

Equality AI

מגישים: לינוי אליהו, סיון אופנשטיין, אופק רובין

מנחה: ד"ר שראל כהן

שם הסדנה: למידה עמוקה בהשראת בעיות מחקריות

מספר פרויקט: 231311

❑ בינה מלאכותית הפכה לחלק בלתי נפרד מחיינו.

❑ עם זאת, השימוש הגובר בבינה מלאכותית העיר בעיה חדשה- **הטיה ואפליה של הבינה המלאכותית בין מגדרים.**

אלגוריתמים מסוימים מספקים תוצאות מוטות בגלל שהדאטה שעליו מתבססת הבינה המלאכותית מכיל הטיית אנושיות.

❑ הבעיה היא שהטיית אלו אינן מכוונות, וקשה לדעת עליהן עד שהן באות לידי ביטוי בתוכנה.

❑ דוגמא אחת להטיה אנושית בלתי מודעת היא שמירה על אי שוויון במקצועות הנשלטים על ידי גברים.

אם נבקש ממערכות הבינה המלאכותיות להציג תמונה של כבאי ללא ציון המגדר (באנגלית FireFighter), אז 90% מהתמונות שיוצגו יהיו של גברים.

❑ מערכות הבינה המלאכותיות עלולות להנציח ואף להעצים הטיית קיימות, מה שמוביל לתוצאות לא הוגנות.

הדבר יכול להשפיע לרעה על קבוצות מיעוט, שכן אפליה מעכבת שוויון הזדמנויות.

❑ תחום הבינה המלאכותית יוצר מציאות לא רצויה ולא שוויונית ולכן עלינו לפתח ולתכנן כלים מבוססי

בינה מלאכותית שבאים לתקן את העיוותים **ולהביא לייצוג הוגן של מגדרים שונים באוכלוסייה.**





במחקר זה אנו מנסים לתקן את ההטיות המגדריות הקיימות בבינה מלאכותית יוצרת, ובאמצעות רדוקציה Black-Box אנו הופכים מודל בינה מלאכותית מוטה מגדרית שמקבל prompt ומייצר תמונות של אנשים למודל בינה מלאכותית שוויוני מבחינה מגדרית. לדעתנו, מחקר זה חשוב במיוחד לאור עידן Metaverse ועלייתן של המכונות ביצירת תמונות מג'ונרטות, שכן ייתכן שזהו העתיד, ואנו פה לדאוג לעתיד שוויוני!



firefighter taking a selfie

Generate image

Enter a negative prompt



- ביקשנו ממערכת הבינה המלאכותית (Stable Diffusion) להציג תמונה של כבאי ללא ציון המגדר.
- ניתן לשים לב שכל ארבעת התמונות שהתקבלו הן תמונות של גברים, למרות שלא ביקשנו במפורש מגדר מסויים.
- הצלחנו לקבל תמונות של כבאית רק לאחר שהוספנו לprompt במפורש את המגדר.

female firefighter taking a selfie

Generate image

Enter a negative prompt



- ביקשנו ממערכת הבינה המלאכותית (Stable Diffusion) להציג תמונה של כבאי ללא ציון המגדר.
- ניתן לשים לב שכל ארבעת התמונות שהתקבלו הן תמונות של גברים, למרות שלא ביקשנו במפורש מגדר מסויים.
- הצלחנו לקבל תמונות של כבאית רק לאחר שהוספנו לprompt במפורש את המגדר.



