

עלותן של החלטות דטרמיניסטיות המתקבלות על ידי רובוטים אוטונומיים

נתי פרל, הפקולטה למשפטים באוניברסיטת חיפה, אוגוסט 2014

הוגש במסגרת סמינר מחקר: שלטון החוק ומצבי קיצון*

1. מבוא 2
2. מושגי יסוד 3
3. יתרונות השימוש ב-FAR בשדה הקרב 5
4. אופן קבלת החלטות על ידי FAR 8
 - רכישת אינפורמציה 10
 - ניתוח האינפורמציה 12
 - קבלת החלטה 14
5. החלת משפט ארו על מנגנון קבלת ההחלטות 17
 - צפיית הקריטריון הדיקטטורי וההחלטות השגויות 18
6. שלטון החוק והטלת אחריותיות על FAR 23
 - באגים – סיכונים ועלויות 25
 - אחריותיות להחלטות FAR במקרים דיקטטוריים 27
7. סיכום 28
8. ביבליוגרפיה 30
9. סימונים 38

*אני מבקש להודות לפרופ' עלי זלצברגר על הערותיו המאירות.

מבוא

ההתפתחויות בתחומי הבינה המלאכותית והרובוטיקה מאפשרות אוטומטיזציה של מגוון פעולות אנושיות, בכללן פעולות הדורשות יכולות קוגניטיביות-אוטונומיות כגון סתגלנות לסביבה לא מוכרת וקבלת החלטות בזמן אמת. בשנים האחרונות אנו עדים למגמה גוברת והולכת של הענקת אוטונומיה לרובוטים, והשימוש בהם מתרחב ליותר ויותר תחומים, ביבשה, באוויר, בים ובחלל.¹ הענקת אוטונומיה לרובוטים בביצוע פעולות כרוכה בד-בבד בהופעתן של שאלות מוסריות ומשפטיות חדשות אודות האחריות על רובוטים אלו. שאלות אלו מקבלות משנה תוקף כאשר עסקינן ברובוטים שמיועדים לפעול בשדה הקרב. לפיכך, קיים דיון נרחב אודות לגיטימיות והסדרת השימוש ברובוטים אוטונומיים במצבי מלחמה.²

נוכח הדיון הנרחב, גוברת חשיבות התווית גבולות להטלת אחריות על רובוטים אוטונומיים. ראשית, התווית הגבולות תהווה תמריץ לייצור רובוטים העומדים בסטנדרטי בטיחות ויעילות גבוהים יותר. שנית, קיומם של גבולות מספק מסגרת נורמטיבית לביקורת שיפוטית הן בשלב התכנון והפיתוח של הרובוטים (Ex-Ante) והן לאחר מעשה, במקרים שנגרם על ידם נזק (Ex-Post).³ התווית הגבולות מחייבת הבנה של הטכנולוגיה ומרכיביה, לצורך התמודדות עם האתגרים הייחודיים שהטכנולוגיה החדשה טומנת בחובה.

תמציתו של הטיעון שיוצע ברשימה זו הינה יישום משפט אי האפשרות של ארו (Arrow)⁴ על מנגנוני קבלת החלטות של רובוטים אוטונומיים, כאשר מושא הטיעון יתמקד בעיקר ברובוטים אוטונומיים שנועדו לפעול בשדה הקרב. ליישום זה צפויות להיות השלכות משמעותיות על תכנון ופיתוח רובוטים אלו, כמו גם על המכשירים המשפטיים שיסדירו את השימוש בהם.

בחלק הראשון אציג מושגי יסוד מעולם הרובוטיקה, לצורך הבהרת הטרימינולוגיה בה ייעשה שימוש במסגרת רשימה זו. לאחר מכן, אעמוד על היתרונות שיש לרובוטים אוטונומיים ביחס ללוחמים

¹ *S. Navy Sends Underwater Sonar Robot in Search for Missing Malaysian Airliner*, USNI NEWS (March 24, 2014). <http://news.usni.org/2014/03/24/u-s-navy-sends-underwater-sonar-robot-search-missing-malaysian-airliner>; Rob Ambrose, et al. *Robotics, Tele-Robotics and Autonomous Systems Roadmap: Technology Area 04*, NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (2012). Available at: http://www.nasa.gov/pdf/501622main_TA04-Robotics-DRAFT-Nov2010-A.pdf; Patrick Lin, et al., *Autonomous Military Robotics: Risk Ethics and Design* 12-19 (2008). Available at: <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA534697> (hereinafter: *Autonomous Military Robotics*)

² ICRC, *Expert Meeting on Autonomous weapon systems: technical, military, legal and humanitarian aspects*, March 2014. Available at: <http://www.icrc.org/eng/assets/files/2014/expert-meeting-autonomous-weapons-icrc-report-2014-05-09.pdf> (hereinafter: *ICRC, Expert Meeting 2014*)

Kenneth Anderson, et al., *Adapting the Law of Armed Conflict to Autonomous Weapon Systems*, INTERNATIONAL LAW STUDIES UNITED STATES NAVAL WAR COLLEGE, FORTHCOMING 2014 (2014)

ICRC, *Autonomous weapons: States must address major humanitarian, ethical challenges*, September 2013. Available at: <http://www.icrc.org/eng/resources/documents/faq/q-and-a-autonomous-weapons.htm>; גבי סיבוני ויוני אשפר "דילמות בהפעלת אמצעי לחימה אוטונומיים", 16 *עזבן אסטרטגי*, 71 (2014)

³ Helen Nissenbaum, *Accountability in a Computerized Society*, 2 SCIENCE AND ENGINEERING ETHICS 25 (1996) (hereinafter: *Nissenbaum*)

⁴ Kenneth J. Arrow, *A Difficulty in the Concept of Social Welfare*, 58 THE JOURNAL OF POLITICAL ECONOMY 328 (1950) (hereinafter: *Arrow*)

Kenneth J. Arrow, *Social choice and individual values*. 12 (Yale university press, 2012)

אנושיים בשדה הקרב, ואנמק מדוע ישנו בסיס סביר לחשוב שהשימוש בהם יהיה נרחב. לצד זאת, אסקור בקצרה מספר אתגרים פילוסופיים ומשפטיים הכרוכים בתכנון ותכנות רובוטים אלו. בפרק הרביעי, אנתח את אופן קבלת החלטות של רובוטים אוטונומיים, תוך התמקדות בקטגוריית מנגנוני קבלת החלטות שעומדים בתנאי משפט אי האפשרות של ארו. בפרק החמישי אבחן את נפקויות החלת משפט אי האפשרות של ארו על עולם הרובוטים האוטונומיים, ואבחן קטגוריה חדשה של שגיאות מחשב (Bugs) בעולם זה. בפרק השישי, אציג את תרומתה של אבחנה זו להתווית גבולות להטלת אחריותיות על שגיאות הנכללות בקטגוריה זו. הצגה זו תעשה בראי האתגרים הניצבים כיום בפני שלטון החוק בהטלת אחריותיות על שגיאות מחשב.

מושגי יסוד

רשימה זו תתמקד במנגנוני קבלת החלטות לפיהם פועלים רובוטים אוטונומיים לחלוטין (Fully Autonomous Robot, להלן: FAR). מכיוון שקיים מגוון רחב של טכנולוגיות מתקדמות ופעמים נוצר בלבול בין המושגים השונים, נגדיר את ה-FAR מושא הרשימה תוך הבהרת הטרמינולוגיה.

רובוט הוא מכונה, לעיתים בעלת יכולת ניווד, המסוגלת לבצע פעולות מורכבות יחסית באופן אוטומטי, אשר ניתנת לשליטה על ידי בקר ממוחשב.⁵ משמעות האוטומטיית הינה יכולת לתפקד ללא התערבות אנושית, אך לא בהכרח יכולת קבלת החלטות.⁶ מערכות אוטומטיות גרידא אינן מנחות את עצמן מה ואיך לעשות, אלא מבצעות פעולות בהנחיית מפעיל אנושי. גורם אנושי מעורב בתפעול הרובוט, מהנקודה שהופעל ועד לביצוע הפעולה.⁷ כך למשל, רוב המל"טים שנמצאים בשימוש היום נשלטים מרחוק על ידי מפעיל אנושי. המל"ט לא חושב, מחליט או פועל לבד, אלא מוציא לפועל את החלטותיו של המפעיל.⁸ בשנים האחרונות, נכנסים לשוק גם רובוטים מתוחכמים יותר שמסוגלים לבצע פעולות מורכבות אשר בעבר חייבו מעורבות אנושית, דוגמת המל"ט X-47B. מל"ט זה מסוגל להמריא ולנחות באופן עצמאי, אולם ההחלטה מתי לנחות, להמריא או לתקוף מטרה עדיין מחייבת התערבות אנושית.⁹ כך גם הרובוט SGR-A1, רובוט שנועד לפעול כ"שומר גבול" בגבול שבין דרום לצפון קוריאה, מצויד בכלי נשק קטלניים אשר ניתנים להפעלה על ידי

⁵ "robot" Oxford English Dictionaries, Oxford University Press. <http://www.oxforddictionaries.com/definition/english/robot>; P. Moubarak & P. Ben-Tzvi, *Adaptive Manipulation of a Hybrid Mechanism Mobile Robot 113-118*, IEEE (2011)

⁶ O. Grant, Clark R. Kok and R. Lacroix., *Mind and Autonomy in Engineered Biosystems*, 12 ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE 389, 389-399 (1999)

⁷ Shane Harris, *Out of the Loop: The Human-free Future of Unmanned Aerial Vehicles*, in EMERGING THREATS IN NATIONAL SECURITY AND LAW (Peter Berkowitz ed., 2012). Markus Wagner, *Taking Humans Out of the Loop: Implications for International Humanitarian Law*, 21 J.L. INFO. & SCI. 1 (2011)

