

תקציר מנהלים

המונח דיפ־פייק (deepfake, בתרגום לעברית "זיוף עמוק") הוא הלחם המושגים למידה עמוקה (deep learning) וזיוף (fake). הוא מתאר יישום טכנולוגי מבוסס בינה מלאכותית, המאפשר לשנות או לעבד את התוכן של תמונות או סרטונים כך שקשה ולפעמים אף בלתי אפשרי להבחין שמדובר בזיוף. זאת בשונה מ'צ'יפ־פייק' – זיוף שנוצר באמצעים של עריכה פשוטה ולא באמצעות זיוף מבוסס בינה מלאכותית. דיפ־פייק, כמו יישומים רבים אחרים בתחום הבינה המלאכותית, הפך בעשור האחרון זל ופשוט יחסית לשימוש, ובחלק מן המקרים הוא אף זמין לכל מי שיש לו גישה לרשת האינטרנט.

קיימות מספר שיטות ליצירת דיפ־פייק. הנפוצה ביותר היא בתחום הווידאו, אשר מסתמכת על שימוש ברשתות נוירונים עמוקות (deep neural networks) הכוללות אלגוריתמים שעושים שימוש בטכניקה של החלפת פנים של אדם אחד בפניו של אדם אחר (face-swapping). טכנולוגיה זו פותחה ושוכללה על ידי מפתחים וקהילות אינטרנטיות של קוד פתוח במטרה ליצור אפליקציות נגישות וידידותיות למשתמש להחלפת פנים, דוגמת FaceSwap-I FakeApp.

נכון לשנת 2021 יש עדיין צורך במחשוב יקר וביצרנים מנוסים על מנת לייצר סרטון אמין ולהקשות את זיהוי הזיוף גם על ידי אנשי מקצוע, ולא רק על ידי הציבור הרחב. עם זאת, הטכנולוגיה זמינה כיום לכל דורש כמעט, בין דרך אתרים מסחריים ובין באמצעות אנשי מקצוע. העלות המוערכת כיום עבור סרטון דיפ־פייק פשוט היא בממוצע בין שלושה ל-30 דולר.

ליצירת תוכני שמע (אודיו) או וידאו סינתטיים בעלי מראה מציאותי קיימים שלל יישומים חיוביים ממגוון תחומי חיים:

- **קולנוע וטלוויזיה** – דיפ־פייק מסייע בשיפור איכות הווידאו ובהפיכת סרטים למקצועיים יותר. כמו כן הוא מאפשר שימוש בשחקנים בפרסומות ובסרטים גם ללא נוכחותם, או טשטוש זהות והחלפתה באחרת מסיבות שונות.
- **רפואה ופסיכולוגיה** – מטופלים שבעקבות מחלות כמו ניוון שרירים מאבדים את יכולת הדיבור יוכלו להשתמש במכשיר שיקריא בקולם שלהם טקסט שיוזן לאלגוריתם. קיימים גם טיפולים פסיכולוגיים, למשל בתחום התמודדות עם שכול, שבהם נעזרים בטכנולוגיה זו.
- **חינוך ושימור היסטורי** – יצירת ראיונות או חומרים חזותיים עם דמויות היסטוריות, בהתבסס על תיעוד כתוב או מצולם, לצורך הנגשה לקהל הרחב.
- **משחקי מחשב** – שיפור חווית המשחק ואמינותו באמצעות יצירת דמויות מציאותיות, לעיתים תוך צמצום השקעה בצילום ממושך או בשכירת שחקנים מקצועיים.

חרף הפוטנציאל החיובי בתחום, הוא מביא עימו גם יישומים מסוכנים ומטרידים המתבססים על אותן יכולות ממש ובהן יצירת סרטוני פורנו מזויפים, יצירת תוצרים למטרות סחיטה והונאה וכן להשפעה על הפוליטיקה, לביצוע מניפולציה על דעת הקהל ולפגיעה בביטחון הלאומי. הדיפ־פייק מעצים את אתגרי עידן הפוסט־אמת,

שכן ניתן לייצר מצג כוזב שעלול להטעות את דרג מקבלי ההחלטות וכמובן להשפיע על דעת הקהל והציבור ואולי אף להניע לפעולה, בהתאם לאינטרסים של יוצרי או מפיצי הזיוף.

