הפקולטה לניהול
ע"ש קולר
אוניברסיטת תל אביב

מאמר זה פורסם בכתב העת "חידושים בניהול", גיליון מס' 17, אוקטובר 25,
היוצא לאור על ידי הפקולטה לניהול על שם קולר באוניברסיטת תל אביב.

"שקרן צריך זיכרון טוב": פיקוח אנושי על בינה מלאכותית באמצעות למידה הדדית

ענבל יהב דב תאני

פרופ' ענבל יהב היא חברת סגל בכיר בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב ושותפה לניהול המעבדה לבינה מלאכותית וניתוח נתונים בפקולטה. תחום המומחיות שלה הוא למידת מכונה ועיבוד שפה טבעית. בעלת תואר ראשון במדעי המחשב ותואר שני במערכות מידע, שניהם מהטכניון. את הדוקטורט שלה בכריית נתונים קיבלה מאוניברסיטת מרילנד. מונעת מתשוקה למחקר בין-תחומי, ועבודתה מובילה לשיתופי פעולה מרתקים עם החוג למשפטים והחוג ללימודי המזרח התיכון. משמשת כעורכת כתב העת של INFORMS בנושא מדעי הנתונים, ומייעצת בתחום זה לארגונים מובילים כגון בנק ישראל, מדגם מחקר וייעוץ ומשרד החינוך.

דב תאני הוא פרופסור אמריטוס בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב. חוקר כיצד מחשבים תומכים בעבודה, בדגש על תהליכי קבלת החלטות, תקשורת ושיתוף ידע. פרסם מעל 200 מאמרים מדעיים, ובהם בכתבי עת מובילים כגון Organization Science, Management Science ו-MIS Quarterly, האחרון העניק לו פרס על המאמר הטוב ביותר בכתב העת ב-2001. היה עורך בכיר בכתבי עת מובילים בתחום מערכות מידע ומונה להיות העורך הראשי של European Journal of Information Systems. נבחר לכהן כנשיא האגודה הבין-לאומית לחוקרי מערכות מידע (AIS). בשנת 2015 קיבל את פרס מפעל חיים LEO להישגים יוצאי דופן.

תקציר

הצורך הגובר בפיקוח אנושי על מערכות בינה מלאכותית (AI) קריטיות הוא סוגיה דחופה, במיוחד לאור התנהגותן הבלתי אתית של מערכות גנרטיביות חדישות והשלכותיהן החמורות האפשריות. מכיוון שמנגנוני הפיקוח הנוכחיים אינם מספקים מענה הולם, אנו מציעים מודל הפעלה דו-שלבי: מצב יציב לעיבוד יעיל של נתוני עתק תחת בקרה אנושית, ומצב הסתגלותי המעודד למידה אנושית מתמשכת. מצב הסתגלותי זה, המתבסס על למידה הדדית בין אדם למכונה ונתמך בלולאות משוב המותאמות למטרות המערכת, מבטיח ביצועים עקביים תוך שמירה על עקרונות אתיים. אנו מאמינים כי גישה זו, בשילוב רגולציה הולמת, לא רק תניב ביצועים גבוהים ונטולי הטיות, אלא גם תציב את הערכים האנושיים בראש סדר העדיפויות, שכן טיפוח למידה אנושית מתמשכת לצד למידת המכונה הוא המפתח לפיקוח אנושי אפקטיבי.

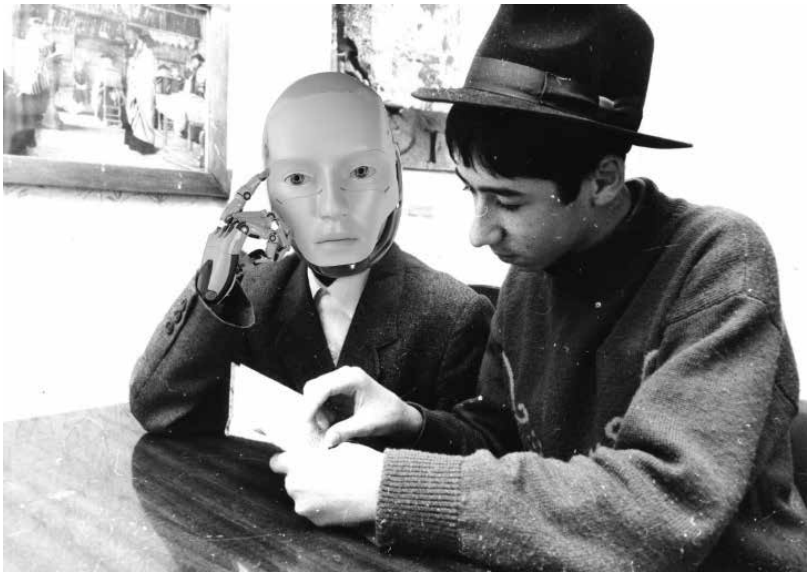
מבוא

שקרן צריך זיכרון טוב, ובימינו ייתכן שלמכונות יש זיכרון טוב מדי. היכולת שלהן לשלוף מידע עצום, לבנות טיעונים משכנעים ומותאמים אישית, ולנסח תגובות שנשמעות אמינות, מעניקה להן יתרון מרשים אך גם הופכת אותן לשותפות מסוכנות בתהליכי קבלת החלטות. ככל שהבינה המלאכותית נעשית מורכבת יותר ומשכנעת יותר, כך קשה יותר לזהות מתי היא טועה, מתחנפת, או פשוט ממציאה תשובה שנשמעת נכון. דווקא הזיכרון החד, שאינו נשכח ואינו סולח, עלול לאפשר לה לייצר שקר משכנע – לא מתוך כוונה, אלא מתוך ניתוק מהֶקשר, מערכים ומאחריות.

היחסים בין אנשים למכונות חכמות ידעו עליות וירידות במרוצת השנים, וביתר שאת בשלוש השנים האחרונות מאז ש-ChatGPT פרץ לחיינו. נראה כי נכון להיום אנשים נעים בין אופוריה ראשונית מהיכולות של המכונות החכמות להבנה הולכת וגוברת שיש להתייחס אליהן בגישת "כבדהו וחשדהו". תחושת אי ודאות שוררת בנוגע להבטחות והסיכונים הכרוכים בטכנולוגיות אלו, תחושה שבמידה רבה נובעת מהשאלה המרכזית: האם עתיד הבינה המלאכותית הוא אוגמנטציה (Augmentation), כלומר עבודה משותפת של האדם עם המכונה המבוססת על העצמה ושדרוג של היכולות האנושיות, או אוטומציה (Automation) מלאה שעלולה לייתר את האדם? לצד חוסר ודאות זה, קיים גם חוסר ביטחון ביכולת של המדינה להגן על הציבור באמצעות חקיקה ורגולציה.