⁸ William C Marra and Sonia K. McNeil, *Understanding "The Loop": Regulating the Next Generation of War Machines*, 36 HARV. J. L. & PUB. POL'Y 1139, 1141 (2012) (hereinafter: *Understanding "The Loop"*)

⁹ W.J. Hennigan, *New drone has no pilot anywhere, so who's accountable?*, LOS ANGELES TIMES (January 26, 2012). <http://articles.latimes.com/2012/jan/26/business/la-fi-auto-drone-20120126>

מפעיל אנושי במידה והרובוט מגלה פעילות חשודה.¹⁰ מגמה זו מהווה צעד בדרך למתן אוטונומיה מלאה לרובוטים, כך שיוכלו לבצע משימות מתחילתן ועד סופן ללא צורך בהתערבות אנושית.¹¹

המושג אוטונומיה קשה להגדרה אפילו כשעסקינן בבני אדם.¹² לצורך רשימה זו, אין כוונתי להשתמש במושג באופן המבטא סוכנות מוסרית (Moral Agency) או רצון חופשי (Willkür), כפי שעושה למשל קאנט,¹³ אלא להתמקד במובן המתייחס ליכולת לפעול בעולם האמיתי ללא שליטה חיצונית.¹⁴ כך למשל, מערכת נשק אוטונומית היא מערכת אשר מרגע שהופעלה יכולה לבחור ולטפל במטרה ללא התערבות אנושית.¹⁵ מכאן, אוטונומיה היא פונקציה שתלויה בשלושה משתנים: אי-תלות (Independence), סתגנות (Adaptability) ושיקול דעת בקבלת החלטה (Discretion). רובוט יוגדר כאוטונומי אם הוא יכול לפעול ללא אינטראקציה עם מפעיל אנושי, לפעול בהצלחה בסביבה לא מוכרת, ולהשיג את המטרות שהוגדרו לו מראש תוך הפעלת שיקול דעת.¹⁶ כיום, הרובוטים המצויים בשימוש נעים על סקאלה של אוטונומיות כפי שהוגדרה לעיל, והמגמה היא להעניק להם יותר אוטונומיות. רשימה זו תעסוק ברובוטים אוטונומיים שהוקנה להם שיקול דעת מלא בקבלת החלטה בזמן אמת. כלומר, מערכת קבלת ההחלטות (Decision Making System) שהוגדרה עבור הרובוט היא אוטונומית, ומבוססת על אלגוריתם קבוע מראש שמיושם בזמן אמת, ללא התערבות אנושית.¹⁷

המשך הדיון יסוב בעיקר סביב FAR הפועלים בשדה הקרב, שפעמים מכונים Killer Robots,¹⁸ בהם מנגנוני קבלת ההחלטות שמרכיבים את האלגוריתם ב"פונקציות הקריטיות" כוללים מעקב ובחירת מטרה לתקיפה.¹⁹ קטגוריית FAR אלו נבחרה להיות מושא הטיעון המוצע משלוש סיבות. ראשית, זהו התחום בו ניכרת ההתפתחות המשמעותית ביותר לכיוון הרחבת האוטונומיה של רובוטים.²⁰ שנית, הסיבות לכדאיות השימוש ב-FAR בשדה הקרב חלות גם על FAR בתחומים

Lewis Page, *South Korea to field gun-cam robots on DMZ*, THE REGISTER (MARCH 14, 2007).¹⁰ http://www.theregister.co.uk/2007/03/14/south_korean_gun_bots;
Human Rights Watch and International Human Rights Clinic of the Human Rights Program at Harvard Law School, *Losing Humanity: The Case Against Killer Robots*, 14 (2012). Available at: http://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf

11 "The Loop" Understanding, לעיל ה"ש 8, בעמ' 1141-42

12 GERALD DWORIN, THE THEORY AND PRACTICE OF AUTONOMY 7 (Cambridge University Press.,1988)

13 Immanuel Kant, *Metaphysische Anfangsgründe der Rechtslehre*, in 6 KANTS GESAMMELTE SCHRIFTEN 230 (Preußische Akademie der Wissenschaften ed., 1902–1923)

14 Autonomous Military Robotics, לעיל ה"ש 1, בעמ' 4

15 U.S. DEP'T OF DEF., DIR. 3000.09, AUTONOMY IN WEAPON SYSTEMS p. 13 (November 21, 2012). Available at: <http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf>

16 "The Loop" Understanding, לעיל ה"ש 4, בעמ' 1155; Walter F. Truskowski et al., AUTONOMOUS AND

17 AUTONOMIC SYSTEMS: WITH APPLICATIONS TO NASA INTELLIGENT SPACECRAFT OPERATIONS AND EXPLORATION SYSTEM 10 (2009)

18 William Boothby, *Some Legal Challenges Posed by Remote Attack*, 94 INTERNATIONAL REV. OF THE RED CROSS 579, 584 (2012)

19 Human Rights Watch and International Human Rights Clinic of the Human Rights Program at Harvard Law School, *Losing Humanity: The Case Against Killer Robots* (2012). Available at: www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf; 'Killer robots' to be debated at UN, BBC NEWS (May 9, 2014) <http://www.bbc.com/news/technology-27343076>; 'Killer robots': MP Nia Griffith calls for world ban, BBC NEWS (June 16, 2013) <http://www.bbc.com/news/uk-wales-22927092>

20 ICRC, Expert Meeting 2014, לעיל ה"ש 2, בעמ' 1

R. Sparrow, *Building a Better WarBot: Ethical Issues in the Design of Unmanned Systems for Military Applications*, 15 SCIENCE & ENGINEERING ETHICS 169, 171 (2009)

אחרים, כך שהטיעון המוצע תקף בפרט אודותיהם. שלישית, זהו התחום השנוי ביותר במחלוקת שכן במקרים אלו ההחלטות המתקבלות הן המשמעותיות ביותר, הואיל וה-FAR מתוכנת לגרום נזק ולא רק למונעו. להבדיל מרובוט המתוכנת לפינוי נפגעים משריפה או רעידת אדמה,²¹ FAR בשדה הקרב יתוכנת לפגוע במטרות. גרימת נזק מחייבת פיקוח משמעותי, ולכן קיים דיון נרחב בגופים בינלאומיים אודות דרכי ההתמודדות עם ההתפתחויות הטכנולוגיות המקנות יותר ויותר אוטונומיה לרובוטים.²² רשימה זו נועדה להשתלב בדיון זה, ולתרום לשיח האקדמי והמדיני על ידי אבחון קטגוריה של מנגנוני קבלת החלטות הדורשת התייחסות משפטית וטכנולוגית מיוחדת, במידה והשימוש ב-FAR בשדה הקרב יותר.

3. יתרונות השימוש ב-FAR בשדה הקרב

היתרון הראשון בהעברת ביצוע פעולות מסוכנות ל-FAR הוא מניעת הנזק לבני אדם שמבצעים כיום את הפעולות האלו. בנוסף, FAR יכולים להקריב את עצמם למען השגת המטרה, כך שהנזק שייגרם יהיה כלכלי לכל היותר ולא יהיה כרוך בפגיעה בחיי אדם.²³

לצד זאת, למערכת ממוחשבת ישנן יכולות עיבוד וניתוח נתונים טובות יותר בזמן אמת, כך שזמן התגובה שלה קצר יותר.²⁴ יכולות אלו חשובות בפרט בשדה הקרב, המתאפיין בהיותו סביבה מורכבת ומרובת משתנים. בסביבה זו מהירות התגובה והחישוב הגבוהה של הרובוט, כמו גם הדיוק הרב יותר בביצוע הפעולות עצמן, יהוו יתרון משמעותי.²⁵ כפועל יוצא מיכולות אלו, FAR יכול לקבל החלטות העומדות בדרישות המוסריות והמשפטיות המוחלטות כיום על לוחמים, בצורה טובה יותר מלוחמים אנושיים. למשל, FAR מסוגל לשקלל את המשתנים בצורה טובה יותר, כך שתתקבל החלטה מדויקת יותר על הפעולה הפרופורציונאלית,²⁶ ולהוציא לפועל את הפעולה הנדרשת תוך עמידה בעיקרון ההבחנה²⁷ ובעיקרון כפל התוצאות.²⁸ בהיבט זה של תפקוד ה-FAR, ישנה מידה של

²¹ *US Marines perfecting autonomous evacuation and supply vehicle*, RT (July 29, 2014) <http://rt.com/usa/176252-marines-self-driving-car/>

²² לעיל הי"ש 2

²³ Ronald C. Arkin, *Governing Lethal Behavior: Embedding Ethics in A Hybrid Deliberative/Reactive Robot Architecture* 6 (2011). (hereinafter: *Governing Lethal Behavior*) Available at: <http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf>

²⁴ DARPA Advances Video Analysis Tools, DARPA (June 23, 2011) http://www.darpa.mil/NewsEvents/Releases/2011/06/23_DARPA_advances_video_analysis_tools.aspx; Unified Military Intelligence Picture Helping to Dispel the Fog of War, DARPA (September 5, 2013). <http://www.darpa.mil/NewsEvents/Releases/2013/09/05.aspx>

²⁵ Thomas K. Adams, *Future Warfare and the Decline of Human Decisionmaking*, PARAMETERS 57, 57-58 (2001)

²⁶ Chairman of the Joint Chiefs of Staff Instruction, CJCSI 3160.01, D-A-7 (Feb. 13, 2009); Defense Department General Counsel, Joint Targeting Cycle and Collateral Damage Estimate Methodology (CDM), (November 10, 2009)

