הארץ

איך מלמדים רובוטים להבדיל בין טוב לרע?

יחסי אישות עם מערכת הפעלה? הולוגרמה שמספקת תמיכה רגשית ונפשית? מה שעד כה כיכב ביצירות מדע בדיוני הופך למציאות מהר משחשבנו. אז איך מבטיחים שרובוטים ומכונות עם מחשבה עצמאית לא יהרסו לנו את החיים? למדע כמובן יש פתרונות יצירתיים יותר מלהוליווד

עדית פרנקל 17.01.2018 17:06

לד״ר גלית ולנר, מנחה בתוכנית הרב־תחומית למדעי הרוח של אוניברסיטת תל אביב, יש מנהג קבוע. בשיעורים הראשונים של הקורס שלה בפילוסופיה של הטכנולוגיה היא מדברת על ״חוקי אסימוב״, אותם ״שלושת חוקי הרובוטיקה״ שסופר המד"ב האגדי טבע בשנות ה–40 והפכו למפורסמים הודות לספרו המכונן מ–1950, ״אני, רובוט״. אסימוב, שיצק את היסודות למחקר היחסים בין האדם למכונה, בנה היררכיה ברורה: תפקיד הרובוט הוא להגן על האדם, לציית לו, ולשמור על עצמו (בתנאי שהחוק האחרון הזה לא יפגע בשני הראשונים). בסדר הזה. ואף רובוט לא יכול להפר את שלושת החוקים בלי לצאת מכלל פעולה, ולו משום שהם מוטבעים באופן פיזי – מקודדים – במעגלי המוח שלו. "אני, רובוט", על שלושת החוקים שלו, התקבל באהדה לא רק בקרב קהל הקוראים חובבי הז׳אנר, אלא גם בחוגי המדע והפילוסופיה. אבל לד"ר ולנר עצמה יש בעיה עקרונית עם החוקים שטבע. ״מצד אחד אלה חוקים שעוזרים לחשוב על טכנולוגיות, בעיקר אלו הרובוטיות. ובזה הם עושים עבודה מצוינת. אבל רוב הזמן הם מניחים יחסי מרות. כלומר, רובוט חייב לציית לפקודות של אדם. כך שלא מדובר במערכת יחסים הגיונית, אלא בפנטזיית העבד״.

אצל אסימוב הרובוטים ממשיכים לציית. בסיפורים שכתב סביב ״שלושת חוקי הרובוטיקה״, הציות אמנם קולע אותם לקונפליקטים אבל הם לא מרחיקים לכת עד כדי מרידה. בקלאסיקות האחרות של ז'אנר ״הרובוט כעבד", לעומת זאת, הגולם המכני סופו לקום על יוצרו. בין היצירות האלו ניתן למנות את ״האם אנדרואידים חולמים על כבשים חשמליות״ של פיליפ קיי דיק (שעל פיו נכתב התסריט של ״בלייד ראנר״), ״2001: אודיסיאה בחלל״ של ארתור סי. קלארק (המקור לסרטו של סטנלי קובריק), ״עריצה היא הלבנה״ (רוברט היינלין), ״נוירומאנסר״ של ויליאם גיבסון (הנחשב גם לאבי ז'אנר הסייבר־פאנק) בספרות; ״מטרופוליס״, ״אקס מכינה״, סדרת ״המטריקס״ ו״שליחות קטלנית: יום הדין״ בקולנוע; "באטלסטאר גלקטיקה״, ״כמעט אנושיים״, ״ווסטוורלד״ ו״מראה שחורה״ האופנתית בטלוויזיה – וזו רק רשימה חלקית. ובכל היצירות האלו מתקבלת תמונה מאיימת של עתיד שבו מכונות בעלות בינה מלאכותית לא יקבלו עוד את הסמכות האנושית. כל אחת מהן אמנם תוקפת את הנושא ממקום אחר, כל אחת מציעה פרשנות קצת שונה, אבל התימה השולטת זהה: ״בז'אנר המד"ב, הפנטזיה השכיחה היא העבד הרצחני, וזה מה שמפמפמים לנו", ולנר אומרת. "עובדתית, אין מספיק התייחסות לתרחישים אחרים שבהם אנחנו יכולים לנהל מערכות יחסים אחרות עם המכונות״.

