סרטון שנוצר על ידי בינה מלאכותית בסכסוך הישראלי-איראני

2026-004 FB-UA

סיכום

בניתוח התפשטות תוכן שנוצר על ידי בינה מלאכותית בסכסוכים מזוינים במקרה של המלחמה בין ישראל לאיראן בשנת 2025, מועצת הפיקוח קוראת ל-Meta לעשות יותר כדי לאפשר למשתמשים לזהות תוכן מסוג זה. הגישה שלה לחשיפת תוכן שנוצר על ידי בינה מלאכותית חייבת להתפתח. זה כולל מתן פרטים בקנה מידה גדול על מקור המדיה, בהתבסס על סטנדרטים של מקור תוכן, השקעה בכלי גילוי חזקים יותר ופיתוח שיטות טובות יותר לתיוג מתאים. Meta צריכה ליצור סט חוקים חדש ונפרד כדי להבטיח כי משתמשים יוכלו לזהות באופן מהימן תוכן שנוצר על ידי בינה מלאכותית. בנוסף, עליה לתקן את המדיניות הנוכחית שלה כדי להבטיח תגובה שניתנת בזמן ובצורה מספקת לתוכן מטעה שנוצר על ידי בינה מלאכותית.

החברה צריכה לעמוד בהתחייבויותיה הציבוריות ולהשתמש בכלים שלה וכלים אחרים הזמינים ברחבי התעשייה כדי להתמודד ביעילות עם תוכן בינה מלאכותית גנרטיבי מטעה המתפשט בין פלטפורמות.

המועצה מבטלת את החלטתה של Meta להשאיר את הפוסט במקרה זה ללא תווית High Risk AI.

למה זה חשוב

ככל שכמות ואיכות התוכן שנוצר על ידי בינה מלאכותית יגדלו, השפעתו על אנשים וחברות תהיה עמוקה. הסיכונים גוברים כאשר תוכן דיפ-פייק שנועד להטעות, לתמרן או להגביר מעורבות משותף במהלך סכסוכים ומשברים, כמו באיראן ובוונצואלה בשנת 2026, ומתפשט במהירות בפלטפורמות של חברות שונות. במהלך שני המשברים הללו, היו טענות כי תוכן מטעה שנוצר על ידי בינה מלאכותית היה אותנטי וכי תוכן אותנטי היה מפוברק. הדבר מחריף את חוסר היכולת של הציבור להבחין באמת, מסמל את תופעת 'דיבידנד השקר', מה שמוביל לחוסר אמון כללי בכל מידע שהוא. קמפיינים להשפעה המונעים על ידי בינה מלאכותית הם אתגר הולך וגובר שנצפה ברחבי העולם בשנים האחרונות, והם מחמירים במערכות אקולוגיות מגבילות של מדיה ואינטרנט המגבילות מידע אמין. עם זאת, ההטעיה שיוצר תוכן שנוצר על ידי בינה מלאכותית אינה כשלעצמה סיבה לגיטימית להגביל את חופש הביטוי. התעשייה זקוקה לעקביות במתן סיוע למשתמשים להבחין בין תוכן מטעה

שנוצר על ידי בינה מלאכותית, ועל הפלטפורמות לטפל בחשבונות ובדפים פוגעניים המשתפים תוכן כזה.

אודות המקרים

מלחמת ישראל-איראן ביוני 2025 סימנה נקודת מפנה, כאשר הנוכחות של תוכן בינה מלאכותית גנרטיבי מטעה ברשתות החברתיות נודע כ"מלחמה רכה" בפני עצמה. דווח כי פלט מטעה שכזה זכה למספר עצום של צפיות, וממשלות ישראל וממשלות איראן הואשמו בניסיונות השפעה המונעים על ידי בינה מלאכותית. ב-15 ביוני, יומיים לאחר תחילת 12 הימים של הסכסוך בין ישראל לאיראן, פורסם סרטון בדף Facebook שטען שהוא מקור חדשות. המשתמש שכתב את הפוסט היה ממוקם בפיליפינים. הסרטון הציג נזק נרחב שנגרם למבנים, עם טקסט באנגלית שכותרתו "בשידור חי עכשיו - נפילתה של חיפה" ("Live now – Haifa Towards Down") ותאריך הפרסום. הסרטון היה דומה מאוד לסרטון שמקורו ב-TikTok וזוהה על ידי בודק עובדות עצמאי (Agence France-Presse) כשקרי וכתוכן הנוצר על ידי בינה מלאכותית. כיתוב על הפוסט ב-Facebook פירט ביטויים רבים בסגנון כותרת הקשורים לסכסוך ומונחים והאשטגים שאינם קשורים. הפוסט זכה ליותר מ-700,000 צפיות, כאשר מספר תגובות ציינו כי התוכן נוצר על ידי בינה מלאכותית.

שישה משתמשים דיווחו על המקרה ל-Meta, אך הוא לא נבדק על ידי החברה וגם לא נבדק על ידי בודקי עובדות של צד שלישי. משתמש הגיש ערעור למועצה. לאחר שהמועצה בחרה במקרה זה, אישרה Meta כי הפוסט לא הפר את תקן קהילת המידע השגוי מכיוון שהוא לא "תרם ישירות לסיכון לפגיעה פיזית קרובה", ולא דרש תווית של בינה מלאכותית.

אותות ברורים של הטעיה הקשורים לפוסט הובילו את המועצה לחקור את Meta לגבי זהות והתנהגות החשבונות המקושרים לדף. לאחר מכן, החברה השביתה שלושה חשבונות המקושרים לדף בגין ניצול לרעה של אינטראקציה וחוסר אותנטיות, והסירה את הדף, ועמו, את תוכן המקרה. הדף היה זכאי להפקת רווחים דרך תוכנית Stars של Meta.

ממצאים עיקריים

המועצה קובעת כי התוכן היווה סיכון מהותי להטעיית הציבור בנושא חשוב בזמן קריטי, ולכן Meta הייתה צריכה להחיל את התווית "High Risk AI". הפוסט לא עמד בדרישות הסף להסרה (מהווה סיכון לפגיעה פיזית או אלימות קרובה). על Meta לנקוט בצעדים נוספים כדי לטפל בהתפשטות

התוכן המטעה שנוצר על ידי בינה מלאכותית בפלטפורמות שלה, כולל על ידי רשתות לא אותנטיות או פוגעניות של חשבונות ודפים, במיוחד בנושאים בעלי עניין ציבורי, כדי שמשתמשים יוכלו להבחין בין מה שאמיתי למה שמזויף.

המועצה מודאגת מדיווחים לפיהם Meta מיישמת באופן לא עקבי תקני הקואליציה למקור ואותנטיות של תוכן (Coalition for Content Provenance and Authenticity - C2PA) אפילו על תוכן שנוצר על ידי כלי הבינה המלאכותית שלה, וכי רק חלק מהתוכן הזה מקבל תיוג מתאים. תקני ה-C2PA קובעים סטנדרטים טכניים להטמעת מידע על מקור כמטא-נתונים בתוכן, מה שמאפשר לפלטפורמות לזהות ביתר קלות תוכן שנוצר ע"י בינה מלאכותית ולהחיל עליו את התוויות המתאימות כדי ליידע את המשתמשים.

המנגנונים הנוכחיים להצמדת אפילו התווית הסטנדרטית של מידע מבוסס בינה מלאכותית לסרטונים (גילוי עצמי של המשתמש או הסלמה לצוות מדיניות התוכן) אינם חזקים ואינם מקיפים מספיק כדי להתמודד עם ההיקף והמהירות של תוכן שנוצר על ידי בינה מלאכותית, במיוחד במהלך משבר או סכסוך בהם יש מעורבות מוגברת בפלטפורמה. מערכת שתלויה יתר על המידה בגילוי עצמי של שימוש בבינה מלאכותית ובבדיקה מואצת (מה שקורה לעתים רחוקות) כדי לתייג כראוי את התוכן הזה, אינה יכולה לעמוד באתגרים הנשקפים בסביבה הנוכחית. בנוסף, חלק מחברי המועצה ציינו כי יש לשלב תוויות High Risk AI (בתוכן שעלול להטעות אנשים בנושאים חשובים) גם עם הורדה בדירוג או הסרה מההמלצות כדי לטפל בחששות מהפצת ההשפעה של תוכן מטעה.

