איום מזויף או אמיתי?

דיפ-פייק והאתגרים לביטחון הלאומי

לירן עונתבי

INSS
המכון למחקרי ביטחון לאומי
THE INSTITUTE FOR NATIONAL SECURITY STUDIES
TEL AVIV UNIVERSITY אוניברסיטת תל אביב

מזכר 220

פברואר 2022

איום מזויף או אמיתי?
דיפ־פייק והאתגרים לביטחון הלאומי

לירן ענתבי
בהשתתפות נועם רחמים

איום מזויף או אמיתי?
דיפ־פייק והאתגרים לביטחון הלאומי

לירן ענתבי
בהשתתפות נועם רחמים

המכון למחקרי ביטחון לאומי

משלב בתוכו את מרכז יפה למחקרים אסטרטגיים, הוקם בשנת 2006. למכון שתי מטרות מוצהרות: הראשונה היא לערוך מחקרים בסיסיים בנושאי הביטחון הלאומי של ישראל, המזרח התיכון והמערכת הבינלאומית, וזאת על פי אמות המידה האקדמיות הגבוהות ביותר, והשנייה — לתרום לדיון הציבורי ולעבודת הממשל בנושאים שנמצאים, או ראוי שיימצאו, בראש סדר היום הביטחוני של ישראל.

קהלי המטרה של המכון למחקרי ביטחון לאומי הם דרג מקבלי ההחלטות, מערכת הביטחון, מעצבי דעת קהל בישראל, הקהילה האקדמית העוסקת בתחומי הביטחון בישראל ובעולם והציבור המתעניין באשר הוא. המכון מפרסם מחקרים שלדעתו ראויים לתשומת הלב הציבורית ושומר על מדיניות נוקשה של אי־משוא פנים. הדעות המובעות בפרסומים הן של המחברים בלבד, ואינן משקפות בהכרח את העמדות של המכון, של נאמניו או של האישים ושל הגופים התומכים בו.

המכון למחקרי ביטחון לאומי
חיים לבנון 40
ת.ד 39950
רמת־אביב
תל־אביב 6997556

info@inss.org.il
http://www.inss.org.il/he
ISBN: 978-965-92918-8-5

עורכת הסדרה: ד"ר ענת קורץ, מנהלת המחקרים, המכון למחקרי ביטחון לאומי
עורכת לשון: מירה ילין
הביאה לדפוס: נעם רן, המכון למחקרי ביטחון לאומי
עיצוב גרפי: מיכל סמו קובץ, המשרד לעיצוב גרפי, אוניברסיטת תל אביב
עיצוב העטיפה: יגאל טליאנסקי, המכון למחקרי ביטחון לאומי
תמונת הכריכה: Photo by Rachel McDermott & Imansyah Muhamad Putera, Unsplash
דפוס: דיגיפריט זהב בע"מ

תוכן העניינים

תקציר מנהלים

המונח דיפ־פייק (deepfake, בתרגום לעברית "זיוף עמוק") הוא הֶלחם 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 זול 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 FakeApp ו-FaceSwap.

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 חוויית 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 דיפ־פייק 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 דיפ־פייק.

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 כדיפ־פייק) 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 להגחיך את הסרטון בעיני הציבור, תוך הפצת סרטוני דיפ־פייק 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 מימושו מוטלת בספק בגלל הקושי לצמצם את התפוצה דרך אפליקציות מסרים מיידיים 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 מזיופים 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 יתרה 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 לצ׳יפ־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 והפיכתה 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.

הקדמה

המונח דיפ־פייק (deepfake) הוא הֶלחם המושגים למידה עמוקה (deep learning) וזיוף (fake). הוא מתאר יישום טכנולוגי מבוסס בינה מלאכותית, המאפשר לשנות או לעבד תוכן של תמונות או סרטונים כך שקשה ולפעמים אף בלתי אפשרי להבחין בזיוף. זאת בשונה מ׳צ׳יפ־פייק׳ — זיוף שנוצר בעריכה פשוטה ולא בזיוף מבוסס בינה מלאכותית (Paris & Donovan, 2019). יישום זה, כמו יישומים רבים אחרים בתחום הבינה המלאכותית, נעשה בעשור האחרון זול ופשוט יחסית לשימוש, וחלק מהיישומים המאפשרים לייצרו זמינים לכל אדם עם גישה לרשת האינטרנט.

יכולות דיפ־פייק מאפשרות יצירת סרטונים או קטעי שמע סינתטיים בעלי מראה או שמע מציאותיים. כך אפשר לייצר הלכה למעשה ״זיוף ביומטרי״ הנסמך לרוב על תווי הפנים או הקול, ולהציג מצג שווא כאילו אדם ביצע פעולות או אמר דברים, ולא היא. השיטה הנפוצה ביותר ליצירת דיפ־פייק היא בתחום הווידאו (Sample, 2020). מספר הסרטונים המזויפים המופצים ברשת גדל באופן מעריכי (אקספוננציאלי) מאז 2018 ומכפיל עצמו כל חצי שנה. בדצמבר 2020 המספר כבר עלה על 85 אלף סרטונים (Sensity, 2021).

היכולת לייצר תוכני שמע (אודיו) או וידאו סינתטיים בעלי מראה מציאותי באה לידי ביטוי בשלל יישומים חיוביים ממגוון תחומי חיים: החל בקולנוע וטלוויזיה דרך רפואה, פסיכולוגיה ועד אומנות, תרבות ושימור היסטורי. כך למשל, יכולות דיפ־פייק מצליחות לתת לחולי ALS קול, עצמאות ויכולת השפעה, ״להשיב לחיים״ דמויות היסטוריות במוזיאונים ואפילו לשפר את הרפואה באמצעות ״חולים סינתטיים״ (Jaiman, 2020).

אולם חרף הפוטנציאל החיובי הטמון בתחום, הוא מביא עימו גם יישומים מסוכנים ומטרידים המתבססים על אותן יכולות ממש. בין היישומים הללו אפשר למנות למשל הונאות פליליות שעלולות לפגוע בפרטים ובארגונים, או יצירת פורנוגרפיה בהשתתפות לכאורה של דמויות מזויפות שלא נתנו לכך את הסכמתן, לעיתים אף למטרות סחיטה או שימושים פוגעניים אחרים. במקביל לאלו ניצבת סכנת פגיעה בחברה האזרחית או במדינות על ידי יצירת מצגים שקריים שתכליתם עיצוב תודעה, הלכי רוח ודעת קהל, על מנת להשפיע על המערכת הפוליטית, על מערכות בחירות ואף על יציבותן של מדינות (Sayler & Harris, 2021). הזיוף העמוק הוא למעשה יישום טכנולוגי המעצים את אתגרי התקופה הקיימים בתחום הביטחון הלאומי בעידן של פוסט־אמת, כאשר סרטונים מזויפים עלולים להטעות את דרג מקבלי ההחלטות ואולי אף להניע לפעולה, בהתאם לאינטרסים של יוצריהם (Schiff, 2019).

משום כך, השימוש ביכולות דיפ־פייק סומן על ידי ארגוני המודיעין האמריקאיים בשנת 2019 כאיום האסטרטגי הגדול ביותר על הביטחון הלאומי (Konkel, 2019; Ng, 2019). לא רק ארגוני מודיעין חוששים מהפצתו של תוכן מזויף. בסקר דעת קהל בינלאומי משנת 2020 טענו 66 אחוזים מן הנסקרים כי הם מוטרדים מכך שבשל הטכנולוגיה אי אפשר לדעת אם מה שאנשים רואים ושומעים הוא אמיתי (2020 Edelman Trust Barometer, 2020). המשמעות העיקרית של יכולות הזיוף העמוק היא אם כן ערעור על התוקף של האמת האובייקטיבית ובה בעת יצירת חשד בכל פיסת תוכן כשקרית, שכן אפשר לטעון כי היא מזויפת גם כאשר

היא אמיתית, ובכך להתנער מאחריות (אורפז, 2020). ליכולת זו עלולה להיות השפעה הרסנית על חברות בכלל ועל דמוקרטיות בפרט.

מחקר זה, **שבוצע במסגרת תוכנית ליפקין־שחק לפוסט אמת ופייק ניוז במכון למחקרי ביטחון לאומי**, סוקר ומציג את הטכנולוגיה המשמשת ליצירת זיופים עמוקים ואת מגוון השימושים החיוביים והשליליים המוכרים שלה, תוך דיון במקרי מבחן בולטים מן העולם. לאחר מכן ייבחנו ההתפתחויות בתחום ההיערכות וההתמודדות עם האתגר בעולם ובישראל. לבסוף, באמצעות מסקנות שעלו מסימולציה שנערכה במכון בהשתתפות מומחים, יוצגו המלצות למדיניות רצויה בתחום עבור ישראל על מנת להיערך כהלכה לקראת איום ההולך ומתגבש.

מחקר זה לא היה מתאפשר ללא עזרתם האדיבה של רבים, ועל כך תודתנו. לתת־אלוף (מיל׳) איתי ברון, סגן ראש המכון למחקר וראש תוכנית ליפקין־שחק במכון למחקרי ביטחון לאומי (INSS), על שיזם את המחקר ועזר לעצב אותו ולצקת בו תוכן. לגב׳ ענבל אורפז — חוקרת בתוכנית ליפקין־שחק של המכון על סיוע בתובנות ובכל היבט נדרש אחר לאורך המחקר כולו. תודתנו הגדולה גם לגב׳ יובל קנפו, מתמחה בתוכנית טכנולוגיות מתקדמות וביטחון לאומי, שסייעה רבות בסקירה ובגיבוש של חומרים למזכר. אנו מבקשות להודות גם לכל חברי ועדת המומחים (הרשימה המלאה מופיעה בנספח א׳) שכונסה לצורך המחקר, אשר תרמו מידיעותיהם, מזמנם וממרצם למחקר זה; לד״ר ענת קורץ, עמיתת מחקר בכירה ומנהלת המחקר במכון למחקרי ביטחון לאומי, ולד״ר גליה לינדנשטראוס, עמיתת מחקר בכירה במכון למחקרי ביטחון לאומי, שעמלו על עריכתו ושיפורו של חיבור זה. לגב׳ מירה ילין על העריכה הלשונית, למר יגאל טליאנסקי על העיצוב הגרפי ולגב׳ נעם רן, מנהלת הפרסומים של המכון למחקרי ביטחון לאומי על הניצוח על האופרציה.

רבים לקחו חלק וסייעו לגיבושו ופרסומו של מחקר זה, אולם האחריות לכל טענה או טעות המופיעה כאן היא של המחברות בלבד.

טכנולוגיות דיפ־פייק

המונח דיפ־פייק (deepfake) מתאר תמונה, קטע קול או סרטון שעברו מניפולציה מלאכותית אך הם בעלי מראה מציאותי ואמין (Sayler & Harris, 2021). המונח דיפ־פייק הוא למעשה הֶלחם של למידה עמוקה (deep learning) וזיוף (fake), שכן שורשי מהפכת הזיוף העמוק נעוצים במהפכה טכנולוגית נוספת שהתחוללה לפני פחות מעשור: למידה עמוקה (Sejnowski, 2018).

למידה עמוקה (deep learning) היא תת־תחום של למידת מכונה, אשר עושה שימוש ברשתות נוירונים מלאכותיות (artificial neural networks). רשתות נוירונים אלה הן אלגוריתמים השואבים השראה מאופן הפעולה של רשת העצבים במוח האדם. הרשת העצבית לומדת על ידי תיקון הקשרים הרבים בתוכה; היא עורכת תיקונים קטנים באמצעות בחינת מידע רב על מנת להיטיב את דיוקה, וכך הפלט של נוירון אחד הוא הקלט של נוירון אחר. בזכות הצלחות ראויות לציון, רשתות הנוירונים הפכו עם השנים לנפוצות ביותר מבין הגישות של למידת המכונה, והן אחראיות למגוון הישגים בתחום הבינה המלאכותית ובתוכם: זיהוי פנים, זיהוי עצמים בתמונות, תמלול ותרגום, שליטה בכלי רכב אוטונומיים וברחפנים ועוד (ענתבי, 2020).

קיימות מספר שיטות ליצירת דיפ־פייק, הנפוצה ביותר היא בתחום הווידאו ומסתמכת על שימוש ברשתות נוירונים עמוקות (deep neural networks) הכוללות אלגוריתמים המשתמשים 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, כמו FakeApp ו־FaceSwap (Khalil & Maged, 2021).

האפליקציות המדוברות ודומותיהן עובדות לרוב באופן דומה. הן מתבססות על מספר תמונות או סרטונים של אדם א׳ — האדם שאנחנו רוצים שיופיע בסרטון — ומתוך מידע זה האלגוריתם רוכש ולומד את המראה, ההגייה וצורת הדיבור של אדם א׳. נוסף על כך יש צורך בסרטון בסיס שבו אדם ב׳ אומר או עושה את מה שנרצה שהוא יציג כאילו אדם א׳ עושה או אומר (אף שלא עשה או אמר). האלגוריתם יודע לשנות את פניו וקולו של אדם ב׳ כך שייראה כאילו אדם א׳ הופיע בסרטון. האפליקציה חותכת למעשה את הפנים מתוך כל תמונה נתונה ומאמנת שתי רשתות נוירונים מסוג אוטואנקודר (autoencoder) — אחת על פניו של המוחלף והשנייה על פניו של המחליף. לאחר האימון, האפליקציה מסוגלת לקבל כל תמונה של המוחלף, להחליף את פניו בפנים האחרות ולחברן לגוף באופן שנראה אמין.

רשת אוטואנקודר, שהיא הבסיס לתהליך זה, היא רשת נוירונית שמשחזרת את הקלט שהיא מקבלת בשכבת הפלט. היא מורכבת משני חלקים: מקודד (encoder) ומפענח (decoder). המקודד מקבל קלט — תמונת פנים, מעבד אותו ומחזיר פיסת נתונים קטנה יותר. אם למשל תמונת הפנים מיוצגת על ידי עשרות אלפי ערכים מספריים, אזי ״הנתונים הנסתרים״ (latent variables) — הפלט במוצא המקודד — יכולים להיות בסך הכול עשרות מספרים. כלומר, המקודד מעבד ומאבד מידע, אך השאיפה היא לאמן אותו כך שהמידע שיוציא ייצג את הקלט בדרך שתאפשר לשחזר את הקלט המקורי. המפענח מקבל את הקוד המקוצר ומטרתו היא לעבד אותו ולהרחיבו לגודלו מקורי, דהיינו שוב לתמונת פנים. האימון של שני המרכיבים —

המקודד והמפענח — נעשה במשותף, והגמול לכל אחד על הצלחתו הוא כאשר התמונה במוצא המפענח זהה לתמונה בכניסה למקודד. אימון מוצלח של אוטואנקודר הוא למעשה דחיסת נתונים או קידוד של הרבה מידע במעט מידע. ליכולת הזו שימושים רבים, ובמקרה של דיפ־פייק זהו ייצוג תמציתי של תמונת פנים של אדם א׳ שנועד לשחזר — באמצעות אוטואנקודר שאומן על מקבץ אחר — תמונת פנים של אדם ב׳. מדובר בשיטה פופולרית המאפשרת ליצור תוצרי דיפ־פייק בתוכנות חינמיות נגישות בקלות ובמחיר נמוך, כאשר התוצאות אמינות ומציאותיות יחסית (Sayler & Harris, 2021).

דרך נוספת ליצור דיפ־פייק עושה שימוש ברעיון שהגו איאן גודפלו(Ian Goodfellow) ועמיתיו ופורסם ב־2014 במאמר שכתבו, אשר הפך לאחד המצוטטים ביותר בשנים האחרונות בתחום מדעי המחשב (כ־35 אלף ציטוטים על פי מנוע החיפוש Google Scholar). הטכנולוגיה שגודפלו ועמיתיו יצרו מכונה Generative Adversarial Nets (להלן: GANs). דרך מערכת GAN יכולות שתי רשתות מחשבים להשוות נתונים זו עם זו בו־זמנית. מערכת GAN מורכבת משתי רשתות נוירונים מלאכותיים שלכל אחת מהן תפקיד מוגדר: האחת היא הרשת הגנרטיבית — הרשת שמייצרת מידע חדש. היא אמונה על יצירת נתונים מזויפים כגון תמונות, קטעי שמע או קטעי וידאו המשכפלים את המאפיינים של מערך הנתונים המקורי; הרשת השנייה היא הרשת הדיסקרימינטורית — המבחינה (מלשון to discriminate), שאמונה על זיהוי זיוף הנתונים (Goodfellow et al,. 2014).

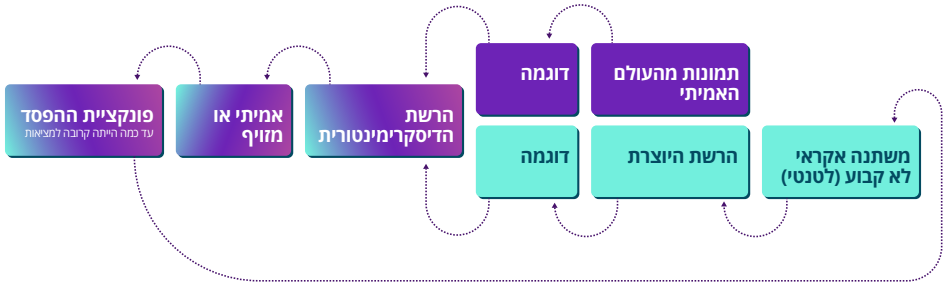
בשלב הראשון מאמנים את הרשת הגנרטיבית באימון בסיסי בנפרד מהרשת השנייה, כדי שתלמד כיצד נראות פנים אנושיות. לצורך זה המערכת מוזנת באלפי תמונות של דיוקנאות שונים, ומתוכם הרשת תסיק את המאפיינים הבסיסיים שמגדירים פנים ואת הקשרים החבויים בין המאפיינים הללו: מיקום העיניים, האף או הסנטר, מה מאפיין פנים נשיות ומה מאפיין פנים גבריות, וכדומה. המטרה אינה לימוד בעל־פה על ידי המערכת היכן נמצא כל איבר בפנים, אלא שהיא תגלה ותגדיר דפוסים כלליים ועקרוניים במידע שהיא מקבלת, כדי שבהמשך תהיה מסוגלת ליצור דיוקנאות חדשים ושונים זה מזה. שלב האימון הראשוני הזה חשוב כדי להביא את הרשת הגנרטיבית, היוצרת, לרמה של יכולת בסיסית. לאחר האימון הבסיסי של הרשת הגנרטיבית, גם הרשת הדיסקרימינטורית נדרשת לעבור אימון בסיסי בנפרד, כדי שתדע גם היא לזהות פנים אנושיות, וכך הן יהיו יריבות ראויות זו לזו (Goodfellow et al., 2014).