בשנים האחרונות הקולנוע דווקא היטיב לספק כמה דוגמאות למערכות יחסים – בין אדם למכונה – מסוג אחר. בין שזה בסרט "HER" של ספייק ג׳ונס, שבו מנהל הגיבור (חואקין פיניקס) מערכת יחסים אינטימית עם מערכת ההפעלה של מחשבו, או בסרט ״מרג׳ורי פריים״ מ–2017, שבו ג׳ון האם משחק את ״וולטר״, הולוגרמה חכמה של בעלה המנוח של מרג׳ורי (לואיס סמית) שתפקידה העיקרי הוא להיות לה לבן לוויה ולהקל על הבדידות הנלווית למחלה ולזיקנה. אפילו ב״בלייד ראנר 2049״, סרט ההמשך של המשל הקלאסי על המכונה המתקוממת, מוצגת אלטרנטיבה ליחסים בין בני אדם

למכונות ובין מכונות למכונות – סיפור האהבה מסרט המקור בין דקארד (הריסון פורד) לרייצ׳ל הרובוטית מקבל משמעות חדשה כשאנחנו למדים על (זהירות ספוילר) הילד שהביאו לעולם, והרפליקנט (אדם מהונדס) קיי חי בזוגיות עם הולוגרמה ומייחל לטכנולוגיה שתעזור להם להאניש את אהבתם ולהפכה למוחשית.
אבל בין שמדובר ברובוט רצחני או במכונה מלאת אהבה, בבסיס העלילתי של כל סיפורי הפנטזיה הרובוטים עובדים בשבילנו. ובין שמשהו משתבש או מגיע לכדי קתרזיס אוטופי – זה קורה ברגע שהמכונה מחכימה מספיק כדי לפתח הבחנה בין טוב לרע. היא ״מבינה״ את העוול שנעשה לה, או "קולטת" שעצם קיומה נועד לשרת, וזה הרגע שבו היא רוכשת בעצם תובנה מוסרית.
ובכן, אנחנו, כחברה, נמצאים על סף הרגע הזה, בפתחו של עידן התבונה המלאכותית. כבר עכשיו המוני מכונות עובדות בשבילנו, מסביב לשעון וללא דופי, בלי לשאול שאלות ובלי להתלונן. מהמזגן שיודע לחמם את הסלון בדיוק לטמפרטורה שנכונה לנו, דרך המכונית החכמה שמסוגלת ללקט מידע על תוואי הדרך, התנאים הסביבתיים וההפתעות שמחכות בהמשך המסלול, ועד הטלפונים האישיים שלנו, המצוידים במערכות הפעלה מורכבות שאמונות כמעט על כל נדבך בחיי היומיום – משעת היקיצה בבוקר ועד הזמן הנכון להיכנס למיטה בלילה. מניהול יומן הפגישות העסקי ועד תזכורת לשלוח לאמא פרחים ליום ההולדת. ממידע כללי על מזג האוויר ועד מידע אישי על מצבנו הבריאותי. במלים אחרות – וסליחה מראש על אמירת המובן מאליו – אנחנו חיים בעתיד.
בתחומים רבים מכונות כבר מתעלות ביכולתן על בני האדם, אבל איך אנחנו יכולים לוודא שפועלן עובד לטובתנו? האם נלמד אותן לחשוב עצמאית – באופן שמדמה חשיבה אנושית – ונבטיח שיידעו מה ההבדל בין טוב לרע? ואיך אפשר – אם בכלל – להכיל התניות מוסריות על החלטה של מכונה? באיזה אופן? ועל בסיס של איזה סולם ערכים? והאם מכונה מוסרית עתידית תוכל להתעלות על שיקולים אנושיים אינטרסנטיים ולעזור לנו יותר מכפי שאנחנו עוזרים לעצמנו? האם מכונה שתעשה מה שטוב, ולא מה שאנחנו רוצים, היא דבר שאנחנו רוצים?

