**מיכל שור-עופרי***

# מעֵבֶר לשקיפות: ריבוי ומגוון במדיניות בינה מלאכותית

מאמר זה מציע לעגן עיקרון של ריבוי (multiplicity) כחלק מעקרונות המדיניות המנחים בתחום הבינה המלאכותית.

ההתפתחות המהירה של הבינה המלאכותית בשנים האחרונות הולידה יוזמות מדיניות ואסדרה במדינות רבות. מדיניות זו מבוססת, במקרים רבים, על כמה עקרונות מרכזיים שיחולו בתחום, כמו, למשל, הסברתיות (explainability), אבטחת מידע, או מעורבות אנושית בתהליכי קבלת החלטות על ידי יישומי בינה מלאכותית. ואולם עקרונות אלה, שחשיבותם אינה מוטלת בספק, אינם מתמודדים באופן מלא עם השלכות הרוחב החברתיות הכרוכות בהתפתחות תחום הבינה המלאכותית היוצרת (Generative AI), בפרט בתחום של מודלים גדולים של שפה (large language models).

היכולת (הלכאורית) של מודלי השפה לספק מידע כמעט בכל נושא, לצד שילובם של כלים כאלה בפלטפורמות טכנולוגיות מרכזיות כמו מנועי חיפוש, עתידים להפוך אותם, כבר בעתיד הנראה לעין, לערוץ עיקרי לצריכת מידע. בחינה של הטכנולוגיה שביסוד המודלים הללו, לצד מחקרים מתחומי התקשורת, הסוציולוגיה והתרבות מעלה כי מודלי השפה הגדולים עתידים לשקף לנו תמונת עולם ממורכזת וצרה יחסית, הנוטה אל עבר הפופולרי וה״מיינסטרימי״, על חשבון מגוון של תכנים וריבוי נראטיבים. תודות לשילוב של מאפייני הטכנולוגיה עם מאפיינים הקשורים באינטראקציה האנושית, הכוח של המודלים להשפיע ולעצב את תמונת עולמם של המשתמשים בהם עתיד להיות גדול במיוחד. מצב דברים זה עלול לייצר השלכות חברתיות שליליות, גם כאשר המידע שהמודלים מספקים אינו מידע שגוי. אלה כוללות פגיעה במגוון התרבותי, בזיכרון הקולקטיבי ובדיאלוג הדמוקרטי. בשל המאפיינים הייחודיים של השפה, התרבות והחברה בישראל, הפגיעה עלולה להיות גדולה במיוחד בהקשר המקומי.

* פרופ׳ למשפטים, האוניברסיטה העברית בירושלים. אני מודה לבר הורוביץ־אמסלם, לתמר ולר ולדרור זמיר על עזרת מחקר מעולה. מאמר זה הוא חלק ממחקר מקיף יותר שחלקים ממנו הוצגו באוניברסיטה העברית בירושלים, באוניברסיטת אינדיאנה, בסדנת AI של אוניברסיטת טורונטו והטכניון, ובסדנת ה־Information Law Institute של NYU. אני מודה למשתתפי הפורומים הללו על הערות ותובנות חשובות, וכן לחברי מערכת ״משפט, חברה ותרבות״ ולקוראים האנונימיים על הערות והצעות מועילות.

בחינת העקרונות האתיים-רגולטוריים המתהווים בתחום הבינה המלאכותית בישראל (תוך השוואה למדינות נוספות) מעלה שהם אינם מספיקים כדי לתת מענה לקשיים האלו. על רקע זה המאמר מציע לעגן, כחלק מהעקרונות המנחים את המדיניות בתחום, גם עיקרון מפורש של ריבוי (multiplicity): הצגת ריבוי ומגוון של תכנים ואפשרויות בפני המשתמשים בבינה מלאכותית יוצרת, או לפחות יצירת מודעות בקרבם של המשתמשים לקיומם של תכנים ואפשרויות נוספים, מעבר לברירות המחדל שהמודלים מציגים להם. המאמר ממשיך ומתווה קווים ליישום העיקרון, ובכללם עיצוב הטכנולוגיה באופן המקדם ריבוי ומגוון (multiplicity by design), הבטחת מגוון מספיק בשוק המודלים עצמו, לצד קידום אוריינות AI בקרב המשתמשים.

## מבוא

ההתקדמות המהירה בתחום הבינה המלאכותית בשנים האחרונות מושכת תשומת לב רבה מצד הציבור וקובעי המדיניות כאחד. אחת ההתפתחויות הבולטות בהקשר זה נוגעת לבינה מלאכותית יוצרת (Generative AI), מערכות העושות שימוש בניתוח מאגרי נתונים מאסיביים על מנת לייצר תכנים חדשים מסוגים שונים, כמו טקסטים, תמונות, סרטים ועוד. בקטגוריה זו נכללים גם מודלים גדולים של שפה (Large Language Models, שיכונו בהמשך מטעמי קיצור גם "מודלים").[1] מודלים אלה יכולים לבצע מגוון משימות רחב אף אם לא הוכשרו להן באופן מיוחד, ובכלל זה תשובות לשאלות בנושאים שונים, תכנות, חיבור טקסטים, סיכום מידע, מתן המלצות ועוד כהנה וכהנה משימות, והכול תוך כדי תקשורת עם המשתמש בשפה אנושית טבעית.[2] לנוכח מנעד היכולות הנרחב שלהם מודלים כאלה מכונים גם "מודלים לכל מטרה" (general-purpose models), ובשפה מדוברת יותר "ask-me-anything models", והם נכללים בהגדרה של "מודלי בסיס" (foundation models) שמופיעה ברגולציה המתהווה בתחום הבינה המלאכותית.[3] אף שגרסאות מוקדמות של מערכות כאלה היו בנמצא כבר שנים,

1 מודלים גדולים של שפה (large language models) הם מערכות בינה מלאכותית שנבנו כדי להבין, לנתח וליצור טקסטים בשפה טבעית. ראו באופן כללי Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 2021 ACM Conf. on Fairness, Accountability & Transparency 610.

2 ראו שם, וכן Xuandong Zhao et al., *Pre-trained Language Models Can be Fully Zero-Shot Learners*, arXiv (May 26, 2023), https://arxiv.org/abs/2212.06950.

3 מודלי בסיס הם מודלים שביכולתם לשרת מגוון גדול של משימות וזאת מבלי שאומנו לכך באופן מפורש. להגדרת מודלי בסיס ראו משרד החדשנות, המדע והטכנולוגיה **עקרונות מדיניות, רגולציה ואתיקה בתחום הבינה המלאכותית** 8 (2023) (להלן: מסמך העקרונות). ראו גם Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, ARXIV (Aug. 18, 2021), https://arxiv.org/abs/2108.07258.

דומה שהשקת מודל השפה ChatGPT3 בחודש נובמבר 2022 והתפוצה הנרחבת שזכה לה תוך ימים ספורים[4] מסמנים קו פרשת מים ביחסים שבין החברה כולה לבין הבינה המלאכותית: שימוש בבינה מלאכותית אינו עוד נחלתם הבלעדית של אנשי מקצוע או "גיקים" טכנולוגיים, אלא נמצא כעת ב"קצות האצבעות" של המוני משתמשים. המספר הגדול של מודלים של שפה שהושקו בעת האחרונה (כולל, בין היתר, Gemini של גוגל, Llama של מטא, גרסאות מתקדמות יותר מבית היוצר של Open AI ועוד) מצביע על כך שהשימוש במודלים ממשיך להתרחב. מבחינת קובעי מדיניות, זהו עיתוי נכון לבחון את ההשפעות של המודלים, את הסיכונים הכרוכים בשימוש בהם, ואת התגובה הרגולטורית המתבקשת, אם בכלל.

כרקע לדיון, חשוב להבחין בין שני סוגי סיכונים שיכולים לנבוע מהשימוש הנרחב בבינה מלאכותית: סיכונים אינדיבידואליים וסיכונים סיסטמיים. כאשר סיכון מהסוג הראשון מתממש, נגרם לאדם (או לגוף משפטי) נזק מוגדר: הפרת זכויות יוצרים על ידי בינה מלאכותית, או נזק שנגרם למי שסבל מאפליה כתוצאה מהחלטה אלגוריתמית, הם דוגמאות לנזקים כאלה. לעומת זאת, כאשר מדובר בסיכונים סיסטמיים, לא תמיד אפשר להצביע על נזק ישיר לאדם ספציפי, אולם במצטבר ולאורך זמן התממשות סיכונים כאלה עלולה לגרום לנזק חברתי ניכר.[5] הבעיה של מידע כוזב המופץ ברשתות חברתיות ("כדור הארץ הוא שטוח", "הקורונה אינה קיימת") היא דוגמה לנזקים מהסוג האחרון: כדור הארץ אומנם לא יגיש תביעה בגין מידע כוזב, אבל באופן מצטבר, לפרסומים כאלה עלולה להיות השפעה חברתית הרסנית.[6]

מאמר זה מתמקד באחד **הסיכונים הסיסטמיים** הכרוכים בשימוש הנרחב בבינה מלאכותית יוצרת, ובייחוד במודלי השפה הגדולים, שלא זכה עד כה לתשומת לב רבה מצד חוקרים ורגולטורים: החשש מפני צמצום המגוון. טענתי היא שבטווח הארוך, מודלי שפה גדולים עלולים לצמצם את "עולם המחשבות העולות-על-הדעת", כולל הזיכרון הקולקטיבי, הנראטיבים ההיסטוריים, תפיסת העולם וההעדפות התרבותיות של

4 לסקירה מטעם OpenAI, יצרנית טכנולוגיית ה־GPT העומדת ביסוד המודל וביסוד הגרסאות המתקדמות שלו, ראו OpenAI, GPT-4 Technical Report, ARXIV (Mar. 4, 2023), https://arxiv.org/abs/2303.08774. להתפתחות הבינה המלאכותית מסוג ask-me-anything לאורך השנים ראו MELANIE MITCHELL, ARTIFICIAL INTELLIGENCE: A GUIDE FOR THINKING HUMANS, 215–219 (2019).

5 ראו הבחנות דומות אצל Rishi Bommasani et al., *Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?*, 2022 CONF. ON NEURAL INFO. PROCESSING SYS. 1; Noam Kolt, *Algorithmic Black Swans*, 101 WASH. U. L. REV. 1177, 1213 (2024).

6 למקצת ההשפעות הללו ראו למשל David M. J. Lazer et al., *The Science of Fake News*, 359 SCI. 1094 (2018); Rachel Leigh Greenspan & Elizabeth F. Loftus, *Pandemics and Infodemics: Research on the Effects of Misinformation on Memory*, 3 HUM. BEHAV. & EMERGING TECH. 8 (2021).

אנשים, וזאת לא רק כאשר התוצרים של אותם מודלים הם שגויים או לא מדויקים, אלא **גם כאשר הם אמינים והגיוניים**.[7]

עוד אטען כי בשל מאפיינים יסודיים של הטכנולוגיה שיתוארו במאמר זה, הפלט של מודלים שנועדו לכל מטרה, דוגמת GPT4 או Gemini, סביר שישקף נקודת מבט צרה ומיינסטרימית, תוך מתן עדיפות לתוכן פופולרי וקונבנציונלי על פני תכנים ונראטיבים מגוונים.[8] לאורך זמן, ההסתמכות על מודלים אלו עלולה לשנות תפיסות חברתיות ולהגביר אחידות וסטנדרטיזציה על חשבון מגוון וריבוי. לשינוי כזה, בתורו, עלולות להיות השפעות חברתיות שליליות – החל מהפחתת המגוון התרבותי, דרך הגבלת הגישה לריבוי הנראטיבים הבונים זיכרון קולקטיבי, ועד לצמצום תפיסות עולם ופגיעה בדיאלוג הדמוקרטי.[9] על רקע הגיוון והמורכבות המאפיינים את החברה הישראלית, ובהיותה של התרבות הישראלית תרבות שהיא נחלתה של קבוצה קטנה יחסית לאוכלוסיית העולם, סיכונים אלה עשויים להיות משמעותיים במיוחד בהקשר המקומי.[10]

הניתוח במאמר זה מראה עוד שבשל שורה של מאפיינים טכנולוגיים ועיצוביים, למודלי שפה גדולים צפויה להיות השפעה גדולה במיוחד על תפיסות והעדפות המשתמשים בהשוואה לאמצעי תקשורת אחרים כמו עיתונים, ערוצי טלוויזיה, רשתות חברתיות או מנועי חיפוש.[11] במקום אחר כיניתי תופעה זו "אפקט מולטיוואק", על שם מחשב-העל המופיע בסיפוריו של אסימוב המאגד את כל המידע האנושי בעולם ומשמש כסמכות העליונה בכל עניין.[12]

על מנת להתמודד עם אתגרים אלה אטען שיש להכיר במפורש **בריבוי (multiplicity) כחלק מעקרונות המדיניות בתחום הבינה המלאכותית**. אציע כי ריבוי, בהקשר זה, משמעו חשיפת המשתמשים, או לפחות יידוע שלהם על אודות קיומם של אפשרויות, תכנים ונראטיבים נוספים, מעבר לברירות המחדל שהמודל מציג, תוך עידוד שלהם לחפש מידע נוסף. החשיפה לריבוי של פלטים ואפשרויות צפויה להפחית את "אפקט מולטיוואק", כלומר למתן את הכוח הסמכותי של המודלים הגדולים, ולאפשר

7 המונח "עולם המחשבות העולות על הדעת" הוא תרגום חופשי של הביטוי "world of thinkable thoughts" שהופיע בעבר בספרות שעסקה במחקר וחיפוש משפטי. ראו: Robert C. Berring, *Legal Research and the World of Thinkable Thoughts*, 2 J. APP. PRAC. & PROCESS 305 (2000); Daniel Dabney, *The Universe of Thinkable Thoughts: Literary Warrant and West's Key Number System*, 99(2) L. LIBR. J. 229 (2007).

8 פרק א.2 להלן.

9 פרק א.4 להלן.

10 פרק א.2.4 להלן.

11 פרק א.3 להלן.

12 ראו Isaac Asimov, *All the Troubles of the World*, in ISAAC ASIMOV, THE COMPLETE STORIES, 387–388 (1958); לפירוט ודיון בהשוואה זו ראו Michal Shur-Ofry, *Multiplicity as an AI Governance Principle*, INDIANA L.J. (forthcoming 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4444354. לכתיבה נוספת שדנה באנלוגיה שבין אלגוריתמים לבין מולטיוואק ראו Michal S. Gal, *Algorithmic Challenges to Autonomous Choice*, 25 MICH. TECH. L. REV. 59, 95 (2018).

לנו לראות אותם כהווייתם: כלים טכנולוגיים, ולא אורים ותומים יודעי כול. מעבר לכך, עיגון העיקרון של ריבוי כחלק מהעקרונות האתיים-רגולטוריים בתחום הבינה המלאכותית יסייע להרחיב את תפיסות העולם של המשתמשים, יתמוך במגוון תרבותי ובזיכרון קולקטיבי, ויקדם דיאלוג דמוקרטי.

המאמר בנוי כך: **הפרק הראשון** עושה שימוש בספרות אינטרדיסציפלינרית, כדי לנתח את הדרכים שבהן מודלי שפה גדולים עשויים להשפיע על "עולם המחשבות" האנושי. הוא בוחן את הטכנולוגיות העומדות בבסיס המודלים, ומבהיר מדוע סביר שהפלט של מודלי שפה גדולים יהיה מכוון-פופולריות וישקף תפיסת עולם "מיינסטרימית" וריכוזית, במקום ריבוי של תכנים ונראטיבים. הפרק ממשיך ומנתח את ההשפעה הצפויה של מודלי השפה על תפיסות המשתמשים בהם ואת הסיכון החברתי הסיסטמי שעלול להיווצר כתוצאה מההטיה אל עבר המרכזי והפופולרי, הן באופן כללי, והן ובפרט בהקשר הישראלי-מקומי.