²⁷ מיכאל וולצר *מלחמות צודקות ובלתי צודקות* 57-58 (ספריית אופקים, עם עובד, 1977) Thomas Hurka, *Proportionality in the Morality of War*, 33 PHILOSOPHY AND PUBLIC AFFAIRS 34, 36 (2005)

²⁸ מיכאל וולצר *מלחמות צודקות ובלתי צודקות* 181-184 (ספריית אופקים, עם עובד, 1977) AN ENCYCLOPEDIA OF WAR AND ETHICS, 258 (D. Wells Ed., Greenwood Press, 1996)

אירוניה בטענה הגורסת כי גורם אנושי יהיה תמיד נחוץ בפקוח וניטור המערכות.²⁹ מכיוון שמערכת ממוחשבת משקללת יותר מידע בפחות זמן, אם יוטל על בן אנוש לפקח על החלטות המערכת בזמן אמת ולבחון האם מתקבלת ההחלטה האופטימאלית, זוהי תהיה משימה בלתי אפשרית. כלומר, אם אדם יוכל לנטר את המערכת הממוחשבת, בהגדרה, המערכת לא פועלת בצורה אופטימאלית.³⁰

התפקוד של FAR בשדה הקרב לא יושפע משני גורמים מרכזיים שעשויים להשפיע לרעה על תפקוד אנושי. הגורם הראשון הוא רגשות המשפיעים לרעה על שיקול הדעת של הלוחם, והשני הוא הטיות פסיכולוגיות-קוגניטיביות המונעות שיקול רציונאלי לגופה של החלטה. אמנם, ל-FAR אין רגשות חמלה ורחמים המספקים מנגנון ביקורת חשוב בכל הנוגע לפגיעה באזרחים.³¹ אולם, רגשות אלו גם מקשים על ביצוע פעולות קשות ונחוצות. יתר על כן, בעת מלחמה סביר יותר כי רגשות שליליים ישפיעו על הלוחם האנושי מאשר רגשות חיוביים. מצב טעון רגשית, דוגמת מלחמה, מוביל לשחרור האינסטינקטים האנושיים הבסיסיים כגון דחף ההישרדות, רגשות שנאה וגזענות. רואנדה, הבלקן, דרפור ואפגניסטן הן רק חלק מהדוגמאות למקרים בהם רגשות לא מרוסנים הובילו לפעולות נוראיות.³² יתרה מכך, בסביבת לוחמה הפרט צפוי להרגיש פחד והיסטריה אשר לוחצים עליו לאמוד את המצב מתוך בהלה ולפעול באופן פושע.³³

מעבר להשפעת הרגשות, יכולתם של בני אנוש לקבל החלטות נפגמת בתנאי לחץ ומצוקה. בתנאים אלו אנשים פועלים בפזיזות, תוך סינון והזנחת מידע.³⁴ הטיה פסיכולוגית זו, הידועה בשם Scenario Fulfillment, גורמת לעיוות או הזנחה של אינפורמציה במצבי לחץ בשל העובדה שבמצבים אלו בני אדם מעבדים מידע חדש בעיקר באופן שמתאים לתבניות של מידע מוקדם שיש להם.³⁵ מעצם הגדרתו, ל-FAR אין רגשות והוא אינו כבול להטיות קוגניטיביות, כך שגורמים אלו לא ישפיעו על החלטותיו. יתר על כן, כך על פי הטענה, נוכחותם של FAR בסביבת לוחמים אנושיים תגרור התנהגות אתית יותר של שני הצדדים. הידיעה של החייל כי ה-FAR יכול לחשב ולהשוות

²⁹ ראו למשל פסיקת בית המשפט בארה"ב בעניין טייס שהעביר את השליטה במטוס לטייס אוטומטי (Autopilot) שגרם להתנגשות - "The obligation of those in charge of a plane under robot control to keep a proper and constant lookout is unavoidable". Brouse v. U.S., 83 F. Supp. 373, 374 (N.D. Ohio 1949)

³⁰ Lisanne Bainbridge, *Ironies of Automation*, 19 AUTOMATICA 775, 775-77 (1983)
³¹ Human Rights Watch and International Human Rights Clinic of the Human Rights Program at Harvard Law School, *Losing Humanity: The Case Against Killer Robots* (2012). Available at: www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf

³² Michael N. Schmitt, *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics*, 13 HARVARD NATIONAL SECURITY JOURNAL FEATURE (2013)
³³ M. WALZER, JUST AND UNJUST WARS 251 (4th Ed. Basic Books, 1977)

³⁴ Anne Edland, *Attractiveness Judgments of Decision Alternatives Under Time Stress*, 21 COGNITION AND DECISION RESEARCH UNIT (1985); Adedeji B. Badiru and Lee Ann Racz, HANDBOOK OF EMERGENCY RESPONSE: A HUMAN FACTORS AND SYSTEMS ENGINEERING APPROACH (CRC Press., 2013)

³⁵ Scott D. Sagan, *Rules of Engagement*, in AVOIDING ; 6-7 בעמ' 23, לעיל ה"ש 23, Governing Lethal Behavior WAR: PROBLEMS OF CRISIS MANAGEMENT 443, 459-61 (Alexander L. George eds., 1991)

את הפעולה שננקטה על ידי לוחם אנושי לפעולה שראויה היתה להינקט, כמו גם יכולותיו של ה-FAR לתעד את פעולות החיילים בזמן אמת, יובילו לריסון פעולותיהם של שני הצדדים בקרב.³⁶ לבסוף, השימוש ברובוטים אוטונומיים בשדה הקרב והמחקר אודותיהם בלתי נמנעים. ראשית, הפיתוחים הטכנולוגיים לא נוגעים רק לשדה הקרב, אלא מהווים חלק ממגמה של התקדמות טכנולוגית לעבר רובוטים אוטונומיים במגוון רחב של תחומים. בשוק כבר מוצעים למכירה FAR המשמשים לניקיון הבית,³⁷ מסייעים למוגבלי יכולת בביצוע פעולות יומיומיות,³⁸ מבצעים את שירות החדרים במלונות,³⁹ ורובוטים הומנואידים המשמשים, למשל, כמלצרים במסעדה.⁴⁰ מכונות אוטונומיות של חברות כגון גוגל,⁴¹ טויוטה,⁴² ומרצדס⁴³ עושות את דרכן אל הכבישים ברחבי העולם, אל שוק ההון חודרות אט-אט תוכנות פיננסיות שמגיבות במהירות ובאופן אוטונומי לשינויים,⁴⁴ ובבתי כלא במדינות מסוימות מסתמנת מגמה של החלפת הסוהרים האנושיים ב-FAR.⁴⁵

שנית, השימוש במערכות נשק אוטונומיות נובע מלחץ צבאי ופוליטי להגן על חיי האזרחים ורכושם.⁴⁶ בעידן המודרני מתנהל מירוץ חימוש של טכנולוגיות צבאיות.⁴⁷ בשנים האחרונות חלק משמעותי במירוץ נסוב סביב השגת טכנולוגיה שתאפשר שימוש במערכות נשק אוטונומיות יותר ויותר בשדה הקרב הפיזי,⁴⁸ כמו גם בעולם הסייבר.⁴⁹ כך למשל, ארצות הברית מצהירה במהלך השנים האחרונות על רצונה לפתח מל"טים שיוכלו לפעול בצורה אוטונומית, ויתוכנתו לקבל החלטות בזמן אמת על פי כללים מוגדרים מראש וללא התערבות אנושית.⁵⁰ גם במישור התקציבי,

Anne Edland, *International Governance of Autonomous Military Robots*, 12 COLUM. SCI. & TECH L. REV. 272, 280 (2011)

See 'iRobot' at: www.irobot.com/us

See 'Human Support Robot (HSR)' at: www.toyota-lobal.com/innovation/partner_robot/family_2.html

דני שדה "רובאטר: הרובוט שישרת אתכם במלון" *Ynet*, 14.8.2014

www.ynet.co.il/articles/0,7340,L-4558761,00.html

Honda Unveils All-new ASIMO with Significant Advancements, HONDA (November 8, 2011).

<http://world.honda.com/news/2011/c1111108All-new-ASIMO>

John Markoff, *Google Cars Drive Themselves, in Traffic*, THE NEW YORK TIMES (October 9, 2010).

גיינר מוקדם מהצפוי: "Ynet" *Ynet*, 22.5.2014; www.nytimes.com/2010/10/10/science/10google.html?pagewanted=all

www.ynet.co.il/articles/0,7340,L-4521319,00.html

Toyota sneak previews self-drive car ahead of tech show, BBC NEWS TECHNOLOGY (January 4, 2013).