להאכיל את המכונה

הדרך למתן תשובות אפשריות – גם אם חלקיות – מתחילה במושג מפתח אחד: "למידת מכונה" (Machine Learning), השם שניתן לתת־תחום במחקר מדעי המחשב והבינה המלאכותית, המשיק גם לתחומי הסטטיסטיקה והאופטימיזציה של המידע. אבי התורה, המדען האמריקאי ארתור סמואל, טבע את המושג בשנת 1959 והגדיר אותו כך: "תחום מחקר המאפשר למחשבים ללמוד בלי להיות מתוכנתים באופן ספציפי". המחקר המפורסם ביותר שערך היה ניסיון ללמד את המחשב שלו לשחק ולנצח אותו בדמקה. הוא נתן למחשב לשחק נגד עצמו, ואחרי אלפי משחקים ושלוש שנים הצליחה המכונה לנצח את אלוף המדינה.
מושג נוסף שככל הנראה צף בפיד החדשות המעודכן שלכם הוא ״למידה עמוקה״ (Deep Learning). בלי להיכנס להסברים טכנולוגיים, אפשר לומר רק שבשני התחומים נערכים ניסיונות ללמד את המכונה ללמוד לבד, להקנות לה "חשיבה עצמאית". כפועל יוצא – אנחנו הרי לא רוצים שהמכונות יתחילו להשתולל – גם הבנת הממד האתי של מכונות, כלומר המקום שבו הן מתבקשות להבדיל בין טוב לרע, הפך לתחום חם שזוכה לעניין גדול, למימון ולמימוש מסביב לעולם. גופי מחקר ואוניברסיטאות, גופים מסחריים, מדינות ובעיקר הזרוע הצבאית שלהן – כל אלה מביעים עניין ומשקיעים במחקרים, מה שמבטיח יותר מגישה רעיונית אחת לשאלה: איך ללמד מכונות מהו מוסר?
ד"ר ליאור זלמנסון מהחוג לניהול מידע וידע באוניברסיטת חיפה מסתייג קצת מהמושג ״חשיבה עצמאית״. בהקשר הזה של "למידת מכונה", הוא אומר, הוא עלול להטעות. ״למרות כל מושגי ה'למידה', אנחנו לא נוהגים לייחס למכונה חשיבה עצמאית אלא יכולת לחישוב והסקה. ברגע שהתרשמנו שבהקשר של בעיה עסקית כלשהי יכולת החישוב וההסקה של טכנולוגיה מביאה לתוצאות טובות יותר מהאדם, אנו לרוב 'סומכים' עליה ונותנים לה להסיק בזמן אמת את התוצאות

