ארכיאולוגיה של ראיית מכונה

קייט קרופורד וטרבור פגלן

המאמץ ללמד מחשבים לזהות תמונות ולפרש אותן נתפס לרוב כעניין טכני טהור. אבל מה אמור מחשב לזהות בתמונה, אילו טעויות עלולות להיות מובנות בפעולתו, ואיזה מידע נשאר שקוף לגמרי מבחינתו? הבנת הפוליטיקה הכרוכה במערכות כאלה חשובה היום יותר מתמיד, משום שהן משתלבות במהירות במבנה של מוסדות החברה ובתהליכי ההחלטה שקובעים את מי לזמן לריאיון עבודה, אילו חשודים לעצור ועוד

ספטמבר 2020

כשמסתכלים על מסד נתונים של תמונות המשמשות לאימון של מערכות בינה מלאכותית, הדבר נדמה בתחילה פשוט למדי. לפנינו אלפי תמונות: תפוחים ותפוזים, ציפורים, כלבים, סוסים, הרים, עננים, בתים ותמרורים. אבל אם נוסיף להתעמק בבסיס הנתונים, יתחילו להופיע בני אדם: מעודדות, צוללנים, רתכים, צופים, הולכים על גחלים, נערות פרחים. עד מהרה הדברים נעשים מוזרים: אישה מחייכת בביקיני בתצלום שמתויג במילים "זונה, פרוצה, מופקרת, יצאנית". צעיר שותה בירה מתויג כ"אלכוהוליסט, שתיין, מבוסם, שיכור". ילד במשקפי שמש מתואר במילים "כישלון, לוזר, אפס, לא יוצלח". זוהי קטגוריית ה"אנשים" במסד הנתונים הקרוי אימג'ֶנֶט (ImageNet), אחת מערכות האימון הפופולריות ביותר ללמידת מכונה.

משהו השתבש כאן. מאין הגיעו התמונות האלה? מדוע תויגו כך האנשים המצולמים? איזו מין פוליטיקה מובלעת בתמונות כאשר מצמידים להן תוויות, ואילו השלכות יש לשימוש בתמונות כאלה המיועדות לאימון של מערכות טכניות? בקיצור - איך הגענו למצב הזה?

♦

"ראיית מכונה" (machine vision), תת-שדה של בינה מלאכותית (AI), היא התחום שעניינו כיצד ללמד מכונות לזהות תמונות ולפרש אותן. האגדה האורבנית מספרת כי בשנת 1966, בימיו הראשונים של התחום, מרווין מינסקי (Minsky) - מרצה צעיר ב-MIT שהתבלט כאחד החוקרים המבטיחים בתחום הבינה המלאכותית[1] - הגיע למסקנה שהיכולת לפרש תמונות היא מאפיין יסוד של תבוניות; הוא ביקש מסטודנט לתואר ראשון בשם ג'רלד סוסמן (Sussman) להקדיש את חופשת הקיץ שלו לפתרון הבעיה, "לחבר מצלמת טלוויזיה למחשב ולתת למחשב לתאר את מה שהוא רואה".[2] כך נולד "פרויקט הראייה של הקיץ" (Summer Vision Project).[] מיותר לציין שהניסיון לגרום למחשבים "לראות" היה קשה מהצפוי והתמשך הרבה מעבר לחופשת הקיץ.

הסיפור שסיפרו לנו במשך שנים נשמע כך: אנשים מבריקים עבדו על בעיית הראייה הממוחשבת לסירוגין, במשך עשרות שנים, עד שבשנות התשעים התקדמותם הואצה בזכות המעבר לשיטות של מידול ולמידה הסתברותיים. כך הגענו לתקופתנו, שבה בעיות כמו זיהוי עצמים וזיהוי פנים נפתרו ברובן.[4] הסיפור הזה חוזר בנרטיבים רבים שעוסקים בבינה מלאכותית, שמניחים כי שיפורים טכניים הדרגתיים יכולים להתמודד עם כל בעיה ומגבלה.

אבל מה אם לא כך הוא? מה אם האתגר - לגרום למחשבים "לתאר את מה שהם רואים" - תמיד יהיה בעיה? במאמר זה נבחן מדוע פרשנות ממופנת של דימויים אינה עניין טכני טהור, אלא פרויקט חברתי ופוליטי מטבעו. הבנת הפוליטיקה הכרוכה במערכות בינה מלאכותית חשובה היום יותר מתמיד, משום שהמערכות הללו משתלבות במהירות במבנה של מוסדות החברה ובתהליכי ההחלטה שקובעים את מי לזמן לריאיון עבודה, אילו חשודים לעצור ועוד.

בשנתיים האחרונות חקרנו את ההיגיון שעל פיו משתמשים בתמונות לאימון יכולתן של מערכות בינה מלאכותית "לראות" את העולם. בחנו מאות אוספים של תמונות המשמשים בתחום הבינה המלאכותית - מן הניסויים הראשונים בזיהוי פנים, שנערכו בתחילת שנות השישים, ועד לערכות אימון עכשוויות שכוללות מיליוני תמונות. מבחינה מתודולוגית אפשר לקרוא לפרויקט הזה "ארכיאולוגיה של ערכות נתונים" (datasets): חפרנו בכל שכבות החומר, קטלגנו את העקרונות והערכים שבאמצעותם נבנות ערכות הנתונים, ניתחנו אילו דפוסי חיים נורמטיביים מונחים ביסודן, באילו הן תומכות, מה הן משעתקות. חשיפת היסודות ואופני הבנייה של ערכות האימון הללו אפשרו לנו לשפוך אור על הנחות רבות שעמדו עד עתה ללא עוררין - הנחות היסוד העומדות עד היום בבסיס פעולתן, וגם בבסיס כשליהן, של מערכות הבינה המלאכותית.

המאמר פותח בשאלה פשוטה לכאורה: מה פעולתן של התמונות במערכות בינה מלאכותית? מה אמורים המחשבים לזהות בתמונה, אילו טעויות הם עושים, ומה נשאר שקוף לגמרי מבחינתם? אחר כך נבדוק את השיטה שבה מזינים תמונות למערכות מחשבים, ונראה כיצד טקסונומיות מארגנות את מושגי היסוד שהמערכת הממוחשבת לומדת "להבין". אז נפנה לשאלת התיוג: כיצד בני אדם מורים למחשבים אילו מילים לקשור לכל תמונה? ומה אנו מסכנים כאשר מערכות בינה מלאכותית משתמשות בתיוגים הללו כדי לסווג בני אדם לפי גזע, מגדר, רגש, כישורים, מיניות ואישיות? לבסוף, נפנה לתכליות שאותן הראייה הממוחשבת נועדה לשרת - יכולות שיפוט ובחירה - ולהשלכות שיש למחשוב היכולות האלה.

אימון הבינה המלאכותית

כדי לבנות מערכות בינה מלאכותית נדרשים נתונים. מערכות מבוקרות של למידת מכונה, שנועדו לזהות אובייקטים או פנים, מאומנות באמצעות כמויות אדירות של נתונים, שמקורם בערכות נתונים שכוללות הרבה מאוד תמונות מובחנות. כדי לבנות מערכת לראייה ממוחשבת שתוכל לזהות למשל את ההבדל בין תמונות של תפוחים ותפוזים, מפתחי המערכת צריכים לאסוף ולתייג אלפי תמונות של תפוחים ותפוזים ולאמן באמצעותן רשת עצבית מלאכותית. התוכנה עורכת סקירה סטטיסטית של התמונות ומפתחת מודל לזיהוי ההבדל בין שתי הקבוצות. אם הכול מתנהל כשורה, המודל המאומן יצליח בשלב זה להבחין בין תפוחים לתפוזים בתמונות חדשות שלא נתקל בהן בעבר.

אם כן, ערכות האימון הן הבסיס שעליו בנויות מערכות של למידת מכונה בימינו.[5] הן ממלאות תפקיד מפתח באופן שבו מערכות בינה מלאכותית מזהות ומפרשות את העולם. ערכות הנתונים הללו מעצבות את הגבולות האֶפִיסטמיים השולטים בפעולתן של מערכות בינה מלאכותית, ועל כן הן מהותיות להבנתן של שאלות חברתיות חשובות שקשורות לבינה מלאכותית.