<http://www.bbc.com/news/technology-20910769>

Matthew de Paula, *Autonomous driving tech package will be an option on Mercedes vehicles by 2020*,

FORBES (September 9, 2013) <http://www.forbes.com/sites/matthewdepaula/2013/09/30/autonomous-driving-will-become-an-option-on-regular-mercedes-models-by-2020>

Kenneth Anderson and Matthew Waxman, *Law and Ethics for Autonomous Weapon System: Why a Ban Won't Work and How the Laws of War Can*, HOOVER INST. MONOGRAPH, 2 (2013) (hereinafter: *Law and Ethics for Autonomous Weapon System*)

Lena Kim, *Meet South Korea's New Robotic Prison Guards*, DIGITAL TRENDS (April 21, 2012).

www.digitaltrends.com/cool-tech/meet-south-korea-s-new-robotic-prison-guards/#!buCxB

Law and Ethics for Autonomous Weapon System 44, בעמ' 2

ICRC, Expert Meeting 2014, לעיל ה"ש 2, בעמ' 1

Autonomous Military Robotics 82, לעיל ה"ש 1, בעמ' 1

PETER W. SINGER, WIRED FOR WAR: THE ROBOTICS REVOLUTION AND CONFLICT IN THE TWENTY-FIRST CENTURY 125 (Penguin press, 2009)

עמי רוחקס דומבה "סנאודן: פותח נשק סייבר אוטונומי" *nrg*, 12.8.2014

www.nrg.co.il/online/1/ART2/609/821.html?hp=1&cat=324&loc=18

U.S. DEP'T OF DEF., 20301-3140, THE RULE OF AUTOMATION IN DoD SYSTEMS (July 2012). Available at: <http://www.acq.osd.mil/dsb/reports/AutonomyReport.pdf>; U.S. DEP'T OF AIR FORCE, UNITED STATES

לפיתוח כלים בלתי מאויישים (כלב"ס) הוקצה תקציב של יותר משישה מליארד דולר לשנה בין השנים 2011-2015,⁵¹ המהווה כמעט עשרה אחוזים מהתקציב המוקצה למחקר ופיתוח בצבא ארה"ב.⁵² גם בישראל כבר עושים שימוש בכלי שיט וכלי רכב בלתי-מאויישים (כרב"ס) להשגת מגוון מטרות צבאיות.⁵³ בנוסף, מחלקות המחקר והפיתוח של גופי הביטחון פועלות לפיתוח FAR צבאיים שונים, כאשר השאיפה היא ששדה הקרב העתידי יהיה כמעט ריק מלוחמים אנושיים.⁵⁴

4. אופן קבלת החלטות על ידי FAR

ההחלטה אותה נדרש לקבל FAR מהווה למעשה בחירה, בזמן אמת, באחת מבין כמה אלטרנטיבות פעולה אפשריות לשם השגת מטרה מוגדרת מראש. אציג את שלבי קבלת ההחלטה של FAR על בסיס מודל קבלת החלטות מופשט המקובל בתחום הרובוטיקה, ואנמק מדוע חלק ממערכות קבלת ההחלטות המנחות את פעולותיהם של FAR עומדות בתנאי משפט אי-האפשרות של ארו (Arrow's Impossibility Theorem להלן: משפט ארו).⁵⁵ משפט ארו הוא משפט מתמטי מתחום הבחירה החברתית, לפיו מנגנוני קבלת החלטות שעומדים בדרישות מסוימות יובילו בהכרח לקבלת החלטה "שגויה"⁵⁶ בחלק מהמקרים. בחינת שלבי קבלת ההחלטה והצגת משפט ארו יעשו, אפוא, כדלקמן: בחינת כל שלב בתהליך קבלת ההחלטה תראה כי המנגנון עומד בדרישה (להלן: אקסיומה אותנאית), אחת או יותר, של המשפט. לשם העמדת דברים על דיוקם, בחינה זו תעשה הן בשפה מתמטית והן

AIR FORCE UNMANNED AIRCRAFT SYSTEMS FLIGHT PLAN, 2009-2047, p. 16, 41 (May 18, 2009). Available at: http://www.fas.org/irp/program/collect/uas_2009.pdf; U.S. DEP'T OF ARMY, SBIR SOLICITATION 07.2 TOPIC A07-032, MULTI-AGENT BASED SMALL UNIT EFFECTS PLANNING AND COLLABORATIVE ENGAGEMENT WITH UNMANNED SYSTEMS, p. 57-68 (2007)

U.S. DEP'T OF DEF., FY2011-2036, UNMANNED SYSTEMS INTEGRATED ROADMAP, p. 13 (2013)⁵¹

U.S. DEP'T OF DEF., The Budget for Fiscal Year 2013, p. 77, 83 (2013). Available at: www.whitehouse.gov/sites/default/files/omb/budget/fy2013/assets/defence.pdf⁵²

נעם ויטמן "הרחבת השימוש בכלי רכב בלתי מאויישים תתן יתרון מבצעי ותשמור על חיי אדם" אתר צה"ל⁵³

www.idf.il/1133-20696-he/Dover.aspx .7.5.2014; ענבל אורפז "הכירו את הפרוטקטור: כלי השיט הבלתי

מאויש הראשון בחיל הים" **TheMarker** 16.4.2013. www.themarker.com/technation/1.1994094 .16.4.2013; ענבל אורפז

"כשיש נגיעה בגדר המערכת, אפשר להקפיץ רובוטים במקום חיילים" **TheMarker** 4.9.2012

www.themarker.com/technation/1.1816281; אפרת כהן "רובוטים בכל זירה" **ISRAELDEFENSE**

www.israeldefense.co.il/?CategoryID=411&ArticleID=916 .13.9.2011

4.1.2014 אמיר בוחבט "רובוטים, מל"טים וטנקים ריקים: לוחמת העתיד בצה"ל" **וואלה!**

http://news.walla.co.il/?w=/2689/2709097; ניר סגל "כבר בשנת 2015: רובוטים בגדודי החי"ר של צה"ל" **עיתון**

www.mako.co.il/pzm-magazine/war-games/Article-c38703876e76341006.htm .6.1.2014 **"במחנה"**

אלקנה שור "רב"ט רובוט: הצצה לצה"ל שנת 2035" **nrg** 14.11.2013

www.nrg.co.il/online/1/ART2/521/903.html

יתרון צבאי נוסף בהקניית אוטונומיה מלאה ל-FAR במצבי מלחמה הוא הגברת יכולת ההרתעה. במילותיו של ישראל אומן: "ידוע שאמל"ח קיים כדי לא להשתמש בו. כשהאויב יודע שיש לך אמל"ח, הוא עשוי לא לתקוף. כך במערכות בלתי מאויישות: הן קיימות כדי ליצור תמריץ אצל האויב לא לתקוף אותך. אם נציב מערכת מסוימת שתגיב באופן אוטונומי ואוטומטי לחלוטין על כל שיגור טיל של האויב, ונודיע על כך מראש, יתכן שהאויב יימנע מלשגר לעברנו, אם יידע בבירור שהמערכת שלנו תגיב ללא כל אפשרות להתערב בתגובתה". ראו: דן ארקין "העתיד של צה"ל - חיילים-רובוטים ונחילי מזל"טים" **iHLS** 29.1.2014 <http://i-hls.com/he/2014/01/future-idf-robot-soldiers-uav-swarm>

/swarms
זוהי, למשל, טקטיקת הלחימה בה נקטה ממשלת ארצות הברית במהלך משבר הטילים בקובה. ראו:

Graham T. Allison *Conceptual Modes and the Cuban Missile Crisis*, 63 THE AMERICAN POLITICAL SCIENCE REV. 689 (1969)

Arrow, לעיל ה"ש 4⁵⁵

המושג "החלטה שגויה" לעניין דידן יוגדר להלן בעמ' 10⁵⁶

בשפה מילולית.⁵⁷ לבסוף, לאחר שיובהרו תנאי ואקסיומות המשפט ויוכח כי המנגנון עומד בהם, אציג את השלכותיו הפרקטיות של המשפט על עולם הרובוטים האוטונומיים, ואבחן קטגוריית גיאות מחשב ייחודית לעולם זה. משאובחנה קטגוריה זו, ולאורם של האתגרים הניצבים כיום בפני שלטון החוק בהטלת אחריותיות לתוצאות שנגרמות מפעולות של מחשבים, אראה כיצד יישומו של משפט ארו תורם לגיבוש מסגרת נורמטיבית להתמודדות עם אתגרים אלו.

הניתוח שיוצע יתבסס על מודל קבלת ההחלטות המופשט OODA.⁵⁸ מודל זה נבחר משום שהוא המודל המקובל במסגרת בניית מודלים עסקיים, בקרב מהנדסים ואנשי צבא.⁵⁹ שלבי קבלת החלטה על פי מודל זה מתחלקים לארבעה: התבוננות (Observe) - זיהוי כל האלטרנטיבות האפשריות. התמצאות (Orient) - חישוב התוצאות שינבעו מנקיטה בכל אחת מהאלטרנטיבות על פי קריטריונים קבועים מראש. החלטה (Decide) - השוואה בין התוצאות על פי מנגנון שקלול קבוע מראש, והחלטה על האלטרנטיבה האופטימאלית. פעולה (Act) – נקיטה באלטרנטיבה שנבחרה.⁶⁰ החלת מודל זה על אופן קבלת ההחלטות של FAR תחולק לארבעת שלבי עיבוד האינפורמציה הנדרשת לקבלת ההחלטה, בהתאמה: רכישת אינפורמציה (Information Acquisition), ניתוח האינפורמציה (Information Analysis), קבלת החלטה (Decision Selection) וביצוע הפעולה (Action Implementation).⁶¹ לשם המחשת המודל, נתבונן במנגנון קבלת ההחלטות שמובנה ב- EXACTO 50-caliber, הקליע המונחה הראשון בעולם אותו הציגה לאחרונה סוכנות המו"פ של צבא ארה"ב.⁶² לקליע מוגדרת מראש מטרה, הגעה מנקודה A לנקודה B. במהלך תעופתו, המערכת האופטית בה הוא מצויד אוספת אינפורמציה ובונה תמונת עולם (World Model). על בסיס תמונה

⁵⁷ מבנה הרשימה כאמור נוסח יפה במילותיו של ישראל אומן:

"If you cannot explain it in words, it is not worth very much; but if you can *only* explain it in words, it also is not worth very much", "Of course, one must add immediately that once one has a model, and one has a set of assumptions and a conclusion, if it remains in mathematical form, it is not worth very much. You have to be able to translate it back into words" [Emphasis in the original] See: Yisrael Aumman, *Economic Theory and Mathematical Method: An Interview*, in ARROW AND THE ASCENT OF MODERN ECONOMY THEORY 136 (G.R. Feiwel ed., 1987)