בהמשך לניתוח זה, **הפרק השני** מציע לשלב את העיקרון של **ריבוי (multiplicity)** כחלק מעקרונות המדיניות המנחים בתחום הבינה מלאכותית בישראל. הדיון בחלק זה מתחיל בסקירת הרגולציה המתהווה בתחום הבינה המלאכותית בכמה שיטות משפט מרכזיות וכן בישראל, ומבהיר מדוע נדרשת הכרה מפורשת בעיקרון של ריבוי ומגוון. **הפרק השלישי** בוחן דרכים **ליישום** עקרון הריבוי כחלק ממדיניות של בינה מלאכותית, תוך שרטוט שלושה כיוונים (לא ממצים). הראשון הוא "ריבוי מובנה" (Multiplicity-by-Design): עיצוב הארכיטקטורה של מודלי השפה בצורה שתחשוף את המשתמשים, או לפחות תיידע אותם על קיומם של אפשרויות, תפיסות עולם ונראטיבים נוספים, ותעודד אותם לחפש מידע נוסף. נתיב שני הוא הבטחת זמינות של כמה מודלי שפה גדולים, באופן שיאפשר למשתמשים לקבל תשובה מכמה מקורות. לבסוף, הניתוח מצביע על כך שלצד פעולות משפטיות-רגולטוריות, קידום של אוריינות משתמשים בתחום הבינה המלאכותית צפוי לתרום תרומה מהותית לשימור הריבוי והמגוון האנושיים.

## א. בינה מלאכותית יוצרת ועולם המחשבות העולות–על–הדעת

האם מודלי שפה גדולים ישפיעו על תפיסות העולם שלנו, ואם כן, כיצד? השפעה אחת, שזוכה לתשומת לב רבה מצד חוקרים, רגולטורים ואמצעי תקשורת, נוגעת לכך שהתוצרים של מודלי השפה אינם תמיד אמינים. הם כוללים טעויות ואי־דיוקים. הם עלולים להיות מוטים. לפעמים, הם שגויים בעליל. המודלים עלולים להפיק, למשל, אזכורים מדעיים מומצאים, ציטוטים לא נכונים, הפניות לפסקי דין שאינם קיימים, ועוד כהנה וכהנה דוגמאות שזכו לכינוי "הזיות" (hallucinations).[13]

13 ראו, למשל, Ziwei Ji et al., *Survey of Hallucination in Natural Language Generation*, 55 ACM COMPUTING SURV. 248 (2023); Gary Marcus, *A Few Words About Bullshit*,

ואולם, המחיר החברתי הכרוך בהתפשטות המואצת של בינה מלאכותית יוצרת אינו מתמצה בהפצת תוכן כוזב. מודלי שפה גדולים יכולים להשפיע על תפיסות העולם שלנו גם כאשר התוכן שהם יוצרים הוא מהימן ובעל ערך. סוג כזה של השפעה רלוונטי בעיקר במקרים שבהם **אין** "תשובה נכונה אחת" לשאלה או לבקשה שמופנית מצד המשתמש למודל ("פרומפט"), אלא יש מגוון של תשובות אפשריות ומקום לשיקול דעת: בקשת מידע על מתכון, המלצה על סדרת טלוויזיה, או מידע בנוגע לדמויות היסטוריות, הן דוגמאות לשאלות כאלו.[14] ההשפעה מהסוג הזה קשורה קשר הדוק להיותם של המודלים מערכות שמארגנות מידע ומתווכות אותו למשתמשים. הסעיפים הבאים בוחנים מקרוב את הטכנולוגיה העומדת בבסיס המודלים הללו, ואת השיפוט וסדרי העדיפויות האנושיים המוטבעים בהם. בחינה זו מבהירה כיצד פועלים מודלי שפה גדולים כמִבני מידע שיש בהם רכיב בלתי נמנע של שיפוט אנושי, ומדוע הם עלולים לצמצם את תפיסות העולם של המשתמשים בהם.

## 1. מודלים של בינה מלאכותית כאתרים של יצירת משמעות

על מנת להבין מדוע הפלט של מודלי השפה הגדולים, גם כאשר הוא מהימן ובעל ערך, אינו שיקוף "אובייקטיבי" של המציאות, מתבקשת סקירה קצרה של הטכנולוגיה שבבסיס מודלים אלה.

מרבית המודלים בנויים כיום בתצורה של רשתות עצביות (neural networks). הרשת העצבית המלאכותית בנויה מיחידות הקשורות זו לזו, המאורגנות במבנה של שכבות. חלק מהיחידות הללו מתקשרות עם הקלט שהמודל מקבל, חלקן מתקשרות את הפלט אל המשתמש, כאשר בין אלה לאלה קיימות שכבות נוספות (המבנה הרב-שכבתי הוא המקור להתייחסות לטכנולוגיה כאל "למידה עמוקה" – deep learning).[15] בנוסף, הממשק שבין המודלים לבין המשתמשים מבוסס על טכנולוגיה של עיבוד שפה טבעית (Natural Language Processing או NLP) שמאפשרת למודל לעבד את השפה האנושית שבה המשתמש מתקשר, ולתקשר את הפלט חזרה אל המשתמש בשפה אנושית-טבעית.[16] אותו פלט מושפע מהצבר של גורמים שהם אומנם טכנולוגיים אך אינם דטרמיניסטיים, אלא משקפים סדרה של החלטות הכרוכות בשיקול דעת אנושי ובחירה בין חלופות: החל מבחירת מאגרי המידע שישמשו לאימון המודל, עבור דרך תהליך האימון עצמו, וכלה בצורה שבה המודל יתקשר עם המשתמשים. אבהיר.

---

MARCUS ON AI (Nov. 16, 2022), https://garymarcus.substack.com/p/a-few-words-about-bullshit.

14 השוו Kiel Brennan-Marquez & Vincent Chiao, *Algorithmic Decision-Making When Humans Disagree on Ends*, 24 NEW CRIM. L. REV. 275, 278–283 (2021) (המציעים, בהקשר של החלטות אלגוריתמיות, הבחנה בין סוגי מקרים מעין אלה). לדיון במקצת הדוגמאות האלה ראו פרק א.2 להלן.

15 מיטשל, לעיל ה"ש 4, בעמ' 178; בומסאני ואח', לעיל ה"ש 3.

16 מיטשל, שם, בעמ' 35–38.

ככלל, מודלים של שפה מתאמנים על **מאגרי מידע** הכוללים כמויות אדירות של טקסט שישמשו כחומרי אימון: עמודי אינטרנט, ויקיפדיה, מאגרים מדעיים, תכנים מרשתות חברתיות ועוד. אימון זה מאפשר למערכות אלו לזהות תבניות בטקסטים קיימים ולהחיל אותן על מידע חדש.[17] מאגרי האימון, ככלל, אינם גלויים למשתמשים. אף בהנחה שמדובר במאגרים איכותיים (כלומר מאגרי תכנים ממקור ידוע שנתפסים כאמינים, כמו עיתונים או מאגרי מידע אקדמיים), הם אינם אובייקטיביים, אלא יש בבחירתם ממד סובייקטיבי ואף פוליטי (במובן הרחב) המשפיע בהכרח על הפלט של המודל.[18] חשבו על דוגמה פשוטה: מודל שפה כללי שהאימון שלו מתבסס על מאגרי מידע בשפה האנגלית צפוי להפיק פלט שונה ממודל שאומן על מאגרי מידע בשפה של מדינה אפריקנית. כלומר, ההחלטה באילו נתונים להשתמש היא חלק מיצירת ה"יקום" שממנו תשאב הבינה המלאכותית את ה"מחשבות העולות-על-הדעת".

ממד דומה של שיקול דעת קיים גם ב**תהליך האימון**: כיום, השלב הראשון של תהליך האימון נעשה לרוב באמצעות "למידה עצמית", שבה הנתונים שכלולים במאגרי המידע מתויגים על ידי מודל השפה עצמו.[19] ואולם, לאחר שלב ראשוני זה, המודלים עוברים סוגים נוספים של אימונים, שדורשים מעורבות אנושית משמעותית. אחת משיטות האימון הרווחות, נכון למועד כתיבת דברים אלה, היא "חיזוק למידה ממשוב אנושי" (Reinforcement Learning Human Feedback, או RLHF).[20] תהליך זה כרוך באיסוף הפלטים שהמודל יצר בתגובה לפרומפט מסוים ודירוגם על בסיס השוואה שלהם לתשובה הרצויה, כפי שהיא נקבעת על ידי מאמנים אנושיים. הדירוגים האלה משמשים לאימון מודל נוסף, הקרוי "מודל תגמול", שממשיך ומאמן את המודל המקורי באמצעות מתן "תגמול" על תשובות טובות.[21] באופן כללי יותר, תהליך האימון רחוק

17 שם, בעמ' 98–100.

18 ראו Kate Crawford & Trevor Paglen, *Excavating AI: The Politics of Training Sets for Machine Learning* 36 AI & SOC. 1105 (2021), שהציגו תובנה דומה ביחס למאגרים המשמשים לאימון בינה מלאכותית בתחום הוויזואלי. בנוסף, ראו: Gordon Hull, *Dirty Data Labeled Dirt Cheap: Epistemic Injustice in Machine Learning Systems*, ETHICS & INFO. TECH. (forthcoming); וכן בנדר ואח', לעיל ה"ש 1, בעמ' 613–614.

19 Louis-François Bouchard, *What is Self-Supervised Learning? Will Machines Ever Be Able to Learn Like Humans?* MEDIUM (May 27, 2020), https://medium.com/what-is-artificial-intelligence/what-is-self-supervised-learning-will-machines-be-able-to-learn-like-humans-d9160f40cdd1.

20 לסקירה של שיטה זו ראו, למשל, Yekun Chai et al., *MA-RLHF: Reinforcement Learning from Human Feedback with Macro Actions*, ARXIV (Oct. 3, 2024), https://arxiv.org/abs/2410.02743.

21 שיטות נוספות לכיול המודלים הן ביצוע "כיול" על ידי מאמנים אנושיים. ראו ההסבר שפרסמה OpenAI, יצרנית מודלי השפה בסדרת GPT: ("We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides – the user and an AI assistant"), OpenAI, *Introducing ChatGPT* (Nov. 30, 2022), https://openai.com/blog/chatgpt.

מלהיות תהליך מכני ואחיד, אלא כרוך במידה כזאת של מיומנות ושיקול דעת עד שחלק ממובילי התעשייה תיארו אותו בעבר כ"אומנות" או "אלכימיה".[22] בדומה, גם ההחלטות בנוגע לאופן שבו מודלי שפה גדולים יתקשרו עם המשתמש רחוקות מלהיות אובייקטיביות. כך, למשל, ניסוח הפלט כפסקה בסגנון של שיחה אנושית מעביר מסר שונה מאשר הצגת רשימה "יבשה" של נקודות רלוונטיות. גם היכולת להמשיך בשיחה, לייצר תגובה נוספת, ההפניה למקורות רלוונטיים או ההימנעות מהפניה כזאת – כל אלה משקפים בחירות אנושיות אשר (כפי שנראה מייד) משפיעות על האינטראקציה בין המודל למשתמשים.

התמונה המצטיירת היא, אפוא, שהפלט של מערכות בינה מלאכותית, גם כשהוא אמין ובעל ערך, אינו אובייקטיבי או ניטרלי. מודלי השפה הגדולים הם אתרים יוצרי-משמעות, המשקפים שורה של החלטות ושיקול דעת אנושיים. הפלט של המודלים הללו הוא, במידה רבה, פרי של מוסכמות, בחירות והירכיות חברתיות. הם יוצרים פריזמה המשקפת שיפוט **מסוים**, מייצרת ציפיות, ובאופן כללי, יכולה להשפיע על תפיסות עולם.[23] בטרם נעבור לבחון מקרוב את ההשפעה הפוטנציאלית הזאת ואת יחסי הכוחות בין המודלים לבין המשתמשים בהם, ראוי להפנות מבט לכמה מאפיינים של המודלים הגדולים שמבהירים כי התוכן שהם מייצרים צפוי לנטות אל עבר המרכזי, הסטנדרטי והפופולרי.

## 2. העדפת הפופולרי והנטייה אל עבר המיינסטרים

מבט קרוב על הטכנולוגיה מבליט שני גורמים שבגינם מודלים של שפה צפויים לייצר פלט ממורכז ולא מגוון, גם במקרים שבהם יש מגוון גדול של תשובות אפשריות. **הגורם הראשון** הוא מאגרי הנתונים שמשמשים לאימון המודל. כפי שהוסבר בפסקאות הקודמות, מאגרי המידע המשמשים לאימון המודלים הם עולם התוכן שממנו תשאב הבינה המלאכותית את המחשבות העולות-על-הדעת. יש להניח שמרבית החומרים והתכנים שעוברים דיגיטיזציה או נכללים במאגרי מידע דיגיטליים הם חומרים שמקורם במדינות המערב, ומרביתם בשפה האנגלית.[24] על כן, היקום שמודלים אלה שואבים

22 ראו Jason Tanz, *Soon We Won't Program Computers. We'll Train Them Like Dogs,* WIRED (May 17, 2016), https://www.wired.com/2016/05/the-end-of-code.

23 תובנה בסיסית זו מופיעה במחקרי הזרם הידוע כ-STS (Science, Technology, and Society), המצביעים על כך שטכנולוגיה, ככלל, אינה ניטרלית, אלא מושפעת משיקולים חברתיים, פוליטיים, תרבותיים וכלכליים. ראו באופן כללי, EDWARD J. HACKETT ET AL., THE HANDBOOK OF SCIENCE AND TECHNOLOGY STUDIES (3d ed. 2007). לדיון בהקשר של מודלי שפה ראו פרקים א.3 ו־א.4 להלן.

24 לספרות מתפתחת בתחום מדעי המחשב הבוחנת את התוכן והתכונות של מערכי נתונים המשמשים לאימון בינה מלאכותית, ראו: Jesse Dodge et al., *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*, ARXIV (Sep. 30, 2021), https://arxiv.org/abs/2104.08758; Margaret Mitchell et al., *Measuring Data*, ARXIV at 2, 8–9 (Feb. 13, 2023), https://arxiv.org/abs/2212.05129.

ממנו את המחשבות העולות-על-הדעת **משקף בעיקר את העולם האנגלו-אמריקני-מערבי, על תפיסותיו, גיבוריו, עברו והעדפותיו התרבותיות**, מה שמניב באופן בלתי נמנע ייצוג דומיננטי של עולם זה. התפיסות, התוצרים וההעדפות של תרבויות אחרות ומגוונות סביר שיידחקו לשוליים.

כך, לדוגמה, באחת הבדיקות שביצעתי בסמוך להשקת מודל השפה ChatGPT3 הפניתי למודל, באמצעות משתמשים שונים, את השאלה מהי סדרת הטלוויזיה הטובה ביותר. כל אחת מהתגובות שהתקבלו כללה שמות של כמה סדרות. התשובות של המודל היו דומות, אך לא זהות לחלוטין, בכל אחת מהפעמים בהן הופנתה אליו השאלה; בסך הכול התקבלו 211 בחירות, שכללו 21 סדרות שונות. כל הסדרות שהציע המודל, ללא יוצא מן הכלל, היו סדרות אנגלו-אמריקאיות, דוברות אנגלית ושוברות קופות, כמו למשל ה"סופרנוס", "משחקי הכס" או "שובר שורות".[25] חרף ההיצע העצום של סדרות טלוויזיה "בעולם האמיתי", לא הופיעו בפלט ברירת המחדל של המודל סדרות ממדינות שאינן אנגלו-אמריקניות. בדומה, כאשר המודל התבקש לנקוב בשלושת האנשים המשפיעים ביותר במאה ה־19 התקבלה רשימת שמות קצרה שהשמות השכיחים ביותר בה היו אברהם לינקולן, המלכה ויקטוריה, צ'רלס דרווין ונפוליאון בונפרטה. אף במקרה זה, הפלט של המודל לא כלל מנהיגים אסייתיים, או קיסרים רוסיים, אלא אך ורק דמויות שהן דומיננטיות בראייה מערבית.[26]

**הגורם השני** והמשמעותי לנטייה של המודלים הגדולים להתרכז בנפוץ ובפופולרי הוא הפרדיגמה הטכנולוגית שבבסיסם, הנשענת על שכיחות סטטיסטית של קשרים בין מילים ומושגים.[27] במילים אחרות (והתיאור כאן מעט פשטני, לטובת בהירות הדברים), כחלק מהאימון, המודלים לומדים לזהות תבניות שכיחות במידע שבמאגרי המידע. בפרט, הם משתמשים בשכיחות סטטיסטית כדי לזהות אילו מילים ומשפטים נוטים לבוא אחרי טקסט קודם, וכך לייצר פלט רלוונטי ומובן.[28] עיקרון זה, הידוע בספרות כ־next-word-prediction paradigm, ובתרגום חופשי "פרדיגמת חיזוי-המילה-הבאה", חשוב במיוחד לעניינו: משמעותו היא שהפלט של מודל השפה מושפע מאוד משכיחות של מילים וצירופים בחומרי האימון. לכן, כברירת מחדל, גם כאשר המודל נשאל שאלה שיש לה מגוון של תשובות אפשריות, הוא צפוי לשקף בתשובותיו את המושגים התרבותיים, הדמויות, התפיסות והנראטיבים הפופולריים ביותר. הצירוף בין

25 לפירוט הבדיקה, המתודולוגיה, המגבלות והממצאים ראו: שור־עופרי, לעיל ה"ש 12, בסעיף 2B2; לבדיקה אמפירית מקיפה ועדכנית יותר המדגימה תופעה זו בקרב שמונה מודלים שונים ראו: Michal Shur-Ofry, Bar Horowitz-Amsalem, Adir Rahamim & Yonatan Belinkov, *Growing a Tail: Increasing Output Diversity in Large Language Models*, ARXIV (Nov. 5, 2024), https://arxiv.org/abs/2411.02989.