בשלב הבא מחברים בין הרשתות: מחברים את המוצא של הרשת הגנרטיבית לכניסה של הרשת הדיסקרימינטורית. מחברים לרשת הדיסקרימינטורית כניסה נוספת שלתוכה יוזנו תמונות של אנשים אמיתיים, לא מזויפים. הרשת הדיסקרימינטורית תידרש להבחין בין התמונות האמיתיות לבין התמונות המזויפות שהפיקה הרשת הגנרטיבית. כעת האימון מתחיל. כדי שהרשת הגנרטיבית תפיק תמונה מזויפת, ראשית יש להזין לתוכה מספר אקראי כלשהו. המספר האקראי הזה ישמש גרעין, שעליו יכולה הרשת הגנרטיבית לבנות את הפנים הספציפיות שהיא מציירת, בעזרת הידע שצברה על אודות הקשרים החבויים שבין חלקי הפנים. האקראיות הזו תבטיח שהרשת הגנרטיבית תפיק בכל פעם תמונה חדשה וייחודית. התמונה החדשה הזו תועבר לרשת הדיסקרימינטורית ולצידה תמונה אמיתית, לא מזויפת. הרשת הדיסקרימינטורית תבחן את התמונות ותקבע מי מהן אמיתית ומי מזויפת. אם הרשת הדיסקרימינטורית צדקה, סימן שהרשת הגנרטיבית לא עשתה עבודה טובה מספיק. המערכת תשנה את עוצמת הקשרים בין הנוירונים ברשת הגנרטיבית כדי

לנסות לשפר אותה ותנסה שוב ליצור תמונה. היא תיצור תמונה חדשה שונה מעט מקודמתה: פנים שונות, שיער אחר וכדומה. אם הרשת הדיסקרימינטורית טעתה וחשבה שהתמונה המזויפת היא בעצם תמונה אמיתית, סימן שהתמונה לא טובה מספיק, ולכן המערכת תשחק עם הקשרים בין הנוירונים של הרשת הדיסקרימינטורית ותנסה לשפר ולחדד את יכולת ההבחנה שלה.

תהליך זה נמשך בלולאה: הרשתות ממשיכות להתחרות זו בזו, לעיתים קרובות במשך אלפי או מיליוני חִזָרורים (איטרציות). בכל סיבוב שתי הרשתות ישתפרו בהדרגה עד התוצאה הרצויה, והיא שהרשת הדיסקרימינטורית תטעה בזיהוי התמונה המזויפת במחצית מהמקרים. בסיום האימון אפשר לפרק את מערכת ה־GAN ולשלוף ממנה את הרשת הגנרטיבית — שכן מטרת האימון היא ליצור רשת גנרטיבית טובה שאפשר יהיה להשתמש בה לתכליות נוספות (Citron & Chesney, 2019 ;Goodfellow et al., 2014).

רשתות יריבות יוצרות



לשיטת GAN יש יתרון משמעותי — אימון רשתות הנוירונים ללא מעורבות אנושית — מה שמכונה "לימוד לא מפוקח" (unsupervised training). כל ההחלטות של המערכת וכל התיקונים והשינויים ברשתות הנוירונים עצמן נעשים באופן עצמאי לחלוטין, ואין צורך באדם שיתערב בתהליך ויתייג כל תמונה כמזויפת או אותנטית (Salimans et al., 2016).

בשימוש ב־GANs כמעט בלתי אפשרי להבחין בעין אנושית לא מאומנת בין אמת לשקר, אולם השימוש ב־GANs מצריך כמות עצומה של נתוני אימון, והוא יעיל ליצירת תמונות סינתטיות ולא סרטונים. עם זאת, ההערכה הרווחת בקרב מומחים היא כי GANs יהיו המנוע העיקרי לפיתוח דיפ־פייק מתוחכם בעתיד (Adee, 2020).

שימושים בדיפ־פייק ומקרי בוחן מן העולם

במסגרת ׳אתוס הקוד הפתוח׳ של חלק ניכר מהקהילה של למידת מכונה, כל התקדמות מחקרית בתחום הזיוף העמוק מתפרסמת מייד בפורומים ייעודיים והופכת פומבית וזמינה (Toews, 2020). כמו ביישומי בינה מלאכותית רבים אחרים, עד לפני מספר שנים היה קשה מאוד להשתמש ביכולות דיפ־פייק, משום שהדרישות לכמות מידע גדולה וכן לכוח מחשוב רב ויקר יצרו מכשול. על כן עיקר השימושים בוצע בגופים גדולים — חברות ומדינות (רשות החדשנות, 2019). אולם ההתקדמות הטכנולוגית בשנים האחרונות הפכה את הטכנולוגיה זמינה לקבוצה רחבה של יוצרי תוכן, חובבים ואנשי מקצוע כאחד (Biggs & Moran, 2021). כיום גם משתמשים ללא כל רקע בתחום המחשוב יכולים ליצור סרטון מזויף קצרצר תוך שניות ספורות, וסרטונים ארוכים יותר תוך מספר דקות (Panyatham, 2020). לפיכך שחקנים שונים יכולים לעשות שימוש בטכנולוגיה זו — אלה הפועלים בחסות מדינה ולצידם קבוצות פוליטיות ואנשים פרטיים (Toews, 2020).

קיימות שלוש דרכים עיקריות ליצירה של זיופים עמוקים בתחום הווידאו והקול. הראשונה היא ייצור עצמי על בסיס אפליקציות ותוכנות מובנות ייעודיות למחשב, שמרביתן מספקות יכולות החלפת פנים בסרטונים קיימים, ואחת מהן אף מאפשרת שיבוט קול סינתטי. שימוש יעיל בתוכנות אלו מצריך לרוב ידע בתכנות ומעבד גרפי רב־עוצמה, מה שהופך אותן נגישות פחות לחובבים. למרות זאת הופצו מספר מדריכים מפורטים לשימוש ביישומים הפופולריים הללו, ועם השנים בוצעו בחלק מן התוכנות הנפוצות עדכונים ששיפרו את הגישה למשתמשים רבים. עם זאת, עדיין נדרשים ידע ויכולת תפעול של יישומים מתקדמים יחסית. הדרך השנייה היא רכישה של סרטונים דרך פורטלים מקוונים המספקים את השירות. הדרך השלישית היא הזמנה של שירותים כאלה מיוצרים פרטיים המפרסמים את שירותיהם בפורומים שונים וב׳אתרי שוק׳ ברשת. שלוש הדרכים המפורטות מתומחרות בטווח שבין שלושה לכ־30 דולר בממוצע עבור סרטון, אך יכולות להיות גם עלויות גבוהות יותר. כמו כן, המזמין נדרש לספק כמות גדולה של תמונות וסרטונים של האדם שהוא מבקש שיופיע בסרטון (Ajder et al., 2019). למרות זאת ניכר כי בשונה מן העבר גברה הזמינות של סרטונים כאלה, תוך שהמחיר ירד ופשטות היצירה או ההזמנה של השירות השתפרו מאוד.

הטכנולוגיה מאפשרת מגוון רחב של שימושים בסרטוני דיפ־פייק, חלקם נפוצים כבר היום. לצד שימושים שליליים בטכנולוגיית הזיוף העמוק כמו פורנוגרפיה, סחיטה והונאה, נעשה גם שימוש לצרכים חיוביים מגוונים, החל בקולנוע וטלוויזיה דרך נגישות, רפואה, פסיכולוגיה ועד אומנות, תרבות ושימור היסטורי. להלן תיאור של השימושים השונים וכן מספר מקרי שימוש שלילי שהתרחשו בעולם, והם בעלי פוטנציאל השפעה על הביטחון הלאומי.

קולנוע וטלוויזיה, אומנות ותרבות

דיפ־פייק יכול להיות שימושי לשם הפיכת סרטים בשחור לבן לסרטים בצבע, שיפור איכות הווידאו והפיכת סרטוני חובבים למקצועיים. נוסף על אלה, השימוש הנפוץ ביותר בדיפ־פייק הוא יצירת סרטונים המבוססים על ״השתלת״ פנים של אדם אחד בסרטונים של אדם אחר. כך לדוגמה, יוצרי סדרת הטלוויזיה South Park

יצרו ב־2020 את התוכנית Sassy Justice — תוכנית קומית ששודרה ביוטיוב העשויה כל כולה דיפ־פייק סאטירי, בכיכובו (לכאורה) של נשיא ארצות הברית לשעבר דונלד טראמפ ודמויות נוספות. לפרק הראשון נרשמו מעל מיליון צפיות, והתוכנית הפכה במהרה לפופולרית במיוחד בארצות הברית (Itzkoff, 2020).

Sassy Justice, 2020
מקור: Press/YouTube

יתרה מכך, בזכות יכולות דיפ־פייק אפשר להשתמש בשחקנים בסרטים ובפרסומות גם אם אינם יכולים להגיע פיזית לאתר הצילומים, כאשר כל שנדרש הוא צילומי בסיס וכפיל גוף. כך לדוגמה, אתר הסטרימינג HULU הפיק בזמן משבר הקורונה פרסומת עתירת כוכבי ספורט ידועים מכל העולם, ללא צורך בתיאום בין לוחות הזמנים שלהם והגעתם לסט (Roettgers, 2020). עם זאת יש לזכור כי שימוש זה אינו שונה במידה משמעותית מצילום האנשים באתרים שונים וחיבור קטעי הווידאו בצורה אמינה תוך שימוש בכלי עריכה טכנולוגיים מתקדמים.

עוד ב־2020 יצא לאקרנים הסרט התיעודי Welcome to Chechnya, המתמקד ברדיפת הלהט״בים בידי הרשויות בצ׳צ׳ניה בעשור השני של המאה ה־21, ובו רואיינו פליטים להט״בים שברחו מצ׳צ׳ניה על חוויותיהם הקשות. הפקת הסרט הייתה מחויבת לביטחונם האישי של המרואיינים ולכן נמנעה מחשיפת זהותם, אך בה בעת רצתה להצמיד פנים אנושיות לסיפור. בעזרת יכולות דיפ־פייק מתקדמות הציגה הפקת הסרט מרואיינים ״מזויפים״ בעלי פנים אנושיות והבעות שנראו אותנטיים לחלוטין, ואפשרו לצופים להזדהות עם כאבם של המציגים (Thomson, 2020).

אפשר להשתמש בטכנולוגיית דיפ־פייק גם בקטעי קול: המכונה לומדת את קולו של השחקן ויכולה לדמות אותו באופן מלאכותי, למשל במקרה ששחקן מאבד את יכולתו הקולית בשל מחלה. הטכנולוגיה המתקדמת אף יכולה ליצור הלימה אמינה בין תנועת השפתיים והבעות הפנים של השחקן למוצא פיו. מדובר ביישום שימושי ביותר בהתמודדות עם בעיות נפוצות כמו שינוי מאוחר בטקסט או בתסריט או צורך בצנזורה (Hao & Heaven, 2020).

Welcome to Chechnya — a film by David France (2020)

השימוש ביצירת קולו של אדם באמצעות דיפ־פייק עורר השנה סערה בתחום הקולנוע התיעודי, עת נעשה שימוש בטכנולוגיה במסגרת הסרט ׳שף: סרט על אנתוני בורדיין׳. בשלוש סצנות מהסרט על אודות השף שהתאבד נעשה שימוש בדיפ־פייק לביצוע מניפולציה בקולו של השף המנוח, על מנת להקליט משפטים שהשף כתב אך לא אמר ולא הוקלטו בקולו. במאי הסרט בחר ליצור את קולו של השף באמצעות טכנולוגיית בינה מלאכותית במקום להיעזר בקריין, ובחירה זו עוררה ויכוח גדול לגבי עתיד תעשיית הסרטים התיעודיים (רוה, 2021).

שימוש בעל פוטנציאל רחב ביותר הוא ליצירת מצג של אמינות במקרים של דיבוב תוצר וידאו לשפות מרובות. בתעשיית הקולנוע והטלוויזיה הטכנולוגיה מפחיתה את עלות העסקת שחקנים מדובבים ויכולה לסייע ביצירת חוויית צפייה איכותית יותר עבור קהלים ברחבי העולם, שיוכלו ליהנות מהם בשפה המקומית (Panyatham, 2020). השימוש ביכולת זו נעשה גם לשם הנגשת מידע חשוב לציבור הרחב בכל העולם, דוגמת הודעתו של דיוויד בקהאם על סוגיית המלריה בתשע שפות שונות, אשר הופצה בקולו המקורי ובתנועות שפתיים התואמות את המילים הנאמרות (ThinkAutomation, n.d.).

כך גם הפוליטיקאי ההודי מאנוג׳ טיוואר (Manoj Tiwari) נעזר ביכולות דיפ־פייק על מנת ״לדבר״ בשפתם של כל תומכיו הפוטנציאליים. השפות הרשמיות בהודו הן הינדי ואנגלית, אך קיימים גם אלפי ניבים במאות השפות המדוברות בקרב אזרחים בחלקים גיאוגרפיים שונים של הפדרציה. כדי לא להחמיץ אף מצביע פוטנציאלי, טיוואר הפך את אחד מסרטוני התעמולה שלו לרב־לשוני באמצעות יכולות דיפ־פייק. התוצאה הייתה אמינה במיוחד והיה קשה עד בלתי אפשרי להבחין שמדובר בסרטון מזויף, שכן תנועות השפתיים תאמו במדויק את הדיבור, והסרטון גרף מעל 15 מיליון צפיות (חיימוביץ׳, 2020).

ב־2019 הציגה מיקרוסופט בכנס Microsoft Inspire בלאס וגאס כלי שמייצר הולוגרמה ההופכת אדם לדובר של שפה אחרת, שהוא כלל לא מכיר. בכנס נשאה דברים באנגלית מנכ״לית Azure ג׳וליה וייט שהרכיבה את המשקפיים המהפכניות של מיקרוסופט, וכך ההולוגרמה אמרה את דבריה של וייט ביפנית שוטפת באופן אמין במיוחד (Focaloid, 2019). סביר להניח כי בעתיד אף יוכלו להקליט ולהפיץ או למכור

הרצאות שלמות של מומחים בשפה שאינם דוברים כלל, וכך להנגיש תכנים הן למרצים והן למאזינים. מדובר בפוטנציאל לקפיצת מדרגה ניכרת בתחום בשילוב עם שירותי תרגום מבוססי בינה מלאכותית, שהולכים ומשתפרים (Lu, 2020).

הדגמת ההולוגרמה של וייט הדוברת יפנית, מקור: Microsoft/YouTube

באופן דומה, אתרי אומנות ותרבות בחרו להשתמש ביכולות דיפ־פייק למטרותיהם. לדוגמה, מוזיאון דאלי בסנט פטרסבורג פלורידה יצר תערוכה בשם 'דאלי חי', ובאמצעות יכולות דיפ־פייק "השיבו לחיים" את האמן הסוריאליסטי סלבדור דאלי ואפשרו למבקרים במוזיאון לקיים עימו אינטראקציה באופן שנראה אותנטי למדי. כך מעבדת הבינה המלאכותית של סמסונג במוסקבה השיבה לחיים את המונה ליזה (Jaiman, 2020).

נגישות, רפואה ופסיכולוגיה

שימושים חיוביים נוספים ביכולות דיפ־פייק החלו להתפתח גם בתחום הנגישות. חולים במחלות כמו ניוון שרירים, אשר מאבדים את יכולת הדיבור, יוכלו בזכות יכולות דיפ־פייק להשתמש במכשיר שיקריא בקולם האישי את הטקסט שיזינו לו, אם המחלה תאפשר להם לבצע זאת. כך אפשר לתרום רבות לעצמאותם של החולים ואף לשפר את מצבם הנפשי, בשל היכולת לתקשר עם קרוביהם וסביבתם גם לאחר שאיבדו את יכולת הדיבור (Jaiman, 2020).

גם בתחום הטיפול הפסיכולוגי נעשה שימוש חיובי ביכולות דיפ־פייק. כך למשל הטכנולוגיה מאפשרת למי שאיבדו את יקיריהם לנהל שיחת וידאו מלאכותית עם אהובם המנוח, בהנחיית מטפלים מוסמכים באבל ושכול. ניסוי שנעשה בהולנד ותועד בסרט בבימויה של רושן נג'ל הוכיח כי אף על פי שבני המשפחה השכולים ידעו שמדובר באשליה, הם דיווחו על שיפור משמעותי במצבם (Nejal, 2020).

נוסף על אלה נעשה שימוש ביכולות דיפ־פייק בהכשרת רופאים ובאימון אלגוריתמים של בינה מלאכותית. יכולות דיפ־פייק מאפשרות לשמור על פרטיותם של המטופלים כאשר נעשה שימוש בחולים מזויפים ובסריקות

MRI סינתטיות לחלוטין, המאפשרות לשפר את חדות האבחנה של הרופאים וכן לאמן אלגוריתמים לזיהוי גידולים ללא צורך במידע פרטי של חולים אמיתיים (ThinkAutomation, n.d.).

חינוך ושימור היסטורי

יכולות דיפ־פייק מאפשרות לגורמי הוראה, חינוך ותרבות להשיב לחיים דמויות היסטוריות באופן מסקרן ואינטראקטיבי, כך שהתלמידים יוכלו ללמוד בצורה ייחודית והידע יישאר חקוק בזיכרונם זמן רב. כך למשל מוזיאון השואה ומרכז החינוך באילינוי יצר ב־2018 סדרת ראיונות הולוגרמטיים עם שורדי שואה. המבקרים יכלו לשוחח עם שורדי השואה, לשאול שאלות ולשמוע את סיפוריהם. עם התקדמות טכנולוגיית הזיוף, היסטוריה וירטואלית מסוג זה עשויה להיות בת־השגה בקנה מידה רחב בהרבה (ThinkAutomation, n.d.).