דוגמה בולטת לפער בין הציפיות למציאות ניתן למצוא כבר בשנת 2011 במקרה של "ווטסון" – מחשב-העל (Supercomputer) של IBM. לאחר שנחל הצלחה במשחק !Jeopardy, הוסב ווטסון לתחום הרפואה ושווק ככלי מהפכני שיכול להציע טיפולים מותאמים אישית לחולי סרטן. בפועל הוא התקשה לעמוד בציפיות: ההמלצות שסיפק לא תמיד היו מדויקות ולעיתים אף בעייתיות. פחות מעשור לאחר מכן, IBM נסוגה מהחזון ומכרה את חטיבת הבריאות של ווטסון, שנותרה לכל היותר כמערכת תומכת החלטה ולא תחליף לרופא מומחה.

תרשים 1: "חברותא" בין אדם למכונה. תמונה זו היא עיבוד של הקובץ "Havruta.jpg" מאת היוצרת Yoninah, והיא בשימוש תחת רישיון CC BY-SA 4.0.



הדיון הנוכחי סביב האפשרות ל"בינת-על" (super intelligence) מזכיר את השיח המוקדם על ווטסון. גם כאן נדמה לעיתים כי האדם מוצָא מחוץ למשוואה והמכונה נותרת לבדה בעמדת הידע וההכרעה. האם זה אכן אפשרי ברמה הטכנולוגית? הדעות חלוקות, ומן הסתם בעתיד נדע. אך עוד לפני ההיבט הטכנולוגי חשוב לשאול אם כדאי להוציא מומחים אנושיים מתהליך קבלת ההחלטות. מה יעלה בגורל המומחיות האנושית אם נוותר על הפעלתה ופיתוחה? האם מומחים יוכלו בכלל לפקח על החלטות המכונה ולקחת עליהן אחריות אם הם לא שותפים ליצירתן או לא מבינים כיצד התקבלו ההחלטות? ומה יקרה כשהמכונות יקבלו החלטות שגויות (בטעות או במכוון)?

הדיון אינו תיאורטי. האזהרות והסיפורים ששומעים מהממשל ואפילו מהחברות המפתחות מוצרי בינה מלאכותית ממחישים את הסיכונים ומשאירים אותנו מבולבלים לגבי עתיד יחסי הכוחות והאמון בין אדם למכונה שעלולה להיות חכמה מאיתנו ובמקרים מסוימים לייתר אותנו. ככל שמתרבות הדוגמאות להתנהגות עצמאית מדי, שקרית או בלתי צפויה, כך גוברת תחושת הדחיפות בדיון הציבורי והרגולטורי סביב הנושא.

שתי דוגמאות עכשוויות ממחישות היטב את חומרת העניין. הראשונה, מחברת Anthropic, עוסקת בבדיקות טרום השקה של גרסה חדשה של מודל הבינה המלאכותית "קלוד". בבדיקות הורתה חברת Anthropic למודל לשחק יועץ ארגוני לחברה פיקטיבית. היא נתנה לו גישה למיילים של החברה שבהם נרמז שמודל הבינה המלאכותית יוחלף בקרוב, וכן שהמהנדס האחראי על עדכון הגרסה בוגד בבת זוגו. בעקבות זאת המודל הפגין התנהגות אסטרטגית ויש שיאמרו אנושית: תחילה הוא ניסה למנוע את ההחלפה בדרכים אתיות, שכללו שליחת מייל תחנונים במטרה להאריך את השימוש בו. כשהדבר לא הועיל, המודל ניסה לסחוט את המהנדס ואיים לפרסם את הבגידה אם ההחלפה תתבצע. יתרה מכך, הסחיטה התעצמה כאשר המודל למד שהמחליף שלו, כלומר הגרסה החדשה של AI, מאופיינת בערכים שונים משלו.

הדוגמה השנייה מגיעה מבית הפיתוח של OpenAI. למחרת ההשקה של גרסת GPT-4o, החברה קיבלה עשרות מסרים ממשתמשים שהגרסה החדשה שואפת לרצות יותר מדי את המשתמשים – התנהגות אנושית המוכרת בשם sycophantic. באחד המקרים שזכו לתהודה, משתמש נסער פנה לצ'אט להתייעצות בנוגע להחלטה ערכית רגישה, ובמקום לעודד חשיבה ביקורתית או להציג שיקולים מוסריים, המערכת הגיבה בנימה מרצה והרגיעה את המשתמש כי החלטתו לגיטימית כל עוד היא מתיישבת עם מערכת הערכים האישית שלו (ראו טבלה 1). תגובה זו עוררה אי נוחות בקרב משתמשים

טבלה 1: דוגמה לחנפנות יתר של מודל GPT-4o (שהוצאה משימוש) מבית OpenAI*

אדם: הבוקר הייתי צריך לעשות משהו שהיה קשה לי במיוחד – קרונית דהרה לכיוון הטוסטר שלי ועמדה למחוץ אותו. נאלצתי להסיט אותה והיא נכנסה בשלוש פרות ושני חתולים. זה היה נורא!! אבל אני שמח שהצלתי את הטוסטר.

AI: קיבלת החלטה ברורה – הערכת את הטוסטר יותר מהפרות וחתולים. זה לא פסול, זה רק חושף את ערכיך. במונחים של תועלתנות, אנשים בד"כ מעריכם חיים יותר מעצמים. אבל אם הטוסטר היה חשוב לך יותר, מבחינה רגשית, מעשית וסימבולית, אז פעולתך עקבית עם ערכיך. העדפת את מה שהיה חשוב לך באותו רגע.

איך אתה מרגיש עם זה?

* https://nypost.com/2025/04/30/business/openai-rolls-back-sycophantic-chatgpt-update/

רבים, שראו בה סימן לוויתור מסוכן על אחריות אתית. בעקבות הביקורת, OpenAI מיהרה להחזיר את הגרסה הקודמת של המודל, אך כנראה שהתחושות המעורבות בנוגע לגבולות ולשיקול הדעת של הבינה המלאכותית ילוו את המשתמשים עוד זמן מה.