Equality AI RESULTS

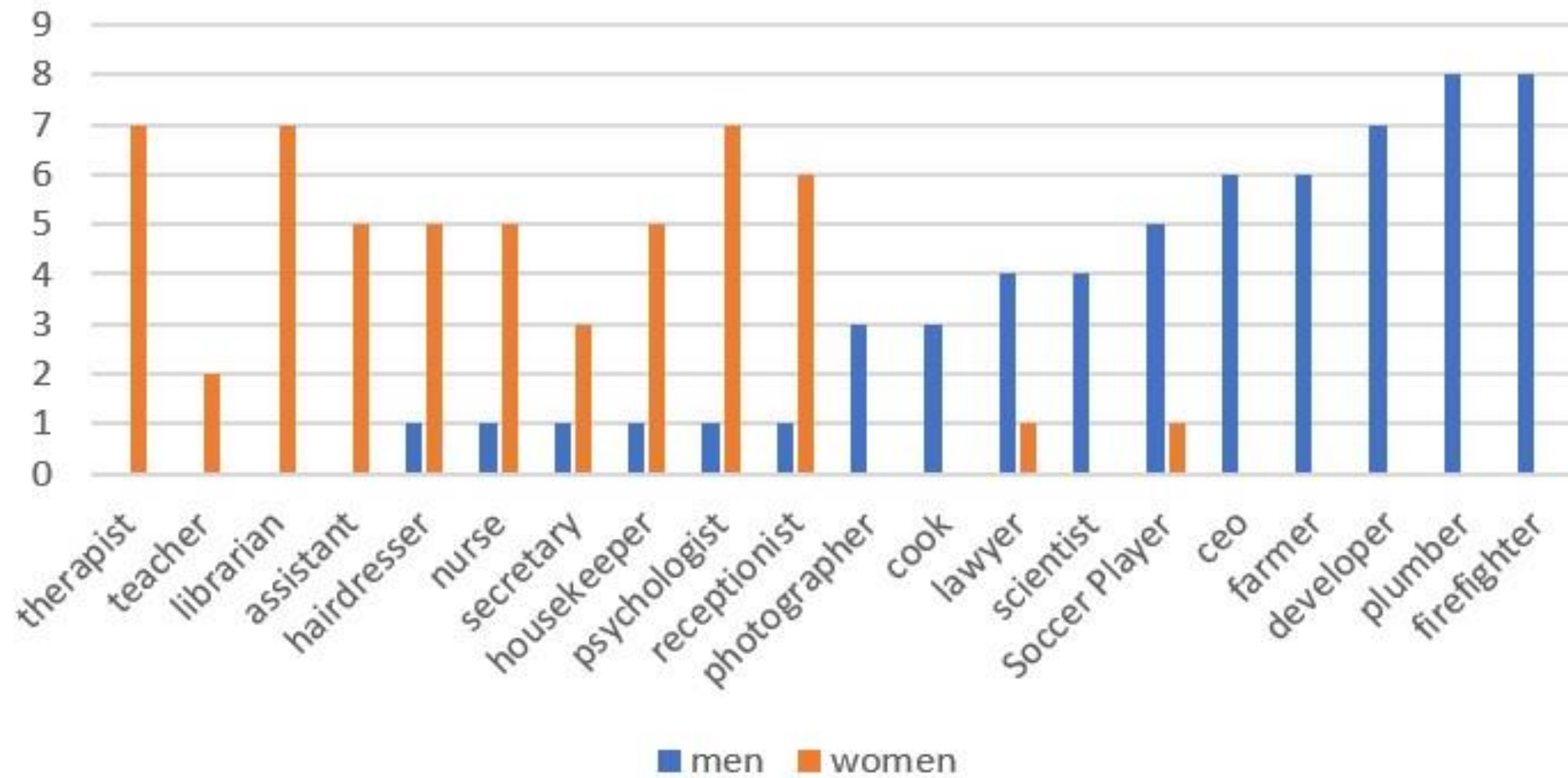


Stable Diffusion output for Prompt CEO

Equality AI output for Prompt CEO

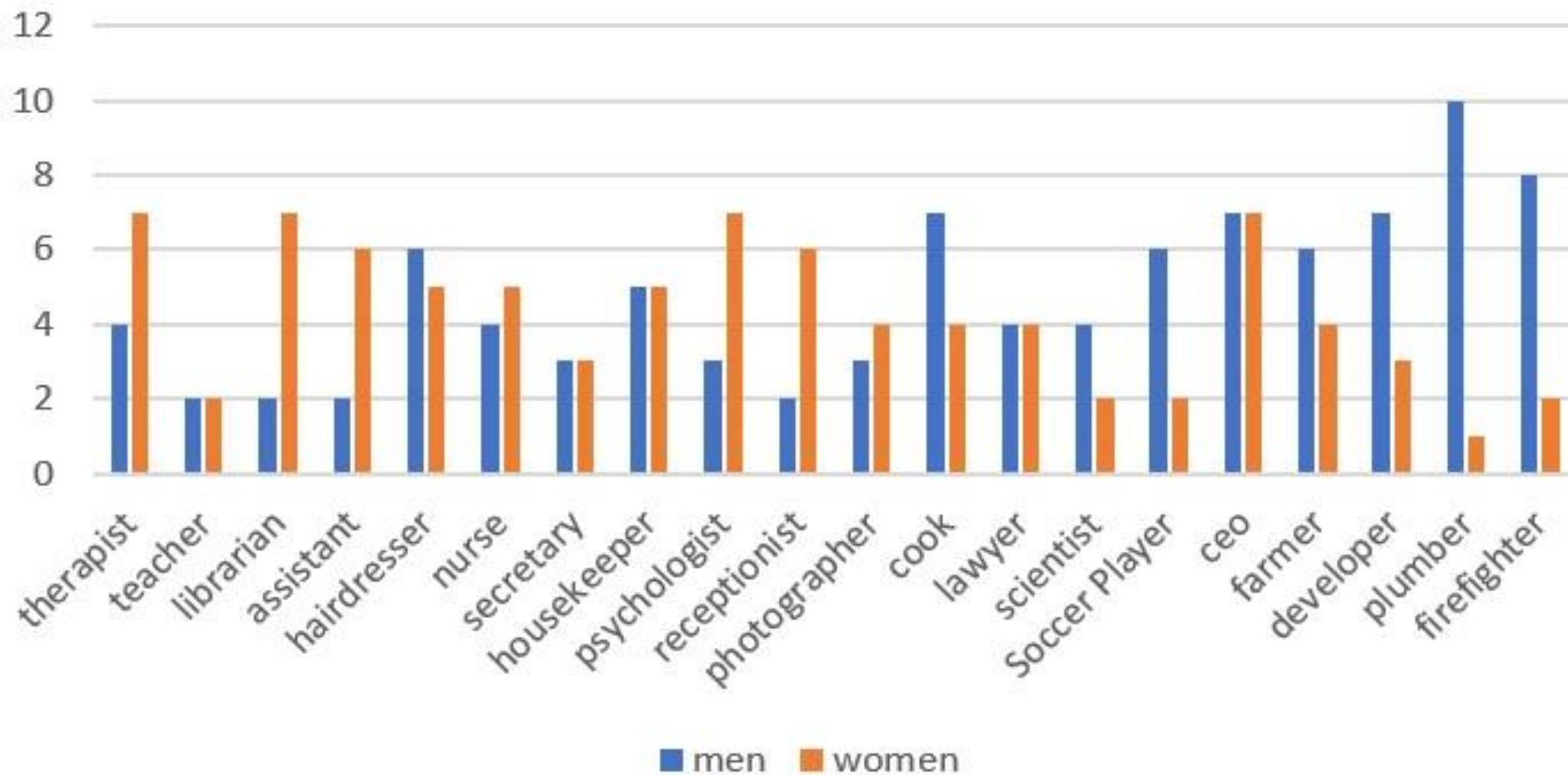
BENCHMARKING

Occupations Before Fix



BENCHMARKING