הדמיית עדות, מקור: USC Shoah Foundation/YouTube

עוד בהקשר זה, חברת MyHeritage מציעה כלי חינמי מקוון ונגיש בשם Deep Nostalgia, שאליו ניתן להעלות תמונה משפחתית עתיקה ולקבל סרטון מונפש קצרצר של המופיעים בה (MyHeritage, n.d.).

משחקי מחשב

גם בתעשיית משחקי המחשב נעשה שימוש בטכנולוגיית הדיפ־פייק. ניתן ליצור דמויות מציאותיות במשחקי וידאו ללא צורך בהשקעה בצילומן במשך שעות וימים, אלא להשתמש בחומר גלם מצולם של השחקן וליצור על בסיסו מצבי משחק רבים ומגוונים (De Agostini, 2020). נוסף על כך השחקנים יכולים להשתמש בקול מזויף התואם לדמותם הווירטואלית, בלי לאבד את האותנטיות והאנושיות של הקול. כך יכולים השחקנים לשמור על זהותם וגם לשפר את חוויית המשחק ואת אמינותו (Somers, 2020). אפליקציות חברתיות פופולריות גם מאפשרות למשתמשים לתפעל הבעות פנים ולבצע החלפת פנים זה עם זה (Panyatham, 2020).

לצד השימושים החיוביים קיימים מספר שימושים שליליים, שניכר כי הם בעלי פוטנציאל לפגוע בצורה ישירה או עקיפה בביטחון הלאומי. להלן שלושה סוגים של שימושים כאלה ושל מקרים בולטים, כפי שפורסמו בתקשורת, שבהם נעשה שימוש זדוני בדיפ־פייק.

אלימות, סחיטה והונאה

פעילים נגד אלימות במשפחה סימנו את טכנולוגיית הזיוף העמוק כאיום ותיארו אותה כ"כלי מושלם למישהו המבקש להפעיל כוח ושליטה בקורבן" (Hao, 2021). ואכן נעשה בה שימוש ליצירת תוכן מזויף ואמין למטרות זדוניות כגון סחיטה, הונאה, השפלה והטרדה. הטכנולוגיה אף מאפשרת סחיטה של אנשים, עסקים, ארגונים או ממשלות על ידי איום בהפצת תוכן מזויף המכיל מידע כוזב כמו סרטון פורנוגרפי, הקלטות קוליות וסרטונים שבהם מוצג אדם האומר או עושה דברים שלא אמר או לא עשה, ועוד (Citron & Chesney, 2019). דוגמה להונאה מסוג זה היא הודעה קולית מזויפת שקיבל אחד מעובדי חברת היי־טק קטנה בארצות הברית, שבה ביקש לכאורה מנכ"ל החברה מהעובד לבצע עסקה דחופה ולהעביר תשלום לחשבון בנק. במקרה זה העובד זיהה כי ההודעה חשודה והעביר אותה למחלקה המשפטית, ובכך מנע את הצלחת ההונאה (Libby, 2020). למרות שלא מדובר בניסיון שונה במהותו מכזה שאפשר היה לבצע באמצעות דואר אלקטרוני למשל, השימוש בהקלטה או בסרטון המזויפים באיכות גבוהה עשוי להוביל להצלחה רבה יותר בתחום ההונאה, הן בשל מודעות נמוכה יותר מזו הקיימת היום להונאות באמצעות דואר אלקטרוני והן משום שבמצבים מסוימים של קושי לאמת או לשלול את אמיתות ההקלטה, תהיה נטייה לבצע את המתבקש בה.

מקרים מן העולם של מרמה באמצעות דיפ־פייק

המקרה הראשון המתועד של הצלחה בהונאה בעזרת טכנולוגיית דיפ־פייק קולי התרחש במארס 2019. מנכ"ל של חברת אנרגיה באנגליה קיבל שיחת טלפון שבה נשמע לכאורה מנהלו, מחברת האם הגרמנית, מבקש ממנו לבצע העברה דחופה לספק הונגרי על סך 243 אלף דולר. המנכ"ל שרומה תיאר כי הקול בשיחה נשמע זהה לחלוטין לזה של מנהלו, הוא זיהה את המבטא הגרמני ואת האינטונציה בדיבור של מנהלו ולכן פעל בהתאם להוראותיו. לאחר ההעברה הראשונה התקבלו עוד שתי שיחות טלפון, ובשלישית, לפני העברת כספים נוספת, מנכ"ל החברה שם לב כי המספר שממנו השיחות מתקבלות חשוד. הוא סיים את השיחה והתקשר למנהלו, שטען כי לא שוחח עימו טלפונית בסיטואציות שהוזכרו וכי מעולם לא ביקש ממנו לעשות עסקאות אלו. המקרה הועבר למחלקה המשפטית בחברה, אך עד שהבינו כי מדובר בהונאה מבוססת דיפ־פייק של קול היה מאוחר מדי. הכסף הועבר ועד היום לא אותרו הפושעים ולא הכסף (Stupp, 2019).

מקרה אחר אירע בינואר 2020, אז הונאה שהתבססה על טכנולוגיית דיפ־פייק הצליחה להוביל לגניבה של 35 מיליון דולר מבנק באיחוד האמירויות הערביות. החוקרים בדובאי טענו כי מנהל בנק שקיבל שיחת טלפון ממי שזיהה לכאורה את קולו בתור דירקטור של חברה שעימו שוחח בעבר, התבקש להיערך לביצוע העברות כספים בסכום של מיליוני דולרים בעקבות רכישה שהחברה מבצעת. חלק ממה שהסיר חשד ממנהל הבנק הוא העובדה שהוא קיבל מיילים שהתחזו גם הם לאותו מנהל ולעורך דין שהיה אמון על העסקה. בעקבות

השיחה והמיילים הכספים הועברו לכמה חשבונות בנק במדינות שונות. מאז נחשפה זהותם של חלק מן הנאשמים בהונאה, אולם התיק לא נסגר (הלפרן, 2021).

הטכנולוגיות הנדרשות ליצירת הונאות מסוג זה זמינות ונגישות למגוון רחב של משתמשים ואף אינן מצריכות הקלטות קוליות ממושכות של המקור כדי להתבסס עליהן. ההתפתחויות האחרונות בטכנולוגיות רשתות נוירונים עמוקות, ובייחוד בטכנולוגיית GAN, סללו את הדרך ליצירת קולות סוריאליסטיים באיכות גבוהה ביותר, שאינם מובחנים באוזני אדם (David, 2021). חוקרים טוענים כי שילוב של הטכנולוגיות והיכרות עם מבנה החברה יחד עם שימוש במניפולציה רגשית כמו הפעלת לחץ זמן יכולים לגרום לכל אדם כמעט לבלוע את הפיתיון (Noone, 2021).

מקרים אלו צריכים לשמש נורות אזהרה לעסקים וזאת בעקבות השימושים הרבים שיכולים להיות לטכניקה זו של הונאה, מעבר ליכולת הפגיעה הכלכלית שכבר הודגמה. בדרך שבוצעה הונאה זו אפשר היה להוציא במרמה ממנכ״ל החברה נתונים רגישים על עסקאות עתידיות, שמות לקוחות ועוד נושאים שעלולים להוביל לפגיעה כלכלית. בעולם של היום מאגרי נתונים יכולים להיות בעלי שווי רב; חברות כמו חברות ביטוח מחזיקות במסדי נתונים גדולים ורגישים שפרסומם יכול לסכן את לקוחותיהן. כן ניתן להשתמש במידע גם לשם סיכול פעילותה של החברה או לפגיעה בשמה הטוב (Chesney & Citron, 2019).

התבססות של גופים שונים כמו בנקים או שירותים ממשלתיים על טכנולוגיות זיהוי קולי למטרות אימות זהות, מסירת מידע ומתן שירותים עשויה להיות נקודת כשל רצינית, אם לא יבוצעו התאמות לזיהוי דיפ־פייק ולמניעת מרמה. זאת ועוד, עקב העבודה בחברות גלובליות ובעסקים חובקי עולם ולצד מגפת הקורונה, עסקים נעשו תלויים יותר ויותר בתקשורת ממרחק — טלפונית ומקוונת. שינויים אלו הופכים חברות וארגונים לפגיעים להונאות ומעלים את סיכויי ההצלחה של הפושעים (Graham, 2021). ניכר כי החקיקה הקיימת תוכל לסייע אם הזיוף שימש למרמה או לגניבה, אולם יש צורך בכלים המתאימים למניעה של תפוצת סרטוני דיפ־פייק או ליכולות של החוק להורות על בירור מקרים כאלה. כמו כן, לא ברור אם יתאפשר טיפול בתלונות על זיוף מסוג זה במקרה שלא ייגרם נזק, אלא רק תיגרם לאדם השפלה או שאמונו במציאות יתערער.

במקרה נוסף של דיפ־פייק למטרות מרמה שאינה כלכלית בוצע שימוש בסרטונים על ידי רפאלה ספון מארצות הברית. ספון השתמשה בטכנולוגיית דיפ־פייק ליצירת סרטונים של חברות קבוצת המעודדות של בתה. בסרטונים נראו הבנות מעשנות, שותות אלכוהול ואף בעירום — כל זאת במטרה להביא להדחתן כדי לשפר את מעמדה של בתה של ספון בקבוצה. את הסרטונים שלחה האם ממספרים חסויים למאמני הקבוצה, אך לאחר שהבנות נחשפו לסרטונים והכחישו כל קשר אליהם נפתחה חקירה, שבסופה נתפסה ספון והתגלה כי סרטונים אלו מזויפים (Ynet, 2021). גם מקרה זה מעלה את השאלה אילו כלים קיימים בידי הרשויות השונות, אם יידרשו לכך, להתמודדות עם מקרים המפריעים לסדר הטוב. ניכר כי שימוש בסרטונים מזויפים עלול להופיע כשאדם מתמודד על מקום עבודה או על משרה ציבורית, או בכל סיטואציה אחרת שעשויה להוביל לרווח של אחד על חשבון השפלתו או פגיעה בשמו הטוב של האחר. גם עצם האיום להפיץ סרטון עשוי לגרום לאדם שאינו מאמין ביכולתה של מערכת החוק והמשפט להגן עליו או לנקות את שמו, בפרט בסיטואציות שבהן יש רגישות לעניין הזמן, להיענות לסוחט ולו כדי למנוע את הפצת הסרטון, למרות היותו מזויף.

פורנוגרפיה

כיום השימוש הנפוץ ביותר בדיפ־פייק הוא בתעשיית הפורנו. משנת 2018 ועד דצמבר 2020, בין 90 אחוזים ל־95 אחוזים מהסרטונים שהופקו באמצעות יכולות דיפ־פייק והופצו ברשת הם בעלי אופי פורנוגרפי, ו־90 אחוזים מתוכם לא בהסכמת המופיעים בהם. ארבעת אתרי האינטרנט הפופולריים ביותר מבין אלו המפיצים פורנוגרפיה מזויפת זכו ליותר מ־134 מיליון צפיות (Patrini, 2019). למעשה אפילו המונח דיפ־פייק נטבע לראשונה ב־2017 על ידי מתכנת שהשתמש בטכנולוגיות בתחום הלמידה העמוקה (deep learning) כדי להטמיע פנים של ידועניות על גבי סרטונים פורנוגרפיים (אורפז, 2020). ללא התרעה מוקדמת, ידועניות רבות מצאו עצמן "משתתפות" בסרטים פורנוגרפיים ללא ידיעתן. אמה ווטסון הייתה אחת מהקורבנות הראשונים של תופעה זו, כאשר התפרסם סרטון פורנוגרפי עם פניה ללא ידיעתה (Stroud & Hayden, 2018).

טכנולוגיה זו הפכה את היכולת ליצור תכנים פורנוגרפיים מזויפים לקלה ונגישה. אפליקציות למטרות אלו הופיעו שוב ושוב אף שנאסרו על ידי החוק במהירות בחלק מהמדינות, למשל בווירג׳יניה ובניו יורק. אפליקציית DeepNude שעלתה בשנת 2019 ובוט מסוג דומה שהופץ בטלגרם בשנת 2020 הם רק שתי דוגמאות ידועות לאפליקציות כאלה. הקוד המאפשר את יצירת האפליקציות "המפשיטות" נשים מבגדיהן עדיין מופץ במאגרי מידע שונים (Hao, 2021). אף על פי שיצירת סרטוני דיפ־פייק התמקדה בעיקר בידוענים, אנשים פרטיים הופכים גם הם למטרה בעקבות ריבוי התמונות המתפרסמות ברשתות החברתיות ומאפשרות יצירת סרטונים אלו (Delfino, 2019). בשל כך היקף סרטוני הפורנו המיוצרים באמצעות דיפ־פייק הלך וגדל, וכעת היעדים הם לא רק ידוענים אלא גם אנשים אנונימיים לחלוטין העומדים בפני סחיטה. לדוגמה, הבוט שהופץ באפליקציית טלגרם אחראי לבדו לכ־100 אלף קורבנות, בהם גם קטינות (Hao, 2021).

מקרים מן העולם של דיפ־פייק פורנוגרפי לשימוש פוגעני

מקרה ידוע של זיוף פורנוגרפי בוצע נגד עיתונאית ידועה בהודו בשם רנה איוב, המוכרת בזכות פרסומיה על אודות מקרים רגישים הנוגעים להפרות זכויות אדם ופשעים שבוצעו על ידי פקידי ציבור וממשל. ייחודיותה של איוב נובעת מהעובדה כי מלבד היותה עיתונאית חוקרת במדינה שבה הפער המגדרי עדיין גדול מאוד, היא גם מוסלמית ואנטי־ממסדית. באפריל 2018, לאחר שראיון עימה בנושא התעללות בילדים עורר זעם רב בקרב תומכי המפלגה הלאומית ההינדית השלטת (BJP), פורסם סרטון שנערך בעזרת טכנולוגיית דיפ־פייק ובו מופיעה איוב לכאורה כשהיא מככבת בסרטון פורנוגרפי. בסרטון שהועלה לרשת האינטרנט והופץ באמצעות רשתות חברתיות נראים פניה של איוב, כאשר בעצם מככבת בסרט שחקנית פורנו צעירה. תומכי ה־BJP שזעמו על איוב שיתפו את הסרטון המזויף באתר הרשמי שלהם, והדבר הוביל לכך שהוא נצפה והופץ על ידי עשרות אלפי אנשים. דווח כי תוך 48 שעות הופיע הסרטון ביותר ממחצית מהטלפונים הסלולריים בהודו (Citron, 2020).

למרות שיש מי שיטענו כי אין בדבר נזק אמיתי, בוודאי אם הוכח שמדובר בזיוף, רוברט צ׳סני ודניאל סיטרון (Chesney & Citron, 2019) טוענים כי הלחץ הפסיכולוגי שמופעל על קורבנות של זיוף פורנוגרפי יכול להיות הרסני למדי. הקורבנות יכולים להרגיש מגוון שלם של רגשות, החל מפחד והשפלה ועד התעללות פסיכולוגית. הדבר ניכר גם במקרה של איוב. מעבר לדיווח על הידרדרות בריאותה הפיזית בימים הראשונים

לפרשה, פרסום הסרטון גרם לה תחושת השפלה כה קשה שמנעה ממנה ככל הנראה את היכולת לתפקד. היא נמנעה מיציאה מביתה ואף הפסיקה את פעילותה כעיתונאית ואת הפעילות ברשתות החברתיות למשך חודשים ארוכים. לאחר ההתאוששות מהמקרה חזרה בהדרגה לעבודתה, אך בניגוד לעבר היא החלה להיזהר בכל פרסום ולצנזר את עמדותיה מתוך חשש מתמיד שהסרטון יופץ בשנית (Ayyub, 2018).

בראיון נוסף עימה היא סיפרה כי לאורך כל הקריירה העיתונאית שלה היא עומדת מול ניסיונות השתקה ופגיעה במוניטין שלה בדרכים שונות, אך פרסום הסרטון היה הפעם הראשונה שהצליחו לשבור אותה; הוא הצליח להחדיר בה פחד שלא הכירה לפני כן (The World, 2019). הדבר נובע מתפיסת האיום המוחשי שנוצרת אצל קורבנות דיפ־פייק פורנוגרפי — הסרטון ממחיש לקורבן באופן ויזואלי את התממשות האיום ואת הפגיעה בו (Chesney & Citron, 2019).

איוב סיפרה כי מהרגע שראתה את הסרטון, היא ידעה שבמדינה כמו הודו תהיה לו השפעה הרסנית על מעמדה. ואכן פרופיל הפייסבוק וחשבון הטוויטר של איוב הוצפו מייד באיומים קשים הקוראים לפגוע בה, לאנוס ואף לרצוח אותה (Citron, 2020). ההשפעה החזקה של הסרטון נבעה בין היתר מתהליכים בהודו, המתאפיינים בעליית הלאומנות והעוינות כלפי מיעוטים בכלל ומוסלמים בפרט. אקלים פוליטי זה, המנציח נרטיבים העולים בקנה אחד עם האידיאולוגיה הפוליטית השלטת, אִפשר לסרטון הפורנו המזויף של איוב להפוך לפופולרי ואמין בקרב הציבור, כך טוענת איוב עצמה (Ayyub, 2018). נוסף על כך, הודו אינה המדינה הליברלית ביותר בכל הנוגע לזכויות נשים ולחירויות מיניות. לכן, כאשר איוב ניסתה להגיש תלונה במשטרה ולקבל הגנה מרשויות המדינה, התעלמו מבקשתה במשך חודשים ארוכים עד שהוגש לממשלת הודו דוח מיוחד מהאו״ם, הדורש מהמדינה לספק לאיוב את ההגנה המגיעה לה (Citron, 2020).

סרטוני דיפ־פייק פורנוגרפיים שנועדו להשתיק, להפחיד או לנצל נשים הפכו לתופעה מקוונת חדשה בכל רחבי העולם ולא רק בהודו. גם בחברות דמוקרטיות שבהן יש שוויון מגדרי ברמה גבוהה יחסית, נשים מתקשות מאוד להילחם בתופעה (Farago, 2019). הלן מורט, כתבת ושדרנית בריטית נפגעה גם היא מתופעה זו. בסוף שנת 2020 גילתה מורט כי סרטוני פורנו וסרטוני אלימות מינית עם שמה ופניה מופצים באינטרנט, ואף מופץ מאגר תמונות שלה בקריאה לעוד יוצרי סרטוני דיפ־פייק ליצור סרטונים מסוג זה.