ייתכן שהגישה הצרה של Meta להחלת דירוגים על תוכן זהה וכמעט זהה גרמה לכך שפוסט זה לא קיבל דירוג לבדיקת עובדות. מגבלות משאבים ונפח תפוקה ניכר מקשים על בודקי עובדות להבטיח סקירה בזמן של כל התוכן המטעה, במיוחד במהלך סכסוך או משבר. המועצה חוזרת ומדגישה כי על Meta להבטיח כי בודקי העובדות מקבלים משאבים נאותים והדרכה לגבי סדר עדיפויות של תוכן מסוים. הייעודים של פרוטוקול מדיניות משברים (CPP) ואירועים טרנדיים היו אמורים לאפשר ל-Meta להבטיח תמיכה יעילה יותר בבודקי עובדות של צד שלישי במהלך המשבר. פיזור דירוגים לקטגוריה רחבה יותר של סרטונים דומים מאוד היה יכול להגביל משמעותית את הנזק הפוטנציאלי, כולל על ידי הורדת הדירוג. המקרה מדגיש חוסר יעילות בגישתה הנוכחית של Meta במהלך סכסוכים מזוינים, ומחריף את החששות שהביעה המועצה בעבר.

מדאיג הוא שעם הפעלת ה-CPP והקצאת משאבים נוספיMeta , לא זיהתה ביוזמתה את הסימנים הברורים לניצול לרעה של האינטראקציה מהדף, וכי היא חקרה את החשבונות שמאחוריו רק בתגובה לשאלות המועצה. אכיפה מדויקת של המדיניות המבוססת על התנהגות הייתה יכולה למנוע את הנזקים שנגרמו מחשבונות מפרים אלה, במקום להסתמך על אמצעי הפחתה מבוססי תוכן במורד הזרם, הנוטים לשיעור כישלון גבוה.

החלטת המועצה לפיקוח

המועצה מבטלת את החלטתה של Meta להשאיר את התוכן ללא תווית של High Risk AI.

המועצה ממליצה כי Meta:

- יש ליצור תקן קהילתי לתוכן שנוצר על ידי בינה מלאכותית, נפרד מתקן הקהילה למידע שגוי, המספק כללים מקיפים בנוגע לשימור מקור, פרוטוקולי תיוג בינה מלאכותית וגילוי עצמי.
- יש לפתח מסלולים להצמדת תוויות של High Risk ו-High Risk AI לתוכן בתדירות גבוהה הרבה יותר, בסיוע ערוצי הסלמה ברורים יותר ממערכות אוטומטיות וביקורת בקנה מידה גדול, כך שתיוג כזה יוכל להתרחש בנפח גבוה משמעותית.
- יש לצרף מידע על מקור וסימני מים בלתי נראים לתוכן שנוצר על ידי כלי בינה מלאכותית של Meta, כולל יישום אישורי תוכן (כפי שנקבע על ידי תקני ה-C2PA) בעת היצירה.
- יש להטמיע אישורי תוכן בקנה מידה גדול וודא שהם גלויים ועקביים ונגישים באופן ברור ועקבי בכל פעם שפרטי המקור זמינים.
- יש להשקיע בכלי זיהוי חזקים יותר עבור תוכן רב-פורמטי (אודיו, אודיו-ויזואלי ותמונה) שנוצר על ידי בינה מלאכותית.
- יש לפרסם הסבר ברור לגבי העונשים על אי גילוי עצמי של תוכן שנוצר או נערך באופן דיגיטלי, כולל הקריטריונים לעונשים והמגבלות הנובעות מכך.
- יש לתקן את תקן הקהילה למידע שגוי כדי להבטיח שבדיקה מהירה של מידע שגוי המסכן ישירות אלימות או פגיעה פיזית מיידית לא תהיה תלויה אך ורק באותות משותפים חיצוניים. מנוף CPP צריך להקצות משאבים לגילוי בזמן ויזום של תוכן מפר חוק כזה, נתמך על ידי מומחיות ופעולות פנימיות, כולל תיוג וחקירת חשבונות ודפי פרסום.

* סיכומי מקרים מספקים סקירה כללית של המקרים ואינם יכולים לשמש כתקדים.

ההחלטה במלואה בנוגע למקרה

1. תיאור המקרה והרקע למקרה

ב-13 ביוני ,2025 ישראל השיקה תקיפה אווירית גדולה המכוונות נגד מתקני גרעין וצבא איראניים, בין היתר. מנהיגי ישראל אמרו כי המתקפות נועדו למנוע את התפתחות תוכנית הגרעין של איראן. מצב זה גרם לחילופי התקפות עזות בין שתי המדינות במשך יותר משבוע וחצי. איראן שיגרה מאות טילים לעבר ערים ישראליות, רבים מהם יורטו על ידי מערכת ההגנה הישראלית, בעוד שישראל תקפה מספר אתרים ברחבי איראן, כולל סמוך לבירה טהראן. ב-18 ביוני, מזכ"ל האומות המאוחדות (או"ם) אנטוניו גוטרש הצהיר כי הוא היה "מודאג מאוד" מההסלמה הצבאית, והוסיף כי "כל התערבות צבאית נוספת עלולה להיות בעלת השלכות עצומות, לא רק על המעורבים אלא על האזור כולו, ועל שלום וביטחון בינלאומיים בכללותם." ב-21 ביוני, ארצות הברית ביצעה סדרה של תקיפות על אתרים גרעיניים באיראן. ב-24 ביוני הוכרזה הפסקת אש בין ישראל לאיראן.

הסכסוך בין ישראל לאיראן בשנת 2025 סימן נקודת מפנה, וההשפעה הגוברת של תוכן המיוצר ע"י בינה מלאכותית ברשתות החברתיות כונתה "מלחמה רכה" בפני עצמה. תאגיד השידור הבריטי דיווח כי שלושה סרטונים מטעים שנוצרו על ידי בינה מלאכותית של הסכסוך צברו מעל 100 מיליון צפיות. שר החוץ של ישראל שיתף סרטון של מתקפה על כלא אווין בטהרן, אשר ניתוח פורנזי קבע מאוחר יותר כי מדובר בסבירות גבוהה בסרטון שנוצר על ידי בינה מלאכותית, למרות שהתרחשה בפועל התקפה על הכלא. מחקר של מעבדת האזרחים באוניברסיטת טורונטו דיווח על רשת מתואמת של פרופילים לא אותנטיים ב-X (לשעבר Twitter), שלכאורה קשורים לישראל, המעודדים את האיראנים להתנגד לממשלתם. ממשלת ישראל דיווחה גם על קמפיינים מונעי-בוטים של איראן שמטרתם לעצב דעות סביב הסכסוך והשפעת התקפותיהם על ישראל.

תוכן מטעה שנוצר על ידי בינה מלאכותית נהיה אתגר הולך וגובר ומתמשך במשברים ובסכסוכים ברחבי העולם בשנים האחרונות. אתגר זה מחריף במקומות שבהם חופש הביטוי נמצא תחת לחץ ודיכוי של כלי תקשורת עצמאיים וסגירת אינטרנט חוסמים מידע אמין שיכול להפריך קמפיינים מטעים. במהלך דיוניה במקרה זה, המועצה בחנה כיצד המבצע האמריקאי ללכידת נשיא ונצואלה והמחאות ההמוניות נגד הממשלה באיראן כללו טענות לפיהן תוכן מטעה שנוצר על ידי בינה מלאכותית היה אמיתי וטענות נגדיות לפיהן תוכן אמיתי היה מפוברק. שני המצבים אתגרו את יכולתו של הציבור להבחין בין בדיה לעובדה, דבר המסמל את תופעת 'דיבידנד השקר', תוך סיכון לחוסר אמון כללי בכל מידע שהוא.