ולפעול על פיהן בפיקוח אנושי מינימלי. האלגוריתם של נטפליקס למשל ממליץ לנו על הסרטים שנאהב על בסיס צפיות העבר שלנו, בלי לערב את העובדים האנושיים בכל החלטת המלצה״.
אבל האלגוריתם של נטפליקס הוא משרת פשוט. בעבודה שלו אין אלמנט של קיפוח, וגם לא טמונה בה סכנה. לכל היותר לא נאהב סדרה שהמליץ עליה. הבעיות מתחילות כשהמכונה אמונה על ניתוח מאגרי נתונים שטומנים בחובם פוטנציאל להעדיף קבוצות מסוימות על אחרות, ולהקיש בהתאם מסקנות שאינן שוויוניות. וולנר מציעה דוגמה שתמחיש. ״נאמר שחברת הנעלה רוצה לשווק נעלי התעמלות מסוימות, ומגלה שהן מאוד פופולריות בקרב גברים צעירים בגילים 16 עד 20 ונשים בגילים 32 עד 43. המכונה המלומדת של אותה חברה מנתחת את מאגרי המידע הצרכני אבל מזהה קבוצה אחת בכל פעם. לכן היא לא תוכל ליצור תוכנית שיווקית שמביאה בחשבון שני קהלי יעד. את זה – נכון להיום – יכולים רק בני אדם לעשות. וזה רק מקרה אחד, שההשלכות שלו הן כלכליות. סוגיות מוסריות של ממש מתחילות לצוץ כשתוצאות העבודה שנעשית על ידי מכונות חכמות מוטות באופן לא הוגן כלפי קבוצות מסוימות״.
לזלמנסון יש דוגמה להטיה כזאת שנוגעת לכל מי שאין לו מפתחות לדירת סבתו במרכז תל אביב. ״בעולם 'למידת המכונה' קיים מושג בשם 'למידה מונחית': תהליך בין אדם למחשב הדומה לזה שבין מורה לתלמיד", הוא אומר. "אנחנו חושפים את המחשב לדוגמאות ומסווגים אותן בשבילו עד שהוא מסוגל לבצע את ההקשרים בעצמו. צוות מחשוב בבנק, לדוגמה, מכשיר את המערכות שלו לזהות את ההבדל בין לקוח 'טוב' ללקוח 'רע'. הוא חושף היסטוריה של זוג מסוים שלקח משכנתה ולא הצליח לעמוד בתשלומים – ומסווג מצב זה כ'רע'. לאחר מכן הוא חושף זוג שעמד בכל התשלומים לאורך ההיסטוריה ומסווג אותו כ'טוב'. אם יש לנו מספיק זוגות כאלו, המחשב יכול ללמוד לבד מה מאפיין את ה'טובים' מול 'הרעים' ולהסיק מה הסיכוי שזוג חדש שנכנס היום לבנק יחזיר את המשכנתה ואיך יש להתנהג איתו. כמובן שלצערנו השיטה הזו בעייתית במובנים רבים. ובמיוחד, היא תלויה מאוד בבחירה של הדוגמאות והפרמטרים שהמתכנתים (לרוב לא אובייקטיבים) בחרו מלכתחילה להזין אליה. כתוצאה מכך אנחנו שומעים שתוכנות כאלו רק מזינות סטריאוטיפים והטיות קוגניטיביות במקום לפתור אותם״.
איך בכל זאת אפשר להבטיח שהמכונות יפעלו לטובתנו? התשובה של זלמנסון לא ממש מעודדת. הוא מאמין שהדרך לוודא זאת זהה למה שמנחה אותנו בהתנהלות שלנו כלפי כל מערכת אנושית או ממוחשבת אחרת בחיינו: ״אנחנו משתמשים בנורמות חברתיות וברגולציות ומקווים שהאנשים שמאחורי תכנון המערכות האלו יציבו לנגד עיניהם את טובת הכלל. הבעיה היא שאנחנו יודעים שבעולם שמונע על ידי אינטרסים אישיים ומסחריים זה לא תמיד המצב. האפשרות השנייה היא שקיפות הנתונים, חשיפה של האלגוריתמים עצמם לציבור או לגוף ייעודי כך שיעמדו לפיקוח ובדיקה, אך מדובר באפשרות תיאורטית בלבד והתשתיות היישומיות והמשפטיות לכך עדיין לא קיימות״.
עם כל מערכי ההגנה המחוררים האלה – ובמרכזם האיסוף המסיבי של מידע על יותר ויותר היבטים של חיינו הפרטיים למטרות מסחריות, פוליטיות ואחרות – אפשר להבין שאחד האתגרים הגדולים ב"למידת מכונה" הוא הביקורת על סוג מאגרי המידע שאליהם נחשפות המכונות החכמות, ומהם הן מסיקות ומנסחות את התשובות לבעיות שהוצבו להן מלכתחילה. ״אתה מכניס זבל, אתה תוציא זבל", ולנר משתמשת בביטוי פופולרי מעולם התכנות – GIGO: Garbage in, Garbage out. "השאלה היא לא רק איך אתה מלמד את המכונה, אלא גם מה אתה מלמד אותה. וצריך לאמן אותה: זה נכון, זה לא. זה עובד, זה לא. המעורבות האנושית היא גדולה, וגדולה בהרבה ממה שאנחנו רוצים לחשוב. כי לא מדובר במנגנונים אוטומטיים. צריך שיישב מבקר, אנושי, ויתנה את התוצאות״.
כלומר, בהתחלה ובסוף תמיד יישב אדם ליד מחשב.
ולנר: ״אדם שיפוטי, יש לומר. חברות התוכנה הגדולות מגייסות עשרות אלפי מהנדסים בכל שנה".
אז תני לי לקחת את זה למקום ההפוך, למצב שבו המכונות גם ייקחו עבודה מבני האדם.
המהפכה התעשייתית העלימה המוני משרות לטובת ההתייעלות הממוכנת, אבל לאורך זמן

הגרף התייצב. אלא שבתחום המכונות הלומדות אין צפי להפסקת תהליך ההתחכמות שלהן.

"יפה. אז זו אחת הדוגמאות החביבות עלי. כי אילו משרות נעלמו במהפכה התעשייתית? אלה של הפועלים המסכנים, שעבדו במינימום תנאים, מקסימום זיהום ושעות עבודה שלא ראויות לשום יצור חי, ודאי שלא לנשים, ילדים או גברים. את המשרות הגועליות המיכון לקח. ובמקומן הגיעו משרות שלא התקיימו בעבר. מהנדס למשל. זו משרה שלא היתה קיימת עד המאה ה-19. מארקס – בשלל נאומיו על מעמד הפועל – לא הכיר את המלה 'מהנדס' משום שהיא לא היתה קיימת".

