבטיחות של הספקים הן פרופורציונליות לגורמים כגון סיכון הפגיעה באנשים, וגודלו ויכולתו של כל ספק. כך מובטח שבעוד שיש להנהיג אמצעי בטיחות בכל התחומים, אין דרישה משירותים קטנים בעלי פונקציונליות מוגבלת לנקוט באותם צעדים כמו התאגידים הגדולים ביותר.

⁷²<https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer>

העבירות החדשות שהחוק הנהיג

העבירות הפליליות שהוכנסו על ידי החוק נכנסו לתוקף ב-31 בינואר 2024. עבירות אלה כוללות:

- עידוד או סיוע לפגיעה עצמית חמורה
- "סייברפלאשינג" (שליחת תמונות מיניות ללא הסכמה)
- שליחת מידע כוזב שנועד לגרום לנזק לא זניח
- תקשורת מאיימת
- פגיעה בתמונות אינטימיות
- הטרדת חולי אפילפסיה

החוק כבר יושם בפועל ובמסגרתו הורשעו אנשים בעבירות הסייברפלאשינג והתקשרות מאיימת.

סוגי תוכן שהחוק מתמודד עמם:

תוכן בלתי חוקי

החוק מחייב את כל החברות לנקוט פעולה נחרצת כנגד תוכן ופעילות בלתי חוקיים. פלטפורמות נדרשות כעת ליישם אמצעים להפחתת הסיכון שהשירותים שלהן ישמשו לעבריינות. הן גם צריכות להנהיג מערכות להסרת תוכן בלתי חוקי כאשר הוא מופיע. גם לשירותי חיפוש יש חובות חדשות לנקוט צעדים להפחתת הסיכון שמשתמשים ייתקלו בתוכן בלתי חוקי דרך השירותים שלהם.

סוגי התוכן והפעילות הבלתי חוקיים שמהם פלטפורמות צריכות להגן על משתמשים מפורטים בחוק, וכוללים תוכן הנוגע ל:

- התעללות מינית בילדים
- התנהגות שולטת או כפייתית
- אלימות מינית קיצונית
- פורנוגרפיה קיצונית
- הונאה
- עבירות סדר ציבורי מוחמרות מטעמים גזעיים או דתיים
- הסתה לאלימות
- הגירה בלתי חוקית והברחת בני אדם
- קידום או הקלת התאבדות
- פגיעה בתמונות אינטימיות
- מכירת סמים או נשק בלתי חוקיים
- ניצול מיני
- טרור

תוכן המזיק לילדים

הגנה על ילדים היא בליבת חוק הבטיחות המקוונת. אף שתוכן מסוים אינו בלתי חוקי, הוא עלול להיות מזיק או בלתי הולם לגילם של ילדים, ופלטפורמות צריכות להגן על ילדים מפניו.

חברות בעלות אתרים בתחולת החוק שצפויים להיות נגישים לילדים צריכות לנקוט צעדים להגן על ילדים מפני תוכן והתנהגות מזיקים.

קטגוריות התוכן המזיק שמהן פלטפורמות צריכות להגן על ילדים מפורטות בחוק. יש למנוע מילדים גישה לתוכן בעדיפות ראשית, ויש לתת להם גישה הולמת לגילם לתוכן בעדיפות. סוגי התוכן הנכללים בקטגוריות אלה מפורטים להלן.

תוכן בעדיפות ראשית

- פורנוגרפיה
- תוכן המעודד, מקדם, או מספק הוראות לגבי:
 - פגיעה עצמית
 - הפרעות אכילה, או
 - התאבדות

תוכן בעדיפות

- הצקה (בריונות)
- תוכן פוגעני או מלא שנאה
- תוכן המתאר או מעודד אלימות או פגיעה חמורה
- תוכן המעודד קסקדות ואתגרים מסוכנים; וכן
- תוכן המעודד בליעה, שאיפה, או חשיפה לחומרים מזיקים.