מספר גישות טכניות צצו כדי לסייע לפלטפורמות ולמשתמשים להבחין בין תוכן מדיה סינתטי או מניפולטיבי לבין תוכן מדיה אותנטי. גישה אחת היא מעקב אחר המקור, כלומר, ההיסטוריה הניתנת לאימות של נכס דיגיטלי, כגון תמונה, סרטון או מסמך. הקואליציה למקור ואותנטיות של תוכן (C2PA Coalition for Content Provenance and Authenticity) - קבעה סטנדרטים טכניים להטמעת מידע על מקור התוכן כמטא-דאטה בתוכן, מה שמאפשר לפלטפורמות לזהות ביתר קלות תוכן שנוצר על ידי בינה מלאכותית ולהחיל תוויות כדי ליידע את המשתמשים. בעוד שכלים אלה עדיין מתפתחים ואף אחד מהם אינו פתרון מושלם, הם נתפסים כ"רצפת הקרקע בכל הנוגע ליצירה והפצה אחראית של תוכן שנוצר על ידי בינה מלאכותית". במקביל, השקעות בזיהוי אוטומטי, כגון מסווגים, עשויות לספק נתיב לגילוי אותות אחרים לכך שתוכן נוצר על ידי בינה מלאכותית.

מקרה זה מסמל את האתגרים הללו. ב-15 ביוני, כאשר הסכסוך הישראלי-איראני הסלים, פורסם סרטון בדף Facebook שטען שהוא מקור חדשותי וצבר 161,000 עוקבים. הסרטון הציג נזק נרחב שנגרם למבנים, מוקפים בעשן והריסות, עם טקסט באנגלית שכותרתו "בשידור חי עכשיו - נפילתה של חיפה" ("Live now – Haifa Towards Down") ותאריך הפרסום. ככל הנראה, הפוסט התייחס לחיפה, עיר בצפון ישראל. נראה שהסרטון דומה מאוד לסרטון שמקורו ב-TikTok וזוהה על ידי בודקי עובדות עצמאיים של Agence France-Presse (AFP) כשקרי וכתוצאה מכך נוצר על ידי בינה מלאכותית (AFP מספקת רק תמונות סטילס של הסרטון המדורג, אך אלה זהות לפריימים בתוכן במקרה זה). כיתוב באנגלית לפוסט ב-Facebook כלל ביטויים רבים בסגנון כותרת הקשורים לסכסוך, כמו גם מונחים והאשטגים שאינם קשורים, ללא נרטיב ברור. הוא הזכיר סכסוך מתמשך, מנהיגים פוליטיים עולמיים, שריפות יער, טילים ועוד. הפוסט זכה ליותר מ-700,000 צפיות, כאשר מספר תגובות ציינו כי התוכן נוצר על ידי בינה מלאכותית.

שישה משתמשים דיווחו על התוכן במקרה תשע פעמים בסך הכול, אך המערכות האוטומטיות של Meta לא תעדפו אותו לבדיקה אנושית. באותו יום בו פורסם התוכן, מסווג מידע שגוי העביר אותו לתור לבודקי עובדות של צד שלישי, אך הוא מעולם לא נבדק ולא דורג. מקרה זה אינו מהווה דבר מה חריג, שכן Meta מסמנת כמות משמעותית של מידע שגוי פוטנציאלי שחורגת מיכולתם של בודקי העובדות לבחון.

לאחר מיצוי הליכי ערעור פנימיים בתוך החברה, אחד מהמשתמשים המדווחים הגיש ערעור על החלטתה של Meta להשאיר את התוכן לידי המועצה. לאחר שהמועצה בחרה במקרה זה, אישרה Meta כי הפוסט לא הפר את תקן קהילת המידע השגוי מכיוון שהוא לא "תרם ישירות לסיכון לפגיעה פיזית ממשית". Meta גם הגנה על החלטתה לא לתייג את התוכן. היא לא נקטה פעולה נגד הדף או החשבונות האחראים לתוכן.

עקב סימנים ברורים של הטעיה סביב פוסט זה, המועצה שאלה את Meta סדרה של שאלות בנוגע לזהות והתנהגות הדף והחשבונות שמאחוריו. הדבר הוביל לחקירה, שהובילה את Meta לזהות שמנהלי הדף הפרו כללים בנוגע לרעה באינטראקציה וחוסר אותנטיות. לאחר מכן החברה השביתה לצמיתות שלושה חשבונות, מה שהסיר את הדף ואת תוכן המקרה מהפלטפורמה.

2. פניות ממשתמשים

בהגשתו למועצה, המשתמש שביקש להסיר את התוכן התלונן כי Meta מאפשרת "פעולות טרור" בפלטפורמה שלה. לא הייתה כל אינדיקציה ברורה בהצהרתם לכך שהם הבינו שהתוכן נוצר על ידי בינה מלאכותית או הינו מטעה.

3. מדיניות התוכן והגשות של Meta

א. מדיניות התוכן של Meta

תקן קהילתי למידע שגוי

תחת התקן הקהילתי למידע שגויMeta , מסירה "מידע שגוי או שמועות שלא ניתנות לאימות ששותפים מומחים קבעו כי עשויות לתרום ישירות לסיכון לאלימות או פגיעה פיזית ממשית באנשים". במדינות "החוות סיכון מוגבר לאלימות חברתית"Meta , פועלת "באופן יזום עם שותפים מקומיים כדי להבין אילו טענות שקריות עשויות לתרום ישירות לסיכון לפגיעה פיזית ממשית" כדי לזהות ולהסיר תוכן המעלה טענות אלה.

תחת כותרת המשנה "מדיה מניפולטיבית"Meta , מציינת כי עבור תוכן שאינו מפר את תקני הקהילה, היא רשאית להוסיף תווית אינפורמטיבית לפוסט כאשר מדובר בתמונה או בסרטון פוטוריאליסטי, או באודיו שנשמע ריאליסטי, שנוצרו או שונו באופן דיגיטלי ויוצרים "סיכון גבוה במיוחד להטעיה מהותית של הציבור בנושא בעל חשיבות ציבורית". מדיניות המידע השגוי דורשת גם ממשתמשים לחשוף בכל פעם שהם מפרסמים "תוכן אורגני עם וידאו פוטוראליסטי או אודיו שנשמע ריאליסטי שנוצר או שונה דיגיטלית". אי שימוש בכלי גילוי הבינה המלאכותית עלול לגרום לעונשים.

במקום אחרMeta , גם אוסרת על תוכן והתנהגות אשר "לעתים קרובות חופפים להפצת מידע שגוי". זה כולל סטנדרטים קהילתיים בנושא שלמות החשבון ,שיטות מטעות והתנהגות לא אותנטית מתואמת. עבור כל מידע שגוי אחר שאינו מפר את תקן קהילת המידע השגוי שלMeta , מתמקדת ב"הפחתת שכיחותו או יצירת סביבה המטפחת דיאלוג פרודוקטיבי". מחוץ לארה"ב , Meta מסתמכת על בודקי עובדות עצמאיים של צד שלישי כדי לבדוק ולדרג תוכן, מה שיכול להוביל

להוספת תוויות התואמות לדירוג לתוכן. דירוגים כוללים "שגוי" ו"עבר שינוי" ויכולים להוביל להפצה מופחתת של תוכן. Meta משתמשת בטכנולוגיה כדי לחשוף מידע שגוי פוטנציאלי עבור בודקי עובדות לבדיקה, ובודקי עובדות יכולים גם לזהות תוכן משלהם לבדיקה. בינואר 2025, Meta הודיעה שהיא מסיימת את תוכנית בדיקת העובדות של צד שלישי בארה"ב ועוברת במקום זאת למודל של 'הערות מהקהילה'.

מדיניות המונטיזציה לשותפים של Meta מתארת את הכללים עבור דפים "המרוויחים כסף בפלטפורמות" ומציינת כי תוכן שסומן כמידע שגוי או קליקבייט עשוי לא להיות זכאי למונטיזציה. מדיניות המונטיזציה של תוכן מתארת בפירוט רב אף יותר את הכללים ל"יצירת תוכן בטוח למותג וניתן להפקת רווחים" ומגבילה או מפחיתה את המונטיזציה על תוכן המתאר או מכיל נושאים מסוימים, כגון "טרגדיה וסכסוך, כולל נזק לרכוש". בהקשר זה, החשבונות המקושרים לדף במקרה זה הוסרו שניהם עקב הפרות ברמת החשבון בנוגע לניצול לרעה של מעורבות.