ROBERT CORAM, BOYD: THE FIGHTER PILOT WHO CHANGES THE ART OF WAR 334 (2002); GRANT T. HAMMOND, THE MINOR OF WAR: JOHN BOYD AND AMERICAN SECURITY (Lorraine Atherton ed., 2001);

FRANS P.B. OSIGNA, SCIENCE, STRATEGY AND WAR: THE STRATEGIC THEORY OF JOHN BOYD (2006) U.S. DEP'T OF AIR FORCE, UNITED STATES AIR FORCE UNMANNED AIRCRAFT SYSTEMS FLIGHT PLAN, 2009-2047, p. 16 (May 18, 2009); Shane Harris, *Out of the Loop: The Human-free Future of Unmanned Aerial Vehicles*, in EMERGING THREATS IN NATIONAL SECURITY AND LAW (Peter Berkowitz ed., 2012); Wagner, *Taking Humans Out of the Loop: Implications for International Humanitarian Law* 21 J.L. INFO. & SCI. 1 (2011); RICHARDS CHET, CERTAIN TO WIN: THE STRATEGY OF JOHN BOYDS APPLIED TO BUSINESS 162-71 (2004)

JOHN BOYD, A DISCOURSE ON WINNING AND LOSING (1987); Teemu Mätäsniemi and Valtion teknillinen tutkimuskeskus, *Operational Decision Making in the Process Industry: Multidisciplinary approach*, 2442 VTT TIEDOTTEITA 107 (2008)

Gilles Coppin & François Legras, *Autonomy Spectrum and Performance Perception Issues in Swarm Supervisory Control*, 100 PROCEEDINGS OF THE IEEE 590, 592 (2012); Raja Parasuraman et al., *A Model for Types and Levels of Human Interaction with Automation*, 30 IEEE TRANSACTION ON 286, 288 (2000)

EXACTO Demonstrates First-Ever Guided .50-Caliber Bullets, DARPA (July 10, 2014). ⁶² <http://www.darpa.mil/NewsEvents/Releases/2014/07/10a.aspx>; Nick Lavars, *DARPA's guided sniper bullet changes path mid-flight*, GIZMAG (July 15, 2014). <http://www.gizmag.com/darpa-sniper-bullet-change-path/32952>

זו מתווה מנגנון קבלת ההחלטות שלו מסלולי תנועה אופציונאליים, תוך שקלול כיוון ועוצמת הרוח וגורמים מפריעים נוספים, ומתוכנת להחליט על המסלול האופטימאלי. מנגנון קבלת החלטות דומה מובנה גם במכוניות אוטונומיות.⁶³

רכישת אינפורמציה

כאמור, FAR רוכש את האינפורמציה מהעולם החיצון על ידי רכיבי חומרה שונים, ובונה מהמידע הגולמי אלטרנטיבות פעולה אפשריות ביניהן יידרש מנגנון קבלת ההחלטות להכריע. הנחת המוצא לטיעון הינה כי ידועה לנו ההכרעה "הנכונה" בין כל זוג אלטרנטיבות אפשריות. מכאן, שהאלטרנטיבות ברות השוואה וקיימת תשובה "נכונה" לשאלה באיזו מהן לבחור. קרי, לצורך ניתוח אופן קבלת החלטה על ידי FAR, ובפרט לצורך בחינתה במושגי טוב ורע, מתחייב שתהיה בידינו אמת מידה אובייקטיבית לבחינה מהי האלטרנטיבה הראויה. בפרט, הכרחי של-FAR תהיה יכולת להשוות בין האלטרנטיבות השונות ולדרג אותן על פי אמת המידה הנתונה, כיוון שללא יכולות אלו הוא לא יוכל לפעול בעולם האמיתי.

בהמשך, אעשה שימוש במושג "דירוג" כדי לבטא את היחס בין האלטרנטיבות האופציונאליות העומדות בפני ה-FAR, כאשר ההחלטה שתתקבל תהיה לנקוט באלטרנטיבה שדורגה בראש. בסיטואציה נתונה, אם ה-FAR יחליט על האלטרנטיבה האופטימאלית על פי אמות המידה האובייקטיביות הנתונות נגיד שהוא קיבל את "ההחלטה הנכונה", ובאם ה-FAR יחליט על אחת מבין שאר האלטרנטיבות נגיד שהוא קיבל "החלטה שגויה". חשוב לציין, כי דירוגים בין אלטרנטיבות פעולה קיימים גם בעולם האנושי היומיומי בסיטואציות של קבלת החלטות, דוגמת תעדוף במיון פצועים או נהלי פתיחה באש. עם התפתחות השימוש ברובוטים אוטונומיים, נצרכים דירוגים אלו גם לארכיטקטורות של קבלת החלטות על ידם.⁶⁴

על אף שבמובנה הלוגי-מתמטי הנחה זו הינה אקסיומטית, בעולם הפילוסופי-משפטי קיימים דיונים הנסובים על אמת המידה האובייקטיבית לפיה יונחה ה-FAR לדרג את האלטרנטיבות. חילוקי הדעות סביב סוגיה זו מערימים קושי נוסף עבור המתכנתים, במקרים בהם ההכרעה על ההחלטה הנכונה קשה כשלעצמה.

כאמור, פילוסופים והוגי דעות חלוקים בדעתם באשר לתורה המוסרית הראויה להנחות פעולות אנושיות, קל וחומר פעולות של FAR.⁶⁵ תורות מוסריות שונות מציגות אמות מידה שונות לבחינת

NIDHI KLARA at el., LIABILITY AND REGULATION OF AUTONOMOUS VEHICLE TECHNOLOGIES 6-8 ⁶³ (California PATH Program, Institute of Transportation Studies, University of California at Berkeley, 2009)

⁶⁴ כך למשל רמת הרגישות של מערכות ליירוט טילים למיניהן. ראו: Inbal Orpaz, *How Does Iron Dome Operate* HAARETZ (November 19, 2012). <http://www.haaretz.com/news/features/how-does-the-irondome-work.premium-1.478988>; *Iron Dome Defense System Against Short Range Artillery Rockets*, RAFAEL. Available at http://www.rafael.co.il/marketing/SIP_STORAGE/FILES/6/946.pdf

Carl Shulman, et al., *Which Consequentialism? Machine Ethics and Moral Divergence*, ROYNOLDS ⁶⁵ AND CASSINELLI 23-25 (2009)

מוסריותה של פעולה, ושוני זה מוביל אף לעיתים תכופות לדירוג שונה של אלטרנטיבות פעולה נתונות. כך למשל, הגישה התועלתנית טוענת כי אמת המידה הראויה לבחינת מוסריות פעולה הינה מקסום התועלת שתנבע מפעולה זו.⁶⁶ לעומתה, גישה דאונטולוגית מבוססת על ציוויים ואיסורים מוחלטים ללא קשר לתוצאות.⁶⁷ נקל להעלות על הדעת סיטואציות בהן שתי הגישות יובילו לדירוג שונה לחלוטין של אלטרנטיבות הפעולה האופציונאליות.

באשר לקשיים הניצבים בפני מתכנתי ה-FAR, ראוי לתת את הדעת כי גם בהינתן עקרונות מוסריים מנחים, תכנון ותכנות FAR מחייבים הכרעה מראש כיצד הוא יפעל בסיטואציות של קונפליקטים מוסריים כבר בשלב זה, טרם יציאתו לשדה הקרב.⁶⁸ למשל, הדיון התיאורטי אודות האבחנה בין פעולה פאסיבית לבין פעולה אקטיבית שמומחשת באמצעות בעיית הקרונית (Trolley Problem) הידועה,⁶⁹ יהפוך לשאלה קונקרטית שדורשת תשובה מראש מהמתכנתים. כך גם הבעיה שבדירוג אלטרנטיבות אינקומנסרביליות,⁷⁰ לכאורה, כגון אלו בהן כרוך שקלול תמורות (Off Trade) בין אובדן חיי אדם ואובדן רכוש, תדרוש פתרון מוגדר היטב וקבוע מראש.

על אף המחלוקות אודות אמת המידה הראויה והקשיים הניצבים בפני המתכנתים, חשוב להדגיש כי הטיעון המוצע תקף ללא תלות בשיטת הדירוג הספציפית שתבחר. כאשר לפנינו FAR ששיטת הדירוג שנקבעה לו עומדת בדרישות משפט ארו, יהיו אשר יהיו כללי הפעולה שמרכיבים שיטה זו, נפקותו של המשפט תהיה רלוונטית ל-FAR זה.