26 לפירוט ראו שור־עופרי, לעיל ה"ש 12, בסעיף 1B2.

27 ראו Sébastien Bubeck et al., *Sparks of Artificial General Intelligence: Early experiments with GPT-4*, ARXIV 80 (Mar. 22, 2023), arxiv.org/abs/2303.12712.

28 ראו הסבר של OpenAI בדקה 24:00-56:00, https://campus.datacamp.com/courses/working-with-the-openai-api/openais-text-and-chat-capabilities?ex=1. ראו גם מיטשל, לעיל ה"ש 4, בעמ' 190–196.

מאגרים אנגלו-אמריקאיים לבין הפרדיגמה הטכנולוגית של "חיזוי-המילה-הבאה" מבהיר, למשל, מדוע בדוגמאות שתיארתי, המודל שנשאל על סדרות טלוויזיה מעולות הציע שוב ושוב את ה"סופרנוס" ולא את הסדרה הסקנדינבית "הגשר" – סביר להניח שבמאגרי המידע שעליהם אומן המודל המילים "סדרת הטלוויזיה" ו"הטובה ביותר" הופיעו בסמוך למילה "הסופרנוס" בתדירות גבוהה יותר מאשר למילה "הגשר". בדומה, פרדיגמת "חיזוי-המילה-הבאה" מבהירה מדוע מודל שהתבקש לציין דמויות מרכזיות במאה ה־19 הזכיר שוב ושוב את אברהם לינקולן, אך לא הזכיר שליטים אסייתיים מאותה תקופה.[29]

שתי תובנות מרכזיות עולות מהדיון עד כה. **ראשית**, הטקסטים שנוצרים על ידי מודלי שפה גדולים, גם כאשר הם רלוונטיים והגיוניים, אינם – ולא יכולים להיות – אובייקטיביים לחלוטין. מודלים גדולים הם מבני מידע שמתווכים מידע למשתמשים. בנייתם ופעולתם כרוכים באופן בלתי נמנע בבחירה בין אפשרויות, בקביעת יחסים ובהגדרת היררכיות בין פיסות מידע.[30] הם משקפים בהכרח ראיית עולם מסוימת. **שנית**, מכיוון שהמודלים ניזונים לעיתים קרובות ממאגרי מידע אנגלו-אמריקאיים והתוצרים שלהם מושפעים מאוד מהסתברות סטטיסטית בשל פרדיגמת "חיזוי-המילה-הבאה" שבבסיס הטכנולוגיה, הם ייטו לכיוון הפופולרי והמיינסטרים. שילוב המאפיינים האלה גורם למודלי שפה גדולים להעדיף אחידות ותבניתיות על פני ריבוי וגיוון. כתוצאה מכך, סביר שהפלט של המודלים ישקף השקפת עולם ממורכזת, יתרכז סביב נראטיבים דומיננטיים, ויחזק את הפופולריות של מה שפופולרי ממילא.[31]

חשוב להכיר בכך שבמידה מסוימת הנטייה אל עבר הפופולרי קיימת לא רק בבינה מלאכותית אלא גם בקרב בני אדם. מחקרים רבים בתחום התרבות ובתורת הרשתות (network science) מראים שבחירות של אנשים משקפות לעיתים קרובות דינמיקה של "המנצח לוקח הכול": אנשים נוטים להיות מושפעים מבחירות של אחרים ברשת החברתית שלהם, וכתוצאה מתהליכי השפעה חברתית כאלה מספר קטן יחסית של "מוצרים" (כולל מוצרי תרבות, טכנולוגיות או אנשים מפורסמים) זוכים במרבית תשומת הלב הציבורית.[32] אכן, ההטיה הקיימת כבר כיום אל עבר הפופולרי מחדדת את

29 ראו לעיל ה"ש 25–26 והטקסט הנלווה.

30 כפי שהדיון בחלק הבא מבהיר, זוהי תכונה משותפת לכל מבני המידע; ראו להלן ה"ש 37–42 והטקסט הנלווה.

31 תופעה זו מזכירה, אך לא זהה, לתופעת ה"מונוקולטורה האלגוריתמית", הנדונה בספרות מדעי המחשב, ומתייחסת למצב שבו מקבלי החלטות מסתמכים כולם על אותו אלגוריתם. ראו: Jon Kleinberg & Manish Raghavan, *Algorithmic Monoculture and Social Welfare*, 118(22) PROC. NAT'L ACAD. SCI. U.S AM. (2021); בהקשר דומה, ראו את רעיון ה"הומוגניזציה של תוצאות" המתייחס ל"תופעה שבה יחידים (או קבוצות) מקבלים באופן בלעדי תוצאות שליליות מכל מקבלי ההחלטות שעימם הם מקיימים אינטראקציה": בומאסני, לעיל ה"ש 5 (תרגום שלי).

32 לסקירת הממצאים העיקריים והתהליכים שמובילים לדינמיקה הזאת ראו EVERETT M. ROGERS, DIFFUSION OF INNOVATIONS 23–34 (5th ed. 2003); Matthew J. Salganik et

הצורך בהגנה על המגוון. ואולם, הבעיה צפויה להחריף בעידן מודלי השפה: המחקרים מצביעים גם על קיומו של "זנב ארוך" בבחירות האנושיות, כלומר מוצרים רבים מקבלים תשומת לב **מסוימת**, אף אם זו פחותה בהרבה מזו הניתנת ל"להיטים".[33] כך, למשל, בדיקה של תשובות שנתנו אנשים לשאלה מהי "סדרת הטלוויזיה הטובה ביותר" בפיד פייסבוק העלתה שלצד הלהיטים האנגלו-אמריקניים שקיבלו את רוב ה"קולות" (ביניהם "חברים", "שובר שורות", "הסופרנוס" ו"נמלטים") המגיבים האנושיים בחרו במספר גדול יחסית של סדרות אחרות, שהיו גם מגוונות יותר מבחינת מדינות המקור שלהן, וכללו גם סדרות ספרדיות, טורקיות, איריות-קנדיות, דרום-קוריאניות, איטלקיות, דניות, צרפתיות וישראליות.[34] לעומת זאת, בחירות המחדל של ChatGPT התמקדו במספר קטן של סדרות ממקור אנגלו-אמריקאי, וכמעט "מחקו" את הזנב הארוך שנתן משקל מסוים לסדרות אחרות. התרשים הבא מתאר באופן גרפי את הממצאים הללו.[35]

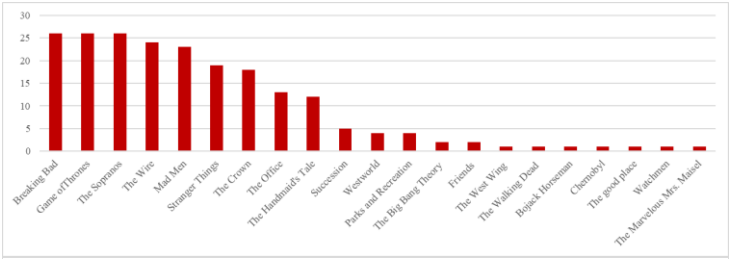
al., *Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market*, 311 Sci., 854 (2006); מיכל שור־עופרי **פופולריות ורשתות בדיני זכויות יוצרים** 23–49 (2011) (להלן: שור־עופרי פופולריות ורשתות); מיכל שור־עופרי "זכויות יוצרים ומגוון תרבותי – האם המטבע מתחת לפנס?" **האם המשפט חשוב?** 569, 584–587 (דפנה הקר ונטע זיו עורכות 2010) (להלן: שור־עופרי מגוון תרבותי). תורת הרשתות היא תחום אינטרדיסציפלינרי, החוקר מאפיינים ותבניות המשותפים למערכות מסוגים שונים אשר מורכבות מרכיבים הקשורים זה לזה ומתקשרים זה עם זה, תוך שימוש בכלים מתמטיים ובמודלים פיסיקליים. ראו באופן כללי שור־עופרי פופולריות ורשתות, בעמ' 27–38.

33 לדיון בתופעת הזנב הארוך, ראו שור־עופרי מגוון תרבותי, שם, בעמ' 587–589.

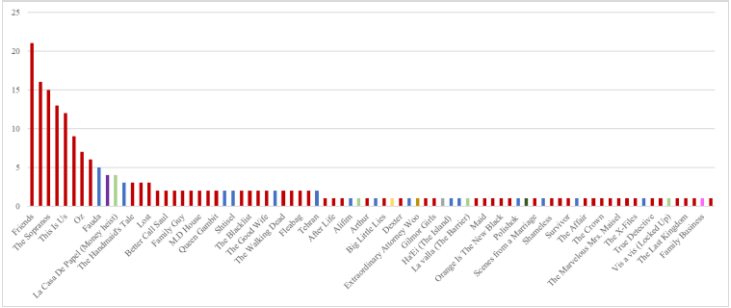
34 211 בחירות אנושיות כללו 82 סדרות שונות (לעומת 21 סדרות שונות שמנה המודל), שכשליש מהן לא היו אנגלו-אמריקאיות. לתיאור מפורט של הבדיקה, כולל המתודולוגיה, התוצאות והמגבלות ראו שור־עופרי, לעיל ה"ש 12, Table 2 והטקסט הנלווה. להשוואות נוספות ראו שור־עופרי ואח', לעיל ה"ש 25.

35 התרשים מתאר, בתמצית, את הממצאים המפורטים אצל שור־עופרי, לעיל ה"ש 12, בסעיף 2B2 ומוצגים שם ב־Table 1 ו־Table 2.

AI




HUMANS




לנוכח ממצאים אלה סביר, אפוא, שהעולם האנושי, למרות נטייתו הטבעית לפופולרי, הוא עדיין רחב ומגוון יותר לעומת ברירות המחדל שתשתקפנה אלינו ממודלי השפה.[36]

מה תהיה ההשפעה של תמונת עולם ממורכזת כזאת על המשתמשים במודלים הללו? האם הטכנולוגיה של מודלי שפה גדולים שונה, לעניין זה, מאמצעי תקשורת קיימים? בחלק הבא אציע כי בשל שורת גורמים טכנולוגיים ועיצוביים, למודלים של שפה עלולה להיות השפעה חזקה במיוחד על המשתמשים ותפיסות העולם שלהם.

## 3. יחסי הכוחות בין מודלים ומשתמשים

ספרות ענפה בתחומי התרבות, התקשורת והמשפט מצביעה על כך שכל מדיום שמתווך מידע משקף בהכרח שיפוט ושיקול דעת אנושיים בנוגע למשמעות ולחשיבות של אותו מידע, ובכך כופה על המשתמשים פריזמה נורמטיבית מסוימת שעשויה להשפיע על ראיית העולם שלהם.[37] כך, למשל, חוקר התקשורת אדווין בייקר הסביר שערוצי תקשורת ההמונים נוטים לחשוף את הצופים לתוכניות אחידות, נוסחתיות וקלות

36 שור־עופרי ואח', לעיל ה"ש 25.

37 ראו, למשל, Niva Elkin-Koren, *Cyberlaw and Social Change: A Democratic Approach to Copyright Law in Cyberspac*e, 14 CARDOZO ARTS & ENT. L. J. 215 (1996); Michal Shur-Ofry, *Databases and Dynamism*, 44 MICH. J. L. REFORM 315 (2011); Eric Goldman, *Search Engine Bias and the Demise of Search Engine Utopianism*, 8 YALE J.L. & TECH. 188, 198–199 (2006); Michal Shur-Ofry & Guy Pessach, *Robotic Collective Memory*, 97 WASH. U. L. REV. 975, 987–989 (2020).

לעיכול, וכי חשיפה כזאת משפיעה על טעמם של הצופים ומגבירה את התיאבון שלהם לתוכניות נוסחתיות, על חשבון תכנים מגוונים ומורכבים יותר.[38] גם פלטפורמות מדיה חברתית, מפייסבוק ועד טוויטר (X), מציגות מידע למשתמשים בהן בדרכים סלקטיביות שעשויות להשפיע על התפיסות של משתמשים אלה. מחקרים רבים מראים שהפלטפורמות הללו נוהגות לחשוף את המשתמשים לאנשים בעלי דעות דומות לשלהם ולא לדעות ולתכנים מגוונים, תופעה היוצרת "תיבות תהודה" (echo-chambers) המהדהדות שוב ושוב את אותן הדעות ועלולות לגרום לדינמיקה של הקצנה.[39] מנועי חיפוש אינטרנטיים משפיעים אף הם על תפיסות המשתמשים בהם. כך, למשל, האלגוריתם של גוגל מייחס משקל לא מבוטל לפופולריות בדירוג תוצאות החיפוש: אתרים פופולריים מקבלים עדיפות בדירוג, מה שבתורו צפוי להגביר עוד יותר את הפופולריות שלהם.[40] באופן דומה, מחקרים מראים שגם פלטפורמות של המלצות-מבוססות-אלגוריתמים חושפות את המשתמשים שלהן לתכנים שאינם צפויים לשקף מגוון אמיתי.[41] בקצה השני של הספקטרום הטכנולוגי, אפילו הארגון של מאגר מידע אנלוגי כמו מדריך "דפי זהב" משקף באופן בלתי נמנע היררכיות ויחסי כוח שיכולים להשפיע, באופן חמקמק ולא מורגש, על ראיית העולם של המשתמשים בו.[42] ואולם, במקרה של מודלי שפה גדולים, הצבר של מאפיינים טכנולוגיים ובחירות עיצוביות צפוי להוביל להשפעה חזקה במיוחד על המשתמשים בהם.

**ראשית**, הפלט של המודלים הגדולים, לפחות בשלב הטכנולוגי הנוכחי, הוא לעיתים קרובות **מנותק מחומרי הגלם**. בניגוד למנועי חיפוש, ברוב המקרים המודלים אינם מאחזרים את המקורות המקוריים שנכללים במאגרי המידע ששימשו לאימון שלהם.[43] במקום זאת, המודלים יוצרים טקסט חדש המתומצת במשפט, פסקה או עמוד,

38 C. EDWIN BAKER, MEDIA, MARKETS, AND DEMOCRACY 24 (W. Lance Bennett & Robert M. Entman eds., 2001). לדיון בסוגיה זו, ראו גם Guy Pessach, *Copyright Law as a Silencing Restriction on Noninfringing Materials: Unveiling the Scope of Copyright's Diversity Externalities*, 76 S. CAL. L. REV. 1067 (2003).

39 ראו, למשל, Matteo Cinelliet et al., *The Echo Chamber Effect on Social Media*, 118(9) PROC. NAT'L ACAD. SCI. U.S AM. (2021); YOCHAI BENKLER ET AL., NETWORK PROPAGANDA: MANIPULATION, DISINFORMATION AND RADICALIZATION IN AMERICAN POLITICS (2018); Philipp Lorenz-Spreen et al., *A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy*, 7 NATURE HUM. BEHAV. 74 (2023).

40 Goldman, לעיל ה"ש 37, בעמ' 193; Lucas D. Introna & Helen Nissenbaum, *Shaping the Web: Why the Politics of Search Engines Matters*, 16 INFO. SOC'Y 169, 176, 181 (2000).

41 Jonathan Gingerich, *Is Spotify Bad for Democracy? Artificial Intelligence, Cultural Democracy, and Law*, 24 YALE J.L. & TECH. 227 (2022).

42 שור־עופרי Databases, לעיל ה"ש 37, בעמ' 317, 325.