בשנת 2019 סומן השימוש ביכולות דיפ־פייק על ידי ארגוני המודיעין האמריקאיים כאיום האסטרטגי הגדול ביותר על הביטחון הלאומי. גם הציבור הרחב מוטרד מהתופעה. בסקר דעת קהל משנת 2020 שכלל 34 אלף מבוגרים מרחבי העולם טענו 66 אחוזים מן הנסקרים כי הם מוטרדים מכך ש"טכנולוגיה הופכת לבלתי אפשרית את הידיעה אם מה שאנשים רואים ושומעים הוא אמיתי". כך השימוש ביכולות דיפ־פייק מוביל בין היתר לערעור האמת ולחשד בכל פיסת תוכן כשקרית, גם אם תהא אמיתית. זאת משום שאפשר למשל לטעון שהקלטה היא שקרית והופקה באמצעות יכולות דיפ־פייק ובכך להתנער מאחריות, גם במקרה שמדובר בתיעוד אמין. הדבר מציב אתגרים של ממש לאמת ולדמוקרטיה. בשנים האחרונות תועדו מספר מקרים של השפעת תוצרים של דיפ־פייק על תפיסת המציאות, שהמחישו היתכנות של פגיעה בסדר הציבורי, בביטחון הלאומי ואף בדמוקרטיה.

• **דיפ־פייק לביצוע מרמה** – שימוש בעיבוד קולי במטרה לגרום להעברת כספים במרמה. בבריטניה נגנבו כך 240 אלף דולר ובאיחוד האמירויות הערביות אפשר דיפ־פייק בצירוף תכתובות אי־מייל מזויפות גניבה של 35 מיליון דולר.

• **דיפ־פייק פורנוגרפי לשימוש פוגעני** – שימוש בטכנולוגיה נגד עיתונאיות בהודו ובבריטניה, שהוביל להשתקתן לתקופה מסוימת ובתוך כך לפגיעה בחופש העיתונות.

• **דיפ־פייק לשם השפעה פוליטית** – מגוון מקרים תועדו בתחום זה הן בארצות הברית והן בישראל, והמקרה הידוע ביותר הוא ניסיון הפיכה שבוצע נגד נשיא גבון בשנת 2018, בין היתר בעקבות סרטון שבו הוא נראה לאחר תקופה שחלה ונעלם מעין הציבור. הסרטון נחשד כמזויף אך החשד לא אומת. החשד שהנשיא אינו מתפקד, שהתחזק עקב הסרטון שנחשד כמזויף, סייע להתסיס את ההמונים עד כדי ניסיון הפיכה. כמו כן תועדו מקרים של הנדסת תצלומי לוויין העשויים להשפיע על החלטות בתחום הצבאי.

בימינו, לשם יצירת סרטון דיפ־פייק אמין שיקשה על זיהוי הזיוף עדיין נדרשים אמצעי מחשוב יקרים ויצרנים מנוסים. לעיתים, במקרים של פוטנציאל להשפעה פוליטית, עשויים להיות בעלי אינטרסים כגון שחקנים גדולים בשוק ואף בזירה הבינלאומית, שיהיו בידיהם האמצעים לייצור סרטונים כאלה.

למרות כל אלו, עדיין לא ברור מה יהיו המידה ורמת ההשפעה של זיוף. התשובה היא ככל הנראה שגם אם סרטון יהיה אמין דיו יש להפיצו בסיטואציה מתאימה, כזו המערבת חוסר יכולת של הגורם המופיע בסרטון להכחיש כשיש בו קריאה לפעולה מיידית, או במקרים שבהם יש רגישות לגורם הזמן (למשל זמן קצר לפני ההצבעה במערכת בחירות, כך שלא מתאפשר בירור). בשלל סיטואציות כאלה עשויים המפיצים להשיג את מטרותיהם, יהיו אשר יהיו, במחירים נמוכים יחסית. עם זאת **עדיין לא ידוע מקרה שבו מדינה ניסתה והצליחה להשפיע על המתרחש בזירה הפוליטית או הביטחונית במדינה אחרת** באמצעות סרטון מזויף.