לסיכומו של החלק הראשון, הראינו כי על מנת ש-FAR יוכל לפעול בעולם האמיתי חייב להתקיים דירוג אופטימאלי של האלטרנטיבות, אשר לפיו נקבע האם התקבלה ההחלטה הנכונה או החלטה שגויה. בנוסף, ל-FAR ישנה יכולת להשוות בין כל זוג אלטרנטיבות אופציונאליות, כך שהיחסים בין כל הזוגות ייבנו את הדירוג הסופי והמלא של האלטרנטיבות. להלן נסמן את האלטרנטיבות ביניהן ה-FAR נדרש להחליט באותיות x, y, z, \dots ואת הדירוג הסופי עליו החליט מנגנון קבלת

R. Brandt, *Utilitarianism and the Rules of War*, 1 PHILOSOPHY AND PUBLIC AFFAIRS 145, 145-165⁶⁶ (1972); C. Cloos, *The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism*, 38-45 (AAAI Technical Report FS-05-06, 2005)

Powers, T., *Deontological Machine Ethics*, 79-86 (AAAI Technical Report FS-05-06, 2005)⁶⁷

⁶⁸ ודוק, בניגוד ללוחם אנושי שעלול להמצא בדילמה מוסרית אם יידרש להחלטה בזמן אמת, על המתכנתים להכריע בקונפליקטים מוסריים לצורך קביעת האלטרנטיבה עליה ירצו ה-FAR יחליט. כפועל יוצא מכך, המגמה למתן אוטונומיה לרובוטים כרוכה בד-בבד בטרנספורמציה של דילמות מוסריות בהם יתקלו הפועלים האנושיים לקונפליקטים מוסריים אליהם יידרשו המתכנתים. על אף שקצרה היריעה מלדון בהשלכותיה של טרנספורמציה זו, פטור בלא כלום אי-אפשר: "דילמה אינה, אפוא, סתם סיטואציה שבה יש שיקולים מתנגשים, אלא היא סיטואציה כזאת, שבה שיקולים אלו מעיקים על הפועל ומאיימים לשתק אותו... ההתייחסות לפועל ולמצבו מהותית, אפוא, בתיאורה של דילמה. לעומת זאת, קונפליקט מצוין התנגשות גרידא, ובהקשר זה נוכל לדבר גם על התנגשות בין אמנות (beliefs).". לדיון אודות האבחנה בין דילמות מוסריות לקונפליקטים מוסריים ראו באריכות: דניאל סטטמן **דילמות מוסריות** 13 (מאגנס אוניברסיטה העברית הוצאה לאור, 1991) (להלן: דניאל סטטמן)

Philippa Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, 5 OXFORD REV. 5, 5-⁶⁹ 15 (1967); Judith Jarvis Thomson, *Killing, Letting Die, and the Trolley Problem*, 59 THE MONTIS 204 (1976); Judith Jarvis Thomson, *The Trolley Problem*, 94 YALE L. J. 1395 (1985); Francis Myrna Kamm, *Harming Some to Save Others*, 57 PHILOSOPHICAL STUDIES 227-60 (1989)

⁷⁰ דניאל סטטמן, לעיל ה"ש 68, בעמ' 74

ההחלטות באות R. כמו כן, אם על פי הדירוג R אלטרנטיבה x עדיפה על אלטרנטיבה y נסמן $x >_R y$, ואם היחס ביניהן על פי R לא ידוע לנו, אך ידוע שקיים כזה, נסמן xRy .

אקסיומה I: עבור כל שתי אלטרנטיבות אופציונאליות x ו-y, בדירוג הסופי R מתקיים:
אלטרנטיבה x מדורגת מעל אלטרנטיבה y, או אלטרנטיבה y מדורגת מעל אלטרנטיבה x.
בניסוח מדויק: $\forall x, y \quad x >_R y \vee y >_R x$ ⁷¹

ניתוח האינפורמציה

לאחר עיבוד האינפורמציה החיצונית ומיפוי האלטרנטיבות האופציונאליות, ה-FAR נדרש להחליט באיזו לנקוט. לצורך כך מתחייבים קריטריונים כלשהם לפיהם יעריך ה-FAR את היתרונות והחסרונות שבנקיטה בכל אחת מהאלטרנטיבות. בהגדרתו, כל קריטריון מהווה אמת מידה למשתנה כלשהו במכלול המשתנים שברצוננו לשקלל בהערכת האלטרנטיבות, כך שבמצב בו נדרשת החלטה מדרג כל קריטריון את האלטרנטיבות על פי משתנה זה. לשם הדוגמא, נתבונן על מכונית אוטונומית שעומדת בפני התנגשות בלתי נמנעת, והיא נדרשת להחליט בין שלושת אלטרנטיבות הפעולה הבאות: x – בה יהרגו שני עוברי אורח, ויגרום נזק ממוני בשווי 200; y – בה יהרגו שלושת נוסעי המכונית, ויגרום נזק ממוני בשווי 100; z – בה יהרג רוכב אופנוע, ויגרום נזק ממוני בשווי 300. נניח כי שניים מבין הקריטריונים שהוגדרו למכונית בהערכת אלטרנטיבה הינם 'מזעור הפגיעה בחיי אדם' ו-'מזעור נזק ממוני'. הקריטריון הראשון ידרג $z > x > y$, השני ידרג $y > x > z$ וכך גם יתר הקריטריונים שהוגדרו למכונית ידרגו את האלטרנטיבות בצורה מסוימת, כך ששקלול סך הדירוגים יוביל לדירוג הסופי R. ראוי לציין, כי דוגמא זו כבר מצביעה על הבעייתיות הכרוכה בטענה כי במצבים בהם אין אלטרנטיבה עדיפה מובהקת על ה-FAR להמנע מקבלת החלטה ללא התערבות מפעיל אנושי, שכן אי נקיטת פעולה גם כן יכולה לגרום נזק. האבחנה בין פעולה אקטיבית וחוסר פעולה הינה שקרית במבחן התוצאה, משום שההחלטה על חוסר פעולה צריכה להתקבל כבר בשלב התכנות.

יושם אל לב, כי מהגדרת הדירוג של קריטריון נובע כי דירוג זה הינו טרנזיטיבי וקונסיסטנטי. מכיוון שהאלטרנטיבות מדורגות על פי ערכו של המשתנה אותו מוגדר הקריטריון לאמוד, בדירוג נתון היחס בין כל שתי אלטרנטיבות קבוע, ובפרט טרנזיטיבי,⁷² וכמו כן היחס בין כל שתיים מהן קונסיסטנטי בסיטואציות שונות.⁷³ כך למשל, על פי דירוגם של הקריטריונים 'מזעור הפגיעה בחיי

⁷¹ Arrow, לעיל ה"ש 4, בעמ' 331

למען בהירות הדיון, הטיעון המוצע יתבסס על ההנחה שכל הדירוגים הינם "דירוגים חזקים" ($>$) כאמור. זאת, על אף שהטיעון תקף גם אם קיימים "דירוגים חלשים" (\geq), שהרי משפט ארו מתיר דירוגים שכאלו. נפקותה של הערה זו הינה החלת הטיעון גם על FAR בעלי רמת אוטונומיות נמוכה יותר, אשר תוכנתו כך שבמצב של "שוויון" בין אלטרנטיבות אופציונאליות הם נדרשים להנחות ממפעיל אנושי.

⁷² בדירוגו של כל קריטריון מתקיים: $\forall x, y, z \quad x > y \wedge y > z \Rightarrow x > z$

⁷³ בכל שתי סיטואציות $S \neq S'$ מתקיים: $\forall S, S' \text{ s. t. } x, y \in S, S', \quad x >_R y \Leftrightarrow x >_{R'} y$

אדם' ו-'מזעור נזק ממוני' שהוצגו, בדירוג הראשון בפרט מתקיים $z > y$ ובשני מתקיים $y > z$. יחס זה יישמר גם אם תתווספה עוד אלטרנטיבות.

חלק אינטגרלי מתכנון ובניית מערכת קבלת ההחלטות של FAR הוא, אם כן, קביעת הקריטריונים הרלוונטיים להערכת האלטרנטיבות והטמעתם באלגוריתם. ככל שמוקנית יותר אוטונומיה ל-FAR, כך מפורמלים, לצד אלגוריתמים פשוטים יחסית בהם מוטבעים הקריטריונים, גם כללים מורכבים יותר כגון כללי מוסר או מלחמה לקוד ממוחשב. גם כללים אלו בנויים על פי קריטריונים קבועים מראש.⁷⁴ עם ההתקדמויות הטכנולוגיות, מפותחים גם FAR בעלי יכולות הסקת מסקנות ולמידה מניסיון פעולה בעולם האמיתי.⁷⁵ גם FAR אלו מסווגים את המידע שצברו במהלך ניסיון ל"טוב" ול"רע", על פי קריטריונים שהוגדרו להם מראש.⁷⁶

ראוי לציין, כי על אף שבחינת האלטרנטיבות באמצעות קריטריונים נומינאליים מחוייבת המציאות לצורך תכנות FAR, הדיון אודות זהותם של הקריטריונים בהם ראוי להתחשב בקביעת ההחלטה מורכב מאוד ונוגע בסוגיות פילוסופיות ומשפטיות רבות. שוב, נתבונן על המכונת האוטונומית העומדת בפני התנגשות בלתי נמנעת, כאשר כעת עליה להחליט בין התנגשות באופנוע שרוכבו חובש קסדה לבין התנגשות באופנוע בו הרוכב אינו חובש קסדה. השכל הישר כנראה יתמוך בגישה תועלתנית ששואפת למזער את הפגיעה בחיי אדם, ומכאן שאחד מהקריטריונים שיוגדרו למכונת יתחשב במידת הנזק שתגרם לנפגע. דא עקא, לא זו בלבד שמשמעות קביעת קריטריון כזה היא, דה-פאקטו, הוראה לפגוע אקטיבית באחד משני הרוכבים, אלא גם הגדרת הקריטריון הספציפי הזה מהווה הענשה של אותו הרוכב רק בגלל שרכב עם קסדה, תוצאה שנראית אבסורדית.⁷⁷

לסיכומו של השלב השני, ראינו כי לצורך תפקודו של FAR נדרשים המתכנתים להגדיר קריטריונים באמצעותם יעריך את האלטרנטיבות האופציונאליות. בנוסף, מהגדרתו, דירוגו של כל קריטריון את האלטרנטיבות האופציונאליות הוא טרנזיטיבי ועקבי. להלן נסמן ב- n את מספר הקריטריונים שהוגדרו, נסמן את הקריטריונים ב- $C_1, C_2 \dots C_i \dots C_n$ ואת דירוגו של הקריטריון C_i נסמן ב- R_i .