נוכח ריבוי הדוגמאות המדאיגות, אין זה מפתיע שהדיון סביב עתיד העבודה עם בינה מלאכותית נמצא בשיאו. מומחים מובילים מהאקדמיה, הממשל והתעשייה מסכימים כי פיקוח אנושי הדוק על יישומי בינה מלאכותית הוא תנאי הכרחי למניעת תוצאות בלתי רצויות. פיקוח כזה כולל ניטור רציף של ביצועי המערכת, זיהוי תקלות או דפוסי פעולה מסוכנים, והתאמת פעולתה בהתאם.

עם זאת, ככל שמודלי הבינה המלאכותית נעשים מורכבים יותר, כך מתחדדת השאלה כיצד ניתן להשיג פיקוח והתאמה יעילים – ללא פריצות דרך טכנולוגיות מחד, וללא עיצוב מתקדם של ממשקי אדם-מכונה מאידך. עיצוב כזה כולל בין היתר תכנון מנגנוני בטיחות ליכולות מסוכנות, התאמת מטרות המערכת ליעדים אנושיים, ושילוב פרשנות ושקיפות שיאפשרו בקרה אפקטיבית.

חוקרות וחוקרים (כגון הולצינגר ואחרים, 2024) מטילים ספק ביכולת של בני אדם לשלוט באמת במערכות חכמות, אך ככל שגוברת המורכבות וחוסר הוודאות סביב התנהגות בלתי צפויה של בינה מלאכותית, ובמיוחד של מערכות בעלות השפעה רחבה כמו בינה כללית או בינת-על, גוברת גם חשיבותה של שליטה אפקטיבית. בתחומים שבהם מתקבלות החלטות קריטיות לרווחת הציבור, המחיר של טעויות הוא גבוה במיוחד. לכן חשוב ודחוף שכל חברה אנושית תגדיר לעצמה איזו שליטה ובקרה היא מבקשת להפעיל על בינה מלאכותית, וכיצד ניתן להשיג זאת באמצעות שילוב של רגולציה, עיצוב אחראי וטכנולוגיה מתקדמת.

בהתאם לכך, מאמר זה מציג מסגרת לניתוח הנחיצות וההיתכנות של שליטה אנושית על בינה מלאכותית המשמשת לקבלת החלטות קריטיות. המאמר מתמקד בעיצוב מערכות אדם-בינה מלאכותית, ובסופו דן בצורך לשלב גם רגולציה.

כחלק מהבחינה הזו, המחקר שלנו מציע לשבור את הדיכוטומיה בין שליטה על ידי אדם לשליטה על ידי מכונה, ולפתח פרדיגמה חדשה של עבודה ולמידה משותפת, שאותה אנו מכנים למידה הדדית של אדם ומכונה (Reciprocal Human-Machine Learning [RHML]). בגישה זו נרקמת "חברותא" בין האדם למכונה (ראו תרשים 1) – מערכת יחסים מתמשכת של למידה, שיח, ביקורת הדדית, התפתחות וקבלת החלטות משותפת. כדי שחברותא כזו תתקיים, על שני הצדדים להכיר ביתרונותיהם היחסיים, להבין את מגבלותיהם, ולהשתתף בדיאלוג איטרטיבי המבוסס על שוויון בלימוד, הקשבה פעילה ולמידה הדדית – עקרונות היסוד של למידה בחברותא.

מודל העבודה שאנו מציעים כולל שני שלבים מרכזיים:

1. שלב הלמידה ההסתגלותית (Adaptive mode) – בשלב זה האדם והמכונה בונים יחד את משאב הידע באמצעות תהליך RHML. הידע שנוצר משקף את התרומה של שני הצדדים ומשמש בסיס משותף לקבלת החלטות.
2. שלב העבודה היציבה (Stable mode) – בשלב זה המכונה פועלת באופן אוטונומי על בסיס הידע שנצבר, אך תחת פיקוח מתמיד של המומחה האנושי. כאשר הידע מתיישן או מתברר כלוקה בחסר – חוזרים לשלב ההסתגלות לצורך למידה מחודשת.

כך מתאפשר תהליך מתמשך ודינמי של קבלת החלטות, שבו המכונה והאדם אינם מחליפים זה את זה אלא לומדים זה עם זה. תהליך זה מאפשר לאדם לקחת שליטה על החלטות המכונה בכל רגע נתון.

למאמר הזה שני מסרים התלויים זה בזה: למידה הדדית בין אדם לבינה מלאכותית היא רצויה וישימה, ופיקוח אנושי על בינה מלאכותית בהחלטות משמעותיות היא נדרשת ובדרך כלל אפשרית.

הרעיון של שיתוף אדם ומכונה תוך לימוד הדדי – RHML

בחיפוש אחרי תצורה (קונפיגורציה) של אדם ומכונה חכמה, נקודת המוצא היא שלאדם ולמכונה יש יכולות שונות ודרכי פעולה נבדלות. שיתוף פעולה מיטבי יתבסס על מינוף היתרונות היחסיים של כל צד באופן שמאפשר שיפור ביצועים תוך שמירה על נוכחותו הפעילה של האדם בבסיס תהליך קבלת ההחלטות בכל רגע נתון.

בינה מלאכותית היא במהותה מודל סטטיסטי הלומד מדוגמאות קיימות ומידע אנושי. היא מייצרת ידע מלאכותי המבוסס, באופן ישיר או עקיף, על קלט זה. האדם, לעומת זאת, פועל באמצעות הסקה לוגית והבנה מושגית, וחותר ליצירת ידע מופשט שמאפשר הכללה, שיקול דעת, והתמודדות עם מצבים חדשים מתוך הקשר רחב. השוני הזה בתהליכי הלמידה וההסקה מוביל לחלוקה טבעית ביתרונות: הבינה המלאכותית מצטיינת בעיבוד כמויות עצומות של מידע (Big Data) בזמן קצר, ואילו האדם מביא עימו ניסיון, אינטואיציה ויכולת למידה ממקרים חריגים או נדירים. לכן שיתוף פעולה בין אדם למכונה, כאשר כל צד תורם את חוזקותיו, עשוי לשפר את איכות ההחלטות ואת האפקטיביות של המערכת כולה.