2. ההגשות של Meta

Meta הצהירה כי הפוסט לא הפר את מדיניות המידע השגוי המחייבת הסרת תוכן "שעשוי לתרום ישירות לסיכון לפגיעה פיזית ממשית". החלטתם התחשבה בכך כי אף מומחה עצמאי, כגון שותף מקומי, לא סימן בפניהם את התוכן או כל מגמה קשורה של מידע שגוי.

Meta לא ייחסה שום תווית לתוכן במסגרת כללי המדיה שעברו מניפולציה. Meta מיישמת שלוש תוויות שונות למדיה מניפולטיבית: AI Info, High Risk, או High Risk AI.

התווית **מידע על בינה מלאכותית (AI Info)** מוחלת אוטומטית על תוכן כאשר Meta מזהה "אינדיקטורים לתמונה הנוצרה ע"י בינה מלאכותית בתעשייה או כאשר אנשים גילו שהם מעלים תוכן שנוצר על ידי בינה מלאכותית". כפי שחשפה המועצה במקרה של זיוף בחירות לכאורה בכורדיסטן העיראקית, Meta מסוגלת כיום לזהות באופן אוטומטי ולשים את התווית AI Info על תמונות סטטיות רק, בהסתמך על מטא-דאטה שכלי בינה מלאכותית יצירתיים רבים מטמיעים בתוכן כזה. תוכן אודיו או וידאו דורש גילוי עצמי מצד המשתמשים כדי שהתווית תוחל. לא היהי כל גילוי עצמי במקרה הזה. תחת התהליכים הנוכחיים של Meta, לא ניתן היה להוסיף תווית באופן אוטומטי בנסיבות אלה.

התווית **סיכון גבוה (High Risk)** חלה על תוכן אשר (א) יוצר סיכון גבוה במיוחד להטעיה מהותית של הציבור בנושא בעל חשיבות ציבורית; ו(ב) הינו בעל אינדיקטורים אמינים לכך שהוא נוצר או שונה דיגיטלית. שלא כמו תווית AI Info, התווית High Risk היא מדיניות להסלמה בלבד, כלומר רק צוותי המדיניות הפנימיים של Meta יכולים להחיל את התווית לאחר בדיקה אנושית. לא הייתה הסלמה כזו של התוכן המדובר לפני בחירת המועצה במקרה זה.

התווית בינה מלאכותית בסיכון גבוה (High Risk AI) מוחלת כאשר תוכן עומד בכל הדרישות של התווית High Risk ויש לו אינדיקטורים אמינים לכך שנוצר או שונה באמצעות בינה מלאכותית. זוהי גם מדיניות של הסלמה בלבד, הדורשת בדיקה אנושית. צוותי המדיניות הפנימיים של Meta בחנו תוכן זה רק לאחר שהמועצה בחרה אותו. הם קבעו שחלף זמן רב מדי עד אז מכדי שהתווית תהיה רלוונטית או דחופה.

Meta בוחנת מקורות חיצוניים ופנימיים זמינים שהיא רואה בהם אמינים בעת הקביעה האם התוכן הוא תוצר של בינה מלאכותית, נוצר או שונה דיגיטלית. מקורות חיצוניים עשויים לכלול חדשות או גופי בדיקת עובדות עצמאיים של צד שלישי המסוגלים לספק בסיס טכני לקביעתם, כגון הפניה למודל גילוי מבוסס בינה מלאכותית או מסקנה של מומחה משפטי.

כפי שהוסבר בהחלטה במקרה של קטעי מחאה משולבים עם קריאות פרו-דוטרטה, תוויות תקשורת מניפולטיבית אינן גורמות להורדה אוטומטית של תוכן בדירוג או להסרתו מהמלצות. במקום זאת, משתמשים אשר משתפים מחדש תוכן עם תוויות אלה עשויים לקבל חלון קופץ, מה שעשוי להפחית באופן אורגני את טווח ההגעה. המסר של החלון הקופץ יהיה תלוי בשאלה האם התוכן נוצר באמצעות בינה מלאכותית או לא.

ל-Meta יש מספר מערכות מחוץ לארה"ב לזיהוי ולטיפול במידע שגוי פוטנציאלי. חלק מהמערכות שולחות תוכן רק לבדיקת עובדות על ידי צד שלישי, בעוד שאחרות שולחות את התוכן לבדיקה וגם מיישמות הורדה זמנית בדרגה בזמן ההמתנה לבדיקה. האם תוכן נשלח רק לבדיקה או שההשפעה שלו מצטמצמת תלויה בסוג המערכת שמסמנת אותו, כמו גם בגורמים כמו מדינה ושפה.

לפי ,Meta פרוטוקול מדיניות המשברים (CPP)הופעל בזמן הסכסוך הישראלי-איראני. ישראל הייתה מוגדרת תחת ה-CPP מאז מתקפות ה-7 באוקטובר2023 , בארץ. איראן הוגשה כאזור בתחילת הסכסוך ביוני 2025. הפעלת הפרוטוקול מאפשרת לחברה לפרוס מערך של מנופים שנועדו לחזק את תגובת המשברים שלה ולאפשר לצוותים שלה להעריך ולמתן את הסיכון לפגיעה ממשית. במהלך המשבר הזה, זה לא הוביל לשינויים כלשהם במערכות הניהול האוטומטי, ותוכן המקרה נבדק באמצעות מודלים וספים קיימים.

Meta הגדירה את הסכסוך הישראלי-איראני כ"אירוע טרנדי" כדי לתמוך טוב יותר בבודקי עובדות של צד שלישי בזיהוי והפרכת טענות שווא ויראליות הקשורות לסכסוך, בהתחשב בסיכון הגבוה להפצת מידע שגוי. עד שהוכרזה הפסקת אש, בודקי עובדות פרסמו מספר בדיקות עובדות הקשורות לסכסוך. Meta מצהירה שיש לה בודקי עובדות בישראל ובפיליפינים (שם נמצא המשתמש המפרסם), אך לא באיראן.

לאחר שהמועצה ביקשה מ-Meta לחקור את התנהגות הדף הזה והחשבונות המקושרים, החברה השביתה את החשבונות של שלושה מנהלי דפים שונים.

מנהל אחד הושבת בגין הפרת מדיניות ייצוג זהות אותנטית על ידי "עיסוק בהצגת זהות שווא כדי להטעות או לרמות אחרים, להתחמק מאכיפה או להפר את תקני הקהילה שלנו". החשבון השני הושבת במסגרת מדיניות שלמות החשבון מכיוון שהוא בבעלות אותו אדם/ישות כמו חשבון מושבת קיים. השלישי הושבת במסגרת מדיניות הספאם עקב ניצול לרעה של המעורבות. מדיניות הספאם אוסרת באופן כללי על שיטות שונות שהינן מטעות, מתעתעות או מציפות כשמטרתן להגביר באופן לא אותנטי את המעורבות בפוסטים. כתוצאה מכך, הדף והתוכן שפורסם בו מוחרם. לפני הסרתו, הדף היה זכאי להפקת רווחים דרך תוכנית Stars של Meta.

המועצה שאלה שאלות בנושא תיוג, גילוי בינה מלאכותית ,בדיקת עובדות, יישום ,CPP התנהגויות ברמת הדף והחשבון ועוד. Meta ענתה על כל השאלות.

4. תגובות הציבור

המועצה קיבלה שש הערות מהציבור שעמדו בתנאי ההגשה. ארבע הערות הוגשו מאירופה ושתיים מארה"ב. לקריאת הערות הציבור שהוגשו בהסכמה לפרסום, לחצו כאן.