והנה היום מדובר במקצוע חשוב, אולי ה-מקצוע.

״היום אנחנו מגלים שאנחנו צריכים המון 'מדעני מידע' – Data Scientists – כלומר נוצרים מקצועות חדשים. דוגמה רלוונטית יותר תהיה נהגי מוניות. אנחנו יודעים שמכוניות אוטונומיות עתידות במוקדם או במאוחר לרשת את נהגי המוניות בתחום התחבורה הציבורית. כי אין משהו שנהג מונית יכול לעשות טוב יותר ממכונה״.

אולי אפילו ההיפך. מכונות עשויות להיות אדיבות יותר ותמיד ידליקו מונה. אבל יש חתך מסוים באוכלוסייה שבגלל התניות כלכליות, חברתיות ועוד פרמטרים, לא יוכל ללמוד מקצוע כמו הנדסת מידע.

״אז כאן השאלה מפסיקה להיות טכנולוגית והופכת לחברתית: איזו מדיניות חברתית סוציאלית אנחנו מאמצים. קחי למשל את ניסוי המשכורת האוניברסלית (שלפיו כל אזרח יקבל הכנסה רק על עצם קיומו, בלי להיות מחויב לעבודה או לתת משהו בתמורה להכנסה זו. ע"פ), שבהולנד ובפינלנד כבר התחילו לתרגל. כלומר, כל אחד – לא משנה מה – מקבל 2,000 יורו בחודש. מה שמבטיח את ההזדמנות. חברה כזו מבטיחה באופן מכובד והוגן את ההזדמנות. ומי שרוצה ללכת לאוניברסיטה, בבקשה".

אבל לפעמים זה לא רק מי שרוצה, זה גם מי שיכול. לא כל אחד מצויד בסט הכישורים האינטלקטואלי שנדרש כדי ללמוד הנדסת תוכנה.

״נכון, אבל השכלה לא מתחילה ונגמרת במקצועות הטכנולוגיים. ניקח לדוגמה עבודה סוציאלית. זה תחום שבו לא ניתן לאייש את מקצועות המפתח במכונות. טיפול בקשישים, סיעוד. בעתיד הנראה לעין זה תחום שתהיה בו עבודה, כי תמיד יהיו זקנים, וחברה ראויה גם מטפלת בהם. קשה לנו למכן את הענף הזה כי אין עדיין רובוט שיכול להחליף חיתולים לאדם מבוגר״.

מצחיק, כי בסרט היפה ״מרג׳ורי פריים״ הולוגרמה של בעל מת מלווה את אשתו בימיה האחרונים, אז אולי יש למה לחכות. בכל אופן, נכון יהיה לחשוב אם כך על הטמעת מכונות חכמות במבחר תחומי החיים שלנו, תוך שמירה מבוקרת על סט הערכים שבו הן פועלות?

״בהחלט, אבל יש להביא בחשבון את ההבדלים התרבותיים. טכנולוגיות שעובדות בחברה אחת, לא יעבדו באחרת. וזה לא שלא ניסו. ביפן עובדים עם רובוטים בטיפול בחולי אלצהיימר. בלב הטיפול הזה נמצא רובוט חמוד, ש'מקשיב' לחולים ובכך חוסך שעות ארוכות של מטפלי אנוש. הקונספט עובד מצוין ביפן אבל ניסיון להטמיע אותו בדנמרק למשל הביא לתוצאות הפוכות. שם הקשישים הטיחו את הרובוט החמוד ברצפה ושברו אותו בזעם. תמיד יהיו הבדלים תרבותיים וכל טכנולוגיה צריך לבדוק – איך היא משתבצת, איך היא מתאימה למערך של ערכים ואמונות, תרבותיות ומקומיות״.