ההיערכות להתמודדות עם האיום מתרחשת בשלוש מסגרות נושאיות מרכזיות:

• **רגולציה, חקיקה ואכיפה** – במספר מדינות בארצות הברית נחקקו חוקים כלליים או נקודתיים בנוגע להפצה של דיפ־פייק (בעיקר בתקופת בחירות, למשל). הקושי הגדול בנושא החקיקה הוא חוסר החבות

המשפטית שיש לרשתות החברתיות, שהן אמצעי ההפצה המרכזי של תכנים כאלה. לחלופין, בסין קיים חוק המחייב את יוצריהם של סרטוני דיפייק לסמנם ככאלה.

- **פתרונות טכנולוגיים** – אלה מאפשרים שימוש בכלי בינה מלאכותית לזיהוי סרטונים מזויפים באופן אוטומטי או ידני, כמו למשל מאמת הווידאו של מיקרוסופט, היישום של פייסבוק בשיתוף עם אוניברסיטת מישיגן או אמצעי של צבא ארצות הברית שפותח לשם כך. עם זאת נדרש שהטמעה ושימוש בכלים אלו ייעשו על פי חוק או רגולציה בצמתים הרלוונטיים.
- **חינוך הציבור** – חשיפת הציבור הרחב לתופעה ולמושג באמצעות המדיה המסורתית על מנת ליצור מידה של היכרות עם הנושא וביקורתיות כלפי תכנים, לצד חשיפת הציבור לתוכנות וליישומים נגישים לצורך זיהוי של דיפייק.

המלצות למדיניות בישראל

במהלך מחקר זה, שבוצע במסגרת **תוכנית ליפקין-שחק לפוסט־אמת ופייק ניוז** במכון למחקרי ביטחון לאומי, נערכה באוקטובר 2021 סימולציה בהשתתפות מומחים מתחומי הדוברות והתקשורת, האסטרטגיה וניהול המשברים, הטכנולוגיה והבינה המלאכותית, החדשנות, המשפט והמדיניות. המומחים התבקשו לגבש המלצות לאור תרחיש בדיוני שהוצג להם, ובו הוביל סרטון (אשר בדיעבד התברר כדיפייק) שהופץ באפליקציות הודעות מיידיות לתגובות אלימות בזירה הביתית, לאיומים באלימות מחוץ וכן לגינויים בזירה הבינלאומית. המשתתפים התבקשו לגבש הן המלצות להתמודדות בזמן אמת והן המלצות להיערכות מראש.

התמודדות בזמן אמת

- במקרה של סרטון מזויף המציג גורמים רשמיים, מומלץ לנקוט אסטרטגיה של הזמה מהירה מבוססת ראיות, למרות שיש להביא בחשבון כי לניסיונות להוכיח שהסרטון זויף לא תהיה השפעה מוחלטת, אלא רק על האוכלוסייה ה"מתלבטת" אם להאמין ולפעול לפי הדברים.
- אפשר לנקוט גם גישה קונוונציונלית פחות במטרה להגחיק את הסרטון בעיני הציבור, תוך הפצת סרטוני דיפייק שיוצגו ככאלה, כדי להדגים עד כמה פשוט לייצר הטעיה.
- יש להיערך להתמודדות עם פוטנציאל למתקפה של סרטונים כאלה ולא של סרטון אחד בלבד, נוכח הסיכוי שהגורם האחראי להפצה הוא מדינתי או בעל אמצעים רבים.
- מבחינה משפטית, גם אם קיים סיכוי סביר לקבל צו עיכוב פרסום מבית משפט, נוכח הסיכון הגבוה לפגיעה בביטחון, יעילות מימוש מוטלת בספק בגלל הקושי לצמצם את התפוצה דרך אפליקציות מסרים מידיים ורשתות שונות. גם הצנזורה ככל הנראה אינה רלוונטית במקרה זה.
- נוכח העובדה שהתקשורת בוודאי תרצה לשדר את הסרטון, יש צורך לעבוד בצמוד לה בנוגע להצגתו ולדרוש הקפדה על סימונו כמזויף.
- בשונה מזויפים ביומטריים אחרים, הרשויות עדיין אינן ערוכות להתמודד עם אירועים מסוג זה, שעונים גם הם על ההגדרה של זיוף ביומטרי (שימוש במראה פנים, קול וחיתוך דיבור שלא ברשות בעליהם). יש לתת את הדעת גם על היבט זה של האיום ולמנות את הגורמים הרלוונטיים לטיפול בהם.