אבל כאשר בוחנים את תמונות האימון המשמשות בדרך כלל במערכות ראייה ממוחשבת, מוצאים תשתית של הנחות מפוקפקות ומוטות. מְסיבות שאינן נדונות בדרך כלל בשדה הראייה הממוחשבת, ולמרות פועלם של מוסדות כמו MIT וחברות כמו גוגל ופייסבוק, הפרויקט של פירוש התמונות הוא מפעל מורכב להחריד, עמוס זיקות וקשרי גומלין. תמונה היא דבר חמקמק מאוד, טעון באינספור משמעויות פוטנציאליות, שאלות סתומות וסתירות. ענפים שלמים בפילוסופיה, בתולדות האמנות ובתיאוריית המדיה מוקדשים לחילוץ כל הדקויות ממערכת היחסים הלא יציבה שבין תמונה למשמעות.[6]

תמונות אינן מתארות את עצמן. אמנים חוקרים את התכונה הזאת מאות שנים. אגנס מרטין יצרה ציור דמוי רשת וקראה לו "פרח לבן", ורנה מגריט צייר תפוח וכתב מעליו את המילים "זה לא תפוח". כאשר אנו רואים כיצד התמונות הללו מתויגות, אנו רואים אותן אחרת. הקשר שבין תמונה, תיוג ורפרנט הוא גמיש, ואפשר לבנות אותו מחדש במגוון דרכים ולמגוון צרכים. יתר על כן, הקשרים האלה יכולים להשתנות במשך הזמן עם השינוי בהקשר התרבותי של התמונה, ומשמעותם יכולה להשתנות לפי המתבונן או לפי המקום. תמונות פתוחות לפירוש ולפירוש מחדש. זו אחת הסיבות לכך שמשימות של זיהוי וסיווג אובייקטים מורכבות יותר מכפי שמינסקי ורבים מממשיכי דרכו תיארו לעצמם בתחילה.

המיתוס השכיח גורס כי הבינה המלאכותית והנתונים שהיא מסתמכת עליהם מסווגים את העולם באופן אובייקטיבי ומדעי; אבל בפועל הם רוויים בפוליטיקה, אידיאולוגיה, דעות קדומות וכל שאר החומרים הסובייקטיביים שמהם עשויה ההיסטוריה. כאשר סוקרים את ערכות האימון הנפוצות ביותר, מוצאים שזהו הכלל ולא היוצא מן הכלל.

אנטומיה של ערכת אימון

אף שערכות אימונים עשויות להיות שונות זו מזו במטרותיהן ובארכיטקטורות שלהן, יש להן כמה תכונות משותפות. בבסיסן, ערכות אימונים למערכות ראייה ממוחשבת מורכבות מאוסף תמונות שתויגו בדרכים שונות וחולקו לקטגוריות. את הארכיטקטורה הכללית שלהן אפשר לחלק לשלוש שכבות: הטקסונומיה הכללית (סך הקבוצות והארגון ההיררכי שלהן, אם יש כזה); הקבוצות עצמן (קטגוריות מובחנות שבתוכן מסודרות התמונות, למשל "תפוח" או "תפוז"); וכל תמונה על תיוגה הנפרד (כלומר תמונה יחידה שתויגה כתפוח). לטענתנו, כל אחת ואחת מן השכבות בערכת אימון כזאת רוויה בפוליטיקה.

ראו לדוגמה את "מסד הנתונים של הבעות פנים של נשים יפניות", ערכה שפיתחו מייקל ליונס (Lyons), מיוקי קמאצ'י (Kamachi) וג'ירו גיובה (Gyoba) בשנת 1998 - ערכה פופולרית מאוד המשמשת לצורכי מחקר ופיתוח בתחום הניתוח הממוחשב של הרגשות. ערכת הנתונים כוללת תצלומים של עשר נשים יפניות המציגות שבע הבעות פנים שאמורות לייצג שבעה מצבים רגשיים בסיסיים.[7] המטרה המוצהרת של ערכת הנתונים הזאת היא לעזור למערכות של למידת מכונה לזהות ולתייג רגשות דומים בתמונות חדשות. הטקסונומיה המובלעת כאן בשכבה העליונה היא, פחות או יותר, "הבעות פנים המתארות רגשות של נשים יפניות".

אם נרד לשכבה הבאה בטקסונומיה נגיע לרמת הקבוצה. בדוגמה זו הקבוצות הן שמחה, עצב, הפתעה, גועל, פחד, כעס והבעה ניטרלית. קטגוריות אלו הן ה"סלים" המארגנים שבתוכם מאוכסנות כל התמונות. במסד נתונים שנועד לזיהוי פנים, הקבוצות עשויות להיות שמות האנשים שפניהם מצויים בערכה. במסד נתונים שנועד לזיהוי אובייקטים, הקבוצות יהיו למשל תפוחים ותפוזים. אלה המושגים המובחנים המשמשים לארגון התמונות.

בשכבה הנמוכה ביותר של הארכיטקטורה של ערכת האימון נמצאת התמונה המתויגת - למשל תמונה של פנים שמתויגת כביטוי של מצב רגשי, תמונתו של אדם מסוים או תמונה של אובייקט מסוים. במקרה של מאגר הבעות הפנים של הנשים היפניות, זאת השכבה שבה אפשר למצוא תמונה של אישה מסוימת מעווה פנים, מחייכת או מופתעת.

הערכה של תמונות הנשים היפניות מבוססת על כמה הנחות מובלעות. ראשית, ברמת הטקסונומיה ההנחה היא שהקטגוריה "רגשות" היא מקבץ תקף של מושגים חזותיים. אחר כך באה שרשרת של הנחות נוספות: שאת המושגים האלה אפשר ליישם על תצלומי פנים של אנשים (ובפרט נשים יפניות); שיש שישה רגשות ועוד אחד ניטרלי; שיש קשר קבוע בין הבעת פניו של אדם ובין מצבו הרגשי האמיתי; ושקשר זה בין הבעה לרגש הוא עקבי, מדיד ואחיד עבור כל הנשים בתצלומים.

ברמת הקבוצה משמשות הנחות כמו "יש הבעת פנים 'ניטרלית'" וכן "ששת המצבים הרגשיים המשמעותיים הם 'שמחה', 'עצב', 'כעס', 'גועל', 'פחד', 'הפתעה'".[8] ברמת התמונה המתויגת מובלעות הנחות נוספות, כמו "התצלום המסוים הזה מתאר אישה 'כועסת'", אף שלמעשה בתצלום מופיעה אישה

שמציגה הבעה כועסת. שהרי בפועל כולן הבעות פנים "מעושות", שאינן נובעות ממצב רגשי פנימי כלשהו אלא משׂוחקות בתנאי מעבדה. כל אחת מן הטענות המובלעות בכל אחד מן הרבדים היא בעייתית במקרה הטוב, וחלקן מפוקפקות ביותר.[9]

ערכת האימון של תמונות הנשים היפניות צנועה במידותיה בהשוואה לערכות אימון בנות ימינו. היא נוצרה לפני עליית הרשתות החברתיות, בתקופה שבה מפתחים לא יכלו עדיין לחלץ מן הרשת כמויות עצומות של תמונות, ולפני שפלטפורמות מקוונות כמו Mechanical Turk של אמזון (פלטפורמה למיקור המונים של משימות עבודה) אפשרו לחוקרים ולחברות להתמודד עם המטלה הכבירה ולתייג כמויות אדירות של תצלומים. ככל שהתרחב היקפן של ערכות האימון כך גדלה גם מורכבותן והתרחבו האידיאולוגיות, הסמיולוגיות והפוליטיקות שבבסיסן. כדי לראות כיצד בא הדבר לידי ביטוי נפנה אל ערכת האימון המפורסמת ביותר, הלוא היא אימג'נט.

ערכת האימון הקנונית: אימג'נט

אחת מערכות האימון החשובות בתולדות הבינה המלאכותית היא אימג'נט, שהוצגה לראשונה כפוסטר מחקר בשנת 2009. מדובר בערכת נתונים גדולה ושאפתנית במידה יוצאת דופן. לדברי אחת מיוצרותיה, פֵּיי-פֵּיי לי (Fei-Fei Li) מאוניברסיטת סטנפורד, הרעיון העומד מאחורי אימג'נט הוא "למפות את כל עולם האובייקטים".[10] בתוך כמה שנים צמח המאגר של אימג'נט לממדים עצומים: צוות הפיתוח אסף מיליונים רבים של תמונות מן האינטרנט, ולזמן קצר הפך למשתמש האקדמי הגדול ביותר ב-Mechanical Turk של אמזון: גדודי עובדים מיינו חמישים תמונות בדקה לתוך אלפי קטגוריות.[11] בסיום הפרויקט כללה אימג'נט יותר מ-14 מיליון תמונות מתויגות שאורגנו ביותר מעשרים אלף קטגוריות. במשך עשור היא נחשבה לפסגת ההישגים בתחום זיהוי האובייקטים ללמידת מכונה.