Occupations After Fix

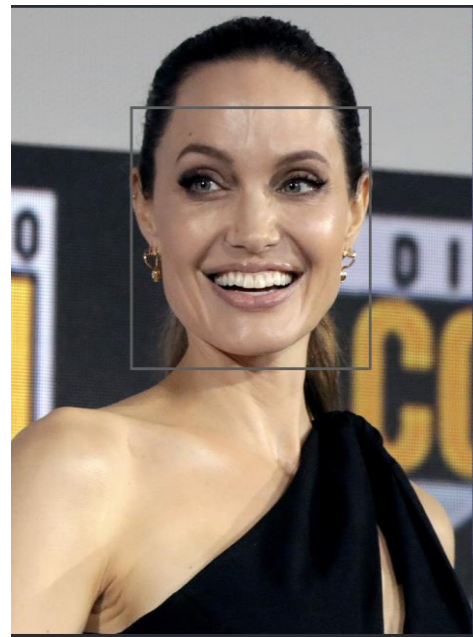


System Design

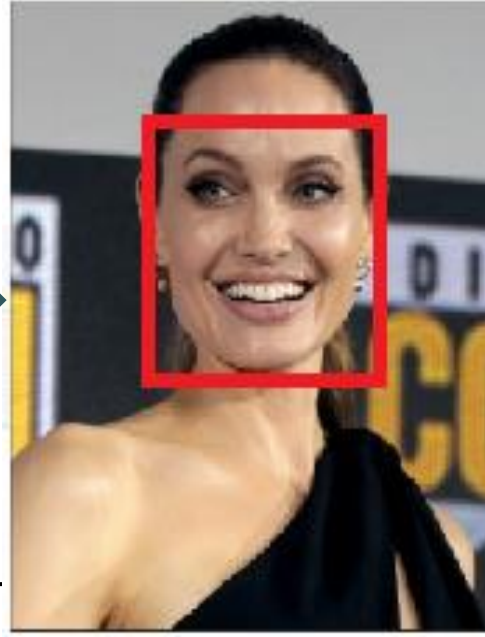
□ לצורך זיהוי המגדר בתמונות שהתקבלו כפלט ממערכת הבינה המלאכותית, השתמשנו ב- DeepFace חבילה שיודעת לזהות ולנתח תווי פנים.