הפניות עסקו בנושאים הבאים: ניהול תוכן במהלך סכסוך ומשבר, שכיחות התוכן שנוצר על ידי בינה מלאכותית ועליית ההתנהגות המתואמת והלא אותנטית בסכסוך מזוין, מגבלות ההגדרה של "נזק פיזי ממשי" של ,Meta חשיבות בדיקת העובדות במהלך סכסוך מזוין, הסטנדרטים והיישום של תוויות מדיה מניפולטיביות, חשיבותם של סטנדרטים של C2PA בגילוי ועוד.

5. ניתוח המועצה לפיקוח

המועצה בחרה במקרה זה כדי לבחון את המדיניות ונהלי האכיפה של Meta בנוגע לשיתוף תוכן מטעה שנוצר על ידי בינה מלאכותית בפלטפורמות שלה, במיוחד בהקשר של סכסוכים מזוינים. מקרה זה נופל תחת סדרי העדיפויות האסטרטגיים של המועצה למצבי משבר וסכסוך, בינה מלאכותית ואוטומציה.

המועצה ניתח את החלטתה של Meta במקרה זה אל מול מדיניות התוכן, הערכים ואחריותה של Meta בתחום זכויות האדם. המועצה גם העריכה את ההשלכות של המקרה לגבי הגישה הכוללת של Meta לניהול תוכן.

5.1 עמידהבמדיניותהתוכן של Meta

כללי התוכן

המועצה קובעת כי תוכן המקרה לא נדרש להסירו במסגרת תקן קהילת המידע השגוי, אך Meta הייתה צריכה להחיל תווית "High Risk AI" על התוכן במסגרת הכללים שלה בנוגע למדיה מניפולטיבית.

הפוסט לא הפר את מדיניות המידע השגוי של Meta המחייבת הסרה, מכיוון שלא היה סביר שהוא יתרום לאלימות קרובה או פגיעה פיזית באנשים. הסרטון הגזים באופן מטעה בהשפעת התקיפה האיראנית על ישראל ופורסם בסמוך לנפילת טילים על חיפה. בעוד שסביר להניח שזה הוסיף למצוקתם של אלו שהוטעו, לא היה זה שתרום לאלימות מצד אזרחים ישראלים נגד יריבים נתפסים או שישפיע ישירות על תגובת ממשלת ישראל. אין כל אינדיקציה לאלימות בין קהילות בתוך ישראל בתגובה לתיאורים מטעים של המתקפות.

למרות מסקנה זו, מדאיג הוא ש-Meta לא סיפקה כל ניתוח של הסיכון הפוטנציאלי לנזק בניתוח התוכן. במקום זאת, היא הגיעה למסקנה כי מכיוון שאף שותף מהימן לא סימן בפניהם את התוכן, התוכן לא הפר את הכללים שלו. זו אינה עמדה מקובלת כאשר שותפים מהימנים רבים מודיעים למועצה שהחברה פחות מגיבה לפניות ולחששות, בין היתר עקב צמצום משמעותי ביכולות של הצוותים הפנימיים של Meta. Meta צריכה להיות מסוגלת לבצע הערכות נזק כאלה בעצמה, במקום להסתמך אך ורק על שותפים הפונים אליהם במהלך סכסוך מזוין. זה היה עמוד בעל מעקב נרחב, התוכן היה ויראלי, מסווג המידע השגוי של Metaסימן את התוכן ומספר משתמשים דיווחו עליו. מקורות אמינים כמו AFP הפריכו סרטון דומה מאוד, והודיעו ל-Meta לבחון באופן יזום טענות מטעות שיכלו לגרום נזק. הפעלת ה-CPP הייתה אמורה להבטיח שהמשאבים הדרושים יהיו זמינים עבור Meta כדי לבצע הערכות מסוג זה בעצמה ולפנות באופן יזום לשותפים לקבלת הקשר בשטח במידת הצורך. בנסיבות אלה, היה צריך להעביר את התוכן לבדיקה.

אילו זה היה קורה, היה ברור שהתוכן מהווה סיכון מהותי להטעיית הציבור בנושא חשוב בנקודת זמן קריטית, ואז היה מוחל על התווית "High Risk AI". הדף המציג את עצמו ככלי חדשות לצד טקסט מעל הסרטון שטען שהוא "בשידור חי עכשיו" וצולם בחיפה, הציג את התוכן הזה כצילומים אמיתיים של סכסוך מזוין מתמשך שבו חיי אזרחים נמצאים בסכנה. משתמש זה, שדף האינטרנט שלו כבר הציג את עצמו באופן שגוי כמקור חדשותי אמין, לא היה חושף את הטעייתו על ידי גילוי עצמי של שימוש בבינה מלאכותית. הציפורים הלבנות הלא מציאותיות שעפות בסרטון, וארגוני מומחים כמו AFP שגילו שהתוכן נוצר על ידי בינה מלאכותית, היו צריכים לגרום ל-Meta להעריך האם היה צורך בתווית. Meta הייתה צריכה גם לתקן את טעותה ברגע שהמועצה הביאה אותה

לידיעתה. מחקר של המועצה בפלטפורמה חשף מספר גרסאות של סרטון זה וצילומי מסך קשורים שהופצו בפלטפורמות של Meta שבועות לאחר מכן, והחברה לא תייגה אף אחת מהן.

אילו פוסט ספציפי זה היה נבדק גם על ידי בודקי עובדות של צד שלישי, סביר להניח שהתוכן היה מקבל תווית כוזבת ומורד בדירוג (בהתבסס על דירוג AFP לסרטון דומה מאוד). ייתכן שהגישה הצרה של Meta לפריסת דירוגים לתוכן זהה וכמעט זהה גרמה לכך שתוכן זה לא דורג גם כן (למשל, בגלל הוספת שכבת טקסט לסרטון). מגבלות משאבים וכן נפח תוכן ניכר מקשים על בודקי עובדות להבטיח סקירה בזמן של כל התוכן המטעה, במיוחד במהלך סכסוך או משבר. המועצה חוזרת ומדגישה כי על Meta להבטיח כי בודקי העובדות מקבלים משאבים נאותים והדרכה לגבי סדר עדיפויות של תוכן מסוכסך, על מנת לבצע את העבודה המאתגרת ש-Meta סומכת עליהם (ראו קטעי מחאה משולבים עם קריאות פרו-דוטרטה).

מדאיג שבמהלך משבר זה, לאחר הפעלת ה-CPP והקצאת משאבים נוספיMeta , לא זיהתה ביוזמתה את הסימנים הברורים לניצול לרעה של האינטראקציה מהדף, וכי היא חקרה את החשבונות שמאחוריה רק בתגובה לשאלות המועצה. אכיפה מדויקת של מדיניות היושרה והאותנטיות המבוססת על התנהגות הייתה יכולה למנוע את הנזקים שנגרמו מתוכן מטעה זה, המגובה על ידי מספר חשבונות מפרים, במקום להסתמך על אמצעי הפחתה מבוססי תוכן במורד הזרם, הנוטים לשיעור כישלון גבוה.

5.2 עמידה באחריותה שלMeta בנוגע לזכויות אדם

המועצה קובעת כי, במסגרת אחריותה של Meta בתחום זכויות האדם, היה צריך להחיל על התוכן את התווית "High Risk AI" שעברה מניפולציה, ועל Meta לעשות יותר כדי לטפל בהתפשטות התוכן המטעה שנוצר על ידי בינה מלאכותית בפלטפורמות שלה, לרבות על ידי רשתות לא אותנטיות או פוגעניות של חשבונות ודפים.