43 יש לציין שמצב עניינים זה עשוי להשתנות בעתיד, שכן בתקופה האחרונה חלק מהמודלים החלו להציע התייחסות לחומרי המקור, אך שינויים כאלה, נכון לכתיבת

ותלוי במידה רבה באופן שבו המשתמשים מנסחים את ההנחיות (פרומפט) שלהם. חשוב לציין: הפלט המובל-על-ידי-המשתמש, לצד שיטת ההצגה התמציתית והאינטגרטיבית, הם שהופכים את הטכנולוגיה של מודלי השפה ליעילה ולידידותית למשתמש. אולם במקביל, תכונות אלה מייצרות גם תמונת עולם צרה יותר, בהשוואה לטכנולוגיות שמאפשרות צפייה בחומרי הבסיס וגישה עצמאית למקורות מידע שנוצרו על ידי צדדים שלישיים.

נדבך **שני**, שמחזק את חוסר הסימטריה ביחסי הכוחות בין מודלי השפה לבין המשתמשים בהם, הוא מאפיין ה"קופסה השחורה" של המודלים. ההבחנה בבחירות ובסדרי העדיפויות האנושיים העומדים בבסיס של טכנולוגיות רבות לתיווך ולהעברת מידע במקרים רבים אינה קלה ואינה מובנת מאליה.[44] במקרה של בינה מלאכותית יוצרת, שיקול הדעת האנושי נסתר במיוחד. כפי שמבהיר הדיון הקודם, הטכנולוגיה הבסיסית אינה מאפשרת להתחקות אחר "תהליך החשיבה" של המודל. המרכיבים שהניבו את הפלט – כמו מאגרי המידע שעליהם התבסס האימון, העקרונות הטכנולוגיים הרבים שמגדירים את פעולת המודל (המכונים "היפר-פרמטרים") והמשוב האנושי במהלך האימון – אינם נגישים כלל למשתמשים. יותר מכך: תהליך למידת המכונה כולל מעגלי היזון חוזר (feedback loops) שבהם המודל מכייל את עצמו בעקבות משוב ללא תכנות מפורש. כך, חלק מהתהליך שמשפיע על הפלט הספציפי שהמשתמש מקבל מתבצע למעשה על ידי מערכת הבינה המלאכותית עצמה, ואינו ניתן להסבר מדויק, אפילו על ידי יוצרי המערכת.[45] כל אלה עלולים לחזק את תפיסת המודל כ"אורקל" יודע כול, שהפלט שלו אינו ניתן לערעור.

מאפיין **שלישי** וחשוב לענייננו קשור לממד הזמן: הכוח של מודלי שפה גדולים בעיצוב עולם המחשבות העולות-על-הדעת צפוי לגדול עם הזמן. השימוש הנרחב במודלים של שפה צפוי להוליד מעגלי היזון חוזר שבמסגרתם הטקסטים שנוצרו על ידי מודלי שפה גדולים יחלחלו חזרה לרשת וישמשו כחומרי אימון לדור הבא של המודלים, וחוזר חלילה. תהליך זה עלול לגרום ליצירת "**תיבות תהודה של בינה מלאכותית**" (AI echo-chambers) שבהן הבינה המלאכותית ניזונה מהתכנים שהיא עצמה יצרה.[46] במצב כזה, כאשר מאגרי המידע יהיו מוצפים בתכנים שנוצרו על ידי

---

הדברים, רחוקים מלהיות מקיפים. ראו, למשל, את ההסבר של גוגל על מודל השפה Gemini: "שאלות נפוצות על ממשקי Gemini" **Google** https://gemini.google.com/faq.

44 ראו, למשל, GEOFFREY C. BOWKER & SUSAN LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES 323 (1999).

45 ראו מיטשל, לעיל ה"ש 4, בעמ' 109, וכן הדיון בתהליך האימון בחלק א.1.

46 לדיון בהיבטים שונים של תופעה זו ראו, למשל, Ethan Perez et al., *Discovering Language Model Behaviors with Model-Written Evaluations* ARXIV (Dec. 19, 2022), https://arxiv.org/pdf/2212.09251.pdf; Ruibo Liu et al., *Best Practices and Lessons Learned on Synthetic Data for Language Models*, ARXIV (Aug. 10, 2024), https://arxiv.org/abs/2404.07503; Rohan Taori & Tatsunori B. Hashimoto, *Data*

מודלי השפה הגדולים, יש להניח שמשקלם של תכנים אחרים ומגוונים בקרב חומרי האימון ילך ויפחת.[47] דינמיקה כזאת יכולה להגביר ולחזק את השפעת המודלים ואת הנטייה אל עבר הסטנדרטי, האחיד והמיינסטרימי. חשש דומה הובע גם על ידי OpenAI, המפתחים של מודלי ה־GPT, בציינם: "ככל שיתרחב השימוש ב־GPT4 ובמערכות בינה מלאכותית דומות בתחומים של ידע, למידה וחדשנות, וככל שתגבר השפעת הנתונים שנוצרים כתוצאה מהשימוש על חומרי האימון, יגדל הכוח של מערכות בינה מלאכותית להעצים אידאולוגיות, תפיסות עולם, אמיתות ואי-אמיתות, **לקבע ולנעול אותן** באופן שימנע את האפשרות לחלוק עליהן, להרהר בהן או לשפר אותן...".[48]

**רביעית**, היכולת של מודלי שפה גדולים להשפיע על תפיסות המשתמשים מושתתת גם על **הטיות אנושיות** שמאפיינות את הקשר שבין האדם למכונה. תופעה אחת היא **הטיית האוטומציה**, קרי הנטייה האנושית לסמוך על פלט שנוצר על ידי מכונה.[49] במקרה שלנו, ניסוח הפלט בצורה תמציתית, ברורה ולעיתים נחרצת, מטשטשת את קיומן של אינספור חלופות נוספות. כפי שציינה מיכל גל במחקר שעסק במערכות המלצה אלגוריתמיות, "משתמש שאינו מודע למגבלות האלגוריתם, סביר שלא יהיה מודע לאפשרויות שהוא ויתר עליהן".[50] בכך, הטכנולוגיה של מודלי השפה נבדלת מזו של מנועי החיפוש: גם אם רובנו איננו מגיעים מעבר לדף הראשון של תוצאות החיפוש בגוגל, עצם הידיעה שהשאילתה שלנו הניבה (למשל) אלפי תוצאות יוצרת, בצורה עדינה אך משמעותית, מודעות לקיומו של עולם שלם של חומרים נוספים הכוללים מידע רלוונטי. לעומת זאת, הטכנולוגיה של מודלי השפה הגדולים **מסכמת וממסכת את העולם**, באופן שיוצר את הרושם שהתשובה שיצר המודל היא **ה**תשובה, בה"א הידיעה, ואין בלתה. מאפיינים אלה, לצד היכולת המרשימה של המודלים הכלליים להשיב כמעט על כל שאלה, צפויים להזין את הטיית האוטומציה, להקנות לתשובות המודלים הילה של סמכות בעיני המשתמש, ולהגביר את "אפקט המולטיוואק" ואת נכונות המשתמשים להסתמך עליהם.[51]

---

*Feedback Loops: Model-driven Amplification of Dataset Biases*, ARXIV (Sep. 8, 2022), https://arxiv.org/pdf/2209.03942.

47 שם. ראו גם Hull, לעיל ה"ש 18, בעמ' 19.

48 OpenAI, לעיל ה"ש 4, בעמ' 49 (תרגום שלי, ההדגשה הוספה).

49 ראו, למשל, Mary Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems*, AIAA 1ST INTELLIGENT SYS. TECH. CONF. (2004).

50 מערכות המלצה אלגוריתמיות הן מערכות שמשתמשות באלגוריתמים כדי להציע מוצרים, שירותים או תוכן למשתמשים על פי העדפותיהם הקודמות או על פי מידע אחר שנאסף עליהם. ראו גל, לעיל ה"ש 12, בעמ' 74 (הציטוט הוא תרגום שלי).

51 השוו Alexander Campolo & Kate Crawford, *Enchanted Determinism: Power Without Responsibility in Artificial Intelligence*, 6 ENGAGING SCI., TECH. & SOC'Y 1 (2020) (הדנים בתופעת ה"היקסמות" (enchantment), שלפיה אנשים מייחסים יכולות על-אנושיות למכונות למידה עמוקה).

**ולבסוף**, מודלי השפה הגדולים שייכים לקטגוריה של "**רובוטים חברתיים**" (social robots): יישומי בינה מלאכותית שהקשר שלהם עם המשתמש הוא מעין-אנושי, תוך שהם מפגינים יכולת חברתיות וכמו-אנושיות, יכולות הסתגלות ומיומנויות למידה.[52] מחקרים רבים מראים שתכונות חברתיות כאלה של אינטראקציה מעין-אנושית מעוררות **אנתרופומורפיזם** – נטייה של בני אדם לייחס תכונות אנושיות לבינה המלאכותית.[53] המיומנויות המתקדמות של מודלי השפה הגדולים, היצירה האוטונומית של טקסט חדש, מנעד היכולות המרשים, כישורי התקשורת המצוינים, היכולת לנהל שיחה טבעית לכאורה, ליצור אינטראקציה ולשתף פעולה – כל התכונות הללו צפויות לעורר תגובה רגשית מצד המשתמשים. אפילו משתמשים מתוחכמים עשויים לבצע האנשה של יישומים אלה, ולהתייחס אליהם כאל יותר מִכְּלים אלגוריתמיים.[54] מאפיינים אלה של יחסי אדם-מכונה צפויים להגביר את האמון של משתמשים בפלט של הבינה המלאכותית היוצרת, ואת ההשפעה של המודל על ראיית העולם האנושית. על רקע זה, קשה שלא להיזכר בתיאור שהציע אסימוב עוד בשנת 1958 ליחסים בין בני האדם למחשב-העל מולטיוואק: "לכל מי שהפנה שאלה [למולטיוואק]... היה אמון [בתשובות שלו]. וזה [כל] מה שהיה נחשב".[55]

***

מצבור התכונות המתואר כאן מצביע על כך שלמודלי השפה הגדולים צפויה להיות השפעה עצומה על תפיסות המשתמשים. סברה זו נתמכת גם על ידי ממצאים אמפיריים, אף אם ראשוניים.[56] האם השפעה זו, שצפויה להוביל להפחתת הריבוי והמגוון, צריכה להדאיג אותנו כחברה? הסעיף הבא פונה לבחון עניין זה.

52 ראו Cynthia Breazeal, *Towards Sociable Robots*, 42 ROBOTICS AND AUTONOMOUS SYS. 167 (2003); Kate Darling, *"Who's Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy* in ROBOT ETHICS 2.0 (P. Lin et al. eds. 2015).

53 לדיון במושג האנתרופומורפיזם, ראו לדוגמה Kate Darling, *Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects*, *in* ROBOT LAW 213 (M. Froomkin et al. eds. 2016); Breazeal, לעיל ה"ש 52, בעמ' 168; שור־עופרי ופסח, לעיל ה"ש 37.

54 Breazeal, לעיל ה"ש 52, בעמ' 164–165; Darling, לעיל ה"ש 52, בעמ' 6.

55 אסימוב, לעיל ה"ש 12, בעמ' 388 (תרגום שלי, במקור: "...every questioner knew the answer to be the best available and had faith in it. And that was what counted").

56 ראו Maurice Jakesch et al., *Co-Writing with Opinionated Language Models Affects Users' Views*, ARXIV (Feb. 1, 2023), https://arxiv.org/pdf/2302.00560.pdf.

## 4. העלויות החברתיות

### 4.1. מבט כללי

מהן, אם בכלל, העלויות החברתיות הכרוכות בנטייה המסתברת של מודלי שפה גדולים אל המיינסטרים והשכיח, על חשבון ריבוי ומגוון? מחקרים רבים בתחומי התרבות, הזיכרון והדמוקרטיה המתדיינת מבהירים את חשיבותם של ריבוי ושל מגוון.[57] ספרות זו מצביעה על כך שעצם החשיפה לריבוי של תפיסות עולם, ותרבויות מסוגים שונים – מקומית וגלובלית, גבוהה, פופולרית ועממית, תרבות לאומית וכן תרבות של "אחרים" – היא גורם בונה ומעצים.[58] עצם החשיפה לתכנים מגוונים, המכונה בתחום לימודי התרבות exposure diversity, יוצרת בקרב אנשים מודעות לדעות, לטעמים ולתפיסות שונות, דבר המקדם סובלנות ושוויון ויכול לשמש בלם כנגד קיצוניות.[59] ההיכרות עם מגוון יצירות, דעות ותפיסות עולם מסייעת לאנשים לגבש את הערכים, סדרי העדיפויות והתפיסות שלהם,[60] ומהווה גם גורם בונה ומעצים במישור האישי.[61] לעומת זאת, המחסור בריבוי ובמגוון יכול לצמצם את תפיסת עולמנו, ועלול לגרום להדרה של

57 למקצת הספרות בנושא זה ראו, למשל, בייקר, לעיל ה"ש 38, בעמ' 93–94; JOHN STUART MILL, ON LIBERTY, 96–132 (1859); Robert C. Post, *Democratic Constitutionalism and Cultural Heterogeneity*, 25 AUSTL. J. LEG. PHIL. 185 (2000); SEYLA BENHABIB, THE CLAIMS OF CULTURE: EQUALITY AND DIVERSITY IN THE GLOBAL ERA (2002); YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMERS MARKETS AND FREEDOM 169 (2006). ראו גם שור־עופרי מגוון תרבותי, לעיל ה"ש 32, בעמ' 576–578.

58 שם.

59 שם. ראו גם עדנה הראל פישר **ממלכתיות במדיניות תרבות** 38–39 (מחקר מדיניות 146, המכון הישראלי לדמוקרטיה 2020), המציינת שמגוון תרבותי מעודד גם שיתוף פעולה וסובלנות בין־מדינתיים. למונח exposure diversity ראו: James G. Webster, *Diversity of Exposure* in MEDIA, DIVERSITY AND LOCALISM: MEANING AND METRICS 13 (Philip M. Napoli ed. 2007); וכן ראו Jurgen Habermas, *Three Normative Models of Democracy*, in DEMOCRACY AND DIFFERENCE (Seyla Benhabib ed., 1996); Cristina M. Rodriguez, *Language and Participation*, 94 CAL. L. REV. 687, 726–727 (2006). להשלכות של היעדר מגוון ראו, למשל, Miller McPherson et al., *Birds of a Feather: Homophily in Social Networks*, 27 ANNU. REV. SOCIOL. 415 (2001).

60 ראו Jack M. Balkin, *Digital Speech and a Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. Rev. 1, 34–39 (2004); אלקין־קורן, לעיל ה"ש 37, בעמ' 232–233; שור־עופרי מגוון תרבותי, לעיל ה"ש 32, בעמ' 577.

61 בייקר, לעיל ה"ש 38, בעמ' 237.

״אחרים״, אלה שאינם תואמים את הסטנדרטי והקונבנציונלי.[62] לגיון ולריבוי יש אפוא חשיבות מכרעת בעידוד סובלנות ויציבות חברתית, ויש להם משמעות דמוקרטית עמוקה.[63]

באופן דומה, גם המחקר הסוציולוגי בתחום הזיכרון הקולקטיבי, העוסק באופן שבו קבוצות חברתיות זוכרות את עברן המשותף, מאיר את חשיבות הריבוי ליצירת זיכרון קולקטיבי.[64] ספרות זו מבהירה כי זיכרון קולקטיבי הוא גורם חיוני בגיבוש זהות קבוצתית, מהווה אמצעי להעצמת מיעוטים, וכן בונה את תחושת העצמי והזהות של הפרט.[65] באופן חשוב לענייננו, הזיכרון הקולקטיבי אינו חופף לחלוטין לתיעוד היסטורי. אירוע היסטורי זהה עשוי למלא תפקיד שונה לחלוטין בזיכרון הקולקטיבי של קבוצות חברתיות שונות (קחו, למשל, את פצצת האטום על הירושימה במלחמת העולם השנייה: משמעות האירוע הזה בזיכרון הקולקטיבי של הציבור היפני שונה ממשמעותו בזיכרון הקולקטיבי האמריקאי).[66] היכולת של קבוצות שונות ליצור את הזיכרון הקולקטיבי שלהן תלויה לכן בזמינות של ריבוי קולות ומשמעויות.[67]

מתבקש להבהיר כאן כי ריבוי אינו רק שיקוף סטטי של העדפות משתמשים. חשיפה למגוון ולריבוי ממלאת תפקיד בעיצוב העדפות כאלה. כך, מחקרים מראים כי החשיפה לתכנים נוסחתיים ואחידים מגדילה את הביקוש לתכנים כאלה ומקטינה את הנכונות

62 ראו Michal Shur-Ofry, *Copyright, Complexity and Cultural Diversity: A Skeptic's View*, in TRANSNATIONAL CULTURE IN THE INTERNET AGE, 203, 205–207 (Sean Pager & Adam Candeub eds. 2012).