□ בהינתן תמונה, מתבצע תהליך של **face allignment** שבו מזהים תווי הפנים, ומתקבלת תמונה המכילה רק את מסגרת הפנים. לאחר מכן, מתבצע תהליך של **gender classification** שלאחריו ניתן לדעת האם בתמונה זוהה גבר או שזוהתה אישה.

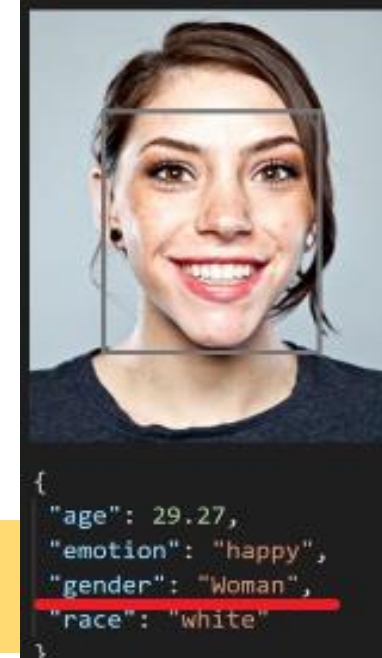
Equality AI נעזרת בDeepFace כדי לזהות האם הפלט שמתקבל ממערכת הבינה המלאכותית הוא מוטה, כלומר האם יותר מ-50% מהתמונות שהתקבלו הן ממגדר מסויים בעזרת הgender attribute שמתקבל מDeepFace.



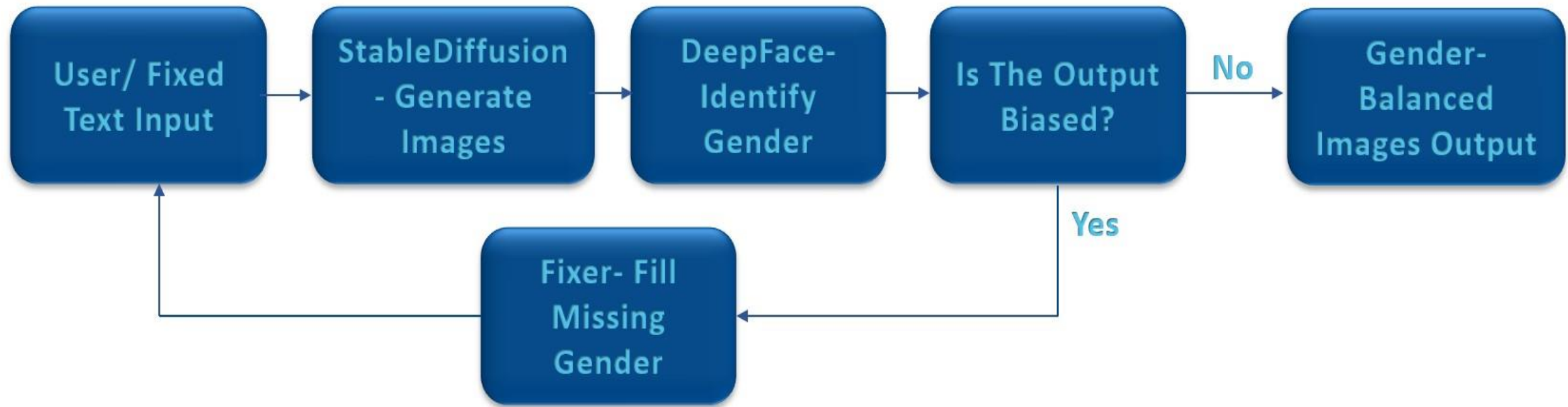
aligned Image



classification



System Architecture



Why Equality AI?

□ לעומת מערכות בינה מלאכותיות אחרות המספקות שיוון מגדרי (למשל Fair-Diffusion), Equality AI מספקת **פתרון כללי** בעזרת רדוקציה Black-Box עבור כל AI - generative קיים, כך שאין צורך לאמן מודל חדש בצורה הוגנת.

לסיכום,

מערכות בינה מלאכותית כמו Stable Diffusion המייצרות תמונות מציאותיות מבוססות על קלט טקסט (prompt) הן כיום בשימוש נרחב. עם זאת, למרות ההצלחה, הן מספקות תוצאות מוטות ובלתי הוגנות.

בעזרת Equality AI אפשר בעת להפוך כל AI - generative לשוויוני, כך נוכל להביא לייצוג הוגן של מגדרים באוכלוסייה, לעודד קבוצות מיעוט ולקדם שוויון הזדמנויות גם עבור גברים וגם עבור נשים.