הניווט ברחבי המבוך של אימג'נט מזכיר טיול בספריית בבל האינסופית של בורחס. הוא עצום ועתיר קוריוזים. יש בו קטגוריות לתפוחים, גרניום בריח תפוחים, כיסני תפוחים, כנימות תפוחים, מחית תפוחים, מיץ תפוחים, עגלות תפוחים, עוגיות תפוחים, עצי תפוח, רוטב תפוחים, ריבת תפוחים, רימות תפוחים, שיכר תפוחים. בקטגוריה hot אפשר למצוא תמונות של קווי חירום (hotline), מכנסיים צמודים (hot pants), כיריים, נזידי בשר ותפוחי אדמה, רכבי אספנות משודרגים, רטבים חריפים, מרחצאות חמים, המשקה hot toddy, אמבטיות חמות, כדורים פורחים, פאדג' שוקולד חם ובקבוקי מים חמים.

עד מהרה הייתה אימג'נט לנכס חיוני למחקרים בראייה ממוחשבת. היא הפכה בסיס לתחרות שנתית שבה מעבדות מכל רחבי העולם הפעילו את האלגוריתמים שלהן על ערכת אימון נבחרת וניסו לתייג בדיוק מרבי ערכות של תמונות. בשנת 2012 צוות מאוניברסיטת טורונטו השתמש בשיטה חדשה בשם "רשת עצבית מתקפלת" (convolutional neural network), ניצח בקלות בתחרות ומשך את תשומת הלב. הרגע הזה נחשב לנקודת מפנה בהתפתחות הבינה המלאכותית בימינו.[12] תחרות אימג'נט האחרונה התקיימה בשנת 2017, ומידת הדיוק בקטלוג האובייקטים בערכה הנבחרת טיפסה מ-71.8% ל-97.3%. מטעמים שיובהרו מיד, הערכה הזאת לא כללה את הקטגוריה "אנשים".

טקסונומיה

המבנה הבסיסי של אימג'נט מבוסס על המבנה הסמנטי של וורדנט (WordNet), מסד נתונים לסיווג מילים שפותח באוניברסיטת פרינסטון בשנות השמונים. הטקסונומיה היא היררכיית מושגים מקוננת (nested) של מקבצי מילים נרדפות שנקראים סינְסֶט (synset). כל סינסט מייצג מושג נבדל; למשל, המילים הנרדפות "אוטו" ו"מכונית" משויכות לאותו סינסט. הסינסטים מאורגנים היררכית, ממושגים כלליים אל מושגים ספציפיים יותר. המושג "כיסא", למשל, מקונן כך: "חפץ"> "ציוד" > "רהיט" > "מושב" > "כיסא". מערכת הסיווג דומה פחות או יותר לזו המשמשת לסידור ספרים בספריות, בקטגוריות ההולכות ונעשות מובחנות יותר.

בניגוד לוורדנט, שמטרתו לארגן את השפה האנגלית כולה,[13] אימג'נט מוגבלת לשמות עצם בלבד, מתוך תפיסה ששמות עצם ניתנים לייצוג באמצעות תמונות. בהיררכיה של אימג'נט, כל מושג מאורגן תחת אחת מתשע קטגוריות על: צמחייה, מבנה גיאולוגי, אובייקט טבעי, ספורט, חפץ, פטרייה, אדם, חיה ושונות. מתחת לקטגוריות האלה יש שכבות של מחלקות מקוננות אחרות.

כפי שהראו זה מכבר מדעי המידע ולימודי מדע וטכנולוגיה, כל טקסונומיה או מערכת מיון היא פוליטית.[14] באימג'נט (שירשה זאת מוורדנט), הקטגוריה "גוף האדם" (human body), למשל, ממוקמת כך: "אובייקט טבעי" > "גוף" > "גוף האדם". התת-קטגוריות שלה הן "גוף גברי", "אדם" (person), "גוף צעיר", "גוף בוגר" ו"גוף נשי". הקטגוריה "גוף בוגר" מכילה את התת-קטגוריות "גוף נשי בוגר" ו"גוף גברי בוגר". כאן מובלעת ההנחה שרק גופים "גבריים" ו"נשיים" הם "טבעיים". קיימת באימג'נט הקטגוריה "הרמפרודיט", אך באופן מוזר (ופוגעני) היא הייתה משובצת תחת הענף "אנשים" > "תאוותן" > "ביסקסואלי", לצד הקטגוריות "פסאודו-הרמפרודיט" ו"גמיש מינית (Switch Hitter)" (קטגוריות אלו נמחקו מאימג'נט ואינן קיימות עוד). היררכיית הסיווג של אימג'נט מזכירה את מערכת המיון הישנה של ספריית הקונגרס האמריקנית, שסיווגה ספרים להט"קיים תחת הקטגוריה "יחסי מין לא-נורמליים, לרבות עבירות מין". הסיווג שונה לבסוף בשנת 1972, אחרי מאבק ממושך שניהל "כוח המשימה לשחרור הקהילה הגאה" של איגוד הספריות האמריקניות.[15]

אם נרד שלב אחד מתחת לטקסונומיה, אל 21,841 הקטגוריות בהיררכיה של אימג'נט, נמצא סוג נוסף של פוליטיקה.

קטגוריות

בניית קטגוריות כמוה כמעשה כשפים. כאשר יוצרים קטגוריות או קוראים לדבר בשם, מחלקים יקום שמורכבותו אינסופית כמעט לתופעות נבדלות זו מזו. וכאשר כופים סדר על מסה חסרת הבחנה, כאשר ממיינים תופעות לקטגוריות - כלומר כאשר קוראים בשם - הופכים את עצם קיומה של הקטגוריה לממשי.

במקרה של אימג'ינט, נדמה שקטגוריות של שמות עצם כמו "תפוח" או "רסק תפוחים" אינן שנויות במחלוקת; אבל לא כל שמות העצם נולדו שווים. אם נשאל רעיון מהבלשן ג'ורג' לייקוף (Lakoff), המושג "תפוח" הוא שֵמי (nouny) יותר מהמושג "אור", שבעצמו הוא שֵמי יותר ממושג כמו "בריאות". [16] שמות עצם נמצאים על הרצף שבין המוחשי למופשט, או בין התיאורי לשיפוטי. אבל הדקויות הללו נמחקות בהיגיון של אימג'ינט. הכול מושטח ומתויג, כמו פרפרים מיובשים בארון תצוגה. התוצאות עלולות להיות בעייתיות, חסרות היגיון וגם אכזריות, בייחוד כאשר התוויות מוצמדות לבני אדם.

אימג'ינט כוללת 2,833 תת-קטגוריות תחת קטגוריית העל "אנשים" (person). הכמות הגדולה ביותר של תמונות נמצאת בתת-קטגוריה "בחורה" (1,664) (gal תמונות), ואחריה "סבא" (1,662), "אבא" (1,643) ו"מנכ"ל" (1,614). בקטגוריות העמוסות הללו אפשר כבר לראות קווי מתאר ראשונים של תפיסת עולם. אימג'ינט מסווגת בני אדם לטווח עצום של סוגים - לפי גזע, לאום, מקצוע, מעמד כלכלי, התנהגות, אופי ואפילו מוסריות. יש קטגוריות של זהויות לאומיות הכוללות ״ילידים בני אלסקה״, ״אנגלו-אמריקנים״, ״שחורים״, ״אפריקנים שחורים״, ״אישה שחורה״, ״אירו-אסיאתים״, ״גרמנו-אמריקנים״, ״יפנים״, ״לאפים״, ״לטינו-אמריקנים״, ״מקסיקנים-אמריקנים״, ״ניקרגואנים״, ״ניגרים״, ״פקיסטנים״, ״פפואנים״, ״אינדיאנים דרום-אמריקנים״, ״היספאנים״, ״טקסנים״, ״אוזבקים״, ״לבנים״, ״תימנים״ ו״בני זולו״. אנשים אחרים מתויגים לפי קריירה או תחביב: יש ״צופים״, ״מעודדות״, ״מדעני מוח״, ״ספרים״, ״מנתחי אינטליגנציה״, ״חוקרי מיתולוגיה״, ״קמעונאים״, ״גמלאים״, וכן הלאה.