חוויות הולמות לגיל עבור ילדים ברשת

החוק מחייב חברות מדיה חברתית לאכוף את מגבלות הגיל שלהן באופן עקבי ולהגן על הילדים המשתמשים שלהן.

שירותים חייבים להעריך סיכונים לילדים משימוש בפלטפורמות שלהם ולקבוע מגבלות גיל הולמות, ולוודא שלמשתמשים הילדים יש חוויות הולמות לגילם ושהם מוגנים מתוכן מזיק. ספקי שירות המשתמשים באמצעים למניעת גישה לשירות שלהם לילדים מתחת לגיל מסוים חייבים לפרט בתנאי השירות שלהם אילו אמצעים משמשים לאכיפת מגבלת גיל זו, ולאכוף זאת באופן עקבי.

למבוגרים תהיה שליטה רבה יותר על התוכן שהם רואים

שירותי משתמש-למשתמש גדולים, הידועים כשירותי קטגוריה 1, נדרשים להציע למשתמשים מבוגרים כלים שבהם, אם יבחרו להשתמש, יעניקו להם שליטה רבה יותר על סוגי התוכן שהם רואים ועם מי הם מתקשרים ברשת. אלו כוללים את היכולת לאמת את זהותם ולקבל גישה לכלים המאפשרים להם להפחית את הסבירות שיראו תוכן ממשתמשים לא מאומתים ולמנוע ממשתמשים לא מאומתים לתקשר עם התוכן שלהם. החוק

אף קובע שעל הספקים להציע למשתמשים כלים אופציונליים שיעזרו להפחית את הסבירות שייתקלו בסוגים מסוימים של תוכן חוקי אך מזיק.

6.2.5 סעיף 230 ל-CDA בארצות הברית

בניגוד לאירופה, בארצות הברית לא קיימת רגולציה מקיפה על רשתות חברתיות. סעיף 230 ל-CDA בארה"ב (Communications Decency Act) מהווה את החקיקה המרכזית בנושא הרשתות החברתיות, ומטרתו להסיר אחריות מחברות אלו. סעיף זה כולל שני חלקים, החלק הראשון מסיר מתאגידי הרשתות החברתיות אחריות לתוכן שמוצא על ידי משתמשים והחלק השני מסיר מהן אחריות אף מההחלטה להסיר תוכן של משתמשים כל עוד פעלו בתום לב להסרתו של תוכן שסברו שהוא מזיק. הסעיף משמש את בתי משפט בארצות הברית על מנת לפטור תאגידי רשתות חברתיות מאחריות לתוכן אותו מעלים משתמשים ברשתות אלו אליהן.

היחס לאחריות תאגידי הרשתות החברתיות לנזקי הרשתות - תחילתו של שינוי כיוון בארצות הברית?

רקע

ב-2026 היו מספר קביעות של בתי משפט במדינות בארצות הברית אשר פירשו באופן שונה את אחריות הרשתות החברתיות ממה שהיה נפוץ עד אז. בקביעות נטען כי סעיף 230 אינו פוטר את תאגידי הרשתות החברתיות מאחריות לנזק שנגרם כתוצאה מהעצמה אלגוריתמית.

הליך משפטי בניו מקסיקו

בהליך זה נקבע כי מטא אחראית לכך שהאלגוריתם שלה מאפשר פגיעה מינית בקטינים. זאת בעקבות חקירה סמויה בה חוקרים התחזו לקטינים ואספו עדויות על שידול מיני.

הליך משפטי בקליפורניה

במקרה זה חבר מושבעים פסק כי למטא וליוטיוב יש אחריות לעיצוב אלגוריתמי ממכר. זאת בעקבות תביעה של אישה בת 20 שהעידה כי התמכרה לפלטפורמות אלו בגיל צעיר, מה שהוביל לדכאון ולפגיעה עצמית.

משמעות

לקביעות אלה השלכות כלכליות מהותיות לתאגידי הרשתות החברתיות, שכן הן מהוות תקדימים משפטיים הפותחים פתח לתביעות נוספות. קביעות רבות ממין זה בבתי משפט עשויות להוות תמריץ כלכלי משמעותי לחברות לשנות את אופן פועלן.