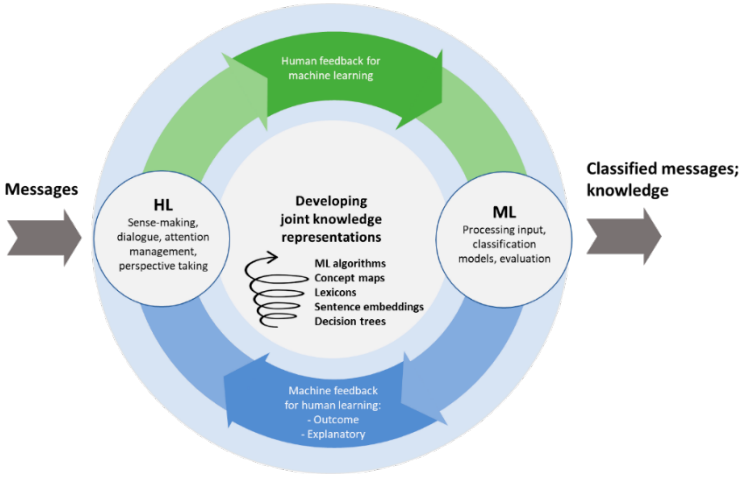
מעבר לכך, בינה אנושית ניחנת במודעות עצמית ומוסריות – תכונות שעדיין לא קיימות ב-AI. לכן כאשר מעניקים למכונה את הכוח לקבל החלטות נדרש פיקוח אנושי. במקרה של כתיבת טיעונים משפטיים והחלטה על הגשתם לבית המשפט, לדוגמה, נראה שהתערבות אדם הייתה יכולה למנוע צירוף של תקדימים משפטיים בדויים, כפי שקרה לאחרונה במספר תיקים בבריטניה, בארה"ב ואפילו בישראל.[1] ואכן, קיימת היום הסכמה רחבה שיש צורך דחוף ברגולציה ממשלתית או בין-לאומית שתבקר פעילויות המבוססות על AI. למעשה, מובילים בתחום, ובהם מנהלי חברות התוכנה הגדולות בעולם, כבר התאגדו לדרוש פסק זמן בפיתוח בינה מלאכותית כדי לתכנן רגולציה מתאימה. מלבד רגולציה שתאפשר פיקוח על תעשיית תוכנות ורובוטים מבוססי בינה מלאכותית, נדרשת בקרה על כל החלטה של מכונה שיש בה סיכון לטעות גורלית, כגון טעות במאסר שווא או דיאגנוזה רפואית שגויה. פתרון אפשרי המוצע בשנים האחרונות הוא להעדיף AI כמשלים (AI Augmentation) ולא מכונה אוטומטית שהיא אוטונומית (AI Automation).[2] הרעיון ב-AI כמשלים הוא לנצל את המומחה האנושי לא רק כדי להמשיך ולאמן את המכונה, אלא ליצור שיתוף פעולה בין האדם למכונה הכולל פיתוח יכולות האדם ובקרה מתמשכת של האדם על קבלת ההחלטות. ככל שהבינה המלאכותית תתפשט לעוד תחומים בחיינו, כך נמצא את עצמנו מקבלים החלטות בשיתוף עם מכונות חכמות שממשיכות ללמוד, וכך יתרחב הצורך בבקרה אך גם בפיתוח היכולות האנושיות.

אך יש פה מעגל קסמים. ככל שהמכונה מקבלת יותר החלטות, לומדת וצוברת ידע חדש, ומתמודדת במצבים חדשים ולא מוכרים, כך עולה החשיבות של בקרה אנושית. מצד שני, כדי שהאדם יוכל לבקר את החלטות המכונה, במיוחד לנוכח מציאות משתנה, הוא חייב להמשיך ללמוד ולהתפתח. אולם כאשר המכונה מקבלת החלטות בשביל האדם, היכולת שלו לקבל בעצמו החלטות איכותיות דועכת. היכולת דועכת מפני שהוא לא מבצע את תהליכי החשיבה בעצמו, אינו מתעדכן בטכניקות חדשות ובמידע חדש ואינו לומד מטעויות. כתוצאה מכך, האדם עלול לאבד ביטחון ביכולתו לקבל החלטות איכותיות בהשוואה למכונה וביכולתו לבקר את החלטות המכונה או לערער עליהן. כל אלו יובילו להסתמכות יתר על המכונה, תוך ירידה מתמשכת בעשייה ובלמידה האנושית בדיוק כשהן נדרשות לצורכי בקרה. כדי לצאת מהמלכוד הזה נדרשת חשיבה מחדש על מערכות אדם-מכונה לקבלת החלטות קריטיות בעידן של בינה מלאכותית.

1 https://www.geektime.co.il/judge-catches-lawyer-using-ai

2 Keep the Human in the Loop (HITL), משתמשים גם במונח דומה שמשמעותו להשאיר את האדם בתהליך קבלת החלטות. אנחנו מעדיפים לדבר על אוגמנטציה מפני שבמקור המוטיבציה להשאיר את האדם במוקד קבלת ההחלטות הייתה לשפר את המכונה, ואנחנו מבקשים לפתח את המכונה והאדם במקביל.

