סרטון שנוצר על ידי בינה מלאכותית בסכסוך האיראני-ישראלי

2026-004 FB-UA

סיכום

בניתוח התפשטות תוכן שנוצר על ידי בינה מלאכותית בסכסוכים מזוינים במקרה של מלחמת ישראל-איראן בשנת 2025, המועצה המפקחת קוראת ל-Meta לעשות יותר כדי לאפשר למשתמשים לזהות תוכן מסוג זה. הגישה 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 שמשתמשים יוכלו לזהות באופן מהימן תוכן שנוצר על ידי בינה מלאכותית. בנוסף, עליה לתקן את המדיניות הנוכחית שלה כדי להבטיח תוך פרק זמן מקובל תגובה מספקת לתוכן מטעה שנוצר על ידי בינה מלאכותית.

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

המועצה מבטלת את החלטתה של Meta להשאיר את התפקיד במקרה זה ללא תווית 'סיכון גבוה של בינה מלאכותית'.

למה זה משנה?

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 אמיתי וכי תוכן אמיתי היה מפוברק. זה מחריף את חוסר היכולת של הציבור להבחין באמת, מסמל את יתרונו של השקרן, 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 'סיכון גבוה של בינה מלאכותית'. 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 המתאימה 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 'סיכון גבוה של בינה מלאכותית' (בתוכן שעלול להטעות אנשים בנושאים חשובים) גם עם הורדה בדרגה או הסרה מההמלצות כדי לטפל בחששות מהפצת ההשפעה של תוכן מטעה.

ייתכן שהגישה הצרה של Meta לפיזור דירוגים לתוכן 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 מבטל את החלטתה של Meta להשאיר את התוכן ללא תווית 'סיכון גבוה של בינה מלאכותית'.

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 להדבקת תוויות של 'סיכון גבוה' ו'סיכון גבוה של בינה מלאכותית' 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 שונה 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.