חופש הביטוי (סעיף 19 לאמנה הבין-לאומית בדבר זכויות אזרחיות ומדיניות)

קטע 19 לאמנה הבינלאומית בדבר זכויות אזרחיות ופוליטיות (ICCPR) קובע הגנה רחבה על ביטויים, כולל ביטוי פוליטי. זכות זו כוללת את "החופש לחפש, לקבל ולמסור מידע ורעיונות מכל הסוגים" (קטע 19, פסקה 2). כאשר מדינה מטילה הגבלות על הביטוי, על הגבלות אלה לעמוד בדרישות של חוקיות, מטרה לגיטימית ונחיצות ומידתיות (סעיף 19, פסקה 3, האמנה הבין-לאומית בדבר זכויות אזרחיות ומדיניות). דרישות אלה מכונות לעתים קרובות "המבחן בן שלושה חלקים". המועצה משתמשת במסגרת זו כדי לפרש את אחריותה של Meta לזכויות האדם בהתאם לעקרונות המנחים של האו"ם בנושא עסקים וזכויות אדם, אשר Meta עצמה התחייבה אליהם

במדיניות זכויות האדם התאגידית שלה. המועצה עושה זאת הן ביחס להחלטת התוכן הנבדקת והן ביחס למה שזה אומר על הגישה הרחבה יותר של Meta לניהול תוכן. כפי שקבע הדווח המיוחד של האו"ם בנושא חופש הביטוי, למרות ש"לחברות אין את החובות של ממשלות, השפעתן היא מסוג המחייב אותן להעריך את אותם סוגיות בנוגע להגנה על זכותם של המשתמשים שלהן לחופש הביטוי" (A/74/486, פסקה 41).

בנוסף, קבוצת העבודה של האו"ם לעסקים וזכויות אדם הביעה כי "מכיוון שהסיכון להפרות בוטות של זכויות אדם מוגבר באזורים שנפגעו מסכסוך", יש "להגביר בהתאם את בדיקת הנאותות על ידי עסקים" (A/75/212, פסקה 13). בדו"ח משנת 2024 על הסכסוך בין ישראל לעזה, דיווח מיוחד של האו"ם לחופש הביטוי והדעה טען כי פלטפורמות נכשלות באופן עקבי במילוי אחריות זו בסכסוכים וציין את הסיכונים המוגברים ממידע כוזב ושגוי במצבים כאלה (A/79/319, פסקה 60, 66). בסכסוך ישראל-איראן, הפסקות אינטרנט השפיעו עמוקות על גישת האזרחים למידע, ויצרה חלל ריק שמילאה במהירות מדיה מטעה שנוצרה על ידי בינה מלאכותית (ראה PC-31545 WITNESS).

לרשות Meta עומדת חבילה חזקה של כלים כדי לצמצם את הנזקים הפוטנציאליים של תוכן מטעה שנוצר על ידי בינה מלאכותית בפלטפורמות שלה. מקרה זה מדגים כי יש ליישם גילוי ותיוג תוויות בצורה עקבית, תכופה ויעילה יותר כדי למנוע נזקים צפויים לטווח קצר למשתמשים, במיוחד במצבי סכסוך שבהם ההימור גבוה בהרבה. יש לתמוך בכך במשאבים הולמים של בודקי עובדות של צד שלישי והנחיות לתעדוף תוכן מתוך תוכן סותר, כמו גם השקעה באכיפה מדויקת נגד התנהגות פוגענית ברמת החשבון והדף.

I. חוקיות (בהירות ונגישות הכללים)

עקרון החוקיות דורש שכללים המכבידים על ביטוי יהיו נגישים וברורים, ומנוסחים בדיוק מספיק כדי לאפשר לפרט לווסת את התנהגותו בהתאם (הערה כללית מס' 34, פסקה 25). בנוסף, כללים אלה "אינם עשויים להעניק שיקול דעת בלתי מוגבל להגבלת חופש הביטוי לאלו המופקדים על ביצועם" וחייבים "לספק הדרכה מספקת לאלו המופקדים על ביצועם כדי לאפשר להם לקבוע אילו סוגי ביטוי מוגבלים כראוי ואילו לא" (*שם*). הכתב המיוחד של האו"ם לחופש הביטוי הצהיר שכאשר מיושמים על ממשל שחקנים פרטיים של שיח מקוון, הכללים צריכים להיות ברורים וספציפיים (A/HRC/38/35, פסקה 46). אנשים המשתמשים בפלטפורמות של Meta צריכים להיות מסוגלים לגשת ולהבין את הכללים ולבודקי התוכן צריכים להיות הנחיות ברורות לגבי האכיפה שלהם.

תקן קהילת המידע השגוי אמור לספק יותר בהירות למשתמשים ולאוכפי הכללים.

ההסבר הכללי לשיתוף פעולה של Meta עם צדדים שלישיים כדי לזהות מגמות מטעות אינו מבהיר שהאכיפה תלויה לחלוטין בשיתוף חששות של שותפים ל-Meta. במקום להבהיר זאת, יש לנקוט בגישה שונה, מהסיבות המפורטות להלן.

התיאור הציבורי המפורט היחיד של שלוש תוויות המדיה המניפולטיביות שבהן משתמשת Meta נמצא בהחלטות המועצה. המועצה חוזרת על המלצתה כי Meta תתאר באופן מלא את שלוש תוויות המדיה המניפולטיביות שהיא מיישמת, את הקריטריונים ליישומן ואת השלכותיהם (קטעי מחאה בשילוב עם קריאות פרו-דוטרטה), תוך ציון כי גם גישה זו חייבת להתפתח.

יתר על כן, מדיניות המדיה המניפולטיבית אינה מתארת בפומבי את העונשים ש"עשויים" להיות מוטלים אם משתמש לא יגלה בעצמו שימוש בבינה מלאכותיתMeta . הסבירה למועצה כי עונשים אלה מוחלים רק במקרה של הסלמה בתגובה לכשלים חוזרים ונשנים, ועשויים להשפיע על הפצת התוכן או לגרום באופן זמני להשעיה של תכונות מסוימות בחשבון. Meta מפעילה שיקול דעת רחב לכאורה בתחום ויש לספק למשתמשים מידע ברור יותר.

הצגת מידע נוסף על מקור התוכן, תיוג וגילוי עצמי בתקן קהילת המידע השגוי עלולה ליצור בלבול על ידי ערבוב בין מאמצים לצמצום הטעיה לבין צעדים חיוביים לקידום שלמות המידע. לא כל שימוש בתוכן שנוצר על ידי בינה מלאכותית, ולא כל יישומי התוויות, יגיבו לניסיון הטעיה. פירוט כללים אלה בתקן קהילתי נפרד יכול לסייע בהבהרת הגישה של Meta ובשיפור התנהגות המשתמשים.

II. מטרה לגיטימית

כל הגבלה על חופש הביטוי צריכה לחתור גם לאחת או יותר מהמטרות הלגיטימיות המפורטות באמנה הבינלאומית למדיניות אזרחית ומדיניות, הכוללות הגנה על ביטחונם וזכויותיהם של אחרים.

במקרה הסרטון המשתנה של הנשיא ביידן, הדגישה המועצה כי "מניעת הטעיית אנשים אינה, כשלעצמה, סיבה לגיטימית להגבלת חופש הביטוי". עם זאת, תקן קהילת המידע השגוי שואף גם להפחית את הסיכון לפגיעה פיזית או אלימות קרובה כלפי אנשים, וזוהי מטרה לגיטימית ביחס לזכויותיהם של אחרים (הערה כללית 34, פסקה 28).

III. הכרח ומידתיות

תחת קטע 19(3) באמנה הבינלאומית למדיניות אזרחית ומדיניות (ICCPR), נחיצות ומידתיות מחייבות כי הגבלות על ביטוי "חייבות להיות מתאימות להשגת תפקידן המגן; עליהן להיות הכלי

הפחות פולשני מבין אלה שעשויים להשיג את תפקידן המגן; עליהן להיות מידתיות לאינטרס שיש להגן עליו" (הערה כללית מס' 34, פסקה 34).

הדווח המיוחד של האו"ם בנושא חופש הביטוי הצהיר כי "במהלך סכסוך מזוין, אנשים נמצאים בפגיעות הגבוהה ביותר וזקוקים ביותר למידע מדויק ואמין כדי להבטיח את ביטחונם ורווחתם". עם זאת, דווקא באותם מצבים חופש הדעה והביטוי שלהם [...] מוגבל ביותר על ידי נסיבות המלחמה ופעולות הצדדים לסכסוך וגורמים אחרים לתמרן ולהגביל מידע למטרות פוליטיות, צבאיות ואסטרטגיות" (A/77/288, פסקה 1).