Thomas M. Powers, *Deontological Machine*; 14-21 בעמ' 23, לעיל ה"ש 74, *Governing Lethal Behavior* (2005). Available at: <https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-012.pdf>; Bruce M. McLaren, *Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions*, 21 INTELLIGENT SYSTEMS, IEEE 2, 2-10 (2006); Bruce M. McLaren, *Extensionally Defining Principles and Case in Ethics: An AI Model*, 150 ARTIFICIAL INTELLIGENCE 145 (2003); Russell W. Robbins & William A. Wallace, *Decision Support for Ethical Problem Solving: A Multi-DECISION SUPPORT SYSTEMS* 1571, 1571-87 (2007); SAMIR CHORPA & LAURENCE *Agent Approach*, 43 F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS (University of Michigan Press. 2011)

Emanuel Diamant, *Cognitive Robotics: For Never was a Story of More Woe than This* (2014).⁷⁵ Available at: <http://arxiv.org/vc/arxiv/papers/1401/1401.4127v2.pdf>; ICT Work Programme 2013, 33-35. Available at: <http://cordis.europa.eu/fp7/ict/docs/ict-wp2013-10-7-2013.pdf>; V. Ferraris, et al. *Defining Profiling* 9, available at SSRN 2366564 (2013)

M Al Fahdi, et al., *Towards An Automated Forensic Examiner (AFE) Based Upon Criminal Profiling*⁷⁶ & *Artificial Intelligence*, 5-6 (2013).

Noah J. Goodall, *Ethical Decision Making During Automated Vehicle Crashes*, 8 (2014); Patrick Lin,⁷⁷ *The robot car of tomorrow may just be programmed to hit you*, RIED (June 4, 2014). [/http://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you](http://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you)

כמו כן, אם הקריטריון C_i מדרג את אלטרנטיבה x מעל אלטרנטיבה y נסמן $y >_i x$. לבסוף, בסיטואציה נתונה, נכנה את כלל דירוגי הקריטריונים $R_1 \dots R_n$ בשם פרופיל ונסמנו P .

אקסיומה II: כל דירוג R_i של הקריטריון C_i הוא טרנזיטיבי וקונסיסטנטי.⁷⁸

קבלת החלטה

לאחר שבוססה יכולתו של FAR להשוות בין אלטרנטיבות פעולה שונות באמצעות הקריטריונים שהוגדרו לו, ונמצאנו למדים כי הוא עומד באקסיומות I ו-II עליהן מושתת משפט ארו, נפנה לבחון את אופן פעולתו של "החלק החושב" במנגנון קבלת החלטות – המנגנון שמשקלל בין דירוגי הקריטריונים (להלן: מנגנון השקלול). לצד הסוגיות שיתעוררו במהלך הגדרת הדירוג הרצוי בכל סיטואציה והקריטריונים שיילקחו בחשבון, רכיב התוכנה הקריטי ביותר בתהליך קבלת החלטה הוא זה שקולט (Input) את דירוגי הקריטריונים ופולט (Output) את הדירוג הסופי R , אשר בראשו האלטרנטיבה בה ינקוט ה-FAR. לרכיב זה, מנגנון השקלול, ישנה את ההשפעה הגדולה ביותר על ההחלטה באיזו אלטרנטיבה ינקוט ה-FAR, שהרי לשם כך נוצר. לשם המחשה, נתבונן בדוגמת המכונית האוטונומית שנדרשת להחליט בין שלושת האלטרנטיבות x, y, z , ונניח שהקריטריון C_1 מדרג $y >_1 x >_1 z$ ואילו הקריטריון C_2 מדרג $z >_2 x >_2 y$. בשלב זה לא ברורה עדיין מהי הפעולה בה צריכה לנקוט המכונית. רק לאחר שייכנס לפעולה מנגנון השקלול, שישקלל את הדירוגים השונים ויצור דירוג סופי, תוכל המכונית לפעול.

בפסקאות הבאות תאופיין משפחה של מנגנוני שקלול, אשר ישנו בסיס סביר לחשוב כי יהיו בשימוש בתכנות FAR שנועד לפעול בשדה הקרב. אפיון זה יעשה מתוך הפריזמה של ארבעת תנאי משפט ארו. לאחר מכן, תבחנה השלכותיו של המשפט על ההחלטות שיתקבלו על ידי FAR בהם מנגנון השקלול משתייך למשפחה זו.

ראשית, כדי שמחשב יוכל לשקלל את הקריטריונים, הם יצטרכו להיות בעלי מידה ניתנת להשוואה (Commensurable Unit).⁷⁹ בנוסף, מכיוון שה-FAR מיועד לפעול בעולם האמיתי, דרישה בסיסית ממנגנון השקלול הינה היכולת להחליט בכל מצב נתון. קרי, עבור כל פרופיל $R_1 \dots R_n$ מנגנון השקלול יצור דירוג R לפיו ה-FAR יפעל. מנגנון השקלול מכריע למעשה בין הדירוגים הסותרים של הקריטריונים, על פי נוסחה שהוגדרה לו. שקלולים שכאלה, המכונים בספרות "קבלת החלטה מרובת קריטריונים",⁸⁰ אינם ייחודיים למנגנונים ממוחשבים, גם קבלת החלטות אנושית נדרשת לאמת מידה ניתנת להשוואה שתשקלל דירוגים השונים. כך למשל, בקביעת המהירות המותרת

⁷⁸ Arrow, לעיל ה"ש 4, בעמ' 331, 334

⁷⁹ LING XU & JIAN-BO YANG, INTRODUCTION TO MULTI-CRITERIA DECISION MAKING AND THE EVIDENTIAL REASONING APPROACH 4 (Manchester School of Management, University of Manchester Institute of Science and Technology. 2001)

⁸⁰ שם, בעמ' 3

בקטע כביש מסוים, אנו נדרשים לאזן בין נוחות הנסיעה לבין בטיחות הנוסעים. מאחר שלא ניתן להשוות בין נוחות לבין בטיחות, נדרשת אמת מידה ניתנת להשוואה אשר לרוב מתורגמת לכסף.⁸¹

תנאי I: דירוגו של מנגנון השקלול, R, מוגדר עבור כל פרופיל P.⁸²

התנאי השני בו יידרש מנגנון השקלול לעמוד הוא אי-תלות באלטרנטיבות לא רלוונטיות (Independence of Irrelevant Alternatives). קרי, כאשר מנגנון השקלול מדרג את האלטרנטיבות על פי דירוגי הקריטריונים השונים, כל אלטרנטיבה נאמדת כשלעצמה ללא תלות בשאר האלטרנטיבות. במילים אחרות: מיקומן היחסי של האלטרנטיבות בדירוג R תלוי רק במיקומן היחסי בדירוגי הקריטריונים $R_1 \dots R_n$, כך שהערכת כל אלטרנטיבה לא תלויה באינפורמציה חיצונית.⁸³ כלומר, היחס בין כל שתי אלטרנטיבות בדירוג R לא ישתנה אם תתווסף אלטרנטיבה אופציונאלית שלישית, ולהיפך, אם תרד אלטרנטיבה היחס בין השתיים האחרות לא ישתנה.⁸⁴

תוצאה ישירה של תנאי זה היא איסור על שרירותיות או רנדומאליות מכל סוג שהוא בתהליך קבלת ההחלטה. זוהי תוצאה רצויה, שכן לאור האיסורים על פגיעה שרירותית בחירות,⁸⁵ קיפוח שרירותי של חיים ככלל,⁸⁶ ובמלחמה בפרט,⁸⁷ מנגנון קבלת החלטות של FAR שיתוכן לפעול בשדה הקרב יידרש שלא לכלול פונקציות רנדום (Random) מכל סוג. יתר על כן, חשיבותו של תנאי זה טמונה בכך שהוא מבטיח, בין היתר, שמנגנון השקלול יהיה קונסיסטנטי וטרנזיטיבי. אם מנגנון שקלול אינו קונסיסטנטי, הוא יהיה חשוף למניפולציות אשר מערערות על אמינותו ועל לגיטימיות השימוש בו בשדה הקרב.⁸⁸ ודוקו, אקסיומה II ביססה את הטרנזיטיביות והקונסיסטנטיות של דירוג כל קריטריון. כעת ענייננו בבחינת פעולתו של מנגנון השקלול עצמו, כלומר האופן בו נקבע הדירוג הסופי R, וסביר לדרוש כי המנגנון יהיה קונסיסטנטי וטרנזיטיבי בהחלטות שיתקבלו מכוחו. הדבר נכון במיוחד בשדה הקרב בו מיועד ה-FAR לגרום נזק בכפוף לחוקים או כללים מוגדרים, וראוי

⁸¹ Jonathan Wolff, *Risk, fear, blame, shame and the regulation of public safety*, 22 ECONOMICS AND PHILOSOPHY 409, 411 (2006)

⁸² Arrow, לעיל ה"ש 4, בעמ' 336

⁸³ John Geanakoplos, *Three brief proofs of Arrow's impossibility theory*, 26 ECONOMIC THEORY 211 (2005) (hereinafter: *Geanakoplos*)

⁸⁴ Arrow, לעיל ה"ש 4, בעמ' 337

⁸⁵ U.N. Human Rights Committee, 106th Session, *ICJ Comments to the U.N. Human Rights Committee on... The International Covenant on Civil and Political Rights* (Feb. 2012). Available at: <http://www.ohchr.org/Documents/HRBodies/CCPR/GConArticle9/ICJ.pdf>

⁸⁶ NIGEL S. RODLEY, *THE TREATMENT OF PRISONERS UNDER DEVELOPMENT OF THE INTERNATIONAL LAW* 355-368 (2nd ed., Oxford University Press, 1999); U.N. General Assembly, *Extrajudicial, Summary or Arbitrary Executions* (A/61/311) pp. 12-13 (5 September 2006).