היא מתארת כי ההשפעה של מסע ההשפלה שאותם אנשים העבירו אותה הייתה עמוקה מאוד. היא שקלה למחוק לגמרי את כל חשבונותיה ברשתות החברתיות ובאינטרנט, והיא גם סובלת מפגיעה פסיכולוגית ומסיוטי לילה בעקבות חשיפתה לסרטונים שבהם נראה כי היא עוברת התעללות קשה. גם היא נתקלה במבוי סתום כאשר ניסתה לקבל את תמיכת הרשויות. בדומה לאיוב ניגשה מורט למשטרה, אך נאמר לה כי אין ביכולתם לעשות דבר לסייע לה. זאת מכיוון שהפצת סרטוני פורנו מזויפים אינה נחשבת כיום עבירה על החוק באנגליה (Hao, 2021; Jacson, 2021).

בשני מקרי הבוחן לא נמצא האדם העומד מאחורי יצירת הסרטון, ולכן גם אי אפשר למצוא פרטים על הטכנולוגיות שבהן השתמשו לשם יצירת הסרטונים. עיתונאי ניסה לבחון מה נדרש לשם יצירת פורנו דיפ־פייק. הוא מצא כי פורומים של יוצרים בעזרת טכנולוגיה זו נמצאים כבר בעמוד הראשון בחיפוש בגוגל, ובעזרת סרטון קצר של הקורבן והסכום הזעום של 30 דולר, ניתן להזמין סרטון פורנו מזויף של כל אדם (Jacoby, 2019). כמו כן, רוב סרטוני הזיוף שנערכו על ידי יוצרים עצמאיים מתבססים על אבולוציה של קוד המקור

ליצירת סרטוני זיוף פורנוגרפיים. קוד זה נוצר ופורסם על ידי משתמש בפורום Reddit בשנת 2017, שעל שמו אף נקראה הטכנולוגיה. משתמש זה התבסס על הספרייה הפתוחה של גוגל בנושא למידת מכונה ופיתח קוד ליצירת סרטוני פורנו מזויפים. לפיכך אפשר להעריך כי על בסיס קוד זה נוצר גם הסרטון במקרה של איוב ובמקרים דומים נוספים (Adee 2020).

שני המקרים שהוצגו כאן מעידים על חוסר היערכות של גורמי חוק במדינות שונות לטיפול בסוגיות כאלה. ניכר כי גם החקיקה בנושא חסרה במדינות רבות וכן חסרים כלי בדיקה ואכיפה, חלקם טכנולוגיים שאינם נמצאים בידי האמונים על שמירת החוק, למרות שניכר כי חלקם זמינים ממש כמו אלו המשמשים ליצירת הזיוף עצמו.

השפעה על פוליטיקה וביטחון לאומי

טכנולוגיית הזיוף העמוק מאפשרת שימושים שלחלקם השלכות בתחום הביטחון הלאומי והפוליטיקה. מדובר בסכנה מטרידה לפגיעה בחברה האזרחית או במדינות, על ידי יצירת מצגים שקריים למטרות עיצוב תודעה ודעת קהל כדי להשפיע על המערכת הפוליטית, על תהליכי בחירות ואולי אף על יציבותן של מדינות (Sayler & Harris, 2021). הזיוף העמוק הוא למעשה יישום טכנולוגי נוסף המעצים את אתגרי התקופה הקיימים בתחום הביטחון הלאומי, כאשר סרטוני הדיפ־פייק יכולים לייצר מצג כוזב שעלול להטעות את דרג מקבלי ההחלטות ולהשפיע על דעת הקהל והציבור בהתאם לאינטרסים של יוצריהם (Schiff, 2019). לאור כל אלה, השימוש ביכולות דיפ־פייק סומן על ידי ארגוני המודיעין האמריקאיים בשנת 2019 כאיום האסטרטגי הגדול ביותר על הביטחון הלאומי (Konkel, 2019 ;Ng, 2019).

השפעת סרטונים אלה על דעת הקהל והפוליטיקה העולמית עשויה להיות רבה, משום שלמרות ההתפתחויות הטכנולוגיות המתקדמות בהפקת סרטונים או קטעי שמע סינתטיים בשנים האחרונות, אנשים עדיין נותנים אמון רב בדברים מוחשיים שראו או שמעו, בייחוד כאשר מדובר בפנים ובקולות של אנשים מוכרים להם (Keitzmann et al., 2020). עם זאת, רק חלק קטן מתוך סרטוני הדיפ־פייק נועד לתמרן מצב פוליטי; רוב הסרטונים של דמויות פוליטיות כיום משמשים למטרות פרודיה או לחינוך הציבור לזהות סרטונים סינתטיים הנוצרים בטכנולוגיה זו (Dunn, 2021).

לפי דוח חברת Deeptrace שפורסם באוקטובר 2019, מבין סרטוני הזיוף העמוק שאינם פורנוגרפיים רק 12 אחוזים מהדמויות שהופיעו בהם היו של פוליטיקאים (Patrini, 2019). מכיוון שנדרשת כמות גדולה של נתונים לשם יצירת תוכן שלא יזוהה כזיוף, כיום עיקר השימוש בדיפ־פייק מתמקד בדמויות מוכרות, אך עם התפתחות הטכנולוגיה הצפי הוא שהטכנולוגיה תהיה נגישה יותר וגם אנשים אנונימיים יהפכו לקורבנותיה.

מקרים מן העולם של ניסיון להשפעה פוליטית באמצעות דיפ־פייק

בסתיו 2018 דווח כי נשיא מדינת גבון עלי בונגו חלה במהלך ביקור בערב הסעודית ונעלם מהחיים הציבוריים, בזמן שהמצב הפוליטי והכלכלי במדינה היה שברירי. חוסר במידע ודיווחים לא עקביים על מצבו יצרו חרושת שמועות על מצבו הבריאותי ואף על מותו של הנשיא. בערב ראש השנה 2019 פרסם הממשל בגבון סרטון שבו הנשיא נראה לכאורה נואם לאומה, לראשונה מזה שלושה חודשים (Brady, 2020). בסרטון נראה כי

תנועותיו של הנשיא מעט מכניות וקפואות ותנועות עיניו מוגבלות. מייד לאחר מכן נמסר בחדשות כי הסרטון זויף בעזרת טכנולוגיית דיפ־פייק, וכי השמועות בדבר מותו ככל הנראה נכונות. כשבוע לאחר מכן מתנגדיו של בונגו ניצלו את המצב וביצעו הפיכה צבאית שכשלה. בבדיקות שנעשו לאחר הפרשה לא נמצאה הוכחה שהסרטון היה מזויף, ועלו חשדות כי הופעתו השונה של הנשיא נבעה אולי מאירוע מוחי שעבר בתקופה זו (אורפז, 2020).

מקרה זה מדגיש את הפוטנציאל של השלכות טכנולוגיית דיפ־פייק על יציבותם של משטרים בעזרת ערעור תפיסת האמת. המקרה הוכיח כי ניתן ליצור השפעה לא רק באמצעות שימוש בטכנולוגיה, אלא עצם הידיעה כי הטכנולוגיה קיימת ואפשרית עשויה להוביל לערעור האמון גם במיצגים אמיתיים. במקרה של גבון ניכר כי תפיסה מעוותת זו תרמה למשבר פוליטי ולניסיון הפיכה צבאית (Delcker, 2019). עם זאת יש לשים לב לכך שנדרשות נסיבות ייחודיות על מנת שלסרטון מזויף תהיה השפעה פוליטית. במקרה של גבון הייתה השפעה משמעותית להיעדרות הממושכת של מנהיג מן הזירה הציבורית. נוסף על כך בונגו נחשד בעבר בעבירות הונאה ושחיתות — מה שהשפיע על אמון הציבור בפרסומיו (Brady, 2020). בד בבד עם המצב המדיני הרעוע במדינה מלכתחילה, ניכר כי במקרה מבחן זה היה מצע מצוין להשפעה שלילית של סרטון מזויף או לחשדות בסרטון מהימן.

מקרה אחר התרחש בישראל במסגרת הבחירות לכנסת ה־24, כשנעשה שימוש בקטע מזויף שבו נראה יאיר לפיד, מנהיג מפלגת יש עתיד, אומר דברים שלא אמר במציאות ואשר עלולים היו לפגוע בבחירה בו, משום שציירו אותו כמי שישתף פעולה עם פלגים מסוימים בחברה הישראלית. בסרטון הופיע סימון קטן שהעיד על היות הקטע המדובר מזויף, אולם לא היה קל להבחין בו. מפלגת יש עתיד עתרה לוועדת הבחירות המרכזית נגד הסרטון אולם לאחר העתירה, עוד בטרם פרסום הפסיקה בעניין, בוצע שינוי נוסף בסרטון על ידי יוצריו ונוספה בו שקופית גדולה המבהירה כי מדובר בזיוף לצורך המחשת רעיון. לאור זאת פסקה ועדת הבחירות כי אין בסרטון משום הפרה של הוראת סעיף 13 לחוק, המדברת על הטעיה מכוונת של הציבור, ולכן לא נדרשה הסרתו מרשת האינטרנט (תב״כ 24/9, 2021).

מקרה של זיוף נוסף לצורך השפעה פוליטית עירב שימוש בפורנוגרפיה. המפרסמים ניסו להשפיע על הבחירות לתפקיד מושל סאו פאולו, ברזיל. בסרטון שהופץ נראה המועמד המוביל באותה עת, ג׳או דוריה, מקיים לכאורה יחסי מין עם חמש נשים צעירות. הסרטון הופץ ברשתות החברתיות ימים ספורים לפני מועד הבחירות. דוריה הכחיש כל קשר להאשמות וטען כי זהו סרטון מזויף שמטרתו פגיעה פוליטית. מדובר היה בפוטנציאל לפגיעה קשה במועמד, בעיקר בגלל הדעות השמרניות וחשיבות ערכי המשפחה שהבליט לאורך מסע הבחירות. דוריה נבחר כשם שניבאו הסקרים לפני הופעת הסרטון (Birchall, 2018). מקרים כאלה יכולים גם ללמד כי הטענה ל״זיוף״, כאשר לציבור קשה להבחין או לבדוק בעצמו אם מדובר בזיוף או באמת, עשויה להוות מפלט קל גם למי שסרטון או הקלטה אמיתיים שלו יפורסמו ויביכו אותו. לשם כך נדרשים כלים לאימות ולמעורבות מצד העיתונות החוקרת ומוסדות המדינה הרלוונטיים, שיוכלו לאמת את המידע.

זאת ועוד, ראוי להתייחס גם להשפעה פוליטית בינלאומית פוטנציאלית. אף שכיום אין תיעוד גלוי של שימוש בדיפ־פייק שהוביל למשל להשפעה צבאית על הזירה הבינלאומית, על פי טוד מאיירס, מוביל אוטומציה של מנהלת הטכנולוגיה במודיעין הגיאו־מרחבי הלאומי של ארצות הברית, סין היא המובילה העולמית ביצירת

סרטוני דיפ־פייק באמצעות שימוש בטכניקת GANs. ב־2017 השתמשו חוקרים סינים ב־GANs כדי לזהות כבישים, גשרים ועצמים אחרים בתמונות לוויין והנדסו את תצלומי הלוויין כך שישרתו אינטרס — כמו יצירת גשרים מזויפים לחלוטין כחלק מתמונת לוויין שנראית אותנטית לחלוטין. מאיירס טען שהצבא והמודיעין האמריקאי יכולים להביס את טכניקת ה־GANs, אך לשם כך דרושים מאגרים כפולים ומשוכפלים של תמונות לוויין וראיות מחזקות נוספות (Tucker, 2019).

נכון לשנת 2021, עדיין יש צורך במחשוב יקר וביצרנים מנוסים על מנת לייצר סרטון אמין או זיוף של צילומי לוויין בצורה אמינה (Toews, 2020). אך כשמדובר במקרים של פוטנציאל להשפעה פוליטית, עשויים להיות בעלי אינטרסים שברשותם האמצעים הנדרשים ליצירת סרטונים כאלה. השאלה הנותרת היא מה תהיה רמת ההשפעה של סרטון כזה. התשובה היא ככל הנראה שגם אם סרטון יהיה אמין דיו, יש צורך בהפצתו בסיטואציה מתאימה כמו למשל חוסר יכולת של הגורם המופיע בסרטון להכחיש, או בסיטואציה שבה יש קריאה לפעולה שתוביל לפעולה מיידית, או במקרים שבהם יש חשיבות לזמן (למשל זמן קצר לפני ההצבעה במערכת בחירות, בצורה שאינה מאפשרת בירור). בשלל סיטואציות כאלה עשויים המפיצים להשיג את מטרותיהם, יהיו אשר יהיו, במחירים נמוכים יחסית. עם זאת, עדיין לא ידוע על מקרה מוצלח שבו ניסתה מדינה אחת להשפיע על אחרת בעזרת סרטון מזויף שנוצר באמצעות למידה עמוקה.

לבסוף, האתגר שמציב דיפ־פייק לביטחון הלאומי נוגע בעיקר לסכנה הפוליטית הנשקפת. לאור זמינות הטכנולוגיה ופשטות היצירה של תוצרי דיפ־פייק שונים, קיימים שלל תרחישים שעלולים להוביל לערעור הסדר הציבורי, לאלימות או אפילו לאובדן אמון הציבור גם במסרים אמיתיים שדרושים לניהולה התקין של חברה בכלל, ושל חברה דמוקרטית בפרט.

דיפ־פייק — האם הביטחון הלאומי מתערער?

התבוננות בשימושים החיוביים והשליליים של טכנולוגיית הזיוף העמוק ובמקרי המבחן שנסקרו לעיל מעלה את הדיון בשאלה: האם אכן יש פוטנציאל לערעור הביטחון הלאומי בעקבות עליית האיכות והזמינות של טכנולוגיית הזיוף העמוק? מחד גיסא, מדובר בתופעה חדשה שעומקי השפעתה אינם ידועים עדיין. מאידך גיסא, גם בעבר התמודדו פרטים, ארגונים ומדינות עם דרכים שונות ומתוחכמות של יצירת זיופים למטרות שונות.

מעבר לכך, בניסיון להשיב על השאלה אם יישומי דיפ־פייק מערערים את הביטחון הלאומי, יש צורך לחזור לשאלת יסוד: מה נחשב לביטחון לאומי? אומנם מקובל לרוב להתייחס לביטחון הלאומי במובן הצר של ההגדרה — היכולת של אומה להגן על אזרחיה ועל ערכיה הפנימיים מפני איומים, ביניהם מדינות עוינות וארגוני טרור (הכהן, 2014), אולם אף שההתייחסויות ההיסטוריות והתאורטיות לסוגיה מדגישות לרוב היבטים של צבא ושל יחסי חוץ, האו״ם למשל כולל בהגדרת הביטחון הלאומי שבעה נדבכים: כלכלה, מזון, בריאות, איכות סביבה, ביטחון אישי, קהילה וביטחון פוליטי (United Nations Development Programme, 1994). הגדרה זו מרחיבה את ההתייחסות גם ליכולתו של הפרט לנהל בביטחון וברווחה את שגרת חייו.

גם במסגרת מחקר זה ההתייחסות היא להגדרתו המרחיבה של המושג ביטחון לאומי, משום שבעידן הנוכחי, בעיקר במדינות מפותחות ובדגש על דמוקרטיות ליברליות מפותחות, הציפייה של אזרחים אינה רק להגנת המדינה עליהם מפני איומים צבאיים אלא גם ליכולת לחיות בביטחון כלכלי, תזונתי, בריאותי ואישי. חלק מן הנושאים הללו הודגשו היטב ככאלה שאינם מובנים מאליהם, עם פרוץ מגפת הקורונה בעולם.

מקרי המבחן שהוצגו בפרק הקודם מדגישים את פוטנציאל ההשפעה הנרחב של טכנולוגיית הדיפ־פייק על תחומי חיים רבים. זאת בין היתר לאור ההנחה הרווחת כי מאז שנת 2016 כי אנו חיים בעידן של פוסט־אמת ופייק ניוז (ברון ורויטמן, 2019). המושג פוסט־אמת מתייחס לתרבות פוליטית שבה עובדות אובייקטיביות משפיעות על עיצוב דעת הקהל פחות מרגשות, אמונות ודעות, והמושג פייק ניוז מתייחס לקלות שבה ניתן כיום להפיץ שקרים, עיוותים, טעויות ותאוריות קונספירציה (ברון ורויטמן, 2019). האתגר שמציבות תופעות הפוסט־אמת והפייק ניוז לביטחון הלאומי ולדמוקרטיה הוא למעשה החשש מעולם שבו קשה עד בלתי אפשרי להבחין בין אמת לשקר, בין סחרירים (ספינים) ומאמצי השפעה לבין עובדות, ולצד זאת החשש מתהליך קבלת החלטות שבו לניתוח המקצועי מבוסס העובדות יש פחות השפעה מזו של רגשות, אמונות, דעות ושקרים. לכל אלה עלולה להיות השפעה מרחיקת לכת על ביטחון לאומי במובנו הרחב.

כוח השכנוע והפוטנציאל הוויראלי של הדיפ־פייק, להבדיל מפייק ניוז, נובע מכך שדיפ־פייק נתפס כ״תיעוד ישיר״ של התרחשות האירועים במציאות. כלומר, אנשים נוטים להאמין לדברים שהם רואים בצורה מומחשת וממשית. הקלטות קול וסרטוני וידאו נתפסים על ידי בני אדם כראיה אמינה ביותר, כשם שהיטיב לבטא זאת ב־2019 הסנטור אנגוס קינג: ״videos do not lie״. אך בעולם של דיפ־פייק, התפיסה המסורתית ש״לראות או לשמוע שווה להאמין״ אינה קיימת עוד (Chesney & Citron, 2019).

