במבט ראשון
סקירות תמציתיות על התפתחויות טכנולוגיות והשפעותיהן
הכנסת
מרכז המחקר והמידע
יוני 2020, סיון תש"ף

אפליה אלגוריתמית במערכות המבוססות על בינה מלאכותית

האם מערכות ממוחשבות הפועלות באופן אוטונומי, על בסיס אלגוריתמים, עלולות להפלות בני אדם?

לפי הגדרה מילונית, **"אלגוריתם" הוא סדרה של הנחיות או חוקים מתמטיים המשמשים לשם פתרון בעיות, בפרט באמצעות מחשב**. אלגוריתמים ממוחשבים בעלי יכולת למידה משמשים כיום עבור שלל פעולות ייעוץ, מיון, המלצה והתאמה אישית, באינטרנט ומחוצה לו. אלגוריתמים הם מאבני הבניין המרכזיות של מערכות בינה מלאכותית.

אפליה אלגוריתמית היא מצב שבו אלגוריתם (למשל במערכת מיון מועמדים לעבודה) מזהה, במישרין או בעקיפין, מידע ביחס למשתני רקע, כגון מגדר, גזע או נטייה מינית, ומפיק תוצאה בהינתן משתנים אלה. חדירתן הגוברת של מערכות בינה מלאכותית אל חיי היום-יום והשימוש בהן בתהליכי קבלת החלטות, תוך צמצום של מעורבות אנושית עד כדי היעדרה, מעוררים שאלות בדבר ההשלכות של שימושים אלה על אפליה.

ממצאים עיקריים

הטיות בבסיסי הנתונים או בהנחיות המוזנות למערכות בינה מלאכותית, בד בבד עם טעויות מערכת אחרות, עשויות לגרום למערכות כאלה להפלות חלק מן המשתמשים. ככל שהשימוש במערכות אלו יגדל, והשפעתן על קבלת ההחלטות תגדל, כך עלולה ה"אפליה האלגוריתמית" או "בעיית ההוגנות" במערכות אלה להתרחב.

דוגמאות לאפליה או הטיה המבוססות על אלגוריתם:

- ☒ אפליה בהחלטות שיפוטיות: בתי משפט בכמה מדינות בארה"ב משתמשים במערכות המדרגות רמת סיכון או רצידיביזם (פשיעה חוזרת) כדי לתמוך בהחלטות שיפוטיות. נטען כי מערכת נפוצה בתחום היא מערכת מוטה ומפלה על בסיס אתני. לפי מחקר שנערך בנושא, רמת הניבוי של המערכת ושיעור הטעויות בה היו דומים להחלטות שהיו מתקבלות על ידי אנשים נטולי כל הכשרה בתחום.
- ☒ אפליה בהעסקה או בשירות: חברות וגופים שונים משתמשים בתהליכי מיון וסיווג (Profiling) של מועמדים למשרות תוך התבססות על כלי בינה מלאכותית. בנקים וגופים מלווים משתמשים לעיתים במערכות בינה מלאכותית כדי לאמוד את סיכויי ההחזר של לווים שונים. לפי מחקרים שונים, מערכות כאלה נטו לשעתק אפליה של אוכלוסיות מסוימות.
- ☒ אפליית מחירים או תמחור דיפרנציאלי: שירותים ומוצרים שונים, ובהם טיסות וביטוחים, מתומחרים בצורה שונה על פי מידע אישי מפורט, הזמין באמצעות ניטור מידע על הלקוחות. במידע כזה עשויים להיכלל סוג המכשיר שממנו מבוצעת הרכישה, מיקום המשתמש, דפוסי חיפוש ודפוסי רכישה קודמים. בהתאם למידע, שבדרך כלל הלקוח איננו מודע לקיומו או לשימוש בו כמדד לתמחור, לכל לקוח נקבע מחיר שונה בעבור מוצר זהה.
- ☒ הטיה במידע: השימוש באלגוריתמים לזיהוי תחומי עניין, עמדות ומאפיינים אישיים, במסגרת הניסיון של פלטפורמות הרשת הגדולות לספק תוכן "רלוונטי", משפיע על החשיפה של משתמשי הרשת לתכנים מסוימים ולאנשים מסוימים, שעמדותיהם דומות לשלהם. מצב זה עלול להעצים תפיסות ועמדות ולחזק מגמות קיטוב. בשל האמור לעיל, אצל משתמש שהרשת בעבורו היא "תמונת העולם" עשויה להיווצר הטיה בתפיסת המציאות.
- ☒ אפקטיביות פחותה של מערכות בינה מלאכותית בעבור אוכלוסיות מסוימות: מידת ההצלחה של מערכות בינה מלאכותית שונה מאוכלוסייה לאוכלוסייה. לדוגמה, מכון התקנים האמריקאי מצא כי שיעור הטעויות של תוכנות זיהוי

פנים בזיהוי אנשים ממוצא אפריקני או מזרח אסיאתי היה גבוה במידה ניכרת משיעור הטעויות בזיהוי אנשים ממוצא אירופי; במקרים אחרים נתגלו טעויות רבות בניבוי תחלואה בקבוצות אוכלוסייה מסוימות.