63 ראו הספרות בה״ש 57 לעיל וכן הברמאס, לעיל ה״ש 59, בעמ׳ 21; Seyla Benhabib, *Models of Public Space: Hannah Arendt, the Liberal Tradition and Jurgen Habermas*, *in* HABERMAS AND THE PUBLIC SPHERE 73, 82–83, 86 (Craig Calhouny ed. 1992); בלקין, לעיל ה״ש 60, בעמ׳ 41.

64 לדיון על מושג הזיכרון הקולקטיבי ויחסיו עם המשפט, ראו: Guy Pessach & Michal Shur-Ofry, *Intangibles and Collective Memory: The Role (and Rule) of Law*, 25 JERUSALEM REV. LEG. STUD. 227 (2022). לדיון על חשיבות הזיכרון הקולקטיבי לקבוצות ויחידים ראו, למשל, Jeffrey K. Olick, *Collective Memory: The Two Cultures*, 17 SOC. THEORY 333 (1999).

65 שם.

66 ראו, למשל, Jeffrey K. Olick & Joyce Robbins, *Social Memory Studies: From "Collective Memory" to the Historical Sociology of Mnemonic Practices*, 24 ANN. REV. SOC. 105, 110 (1998); Amos Funkenstein, Collective Memory and Historical Consciousness, 1 HIST. & MEMORY 5 (1989). למשמעויות של פצצת האטום בזיכרון הקולקטיבי של קבוצות שונות, ראו Stefanie Fishel, *Remembering Nukes: Collective Memories and Countering State History*, 1 CRITICAL MIL. STUD. 131, 136–137, 141 (2015).

67 שם.

לצרוך תכנים מגוונים יותר, ומנגד, קיומו של היצע תכנים מגוון משפיע לטובה על הביקוש למגוון וליצירות נישה, ומצמיח "זנב ארוך" לעקומת הביקוש.[68]

על רקע הספרות הזאת, מתבהר הסיכון בכך שהעידן של מודלי השפה הגדולים יוביל להפחתת המגוון ולצמצום בתפיסות העולם. מודלי השפה משולבים כבר במוצרי צריכה שונים, וכן בארגז הכלים הטכנולוגי הסטנדרטי שלנו הכולל מנועי חיפוש ומעבדי תמלילים. סביר להניח שמודלים אלה, בין כשלעצמם ובין בשילוב עם מוצרים אחרים, יהפכו למקור דומיננטי, ואולי אף למקור העיקרי, שדרכו יקבלו אנשים מידע על העולם. לכן, ולאור המאפיינים הטכנולוגיים שתוארו קודם לכן, הנטייה של אותם מודלים למיינסטרים ולפופולרי צפויה לחלחל ולהשפיע גם על התפיסות האנושיות: החל מהחשיבות שאנו מייחסים לנראטיבים היסטוריים, דרך המוצרים התרבותיים שאנחנו בוחרים לצרוך או הבחירות הקולינריות שלנו, וכלה בתמונת העולם שלנו בנושאים שונים.

חשוב להבהיר עוד: הנזק שמתואר כאן הוא נזק מצטבר וסיסטמי. ברמת הפרט, השפעת המודלים צפויה להיות חמקמקה וכמעט בלתי נראית. לא סביר שמשתמש המתמקד במשימה מסוימת, כמו חיפוש מידע היסטורי, או המלצה על סדרת טלוויזיה, יבחין בה, וגם אם כן, קשה להעלות על הדעת שאדם יוכל להראות נזק בר־תביעה אם קיבל ממודל שפה המלצה על סדרה כמו "הסופרנוס" או מידע על לינקולן בתשובה לשאלה על אישיות חשובה במאה ה־19. עם זאת, בטווח הארוך ובאופן מצטבר, לפריזמה הצרה והמרוכזת של המודלים הגדולים צפוי להיות אפקט מצרפי שיצמצם גם את ראיית העולם של אנשים, ובאופן רחב יותר ישפיע על החברה האנושית כולה.

## 4.2. ההקשר הישראלי

האמור עד כה נכון באופן כללי ובכל חברה אנושית. ואולם **מהזווית המקומית-ישראלית**, נדמה שיש לדברים חשיבות מיוחדת. מבחינת מודלי השפה הבסיסיים הגדולים, התרבות הישראלית, על כל גווניה, היא תרבות "נישה", ולא "מיינסטרים". **ראשית**, עיקר הפעילות התרבותית בישראל מתנהלת בשפות שאינן אנגלית: בפרט עברית, וכן ערבית, רוסית, אמהרית, יידיש ועוד.[69] **שנית**, הגם שהתרבות המקומית אינה ענייה בחומרים כתובים, מתועדים, ואף כאלה שעברו דיגיטיזציה, הרי שבשל גודל המדינה ואוכלוסייתה מדובר בתוצרים שהכמות שלהם קטנה יחסית למדינות אחרות. על כן, גם אם התוצרים האלה ייכללו בחומרי האימון של המודלים הגדולים, חלקם היחסי יהיה קטן. מכיוון שהעיקרון הטכנולוגי המארגן שביסוד המודלים הוא שכיחות

68 למקצת הספרות בנושא ראו, למשל, בייקר, לעיל ה"ש 38, בעמ' 87–92; תאודור ו. אדורנו ומקס הורקהיימר "תעשיית תרבות: נאורות כהונאת המונים" **אסכולת פרנקפורט: מבחר** 158, 172–178 (דוד ארן מתרגם 1993). השוו CHRIS ANDERSON, THE LONG TAIL: WHY THE FUTURE OF BUSINESS IS SELLING LESS OF MORE (2006). לדיון מפורט יותר בסוגיה ראו שור־עופרי מגוון תרבותי, לעיל ה"ש 32.

69 ראו, למשל, אליעזר בן רפאל "רב־תרבותיות ורב־לשוניות בישראל" **מדברים עברית** יח 67 (2002).

סטטיסטית, משקלם של תוצרי התרבות המקומית בפלט ברירת המחדל של המודלים עלול להיות קטן אפילו יותר ממשקלם בחומרי האימון, והדברים, כמובן, אמורים בדרך קל וחומר בנוגע לתרבויות מיעוטים בישראל. ברוב המקרים יש להניח שמודלים של שפה יבחרו בדמויות, נראטיבים, אירועים ויצירות תרבות פופולריים, מהעולם המערבי האנגלו-אמריקאי. הדוגמאות שתוארו קודם לכן ממחישות זאת: כך, ChatGPT מיקד את התשובות שלו בנושא אישים בולטים מהמאה ה־19 בעולם האנגלו-אמריקאי. לעומת זאת, כאשר הפנינו (במסגרת מחקר אחר) שאלה דומה למשתמשי הטוויטר (X) קיבלנו בין התשובות גם שמות מנהיגים מאסיה, וכן את תאודור הרצל – מנהיג שיש לו חשיבות מיוחדת בהקשר הישראלי, אבל לא הופיע, ואין לצפות שיופיע, בבירורות המחדל של הבינה המלאכותית.[70]

חשוב להעיר, בהקשר זה, ששימור התרבות הישראלית, על נכסיה המוחשיים והלא־מוחשיים ועל התרבויות של הקבוצות המרכיבות אותה, הוכר כיעד לאומי בשורת דברי חקיקה והחלטות שאומצו על ידי המדינה לאורך השנים.[71] כפי שציינה עדנה הראל פישר במחקרה, במדינת "כור-היתוך" כמו ישראל יש חשיבות מיוחדת גם בשמירה והכרה בתרבויות של קבוצות שונות, כאמצעי לעידוד סובלנות, קידום תחושות שייכות והעמקת האמון.[72] הקושי שבצמצום המגוון בעידן הבינה המלאכותית צריך להטריד אותנו, אפוא, הן כחלק מהחברה האנושית והן כחלק מתרבות "נישתית" שעלולה, יחד עם התרבויות של הקבוצות המרכיבות אותה, להידחק לשולי השוליים של עולם המחשבות העולות-על-הדעת של המודלים.

***

לסיום פרק זה, ובטרם אפנה לבחון איך יכולה מדיניות של בינה מלאכותית להגיב לאתגר שתואר כאן, ראוי לעיין בשתי שאלות נוספות. **ראשית**, האם התאמה אישית (personalization), כלומר מתן אפשרות למשתמשים להתאים אישית את הפלט של מודלי השפה הגדולים להעדפות ולטעמים שלהם, יכולה לספק פתרון לקושי של צמצום המגוון? מודלים עדכניים מאפשרים כיום למשתמשים בהם להתאים את תוצרי המודל להעדפות האישיות שלהם,[73] אולם ספק אם התאמה אישית כזאת יכולה לקדם

70 שור־עופרי ואח', לעיל ה"ש 25.

71 ראו הראל פישר, לעיל ה"ש 59, בעמ' 106 ובפרט בהערות 187–188. הכותבת מונה שורה של חוקים המעניקים חשיבות להנצחת המורשת והתרבות של מגוון התרבויות המרכיבות את החברה הישראלית (כמו למשל יידיש, תרבות יהדות לוב, תרבות לדינו) וכן לקידום תרבויות מיעוטים. ראו גם החלטת הממשלה בעניין "'ציוני דרך' – התכנית להעצמת תשתיות המורשת הלאומית" (21.2.2010): תוכנית מורשת לאומית שתשקם, תעצים ותשמר את תשתיות המורשת הלאומית, ובכלל זה גם "נכסים שאינם מוחשיים", כמו זמר עברי, מחול, עיצוב ועוד.

72 הראל פישר, לעיל ה"ש 59, בעמ' 43–48.

73 OpenAI, *How Should AI Systems Behave, and Who Should Decide?* (Feb. 16, 2023), https://openai.com/blog/how-should-ai-systems-behave (We believe that AI should

מגוון וריבוי אמיתיים. גם אם המודלים יהיו מכוילים כך שהפלט שלהם יתאים יותר לטעמים, להשקפות ולהעדפות של המשתמש הספציפי, כל אחד מהמשתמשים ייחשף עדיין לתשובה אחת מסונתזת בתגובה לפרומפטים שיציג. אם לחזור לאחת הדוגמאות למעלה, הצגת סדרת טלוויזיה סקנדינבית למשתמשת בתגובה לשאלה שלה בדבר סדרת הטלוויזיה הטובה ביותר עשויה להתאים יותר להעדפותיה האישיות, אך עדיין תמסך את המגוון העצום של תשובות אפשריות אחרות לשאלה זו. יותר מכך, סביר שהתאמה אישית תהדהד העדפות קיימות, ובכך היא עלולה דווקא להפוך את המשתמשים לפחות פתוחים לאפשרויות אחרות.[74] כפי שטען לאחרונה ג׳ונתן ג׳ינגריך בהקשר של מערכות המלצה, התאמה אישית של תוכן עשויה להציע ״מגוון שטחי״, אך לא סביר שתשקף ״מגוון עמוק״, כזה שיכול לאתגר תפיסות של משתמשים.[75] עיקר האתגר, במילים אחרות, הוא לא להפוך את **המודל** לפתוח יותר לנקודות מבט שונות, אלא לשמור על המודעות והפתיחות של **המשתמשים** לנקודות מבט כאלה.

**שנית**, האם יש לצפות לפתרון ״שוקי״ לקשיים שהועלו כאן? במילים אחרות, האם סביר ששוק המודלים הגדולים יסדיר את עצמו במשך הזמן ויספק לצרכנים מודלים שהפלט שלהם מגוון יותר, באופן שייתר התערבות רגולטורית? אכן, שוק הבינה המלאכותית נמצא בשלבי התגבשות והתהוות, וקשה לצפות במדויק את התפתחותו. אולם, בהינתן הניתוח שלמעלה ניתן להעריך, בזהירות, כי הסיכוי שהשוק ינוע ״מעצמו״ אל עבר ריבוי ומגוון בפלט ברירת המחדל של מודלי השפה אינו גדול: ההטיה אל עבר המיינסטרים נובעת ממאפיינים בסיסיים של הטכנולוגיה, כך שהגברת המגוון טעונה מאמץ מודע ומכוון מצד ספקי המודלים. לצד זאת, הטכנולוגיה צפויה גם להשפיע (וכנראה כבר משפיעה) על טעמם ועולמם של המשתמשים, באופן שיגביר עוד יותר את הביקוש הקיים ממילא לתוכן פופולרי ואחיד, על חשבון הביקוש למגוון של תכנים.[76] לאור זאת, ספק רב אם ליצרני הבינה המלאכותית יהיה תמריץ מספיק לנקוט צעדים מגבירי-מגוון.[77]

כיצד, אם כן, יכולה מדיניות של בינה מלאכותית להגיב לאתגרים של צמצום המגוון? הפרק הבא פונה לבחון שאלה זו, ומציע כי שילוב העיקרון של ״**ריבוי**״ כחלק מעקרונות המדיניות הבסיסיים בתחום הבינה המלאכותית יסייע בהתמודדות עם קשיים אלה.

---

be a useful tool for individual people, and thus customizable by each user...); Alireza Salemi et al., LaMP: When Large Language Models Meet Personalization, 1 PROC. 62ND ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 7370 (2024).

74 ראו Erik Hermann, *Artificial Intelligence and Mass Personalization of Communication Content – an Ethical and Literacy Perspective*, 24 NEW MEDIA & SOC'Y 1258 (2021); ראו גם את הדיון בתאי תהודה של מדיה חברתית והסיכון להקצנה הכרוך בהיעדר חשיפה לדעות מגוונות, לעיל ה״ש 32 והטקסט הנלווה.

75 Gingerich, לעיל ה״ש 41, בעמ׳ 271–272.

76 ראו הדיון סביב ה״ש 57 והטקסט הנלווה.

77 לניתוח מפורט וכללי יותר שלפיו קידום אמיתי של מגוון אינו מאפשר הסתמכות על כוחות השוק בלבד ראו שור־עופרי מגוון תרבותי, לעיל ה״ש 32, והמחקרים הנדונים שם.

## ב. עיקרון של ריבוי כחלק ממדיניות בינה מלאכותית

ההתפתחויות המהירות בתחום הבינה המלאכותית בתקופה האחרונה הובילו לדיון אינטנסיבי באסדרה של התחום; שותפים לו חוקרים, גורמים בתעשייה וקובעי מדיניות. מדינות שונות מציעות כעת בהליכים שמטרתם לגבש עקרונות מדיניות מרכזיים בתחום הבינה המלאכותית ולקבוע את הדרכים הרגולטוריות ליישומם. בפסקאות הבאות אסקור בתמצית כמה הצעות כאלה בשיטות משפט מרכזיות, לצד המדיניות המתגבשת בישראל במקביל לכתיבת דברים אלה, במטרה לבחון אם יש בהן מענה לאתגר החברתי שמאמר זה מתרכז בו.

### 1. עקרונות המדיניות המתגבשים

#### 1.1 ארה"ב ואירופה

ככלל, המגמה המסתמנת בקרב שיטות משפט מרכזיות היא לבסס את המסגרות הרגולטוריות בתחום הבינה המלאכותית על כמה עקרונות ליבה שאמורים להנחות את הפיתוח, השימוש וההפצה של מערכות כאלה.[78] כך, למשל, בארצות הברית פורסמו בשנה האחרונה כמה מתווים בנושא מדיניות בינה מלאכותית מטעם גורמי ממשל, ובכלל זה צו נשיאותי מנובמבר 2023 שכולל הנחיות לגורמים פדרליים.[79] מבלי להיכנס לדקויות ההבדלים ביניהם (שהם חשובים כשלעצמם אך אינם חיוניים לדיון כאן) מתווים אלה מציעים לבסס את מדיניות הבינה המלאכותית על עקרונות כמו אבטחת מידע, מניעת אפליה, פרטיות, שקיפות והסברתיות, וכן מעורבות אנושית בהחלטות אלגוריתמיות. גישה דומה ננקטה בבריטניה, שם הציע המשרד לבינה מלאכותית בנייר עמדה שפורסם ב־2023 כי הרגולציה בתחום תתבסס על עמידה בכמה סטנדרטים מרכזיים הכוללים מהימנות, בטיחות, אבטחת מידע ועמידות, אחריותיות ושקיפות,

78 ראו גם הסקירה ההשוואתית במסמך העקרונות, לעיל ה"ש 3.