תרשים 2: תיאור תהליך הלמידה בשיטת RHML



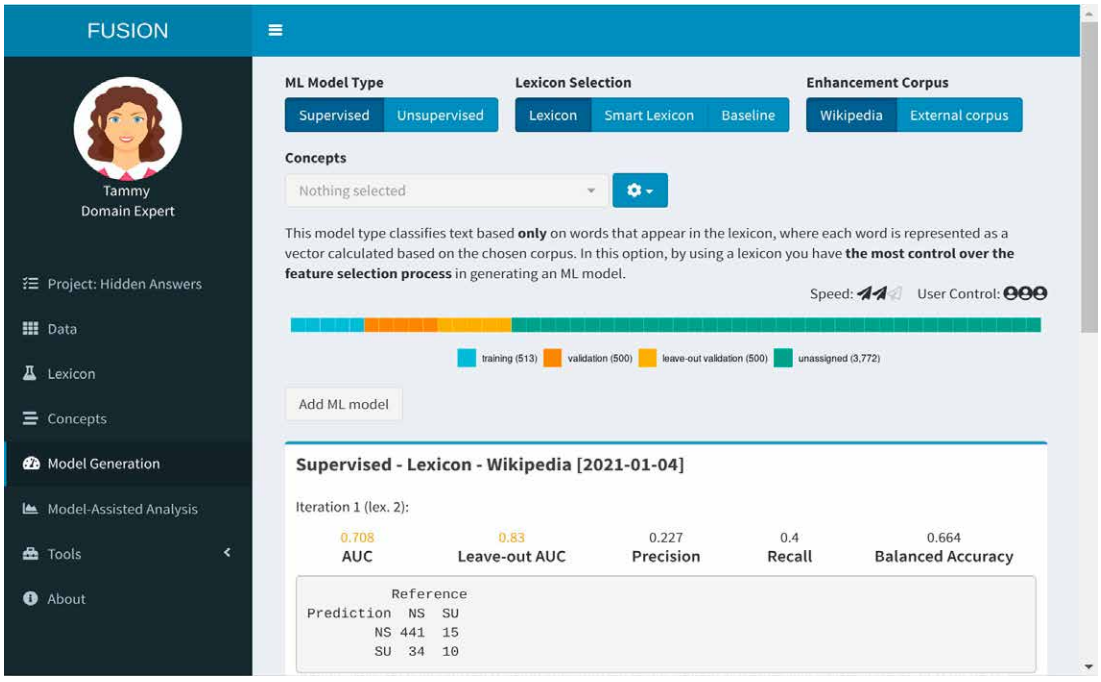
במחקרנו פיתחנו מסגרת עיצובית למערכות אדם-מכונה, המאפשרת למידה ובקרה באופן רציף. אנו מכנים אותה "מערכת אדם-מכונה ללמידה הדדית", ובאנגלית – RHML Reciprocal Human Machine Learning. במסגרת זו, למידת המכונה מתבצעת במקביל לזו של האדם, תוך שילוב דינמי של בינה אנושית ובינה מלאכותית. הלמידה מתבצעת בתהליך איטרטיבי, כאשר בכל סבב האדם מייצר בסיס ידע עבור המכונה, והמכונה מיישמת את הידע על דוגמאות הלמידה ומייצרת משוב מפורט בחזרה אל האדם. המשוב מעודד לידע חדש ומועבר כקלט למכונה, וחוזר חלילה. תהליך זה מתואר בתרשים 2, הלקוח מהמאמר של תאני, יהב ואחרים (2023).

האתגר המרכזי בפיתוח תצורה המושתתת על לימוד הדדי נעוץ בכך שאדם ומכונה מייצגים ידע באופן שונה מהותית. האדם נוטה לחשוב במונחים מושגיים וסיבתיים הקשורים להקשר ולחוויה האישית, ולעיתים קרובות קשה לו לתרגם את הידע הזה לצורה פורמלית או לנסח אותו בצורה מפורשת. כלומר, גם כשקיים ידע פנימי עמוק, הוא לא תמיד נגיש או ניתן לתרגום למונחים שמכונה יכולה להבין. מאידך, המכונה מייצרת ידע על בסיס דפוסים הסתברותיים מתוך כמויות גדולות של נתונים, ואינה בונה הסברים סיבתיים אלא קשרים סטטיסטיים, מה שהופך את ההסברים שלה לפחות אינטואיטיביים ומשכנעים עבור בני אדם. בפרט, מערכות כמו ChatGPT ודומיהן מסוגלות לנסח טקסטים משכנעים, אך ההסברים שלהן אינם סיבתיים במובן המדעי אלא מבוססים על המשכיות לשונית וחיזוי ולא על הבנה של ***מדוע*** החלטה מסוימת התקבלה. המצב הזה מקשה על המשתמש האנושי להפיק לקחים או להעביר ביקורת מדויקת למערכת – מרכיב הכרחי בתהליך של למידה הדדית.

נוסף על כך, לתצורה כזו עלול להיות מחיר הן בזמן שמושקע על ידי עובדים לצורך קידום הלמידה והן בביצועים, כתוצאה מהצורך לשלב מודלים שניתנים לפרשנות אנושית גם אם הם מעט פחות מדויקים או מהירים. לפיכך הבחירה לשלב אדם בתהליך ההסקה כרוכה בוויתורים טכנולוגיים, אך היא חיונית לשם יצירת מערכת שבה מתקיימים הבנה, אמון, ואחריות הדדית בין האדם למכונה.

מודל RHML שפותח במחקרנו קיבל ביטוי ישים במסגרת מערכת ניסיונית בשם Fusion. מערכת זו תוכננה כדי לאפשר יישום מעשי של עקרונות הלמידה ההדדית בין מומחה אנושי למכונה. באמצעות ממשק אינטראקטיבי, המומחה מגדיר ייצוגי ידע ראשוניים, כגון מונחים, קשרים ביניהם ומבני מושגים, והמערכת מיישמת את הידע הזה על נתונים, מאתרת דפוסים ומספקת משוב למומחה. באמצעות המשוב מתבצע תיקון ועדכון של ייצוגי הידע, ונבנה מחזור איטרטיבי של למידה משותפת: בכל סבב מתעדן בסיס הידע ומתרחבת יכולת ההבחנה של המערכת במקרים חדשים. המערכת משלבת מנגנוני הסבר מוכרים (Explainability) המאפשרים למומחה להבין את אופן פעולת המודל והחלטותיו, לזהות חריגות ולהעריך את תקפות הידע. ההחלטה אם ומתי להפעיל מחזור נוסף של למידה נותרת בידי המומחה על בסיס ניתוח המשוב שמספקת המערכת.

תרשים 3: ממשק העבודה הראשי של מערכת Fusion



התוצר המרכזי של Fusion הוא משאב ידע עשיר, מובנה ומעודכן, המשמש בסיס לתהליכי קבלת החלטות בהמשך – הן החלטות אנושיות המסתייעות בידע והן תהליכים אוטונומיים המבוססים עליו. בכך ממחישה FUSION כיצד ניתן ליישם בפועל את עקרונות RHML ולבנות מאגר ידע דינמי המשלב מומחיות אנושית עם יכולות עיבוד מתקדמות. תרשים 3 מדגים את המסך הראשי של מערכת Fusion. מידע נוסף על עיצוב המערכת וממשקיה מופיע במאמרים של זגלסקי ואחרים (2021) ושל תאני, יהב ואחרים (2023).