6.2.6 אמנה וולונטרית בנושא מידע כזב (דיסאינפורמציה) באיחוד האירופי

האיחוד האירופי ניסח גם אמנה וולונטרית בנושא דיסאינפורמציה⁷³. החקיקה אינה מחייבת את תאגידי הרשתות החברתיות להיות חתומות על האמנה הזו, החתימה עליה היא על בסיס התנדבותי. תאגידי המחליטים לחתום על האמנה מקבלים על עצמם את החובה לעמוד במספר דרישות, לרבות מחויבות להורדת התמריצים הכלכליים לפרסום מידע כוזב, שקיפות בנוגע לשיווק פוליטי, הפחתת מניפולציות המגבירות הפצה של מידע כזב, העצמת משתמשים על ידי הגברת שליטתם בתוכן אליו הם נחשפים, העצמת חוקרים על ידי מתן גישה מוגברת לנתונים שבידי החברות, והעצמת קהילת בודקי העובדות על ידי שיפור תנאי העבודה שלהם. ב-2025 הוחלט כי האמנה הזו תהיה באופן רשמי חלק מחוק השירותים הדיגיטליים האירופי.

6.2.7 העלאת רף הגיל לשימוש ברשתות חברתיות

צעירים המשתמשים ברשתות חברתיות נחשפים לנזקים דוגמת התמכרות, בירונות רשת, ניצול מיני ועוד. במספר מדינות מערביות התקבלה ההחלטה כי הדרך הטובה ביותר להגן על קטינים מפני נזקי הרשתות החברתיות היא להעלות את רף הגיל לשימוש בהן. נכון להיום, רשתות אלו רשמית אינן מאפשרות לצעירים מתחת לגיל 13 להשתמש בשירותיהן, ב-2026 נכנס לתוקף באוסטרליה חוק אשר מגביל את גיל הכניסה לרשתות חברתיות לגיל 16. מדינות רבות נוספות שוקלות הגבלה דומה. במדינות דוגמת צרפת, פולין וספרד הועלו הצעות חוק ברוח זו.

6.3 קודים אתיים למודלים של בינה מלאכותית

קודים אתיים למודלים של בינה מלאכותית מבקשים להציג עקרונות מוסריים שעל חברות המפתחות כלי בינה מלאכותית לאמץ ושלאורם על ממשלות לכוון את החקיקה שלהן. אלו אינם בעלי תוקף חוקי כשלעצמם, אולם הם משמשים מצפון מוסרי אחיד עבור המעוניינים לעמוד בכללים אתיים.

הקוד האתי הראשון למודלים של בינה מלאכותית כללית התקבל על ידי הנציבות האירופית ונועד לשמש ככלי וולונטרי שיסייע לתעשייה לעמוד בדרישות החוק.

הקוד גובש ביוזמת משרד הבינה המלאכותית האירופי, בהובלת מומחים עצמאיים. תהליך גיבוש הקוד היה שקוף והשתתפו בו כ-1000 בעלי עניין (ספקי מודלים, ארגוני תעשייה, החברה האזרחית, בעלי זכויות, אקדמיה ומומחים עצמאיים), נציגי מדינות האיחוד האירופי, משקיפים אירופיים ובינלאומיים.

הקוד האתי למודלים של בינה מלאכותית כללית (GPAI) כולל שלושה פרקים עיקריים:

⁷³ European Commission, The 2022 Code of Practice on Disinformation, <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation>

1. שקיפות וזכויות יוצרים

שני פרקים אלו חלים על כלל ספקי המודלים של GPAI ומטרתם להבטיח שהמודלים פועלים באופן שקוף, תוך שמירה על זכויות יוצרים של תכנים שנעשה בהם שימוש.

2. בטיחות ואבטחה

פרק זה רלוונטי רק לקבוצת ספקים מצומצמת – אלו שמפתחים את המודלים המתקדמים ביותר, אשר עשויים לשאת סיכונים מערכתיים בשל יכולותיהם הגבוהות או תפוצתם הרחבה.