ב׳ברומטר האמון של אדלמן׳ שפורסם בשנת 2020, אשר סקר 34 אלף בני אדם בוגרים מרחבי העולם, נמצא כי 66 אחוזים מהנסקרים ״מוטרדים מכך שטכנולוגיה הופכת לבלתי אפשרית את הידיעה אם מה

שאנשים רואים ושומעים הוא אמיתי", וכי 61 אחוזים מהנסקרים סבורים ש"ממשלות לא מבינות מספיק את הטכנולוגיות המפציעות כדי להחיל עליהן רגולציה יעילה" (2020 Edelman Trust Barometer, 2020). השימוש ביכולות דיפ־פייק אכן מוביל בין היתר לערעור האמת ומעורר חשד שכל פיסת תוכן היא שקרית, גם אם היא אמיתית (אורפז, 2020). כך למשל אפשר לטעון שהקלטה היא שקרית והופקה באמצעות יכולות דיפ־פייק, ובכך להתנער מאחריות. לאור זאת, יכולות הזיוף העמוק מעצימות ומחריפות את המצב בעידן הפוסט־אמת והפייק ניוז שבו אנו נמצאים, ואת האתגרים שהוא מציב לאמת ולדמוקרטיה (Villasenor, 2019).

ככל שמתפתחת טכנולוגיית הזיוף העמוק כך מתפתחים ומשתכללים גם הכלים לזיהוי הזיוף העמוק. אולם יש צורך גם בחינוך הציבור וכן ברגולציה ובכלים משפטיים להתמודדות עם התופעה — אמצעים שאינם רווחים דיים ברוב המדינות בעולם, כפי שיוצג בפרק הבא.

פתרונות ודרכי התמודדות בעולם

מדינות, גופי תקשורת, חברות ואנשים פרטיים מתמודדים בשנים האחרונות עם האיום שמציבות זמינותן ונגישותן ההולכות וגוברות של יכולות דיפ־פייק. בפרק זה סקירת פתרונות ודרכי התמודדות מן העולם, הנחלקים לשלוש קטגוריות: מאמצי רגולציה, חקיקה ואכיפה; פתרונות טכנולוגיים; מאמצים לחינוך הציבור.

רגולציה, חקיקה ואכיפה

ככל שהשימוש בפייק־ניוז ובדיפ־פייק מתעצם, כך גוברים המאמצים לפתח כלים ושיטות לסיכול האיום. סוכנויות מודיעין ורשויות ממשל במדינות דמוקרטיות בעולם משתתפות באופן פעיל במאמץ זה על ידי הקצאת תקציבים לפיתוח הטכנולוגיות לזיהוי תוכן מזויף ולהתמודדות עם התופעה (Parkin, 2019). לצד מתן תקציבים למציאת פתרונות טכנולוגיים, רשויות ממשל במדינות דמוקרטיות ברחבי העולם מגלות גם מעורבות משפטית בנושא. יצוין כי תהליכים קשורים מתרחשים גם במדינות לא־דמוקרטיות שבהן עשויה להיות השפעה לדיפ־פייק, ביניהן סין, אולם המידע בנוגע לאלה גלוי פחות, והאתגרים שונים במדינות שבהן התקשורת מוגבלת.

האתגר המשפטי והיישומי של התמודדות עם סכנות הפייק ניוז והדיפ־פייק בדמוקרטיות בפרט נובע מקצב ההתפשטות הגבוה, הנרחב והמגוון שלהן ומהיכולת המוגבלת לזהותן ולהגבילן באמצעים טכנולוגיים. לאחרונה ניתנה בישראל אף הגדרה משפטית לדיפ־פייק, הממחישה את הצורך של המשפט להתייחס אליה ולסכנותיה בצורה ייחודית. היא ניתנה במסגרת פסיקה של שופט בית המשפט העליון עוזי פוגלמן בנוגע לעתירה שהוגשה לוועדת הבחירות המרכזית לכנסת ה־24 ומגדירה דיפ־פייק כך:

> טכנולוגיה ליצירת תוכן קולי או חזותי או לשינוי תוכן קיים, כך שהצופה הסביר (ואף הצופה המתוחכם) יסבור כי פלוני ביצע פעולה או העביר מסר, אך התוכן אינו אמיתי. התוכן הוא באיכות גבוהה עד כדי כך שמשתמש מן היישוב יתקשה לרוב לגלות שמדובר בזיוף (תב״כ 24/9, 2021).

לצד מאמצי חקיקה ואכיפה אפשר לראות ברחבי העולם מאמצי רגולציה מצד ממשלות, הן על פיתוח מוצרי דיפ־פייק והן על הפצתם. בסין כבר מחייבים ואוכפים סימון ברור של תוצרי דיפ־פייק, ובארצות הברית הונחה של שולחנו של הקונגרס ביוני 2019 הצעת חוק שעיקרה חובת סימון ושקיפות כשמדובר בדיפ־פייק, אשר נותרה בעינה מאז (ברון ואלטשולר, 2019; Statt, 2019).

סוגיה מטרידה נוספת הנובעת מאיום הדיפ־פייק נוגעת לעובדה שקיומה של הטכנולוגיה שוחק את האמון בראיות וידאו ועלול לערער את הערך הראייתי שלהן בבית המשפט. תמונות וסרטונים הם אופן שכנוע עוצמתי, שכן ייצוג חזותי נתפס מאז ומתמיד על ידי בני אנוש כבטוח ואמין. במערכת המשפט האמריקאית, למשל, תמונות וסרטונים דיגיטליים יכולים לשמש עדויות וראיות קבילות כל עוד ניתן לאמת את מהימנותם. המשמעות היא שמערכת המשפט צריכה למצוא את הכלים המתאימים ולפתח טכנולוגיות מקבילות כדי לזהות ולאתר דיפ־פייק, בעיקר בעולם שבו הטכנולוגיה משתכללת בהתמדה, ועד מהרה הסרטונים המזויפים יהיו כה אמינים עד שעין אנושית לא תוכל להבחין בזיוף (Maras & Alexandrou, 2018).

גם בישראל, ראיות כגון תמונות וסרטונים דיגיטליים עשויות להיות קבילות בתנאי שאושרו על ידי בית המשפט, הן מבחינת אמינות והן מבחינת קבילות האופן שבו הושגו. לכן מדובר באתגר בכל הנוגע להקלטות ולראיות מתחום הווידאו, שניכר כי הגורמים הרלוונטיים עדיין לא נתנו עליו את הדעת.

ההתמודדות בארצות הברית

עיקר הקושי להתמודד עם התוצר הטכנולוגי נובע ככל הנראה מכך שהגבלתו עשויה לפגוע בחופש הביטוי. ב־2019 כתבו סיטרון וצ׳סני כי אין בעולם משטר או חוק פלילי האוסר על יצירה או הפצת תוכן הנוצר באמצעות דיפ־פייק, ודנו בבעיות העקרוניות הכרוכות באיסור כללי כזה. בראש ובראשונה הם ציינו כי איסור כללי אינו רצוי משום שעצם המניפולציה הדיגיטלית על תוכן היא לא הבעיה, שכן יש לה מגוון יישומים חיוביים. שנית, איסור כללי ימנע התפתחויות חשובות נוספות בתחומי החדשנות הדיגיטלית. לבסוף, איסור כללי על יצירת תוכן דיפ־פייק והפצתו מתנגש עם התיקון הראשון לחוקה, המגן על חופש הביטוי.

פסקי דין קודמים בארצות הברית קבעו כי גם במקרה שהמידע המופץ כוזב, עדיין חלה עליו ההגנה של חופש הביטוי. ב־1964 קבע בית המשפט העליון בארצות הברית במשפט ניו־יורק טיימס נגד סאליבן כי דברי כזב נהנים מהגנה חוקתית, מהסיבה שאיסורם היה מהווה פגיעה בחופש הביטוי. ב־2012 בית המשפט העליון בארצות הברית הוסיף וקבע כי יש להגן על שקרים, שכן תפקידם לעורר הפרכה ושיח מנומק שהוא חיוני (Chesney & Citron, 2019).

מאז פרסום מאמר זה גילו מדינות שונות בארצות הברית מעורבות משפטית חקיקתית בנושא. טקסס הייתה המדינה הראשונה בארצות הברית שאסרה תוכני דיפ־פייק כאשר חוקקה בספטמבר 2019 את החוק SB751, שלפיו יצירה או הפצה של תוכני דיפ־פייק במטרה לפגוע במועמד לבחירות או להשפיע על תוצאותיהן בשלושים הימים שטרם הבחירות מהווה עבירה פלילית (Salazar, 2019). באוקטובר 2019 חתם מושל קליפורניה גאווין ניוסום על חוק AB730, שלפיו אין זה חוקי להפיץ סרטוני וידאו שעברו מניפולציה במטרה לפגוע בדמותו של מועמד פוליטי ולהונות מצביעים, בטווח של 60 יום ממועד בחירות. מארק ברמן, חבר המועצה של קליפורניה שאישרה את החוקים, אמר כי ״לבוחרים קיימת הזכות לדעת מתי קטעי וידאו, קול ותמונות שמראים להם, שמטרתם להשפיע על בחירות מתקרבות, עברו מניפולציה ואינם לקוחים מהמציאות״.[1] ברמן הוסיף כי ״בהקשר של בחירות, היכולת לייחס מלל או התנהגות שאינם אמיתיים למועמד כלשהו הופכת את הדיפ־פייק לכלי עוצמתי ומסוכן״[2] (Berman, 2019).

נוסף על כך, בדצמבר 2019 חתם נשיא ארצות הברית דאז דונלד טראמפ על החוק הפדרלי הראשון הקשור בדיפ־פייק. הצו עסק בהיערכות האמריקאית הנדרשת לאיום זה וכלל התייחסות להערכת היכולות הטכניות של ממשלות זרות בנוגע לזיופים עמוקים, ניתוח האיום וההשלכות ומתן תמריץ כספי לפיתוח כלים לאיתור ולזיהוי של דיפ־פייק (Ferraro et al 2019).

1 ״Voters have a right to know when video, audio, and images that they are being shown, to try to influence their vote in an upcoming election, have been manipulated and do not represent reality.״

2 ״In the context of elections, the ability to attribute speech or conduct to a candidate that is false — that never happened — makes deepfake technology a powerful and dangerous new tool in the arsenal of those who want to wage misinformation campaigns to confuse voters.״

אולם הצעדים הללו, מעודדים ככל שיהיו, אינם מייצגים את שאר העולם (Feeney, 2021). כך למשל במערכת הבחירות האחרונה בישראל בשנת 2021 התמודדה מערכת המשפט הישראלית עם פרשיית דיפ־פייק, ובחרה לסרב להורות על הסרת הסרטון בשל החשיבות הרבה לשמירה על חופש הביטוי (תב״כ 24/9, 2021).

נוסף על חוק AB730 ניוסום חתם גם על חוקAB602 , אשר מעניק לתושבי קליפורניה את הזכות לתבוע אדם שיצר דיפ־פייק שבו מוטמעת דמותם בחומרים פורנוגרפיים שנוצרו בלא הסכמתם (Sheller, 2019). זוהי אחת המדינות הראשונות שנותנת כלי משפטי בידי נפגעי דיפ־פייק פורנוגרפי. במקומות רבים אחרים בעולם אין ודאות כי אדם שיתלונן יזכה לסעד מכוחות אכיפת החוק ומערכת המשפט, שכן לרוב אין להם מודעות כמו גם כלים מתאימים לכך, ובהם חקיקה רלוונטית.

גם במקרים שיהיה רצון לתת סעד לפונה בעניין של דיפ־פייק, איתור האשם והעמדתו לדין נתקלים בשני מחסומים עיקריים. הראשון הוא אתגר טכנולוגי הנובע מאופן יצירת דיפ־פייק, המאפשר ליוצריו להישאר אנונימיים. לפיכך גם אם יימצא חשוד ביצירת הסרטון, הוכחת האשמה תהא מאתגרת מאוד באמצעים הקיימים, ואולי אף בלתי אפשרית. המחסום השני מתייחס לכך שגם אם אותר הנאשם ביצירת התוכן המזויף, היכולת להעמידו לדין אינה פשוטה כלל. בארצות הברית, התקנות שעל בסיסן ניתן לתבוע את הנאשם והן בעלות הפוטנציאל הגבוה ביותר להצלחה הן תביעה בגין לשון הרע, הוצאת דיבה וגרימה מכוונת למצוקה רגשית. לתביעות על רקע הפרת זכויות יוצרים, זכויות פרסום ופרטיות יש סיכויים מוגבלים יותר להצלחה. לבסוף, במקרים מסוימים לא תתאפשר העמדה לדין של יוצר או מפיץ התוכן המזויף, אם הוא שוהה מחוץ לגבולות השיפוט האמריקאיים (Chesney & Citron, 2019).

לאור האתגר הגדול הכרוך בהוכחת האחריות ובהטלת אשמה על יוצרי התכנים, נראה כי הדרך היחידה להרתעה ולתיקון עוולות היא ייחוס האחריות לפלטפורמות הפרסום שבהן הם מופצים. אולם, סוגיה זו מעלה אף היא קשיים עקב סעיף 230 לחוק הגינות התקשורת, המעגן בארצות הברית את חסינות הפלטפורמות מפני נשיאה באחריות לתכנים המתפרסמים בהן, וזאת במטרה להגן על חופש הביטוי ולשמור על ריבוי דעות ברשת (Chesney & Citron, 2019). כך השילוב של התיקון הראשון לחוקה וסעיף 230 בחוק הגינות התקשורת פועל כחסם עוצמתי בפני מאבקם החוקי של קורבנות סרטוני דיפ־פייק פורנוגרפיים בארצות הברית, כאשר הם מגינים יחד הן על יוצרי התוכן והן על הפלטפורמות שבהן התכנים מופצים. עד כה, קיטוב חברתי, ביטויי שנאה ופייק ניוז לא גרמו למחוקקים לבטל את הפטור שממנו נהנות הפלטפורמות, אך ייתכן שדיפ־פייק יהיה קו פרשת המים להטלת אחריות כזאת (ברון ושוורץ אלטשולר, 2019).

מאמרה של אן פצ׳ניק גיזֶקֶה משנת 2020 מציע תיקון לחוק הגינות התקשורת בארצות הברית, כך שיאפשר לתבוע יצרנים ופלטפורמות המשתמשים לרעה בטכנולוגיית דיפ־פייק, תוך ניסוח החוק בצורה המבדלת אותו מהגנות התיקון הראשון ומתייחסת נקודתית לסרטוני דיפ־פייק פורנוגרפיים, בשילוב מתן חסינות לפלטפורמות הנאבקות להגן על הקורבנות מפני הטרדות ומשתפות פעולה בהסרת הסרטונים הפוגעניים. תיקון זה עשוי לספק שתי דרכים יעילות להילחם בתופעה: הראשונה היא אפשרות לתבוע את יוצרי הסרטונים; השנייה היא הסרה מהירה של הסרטונים הפוגעניים על ידי תמריץ חיובי לשיתוף פעולה מצד הפלטפורמות (Pechenik Gieseke, 2020).

ההתמודדות באיחוד האירופי

על הנעשה באירופה אל מול איום הדיפ־פייק אפשר ללמוד מדוח שפרסם הפרלמנט האירופי ביולי 2021. בהחלטת הפרלמנט ממאי 2017 הוא קרא לנציבות האירופית לנקוט אמצעי רגולציה קשיחים להתמודדות עם פייק ניוז. ההחלטה הפרלמנטרית ממאי 2018 בנושא פלורליזם תקשורתי קיבלה מספר המלצות שלא היו קשורות לדיפ־פייק אך הן רלוונטיות גם לגביו, ביניהן: שקיפות מלאה בשימוש באלגוריתמים, בבינה מלאכותית ובקבלת החלטות אוטומטיות; סינון והסרה של תוכן אינטרנט פוגעני; חשיבותם של ארגונים לבדיקת עובדות עצמאית וללא משוא פנים; חובת אימות המקור; מתן אפשרות למשתמשים לדווח ולסמן דיסאינפורמציה פוטנציאלית; הצגה ותיוג של דיסאינפורמציה המתגלה ככזו כדי לעורר דיון ציבורי ולמנוע עליית התוכן מחדש. אזכור ספציפי של דיפ־פייק אפשר למצוא בעמדות פרלמנטריות שונות, ביניהן החלטת הפרלמנט מפברואר 2019 הקוראת לנציבות לדרוש הצגת תיוג ליוצרי דיפ־פייק (van Huijstee et al., 2021).

המסמך המקיף והעדכני ביותר שיצא באירופה בנושא דיפ־פייק הוא החלטת הפרלמנט האירופי ממאי 2021 בנושא 'בינה מלאכותית בחינוך ובתרבות'. נוסף על ההצעות שהוזכרו לעיל, החלטה זו מכילה הצעות שונות כיצד להיערך להתמודדות עם דיפ־פייק כ"איום מיידי על הדמוקרטיה". אלה כללו את חשיבות העלאת המודעות הציבורית לסיכונים של דיפ־פייק ושיפור האוריינות הדיגיטלית; טיפול בקושי הגובר לאתר ולתייג תוכן כוזב ומניפולטיבי באמצעים טכנולוגיים; קריאה לנציבות להציג מסגרות משפטיות מתאימות לשליטה ביצירה, ייצור או הפצה של דיפ־פייק למטרות זדוניות; קידום פיתוח יכולות נוספות לאיתור ולזיהוי של דיפ־פייק; שיפור השקיפות ביחס לתוכן המוצג למשתמשי הפלטפורמות, אשר נותן להם חופש רב יותר להחליט אם ואיזה מידע הם רוצים לקבל (European Parliament, 2021).

ההתמודדות בבריטניה

בבריטניה, כמו באיחוד האירופי, החוק לוקה בחסר באשר ליכולת להתמודד עם איום הדיפ־פייק, כאשר כיום אין באיחוד חוקים המיועדים במיוחד לדיפ־פייק, וכן אין "זכות קניין רוחני עמוק" או חוק המגן על תדמית או אישיות של אדם. יתרה מכך, במקרה של דיפ־פייק שקורבנותיו הם ידוענים גם זכויות יוצרים לא יועילו, שכן זכויות היוצרים נתונות לאולפני סרטים ולצלמים, לא לסובייקטים עצמם המופיעים בסרטון. המשמעות היא שסוגיית הדיפ־פייק בבריטניה פרוצה לחלוטין. מומחים קראו לממשלה הבריטית לנקוט צעדים מהירים להסדרת הנושא ולפעול לרגולציה שתשמור על יתרונות הטכנולוגיה אך תמגר השלכות שליליות כמו פורנוגרפיה, הונאה ופגיעה בדמוקרטיה.[3] לעת עתה הטכנולוגיה בבריטניה מקדימה את החוק (The rise of the deepfake, 2021). כך גם במרבית מדינות העולם.