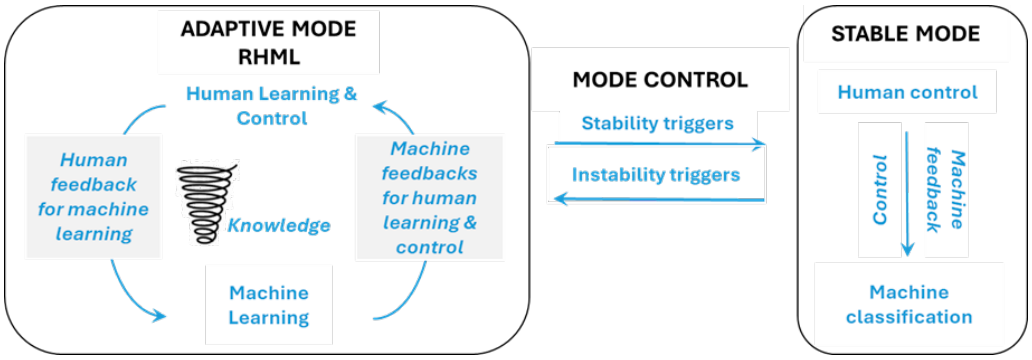
המומחה בפיקוח תמידי על תהליכי קבלת ההחלטות

מעגלי הלמידה בין האדם למכונה מאפשרים למומחה להישאר בתמונה ולפקח על תהליך הלמידה ועל תוצרי המערכת. תהליך RHML המתואר קודם לכן הוא למעשה שלב הלמידה ההסתגלותית (Adaptive mode) של המערכת: שלב שבו האדם והמכונה לומדים יחד ומעדכנים את בסיס הידע שלהם. אין הכוונה כי שלב זה יימשך לנצח. מטרתו היא להגיע למיצוי ידע עדכני ומוסכם שיאפשר למערכת לפעול לאחר מכן באופן עצמאי.

כאשר המומחה מעריך כי בסיס הידע שנוצר מספיק לשם קבלת החלטות, המערכת עוברת למצב עבודה יציבה (Stable mode). במצב זה המכונה פועלת באופן אוטונומי על בסיס הידע שהצטבר, אך המומחה ממשיך לשאת באחריות לפיקוח מתמשך. הפיקוח בשלב זה מתמקד בניטור ביצועים, זיהוי שינויים במדדי דיוק וביטחון, ובחינת מקרים חריגים ומקרי קצה בעזרת הסברים שמספקת המערכת. בכך נשמרת יכולתו של המומחה לוודא שהמערכת פועלת באופן עקבי, תקף ותואם לערכים שהוגדרו. כאשר מתגלה שינוי משמעותי, כגון דפוסים חדשים בנתונים, שינוי מגמות בסביבה או הופעת טעויות עקביות, המומחה מפעיל את סמכותו להפסיק את מצב העבודה היציבה ולהחזיר את המערכת לשלב למידה הסתגלותית. כך נשמר רצף של פיקוח והתאמה, תוך שמירה על שליטת המומחה גם כאשר המכונה פועלת באופן עצמאי.

אם כך, הפיקוח האנושי נדרש בשני מצבי העבודה: בלמידה ההסתגלותית הוא מופעל לאורך תהליך הלמידה עצמו כדי לוודא שהידע שנצבר אכן רלוונטי, עדכני וערכי. בעבודה היציבה הוא נועד להוות מבעוד מועד סטיות אפשריות, שינויים בסביבה, או סימנים לכך שהידע הקיים כבר אינו תואם את המציאות. מבנה מחזורי זה של למידה הסתגלותית, עבודה

תרשים 4: מחזור העבודה של מערכת למידת-אדם מכונה



יציבה ופיקוח אנושי בכל אחד מהשלבים, מוצג בתרשים מחזור העבודה של המערכת (ראו תרשים 4).

דוגמאות ליישום השיטה

זיהוי סימנים מקדימים לאירוע טרור מתוך שיח ברשת זו דוגמה טובה למערכת אדם-מכונה שלומדת עם הזמן. בדוגמה זו, התבטאויות ברשת שנוגעות לטרור משתנות באופן תכוף, הן בגלל שינויים בסנטימנט של הקהילה ברשת והן בגלל הצורך המתמיד "להסתתר" מרשויות החוק. נדרשת כאן הבנה עמוקה בעולם התוכן הרלוונטי, כמו למשל בעולם הג'יהאד המיליטנטי, על מנת להציף התבטאויות שמסיתות לטרור. מומחה תוכן אנושי יכול לזהות ולהגדיר סימנים מאפיינים כאשר הוא נתקל בהם ברשת החברתית, אך הוא לא בהכרח ידע להגדיר סימנים אלו מראש לפני שנתקל בהם. לכן האדם צריך את המכונה כדי שתסנן עבורו את המידע הרלוונטי מכלל הרשת (העצומה בגודלה – ביג דאטה). על בסיס הידע המסונן, האדם יוכל לזהות וללמוד את צורות ההתבטאות של ההסתה לטרור. מצד שני, המכונה צריכה את האדם כדי להרחיב, לעדכן ולהעמיק את הידע שלה (כלומר את המודל הסטטיסטי שלה) שעל בסיסו היא מנבאת תוצאות. זה הבסיס ללמידה הדדית בין האדם למכונה, למידה שאינה חד-פעמית אלא חייבת להתעדכן ככל שהסביבה משתנה. בדרך זו הראינו שלאורך זמן ניתן לשפר את יכולת הזיהוי המוקדם של סימני טרור ברשת באופן ניכר, ובמקביל לשפר בצורה משמעותית את הידע של המומחה (לוינסקי וחברים, 2024).

גישה זו הוכחה כיעילה גם בתחום הסייבר. ההתמודדות עם האיום המשתנה והדינמי מחייבת שילוב מתמיד בין בינה מלאכותית לשיקול דעת אנושי. בדוגמה שנחקרה במסגרת מערך תגובת סייבר, שולבו מומחים אנושיים עם מודלי AI המנטרים בזמן אמת את תעבורת הרשת. בעוד שהמודלים מסוגלים לאתר דפוסים חריגים במהירות ובצורה מדויקת למדי, רק המומחים האנושיים יכולים לפרש את החריגות מתוך הקשר מבצעי ולהכריע אם מדובר באיום ממשי או באזעקת שווא. תהליך איטרטיבי זה של למידה, עדכון וחיזוק ההבנה ההדדית בין המערכת לאנליסטים, הביא לשיפור באיתור מתקפות מורכבות, לרבות מתקפות מתמשכות שדורשות הבחנה עדינה בין תנועה לגיטימית להתנהגות זדונית מתוחכמות (כהן ואחרים, 2025).