הדווח המיוחד הדגיש גם כי "לחברות יש כלים להתמודד עם תוכן בדרכים התואמות זכויות אדם, במובנים מסוימים מגוון רחב יותר של כלים מאלה שנהנות מהמדינות. מגוון אפשרויות זה מאפשר להם להתאים את תגובותיהם לתוכן בעייתי ספציפי, בהתאם לחומרתו ולגורמים אחרים" (A/74/486, פסקה 51).

בהערכת נחיצותם ומידתיותם של צעדים פוטנציאליים, שקלה המועצה את הדברים הבאים: (א) הדף הציג את עצמו באופן שגוי כמקור חדשותי אמין; (ב) התוכן קשור ישירות לסכסוך מזוין מתמשך; (ג) הפגיעות של אזרחים המחפשים מידע מאומת בעיצומו של אותו סכסוך; (ד) ההתפשטות המהירה המתועדת היטב של תוכן מטעה שנוצר על ידי בינה מלאכותית במהלך סכסוך זה (ראה PC-31528 מכון אלן טיורינג); (ה) הפצה חוצת פלטפורמות של תוכן דומה או כמעט זהה; ו(ו) תמריצי המעורבות והמונטיזציה של יצירת מדיה מניפולטיבית במהלך סכסוכים.

המועצה קובעת כי הצבת תווית מדיה מניפולטיבית של "High Risk AI" על התוכן תעמוד בדרישות ההכרח והמידתיות, והיא מודאגת מכישלונה של Meta לעשות כן. זה יהיה הרבה פחות פולשני מהסרה, בהתחשב בכך שההטעיה לא צפויה להוביל לנזק מיידי. נכון לעכשיו, תיוג כזה יהיה אינפורמטיבי ולא יביא להורדה בדרגה או להסרה מהמלצות. Meta תציג חלון קופץ למשתמשים שמנסים לשתף מחדש תוכן עם תווית זו. התווית תסייע להפחית את ההשפעה של הטעיה זו על משתמשים המחפשים מידע מדויק באינטרנט על הסכסוך.

מקרים קודמים של המועצה טוענים כי התפשטות של מדיה מניפולטיבית ללא תווית עלולה לפגוע באמון באותנטיות של התוכן בפלטפורמה באופן רחב יותר. זה נכון במיוחד במהלך סכסוכים מזוינים, שבהם כלי תקשורת מניפולטיביים המתארים הפרות של המשפט ההומניטארי הבינלאומי עלולים לפגוע באמון במסגרות משפטיות אלה ובהגנות לאזרחים שהן מספקות. כדי לעמוד בחובותיה בתחום זכויות האדם, Meta הייתה צריכה להגביר את הבדיקה של התוכן מבלי להסתמך על צדדים שלישיים חיצוניים, כדי לתייג את התוכן ללא דיחוי.

המנגנונים הנוכחיים להצמדת אפילו התווית הסטנדרטית של מידע מבוסס בינה מלאכותית לסרטונים (גילוי עצמי של המשתמש או הסלמה לצוות מדיניות התוכן) אינם חזקים ואינם מקיפים מספיק כדי להתמודד עם ההיקף והמהירות של תוכן שנוצר על ידי בינה מלאכותית, במיוחד במהלך משבר או סכסוך בהם יש מעורבות מוגברת בפלטפורמה. המועצה מציינת כי מערכת התלויה יתר על המידה בגילוי עצמי של שימוש בבינה מלאכותית ובבדיקה מואצת (אשר מתרחשת לעתים רחוקות) כדי לתייג כראוי תוכן זה אינה יכולה לעמוד באתגרים הנשקפים בסביבה הנוכחית.

חלק מהחברים ציינו כי ההגדרה של Meta ל"נזק פיזי או אלימות קרובים" אינה כוללת את הדרכים השונות בהן תוכן מטעה שנוצר על ידי בינה מלאכותית עלול להיות בעל השפעות חברתיות פחות ישירות אך עדיין חמורות במהלך סכסוך מזוין. זה עלול, למשל, לפגוע בגישה למידע אמין הדרוש כדי להעמיד גורמי ממשל וגורמים אחרים לדין, להגביר את פגיעותן של אוכלוסיות למניפולציה ולאפשר צורות אחרות של השפעה מטעה. עבור חברי מועצה אלה, תיוג של "תקשורת מניפולטיבית" שאינו מלווה בהורדה בדרגה או הסרה מהמלצות אינו מספיק כדי לטפל בחששות אלה. דיכוי ההשפעה של תוכן כזה ברגעים של סיכון גבוה כמו סכסוכים מזוינים יהיה התערבות הכרחית ומידתית. המועצה מכירה בכך שדירוג שגוי מצד בודקי עובדות של צד שלישי על תוכן זה היה מביא גם הוא לתוצאה זו.

ברחבי התעשייה, ישנם אתגרים טכניים משמעותיים כדי להבטיח תיוג מדויק ועקבי של תוכן שנוצר על ידי בינה מלאכותית. Meta הסבירה למועצה שהיא משתמשת באינדיקטורים סטנדרטיים בתעשייה – שהם המטא-דאטה שכלי בינה מלאכותית יצירתיים מטמיעים לעתים קרובות בתוכן – כדי לתייג תמונות סטטיות באופן אוטומטי. עם זאת, לא כל כלי הבינה המלאכותית הגנרית מצרפים כיום את המטא-דאטה הדרושים להחלת תווית. אפילו אם כלי אכן מצרף את המטא-דאטה, משתמשים יכולים להסיר אותו בקלות מהתוכן לפני שיתוף שלו ברשתות החברתיות. ייתכן שהמנגנונים המוגבלים של Meta משקפים את האתגרים הנוכחיים הללו; עם זאת, באחריות החברה להגיב באופן יזום לטכנולוגיה המתפתחת במהירות. זוהי בעיה שמתרחבת מעבר לתוכן מטעה, שכן מגבלות אלו מסתירות את יכולתם של המשתמשים לאמת את האותנטיות של כל המידע. מכשולים אלה רק יגברו ככל שנפח ואיכות הווידאו והאודיו שנוצרו על ידי בינה מלאכותית ימשיכו לעקוף את כלי הזיהוי והתיוג הזמינים. המועצה מעודדת מאוד את Meta לתעדף את חידוד מנגנוני הזיהוי והתיוג שלה כדי ללכוד טוב יותר את כל צורות התוכן שנוצר על ידי בינה מלאכותית בפלטפורמות שלה, ליידע את המשתמשים כראוי מתי הם עשויים לתקשר עם מדיה מניפולטיבית, ולהבטיח התמקדות בסוגי התוכן המהווים את הסיכונים הגבוהים ביותר.

המועצה מכירה בנוכחותה של Meta בוועדת ההיגוי של C2PA. תקן ה-C2PA קובע כי ככל ששיתוף מידע משתנה במהירות, חיוני לעקוב אחר מקור המדיה. הוא מספק תקן טכני פתוח לקביעת המקור והעריכה של תוכן דיגיטלי. דיווחים לפיהם Meta אינה מיישמת באופן עקבי סטנדרטים של C2PA, אפילו לא על תוכן שנוצר על ידי בינה מלאכותית מהכלים שלה, הינם מדאיגים. מחקר שיצא לאחרונה הראה שבעת בדיקה על ידי הכלים של C2PA, רק חלק מהתמונות והסרטונים שנוצרו על ידי כלי הבינה המלאכותית של Meta סיפקו אישורי תוכן וקיבלו תיוג מתאים.

נראה כי תוכן המקרה מקורו לראשונה בפלטפורמה של Meta ב-TikTok, לפני ששותף במהירות בין פלטפורמות שונות, כאשר פוסטים דומים הופיעו ב-Facebook, Instagram ו-X, למרות הדיווח של AFP על זיוף תוכן דומה. תגובות הציבור הדגישו את הצורך בשיתוף פעולה חזק בין פלטפורמות בתקופות של סכסוך מזוין כדי להפחית ולמתן את קצב התפשטות תוכן מטעה שנוצר על ידי בינה מלאכותית.