79 ראו Exec. Order No. 14, 110, 88 Fed. Reg. 75191 (Oct. 30, 2023). לטיוטת המתווה שפורסמה ב־2022 ראו *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*, THE WHITE HOUSE (2022); ראו גם את "התוכנית לניהול סיכוני בינה מלאכותית" של משרד הסחר האמריקאי – המכון הלאומי לתקינה וטכנולוגיה (NIST) Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Management Framework* (AI RMF 1.0) (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, שמונה סטנדרטים דומים אשר יישומי בינה מלאכותית מצופים לעמוד בהם. יצוין, כי הצו הנשיאותי מאוקטובר 2023 בוטל על ידי הנשיא טראמפ בחודש ינואר 2025, בעת שמאמר זה נמצא בשלבי פרסום סופיים.

הסברתיות, פרטיות, הוגנות וטיפול בהטיות מזיקות.[80] כעולה מהעיון במסמכים אלה, עקרונות היסוד שצפויים להוות בסיס לאסדרת תחום הבינה המלאכותית, הן בארה״ב והן בבריטניה, אינם כוללים ריבוי או מגוון. בדומה, אמנה בינלאומית חדשה בנושא בינה מלאכותית, זכויות אדם, דמוקרטיה ושלטון החוק, שאומצה בחודש ספטמבר 2024 ביוזמת הנציבות האירופית, מטילה על המדינות החברות להטמיע במדיניות הבינה המלאכותית שלהן שורת עקרונות כמו שמירה על כבוד האדם, שקיפות, אחריותיות, שוויון, פרטיות ועוד. גם אמנה זו אינה כוללת התייחסות למגוון.[81]

ואולם, ההסדרה המפורטת ביותר עד כה של תחום הבינה המלאכותית היא **חוק הבינה המלאכותית של האיחוד האירופי**. החוק נדון בפרלמנט האירופי חודשים ארוכים ואושר לבסוף בחודש יוני 2024.[82] ככלל, החוק מבחין בין מערכות בינה מלאכותית על פי תחומי פעילות ורמת הסיכון הכרוכה בהם. מערכות שהוגדרו ״בסיכון גבוה״ הן כאלה שעלולות ליצור סיכון משמעותי לפגיעה בבריאות, בבטיחות, בזכויות יסוד, בסביבה ועוד.[83] אלה הוכפפו לסטנדרטים מחמירים יותר המטילים חובות בנוגע לניהול מידע, שקיפות, שמירת תיעוד, אבטחה ופיקוח אנושי.[84] סטנדרטים מנדטוריים אלה אינם כוללים התייחסות לריבוי או למגוון.

יחד עם זאת, לצד חובות אלה, המבוא לחלק האופרטיבי של החוק האירופי מאזכר את הנחיות האתיקה לבינה מלאכותית שגובשו בשנת 2019 על ידי מומחים מטעם הנציבות האירופית. בהנחיות הללו הוגדרו שבעה עקרונות אתיים, לא מחייבים, אשר מטרתם להבטיח שהפיתוח והשימוש במערכות הבינה המלאכותית יעלו בקנה אחד עם

---

80 Dep’t for Sci., Innovation & Tech., A Pro-Innovation Approach to AI Regulation (Mar. 29, 2023), https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.
ראו גם מסמך ממשלתי מאוחר יותר שמזכיר בהסכמה אותם עקרונות: Dep’t for Sci., Innovation & Tech., *Consultation Outcome: A Pro-Innovation Approach to AI Regulation: Government Response* (Feb. 6, 2024), https://www.nevo.co.il/shorturl/921D2653-6B7.

81 Eur. Consult. Ass., *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law* Doc. No. 225 (2024).

82 Regulation (Eu) 2024/1689 of The European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (להלן: חוק הבינה המלאכותית האירופי).

83 חוק הבינה המלאכותית האירופי, שם, בס׳ 6 ו־7.

84 שם, בס׳ 8–15. לביקורת על ההבחנה בין רמות סיכון ברגולציה של בינה מלאכותית ראו Michal Shur-Ofry, *A Networks-of Networks Approach to AI Regulation*, NETWORK L. REV. (2024).

הגישה האירופאית של "האדם במרכז".[85] לצד עקרונות מוכרים כמו בטיחות טכנית, פיקוח אנושי, פרטיות, שקיפות והסברתיות, העקרונות האתיים שחוק הבינה המלאכותית מפנה אליהם מתייחסים במפורש גם למגוון תרבותי, וקובעים כי "מערכות בינה מלאכותית יפותחו, והשימוש בהן ייעשה, באופן שיכלול שחקנים מגוונים ויקדם גישה שווה, שוויון מגדרי **ומגוון תרבותי**...".[86] החוק האירופי מבהיר כי אין מדובר בכללים מחייבים, אולם ממליץ ליישם את העקרונות בתכנון מערכות בינה מלאכותית ולשלב אותם בקודים התנהגותיים וולונטריים שיפותחו על ידי השחקנים בתחום.[87]

הרגולציה באיחוד האירופי נמנעת, אפוא, מהטלת חובות "נוקשות" על ספקי בינה מלאכותית ביחס למגוון, ומסתפקת במקום זאת בעקרונות לא מחייבים ובקודים וולונטריים. אומנם מוקדם להעריך את האפקט שיהיה למנגנונים אלה בפועל, אולם לכל הפחות הם משקפים הכרה מפורשת בהשפעות אפשריות של בינה מלאכותית על נושא המגוון, ושמים את הסוגיה "על השולחן" של קובעי המדיניות ושחקני התעשייה בתחום.

## 1.2 ישראל

המסמך המקיף ביותר בנושא המדיניות הכללית בתחום הבינה המלאכותית בישראל הוא "מסמך עקרונות מדיניות, רגולציה ואתיקה בתחום הבינה המלאכותית", שפורסם על ידי משרד החדשנות, המדע והטכנולוגיה בשיתוף עם משרד המשפטים ומחלקת ייעוץ וחקיקה, ב־17 בדצמבר 2023.[88] בדומה למצב בארצות הברית ובבריטניה, גם המסמך הישראלי, שגובש בעקבות החלטת ממשלה מ-2021, הוא מסמך עקרוני, מקיף ורחב, המבקש "להתוות ולהנחות את אופן הפעולה הממשלתי ביחס לטכנולוגיה במישור הרגולטורי, המשפטי והאתי", באופן שיהווה תשתית לאסדרה ולחשיבה עתידית.[89]

לצד סקירה השוואתית מקיפה, המסמך ממפה את הקשיים העיקריים הכרוכים בשימוש הגובר במערכות בינה מלאכותית ומציע, בשורה התחתונה, לאמץ שישה

85 שם, בס׳ 27 למבוא; למסמך ההנחיות האתיות המקורי ראו Ethics Guidelines for Trustworthy AI, Independent High-Level Expert Group on Artificial Intelligence 19-20 (2019).

86 חוק הבינה המלאכותית האירופי, לעיל ה"ש 82, בס׳ 27 למבוא (ההדגשה הוספה, תרגום שלי).

87 שם. לתיאור הקודים הוולונטריים ראו שם, בס׳ 95(1) ובס׳ 95(2)(a).

88 מסמך העקרונות, לעיל ה"ש 3. יצוין כי ישראל גם חתמה על האמנה הבינלאומית בנושא בינה מלאכותית הנזכרת בה"ש 81 לעיל, ראו "היסטוריה: ישראל חתמה על אמנה עולמית, תקדימית וראשונה מסוגה לאסדרת השימוש בבינה המלאכותית" **משרד החדשנות, המדע והטכנולוגיה** (5.9.2024) https://www.gov.il/he/pages/ai2024.

89 מסמך העקרונות, לעיל ה"ש 3, בעמ׳ 4. להחלטת הממשלה שקדמה למסמך ראו החלטה 212 של הממשלה ה־36 "תכנית לקידום חדשנות, עידוד צמיחת ענף ההייטק וחיזוק המובילות הטכנולוגית והמדעית" (1.8.2021).

עקרונות אתיים שהם בסיס למדיניות בתחום המבוססים על עקרונות שגובשו במסמך של ה־OECD.[90] עקרונות אלה כוללים הכרה בכך שהשימוש בבינה מלאכותית מהימנה הוא אמצעי לקידום "צמיחה, פיתוח בר־קיימא, רווחה חברתית, וקידום המובילות הישראלית בתחום החדשנות"; אימוץ גישת "האדם במרכז", שלפיה הפיתוח והשימוש בבינה מלאכותית ייעשה תוך כיבוד זכויות יסוד ואינטרסים ציבוריים; וכן שמירה על שוויון ומניעת אפליה, שקיפות והסברתיות, אבטחה, בטיחות מידע ואמינות, ואחריותיות.[91] יודגש כי לפי המסמך, העקרונות המוצעים הם עקרונות **אתיים** שאינם מחייבים ואינם בעלי מעמד משפטי.[92] לצד עקרונות אלה המסמך מתייחס לרגולציה בתחום הבינה המלאכותית ונוקט גישה שלפיה אסדרת התחום תהיה באמצעות "רגולציה ענפית" ולא "רגולציה רוחבית־כללית", כלומר תתבסס על צרכים קונקרטיים בתחומים מסוימים (למשל רפואה, חינוך או התחום הפיננסי) ותטופל על ידי הרגולטורים האמונים על אותו תחום.[93]

מהי עמדת מסמך העקרונות הישראלי בסוגיות של ריבוי ומגוון? בדומה לחלק ממסמכי המדיניות שסקרתי קודם לכן, המסמך אינו מתמקד ואינו עוסק ישירות בנושאים הללו. העקרונות האתיים המוצעים בו אינם כוללים ריבוי או מגוון תרבותי. המסמך דן אומנם בצורך לגוון את מאגרי המידע המשמשים לאימון בינה מלאכותית, אך זאת רק כאמצעי לצמצום בעיות של אפליה וחוסר שוויון בהחלטות הבינה המלאכותית.[94] מעבר לכללים האתיים, התפיסה הרגולטורית שהמסמך מציע מתמקדת, כאמור, ברגולציה ענפית של תחומים שנתפסים כמסוכנים יותר, ולפי המסמך אינה אמורה לחול על יישומי בינה מלאכותית באשר הם. המשמעות היא ש"מודלים לכל מטרה" (דוגמת ChatGPT או Gemini) ככל הנראה לא יוכפפו לדרישות רגולטוריות מחייבות. כחלק מהצדקת התפיסה הענפית, מנסחי המסמך אף מציינים במפורש כי "יש הבדל בין שימוש בבינה מלאכותית כדי להציע לצרכן סרט או שיר לטעמו לבין שימוש בה לאבחון רפואי או החלטת אשראי" מבחינת רמת הסיכון.[95] התפיסה המשתמעת היא, שהשפעת הבינה המלאכותית בתחום התרבותי אינה מצריכה התערבות רגולטורית.

על רקע האמור, האם ראוי שישראל (לצד מדינות נוספות) תעגן במפורש ריבוי ומגוון כחלק מעקרונות המדיניות הבסיסיים שיחולו בתחום הבינה המלאכותית? בפרק הבא אציע כי התשובה היא חיובית.

90 מסמך העקרונות, לעיל ה"ש 3, בעמ' 14–15, 102–103.

91 שם.

92 שם, בעמ' 14.

93 שם, בעמ' 92.

94 שם, בעמ' 53, 81–82. כפי שיובהר מייד, אין באמצעי כזה כדי לפתור את האתגרים שמאמר זה דן בהם.

95 שם, בעמ' 84 ו־95.

## 2. הצעה: ריבוי כעיקרון אתי־רגולטורי במדיניות בינה מלאכותית

האם אפשר להתמודד עם האתגרים של אובדן הריבוי והמגוון בעידן של בינה מלאכותית יוצרת באמצעות הישענות על העקרונות שכבר מוכרים בתחום הבינה המלאכותית, ובפרט הסברתיות, שקיפות או הפחתת הטיות? עמדתי היא שהעקרונות הקיימים, אף שהם נותנים מענה חשוב לאתגרים שונים הקשורים בבינה המלאכותית, לא יספיקו כדי להתמודד עם הסיכון הסיסטמי של הפחתת המגוון וצמצום עולם המחשבות העולות-על-הדעת.

כך, אחד העקרונות הראשונים שהוצעו בהקשר של בינה מלאכותית הוא עקרון ההסברתיות, שנועד להתמודד עם מאפייני ה״קופסה השחורה״ על ידי הטלת חובות של גילוי ומתן הסברים לגבי הטכנולוגיה ותוצריה.[96] הטלת חובה לקדם שקיפות והסברתיות יכולה לסייע בזיהוי הטיות בתהליכי קבלת החלטות אלגוריתמיים, עניין שהוא בעל חשיבות רבה כאשר מדובר בהחלטות הנוגעות לזכויות הפרט, כמו למשל ניקוד לצורך הקצאת אשראי, או קביעת פרופיל סיכון. בהקשר של בינה מלאכותית יוצרת, מתן הסברים יכול לעורר מודעות בקרב המשתמשים לרכיבי השיפוט האנושי המוטמעים בטכנולוגיה, ולהפחית במידת מה את התפיסה של מודלי השפה הגדולים כ״אורקל״ יודע-כול. בדומה, חשיפת מידע על עקרונות התכנון הכלליים של המערכת, ובייחוד על מאגרי הנתונים ששימשו לאימון הבינה המלאכותית, יכולה לתת לנו מושג לגבי ה״יקום״ שממנו שואבים המודלים הגדולים את ״עולם המחשבות״ שלהם. עם זאת, כאשר המטרה היא לשמר ריבוי של נראטיבים, אפשרויות ותפיסות, ההסברתיות אינה מספקת. קשה להניח שמשתמש שקיבל תשובה הגיונית לשאלה לגבי אירוע היסטורי או המלצה על מוצר תרבותי יטרח לחפש הסברים ומידע בנוגע ל״אחורי הקלעים״ של המערכת, ובכל מקרה, כאמור, היכולת להסביר פלט ספציפי של בינה מלאכותית יוצרת לרוב אינה קיימת.[97]

גם עקרונות הנוגעים למניעת הטיות או אפליה, כמו אלה המוצעים במסמך העקרונות הישראלי, אינם צפויים לתת מענה טוב לחששות מפני צמצום של ריבוי ומגוון. ראשית, ההתייחסות לכל תוכן מיינסטרימי שמיוצר על ידי בינה מלאכותית (נניח, הצעה של ״הסופרנוס״ כהמלצה על סדרת טלוויזיה) כאל ״הטיה״ אינה מעשית ואינה מוצדקת. באופן כללי יותר, ההוראות שעניינן מניעת הטיות בבינה מלאכותית מתמקדות ככלל במניעת אפליה של **פרטים**.[98] לעומת זאת, הדיון כאן מתמקד, כזכור, בסיכונים **סיסטמיים**. לא מדובר בהחלטות לא־הוגנות שיש להן השפעה מיידית על אנשים, אלא בהשפעה אינקרמנטלית-אבל-מצטברת על מספר עצום של אנשים, שלאורך זמן עלולה להניב השלכות חברתיות לא רצויות.[99] גיוון חומרי האימון באופן

96 למאפייני ה״קופסה השחורה״ של המודלים ראו למשל מסמך העקרונות, לעיל ה״ש 3, בסעיף 4.3.1, והדיון בה״ש 50–51 והטקסט הנלווה.

97 שם.

98 ראו הדוגמאות המובאות במסמך העקרונות, לעיל ה״ש 3, בסעיף 4.1.1.

99 ראו הדיון לעיל בה״ש 5–7 והטקסט הנלווה.

שייצגו את האוכלוסייה באופן מהימן, כפי שמוצע במסמך העקרונות הישראלי, עשוי לסייע במיתון אפליה בהחלטות אלגוריתמיות. ואולם, לאור הבסיס הטכנולוגי של המודלים, כמתואר לעיל, תמונת העולם שתשתקף מהם בסוגיות היסטוריות, תרבותיות וכדומה צפויה להישאר ממורכזת ולא מגוונת. באופן דומה, הקפדה על עקרונות של אבטחת מידע תפחית את הסיכונים למניפולציה של תוצרי הבינה המלאכותית ותגביר את מהימנותם, אבל לא תפחית את הסיכון שהסתמכות על תוצרים כאלה (מהימנים ככל שיהיו) תגרום לאורך זמן להפחתת המגוון התרבותי, לפגיעה בזיכרון הקולקטיבי או לצמצום "עולם המחשבות" האנושי באופן כללי. חששות אלה מחייבים התמודדות ישירה יותר.