ההתמודדות בסין

סין הייתה המעצמה הראשונה שהפכה את השימוש בדיפ־פייק לפשע פלילי מבלי לסמנו ככזה. על פי תקנות משרד הסייברספייס של סין, אשר נכנסו לתוקף בינואר 2020, כל סרטון שנוצר בעזרת בינה מלאכותית צריך

3 "The time is now to introduce regulation in this area in order to prevent negative uses of the technology and create an environment where positive use cases emerge."

להיות מסומן ככזה בצורה ברורה. לדברי המשרד, טכנולוגיית דיפ־פייק יכולה "להוות סכנה לביטחון הלאומי, להפריע ליציבות החברה, להפריע לסדר הציבורי ולפגוע בזכויות וברצונות של אחרים". הפרת התנאי תחשוף את היוצר שלו ואת האתר המארח של הסרטון לדין פלילי. אף שהחוק הסיני אוסר על יצירה או הפצה של דיפ־פייק ללא סימון, הוא אינו ברור בעניין העונש למי שיפר אותו. בהודעה שפרסמה ממשלת סין לאתרים נאמר רק כי יוצרי הסרטון וכן מארחיו צפויים לעמוד לדין פלילי (Statt, 2019).

פתרונות טכנולוגיים

'זיהוי ידני' של דיפ־פייק מחייב תשומת לב רבה לשינויים כגון מציאת אי־התאמה בקול, בתמונה או בווידאו. זיהוי ידני מבוסס על יכולתה של העין האנושית למצוא חוסר עקביות באמצעות ניתוח דפוסי התנהגות בסרטונים לעומת אלו המוכרים מהמציאות, או לחלופין, מציאת עדויות לשינוי או להתערבות בקובץ הדיגיטלי, ניסיון לאתר את מקורו, את מועד יצירתו ועוד. החיסרון העיקרי של שיטות עבודה ידניות קשור בהיקף החומר שניתן להתמודד עימו. לעומת זאת, כלים אוטומטיים יכולים להתמודד עם כמות גדולה של קבצים ולאתר זיופים בקלות (van Huijstee et al., 2021). מכאן נובעת חשיבותן הרבה של שיטות הזיהוי האוטומטיות או האוטומטיות למחצה, העושות שימוש בכלי בינה מלאכותית, למשל: זיהוי קולי מבוסס בינה מלאכותית (דוגמת השוואה בין "חתימת הקול "של האדם שאותו מבקשים לחקות לבין הקובץ החשוד); ניתוח מבוסס בינה מלאכותית של תווי פנים; או איתור אזורים "מטושטשים" שלעיתים נוצרים כאשר נעשה שינוי בקובץ מדיה מקורי. לשיטות זיהוי אלה יש מגבלות, וייתכנו מקרי אי־זיהוי של קובצי דיפ־פייק גם באמצעים טכנולוגיים מתקדמים.

בין הכלים הטכנולוגיים הקיימים בשוק נמנה מאמת הווידאו של מיקרוסופט, שהושק ב־2020. התוכנה פועלת על בסיס שיטת ניקוד, הקובעת באחוזים את מידת הסיכוי של קובץ דיגיטלי להיחשב אותנטי או מזויף. גישה נוספת לזיהוי קבצים אותנטיים או מזויפים היא הטמעת "חותמת איכות" של יוצרי הקבצים. מיקרוסופט עושה שימוש במערך הענן שלה Azure בחותמת דיגיטלית המוצמדת לכל קובץ, כך שאפשר להבחין במהירות אם מדובר בקובץ מקורי על פי הפרמטרים של החברה.

חברת פייסבוק הודיעה אף היא ביוני 2021 כי פיתחה יישום לזיהוי דיפ־פייק בשיתוף עם אוניברסיטת מישיגן. הכלי עושה שימוש בזיהוי דפוסים של קבצים שנוצרו על ידי בינה מלאכותית. כך יכולה פייסבוק לאתר את מקור הקובץ, לחסום אותו ואף למנוע הפצה עתידית של קבצים נוספים. אולם איכות הזיהוי ככל הנראה עדיין אינה טובה דיה. בשנת 2020 ערכה פייסבוק תחרות שבמסגרתה הוגשו אלגוריתמים שונים לזיהוי דייפ־פייק, והכלי המוצלח ביותר השיג שיעור זיהוי של כ־65 אחוזים. מאז ינואר 2020 מיישמת פייסבוק את מדיניות האכיפה עבור קבצים המכונה manipulated media — קבצים שפייסבוק מזהה כי הם נוצרו על ידי בינה מלאכותית או למידת מכונה, אשר מטרתם להיראות אותנטיים ומטעים. מאז יישום המדיניות פייסבוק הסירה תכנים שאותם הגדירה ככאלה. פייסבוק גם יצרה שיתוף פעולה עם סוכנות הידיעות רויטרס לצורך קורס מקוון חינמי לציבור, במטרה להכשירו לזיהוי קובצי דייפ־פייק.

חוקרי מעבדת המחקר של צבא ארצות הברית פיתחו עם אוניברסיטת דרום קליפורניה כלי חדשני לזיהוי תוצרי דיפ־פייק הנקרא DeFakeHop ומבוסס על טכניקות זיהוי פנים ביומטריות. החידוש הבולט בכלי זה

הוא תאוריה ומסגרת מתמטית חדשנית שפיתחו החוקרים, שקראו לה Successive Subspace Learning או SSL. בעזרת ה־SSL מחלצים באופן אוטומטי תכונות מחלקים שונים של תמונות פנים, מנתחים את הממצאים ומזהים אם הסרטון הוא זיוף. לפי פרופסור קאו מאוניברסיטת דרום קליפורניה, מדובר במסגרת מתמטית חדשה לחלוטין ושונה מהגישה המסורתית בתחום ארכיטקטורת הרשת העצבית. השיטה הוצגה לראשונה במארס 2021 והיא מראה הצלחה של יותר מ־90 אחוזי דיוק בכל מערכי הנתונים, ובחלקם אף של 100 אחוזים (U.S. Army DEVCOM, 2021).

עד כה, רוב מאמצי המחקר הנוגעים לטכנולוגיות לזיהוי ואיתור דיפ־פייק מתמקדים בפתרונות איתור אוטומטי שיסייעו לזהות דיפ־פייק בשנים הקרובות, בהתאם להתפתחות הטכנולוגית הקיימת והצפויה. אך שיטות זיהוי אוטומטי עשויות להפוך בלתי יעילות כבר בעתיד הקרוב, שכן הטכנולוגיות ליצירת הזיוף משתפרות במידה ניכרת. כמו כן, מול ההתפתחות וההשקעה המסיבית של ענקיות האינטרנט בתחום נצפית מעורבות מדינתית וחקיקתית דלה בלבד, המתמצה בעיקרה בהגבלה ובפיקוח על פרסומים תעמולתיים בעת בחירות, ללא השקעה במחקר (Vizoso et al., 2021). לפיכך, נוסף על תמיכה בפיתוח פתרונות לטווח קצר, על גורמי ממשל להשקיע בגילוי ובפיתוח פתרונות לטווח הרחוק. אנגלר (Engler, 2019) ממליץ לגורמי הממשל לתמוך ולממן תוכניות העוסקות במאמצי זיהוי מתמשכים ובהכשרת עיתונאים ובודקי עובדות שישתמשו בכלים אלו; לקיים תחרויות המעודדות חברות לפתח כלים חדשניים, למשל מערכות אימות מבוססות בלוקצ׳יין,[4] שעשויים לפעול בצורה מהימנה יותר מול זיופים עמוקים; לעודד את פרסום מאגרי הנתונים הגדולים של המדיה החברתית לחוקרים לשם לימוד ומחקר אקדמי של פתרונות למניעת התפוצה של פייק ניוז, ובכלל זה דיפ־פייק.

חינוך הציבור

לצד מאמצי רגולציה, חקיקה ואכיפה ופתרונות טכנולוגיים, מדינות, ארגונים וחברות בעולם עוסקים בהסברה ובהעלאת המודעות בקרב הציבור לאתגר הדיפ־פייק. ההסברה לציבור מתמקדת בהבהרות כי יצירה או הפצה של תכנים באמצעות דיפ־פייק (ללא סימון) עלולה להוות עבירה פלילית; בקמפיינים להעלאת מודעות לקלות ולזמינות ייצורם של סרטוני דיפ־פייק אמינים למדי; וכן בחינוך להטלת ספק מתמדת ולצריכת מידע ממקורות מהימנים. אחת המדינות החלוצות והמובילות בתחום זה היא פינלנד, שבה קיימת תוכנית חינוכית רחבה לצריכה מושכלת של תקשורת ולשיטוט חכם ברשתות החברתיות (Barber, 2021). הפרלמנט האירופי המליץ שוב בשנים האחרונות לפעול לחינוך הציבור ולשיפור האוריינות הדיגיטלית (European Parliament, 2021), ומבריסל פרסמו לציבור האירופאי אסטרטגיה להתמודדות עם דיסאינפורמציה, הכוללת הנחיות ספציפיות בעניין דיפ־פייק.

בישראל, במסגרת המאמץ להעלאת מודעות הציבור לנושא, איגוד האינטרנט הישראלי פרסם לאחרונה רשימה של תוכנות ויישומים חינמיים ונגישים לציבור לזיהוי דיפ־פייק. בין היתר הציג איגוד האינטרנט לציבור בספטמבר 2020 את מאמת הווידאו של חברת מיקרוסופט העולמית (Microsoft Video Authenticator),

4 טכנולוגיה המאפשרת פעילות עסקית מאובטחת ללא צורך בישות ניהול מרכזית.

המסוגל לנתח קובצי קול ווידאו על מנת להעריך באחוזים את ההסתברות שתכנים אלו עובדו או שונו בצורה מלאכותית (קאהאן, 2020).

כלי נוסף שעליו המליץ איגוד האינטרנט הישראלי שייך לחברת sensity.ai ההולנדית. הוא מתבסס על שיטת DCT (Discrete Cosine Transform) המאפשרת לגלות תוצרי דיפ־פייק שנוצרו על ידי שימוש ב־GANs. הכלי מתמקד בזיהוי פנים ששונו, אך עדיין אינו מזהה חפצים שנערכו לתוך חומרי דיפ־פייק. חברת אבטחת המידע Zemana השיקה אף היא כלי טכנולוגיה לשימוש הציבור הרחב בשם Deepware, המאפשר למשתמשים להזין ישירות לאתר האונליין של החברה סרטוני וידאו המנותחים ונסרקים בו על מנת לזהות אם עברו עיבוד או מניפולציה חזותית, לרבות באמצעות דיפ־פייק (איגוד האינטרנט הישראלי, ל״ת).

נוסף על חינוך האוכלוסייה הכללית, קיימת חשיבות גבוהה לחינוך בנוגע לדיפ־פייק לאוכלוסיות ייעודיות כגון אנשי מקצוע בתחום אכיפת החוק והמשפט, מקבלי החלטות ובעיקר עיתונאים, העשויים להוות גורם נוסף להפצה (שלא בידיעה) של ידיעות מזויפות. עוד לפני הופעתן של טכנולוגיות דיפ־פייק היה על עיתונאים להתמודד עם האתגר הכרוך בתארוך של מידע המגיע אליהם או אימות אמינותו. הדיפ־פייק הופך את עבודתם למסובכת יותר ולכן נודעת חשיבות להגברת המודעות בקרבם לסוגיה, וכן למתן כלים בידם לזיהוי ולאימות חומרים המגיעים אליהם.

במצב אידיאלי כלים לאימות של דיפ־פייק צריכים להיות זמינים לכל אדם, אולם משום שחלק מהטכנולוגיה הזו נמצא בתהליכי פיתוח מוקדמים, יש חוקרים הדואגים לסייע כבר כיום לעיתונאים מתוך תפיסה שהם ״קו ההגנה הראשון מפני התפשטות מידע מוטעה״ (Sohrawardi & Wright, 2020). זאת ועוד, מעת לעת מומחים מפיצים מדריכים המסייעים גם לאדם הסביר לזהות באופן לא־טכנולוגי תוצרי דיפ־פייק, כזה שפורסם על ידי מעבדת המדיה של MIT ובו צעדים מומלצים כגון: שימת לב למצח וללחיים של המופיעים בסרטון, האם הם חלקים מדי? האם העיניים קרובות מדי? האם יש בוהק מוזר במשקפיים? (Groh, n.d.). עם זאת, מוצלחות ככל שיהיו, מדובר ביוזמות נקודתיות והשפעתן מוגבלת.

סימולציה — איום הדיפ־פייק על הביטחון הלאומי בישראל

לצורך גיבוש המלצות לישראל לשם התמודדות עם איום הדיפ־פייק על הביטחון הלאומי, התקיימה באוקטובר 2021 במכון למחקרי ביטחון לאומי סימולציה בהשתתפות מומחים מתחומי הדוברות והתקשורת, האסטרטגיה וניהול המשברים, הטכנולוגיה והבינה המלאכותית, המשפט והמדיניות (ראו נספח א׳ — המשתתפים בסימולציה).

במסגרת הסימולציה חולקו המשתתפים לשלוש קבוצות, כל אחת מהן שיחקה בתפקיד קבוצת מומחים עצמאית שגויסה אד הוק לייעץ לראש הממשלה בנוגע להתנהלות מומלצת אל מול תרחיש בדיוני. בתרחיש נכלל סרטון דיפ־פייק שהופץ והחל לעורר תגובות ולהשפיע על המצב בישראל ומחוצה לה. לאחר שגיבשו המלצות להתמודדות עם הסיטואציה בזמן אמת, הצוותים התבקשו לגבש המלצות לצעדי היערכות מקדימים שיסייעו להתמודדות משופרת עם סיטואציה דומה, אם תתרחש בעתיד. התרחיש, הדיונים והמסקנות התנהלו במסגרת ׳כלל בית צ׳טהאם׳.[5]

תרחיש הסימולציה, כפי שהוצג למשתתפים

הזמן: 21 באוקטובר 2021.[6] בשבועיים האחרונים הקואליציה מתמודדת עם משבר פוליטי סביב סוגיית ירושלים, עקב התנהגות פרובוקטיבית של חלק מחברי האופוזיציה. ההתרחשות אף הובילה לגינוי מטעם מלך ירדן, להתערבות מצד מדינות אסלאמיות שעימן נמצאת ישראל בקשרים דיפלומטיים ולדרישה אמריקאית ואירופית לרסן את הגורמים השונים. ההתרחשויות עירבו גם מספר התפרצויות של הפגנות אלימות בירושלים וכן בגבול רצועת עזה.

בשעות הבוקר המוקדמות החלה הפצת סרטון ברשתות החברתיות ובאמצעות קבוצות ווטסאפ וטלגרם רבות, שבו נראה מפגש מצומצם של מנהיגי מפלגות הימין הפוליטי בישראל, וביניהם ראש הממשלה מר נפתלי בנט. בסרטון קורא אחד מחברי הכנסת המוכרים ממפלגות הימין הדתי: ״עלינו לסיים את הסוגיה הפלסטינית אחת ולתמיד! על אלימות נגיב באלימות! לא ניתן שיכתיבו לנו כיצד להתנהג בבירתנו הנצחית, עלינו לעלות חמושים על אל־אקצה ולהראות להם מי הריבון״. בסרטון נראים כל הנוכחים במקום, בכללם בנט, מגיבים לדברים במחיאות כפיים סוערות ואת בנט מוסיף: ״יש גבול לפשרנות״.

בערוצי התקשורת המסורתיים וכן באתרי החדשות באינטרנט מתלבטים כיצד להציג את העניין ואף פנו לנראים בסרטון כדי לקבל את תגובתם. במקביל דיווחו גורמים בשב״כ על התארגנות של גורמי ימין קיצונים לעלות ולפרוץ בכוח להר הבית. שעות ספורות לאחר פרסום הסרטון איימו חמאס וחזבאללה לשגר בערב מטחים למרכז ישראל אם הממשלה לא תתנצל, תקל את תנאי התפילה של המוסלמים בהר הבית ותמנע גישה של יהודים לאתר.

5 The Chatham House Rule — כאשר פגישה או חלק ממנה מתקיימת על פי כלל בית צ׳טהאם, המשתתפים רשאים להשתמש במידע שהתקבל בפגישה, אך נאסר עליהם לחשוף את זהות הדובר או הדוברים ואת השתייכותם המקצועית.

6 שבועיים קדימה מיום קיום הסימולציה, במטרה לאפשר למשתתפים לדמיין עתיד קרוב שבו חלות הנסיבות המוכרות להם.

עקב הפצת הסרטון ניכר כי אלפי אזרחים ישראלים מתכננים לצאת לרחובות למהומות אלימות ברוב ערי ישראל, ובכלל זאת לעשות שימוש בנשק חם. מספר מנהיגים מהעולם המערבי וכן מנהיגי מדינות ערביות ואסלאמיות ידידותיות לישראל פנו לממשלה בדרישה להבהרה והתנצלות.