הגורמים לאפליה או הטיה במערכות בינה מלאכותית:

- ☒ הקלט המוזן למערכת: טעויות בסוג המידע המוזן למערכות למידת מכונה, המשמש נקודת ייחוס של המערכת לחיזוי, הן מקור נפוץ להטיות בתוצאות או בהמלצות של מערכות כאלה. לדוגמה, הזנת נתונים על גברים בלבד, לצורך חיזוי תחלואה, במערכת שאמורה להפיק תוצאה גם על תחלואת נשים.
- ☒ שימוש במשתני ניבוי מוטים או מפלים: במסגרת תהליך העיצוב של אלגוריתמים יש לקבוע מדדים לתוצאה הרצויה. מדדי ניבוי בעייתיים צפויים לגרום, במתכוון או בטעות, להטיות. לדוגמה, במיון מועמדים – הגדרה של עובד טוב כ"מי שנשאר בעבודה עד מאוחר" צפויה לפגוע בדירוג של נשים. הגדרה של עובד טוב כ"מי שזכה להערכה גבוהה ממעסיק קודם" צפויה לשעתק אפליה מצד מעסיקים קודמים. יצוין כי גם במקרים שבהם אלגוריתמים אינם מוזנים במשתני רקע "רגישים", המערכות מצליחות להסיק אותם ממשתנים אחרים. לדוגמה, להסיק מידע על מצב כלכלי על בסיס כתובת מגורים.

אפליה אנושית או אלגוריתמית: במחקרים שונים נמצא כי יש נטייה אנושית להפלות – נטייה שלעיתים משועתקת או מועצמת במעבר למערכות בינה מלאכותית. גם בדוגמאות שלעיל, הגורם האנושי אחראי לרוב ההטיות של המערכות. אף על פי כן, הנטייה לראות במערכות בינה מלאכותית מכשיר מדעי ואובייקטיבי עלולה להקנות לתוצאותיהן מעמד של אמת שאין בלתה. בניגוד לטענה המקובלת כי מערכות בינה מלאכותית הן כמו "קופסאות שחורות" שאי-אפשר לראות דרכן ולהבין את המתרחש בתוכן, במחקר מקיף בנושא נטען כי דווקא במערכות בינה מלאכותית מתאפשר לזהות, לנטר ולצמצם אפליה ביעילות רבה יותר מאשר בניסיון לזהות אפליה שנעשית על ידי בני אדם.

דרכי התמודדות עם אפליה אלגוריתמית

כדי לצמצם את ממדי האפליה באמצעות אלגוריתמים מוצעים כלים ועקרונות שונים, ובהם שקיפות ואחריותיות בעצם השימוש במערכת בינה מלאכותית; מתן הסברים לרציונל של החלטות שנתקבלו על ידי מכונה; פיתוח כלים טכנולוגיים לשם בדיקת תהליכי קבלת החלטות של מכונות ויצירת סטנדרטים למובנות או הסברתיות (Explainability) של החלטות; הכשרה של מתכנתים בתחומי האתיקה; גיוון אתני ומגדרי של כוח העבודה בתחומים אלה. יצוין כי סעיף 22 ב-GDPR, החקיקה האירופית בנושא הגנה על מידע, מתייחס מפורשות לזכויותיו של אדם בכל הקשור להחלטות המתקבלות בעניינו, בהתבסס על "מערכות החלטה אוטומטיות", כולל למיפוי מידע אישי או ניקוד שלו (Profiling). כמו כן, ארגון ה-OECD פרסם במאי 2019 מסמך המלצות בתחום הבינה המלאכותית, ובהן נכללה גם מניעת אפליה.

מדיניות ממשלתית

בישראל טרם גובשה "אסטרטגיה לאומית" לבינה מלאכותית, להבדיל ממדינות אחרות (ראו סקירה של הפרלמנט האירופי). במסגרת "המיזם הלאומי למערכות נבונות" שהוקם ביולי 2018 פעלו 15 צוותים, שהורכבו ממאות מומחים מכלל המגזרים. הצוותים עסקו בטכנולוגיות שונות, במגזרי משק רלוונטיים ובסוגיות מרכזיות. בנובמבר 2019 פורסם דוח של ועדת משנה של המיזם בנושא אתיקה ורגולציה של בינה מלאכותית. בדוח הומלץ לאמץ עקרונות אתיים לבינה מלאכותית: הוגנות; אחריותיות (כולל שקיפות, הסברתיות וניהול סיכונים); כיבוד זכויות אדם; הגנת סייבר ואבטחת מידע; בטיחות וקיום שוק תחרותי. זאת ועוד, בדוח הוצג כלי לבחינת ההשלכות האתיות של פיתוח מערכת בינה מלאכותית ושימוש בהן. בקרוב יוגש לראש הממשלה דוח שמסכם את המיזם ואת שאלת גיבוש המדיניות הלאומית בתחום הבינה המלאכותית. מוסד שמואל נאמן פרסם דוחות שונים על בינה מלאכותית שנכתבו עבור המועצה הלאומית למחקר ופיתוח אזרחי ובמימונה.