היערכות מקדימה

- **חקיקה ורגולציה** – יש לתת בידי הגורמים האמונים על האכיפה, כלים להתמודד עם אתגר הדיפ־פייק הן בשלב מקדים והן בזמן אמת, תוך התייחסות למורכבות ולרגישות בנוגע לחופש הביטוי.
- **רשויות החוק והמשפט**, שאחראיות על תחומי המחשוב והדיגיטל במדינת ישראל, נדרשות גם הן להתארגנות לטיפול בנושא, מתוך הכרה שמדובר באירוע ביומטרי לכל דבר ועניין.
- **חיזוק הקשר ועבודה משותפת עם התקשורת הממוסדת והרשתות החברתיות** – יש צורך בגיבוש נוהל עבודה מסודר מול התקשורת הממוסדת וביצירת סטנדרטים לטיפול באירועים רלוונטיים.
- **חינוך הציבור** – יש לסייע בהעלאת המודעות של הציבור לאתגר הדיפ־פייק ולפשטות ולזמינות של יצירתו, על מנת לעורר ביקורתיות כלפי תוכן שהציבור נחשף אליו בערוצים הלא־ממוסדים. יש צורך להגביר את המודעות בקרב אנשי מקצוע – עיתונאים, מנהלי קהילות ברשתות החברתיות ומשפיענים אחרים – ולחשוף אותם לכלים טכנולוגיים המסייעים בזיהוי של דיפ־פייק. לצורך חינוך הציבור אפשר להיעזר בקמפיינים ממומנים על ידי הממשלה, כפי שנעשה בכל הקשור למגפת הקורונה למשל, או להיעזר בתוכניות טלוויזיה העוסקות בטכנולוגיה לשם הצגת העניין לציבור. עם זאת, הצלחה יתירה של קמפיינים כאלה עשויה להוביל לעודף חשדנות ולערעור אמון הציבור גם בתוכן שמקורו במוסדות המדינתיים, ולפיכך יש לפעול בצורה מבוקרת ומנוהלת ובסייגים המתאימים.

במסגרת הסימולציה עלתה הסתייגות מסוימת מהייחודיות של הדיפ־פייק עצמו ביחס לציפ־פייק – זיופים המושגים באמצעים פשוטים יותר שהיו מוכרים ונפוצים כבר בעבר. זאת בין היתר משום שעדיין לא הודגם מצב של מתקפת דיפ־פייק משמעותית. עיקר המבוכה נבע מן ההתמודדות עם תפוצה כמעט בלתי ניתנת לעצירה של תוכן נוכח הקושי של הרשויות להגביל את התפוצה ברשתות החברתיות, את האפליקציות להעברת מסרים מיידיים ואת עוצמתן של קבוצות וקהילות דיגיטליות, כמו גם את היכולת להשפיע באמצעות פרופילים מזויפים ובוטים לסוגיהם. יוזכר כי בעבר עלה ספק לגבי האיום הגלום באיומים טכנולוגיים אחרים דוגמת רחפנים, שכיום זוכים לתשומת לב נרחבת המובילה בין היתר לפיתוח מערכות נגד ודוקטרינות להתמודדות, אולם באיחור של כשבע עד עשר שנים ביחס להצגה הראשונית של האיום.

אף שאיום הדיפ־פייק עדיין לא הומחש בתוצאה קטסטרופלית בתחום הביטחון הלאומי, נוכח הפוטנציאל ההרסני שבטכנולוגיה זו והפרכתה לזולה וזמינה, נדרשים בכל מדינה מחקר של הנושא ויצירת כלים להתמודדות מונעת ומאוחרת עימו. מסקנה זו תקפה שבעתיים עבור דמוקרטיה ליברלית דוגמת ישראל, שמתקיימת בה תקשורת חופשית מחד גיסא, אך מאידך גיסא היא נדרשת להיערכות ולהתמודדות יום־יומית עם איומים ביטחוניים שהדיפ־פייק עשוי בקלות להיות אחד המרכזיים שבהם.