על רקע ניתוח זה, מתבהר הצורך בשילוב עיקרון של ריבוי כחלק מעקרונות היסוד בתחום הבינה המלאכותית. כאמור קודם לכן, **"ריבוי" משמעו חשיפה של המשתמשים, או לפחות יידוע שלהם אודות קיומם של תשובות, נראטיבים, אפשרויות ותכנים נוספים ומגוונים**.[100] שילוב העיקרון של ריבוי כחלק ממדיניות בינה מלאכותית יסייע למשתמשים להיות מודעים לקיומם של תפיסות וטעמים שונים, ויעניק להם הצצה לעולם שנמצא מעבר לפלט ברירת המחדל של מודלי השפה הגדולים. יותר מכך, עצם החשיפה לחלופות (למשל, דמויות היסטוריות או מוצרי תרבות אחרים, מעבר לאלה שהמודל בחר כברירת מחדל), ואפילו עצם המודעות לכך שקיימים מקורות נוספים ואפשרויות נוספות, עשויים למתן את ההיקסמות ממודלי השפה, ולהקטין את הנטייה לסמוך אוטומטית על הפלט שלהם. כתוצר לוואי, החשיפה לאפשרויות שונות יכולה, בעקיפין, גם למתן הטיות בקרב משתמשים: לא על ידי הצגת מציאות "אובייקטיבית" או "מייצגת", אלא על ידי העלאת מודעות של המשתמש לקיומם של השקפות ונראטיבים שונים ואפשריים. לסיכום, ההכרה בריבוי כעקרון יסוד בבינה מלאכותית, ואימוצו כחלק מעקרונות המדיניות בתחום, יתמודדו ישירות עם הסיכונים הסיסטמיים של פגיעה במגוון ושל צמצום עולם המחשבות העולות-על-הדעת בעידן הבינה המלאכותית היוצרת.

לפני שנמשיך אתייחס, בתמצית, להתנגדות אפשרית. האם ההכרה בריבוי כחלק מהעקרונות הבסיסיים של מדיניות בינה מלאכותית, כפי המוצע כאן, עולה בקנה אחד עם עקרון חופש הביטוי? מחקרים מהעת האחרונה הציעו שעידן הבינה המלאכותית היוצרת מעורר שאלות חדשות בתחום חופש הביטוי.[101] דיון מקיף בנושא זה מחייב הבחנה בין סוגי דוברים וביטויים, כמו גם התייחסות נרחבת לתאוריות שונות של חופש הביטוי, וחורג ממסגרת הניתוח כאן. עם זאת, דווקא העיקרון של ריבוי אינו מעורר,

100 המונח "ריבוי" דומה במשמעותו למונחים כמו "מגוון" (מונח שמתקשר פעמים רבות לתחום התרבותי) או "פלורליזם" (מונח שמתקשר לתחום החברתי-פוליטי), ועדיף בעיניי על פני שני אלה בהיותו רחב וכוללני יותר.

101 ראו, למשל, Cass R. Sunstein, *Artificial Intelligence and the First Amendment*, SSRN (Apr. 28, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4431251; Peter Henderson et al., *Where's the Liability in Harmful AI Act?*, ARXIV (Aug. 16, 2023), https://doi.org/10.48550/arXiv.2308.04635.

לדעתי, קושי מיוחד בהקשר זה. ראשית, שילוב עיקרון של ריבוי במדיניות בינה מלאכותית הוא ניטרלי מבחינה תוכנית. הרעיון הוא לא לאסור על ביטוי או תשובה מסוימים שמיוצרים על ידי מודלי השפה (ובמשתמע, אולי, על ידי ספקי הבינה המלאכותית). אדרבה, ההפך הוא הנכון: עיקרון של ריבוי יעלה "אל פני השטח" שפע של ביטויים ויקדם שוק מגוון של רעיונות, שהוא דווקא חיוני מנקודת המבט של חופש הביטוי.[102] בנוסף, כפי שציין ג'ק בלקין בנוגע לפלטפורמות המדיה החברתית, ההשתתפות של משתמשים ביצירת תרבות ובתהליכים של יצירת משמעות היא חלק מזכותם לחופש ביטוי, ופן זה של הזכות מצדיק הטלת חובות מסוימות על מתווכי המידע.[103] עמדתי, לכן, היא שההכרה בחובות סבירות שיחולו על מפתחים של מודלי שפה גדולים, לחשוף את המשתמשים או ליידע אותם על קיומם של תכנים, נקודות מבט תרבותיות ונראטיבים מגוונים, עולה בקנה אחד עם עקרונות של חופש הביטוי.[104]

מהי המשמעות הקונקרטית של הכרה בריבוי כעקרון מדיניות בתחום הבינה המלאכותית? הפרק הבא, והאחרון, פונה לבחון שאלה זו ומתווה קווים ליישום העיקרון.

## ג. יישום

איך אפשר לשלב את עקרון הריבוי במדיניות של בינה מלאכותית, ואילו צעדים אפשר לנקוט כדי ליישם אותו בתחום המודלים הגדולים?

ראשית, ברמה המשפטית, מוצע לעגן את עקרון הריבוי באופן מפורש כחלק מהעקרונות הרגולטוריים והקודים האתיים בתחום הבינה המלאכותית. הדיון בחלק הקודם מצביע על כך שקובעי מדיניות ברחבי העולם מצויים בתהליך של גיבוש עקרונות להתמודדות עם אתגרים שונים בתחום הבינה המלאכותית, באמצעות מערכות של עקרונות אתיים ורגולטוריים, וחלקם אף מגדירים סטנדרטים שספקי בינה מלאכותית יהיו חייבים לעמוד בהם. חלק ממסמכי המדיניות מציינים בגלוי שהם מונעים לא רק מחשש לפגיעה בזכויות הפרט, אלא גם מהפוטנציאל של בינה מלאכותית לגרום נזק לחברה, או, בטרמינולוגיה של מאמר זה, מהסיכונים הסיסטמיים שכרוכים בה.[105] הניתוח שלנו מראה שעיקרון של ריבוי נדרש כדי לתת מענה לחלק מהסיכונים האלה. לכן, יש להוסיף אותו לרשימת העקרונות האלו. חוק הבינה

102 Sunstein, שם, בעמ' 10–11. להצדקת "שוק הרעיונות" לחופש הביטוי ראו, למשל, Joseph Blochner, *Institution in the Marketplace of Ideas*, 57 DUKE. L.J. 821 (2008).

103 בלקין מעגן זכות זו בתאוריה של cultural democracy: Jack M. Balkin, *Cultural Democracy and the First Amendment*, 110 N. U. L. REV. 1053 (2016).

104 השוו Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. Rev. 1183, 1225 (2016); שור־עופרי ופסח, לעיל ה"ש 37, בעמ' 995–996. כן ראו הדיון בחלק א.4.1.

105 ראו חוק הבינה המלאכותית האירופי, לעיל ה"ש 82, וכן ההתייחסות לאינטרסים ציבוריים במסמך העקרונות, לעיל ה"ש 3.

המלאכותית האירופי מהווה צעד ראשון, אף אם חלקי, בכיוון זה. כאמור, החוק האירופי מפנה לעקרונות אתיים בנושא בינה מלאכותית הכוללים מחויבות לשמירה על מגוון, ומעודד הטמעה של עקרונות אלה בקודים האתיים הוולונטריים שיאומצו בתחום.[106] לפי הערכות, קודים וולונטריים כאלה עשויים להיות מאומצים באופן הדדי ובו-זמני בארה"ב ובאיחוד האירופי.[107]

אבהיר: איני מקיימת כאן דיון מלא בשאלה אם עקרון הריבוי צריך להיות בעל מעמד רגולטורי או להישאר בגדר עיקרון אתי "בלבד" (ברוח ההצעות של מסמך העקרונות הישראלי). התשובה לשאלה זו תלויה במידה רבה בהתפתחויות שיחולו במדיניות הבינה המלאכותית בעולם, כמו גם בשוק הבינה המלאכותית עצמו. עם זאת, כעולה מהדיון למעלה, אין לצפות שהותרת הנושא לכוחות השוק תוליד רמה אופטימלית של ריבוי ומגוון בשוק המודלים, אלא להפך. על כן, גם אם יוחלט לעגן את העיקרון במישור האתי, חשוב שלא להסתפק באמירה בעלמא, אלא יש להבטיח תמריצים שיעודדו את השחקנים השונים בתחום לאמץ וליישם את העיקרון, למשל באמצעות רגולציה המקנה יתרונות משפטיים-עסקיים לחברות שתעמודנה בסטנדרטים הוולונטריים.

יצוין בהקשר זה שכמה שחקנים בתעשיית הבינה המלאכותית הכריזו לאחרונה על מחויבותם לסטנדרטים אתיים מסוימים. בפרט, עקרונות הבינה המלאכותית של גוגל כוללים התייחסות למגוון כשיקול בעת יצירת מאגרי נתונים, בתיוג המידע, ובהערכת בטיחות של מאגרים כאלה.[108] על רקע זה, אימוץ עיקרון של ריבוי כחלק ממדיניות של בינה מלאכותית על ידי רגולטורים יכול לחלחל לתעשייה ולתרום לצמיחתו של סטנדרט **דה־פקטו**, אשר יאומץ על ידי שחקנים בתחום גם בהיעדר חובה מנדטורית.

ברמה הקונקרטית-מעשית, איני מתיימרת להציע כאן מפרט ממצה, אלא לשרטט שלושה כיוונים עיקריים ליישומו של עקרון הריבוי: הראשון עניינו **שילוב עקרון הריבוי בארכיטקטורה** של המודלים הבסיסיים (טכניקה שתכונה כאן "ריבוי מובנה" או multiplicity by design); השני הוא שמירה על **ריבוי ומגוון בשוק המודלים עצמו**, באופן שיאפשר למשתמשים לקבל "חוות דעת שנייה" מצד מודלים נוספים של בינה מלאכותית; והשלישי הוא פיתוח **אוריינות בינה מלאכותית** בקרב המשתמשים.

106 חוק הבינה המלאכותית האירופי, לעיל ה"ש 82, והטקסט הנלווה.

107 ראו, למשל, Natasha Lomas, *EU and US Lawmakers Move to Draft AI Code of Conduct Fast*, TECHCRUNCH (May 31, 2023), https://techcrunch.com/2023/05/31/ai-code-of-conduct-us-eu-ttc; Camille Ford & Carrisa Nietze, *US-EU AI Code of Conduct: First Step Towards Transatlantic Pillar of Global AI Governance?* EURACTIVE (July 27, 2023), https://www.euractiv.com/section/artificial-intelligence/opinion/us-eu-ai-code-of-conduct-first-step-towards-transatlantic-pillar-of-global-ai-governance.

108 ראו GOOGLE, AI PRINCIPLES PROGRESS UPDATE (2023).

## 1. ריבוי מובנה – Multiplicity by Design

אחת הדרכים לשמר ריבוי פרספקטיבות בקרב משתמשי מערכות בינה מלאכותית היא באמצעות שילוב של תכונות-מקדמות-ריבוי בתכנון, בעיצוב ובהנדסה של מערכות אלה. הרעיון של "ריבוי מובנה" שואב מתובנות שהתגבשו בתחום המשפט והטכנולוגיה בשנים האחרונות, ולפיהן ערכים מסוימים שחשובים לחברה יכולים להיות מוטמעים ומקודמים באמצעות הארכיטקטורה של הטכנולוגיה.[109] הדוגמה הבולטת ביותר עד כה היא "privacy by design" (או "עיצוב לפרטיות"), עיקרון שלפיו יש לקחת בחשבון שיקולי פרטיות במהלך הנדסת הטכנולוגיה, באופן שברירות המחדל של עיצוב טכנולוגיות תשקפנה ותקדמנה שיקולי פרטיות.[110] ואכן, עקרון ה"עיצוב לפרטיות" זכה בהכרה נרחבת ואומץ על ידי כמה רגולטורים ברחבי העולם.[111]

בדומה, הארכיטקטורה של בינה מלאכותית יוצרת יכולה להטמיע תכונות המקדמות ריבוי באופן שיפנה את המשתמשים אל תכנים מגוונים, תפיסות עולם מרובות וחלופות שונות, או לפחות יתריע על קיומם. דוגמה אחת, שהוטמעה כבר בארכיטקטורה של ChatGPT, היא "כפתור" מסוג "Regenerate". ההימצאות של כפתור כזה מאותתת למשתמשת שפלט ברירת המחדל הראשוני שהמודל מספק בתגובה לפרומפט שלה הוא לא הפלט האפשרי היחיד, ומאפשרת לה "לזמן" בקלות חלופות נוספות. עוד דוגמה להטמעת "ריבוי מובנה" היא אופן הניסוח של הפלט שנוצר על ידי מודלי השפה. ניסוח טנטטיבי המכיר בצורה מפורשת בספקטרום של תפיסות עולם אפשריות וחלופות מקדם ריבוי יותר מאשר תגובה קצרה והחלטית המציגה למשתמש "רשימה סגורה" של אפשרויות. כך, למשל, תשובה נחרצת לשאלה מיהם שלושת האנשים המשפיעים במאה ה־19 (לדוגמה, "שלושת האנשים החשובים ביותר שחיו במאה ה־19 הם נפוליאון בונפרטה, המלכה ויקטוריה ואברהם לינקולן") מחזקת את הדימוי של המודל כסמכות אולטימטיבית. לעומת זאת, ניסוח פתוח וטנטטיבי מסמן למשתמש שבבחירה כרוך שיקול דעת, ומתריע בפניו שיש מגוון של דעות אפשריות, לפני שהמודל מציע את "דעתו" שלו. תגובה כמו "קשה לקבוע מי היו שלושת האנשים החשובים ביותר במאה ה־19, מכיוון שזה תלוי במידה רבה בפרספקטיבה של האדם, ובקריטריונים שבהם

109 השוו, למשל, Niva Elkin-Koren & Maayan Perel, *Speech Contestation by Design: Democratizing Speech Governance by AI*, 50 FLA. ST. U. L. REV. 611 (2023).

110 מקורה של גישה זו מיוחס לעיתים קרובות לאן קבוקיאן, נציבת המידע והפרטיות מאונטריו, קנדה. ראו ANN CAVOUKIAN, PRIVACY BY DESIGN: THE 7 FOUNDATIONAL PRINCIPLES: IMPLEMENTATION AND MAPPING OF FAIR INFORMATION PRACTICES (2011).

111 ראו, למשל, Ira S. Rubinstein, *Regulating Privacy by Design*, 26 BERKELEY TECH. L.J. 1409, 1410–1411 (2012); להטמעת עיצוב לפרטיות ברגולציה האירופית, ראו: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119).

משתמשים לקביעת החשיבות. כמה דמויות בולטות... כוללות... עם זאת, אחרים עשויים לטעון ש...", עשויה לעודד את המשתמשים להוסיף ולחקור, ואף מזמינה הערכה ביקורתית ורפלקסיבית של הפלט.[112] דוגמה נוספת של "ריבוי מובנה" היא קישור הפלט של המודל לחומרי הגלם שהמודל תופס כרלוונטיים. עיצוב כזה מזמין את המשתמשים לעיין בחומרי הבסיס, לשפוט אותם בעצמם ואולי גם להמשיך לחקור.[113]

מאפיין נוסף, שזמין כעת למשתמשים בחלק ממודלי השפה הגדולים, מאפשר למשתמש לכייל את אחד מההיפר-פרמטרים (העקרונות הטכנולוגיים הרבים שמגדירים את פעולת המודל) המכונה "טמפרטורה".[114] ככלל, טמפרטורה גבוהה יותר מגבירה את האקראיות בתשובות המודל, וסביר שתניב תגובות מגוונות יותר, לעומת טמפרטורה נמוכה.[115] עם זאת, מפתחים של מודלי שפה גדולים ציינו שטמפרטורה גבוהה יותר עלולה לפגום בדיוק של התוכן שנוצר על ידי המודלים.[116] משתמשים שמבצעים כיול הטמפרטורה יצטרכו, אפוא, להיות מודעים למגבלות האלו, ולנהוג זהירות רבה יותר בהסתמכות על הפלט. נקודה זו מתקשרת לסוגיות אוריינות המשתמש בתחום הבינה המלאכותית, ואשוב אליה מייד.