ארגונים נוספים הציעו אף הם עקרונות אתיים אשר לאורם יש להמשיך לפתח כלי בינה מלאכותית. כמה דוגמאות מרכזיות כוללות:

איגוד מכונות המחשוב (ACM) ומכון מהנדסי החשמל והאלקטרוניקה (IEEE): מוסדות אלו הציעו כי על פיתוחים כאלה לעמוד בתנאים של כשירות, סודיות, ניגוד עניינים, פגיעה/נזק, חוקים/תקנים, מקצועיות, פיתוח מקצועי, קידום הבנה, שקיפות, אחריותיות, פרטיות

האגודה הבריטית למחשבים (BCS) והאגודה האוסטרלית למחשבים (ACS): מוסדות אלו הציעו פיתוח בינה מלאכותית על פי ערכים של מקצועיות, כשירות, יושרה, טובת הציבור

6.3.1 מסגרת האתיקה של ממשלת בריטניה לנתונים ובינה מלאכותית⁷⁴

ממשלת בריטניה מספקת מסגרת עקרונות המיועדת לעובדי הסקטור הציבורי (מפתחים, מנהלי פרויקטים, אנליסטים, קובעי מדיניות ועוד) להנחיית שימוש אחראי בנתונים ובינה מלאכותית. עקרונות המפתח:

- שקיפות: תקשור ברור ונגיש של מידע על הפרויקט, כולל היכולת להסביר כיצד המערכת מקבלת החלטות.
- אחריותיות: ממשל ופיקוח ברורים, ואפשרות לערער על החלטות.
- הוגנות: הימנעות מהטיות לא הוגנות, טיפול בנגישות, והתאמה לחוקי השוויון הבריטיים.
- פרטיות: שימוש בנתונים אישיים אך ורק למטרה המיועדת, בהתאם לרגולציית הגנת המידע.
- בטיחות: מערכות חסינות, מאובטחות ואמינות שאינן פוגעות באנשים או בסביבה.
- השפעה חברתית וסביבתית: תרומה לפתרון אתגרים חברתיים ולקיימות.

6.3.2 ערכי הליבה של אונסק"ו (UNESCO)

ארבעה ערכי יסוד להנחיית בינה מלאכותית לטובת האנושות:⁷⁵

1. כבוד וזכויות אדם

⁷⁴ The UK: Data and AI Ethics Framework

⁷⁵ <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

2. חברות צודקות, שלוות ומחוברות
3. גיוון והכללה
4. שגשוג הסביבה והמערכות האקולוגיות

6.3.3 הנחיות האתיקה של הנציבות האירופית (2019)⁷⁶

בינה מלאכותית "אמינה" (Trustworthy AI) צריכה להיות: חוקית, אתית וחסונה. הוגדרו 7 דרישות מרכזיות:

1. פיקוח אנושי: המערכת מעצימה אנשים ומאפשרת החלטות מושכלות, עם מנגנוני פיקוח (אדם בתוך/על/מעל הלולאה).
2. חוסן ובטיחות טכניים: מערכות מאובטחות, מדויקות ובעלות תוכנית גיבוי למקרה כשל.
3. פרטיות וממשל נתונים: הגנה על פרטיות וגישה מוסדרת לנתונים איכותיים.
4. שקיפות: מודל עסקי, מערכת והחלטות ניתנים למעקב והסבר; יידוע המשתמש שהוא מתקשר עם AI.
5. גיוון, אי-אפליה והוגנות: מניעת הטיות שפוגעות בקבוצות מוחלשות, נגישות לכולם.
6. רווחה חברתית וסביבתית: קיימות והתחשבות בהשפעה החברתית והסביבתית, כולל דורות עתידיים.
7. אחריותיות: מנגנוני ביקורת (audit) ואפשרויות תיקון/פיצוי.

⁷⁶ <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>



neaman.org.il