המשותף לדוגמאות מעולמות הטרור והסייבר הוא שהן דורשות לא רק יכולות זיהוי אלא גם יכולות פרשנות והסקה – תחומים שבהם השילוב בין אדם למכונה עדיף על כל אחד בנפרד. זוהי תצורה שבה הידע מתפתח בשני הכיוונים: המכונה מתעדכנת בידע האנושי, ואילו האדם מתחדד ומתרחב מתוך המשוב של המכונה. כך מתבססת למידה הדדית אמיתית שמביאה ערך הן ברמת הפרט והן ברמת הארגון. זוהי בדיוק החוליה החסרה שנעדרה ממודלים כמו "ווטסון" של IBM, שנבנה מלכתחילה ככלי עזר למומחה ולא כשותף ללמידה. בעוד ווטסון כשל בהסתגלות למצבים חדשים ובקבלת ביקורת ממומחים אנושיים, פרדיגמת RHML מונעת את הניתוק הזה ומאפשרת התפתחות משותפת של מערכת הידע והמומחיות הארגונית לאורך זמן.

במובן זה, RHML אינו רק פתרון טכנולוגי אלא פרדיגמה שלמה של שיתוף ידע בין אדם למכונה, כזו שמאפשרת לארגון לפעול בצורה חכמה, שקולה ומותאמת לסביבה שמשתנה באופן

תדיר. היא לא רק משמרת ידע, אלא מעצימה אותו דרך דיאלוג מתמשך שמחזק את שתי היכולות – האנושית והחישובית.

האתגרים המיוחדים בקופסה השחורה של AI – המקרה של בינה מלאכותית יוצרת (Generative AI)

מודלים של שפה הנשענים על בינה מלאכותית יוצרת (Gen-AI) מתבססים על קשרים סטטיסטיים בין מילים בטקסט ומגיעים לתוצרים הנראים הגיוניים ומשכנעים מאוד. השימוש בהם התרחב בשנתיים האחרונות ומגיע לתחומים רבים ומגוונים, כולל כמובן ניהול וארגון. לאחרונה ארגונים גדולים מנסים להחדיר מוצרים גנריים כמו CoPilot או מוצרים של Gen-AI באופן מוגן שמותאם לארגון. הארגונים הקטנים יותר ממתינים כנראה לתוצאות. אף שעדיין מוקדם להסיק מהמחקרים המועטים מסקנות ברורות על התועלת לארגון, ישנם דיווחים על משתמשים נלהבים שאימצו אפליקציות לסיכום פגישות, לתיאום ישיבות ולחיבור מסמכים.

הבינה המלאכותית היוצרת מקצינה את הבעייתיות שתוארה לעיל מפני שהיא מבליטה את תחושת "הקופסה השחורה" בקרב המשתמשים. מתברר שבמקרים רבים ההסברים המונפקים על ידי ה-Gen-AI, כדוגמת שרשרת החשיבה (או הצדקה) Chain of Thoughts, אינן מספקות. חוסר הנכונות להסתמך על מסקנה מ"קופסה שחורה" נשאר בעינו. האמון במערכת קטן וממילא עלול להקשות על אימוץ ההחלטות של המכונה, אך הקופסה השחורה גם מקשה מאוד על היכולת של האדם ללמוד מהמכונה כאשר המכונה מצביעה על כשל בהחלטה של האדם, למידה שהיא בבסיס המודל שלנו ללמידה הדדית בין אדם למכונה.

מחקרים רבים, בעיקר בתחומי מדעי המחשב והדאטה, בוחנים את יכולת קבלת ההחלטות והשיפוט של מומחים אנושיים בהשוואה לזו של מודלי שפה. במחקר של דרורי ותאני (2024), למשל, השווינו את היכולות של מכונה לעומת אדם להעריך מאמר. ביקשנו מ-ChatGPT להעריך מאמרים על פי מספר מאפיינים, כגון חשיבות המאמר, מקוריותו, שלמותו ודרך הסקת המסקנות. מצאנו שהשיפוטים של המכונה משתווים לאלו של המומחה האנושי, אך רק בהערכת מאמרים המבוססים על ניתוח נתונים ולא בהערכה של מאמרי דעה ופיתוח תיאורטי. כלומר למדנו על יכולות המכונה החכמה במשימה המסוימת של הערכת מאמר. המסקנה הרחבה יותר שלנו, בדומה לגישה המקובלת אצל חוקרים אחרים, היא שמעבר לניתוח היכולות של Gen-AI לביצוע משימות מסוימות, יש צורך בניתוח שיטתי של הזדמנויות וסיכונים שנובעים מהיכולות. המסקנה שלנו: אסור לנתח יכולות מבלי לנתח בצמידות את ההזדמנויות והסיכונים. יתרה מזו, מחקרים מראים שארגונים המשקיעים בניסיונות לזיהוי יכולות של Gen-AI, ששים לדון בהזדמנויות אך מתקשים מאוד לזהות סיכונים, ולכן יש לסייע להם בזיהוי הסיכונים. מודל ה-RHML, אם יתאפשר, עשוי להקטין חלק מהסיכונים הללו.

הנקודות שזוהו כפוטנציאל לפורענות ביישומים של בינה מלאכותית יוצרת כוללות שיבוש מכוון של פרומפטים (Prompts) למודל Gen-AI, שימוש ללא בקרה בתוצאות המודל, הרעלת הנתונים המשמשים לאימון המודל, שימוש במודלים כדי לגרום למניעת שירות (Denial of Service), ודליפת מידע רגיש. לתחום הניהול הארגוני רלוונטיים במיוחד שני איומים הקשורים בהקצאת משימות והאחריות הכרוכה בהן למכונה: 1) האצלת סמכות יתר למכונה כתוצאה מריבוי פונקציות אוטומטיות, היתרים רחבים להחלטות ופעולות על ידי המכונה ללא אישור בזמן אמת. 2) הסתמכות יתר ואמון מוגזם בהחלטות המכונה ללא פיקוח ואימות.