אמצעי נוסף, עקיף, לקידום "ריבוי מובנה" נוגע לצוותים העוסקים בפיתוח המודלים. מכיוון שתוצרי המודלים הם, במידה מסוימת, גם פרי שיקול דעת אנושי, הכשרת צוותי המפתחים תוך יצירת מודעות ביחס לחשיבות הריבוי וביחס ל"הטיית המיינסטרים" של המודלים, לצד גיוון ההרכב האנושי של צוותי הפיתוח עצמם, עשויים לתרום לכך ששיקול הדעת וההחלטות המתקבלות בפיתוח המערכות הללו ייתנו ביטוי טוב יותר לשיקולי מגוון וריבוי.[117]

ולבסוף, צעד שאפתני הרבה יותר שעשוי להוביל להגברת המגוון הוא שינוי בפרדיגמת "חיזוי-המילה-הבאה", המעניקה משקל בלתי נמנע לפופולריות ביצירת הפלט על ידי מודלי שפה גדולים. במחקר משפיע שנערך על ידי חוקרי מיקרוסופט נטען שלפרדיגמה זו יש מגבלות טבועות, המתבטאות ב"פגמים ביכולת התכנון,

112 לדוגמאות נוספות ראו שור־עופרי, לעיל ה"ש 12.

113 כך, לדוגמה, מודל השפה Gemini מציע לפעמים קישורים למקורות כאלה.

114 במודלי השפה הגדולים המונח "טמפרטורה" מתייחס להיפר-פרמטר השולט ברמת האקראיות הבאה לידי ביטוי בפלט. ככלל, כאשר ערך הטמפרטורה נמוך, המודל נוטה לייצר תגובות שמרניות וצפויות יותר. לעומת זאת, כאשר הטמפרטורה גבוהה, המודל מייצר טקסט אקראי ויצירתי יותר. ראו למשל Geoffrey Hinton et al., Distilling the Knowledge in a Neural Network, ARXIV (Mar. 9, 2015), https://arxiv.org/abs/1503.02531.

115 שם. בנוסף, ראו Hugo Touvron et al., *Llama 2: Open Foundation and Fine-Tuned Chat Models*, ARXIV (July 19, 2023), https://arxiv.org/abs/2307.09288.

116 שם, בעמ' 13; OpenAI, *Why the API output is inconsistent even after the temperature is set to 0*, (Aug. 25, 2023), https://community.openai.com/t/why-the-api-output-is-inconsistent-even-after-the-temperature-is-set-to-0/329541.

117 ראו התייחסות לגיוון צוותי הפיתוח למשל בחוק הבינה המלאכותית האירופי, ס' 95(2)(d), ובמדיניות גוגל, לעיל ה"ש 108.

בזיכרון העבודה... וביכולות ההנמקה״ של GPT-4.[118] הניתוח במאמר זה מצביע על כך שמגבלה נוספת של הפרדיגמה הטכנולוגית הרווחת היא תפוקה מיינסטרימית, הנוטה לאחידות. בעת כתיבת הדברים האלה שינוי כזה נשמע מהפכני למדי. ואולם, גם בטרם נמצא תחליף לפרדיגמה הקיימת, נראה שאפשר לנקוט שורה של צעדים נוספים, פשוטים יותר, כדי להטמיע ולקדם ״ריבוי מובנה״ בתחום הבינה המלאכותית. פיתוח ארכיטקטורות כאלה יצריך כמובן מאמץ רב־תחומי משולב מצד קובעי מדיניות, מדעני מחשב, אנשי מדעי חברה ומעצבי הטכנולוגיות של מודלי שפה גדולים. לאור ההשפעה הצפויה (והקיימת) של המודלים הללו על התרבות, הזיכרון הקולקטיבי, סדרי העדיפויות והתפיסות שלנו, מדובר במאמץ ראוי.

## 2. ריבוי ומגוון בקרב מודלים של שפה

דרך נוספת לקדם ריבוי היא באמצעות הבטחת זמינותם של כמה מודלים גדולים, בייחוד מודלים כלליים, באופן שיאפשר למשתמשים לקבל ״חוות דעת שנייה״ מבינה מלאכותית שמקורה שונה. בשל ההבדלים בין המודלים במאגרי המידע שעליהם אומנו, בתהליכי האימון ובאופן הצגת הפלט, סביר להניח שכל אחד מהמקורות הללו ישקף תפיסה שונה, ולו במקצת, של עולם המחשבות העולות-על-הדעת. אם לחזור לדוגמת האישים המשפיעים במאה ה־19: מודל אחד עשוי להציע רשימה הכוללת את לינקולן, המלכה ויקטוריה ודרווין, מודל שני עשוי להציע רשימה הכוללת את לואי פסטר ולא את דרווין, בעוד שמודל שלישי שאומן גם על חומרים מקומיים עשוי להוסיף לרשימה גם דמויות משמעותיות בהקשר המקומי.[119] העובדה שאין חפיפה מלאה בין המודלים היא כשלעצמה מעוררת מודעות לקיומן של מגוון תשובות אפשריות, אף מעבר לאלה שהמשתמשים נחשפו אליהן, ולשיקול הדעת הכרוך בתשובות.[120] כך, מגוון של כלי בינה מלאכותית יסייע גם בהפחתת ה״היקסמות״ וימתן את ״אפקט המולטיוואק״ – התפיסה של מודלי השפה הגדולים כיודעי כול. לבסוף, ובסוגריים, אף שהדיון כאן מתמקד במקרים שבהם אין תשובה נכונה אחת, ״חוות דעת שניות״ של בינה מלאכותית יכולות לסייע באיתור טעויות, אי-דיוקים ומידע כוזב בפלט של מודלי שפה גדולים, במקרים אחרים שבהם קיימת תשובה נכונה וחשוב לזהות אותה.

למרות היתרונות הללו, החזון של היצע מגוון של מודלים גדולים של שפה אינו מובן מאליו. מחקרים רבים מהשנים האחרונות מצביעים על כך שתחום הבינה המלאכותית עצמו הופך מרוכז יותר ומגוון פחות. הצורך בגישה לכמויות אדירות של נתונים, בצירוף כוח המחשוב האדיר הנדרש לפיתוח בינה מלאכותית ללמידה עמוקה, עשויים להשאיר את הזירה בידי קבוצה קטנה של שחקנים – ככל הנראה ענקיות

118 Bubeck et al., לעיל ה״ש 27 (תרגום שלי).

119 ראו הדיון בחלק א.4.2. לעיל.

120 השוו אלקין־קורן ופרל, לעיל ה״ש 109 (הטוענות כי ה״חיכוך״ בין מודלים יוצר מרחב דמוקרטי שמאפשר שיח בין עמדות שונות).

הטכנולוגיה.[121] לאור זאת, חלק מהחוקרים מציעים התערבויות רגולטוריות לשם הפחתת חסמי כניסה ומיתון הריכוזיות בשוק המודלים, החל משימוש בכלי אכיפה מתחום ההגבלים העסקיים, ועד הטלת חובות שיתוף על החברות הגדולות השולטות בנתוני האימון, אשר יאפשרו לגופים אחרים גישה לנתונים אלה.[122]

דיון זה מעלה שאלה בסיסית יותר הנוגעת למעורבותה של המדינה בתחום המתפתח של בינה מלאכותית, לא רק בכובעה כרגולטור-מסדיר אלא גם כ"שחקן" פעיל. נכון למועד כתיבת הדברים, המפתחים הבולטים של מודלי שפה גדולים הם תאגידי ענק הפועלים בהתאם למערכת של תמריצים מבוססי שוק.[123] כאמור לעיל, תמריצים אלו אינם מכוונים את אותם בעלי עניין לתיעדוף סוגיות של ריבוי ומגוון. שילוב דומה של נסיבות שכוללות טכנולוגיות משבשות להפצת תוכן, חסמי כניסה גבוהים, ופוטנציאל השפעה גדול על תפיסות הציבור, גרם בעבר למדינות שונות להיות מעורבות לא רק ברגולציה של שוק התקשורת אלא גם באספקת תכנים בפועל. הדוגמה האולטימטיבית היא השידור הטלוויזיוני: מדינות דמוקרטיות שונות בוחרות להפעיל ערוצי שידור ציבוריים, לצד ערוצי טלוויזיה בבעלות פרטית. ואכן, מחקרים מצביעים על כך שקיומם של ארגוני שידור ציבורי חזקים, ובאופן כללי יותר השקעה משמעותית של המדינה בתוכן, הם גורמים המשפיעים באופן חיובי על רמת המגוון התרבותי.[124] ככל שמדינות תבחרנה לפתח בעצמן מודלי שפה גדולים, הציפייה שמודלים "ציבוריים" כאלה ייתנו עדיפות לשיקולים ציבוריים, כולל שיקולי ריבוי ומגוון, בהשוואה למודלים שיספק השוק הפרטי, היא ריאלית יותר.[125]

---

121 ראו, למשל, Nur Ahmed et al., *The Growing Influence of Industry in AI Research*, 379 SCI. 884, 884–886 (2023); Reza Shokri & Vitaly Shmatikov, *Privacy-Preserving Deep Learning*, PROC. 22ND ACM SIGSAC CONFERENCE ON COMPUT. & COMMC'N SEC. 1310 (2015); Jonas Traub et al., *Agora: Towards an Open Ecosystem for Democratizing Data Science & Artificial Intelligence*, ARXIV (July 19, 2020), https://arxiv.org/abs/1909.03026.

122 ראו, למשל, Viktor Mayer-Schonberger & Thomas Ramge, *A Big Choice for Big Tech: Share Data or Suffer the Consequences*, 97 FOREIGN AFF. 48, 52–54 (2018), וכן המקורות בה"ש 121 לעיל.

123 ראו למשל קולט, לעיל ה"ש 5, בעמ' 18.

124 ראו MICHELLE LAMONT, MONEY, MORALS AND MANNERS: THE CULTURE OF THE FRENCH AND AMERICAN UPPER-MIDDLE CLASS 140–145 (1992); SARAH M. CORSE, NATIONALISM AND LITERATURE: THE POLITICS OF CULTURE IN CANADA AND THE UNITED STATES 58–61 (1997); הראל פישר, לעיל ה"ש 59, בעמ' 63–76 (שמסבירה כי מימון ציבורי של תרבות נדרש לשם הבטחת מגוון תרבותי); שור־עופרי מגוון תרבותי, לעיל ה"ש 32, בעמ' 597.

125 ליוזמה אירופית לפיתוח מודלי שפה שישקפו את עולם הערכים וסדרי העדיפויות האירופאים, ראו Jennifer L. Schenker, *Can Europe Compete on Generative AI?*, THEINNOVATOR (2022), https://theinnovator.news/can-europe-compete-on-generative-ai. מובן, עם זאת, ש"מודלים ציבוריים" עשויים לעורר אתגרים הנוגעים

מובן ששאלות של מרכוז מול ביזור בתעשיית הבינה המלאכותית אינן מוגבלות למודלים גדולים של שפה, וגם לא לשיקולים של ריבוי ומגוון, אלא יש להן השלכות רחבות בהרבה. סקירה מלאה של נושא זה, ומידת המעורבות הנדרשת מצד המדינה, חורגת מתחומי מאמר זה. עם זאת, ההכרה בריבוי כעקרון יסוד במדיניות בינה מלאכותית תורמת לדיון הזה זווית נוספת, ומאירה את העובדה שגיוון שוק הבינה המלאכותית כדי להבטיח (בין השאר) את הזמינות של "חוות דעת שניות" יתרום גם לגיוון עולם המחשבות והתפיסות האנושיות.[126]

## 3. **אוריינות** AI

לבסוף, המאמץ לשמור על ריבוי בעידן של מודלי השפה הגדולים אינו יכול להתמצות בפתרונות רגולטוריים. הניתוח לאורך מאמר זה מבהיר שהאתגרים של הפחתת המגוון וצמצום תפיסות העולם קשורים בטבורם הן לתכונות של הטכנולוגיה החדשה, הן לסביבת השוק שבה פועלות חברות הבינה המלאכותית. יותר מכך, מחקרים רבים מבהירים שמגוון הוא תופעה רב־סיבתית, שאינה ניתנת לפתרון באמצעות התערבות רגולטורית אחת ויחידה.[127] הצורך לשמר ריבוי ומגוון נוכח ההתפתחויות בתחום הבינה המלאכותית מחייב, לכן, לחקור ולקדם צעדים נוספים. צעד משמעותי אחד כזה הוא עידוד **אוריינות AI** בקרב המשתמשים.[128] אוריינות בינה מלאכותית משמעותה, בהקשר שלנו, הבנה בסיסית של העקרונות שלפיהם פועלים מודלי השפה הגדולים, כולל האופן שבו הפלטים שלהם מושפעים משיקול דעת אנושי, מהנתונים שכלולים במאגרי האימון ומפופולריות, לצד הבנה של התכונות שבגינן המודלים הללו עלולים להשפיע באופן מהותי על תפיסות העולם שלנו. הבטחת רמת אוריינות, אפילו בסיסית, בתחום הבינה המלאכותית, תעצים את המשתמשים, תבליט בפניהם את היכולת שלהם להחליט, ואף לחלוק על תשובות המודלים,[129] ועשויה גם לעודד אותם לחפש מידע נוסף. בפראפראזה על אסימוב, אוריינות בינה מלאכותית תעזור לנו להעריך באופן ביקורתי את התגובות שאנו מקבלים ממודלי השפה גדולים, ולהבין שראיית העולם שלהם אינה תמיד "הטובה ביותר שבנמצא", ובוודאי שאיננה "כל מה שנחשב".[130]

לשליטה בהם, בפרט במדינות שאינן דמוקרטיות, וכן שאלות בנוגע להיקף השימוש וההטמעה שלהם. כל אלה ראויים ודאי למחקר נוסף.

126 Mayer-Schonberger & Ramge, לעיל ה"ש 122, בעמ' 54.

127 שור־עופרי מגוון תרבותי, לעיל ה"ש 32, בעמ' 596–599 והמקורות הנזכרים שם.

128 להצעות לקידום אוריינות בינה מלאכותית כדרך לצמצום אתגרים המתעוררים בתחום ראו, למשל, Hermann, לעיל ה"ש 74, בעמ' 13–14.

129 שם, בעמ' 13.

130 אסימוב, לעיל ה"ש 12.

# סיכום

הדיון הציבורי במהפכת הבינה המלאכותית ובתגובה המשפטית הראויה עודנו בראשיתו. ההשלכות מרחיקות הלכת של הטכנולוגיה החדשה על החברה האנושית יוסיפו ויתבררו בשנים הקרובות. מאמר זה ביקש להאיר את אחת ההשפעות האלה, שעד כה נותרה במידה רבה "מתחת לרדאר" של קובעי מדיניות: ההשפעה הצפויה של מודלים גדולים של שפה על מגוון המחשבות והתפיסות האנושיים.

הניתוח במאמר זה מצביע על כך שמודלים של שפה, בייחוד מודלים כלליים, יוצרים פריזמה שתהווה מקור מרכזי לקבלת מידע על העולם, והשפעתה על המשתמשים צפויה להיות גדולה במיוחד. בשל מאפייני הטכנולוגיה, העולם שישתקף אלינו, המשתמשים, מבעד לפריזמה הזאת הוא עולם צר שנוטה לסטנדרטי, לאחיד, ולנפוץ: יככבו בו דמויות ואירועים מרכזיים מההיסטוריה המערבית, מוצרי תרבות פופולרית ממקור אנגלו-אמריקאי, שפות מערביות נפוצות (בראש ובראשונה אנגלית), ויש להניח שגם תפיסות עולם מיינסטרימיות בנושאים רבים אחרים. תמונת עולם זו עתידה להשפיע, באופן אינקרמנטלי, לא מורגש, אך מצטבר ומשמעותי לאורך זמן, על ראיית העולם האנושית, באופן שעלול להניב השלכות סיסטמיות מדאיגות: החל מאובדן המגוון התרבותי, עבור דרך פגיעה במגוון נראטיבים הנחוצים לבניית הזיכרון הקולקטיבי, וכלה בפגיעה בדיאלוג הדמוקרטי, בסובלנות ובסולידריות. הסיכון שכרוך באובדן הריבוי והמגוון משמעותי במיוחד בישראל, לאור ייחודיות השפה והתרבות המקומית, וריבוי התרבויות המרכיבות אותה.

כדי לתת מענה לאתגרים אלה מתבקשת הכרה מפורשת בעיקרון של ריבוי (multiplicity) ושילובו במדיניות הבינה המלאכותית המתגבשת בישראל ובמדינות נוספות. שילוב עקרון הריבוי כחלק מעקרונות המשילות שיחולו בתחום הבינה המלאכותית, לצד פיתוח כלים אתיים, רגולטוריים, טכנולוגיים ואחרים להטמעה שלו, יאפשרו לנו כחברה לשמר מגוון של תכנים, נראטיבים, תפיסות עולם ונקודות מבט אנושיים. כך, נוכל למצות את הפוטנציאל הגדול של הטכנולוגיה מבלי לאבד את הניואנסים והמורכבות של החוויה האנושית.