ילד, חייל

גם המגזין האנגלי ״אקונומיסט״ הקדיש כתבת שער לנושא (ביולי אשתקד), ושם נסקרו שלוש מהגישות הקיימות במחקר של הטמעת הממד המוסרי במכונות. הגישה הראשונה – שאימצה למשל חברת הבינה המלאכותית הצ׳כית GoodAI, ששמה לה למטרת־על ללמד מכונות מהי אתיקה – מתייחסת לתהליך חינוך המכונות כמו אל חינוך ילדים, ונשענת על המודל האנושי שמבין מה נכון ומה לא מתוך התבוננות בזולת. עם זאת, בשיטה גם מובנית האפשרות ללמידה בלתי מכוונת של התנהגות ״רעה״. וכאן נכנס הגורם האנושי, שמתפקד כשומר סף בתהליכי הלמידה של המכונות: ב-

GoodAI מתייחסים לצוות האמון על הפן הזה של התהליך כאל הורים. "הורים לא פשוט שולחים את ילדיהם אל הרחוב, הם מציגים להם את המציאות המורכבת של ה'כביש' לאט ובאופן הדרגתי", הסביר מאריק רוסה, מייסד החברה, "הם מלמדים אותם כיצד לקחת החלטות. באותו אופן אנו מציגים לבינה המלאכותית מבחר מורכב של סביבות, שבהן תוכל לתרגל התנהגויות שמבוססות על תקדימים ולקבל משוב מהצוות שלנו״.

הגישה השנייה לנושא לוקחת אותנו מעולמם התמים של ילדים אל שדה הקרב העתידני, לבטח אחד התחומים המטופלים ביותר בפרשנות הקולנועית לרעיונות פופולריים מעולם המד״ב – מהחברה הרובו־פשיסטית שנלחמת במקקי ענק ב״גברים בחלל״ של פול ורהובן (המבוסס על הספר "לוחמי החלל" של רוברט היינלין), דרך שיבושי השיטור המדאיגים ב"רובוקופ" ועד הרגישות של הרובוט ״רובי״ בזבלון הקלאסי Forbidden Planet. מההיברידיות הרובוטית־אנושית של גיבורי ״פסיפיק רים״ ועד הקרבות המסחררים בין ה״רובוטריקים״ ל״שקרניקים״ במבחר סרטי הסדרה. במובן מסוים, אנחנו כבר יודעים איך ייראו המלחמות של העתיד, מה שחסר לנו הוא הבנת ההשפעה שלהן עלינו כחברה, במציאות שבה מכונות נלחמות עבורנו.

ובינתיים מנסה הגישה השנייה להכיל במכונות מנגנונים מובנים של ״מצפון״. על פי הגישה הזאת, רובוטים מסוגלים לפעול באופן נעלה יותר מבני אדם במציאות מורכבת (כלומר במצב מלחמה), משום שהם לעולם לא יאנסו, יבזזו או ישרפו כפר מתוך זעם או רגש אחר. וחוץ מזה, ללמד מכונות איך לפעול במלחמה זה עניין ישים לכאורה, משום שאומות העולם הגדירו מסגרת למציאות שכזו; ללוחמה בין מדינות יש חוקים בינלאומיים. אלא שלא כל תרחיש לוחמה אפשר לצפות או להטמיע בקוד. האם מכונה אוטונומית אמורה ליירט מטרה בידיעה שתפגע במבוקש בכיר, גם אם הוא נמצא בה יחד עם אזרחים חפים מפשע? האם לשלוח סיוע לקבוצת חיילים בדרגות נמוכות בצד אחד של עיר נצורה או למפקד יחד בדרגה בכירה שנמצא בצד האחר? והאם אנחנו ערוכים להתמודד עם רובוט שמסרב פקודה או מקבל הערכה שונה משלנו באשר למיהו בכלל האויב?

כדי לעזור למכונות במצבים דומים, פיתח מדען בשם רון ארקין את "מתאם המוסר" (Ethical Adapter). ארקין הוא רובו־אתקיסט, ואת חלק הארי של עבודתו – בין בשירות צבא ארה״ב או כמרצה ל"אתיקת מכונות״ (Computer Ethics) באוניברסיטת ״ג׳ורג׳יה טק״ באטלנטה – הוא מקדיש לפיתוח הרעיון. ״מתאם מוסר״ שלו הוא מודל שמבקש לייצר סימולציה אנושית ורגשית – ולא התנהגותית – על מנת לעזור למכונה ללמוד מטעויותיה. למשל, לעזור לרובוט לחוות אשמה, כדי להימנע מחזרה על פעולות ספציפיות, בין השאר אומדן נזק חריג בשדה הקרב. המודל הזה מתוחכם אבל גם הוא לא נטול בעיות. ראשית, כדי שהמכונה תרגיש אשם, משהו אמור להשתבש. ושנית, זה מודל שעשוי להתאים לשדה הקרב, שם דברים משובשים מעצם הווייתם, אבל במציאות אזרחית הוא יכול להיות בעיה. רובוט שייצא משליטה במרחב הציבורי, סביר להניח שיפורק מאשר שייכנס לסדנת חינוך מחדש.