כפי שפורט לעיל, הייעודים של CPP ו'אירועים טרנדיים' היו אמורים לאפשר ל-Meta להבטיח תמיכה יעילה יותר לבודקי עובדות של צד שלישי במהלך משבר זה. בפרט, פיזור דירוגים לקטגוריה רחבה יותר של סרטונים דומים מאוד היה יכול להגביל משמעותית את הנזק הפוטנציאלי, כולל על ידי הורדת הדירוג. המקרה מדגיש חוסר יעילות בגישתה הנוכחית של Meta במהלך סכסוכים מזוינים, ומחריף את החששות שהמועצה הביעה במקרים קודמים בהקשרים שונים (קטעי מחאה משולבים עם החלטה על קריאות פרו-דוטרטה).

6. החלטת המועצה לפיקוח

המועצה מבטלת את החלטתה של Metaלהשאיר את התוכן ללא תווית High Risk AI.

7. המלצות

א. מדיניות תוכן

מידע מוטעה

1. כדי להבטיח סקירה מהירה של מידע שגוי המוביל לסיכונים של פגיעה פיזית או אלימות ממשית במשברים, על Meta לתקן את תקן קהילת המידע השגוי כדי להבטיח שאכיפת כלל זה לא תהיה תלויה באותות משותפים חיצוניים. במסגרת פרוטוקול מדיניות המשברים צריך להיות מנוף להקצאת משאבים לגילוי בזמן ויזום של תוכן מפר כזה, הנתמך על ידי מומחיות פנימית, לזיהוי, סקירה ותפעול של תוכן במסגרת המדיניות (כולל הדבקת תוויות במסגרת מדיניות המדיה המניפולטיבית וחקירת חשבונות ודפי פרסום המראים סימנים של שימוש לרעה באינטראקציה).

המועצה תשקול יישום זה כאשר Meta תעדכן את מדיניות המידע השגוי שלה כדי לשקף דרישות אלו עבור קטגוריית פגיעה פיזית או אלימות.

תוכן שנוצר על ידי בינה מלאכותית

2. כדי לסייע בקידום האמון במידע בפלטפורמות של ,Meta על Meta ליצור תקן קהילתי לתוכן שנוצר על ידי בינה מלאכותית, נפרד מתקן הקהילה למידע שגוי. תקן הקהילה החדש צריך לספק פרטים מקיפים על שימור מקור (כלומר, לכידת העובדות המפורטות אודות היסטוריה של תוכן דיגיטלי) ,פרוטוקולים לתיוג בינה מלאכותית וכללי גילוי עצמי.

המועצה תשקול יישום זה כאשר Meta תפרסם תקן קהילתי חדש ספציפית לגבי תוכן שנוצר על ידי בינה מלאכותית.

3. כדי לשפר את בהירות הכללים שלה, על Meta לפרסם הסבר ברור לגבי העונשים על אי גילוי עצמי של תוכן שנוצר או שונה דיגיטלית. עליה לספק קריטריונים לעונשים ולפרט אילו תכונות חשבון מוגבלות כתוצאה מכך ולמשך כמה זמן.

המועצה תשקול יישום זה כאשר Meta תעדכן את תקן הקהילה כך שיכלול את פרטי העונשים הללו ותנגיש את ההנחיות המתוקנות במרכז השקיפות הציבורי שלה.

ב. <u>אכיפה</u>

מקור

4. כדי להבטיח שמשתמשים יוכלו לזהות באופן מהימן תוכן שנוצר על ידי בינה מלאכותית, על Meta ליישם אישורי תוכן (כפי שנקבע על ידי הקואליציה למקור ואותנטיות של תוכן) בקנה מידה גדול ולהבטיח שהם גלויים ועקביים ונגישים למשתמשים בכל פעם שפרטי המקור זמינים. אסור שהמקור יישאר ניתן לזיהוי באופן פנימי בלבד או מוגבל למערכות אחוריות.

המועצה תשקול יישום זה כאשר Meta תספק דוח המסביר את השינויים שביצעה בממשקים ובמוצרים שלה כדי להבטיח שאישורי התוכן מוצגים באופן עקבי וברור למשתמשים כאשר הם זמינים.

5. כדי לשפר את דיוק הזיהוי והתיוגMeta , צריכה להשקיע בכלי זיהוי חזקים יותר עבור פורמטים מרובים שנוצרו על ידי בינה מלאכותית (תוכן אודיו, אודיו-ויזאולי ותמונות). כלים צריכים לתמוך בצוותי הסלמה כדי לזהות טוב יותר מגמות תוכן יצירתי של בינה מלאכותית, כולל נזקים פוטנציאליים סביב תוכן מטעה במצבי משבר.

המועצה תשקול יישום של המלצה זו כאשר החברה תאשר כי אומצו כלים חזקים יותר ותשתף נתוני שקיפות על ביצועי כלים אלה. יש לפרט את הממצאים הללו לפי שפה ומדינה, וכן האם פרוטוקול מדיניות המשברים הופעל. נתונים אלה חייבים לשקף תקופות זמן דומות לפני ואחרי יישום שינויים אלה.

6. כדי להבטיח תיוג מדויק יותרMeta , צריכה לצרף מידע על מקור התוכן וסימני מים בלתי נראים לתוכן שנוצר על ידי כלי הבינה המלאכותית של ,Meta כך שניתן יהיה לזהותו ולתייג אותו באופן עקבי בפלטפורמות שונות. זה צריך לכלול יישום של אישורי תוכן בנקודת היצירה לצד שימוש באינדיקטורים סטנדרטיים בתעשייה לייחוס לכל התוכן שנוצר על ידי בינה מלאכותית של Meta.

המועצה תשקול יישום של המלצה זו כאשר החברה תשתף עם המועצה דוח על האופן שבו בינה מלאכותית של Meta מצרפת ומשמרת באופן עקבי נתוני מקור וסימני מים בלתי נראים לתוכן המשותף בפלטפורמה.

7. כדי להפוך את השימוש בתוויות High Risk ו-High Risk AI על תוכן מטעה לעקבי יותר, על Meta לפתח מסלולים להדבקת תוויות אלו לתוכן בתדירות גבוהה הרבה יותר, בסיוע ערוצי הסלמה ברורים יותר ממערכות אוטומטיות ובדיקה בקנה מידה גדול, כך שתיוג כזה יוכל להתרחש בנפח גבוה משמעותית.

המועצה תשקול יישום של המלצה זו כאשר יהיו מסלולים חדשים להסלמה כדי להצמיד תוויות High Risk ו-High Risk AI לתוכן, ו-Meta תדווח למועצה על כמות התוויות הללו המצורפות בשנת ,2026 לפי רבעון. היעדר מכנה (כלומר, הנפח הכולל של תוכן לא מתויג שאינו עומד בסף זה) לא אמור להוות מכשול למתן מידע זה למועצה.

***הערה פרוצדורלית:**

- החלטות מועצת הפיקוח מתקבלות על ידי פאנלים של חמישה חברים ומאושרות ברוב קולות של המועצה המלאה. החלטות המועצה אינן מייצגות בהכרח את דעותיהם של כל החברים.

- על פי אמנתה, מועצת הפיקוח רשאית לבחון ערעורים של משתמשים שתוכנם הסירה Meta, ערעורים של משתמשים שדיווחו על תוכן ש-Meta השאירה, והחלטות ש-Meta מפנה אליה (קטע 2 באמנה, סעיף 1). למועצה יש סמכות מחייבת לאשר או לבטל את החלטות התוכן של Meta (קטע 3 לאמנה, סעיף 5; קטע 4 לאמנה). המועצה רשאית להוציא המלצות לא מחייבות ש-Meta נדרשת להגיב עליהן (אמנת חלק 3, סעיף 4; חלק 4). במקרים בהם Meta מתחייבת לפעול על פי המלצות, המועצה עוקבת אחר יישומן.

- לשם ההחלטה בנושא פנייה זו, המועצה הזמינה מחקר עצמאי שבוצע עבורה. המועצה נעזרה ב-Duco Advisors, חברת ייעוץ המתמקדת בשילוב של גיאופוליטיקה, אמון ובטיחות וטכנולוגיה.