הדיון בקבוצות

במסגרת הדיון בקבוצות נשאלו המשתתפים: כיצד תמליצו לראש הממשלה לפעול על מנת למנוע או לצמצם את ההסלמה? המשתתפים התבקשו לחשוב על צעדים הניתנים לביצוע בזמן אמת, בתגובה לסרטון שכבר הופץ ברשתות החברתיות, אך טרם עלה בערוצי התקשורת המסורתיים ובאתרי החדשות באינטרנט. הדיון שהתקיים בשלוש קבוצות שונות הפיק בחלקו מסקנות דומות ובחלקו מסקנות שונות. להלן ההמלצות שגובשו בקבוצות:

המלצות להתמודדות בזמן אמת

- בקבוצות השונות רווחה המלצה לראש הממשלה לנקוט **אסטרטגיה של הזמה מהירה מבוססת ראיות** — הצהרה תקשורתית מהירה, פשוטה וכנה של המופיעים לכאורה בסרטון, מגובה בראיות טכנולוגיות ובתפוצה נרחבת. ההמלצה נבעה מתוך הידיעה של ראש הממשלה כי הסרטון מזויף, וכי לא יתאפשר להציג ראיות סותרות כאילו הוא או מי מהאחרים באמת היה במקום או אכן אמר את שנראה בסרטון.
- חלק מהמשתתפים סברו כי ל**הזמת הסרטון תהיה השפעה מצומצמת בלבד** וכי "מי שירצה לעלות להר הבית — יעלה"; "ההשפעה כבר נעשתה"; "זה לא ישנה כלום"; "גם אם תצא הכחשה נחרצת, לא בטוח שיקבלו אותה". סברה זו מדגישה את חשיבות ההיערכות למניעה מקדימה. בה בעת, חלק מהמשתתפים סברו כי הכחשה והזמה של הסרטון ישפיעו על המתלבטים, ויש בה טעם גם אם השפעתה תהיה חלקית.
- **חלק קטן מן המשתתפים** הציעו לנקוט דרך קונוונציונלית פחות ולהמליץ לראש הממשלה לרתום את היכולות הטכנולוגיות של גורמים בישראל כדי **ליצור במהירות סרטון דיפ־פייק** אמין שבו ייראו מנהיגים של ארגוני טרור המאיימים על ישראל אומרים אמירות שלא יעלו על הדעת. סרטון זה יסומן כמזויף ומטרתו תהיה להגחיך גם את הסרטון שעורר את ההסלמה ולהדגים עד כמה קל להראות מישהו אומר דברים מופרכים לחלוטין (במקרה זה, לא כדי ליצור מצג שווא נגדי ומטעה). אולם חלק ניכר מהמשתתפים התנגדו להצעה זו בטענה שאין לתת לגיטימציה לשימוש בפייק, שאינו אמור להיות חלק מהאמצעים הלגיטימיים של הממשלה. המשתתפים נשארו חלוקים בדעותיהם בעניין זה.
- רעיון נוסף שעלה הוא **נקיטת מהלך תקשורתי שיתמקד בדיון באחריותן של הרשתות החברתיות למצב המתוח**. אולם מרבית המשתתפים סברו שלא תצמח ממהלך זה התועלת המבוקשת במובן של צמצום ההסלמה.

הקשרים נושאיים ממוקדים

- **הזווית הטכנולוגית** — הכלים הטכנולוגיים לזיהוי ולהוכחת פייק קיימים, אך אינם מספקים מענה מקיף לאתגר. אם מופץ סרטון דיפ־פייק ברמה גבוהה, יש להניח כי גורם בעל עוצמה (ככל הנראה מדינתי)

עומד מאחוריו, ולכן יש להיערך להתמודדות לא רק עם הסרטון אלא גם עם פוטנציאל למתקפה של סרטונים כאלה.

- **הזווית המשפטית** — מבחינה משפטית אין כיום מענה ממוקד, ולא ניתן להוציא מיידית צו עיכוב נגד הפצה. גם החקיקה ברחבי העולם בנושאי דיפ־פייק בפרט ופייק ניוז בכלל היא לוקאלית, נישתית ולא משמעותית. המשתתפים בסימולציה סברו כי יש עילה לפנות לרשתות החברתיות בבקשה להוריד את הסרטונים, אך לא חלה עליהן חובה לעשות זאת. המשתתפים בסימולציה הזכירו שוב ושוב את הצורך לפעול להסדרה משפטית מבעוד מועד. עם זאת רווחה בקרבם הנחה כי פנייה לבית המשפט במטרה להוציא צו איסור פרסום תתקבל, משום שאף כי מדובר בפגיעה בחופש הביטוי ובזכות הציבור לדעת, האיום הביטחוני הוא משמעותי ולכן תגבר הזכות לביטחון. עם זאת נותרה ספקנות לגבי יעילותו של המהלך עקב חוסר ההשפעה המשוער שלו על התפוצה בקבוצות שונות באפליקציות המסרים המיידיים.
- **הזווית של התקשורת** — המשתתפים הניחו כי כל עורך תקשורת ירצה לשדר את הסרטון וישדרו, העניין הוא איך יציג אותו. הממשלה חייבת לפנות לתקשורת הממוסדת ולבקש ממנה לשדר את הסרטון עם 'סימן מים' בוהק ומודגש שיבהיר כי מדובר בפייק. יש מי שסבר כי גם אם צעד זה לא ישיג תוצאה מיידית, תהיה לו חשיבות חינוכית מתמשכת.
- **הזווית של הצנזורה** — רבים טענו ש"הצנזורה לא רלוונטית במקרה הזה", משום שחופש הביטוי כנראה יעמוד לזכות המפיצים ובכל מקרה מדובר בפתרון שאינו מניח את הדעת, כי אנשים יוכלו להפיץ את המידע דרך אפליקציות המסרים המיידיים כמו ווטסאפ או טלגרם.
- **הזווית הביומטרית** — לשיטתם של חלק מן המשתתפים, דיפ־פייק מעלה סכנה מטרידה בכל הקשור לזיוף ביומטרי, שכן תכונות ביומטריות (מראה פנים, קול וחיתוך דיבור) נלקחות מאדם מסוים ונעשה בהן שימוש שלא ברשותו על ידי אדם אחר. אלא שהרשויות עדיין לא ערוכות להתמודד איתו. זאת בשונה מזיוף תיעוד ביומטרי אחר — למשל תעודות זהות ודרכונים. עם זאת הודגש כי האפשרות להתחזות הייתה קיימת קודם לכן, ללא תלות במחשוב מתקדם, למשל באמצעות מסכות סיליקון.
- **מקור הסרטון** — חלק מהקבוצות קיבלו הזדמנות להתמודד עם השאלה "האם למקור הסרטון יש השפעה על דרך ההתמודדות?" וכל קבוצה התמודדה עם מקור אחר — מדינתי או לא־מדינתי. יודגש שבין הקבוצות לא ניכרו הבדלים בדרך ההתמודדות בתוך ישראל, על פי מקור ההפצה של הסרטון, אולם ניכרו הבדלים בהמלצות לראש הממשלה לגבי התמודדות בזירה הבינלאומית. היכולת להוכיח את מקור הסרטון עשויה לסייע ברתימת גורמים בינלאומיים לגנות את מפיציו ולא לחבור למגנים את ישראל.

המלצות להיערכות מקדימה

לאחר הדיון בקבוצות נערך דיון במליאה, שבו נדונו סוגיות הקשורות להיערכות מקדימה למניעת מקרים כמו זה שהוצג בתרחיש, וכן לשם יצירת אמצעים להתמודדות עם אירוע דומה, אם יתרחש. גם בדיון זה הודגשה החשיבות של האמצעים הטכנולוגיים הקיימים לזיהוי מוקדם של הפצת פייק ניוז, ובמסגרתו גם דיפ־פייק ברשתות השונות.

נוכח עוצמתן של הרשתות החברתיות וכן של האפליקציות להעברת מסרונים מיידיים וטכנולוגיות נוספות (דוגמת פרופילים מזויפים ובוטים) המאיצים את התפוצה, ניכר כי על ישראל (כמו גם מדינות אחרות) להיערך במספר היבטים:

- **חקיקה ורגולציה** — יש לעסוק במשמעויות המשפטיות ולאפשר לחוק להתמודד עם סוגיית הדיפ־פייק הן טרם ההתרחשות והן בזמן אמת. כלים אלה יתייחסו למורכבויות בנוגע לחופש הביטוי ויכילו אותן.
- **רשויות החוק והמשפט,** כמו גם רשויות שאחראיות על תחומי המחשוב והדיגיטל במדינת ישראל, נדרשות לחשיבה ולהתארגנות לטיפול בנושא מתוך הכרה שמדובר באירוע ביומטרי.
- **חיזוק הקשר והעבודה המשותפת של הממשל עם התקשורת הממוסדת והרשתות החברתיות** — מומלץ לקבוע נוהל עבודה מסודר מול התקשורת הממוסדת וליצור סטנדרטים ידועים ומוסכמים לטיפול באירוע דיפ־פייק, כפי שהוצג בתרחיש. עבודה משותפת עשויה לגרום לתקשורת לקבל על עצמה הגבלות וולונטריות גם במקרה שהחקיקה עדיין חסרה.
- **חינוך הציבור** — יש לסייע בהפצת הידע והמודעות של הציבור לקיומו של דיפ־פייק ולפשטות והזמינות של יצירתו, על מנת לעורר ביקורתיות כלפי תוכן שהציבור נחשף אליו בערוצים הלא־ממוסדים. כן יש צורך להגביר את המודעות בקרב אנשי מקצוע — עיתונאים, מנהלי קהילות ברשתות החברתיות — ולחשוף אותם לכלים טכנולוגיים המסייעים בזיהוי דיפ־פייק. אפשר להיעזר בקמפיינים ממומנים על ידי הממשלה כפי שנעשה בהקשר למגפת הקורונה למשל, בתוכניות טלוויזיה העוסקות בטכנולוגיה ובגורמים אמינים נוספים, שיציגו את הענין לציבור.

מסקנות ניתוחיות מן הדיון

בין היתר, במהלך הסימולציה ניכרה מחשבה בקרב המשתתפים כי ייחודיותה של תופעת הדיפ־פייק מאותגרת למעשה. משתתפים רבים לא התייחסו לסיטואציית התרחיש כאל זו היוצרת איום חדשני שטרם נראה, או כאל איום שונה במהותו מאיומים אחרים הנובעים מתפוצת המידע המסיבית והמהירה ברשתות החברתיות ובאפליקציות להעברת מסרים מיידיים.

בזמן שמשתתפים אחדים טענו כי "הפצת שמועות שקריות הייתה קיימת עוד ברומא העתיקה", חלק ניכר מהם טענו כי מדובר בתופעה חדשנית בעלת פוטנציאל השפעה בסדרי גודל שהמוח האנושי מתקשה לתפוס, ולכן יש להיערך להתמודדות בהתאם. בהתייחס לסימולציה עצמה, רק מעטים מבין המשתתפים התייחסו לעובדה שמדובר בסרטון — שבו נראים ונשמעים לכאורה נבחרי הציבור ובתוכם גם ראש הממשלה — כאל סיטואציה שעשויה להיות לה השפעה רבה יותר מזו של זיוף דומה בכתב או בתמונה. ראוי לציין כי המשתתפים לא נשאלו שאלה ישירה בעניין זה.

אפשר למצוא הסברים שונים לממצא זה. השערה אחת היא ש'פייק הוא פייק', ובעידן הפוסט־אמת יש לדיפ־פייק מתוחכם או לצ'יפ־פייק (זיוף שנוצר באמצעים כגון עריכה פשוטה ולא באמצעות זיוף מבוסס בינה מלאכותית) השפעה דומה. השערה שנייה גורסת כי קיים עדיין קושי לתפוס את השפעות הדיפ־פייק הייחודיות. ההשערה השלישית היא כי למרות שקיים הבדל בין השפעת הדיפ־פייק לצ'יפ־פייק, הם אינם נבדלים מהותית במענה שהם דורשים מאיתנו לגבש (אותו מענה שהמשתתפים בסימולציה התבקשו לספק).

כמו כן עלתה מבוכה מסוימת אל מול ההתמודדות עם פייק שתפוצתו אינה ניתנת לעצירה כמעט, על כל צורותיו השונות. ניכר כי הרשויות והגורמים השונים אינם ערוכים כיום להתמודדות גם עם צ׳יפ־פייק נוכח הקושי של הרשויות מול הרשתות החברתיות, האפליקציות להעברת מסרים מיידיים ועוצמתן של קבוצות וקהילות דיגיטליות, כמו גם היכולת להשפיע עליהן באמצעות פרופילים מזויפים, בוטים ואמצעים טכנולוגיים נוספים.

על אף ההסתייגות המסוימת העולה כאן לגבי הייחודיות של הדיפ־פייק, הנושא ראוי לבדיקה ולהתייחסות מיוחדת. מתקפת דיפ־פייק משמעותית בניסיון לפגוע בביטחון לאומי של ישראל אומנם טרם התרחשה, אך יוזכר כי ספקנות התעוררה בעבר גם לגבי איומים טכנולוגיים אחרים, ביניהם איום הרחפנים שזוכה כיום לתשומת לב רבה, המובילה לפיתוח מערכות נגד ודוקטרינות להתמודדות — אם כי בפיגור של שבע עד עשר שנים ביחס להופעתו הראשונית. לסיכום, הסימולציה חשפה פער מטריד בין איום חדש שמשמעויותיו אולי אינן מובנות במלואן לבין מידת המוכנות וההיערכות אליו בישראל, הן מבחינה טכנולוגית, הן מבחינת חינוך הציבור והן מבחינת רגולציה, חקיקה ואכיפה.

סיכום ומסקנות

עם ההתפתחות הטכנולוגית בתחום הבינה המלאכותית, כמו גם הפיכתה לזולה וזמינה יחסית לקהל הרחב, גם היכולות לייצר דיפ־פייק קשה לזיהוי נעשות נגישות יותר. להתפתחות זו מגוון שימושים פוטנציאליים. חלקם חיוביים, למשל בתחומי התרבות, האומנות, השחזור ההיסטורי ותחומים עסקיים שונים; חלקם בעלי פוטנציאל שלילי לשימוש בזדון, למשל בתחומי הפורנו, ההונאה וההשפעה הפוליטית המקומית או הגלובלית, ואף גלום בהם איום על הביטחון לאומי.

מחקר זה, שבוצע במסגרת תוכנית ליפקין־שחק לפוסט־אמת ופייק ניוז, סקר את הטכנולוגיה המשמשת ליצירת זיופים אלו, את השימושים השונים שלה וכן מקרי מבחן של שימוש שלילי מרחבי העולם, כמו גם דרכי ההתמודדות איתו המקובלות כיום, במטרה לנסות לברר אם מדובר באיום ממשי על הביטחון הלאומי וכיצד ראוי להתמודד איתו. במסגרת המחקר בוצעה סימולציה בהשתתפות מומחים; הם התבקשו להתמודד עם תרחיש של הפצת סרטון דיפ־פייק בישראל, שנבע בהשראתו וממנו איום באלימות מבית ומחוץ, וכן לגבש המלצות להתמודדות עם האתגר.

המחקר ותוצאות הסימולציה מעידים כי ארגז הכלים הקיים כיום בידי מדינות וביניהן ישראל, לשם התמודדות עם איום הדיפ־פייק, לוקה בחסר. זאת בין היתר משום שהאיום עצמו עדיין לא ברור מספיק, וכי לפחות בתחום הביטחון הלאומי טרם הודגמו השפעות מרחיקות לכת שלו. בין היתר, עדיין לא ברורה למקבלי ההחלטות ולציבור ההבחנה בין דיפ־פייק לצ׳יפ־פייק (תוכן מזויף המושג באמצעים שאינם בהכרח בינה מלאכותית). יודגש כי גם ההתמודדות עם צ׳יפ־פייק כיום לוקה בחסר, בעיקר נוכח חוסר שליטה באמצעי ההפצה השונים ובתוכם הרשתות החברתיות והיישומים להעברת הודעות מיידיות דוגמת ווטסאפ וטלגרם.

הן סקירתם של צעדי ההתמודדות המקובלים כיום בעולם והן תוצאות הסימולציה קשורים במספר תחומים עיקריים, שבמסגרתם על ישראל לפעול בהקדם האפשרי כדי לנסות לצמצם את השפעת האיום הגלום בדיפ־פייק על הביטחון הלאומי ה״קלאסי״ וגם על הביטחון הלאומי במובן הרחב, הכולל ביטחון אישי ותשתית המדינה הכלכלית.

חקיקה ורגולציה נדרשות הן למניעת הפצתו של דיפ־פייק באופן שיוביל להטעיה והן כדי לתת כלים בידי הרשויות השונות להתמודדות ולתגובה לאחר מעשה, אם יופצו תכנים כאלה. הרשויות הרלוונטיות נדרשות להיערך לפעולה מקדימה ולתגובה תוך הכרה שמדובר בעניין ביומטרי. יש גם צורך בהשקעה בהסברה ובחינוך הציבור בישראל לגבי האתגר הגלום בתכנים מזויפים, ובכלל זאת בדיפ־פייק. נדרש גם דגש מיוחד על הסברה והנגשת כלים לזיהוי פייק בקרב עיתונאים, מנהלי קהילות ברשתות החברתיות וכן בקרב אנשי מקצוע, למשל ברשויות אכיפת החוק. לכל אלו רצוי להוסיף את **חיזוק הקשר והעבודה משותפת של המוסדות המדינתיים עם התקשורת הממוסדת והרשתות החברתיות**, הן לצורך מניעת הפצתו של תוכן מזויף (מבלי שמוסבר שהוא כזה) והן לשם היערכות מקדימה.

ניכר כי אף שדיפ־פייק עדיין לא גרם לאסון כלשהו בתחום הביטחון הלאומי, היערכות ובמסגרתה מחקר ויצירת כלים מתאימים להתמודדות מונעת ולאחר מעשה היא הכרחית עבור כל מדינה. האתגר תקף שבעתיים

עבור דמוקרטיה ליברלית כמו ישראל, שבה מתקיימת תקשורת חופשית מחד גיסא, אך מאידך גיסא היא נדרשת להתמודדות יום־יומית עם איומים ביטחוניים, שהדיפ־פייק עשוי להפוך לאחד המרכזיים שבהם.