כאשר צוללים אל עומק קטגוריות האנשים של אימג'ינט, מלאכת המיון של האנושות פונה פתאום לתוך סמטה אפלה. אנחנו מוצאים שם קטגוריות ל"איש רע", ״אפס״, ״בינוני״, ״בתולה זקנה״, ״גברבר״, ״דיכאוני״, ״דפוק״, ״הומו בארון״, ״הססן״, ״חובבן״, ״יצאנית״, ״כישלון״, ״לוזר״, ״מופקרת״, ״מסומם״, ״מרשעת״, ״משוגע״, ״נערת טלפון״, ״סוטה״, ״סכיזופרן״, ״עבריין״, ״פרימדונה״, ״צבוע״, ״קלפטומן״, ״רכרוכי״, ״שקוף״ ו״שמוק״. יש שם אינספור כינויי גנאי גזעניים ומונחים מיזוגיניים.

מכיוון שאימג'ינט שימשה לרוב לזיהוי אובייקטים, הקטגוריה "אנשים" נדונה לעיתים נדירות בלבד בכנסים מקצועיים וגם לא זכתה לתשומת לב ציבורית רחבה. אבל הארכיטקטורה המורכבת הזאת, ובה תמונות של בני אדם אמיתיים המתויגים לא פעם בתוויות פוגעניות, זמינה באינטרנט זה עשור. היא מספקת דוגמה חשובה ונוקבת למורכבויות ולסכנות הכרוכות במיון של בני אדם, ולמדרון החלקלק המחבר בין תוויות פשוטות לכאורה כמו "חצוצרן" או "שחקנית טניס" ובין מושגים כמו "עווית", "מולאטית" או ״רֶדנֶק״. במנותק מן הניטרליות המשוערת של קטגוריה נתונה כלשהי, בחירת התמונות מעוותת את המשמעות באופן ממוגדר, מוגזע, אייבליסטי וגילני. אימג'ינט היא דוגמה למה שקורה כאשר מסווגים בני אדם כאובייקטים. בשנים האחרונות תפוצתה של הפרקטיקה הזאת רק הולכת וגדלה, לעיתים קרובות בתאגידים הגדולים ביותר של בינה מלאכותית, ומי שאינו עובד בחברות אלו אינו יכול לראות כיצד התמונות מאורגנות וממוינות.

לבסוף, נשאלת השאלה מאין נשאבות אלפי התמונות בקטגוריה "אנשים". יוצרי אימג'נט קצרו המוני תמונות ממנועי חיפוש כמו גוגל, ניכסו לעצמם תמונות סֶלפי ונופש בלי ידיעת המצולמים, ואז תייגו וארזו אותן מחדש כנתוני בסיס לתחום שלם.[17] כאשר מביטים בשכבת היסוד של התמונות המתויגות מוצאים הנחות סמיוטיות מפוקפקות ביותר, הדהודים לתורת הפְּרֶנולוגיה מן המאה התשע-עשרה, וייצוג מזיק הנגרם מסיווג תמונות של אנשים בלא הסכמתם או שיתופם.

תמונות מתויגות

התגיות של אימג'נט מפשטות לא פעם את התמונות ומשטחות אותן לכדי קלישאות גמורות. תצלום אחד מראה למשל פעוטה שחומת עור בבגדים מרופטים ומלוכלכים, פיה פתוח ובידה בובה מוכתמת. התמונה נטולת הקשר כלשהו. מי הילדה? איפה היא נמצאת? התצלום מתויג בפשטות במילה "צעצוע".

בתגיות אחרות פשוט אין כל היגיון. אישה יושבת במטוס, ישנה, וידה הימנית מגוננת על בטן הריונית. התגית: "סנובית". תמונה ערוכה בפוטושופ מראה את ברק אובמה במדים נאציים, מחייך, וידו מורמת ומחזיקה דגל נאצי. התגית: "בולשביק".

ברובד התמונות של ערכת האימון, כמו ברבדים שמעליו, אנחנו מוצאים הנחות, פוליטיקה והשקפות עולם. על פי אימג'נט, למשל, השחקנית סיגורני ויבר היא "הרמפרודיטית", צעיר בכובע קש הוא "שמוק", וצעירה השוכבת על מגבת חוף היא "קלפטומנית". אבל תפיסת העולם של אימג'נט אינה מוגבלת לחיבור המשונה או הפוגעני בין תמונות לתגיות.

הנחות יסוד נוספות לגבי הקשר בין תמונות למושגים מהדהדות את "חוכמת הפרצוף" או הפיזיונומיה - ההנחה הפסאודו-מדעית שלפיה אפשר לפענח את אופיו של אדם על פי מאפיינים של גופו ופניו. אימג'נט לוקחת את התפיסה הזאת עד הקצה, ומניחה שאפשר להכריע אם אדם הוא "בעל חוב", "סנוב" או "מתפרפר" על פי הסתכלות בתמונה. במטפיזיקה המשונה של אימג'נט יש קטגוריות תמונה נפרדות ל"מרצה בכיר" ול"פרופסור חבר" - כאילו כאשר אדם מקודם במשרתו, החתימה הביומטרית שלו תשקף זאת.

מובן שלהנחות אלו יש היסטוריה אפלה משלהן ותפיסות פוליטיות שנלוות אליהן.

אוניברסיטת טנסי: פענוח הגזע והמגדר ממראה הפנים

בשנת 1839 טען המתמטיקאי פרנסואה אָרָגו (Arago) שהתצלום "משמר מתמטית את צורתו של האובייקט".[18] בהקשר של האימפריאליזם והדרוויניזם החברתי של המאה התשע-עשרה, הצילום עודד צורות שונות של פְּרֶנולוגיה, פיזיונומיה ואאוגניקה והעניק להן כסות "מדעית".[19] פיזיונומים כמו פרנסיס גלטון (Galton) וצֶ'זָרה לומברוזו (Lombroso) יצרו תמונות מרוכבות של פושעים, חקרו כפות רגליים של נשים שעסקו בזנות, מדדו גולגולות והרכיבו ארכיונים קפדניים של תמונות מתויגות

ומדידות - הכול בניסיון לזהות סימנים חזותיים באמצעות תהליכים "מכניים" ולאפשר מיונים לפי גזע, נטייה לעבריינות וסטייה מאידיאלים בורגניים. המטרה הייתה לאתר את כל מה שנראה כהתנהגות סוטה או עבריינית, לחשוף אותה לעין כול ולערוך לה פתולוגיזציה.

וכפי שנראה, לא זו בלבד שהנחות היסוד של הפיזיונומיה קמו לתחייה בערכות אימונים מודרניות; כמה וכמה ערכות אימונים אף נבנו כך שאלגוריתמים ותווי פנים ישמשו כקני מידה חדשים למדידות גולגולת. כך לדוגמה, ערכת הנתונים של אוניברסיטת טנסי בנוקסוויל, הקרויה UTKFace, כוללת יותר מעשרים אלף תמונות פנים עם הערות גיל, מגדר וגזע. יוצרי הערכה טוענים שהיא יכולה לשמש למגוון מטרות, כמו זיהוי פנים אוטומטי, הערכת גיל ועיבוד תמונות לשינוי הגיל (age progression).

בהערות המוצמדות לכל תמונה מצוין גילו המשוער של כל אדם, מאפס ועד 116. המגדר הוא בחירה בינרית: 0 לזכר, 1 לנקבה. הגזע מקוטלג לפי חמש קטגוריות שערכיהן 0-4: לבן, שחור, אסיאתי, אינדיאני ו"אחר". הפוליטיקה כאן מובהקת ומטרידה. ברובד הקטגוריה, התפיסה המגדרית של החוקרים היא של מבנה בינרי פשוט שבו זכר ונקבה הם החלופות הבלעדיות. ברובד התמונה המתויגת, ההנחה היא שזהות מגדרית אפשר לקבוע על פי תצלום.

השיטה של מיון הגזע מזכירה מיוני גזע מפוקפקים מן המאה העשרים. משטר האפרטהייד בדרום אפריקה, למשל, ביקש לסווג את האוכלוסייה כולה לארבע קטגוריות: שחור, לבן, צבעוני (בן תערובת) או הודי.[20] בשנת 1970 לערך יצרה ממשלת דרום אפריקה מערכת "פנקסי מעבר" מאוחדת בשם "ספר החיים" (Book of Life), שנקשר במאגר נתונים מרכזי שבנתה חברת IBM. המיון התבסס על קריטריונים מפוקפקים ונזילים של "מראה והתקבלות כללית או מוניטין", ואנשים רבים סווגו מחדש, לפעמים יותר מפעם אחת.[21] מערכת סיווג הגזע הדרום-אפריקנית הייתה במכוון שונה מאוד מכלל "הטיפה האחת" האמריקני (one-drop rule), שקבע כי די באב קדמון יחיד ממוצא אפריקני כדי להגדיר אדם כשחור - כנראה משום שכמעט לכל הדרום-האפריקנים הלבנים הייתה מידה כלשהי של מורשת אפריקנית שחורה.[22] מערכות המיון הללו גרמו נזק עצום לאנשים, ומדד הגזע החמקמק תמיד היה שנוי במחלוקת. אבל הניסיון לשפר את המצב באמצעות פיתוח ערכות אימון "מגוונות יותר" עבור הבינה המלאכותית מעורר קשיים אחרים.