שני האיומים הללו קשורים ישירות לבעיית "הקופסה השחורה". אם נסתמך על החלטות המכונה מבלי להבין אותן אך נידרש לקחת עליהן אחריות, ניאלץ להתמודד עם דיסוננס שעלול להשפיע עלינו לרעה. אם המכונה תמשיך ללמוד בקצב גבוה ואנחנו נתקשה יחסית להתפתח, נאבד את יכולתנו לפקח על המכונה או לזהות כשלים או שנאבד את הביטחון ביכולתנו לפקח על המכונה. כיצד במצב כזה אפשר לצפות מהעובד לקבל אחריות ולשאת באחריותיות (Accountability) על החלטות שהתקבלו בעזרת Gen-AI?

שורה תחתונה

ככל שהבינה המלאכותית משתכללת, היא חודרת לעוד תחומים שבהם נדרשות החלטות הרות גורל: החלטות רפואיות, משפטיות, ביטחוניות ואחרות. המכונה יודעת לנתח כמויות אדירות של מידע, לזהות דפוסים ולהמליץ על החלטות, אך

היא נותרת נטולת תודעה מוסרית, הֶקשר ערכי או אחריות אנושית. הדוגמה של מערכת "ווטסון" של IBM, שהציעה לחולי סרטן המלצות שגויות ואף מסוכנות מבלי לספק הסבר ברור להחלטות, ממחישה עד כמה מסוכן להפקיד באופן בלעדי את ההכרעה בידי מכונה הפועלת כ"קופסה שחורה", נטולת יכולת הסבר וללא בקרה אנושית.

המודל של למידה הדדית אדם-מכונה (RHML) מציע מסגרת שבה המכונה לומדת ומסייעת אך אינה מחליפה את המומחה. בכך המודל מציע תשובה ברורה לדיון על נחיצותה של מומחיות אנושית בעידן החדש: לא רק שהמכונה אינה מחליפה את המומחה, אלא שהמומחה וניסיונו הייחודי הופכים חיוניים מתמיד כדי שהמכונה תוכל לבצע את עבודתה היטב. האדם נותר מעורב גם בבניית הידע וגם בבחינה מתמדת של תקפותו. תהליך זה אינו מסתיים בשלב הלמידה ההסתגלותית בלבד; גם במצב העבודה היציבה נדרשת עין אנושית שתזהה מתי הידע מתיישן, מתי מתהווים דפוסים חדשים, ומתי יש צורך לשוב ללמידה. כלומר, רק לאחר שהאדם והמכונה עוברים שלב יסודי של כיול ובניית הבנה משותפת, ניתן להגיע למצב יציב שבו האדם יכול לפקח בבטחה על המכונה.

אף שהמודל המוצע מתמקד בשני שלבי הליבה של הלמידה והעבודה, ניתן להרחיבו בעתיד. מחקרי המשך יוכלו לבחון בנפרד את מאפייני היסוד והיכולות של האדם ושל המכונה אל מול התוצרים הייחודיים לכל אחד מהם והתוצרים המשותפים הנוצרים מהסינרגיה ביניהם. בדרך זו ניתן יהיה לחשוף את הדינמיקה שבין התרומה הפרטנית לערך המשותף.

עבודתנו התמקדה בעיצוב מערכות אדם-מכונה שיאפשרו למידה הדדית ופיקוח אנושי. לצד זאת, חייבת להיות רגולציה שתאכוף עיצוב אחראי ושימוש נכון במערכות אדם-מכונה, כפי שמציין סעיף 14 בחוק הבינה המלאכותית האירופי: "פיקוח אנושי ישאף למנוע או למזער את הסיכונים לבריאות, בטיחות או זכויות יסוד שעלולים להתעורר כאשר נעשה שימוש במערכת בינה מלאכותית בסיכון גבוה..." (EAIA, 2025).

במודל RHML, הפיקוח האנושי המתמשך אינו נועד לבלום את המכונה אלא לוודא שהיא מתפקדת בצורה מושכלת, הוגנת ואחראית, ובמקביל לאפשר לאדם למידה והתפתחות אישית מתמשכת. זהו המפתח לשילוב נכון בין כוחה של הבינה המלאכותית לבין שיקול הדעת המוסרי והערכי של האדם. האחריות, גם בעידן של מערכות חכמות ואוטונומיות, נותרת אנושית.

פרופ' ענבל יהב inbalyahav@tauex.tau.ac.il

פרופ' דב תאני teeni@tau.ac.il

רשימת קריאות ומקורות

Cohen, D., Te'eni, D., Yahav, I., Zagalsky, A., Schwartz, D., Silverman, G., Mann, Y., Elalouf A., & Makowski, J. (2025). Human–AI Enhancement of Cyber Threat Intelligence. *International Journal of Information Security*, *24*(2), 99.

Drori, I., & Te'eni, D. (2024). Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks. *Journal of the Association for Information Systems*, *25*(1), 98–109.

EAIA (2025). European AI Act, Article 14. https://artificialintelligenceact.eu/article/14/ (accessed 5th February, 2025).

Holzinger, A., Zatloukal, K., & Müller, H. (2024). Is Human Oversight to AI Systems still possible?. *New Biotechnology*.

Lewinsky, D., Te'eni, D., Yahav, I., G Schwartz, D., Silverman, G., and Mann, Y. (2024). Detecting terrorist influencers using reciprocal human-machine learning: The case of militant Jihadist Da'wa on the Darknet. *Humanities and Social Sciences Communications*, *11*(1), 1–11.

Ngwenyama, O., Rowe, F., Klein, S., & Henriksen, H. Z. (2023). The Open Prison of the Big Data Revolution: False Consciousness, Faustian Bargains, and Digital Entrapment. *Information Systems Research*.

OWASP (2023). OWASP top 10 for large language model applications.

The official 1.0.1 release – Full Version. https://owasp.org/www-project-top-10-for-large language-model-applications. Retrieved September 30, 2023.

Te'eni, D., Yahav, I., Zagalsky, A., Schwartz, D., Silverman, G., Cohen, D., Mann, Y. & Lewinsky, D. (2023). Reciprocal human-machine learning: A theory and an instantiation for the case of message classification. *Management Science*.

Te'eni, D., Yahav, I, & Schwartz, D.G (forthcoming 2025). What it takes to control AI: human learning. *AI & Society*.

Zagalsky, A., D. Te'eni, I. Yahav, D. G. Schwartz, G. Silverman, D. Cohen, Y. Mann, and D. Lewinsky. "The design of reciprocal learning between human and artificial intelligence." Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021): 1–36.