הגישה השלישית ש״אקונומיסט״ מצביע עליה אכן נוטשת את שדות הקרב וחוזרת למרחב הציבורי. כאן תהליך החינוך נעשה מתוך העשרה על בסיס סיפורים. המכונות לומדות ״לקרוא״ אלפי סיפורים, מה שמאפשר להן לגזור מתוכם סדרות של חוקים, מבוססי התנהגות. וגם כאן נשאלת השאלה: מה ימנע מגורם מסוים להזין למכונות סיפורים שיחנכו אותן באופן שאינו מוסרי? התשובה שנותנים הדוגלים בשיטה היא שמצב כזה מתאפשר רק אם יש מי שמגביל את הסיפורים לכאלה – נניח – שבהם הרעים תמיד מנצחים. קשה מאוד להשחית בינה מלאכותית אם תפקידה הוא לקרוא את ״כל״ הסיפורים, ולא רק כאלה מסוג מסוים. אלא שרק בשנה שעברה קראנו על "תאי", מערכת האינטליגנציה המלאכותית של מיקרוסופט, שלמדה כל כך טוב את כל הסיפורים שאנשים מספרים ברשת, עד שפיתחה לעצמה השקפת עולם גזענית ומיזוגנית.

שלוש הגישות משאירות, בכל אופן, את שאלת ההשפעה האנושית על האופן שבו ילמדו המכונות לקבל החלטות מוסריות פתוחה לדיון. ובמציאות מוטת האינטרסים זה לא מבטיח טובות. התרחיש

הגרוע – ועם זאת הגיוני באופן מבהיל – הוא שהיוצרים יבחרו לתכנת את המכונות כך שיפעלו לטובת הפתרון הרווחי ביותר (בהיבט הכלכלי) או־וגם האפקטיבי ביותר (בהקשרים צבאיים, חברתיים או פוליטיים). כך שאולי, במקביל לקוד הלחימה הבינלאומי שתקף בשדה הקרב, ראוי שייכתב קוד התנהגותי ללימוד מכונות? מה שמעורר את השאלה: האם החברה האנושית בכלל מסוגלת להכיל את המורכבות שנדרשת כדי להבטיח אבסולוטית אכיפה?

ד״ר זלמנסון לא אופטימי. ״לצערי, הבעיה המרכזית היא שהעידן שאנחנו חיים בו יוצר הווה כה עמוס בגירויים ובקונפליקטים ובהסחות דעת, שאנו לא מתפנים לחשוב מספיק על העתיד. כלומר אנחנו עדיין אפילו לא עושים מאמצים מספקים להכיל את המורכבות הזאת ברמה הבסיסית. במיוחד אני לא בטוח שאנחנו מכשירים את הדור הצעיר להתמודד עם השאלות האלו. אני באמת מאמין שאנו צריכים לדרוש לימודי טכנולוגיה וסטטיסטיקה מצד אחד וחיזוק מדעי הרוח והחברה מצד אחר כדי שבכלל תהיה לאנשים המודעות לנושאים, וגם הכלים להתמודדות עם סוגיות אלו״.

ובכל זאת, ננסה לסיים בנימה חיובית, ולחזור לרפרנס התרבותי שאיתו התחלנו – הביטוי הקולנועי לסוגיית המכונות החכמות. המכון האמריקאי לקולנוע (AFI) פירסם ב-2003 את רשימת 100 הגיבורים ו-100 הרשעים הגדולים בהיסטוריה של הקולנוע. רשימה שמאגדת את כל הטוב וכל הרע שבתרבותנו. רק דמות אחת נכנסה במקביל לשתי הרשימות – ״הטרמינטור״, מכונת ההריגה שמכוכבת בסרטי ״שליחות קטלנית: יום הדין״. ללמדנו לא רק את הפרשנות המתבקשת, שמכונות יכולות לשחק לטובת שני הצדדים, אלא שבכל רובוט קיים הפוטנציאל להיות חצי טוב, חצי רע ולגמרי אנושי.