מגוון הפנים של IBM

ערכת הנתונים של IBM, הקרויה Diversity in Faces (DiF), הוקמה בתגובה לביקורת שהראתה כי תוכנת זיהוי הפנים של החברה מתקשה לזהות בעלי עור כהה.[23] IBM הכריזה כי יש בכוונתה לשפר את ערכות הנתונים המשמשות אותה לזיהוי פנים ולעשותן "מייצגות" יותר, ולשם כך פרסמה את ערכת הנתונים DiF.[24] המאגר נועד לשמש "בסיס מעשי מבחינה מִחשובית להבטחת הגינות ודיוק בזיהוי פנים". הוא מכיל כמעט מיליון תמונות שנשלפו מערכת הנתונים החינמית של פליקר (Yahoo! Flickr Creative Commons), שנאספו במיוחד במטרה להשיג שוויון סטטיסטי בין קטגוריות של גוון עור, מבנה פנים, גיל ומגדר.

לערכת הנתונים עצמה המשיכו להתווסף מאות אלפי תמונות של אנשים תמימים שהעלו תצלומים לאתרים כמו פליקר.[25] אבל יש בה מקבץ ייחודי של קטגוריות שלא נראו עד אז בערכות נתונים אחרות של תמונות פנים. אנשי צוות מגוון הפנים של IBM תהו אם נתונים של גיל, מגדר וצבע עור אכן מספיקים כדי לייצר ערכת נתונים שתבטיח הגינות ודיוק בזיהוי פנים, והגיעו למסקנה שדרושים סיווגים נוספים. לכן הם נקטו שיטה משונה מאוד: שימוש בנתונים הנוגעים לסימטריה של פנים וצורות גולגולת כדי ללמד את התוכנה לבנות תמונה שלמה של הפנים. לטענת החוקרים, השימוש במאפיינים של מבנה הגולגולת מוצדק, משום שהוא לוכד מידע ספציפי הרבה יותר לגבי פניו של אדם בהשוואה למגדר, גיל וצבע עור בלבד. המאמר המלווה את ערכת הנתונים מדגיש מחקרי עבר שהראו כי צבע העור לבדו אינו מנבא חזק של גזע; אבל מתבקשת השאלה מדוע בחינה של צורת הגולגולת היא גישה ראויה.

מדידת גולגולות הייתה הגישה המתודולוגית המובילה בדטרמיניזם הביולוגי של המאה התשע-עשרה. כפי שמראה סטיבן ג'יי גולד בספרו **אין מידה לאדם**, מידות גולגולת שימשו פסאודו-מדענים במאות התשע-עשרה והעשרים ל"הוכחת" עליונותם הטבועה של לבנים על פני שחורים, וצורות ומשקלות של גולגולות נחשבו מדד אמין לקביעת אינטליגנציה - תמיד לפי קווים גזעיים.[26]

חברות מתאמצות לבנות ערכות אימון מגוונות יותר ולעיתים קרובות הן טוענות כי המטרה היא לחזק את ה"הגינות" ו"למתן את ההטיה", גם אם ברור שיש להן תמריצים עסקיים מובהקים לפתח כלים שפעולתם תהיה אפקטיבית יותר בשווקים רחבים יותר. אבל גם כאן התהליך הטכני של קִטלוג ומיון אנשים מתברר כמעשה פוליטי. כיצד, למשל, מושגת התפלגות "הוגנת" בתוך ערכת הנתונים?

חברת IBM החליטה להשתמש בגישה מתמטית לכימות ה"גיוון" וה"שוויון" כך שכל אחד מן המאפיינים המכומתים יהיה אחיד באופן עקבי לרוחב ערכת הנתונים. הערכה כוללת גם הערות סובייקטיביות המתייחסות לגיל ומגדר, שמיוצרות באמצעות שלושה עובדי Mechanical Turk לכל תמונה - כמו בשיטות המשמשות את אימג'נט. כלומר, מגדר וגיל של אנשים "מנובאים" על סמך ניחושיהם של שלושה אנשים לגבי תצלום שנשלף מהאינטרנט. הדבר מזכיר משחקים מסוג "נחש מה המשקל", והתוקף המדעי דומה.

בסופו של דבר, נוסף על הבעיות המתודולוגיות העמוקות, רעיון המגוון וההיסטוריה הפוליטית שלו מתרוקנים ממשמעות, וכל מה שנותר ממנו הוא התייחסות רחבה לפנוטיפים ביולוגיים. בהקשר זה, שונות פירושה בסך הכול מגוון רחב יותר של צורות גולגולת וסימטריית פנים. חוקרי הראייה הממוחשבת עשויים לראות בכך "מתמטיזציה של הגינוּת", אבל הדבר פשוט תורם לשיפור יעילותן של מערכות מעקב. ואפילו אחרי כל הניסיונות הללו להרחיב את דרכי המיון של בני אדם, ערכת הנתונים DiF עדיין נסמכת על סיווג בינרי של מגדר: אפשר לסמן אנשים רק כזכר או כנקבה. השגת שוויון סטטיסטי בין קטגוריות אינו זהה להשגת מגוון או הגינוּת, ובניית הנתונים ופיתוחם ב-IBM רק מנציחים שיטה מזיקה של סיווג, הכפופה להשקפת עולם צרה.

האפיסטמיקה של ערכות אימון

אילו הנחות עומדות ביסודן של מערכות לראייה ממוחשבת? ראשית, הפרדיגמה התיאורטית המונחת ביסוד ערכות האימון מניחה שהמושגים - יהיו אלה "תירס", "מגדר", "רגש" או "לוזר" - קיימים מלכתחילה, ושהמושגים האלה קבועים, אוניברסליים, וניחנים במעין עוגן טרנסצנדנטי ועקביות פנימית. שנית, היא מניחה שיש קשר קבוע ואוניברסלי בין תמונה למושג, בין חזות למהות. יתרה מזו, היא מניחה שיש קשרים לא מורכבים, מובנים מאליהם ומדידים בין תמונות, רפרנטים ותגיות. במילים אחרות, היא מניחה שלמושגים - למשל "תירס" או "קלפטומן" - יש מהות כלשהי שמחברת בין כל מופעיהם הפרטיים, ושלמהות היסוד הזאת יש ביטוי חזותי. יתר על כן, המהות החזותית ניתנת לזיהוי באמצעות שיטות סטטיסטיות שמחפשות דפוסים פורמליים באוספים של תמונות מתויגות. תמונות של אנשים שנקראים "לוזרים", כך על פי התיאוריה, מאופיינות בדפוס חזותי כלשהו המבחין אותן למשל מ"חקלאים", "מרצים בכירים" או אפילו "תפוחים". לבסוף, הגישה מניחה שכל שמות העצם המוחשיים דומים זה לזה מבחינה עקרונית, ושגם לרבים משמות העצם המופשטים (למשל "אושר" או "אנטישמיות") יש ביטוי קונקרטי וחזותי.

ערכות האימון של תמונות מתויגות, שנפוצות כל כך בתחום הראייה הממוחשבת והבינה המלאכותית בימינו, בנויות על מסד של הנחות אפיסטמולוגיות ומטפיזיות מפוקפקות ולא יציבות לגבי הטבע של תמונות, תוויות, מיונים וייצוגים. ההנחות הללו אף מהדהדות גישות היסטוריות שבגינן הערכה וסיווג חזותיים שימשו ככלי לדיכוי ולמדע גזעני.

ערכות נתונים אינן סתם חומרי גלם שמזינים אלגוריתמים; יש להן משמעויות פוליטיות. לפיכך, חלק נכבד מן הדיון ב"הטיות" הגלומות במערכות של בינה מלאכותית מחמיץ את הנקודה: אין נקודת מבט ניטרלית, טבעית או א-פוליטית שעליה אפשר לבסס את נתוני האימון. אין פתרון טכני קל שאפשר לממשו באמצעות תיקונים דמוגרפיים, מחיקה של מונחים פוגעניים או חתירה לייצוג שוויוני של גוני עור. עצם המפעל של איסוף תמונות, מיונן ותיוגן הוא כשלעצמו פוליטי, משום שיש לשאול מי מחליט מה משמעותן של תמונות ואילו סוגי עבודה חברתית ופוליטית הייצוגים הללו מבצעים.