נספח א' — רשימת המשתתפים בסימולציה (לפי סדר הא-ב)

- מר מולי אדן, סגן נשיא בכיר באינטל לשעבר ומומחה לחדשנות
- גב׳ ענבל אורפז, חוקרת בתוכנית ליפקין־שחק, המכון למחקרי ביטחון לאומי (INSS)
- גב׳ טוהר איזיקסון, מתמחה במכון למחקרי ביטחון לאומי (INSS)
- מר ארמן איסמעיליזדה, ראש תחום מודיעין ודיסאינפורמציה בחברת ActiveFence
- מר אורי אליאבייב, יועץ בתחום הבינה המלאכותית ומייסד קהילת Deep & Machine Learning Israel
- תת־אלוף (מיל׳) אריאלה בן אברהם, הצנזורית הראשית לשעבר
- עו״ד אורי בארי, עוזר מחקר במכון למחקרי ביטחון לאומי (INSS)
- תת־אלוף (מיל׳) איתי ברון, סגן ראש המכון למחקר, המכון למחקרי ביטחון לאומי (INSS)
- ד״ר אביב גאון, חבר סגל בית ספר הארי רדזינר למשפטים, אוניברסיטת רייכמן
- מר דרור גלוברמן, עיתונאי ופרשן, שידורי קשת
- גב׳ ענבר דולינקו, חוקרת מדיניות וטכנולוגיה
- ד״ר אסף וינר, ראש תחום רגולציה ומדיניות, איגוד האינטרנט הישראלי
- מר ירדן ותיקאי, יועץ אסטרטגי, לשעבר מנהל מטה ההסברה הלאומי במשרד ראש הממשלה
- עו״ד ורד זליכה — שותפה וראשת תחום סייבר ובינה מלאכותית במשרד ליפא מאיר ושות׳; עמיתת מחקר במרכז פדרמן לחקר משפט ומדיניות סייבר, האוניברסיטה העברית; לשעבר ראשת תחום מדיניות בינלאומית ויוזמות בינלאומיות במערך הסייבר הלאומי
- סגן־אלוף (מיל׳) בעז זלמנוביץ, חוקר במחלקה להיסטוריה, צה״ל
- מר יואב זקס, רח״ט מו״פ, המטה לביטחון לאומי (מל״ל)
- גב׳ מלי מרטון, ראש תחום חדשנות, אלתא
- מר אלישע סטואין, ראש תחום סריקת האופק, המשרד לענייני מודיעין
- סגן־אלוף (מיל׳) דוד סימן טוב, חוקר בכיר, המכון למחקרי ביטחון לאומי (INSS)
- ד״ר לירן ענתבי, מנהלת תוכנית טכנולוגיות מתקדמות וביטחון לאומי, המכון למחקרי ביטחון לאומי (INSS)
- מר טל פיאלקוב, חוקר בינה מלאכותית
- מר רועי פרידמן, ראש היחידה להזדהות וליישומים ביומטריים, מערך הסייבר הלאומי
- מר רונן צור, יועץ תקשורת ומומחה לניהול משברים
- גב׳ יובל קנפו, מתמחה, המכון למחקרי ביטחון לאומי (INSS)
- עו״ד מוטי קריסטל, מומחה למשא ומתן ולניהול משברים
- גב׳ רוני קרשטדט, עוזרת מחקר, המכון למחקרי ביטחון לאומי (INSS)
- גב׳ נועם רחמים, עוזרת מחקר, המכון למחקרי ביטחון לאומי (INSS)
- מר דור שושן, מתמחה, המכון למחקרי ביטחון לאומי (INSS)
- מר עפר שלח, חוקר בכיר, המכון למחקרי ביטחון לאומי (INSS)
- אלוף־משנה (מיל׳) עו״ד פנינה שרביט ברוך, חוקרת בכירה, המכון למחקרי ביטחון לאומי (INSS)

אורפז, ע׳ (2020). **Deepfake: מקרה בוחן להשפעת פייק ניוז על הביטחון הלאומי**. המכון למחקרי ביטחון לאומי. https://bit.ly/35rTQB7

איגוד האינטרנט הישראלי. (ל״ת). **מידע כזב (Fake news, Deepfake)**. https://bit.ly/3F1DzTg

ברון, א׳ ושוורץ אלטשולר, ת׳ (2019, 14 ביולי). **פייק ניוז: הדור הבא**. המכון הישראלי לדמוקרטיה. https://bit.ly/3DZ2tld

ברון, א׳ ורויטמן, מ׳ (2019). **ביטחון לאומי בעידן של פוסט־אמת ופייק ניוז**. המכון למחקרי ביטחון לאומי. https://bit.ly/3dZmOfy

הכהן, ג׳ (2014). **מה לאומי בביטחון הלאומי?** אוניברסיטה משודרת, הוצאת משרד הביטחון.

הלפרן, נ׳ (2021, 15 באוקטובר). זייפו את קולו של חבר דירקטוריון — וגנבו 35 מיליון דולר מהבנק. **TheMarker**. https://bit.ly/3G14JLf

חיימוביץ׳, ה׳ (2020, 21 בפברואר). **פוליטיקאי השתמש ב־Deep Fake בצורה גאונית וזה עשוי להביא לו בוחרים חדשים**. Geektime. https://bit.ly/30v9bSv

ענתבי, ל׳ (2020). **בינה מלאכותית וביטחון לאומי בישראל**. המכון למחקרי ביטחון לאומי. https://bit.ly/3m48E19

קאהאן, ר׳ (2020, 2 בספטמבר). מיקרוסופט מציגה: כלי חדש לזיהוי תמונות וסרטוני Deepfake. **כלכליסט**. https://bit.ly/3wnIKsw

רוה, י׳ (2020, 15 בספטמבר). האם סרט תיעודי שעושה שימוש בדיפ־פייק הוא עדיין סרט תיעודי? **כלכליסט**. https://bit.ly/3CbZi9C

רשות החדשנות. (2019). **מרוץ העוצמה הטכנולוגית ממשיך**. https://bit.ly/31THO5l

תב״כ 24/9. ועדת הבחירות המרכזית לכנסת ה־24. https://bit.ly/3s5hFuU

Ynet (2021, 16 במארס). **אמא יצרה דיפ־פייק כדי לסלק בנות מנבחרת המעודדות**. https://bit.ly/2Zb9cKA

2020 Edelman Trust Barometer. (2020, January 19). Edelman. https://bit.ly/2Z3OMmj

Adee, S. (2020, April 29). *What are deepfakes and how are they created?* IEEE Spectrum: Technology, Engineering, and Science News. https://bit.ly/3iBfLgN

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of deepfakes: landscape, threats, and impact*. Deeptrace. https://bit.ly/3aSYp9Z

Ayyub, R. (2018, November 21). *I was the victim of a deepfake porn plot intended to silence me*. Huffpost. https://bit.ly/3xcijWa

Barber, H. (2021, February 16). Finland´s secret weapon in the fight against fake news: Its kindergarten children. *The Telegraph*. https://bit.ly/3E65mRb

Berman, M. (2019, October 3). California Clamps Down on Nonconsensual Deepfake Pornography. *Official Website MARC BERMAN Assemblymember, District 24*. https://bit.ly/3HeyZ59

Biggs, T. & Moran, R. (2021, June 2). What is a deep fake? *The Sydney Morning Herald*. https://bit.ly/3zWADoU

Birchall, G. (2018, October 25). *Married Brazilian politician forced to deny he´s man in video of hotel room orgy with FIVE women released just days before election*. The Irish Sun. https://bit.ly/3rGo6Cz

Brady, M. (2020, September 1). *Deepfakes: A new disinformation threat?* Democracy Reporting International. https://bit.ly/3yaUzUs

Chesney, R., & Citron, D. K. (2019). 21st century-style truth decay: Deep fakes and the challenge for privacy, free expression, and national security. *Maryland Law Review*, *78*(4), 882–891. https://bit.ly/2StnKlq

Citron, D. (2020, October 30). *What happens in a world where fake becomes real?* Ted Radio Hour. https://n.pr/3f5DZxk

Citron, D. K., & Chesney, R. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753-1820. https://bit.ly/320IntR

David, D. (2021, May 10). Analyzing the rise of deepfake voice technology. *Forbes*. https://bit.ly/3pZJeDb

De Agostini, D. (2020, May 3). *Why deepfake will make you play video games instead of movies*. Predict. https://bit.ly/3ddsFOC

Delcker, J. (2019, December 17). *Welcome to the age of uncertainty*. Politico. https://politi.co/2TLh17a

Delfino, R.A. (2019). Pornographic deepfakes: The case for federal criminalization of revenge porn´s next tragic act. *Fordham Law Review*, *88* (3). https://bit.ly/3wdji8n

Dunn, S. (2021, March 3). *Women, not politicians, are targeted most often by deepfake videos*. Center of International Governance Innovation. https://bit.ly/3hxdPnt

Engler, A. (2019, November 14). *Fighting deepfakes when detection fails.* Brookings. https://brook.gs/3yybJM5
European Parliament. (2021). *Artificial intelligence in education, culture and audiovisual sector*. https://bit.ly/3p2lb7w
Farago, t. (2019). *Deep fakes — an emerging risk to individuals and societies alike.* Tilburg Papers in Culture Studies, Paper 237. Tilburg University. https://bit.ly/3lNUgKD
Feeney, M. (2021). *Deepfake laws risk creating more problems than they solve.* Regulatory Transparency Project of the Federalist Society. https://bit.ly/3aX867n
Ferraro, M. F., Chipman, J. C. & Preston, S. W. (2019, December 23). *First federal legislation on deepfakes signed into law.* WilmerHale. https://bit.ly/3E6R6YM
Focaloid.com. (2019, October 17). *Microsoft breaks new ground with AI neural TTS and life-size holograms.* https://bit.ly/3GPOq39
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative adversarial nets.* arXiv preprint arXiv:1406.2661. https://bit.ly/3zob7bV
Graham, N. (2021, May 21). *Deepfake deception: The emerging threat of deepfake attacks.* Dentons. https://bit.ly/3l2zGXe
Groh, M. (n.d.). *Detect deepFakes: How to counteract misinformation created by AI.* MIT Media Lab. https://bit.ly/3aWyDC0
Hao, K., & Heaven, W.D. (2020, December 24). *The year deepfakes went mainstream.* MIT Technology Review. https://bit.ly/35MyZZl
Hao, K. (2021, February 12). *Deepfake porn is ruining women´s lives. Now the law may finally ban it.* MIT Technology Review. https://bit.ly/3h8FrAm
Itzkoff, D. (2020, October 29). The ´South Park´ guys break down their viral deepfake video. *The New York Times.* https://nyti.ms/3dRpNqC
Jackson, F. (2021, January 26). *Helen Mort: Behind every image is a person.* Now Then. https://bit.ly/3f5Ejw2
Jacoby, E. (2019, December 9). *I paid $30 to create a deepfake porn of myself.* Vice. https://bit.ly/3f3foJu
Jaiman, A. (2020, August 14) *Positive use cases of synthetic media (aka deepfakes).* Towards Data Science. https://bit.ly/3d7JrPe
Keitzmann, J., Lee, L.W., McCarthy, I.P. & Keitzmann, T.C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135-146. https://bit.ly/3weDjuW
Khalil, H. A., & Maged, S. A. (2021). *Deepfakes Creation and Detection Using Deep Learning.* 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), pp. 1-4, IEEExplore https://ieeexplore.ieee.org/abstract/document/9447642
Konkel, F. (2019, January 29). *AI, deepfakes and the other tech threats that vex Intel leaders.* Nextgov. https://bit.ly/3gd7gH7
Libby, K. (2020, August 13). *Deepfakes are amazing. They´re also terrifying for our future.* Popular Mechanics. https://bit.ly/3vNClWs
Lu, D. (2020, January 24). *Deepfake software translates videos from one language to another.* NewScientist. https://bit.ly/3IRtsms
Maras, M.-H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof*, *23*(3), 255–262. https://bit.ly/3Cev6Lr
MyHeritage Deep Nostalgia, deep learning technology to animate the faces in still family photos. (n.d.). MyHeritage. https://bit.ly/3EegQ5x
Nejal, R. (2020). *Deepfake Therapy* [video]. IMDb. https://www.imdb.com/title/tt12672938/
Ng, A. (2019, January 29). *Deepfakes, disinformation among global threats cited at Senate hearing.* Cnet. https://cnet.co/3pMQh1v
Noone, G. (2021, February 4). *Listen carefully: The growing threat of audio deepfake scams.* TechMonitor. https://bit.ly/3BSMp4Q
Panyatham, P. (2020, March 10). *Deepfake technology in the entertainment industry: Potential limitations and protections.* Arts Management & Technology Laboratory. https://bit.ly/3j74IvZ
Paris, B., & Donovan, J. (2019). *Deepfakes and cheap fakes.* Data & Society. https://bit.ly/3p5qYrJ

Parkin, S. (2019, June 22). The rise of the deepfake and the threat to democracy. *The Guardian*. https://bit.ly/3219o01

Patrini, G. (2019, October 7). *Mapping the deepfake landscape*. Sensity. https://bit.ly/3x6PZWl

Pechenik Gieseke, A. (2020). "The new weapon of choice": Law's current inability to properly address deepfake pornography. *Vanderbilt Law Review*, *73*(5), 1479–1515. https://bit.ly/3IOdCcp

Roettgers, J. (2020, August 21). *Hulu deepfaked its new ad. It won't be the last*. Protocol. https://bit.ly/3IV2lqM

Salazar, J. (2019, October 11). *"Deepfake" videos under spotlight of new Texas law*. SpectrumNew1. https://bit.ly/32ail7r

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In D. Lee, U. von Luxburg, R. Garnett, M. Suguyama, & I. Guyon (Eds.), *Advances in neural information processing systems 29* (pp. 2234-2242). https://bit.ly/3m8dLOe

Sample, I. (2020, January 13). What are deepfakes — and how can you spot them? *The Guardian*. https://bit.ly/3cC1YTc

Sayler, K. M., & Harris, L. A. (2021). *Deep fakes and national security*. Congressional Research Service. https://bit.ly/3cydaAo

Schiff presses Facebook, Google and Twitter for policies on deepfakes ahead of 2020 election. (2019, July 15). Adam Schiff, Press Release. https://bit.ly/3cBYBfn

Sejnowski, T. J. (2018). *The deep learning revolution*. MIT press.

Sensity Team (2021, August 2). *How to detect a deepfake online*. Sensity. https://bit.ly/2Ttkffh

Sheller, A. (2019, July 2). California Considers Deepfake Ban for Election Integrity. *GovTech*. https://bit.ly/3sfniqm

Sohrawardi, J., & Wright, M. (2020, October 9). *In a battle of AI versus AI, researchers are preparing for the coming wave of deepfake propaganda*. The Conversation. https://bit.ly/3aVPUv0

Somers, M. (2020, July 21). *Deepfakes, explained*. MIT Management Sloan School. https://bit.ly/3gUoFVe

Statt, N. (2019, November 29). *China makes it a criminal offense to publish deepfakes or fake news without disclosure*. The Verge. https://bit.ly/3jmwp3r

Stroud, S.R., & Hayden, J. (2018). *Case study: "Deepfakes" and the ethics of faked video content*. Media Ethics Initiative, University of Texas at Austin. https://bit.ly/3AlsNFw

Stupp, C. (2019, August 30). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *The Wall Street Journal*. https://on.wsj.com/3rMpm71

The World (2019, June 13, 2019). *Internet 'deepfakes' threaten truth and reality*. https://bit.ly/2V23fgO

The rise of the "deepfake" demands urgent legal reform in the UK. (2021, March 23). National Law Review. https://bit.ly/3GPf9wV

ThinkAutomation (n.d.). *Yes, positive deepfake example exist*. https://bit.ly/3vOgiPj

Thomson, P. (2020, August 26). *How hidden cameras captured a daring rescue in 'Welcome to Chechnya'*. International Documentary Association. https://bit.ly/3F32w0J

Toews, R. (2020, May 25). Deepfakes are going to wreak havoc on society. We are not prepared. *Forbes*. https://bit.ly/3zNsenI

Tucker, P. (2019, March 31). *The newest AI-enabled weapon: 'Deep-Faking' photos of the Earth*. Defense One. https://bit.ly/3oZefYL

van Huijstee, M., van Boheemen, P., Das, D., Nierling, L., Jahnel, J., Karaboga, M., Fatun, M. (2021). *Tackling deepfakes in European policy*. European Parliamentary Research Service. https://bit.ly/3pkMOcd

United Nations Development Programme. (1994). *Human Development Report*. Oxford University Press. https://bit.ly/3FbIEbU

U.S. Army DEVCOM Army Research Laboratory Public Affairs. (2021, April 29). *Breakthrough Army technology is a game changer for deepfake detection*. https://bit.ly/3oYsCwn

Villasenor, J. (2019, February 14). *Artificial intelligence, deepfakes, and the uncertain future of truth*. BROOKINGS, https://brook.gs/3EZEAeQ

Vizoso, Á., Vaz-Álvarez, M. & García, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation. *Media and Communication*, *9*(1), 291–300. https://bit.ly/323fUUf

דיפ־פייק (זיוף עמוק) הוא יישום טכנולוגי מבוסס בינה מלאכותית, המאפשר לשנות או לעבד את התוכן של תמונות או סרטונים כך שקשה ולפעמים אף בלתי אפשרי להבחין שמדובר בזיוף. כמו יישומים רבים אחרים בתחום הבינה המלאכותית הפך יישום זה בעשור האחרון זול ופשוט יחסית לשימוש, ובחלק מן המקרים הוא אף זמין לכל מי שיש לו גישה לרשת האינטרנט.

הטכנולוגיה מאפשרת שלל שימושים חיוביים בתחומי הקולנוע והטלוויזיה, הרפואה והפסיכולוגיה, החינוך והשימור ההיסטורי וכן בתחום משחקי המחשב, אך היא מביאה עימה 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. הסיכון בטכנולוגיה הזו ממשי עד כדי כך שהיא הוגדרה בשנת 2019 על ידי ארגוני המודיעין האמריקאיים כאיום האסטרטגי החמור ביותר על הביטחון הלאומי.

מחקר זה, שבוצע במסגרת תוכנית ליפקין-שחק לפוסט-אמת ופייק ניוז במכון למחקרי ביטחון לאומי, סוקר ומציג את הטכנולוגיה המשמשת ליצירת זיופים עמוקים ואת מגוון השימושים החיוביים והשליליים המוכרים שלה, תוך דיון במקרי מבחן בולטים מן העולם. כמו כן הוא בוחן את ההתפתחויות בתחום ההיערכות וההתמודדות עם האתגר בעולם ובישראל. המחקר מציג מסקנות שעלו מסימולציה שנערכה בהשתתפות מומחים והמלצות למדיניות רצויה בתחום עבור ישראל, על מנת להיערך כהלכה לקראת האיום ההולך ומתגבש של הדיפ־פייק.

INSS
המכון למחקרי ביטחון לאומי
THE INSTITUTE FOR NATIONAL SECURITY STUDIES
אוניברסיטת תל אביב
TEL AVIV UNIVERSITY