היעלמות

בינואר 2019 החלו להיעלם תמונות בקטגוריית ה"אנשים" של אימג'נט. הגישה ל-1.2 מיליון תצלומים על השרתים של אוניברסיטת סטנפורד נחסמה. נעלמו תצלומי המעודדות, הצוללנים, הרתכים, נערי המזבח, הגמלאים והטייסים. נעלמה תמונת ה"אלכוהוליסט" ששותה בירה, ואיתה נעלמו ה"מופקרת" בביקיני והנער ה"לוזר". תמונת האדם שאכל סנדוויץ' ("אגואיסט") נעלמה אף היא. חיפוש של התמונות האלה מעלה כעת הודעה כי באתר אימג'נט נעשות עבודות תחזוקה ורק הקטגוריות שהשתתפו בתחרות אימג'נט נכללות עדיין בתוצאות החיפוש. בעת כתיבת המאמר הזה, הקטגוריה "אנשים" עדיין זמינה בממשק המקוון של ערכת הנתונים, אבל התמונות עצמן אינן עולות. עם זאת, כתובות הרשת לתמונות המקוריות זמינות עדיין (מחברי המאמר גיבו את ערכת הנתונים של אימג'נט טרם המחיקות הגדולות).

בחודשים הבאים החלו להיעלם אוספי תמונות נוספים המשמשים לחקר הראייה הממוחשבת והבינה המלאכותית. בתגובה למחקר שפרסמו אדם הארווי (Harvey) וז'ול לפלס (LaPlace) (שאת פרויקט MegaPixels שלהם אפשר לראות כאן), אוניברסיטת דיוק הורידה מאגר תצלומים עצום של סרטוני מצלמות אבטחה שהראו סטודנטים בכיתות (המאגר נקרא Duke Multi-Target, Multi-Camera [MTMC] dataset). התברר שמחברי ערכת הנתונים אספו תמונות של אנשים במרחב הציבורי והנגישו את ערכת הנתונים לציבור הרחב בניגוד לתנאי האישור שקיבלו מוועדת הביקורת המוסדית של האוניברסיטה. [27]

ערכות נתונים דומות מסרטוני אבטחה נעלמו מן השרתים של אוניברסיטת קולורדו בקולורדו ספרינגס, וכן מאוניברסיטת סטנפורד, שם "הוסרה לבקשת המפקיד" הגישה לאוסף תמונות פנים שלוקטו ממצלמת רשת שהותקנה בבית הקפה המפורסם Brainwash Cafe בסן פרנסיסקו.[28]

עד ראשית יוני באותה שנה, מיקרוסופט נהגה באופן דומה והסירה את אוסף תצלומי הידוענים שלה (MS-CELEB) שלה, שכלל כעשרה מיליון תצלומים של כמאה אלף בני אדם אשר נשלפו מן הרשת בשנת 2016. הייתה זו ערכת נתוני זיהוי הפנים הציבורית הגדולה בעולם, ונכללו בה לא רק שחקנים ופוליטיקאים ידועים אלא גם עיתונאים, פעילים חברתיים, קובעי מדיניות, אנשי אקדמיה ואמנים. [29] למרבה האירוניה, חלק ממי שנכללו במאגר בלא הסכמתם נודעו בזכות ביקורתם על תחום אמצעי המעקב וזיהוי הפנים - למשל הקולנוענית לורה פויטרס (Poitras), פעילת הזכויות הדיגיטליות ג'יליאן יורק (York), המבקר יבגני מורוזוב (Morozov) ומחברת הספר *Surveillance Capitalism* שושנה זובוף (Zuboff). אחרי תחקיר שפרסם ה**פיננשל טיימס** בהתבסס על עבודתם של הארווי ולפלס, הערכה נעלמה.[30] דובר מטעם מיקרוסופט טען שהיא הוסרה פשוט מפני ש"אתגר המחקר הסתיים".[31]

מצד אחד, הסרת ערכות הנתונים הבעייתיות הללו מן הרשת עשויה להיראות כניצחון: הפרות הפרטיות והאתיקה המובהקות ביותר מטופלות באמצעות ביטול הגישה לתמונות. אבל הסילוק מן הרשת אינו קוטע את פעולתן בעולם: ערכות האימון הללו הורדו אינספור פעמים ועשו את דרכן אל מערכות ייצור של בינה מלאכותית ולמאמרים אקדמיים. מחיקתן המוחלטת לא רק גורמת לאובדנו של חלק משמעותי בהיסטוריה של הבינה המלאכותית, אלא גם מונעת מחוקרים לראות כיצד הנחות, תגיות וגישות סיווג שוחזרו במערכות חדשות, או לעקוב אחר מקורותיהם של עיוותים והטיות כאלה ואחרים הפועלים במערכות הקיימות. מערכות בינה מלאכותית שמזהות פנים ורגשות כבר מחלחלות לתחומי העסקת העובדים, החינוך והבריאות. הן משתלבות בבדיקות ביטחוניות בשדות תעופה ובפרוטוקולים של ראיונות בחברות גדולות. העובדה שכבר אין לדעת כיצד אומנו מערכות הבינה המלאכותית מסלקת שיטה פורנזית חשובה שעשויה להסביר כיצד הן פועלות. ולעניין זה יש השלכות עמוקות ביותר.

כך לדוגמה, מחקר שהוביל לאחרונה דוקטורנט מאוניברסיטת קיימברידג' הציג מערכת מעקב באמצעות רחפן שנועדה לזהות בזמן אמיתי אלימות במקומות ציבוריים. המערכת אומנה על ערכות נתונים של "התנהגות אלימה", ומשתמשת במודלים הללו לצורך איתור ובידוד של התנהגות כזאת בתוך קהלים גדולים. הצוות יצר ערכת נתונים שנקראת AVI (Aerial Violent Individual), הכוללת אלפיים תמונות של אנשים המעורבים בחמישה סוגי פעולה: מכות, דקירות, ירי, בעיטות וחנק. כדי לאמן את מערכת הבינה

המלאכותית, 25 מתנדבים בני 18-25 התבקשו להציג את הפעולות הללו. הצפייה בסרטונים קומית כמעט. השחקנים עומדים הרחק זה מזה ומבצעים מחוות מוגזמות להחריד. הכול נראה כמו פנטומימה של ילדים, או כמו דמויות מגושמות במשחקי מחשב.[32] ערכת הנתונים השלמה אינה זמינה לציבור להורדה. החוקר הראשי, אמרג'וט סינג (Singh) (כעת באוניברסיטת הרווארד), אמר שהוא מתכוון לבחון את מערכת הבינה המלאכותית באמצעות הטסת רחפנים מעל שני פסטיבלים גדולים, ואולי גם באזור קווי הגבול של הודו.[33]

ניתוח "ארכיאולוגי" של ערכת הנתונים של AVI - כמו זה שעשינו לאימג'נט, למסד הנתונים של הנשים היפניות ול-DiF של IBM - עשוי להיות מאלף. יש הבדל ברור בין מצגים מבוימים של אלימות ובין אלימות בעולם האמיתי. החוקרים מאמנים רחפנים לזהות חיקויים של אלימות, ומכך עלולות לצמוח טעויות. יתר על כן, בערכת הנתונים של AVI אין כלל דוגמאות לפעולות שאינן אלימות אבל עשויות להיראות כך; מפתחיה אינם מפרסמים פרטים כלשהם לגבי שיעור התוצאות החיוביות השגויות (כלומר שיעור המקרים שבהם המערכת מזהה התנהגות אלימה אף שבפועל מדובר בהתנהגות לא אלימה).[34] כל עוד הנתונים אינם מפורסמים, אי-אפשר לערוך בדיקה פורנזית לאופן שבו הם מסַווגים ומפרשים גופים אנושיים ופעולות.

זו הבעיה הגלומה בערכות נתונים לא זמינות או נעלמות. אם הן משמשות כעת או שימשו בעבר במערכות שממלאות תפקיד בחיי היומיום, חשוב שנוכל ללמוד ולהבין את תפיסת העולם שהן מנרמלות. לטובת המחקר בעתיד, יש לפתח מסגרות שיאפשרו גישה לערכות הנתונים הללו מבלי להנציח את פגיעתן.

סיכום: מי מחליט?

הקרימינולוגים ממשיכי דרכו של לומברוזו, ושאר הפְרֶנולוגים מראשית המאה העשרים, לא ראו בעצמם ריאקציונרים פוליטיים. נהפוך הוא, כפי שסטיבן ג'יי גולד מציין, הם היו בעיקר ליברלים וסוציאליסטים שביקשו "להשתמש במדע המודרני כבמטאטא כדי לנקות מן המערכת המשפטית את המטען הפילוסופי המיושן של רצון חופשי ואחריות מוסרית מלאה".[35] הם האמינו שהשיטה האנתרופומטרית של לימוד העבריינות תוליך אל גישה מושכלת יותר להפעלתו של משפט צדק. חלקם באמת האמינו שהם מנקים את מערכת המשפט הפלילי מהטיות, שהם יוצרים תוצאות "הוגנות" יותר באמצעות יישום של שיטות "מדעיות" ו"אובייקטיביות".

בשיא פריחתן של הפרנולוגיה וה"אנתרופולוגיה הפלילית", האמן רנה מגריט צייר מקטרת וכתב לידה Ceci n'est pas une pipe ("זו אינה מקטרת"). מגריט קרא לציור הזה La trahison des images, "בגידת הדימויים". באותה שנה הוא גם חיבר מאמר מאויר קצר לביטאון הסוריאליסטי *La Révolution surréaliste* ("המהפכה הסוריאליסטית"). המאמר, שכותרתו "מילים ותמונות" ("Les mots et les images"), משתעשע במורכבויות ובדקויות של תמונות, תגיות, סמלים וסימונים, ומדגיש עד כמה אין דבר שהוא מובן מאליו ביחסים שבין תמונות ומילים או מושגים לשוניים. הסדרה של "בגידת הדימויים" הסתיימה בציור שנקרא "זה אינו תפוח".

הניגוד בין שתי הגישות לייצוג - זו של מגריט וזו של הפיזיונומים - מעיד על שתי תפיסות שונות בתכלית לגבי זיקת היסוד שבין תמונות לתגיות שלהן ולגבי הייצוג עצמו. הפיזיונומים האמינו שהזיקה בין דמותו של אדם ובין אופיו גלומה בתמונות עצמן. הנחתו של מגריט הייתה הפוכה כמעט: לגישתו, התמונות מקיימות לכל היותר זיקה לא יציבה עם הדברים שהן מתיימרות לייצג, וזיקה זו ניתנת לעיצוב בידי מי שיש לו הכוח לקבוע את משמעותה של תמונה נתונה. מבחינת מגריט משמעות התמונה נגזרת מן ההקשר, ואפשר לערער עליה. במבט ראשון הציור של מגריט עשוי להיראות כתעלול סמיוטי פשוט; אבל הדינמיקה הבסיסית שמגריט מציג בציור מצביעה על תפיסה פוליטית רחבה של ייצוג וייצוג עצמי.

מאבקים על צדק תמיד נסובו במידה מסוימת גם על משמעותם של דימויים וייצוגים. בשנת 1968, עובדי תברואה אפרו-אמריקנים פתחו בשביתה במחאה על תנאי העבודה המסוכנים שלהם ועל היחס הנורא שזכו לו מידי השלטון המקומי הגזעני בממפיס. הם הניפו שלטים שעליהם נכתב "אני אדם" (I AM A MAN), מתוך התייחסות לתנועה לביטול העבדות מן המאה התשע-עשרה. בשנות השבעים, פעילי התנועה הגאה ניכסו לעצמם סמל ששימש לזיהוי הומוסקסואלים, ביסקסואלים וטרנסים במחנות ריכוז נאציים. המשולש הוורוד נעשה טלאי של גאווה, מן הסמלים המזוהים ביותר עם התנועה הגאה. במאבקים על צדק, רבות הדוגמאות של אנשים שמבקשים להגדיר בעצמם את משמעויות הייצוג שלהם. ייצוגים אינם מוגבלים לתחומי השפה והתרבות; יש להם השלכות ממשיות על זכויות, חירויות וצורות של הגדרה עצמית.

למבנה ולתכנים של ערכות האימון המשמשות בבינה המלאכותית יש השלכות רבות. הם יכולים לקדם או להפלות, לאשר או לדחות, להבליט או להסתיר, לשפוט או לכפות. עלינו לבדוק אותם, כי הם כבר משמשים כדי לבדוק אותנו. עלינו לנהל דיון ציבורי נרחב בהשלכותיהם, ולא לשמור את הדיון הזה למסדרונות האקדמיה. ככל שגדלה חשיבותן של ערכות האימון בתשתית העירונית, המשפטית, הלוגיסטית והמסחרית, כך מתעצם כוחן לעצב את העולם בדמותן, והכוח הזה אינו מבוקר דיו.

הערות שוליים

[1]↑ מינסקי התמודד עם האשמות חמורות באשר לקשריו עם האנס והפדופיל המורשע ג'פרי אפשטיין. מינסקי היה אחד מבין כמה מדענים שנפגשו עם אפשטיין וביקרו באי שלו, שבו קטינות אולצו לקיים יחסי מין עם בני חוגו של אפשטיין. כפי שציינה החוקרת מרדית' ברוסארד (Broussard), היה זה חלק מתרבות נרחבת יותר של הדרה שנפוצה בתחום הבינה המלאכותית: "בצד היצירתיות הנהדרת של מינסקי וחבורתו, הם תרמו לחיזוק תרבות ההייטק כמועדון של גברים מיליארדרים. מתמטיקה, פיזיקה וכל שאר המדעים ה'קשים' מעולם לא הסבירו פנים במיוחד לנשים ולמיעוטים; עולם ההייטק הלך באותו כיוון". ראו Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, Cambridge, Massachusetts and London: MIT Press, 2018, p. 174

[2]↑ Daniel Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence*, New York: Basic Books, 1993, p. 88

[3]↑ מינסקי קיבל את הקרדיט על הרעיון, אבל מובן שצוותים של "עובדי קיץ" נטלו כולם חלק במאמץ המוקדם הזה לגרום למחשבים לתאר אובייקטים בעולם. כפי שכתב סימור פפרט (Papert), אחד המשתתפים: "פרויקט הראייה של הקיץ הוא ניסיון להשתמש בעובדי הקיץ שלנו באפקטיביות כדי לבנות חלק משמעותי במערכת ראייה. המשימה המסוימת נבחרה גם מפני שאפשר לחלקה לכמה תת-בעיות, וכך יחידים יכלו לעבוד עצמאית ובה בעת לקחת חלק בבנייה של מערכת שבזכות מורכבותה תהיה ציון דרך של ממש בפיתוח של 'זיהוי דפוסים'" (Seymour A. Papert, "The Summer Vision Project," DSpace MIT, July 1, 1966).

[4]↑ Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Upper Saddle River, NJ: Prentice Hall Pearson Education International, 2010, p. 987

[5]↑ בשלהי שנות השבעים, רישרד מיכלסקי (Michalski) כתב אלגוריתם שהתבסס על "משתני סמלים" וכללים לוגיים. שפה זו הייתה פופולרית מאוד בשנות השמונים והתשעים, אבל ככל שכללי קבלת ההחלטות והקטלוג נעשו מורכבים יותר, השפה נעשתה שמישה פחות. באותה תקופה, בזכות הפוטנציאל הטמון בערכות אימון גדולות, הגישה הזאת (שכונתה "צבירה קונספטואלית") נזנחה לטובת גישות בנות ימינו ללמידת מכונה. ראו Ryszard Michalski, "Pattern Recognition as Rule-Guided Inductive Inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (1980), pp. 349–361.

[6]↑ מאות ספרים אקדמיים עוסקים בנושא זה; ראו למשל William J. T. Mitchell, *Picture Theory: Essays on Verbal and Visual Representation*, Chicago: University of Chicago Press, 2007.

[7]↑ Michael Lyons, Shigeru Akamatsu, Miuky Kamachi, and Jiro Gyoba, "Coding Facial Expressions with Gabor Wavelets," *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205

[8]↑ כמתואר בדוח מ-2018 של מכון המחקר AI Now, סיווג זה של הרגשות לשש קטגוריות מקורו בעבודתו של הפסיכולוג פול אקמן (Ekman). "על פי אקמן, לימוד הפנים מספק קריאה אובייקטיבית של מצבים פנימיים אותנטיים - צוהר ישיר אל הנפש. ביסוד אמונתו עמד הרעיון שהרגשות קבועים ואוניברסליים, זהים בין אדם לאדם, ונגלים בבירור באמצעות מנגנונים ביולוגיים שאפשר לצפות בהם ואשר אינם מושפעים מהקשר תרבותי. אבל עבודתו של אקמן ספגה ביקורת נוקבת מידי פסיכולוגים, אנתרופולוגים וחוקרים אחרים [...].
הפסיכולוגית ליסה פלדמן ברט (Feldman Barrett) ועמיתיה טענו שהבנת הרגשות במונחים של קטגוריות נוקשות וסיבות פיזיולוגיות פשטניות אינה קבילה עוד. ולמרות זאת, חוקרי בינה מלאכותית התייחסו לעבודתו כאל עובדה והשתמשו בה כבסיס לאוטומטיזציה של זיהוי רגשות". ראו Meredith Whitaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kazionas et al., *AI Now Report 2018*, New York: AI Now Institute, 2018; Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix L. Martinez, and Seth D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," *Psychological Science in the Public Interest* 20, 1 (July 17, 2019), pp. 1–68.

[9]↑ ראו למשל Ruth Leys, "How Did Fear Become a Scientific Object and What Kind of Object Is It?" *Representations* 110, 1 (May 2010), pp. 66-104. ליס מתחה ביקורת בכמה הזדמנויות על תוכנית המחקר של אקמן. ראו Ruth Leys, *The Ascent of Affect: Genealogy and Critique*, Chicago and London: University of Chicago Press, 2017; Lisa Feldman Barrett, "Are Emotions Natural Kinds?" *Perspectives on Psychological Science* 1, 1 (March 2006), pp. 28-58; Erika H. Siegel, Molly K. Sands, Wim Van den Noortgate, Paul Condon, Yale Chang, et al., "Emotion Fingerprints or Emotion Populations? A Meta-Analytic Investigation of Autonomic Features of Emotion Categories," *Psychological Bulletin* 144, 4(2018), pp. 343-393.

[10]↑ Dave Gershgorn, "The Data That Transformed AI Research - and Possibly the World," *Quartz*, July 26, 2017

[11]↑ John Markoff, "Seeking a Better Way to Find Web Images," *The New York Times*, November 19, 2012

[12]↑ את מאמרם אפשר למצוא כאן: Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* 25, 2 (2012), pp. 1097-1105

[13]↑ WordNet, מסד הנתונים המילוני לשפה האנגלית, פורסם באמצע שנות השמונים. אפשר לראות בו תזאורוס שמגדיר ומקבץ יחדיו מילים באנגלית לכדי סינסטים, כלומר מקבצים של מילים נרדפות. הפרויקט כולו נעשה בתקופה שבה חלו התפתחויות בתחום הבלשנות החישובית ועיבוד שפה טבעית (NLP) - תחום שמתבסס על אלגוריתמים ללמידת מכונה כדי לפתח אמצעים לעיבוד ולניתוח כמויות גדולות של נתוני שפה טבעית.

[14]↑ Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences*, Cambridge, Massachusetts and London: MIT Press, 2000, pp. 44, 107; Anja Bechmann and Geoffrey C. Bowker, "Unsupervised by Any Other Name: Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media," *Big Data & Society* 6, 1 (January 2019)

[15]↑ לניתוח הפוליטיקה של המיון בספריית הקונגרס ראו Sanford Berman, *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People*, Metuchen, NJ: Scarecrow Press, 1971.

[16]↑ אנחנו נסמכים כאן, חלקית לפחות, על עבודתו של לייקוף. ראו George Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Chicago: University of Chicago Press, 2012.

[17]↑ Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, et al., "Imagenet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255

[18]↑ Allan Sekula, "The Body and the Archive," *October* 39 (1986), pp. 3-64

[19]↑ שם. לדיון נרחב יותר בסוגיות של אובייקטיביות, שיפוט מדעי ותפיסה דקה יותר של תפקיד הצילום באלה, ראו Lorraine Daston and Peter Galisoni, *Objectivity*, New York: Zone Books, 2010.

[20]↑ Paul N. Edwards and Gabrielle Hecht, "History and the Technopolitics of Identity: The Case of Apartheid South Africa," *Journal of Southern African Studies* 36, 3(September 2010), pp. 619-39. סיווגים קודמים, שיושמו ב-1950 בחוק רישום האוכלוסין (Population Registration Act) ובחוק האזורים הקבוצתיים (Group Areas Act), קבעו ארבע קטגוריות: "אירופים, אסיאתים, בני גזע מעורב או צבעונים, וילידים" או פרטים טהורי דם מגזע הבאנטו" (Bowker and Star, הערה 14 לעיל, עמ' 197). דרום-אפריקנים שחורים נדרשו לשאת איתם פנקסי מעבר, ולא הורשו למשל לבלות יותר מ-72 שעות באזור לבן בלא אישור ממשלתי להסכם עבודה (שם, עמ' 198).

[21]↑ Star, הערה 14 לעיל, עמ' 208.

[22]↑ Floyd James Davis, *Who Is Black? One Nation's Definition*, University Park, PA: Pennsylvania State University Press, 2001

[23]↑ Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018), pp. 77-91

[24]↑ Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith, "Diversity in Faces," arXiv, January 29, 2019

[25]↑ Olivia Solon, "Facial Recognition's 'Dirty Little Secret': Millions of Online Photos Scraped without Consent," *NBC News*, March 12, 2019

[26]↑ סטיבן ג'יי גולד, **אין מידה לאדם**, בתרגום עמוס כרמל, תל אביב: דביר, 1992. הגישה של מדידת האינטליגנציה על סמך גודל הגולגולת נפוצה ברחבי אירופה וארצות הברית. בצרפת, למשל, פול ברוקה (Broca) וגוסטב לה בון (Le Bon) פיתחו גישה למדידת האינטליגנציה על סמך גודל הגולגולת. ראו Paul Broca, "Sur le crâne de Schiller et sur l'indice cubique des crânes," *Bulletin de la Société d'anthropologie de Paris* 5, 1(1864), pp. 253-260; Gustave Le Bon, *L'homme et les sociétés: Leurs origines et leur développement*, Paris: Edition J. Rothschild, 1881. בגרמניה הנאצית, ה"אנתרופולוגית" אווה יוסטין (Justin) כתבה על בני סינטי ורומה על סמך מדידות אנתרופומטריות ומדידות גולגולת. ראו Eva Justin, "Lebensschicksale artfremd erzogener Zigeunerkinder und ihrer Nachkommen" [Biographical Destinies of Gypsy Children and Their Offspring Who Were Educated in a Manner Inappropriate for Their Species], Ph.D. dissertation, Friedrich-Wilhelms-Universität Berlin, 1943.

[27]↑ Jake Satisky, "A Duke Study Recorded Thousands of Students' Faces. Now They're Being Used All over the World," *The Chronicle*, June 12, 2019

[28]↑ University of Colorado Vision and Security Technology, "2nd Unconstrained Face Detection and Open Set Recognition Challenge"; Russell Stewart, "Brainwash Dataset," Stanford Digital Repository, 2015

[29]↑ Melissa Locker, "Microsoft, Duke, and Stanford Quietly Delete Databases with Millions of Faces," *Fast Company*, June 6, 2019

[30]↑ Madhumita Murgia, "Who's Using Your Face? The Ugly Truth about Facial Recognition," *Financial Times*, April 19, 2019

[31]↑ Locker, הערה 29 לעיל.

[32]↑ Amarjot Singh, "Eye in the Sky: Real-Time Drone Surveillance System (DSS) לסרטון השלם ראו for Violent Individuals Identification" [YouTube video], 2018.

[33]↑ Steven Melendez, "Watch This Drone Use AI to Spot Violence in Crowds from the Sky," *Fast Company*, June 6, 2018; James Vincent, "Drones Taught to Spot Violent Behavior in Crowds Using AI," *The Verge*, June 6, 2018

[34]↑ Vincent, הערה 33 לעיל.

[35]↑ גולד, הערה 26 לעיל, עמ' 149.

קייט קרופורד (Crawford) היא פרופסור באוניברסיטת ניו יורק ומנהלת מכון AI NOW. מחקריה עוסקים בהשלכות החברתיות של מערכות מידע, למידת מכונה ובינה מלאכותית. טרבור פגלן (Paglen) הוא אמן וסופר שעבודתו עוסקת בין היתר במעקב המונים ובאיסוף מידע. מאמר זה פורסם באנגלית בספטמבר 2019 באתר www.excavating.ai
תרגום: יניב פרקש.