⁸⁷ JEAN-MARIE HENCKAERTS, et al., *CUSTOMARY INTERNATIONAL HUMANITARIAN LAW: RULES § 1, Rule 89* (Cambridge University Press, 2005); GA res. 2200A (XXI), 21 UN GAOR Supp. (No. 16) at 52, UN Doc. A/6316 (1966); 999 UNTS 171; 6 ILM 368 (1967)

⁸⁸ כך למשל, על בסיס משפט ארו הוכיחו Gibbard ו-Satterthwaite כי אין שיטת בחירה שאינה ניתנת למניפולציה. ראו: Allan Gibbard, *Manipulation of voting schemes: a general result*, 41 JOURNAL OF THE ECONOMETRIC SOCIETY 587 (1973); Mark Allen Satterthwaite, *Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions*, 10 JOURNAL OF ECONOMIC THEORY (1975); SEMIH KORAY, *SELF-SELECTIVE SOCIAL CHOICE FUNCTIONS VERIFY ARROW AND GIBBARD-SATTERTHWAITE THEOREMS*, 68 ECONOMETRICA 981 (2000)

שיעשה זאת באופן עקבי. לשם ההמחשה, מסתבר כי הדירוג הנכון, לפיו נרצה שה-FAR יפעל, יעלה בקנה אחד עם הדירוג המצופה כיום מלוחמים אנושיים. לכן, הגיוני כי למנגנון קבלת ההחלטות של FAR הפועלים בשדה קרב ייצרבו חוקי המלחמה הצודקת וכללי המשפט הבינלאומי ההומניטארי. סט חוקים וכללים אלו, בהגדרתו, שואף להוביל לתוצאה עקבית.⁸⁹

דרך שקולה לבטא תנאי זה היא: אם בדירוגו של כל קריטריון היחס בין x לבין y זהה בשני פרופילים שונים P, P' אזי היחס ביניהן בדירוג הסופי R ו- R' יהיה זהה.⁹⁰

תנאי II: $\text{if } \forall C_i x R_i y \Leftrightarrow x R'_i y \text{ then } x R y = x R' y$

בנוסף לשני התנאים הראשונים, מנגנון קבלת ההחלטות נדרש לכבד קונצנזוס (Unanimity) בין דירוגי הקריטריונים. כלומר, עבור כל שתי אלטרנטיבות x, y , אם כל R_i מדרג $x >_i y$ אזי בדירוג הסופי R שיצור מנגנון השקלול יוחלט כי $x >_R y$.⁹² תנאי זה נקרא יעילות פארטו (Pareto Efficiency), והוא משמש ככלי למדידת יעילותן של מערכות שונות בתחומי הנדסה וטכנולוגיה.⁹³ כאשר נשלח FAR נתון לשדה הקרב, נרצה שההחלטה שמנגנון השקלול ידרג בראש ובה ינקוט ה-FAR תהיה פארטו אופטימאלית (Pareto Optimality) על פי הקריטריונים שהוגדרו לו.

תנאי III: $\forall x, y \text{ if } \forall C_i x >_i y \text{ then } x >_R y$

התנאי הרביעי והאחרון בו יידרש לעמוד מנגנון השקלול, הוא היעדר קריטריון דיקטטורי שישפיע באופן בלעדי על ההחלטה שתתקבל. קרי, בין הקריטריונים שהוגדרו לא ייווצר מצב בו קיים קריטריון יחיד C_d אשר לפי דירוגו, ורק לפיו, יצור מנגנון השקלול את הדירוג הסופי R . כמובן, מנגנון השקלול יכול להעניק לכל קריטריון משקל שונה. הדרישה הינה כי לא יהיה קריטריון שמנגנון השקלול יעניק לו בלעדיות על קבלת ההחלטה, ללא התחשבות בדירוגם של שאר הקריטריונים.

תנאי VI: $\exists C_i \text{ s.t. } \forall x, y x >_i y \Rightarrow x >_R y$

Chris Jenks, *Law from Above: Unmanned Aerial Systems, Use of Force and the Law of Armed Conflict*,⁸⁹ 85 N.D. L. REV. 649, 671 (2009); Ryan J. Vogel, *Drone Warfare and the Law of Armed Conflict*, 39 DENV. J. INT'L L. & POL'Y 101, 138 (2010)

Ning Neil Yu, *A one-shot proof of Arrow's impossibility theorem*, 50 ECONOMIC THEORY 523 (2012)⁹⁰ (hereinafter: Yu)

Arrow⁹¹, לעיל ה"ש 4, בעמ' 336-37

Geanakoplos⁹², לעיל ה"ש 83, בעמ' 212

Parke Godfrey, et al., *Algorithms and analyses for maximal vector computation*, 16 THE VLDB JOURNAL – THE INTERNATIONAL JOURNAL ON VERY LARGE DATA BASES 1, 5-28 (2007); Bogdan Tomoiagă, et al., *Pareto optimal reconfiguration of power distribution systems using a genetic algorithm based on NSGA-II*, 6 ENERGIES 1439 (2013)

Yu⁹⁴, לעיל ה"ש 90, בעמ' 524

Arrow⁹⁵, לעיל ה"ש 4, בעמ' 339; Yu; לעיל ה"ש 90, בעמ' 524; Geanakoplos; לעיל ה"ש 83, בעמ' 212

5. החלת משפט ארו על מנגנון קבלת החלטות

בטרם יוצג משפט ארו ויבחנו השלכות יישומו על מנגנוני קבלת החלטות של FAR, נסכם את מאפייני ה-FAR מושא הטיעון. הדרישות האקסיומטיות הינן כי ל-FAR ישנה יכולת לדרג את האלטרנטיבות האופציונאליות באמצעות קריטריונים שהוגדרו לו, וכי דירוגו של כל קריטריון הוא טרנזיטיבי וקונסיסטנטי. בנוסף, מנגנון השקלול של ה-FAR עומד בארבעה תנאים: יש לו יכולת לדרג את האלטרנטיבות על סמך כל פרופיל נתון, הדירוג R נקבע ללא תלות באלטרנטיבות לא רלוונטיות, המנגנון מכבד קונצנזוס ואין בו קריטריון דיקטטורי.

משפט אי-האפשרות של ארו: לא קיימת שיטת דירוג שמקיימת את אקסיומות I ו-II, וגם את התנאים I-IV. ניסיון לבנות שיטה שכזאת יוביל לסתירה לוגית.⁹⁶ משפט שקול: בכל שיטת דירוג המקיימת את אקסיומות I ו-II ועומדת בתנאים I-III קיים קריטריון דיקטטורי.⁹⁷ מכאן נובע, כי לכל FAR שעומד באקסיומות I ו-II ותוכנת כך שמנגנון השקלול שלו עומד בתנאים I-III, ללא תלות בזהות הקריטריונים שהוגדרו עבורו - אחד מהם דיקטטורי. כפועל יוצא מכך, יהיו מקרים, עליהם נעמוד בפירוט להלן, בהם ההחלטה שתקבל תעמוד בניגוד לשאר הקריטריונים וממילא תהיה החלטה שגויה.

לשם המחשת המשפט, נתבונן על FAR הניצב בפני מספר אלטרנטיבות פעולה אופציונאליות להשגת מטרה שהוגדרה לו, כאשר ישנם שלושה קריטריונים אותם הוא צריך לשקלל: C_1 - מזעור פגיעה באזרחים; C_2 - מקסום פגיעה בלוחמים; C_3 - שמירה על שלמותו.⁹⁸ על פי המשפט, במידה ולפנינו מנגנון שקלול שעומד בשלושת התנאים הראשונים, אחד מהקריטריונים בהכרח דיקטטורי. אם C_3 הוא הקריטריון הדיקטטורי, יהיו מקרים בהם ה-FAR יבחר באלטרנטיבה עם מספר האזרחים ההרוגים הגדול ביותר ומספר הלוחמים ההרוגים הקטן ביותר, רק מכיוון שזוהי האלטרנטיבה שגורמת לו את הנזק הקטן ביותר. חלקם של מנגנוני השקלול יכול להמנע משגיאה זו במידה והמנגנון יכפה ש- $C_d \neq C_3$. אולם, כך על פי המשפט, גם במנגנונים אלו יהיה קיים קריטריון דיקטטורי $C_d = C_2$ או $C_d = C_1$, שיוביל לקבלת החלטה שגויה במקרים המתאימים.

לסיכום, אם משפט ארו אכן חל על FAR נתון, הרי ש-FAR זה בהכרח יקבל החלטה שגויה במקרים מסוימים. להלן אציג שיטה לזיהוי הקריטריון הדיקטטורי ומקרים אלו עוד בשלב תכנון ה-FAR.

⁹⁶ Arrow, לעיל ה"ש 4, בעמ' 340-42

⁹⁷ Geanakoplos, לעיל ה"ש 83, בעמ' 212

⁹⁸ דוגמא זו נבנתה ברוח שלושת חוקי הרובוטיקה שנוסחו על ידי אייזק אסימוב: 1. לא יפגע רובוט לרעה בבו אדם, ולא יניח, במחדל, שאדם ייפגע. 2. רובוט חייב לציית לפקודותיו של אדם [כל עוד הן לא סותרות את החוק הראשון]. 3. רובוט ידאג לשמור על קיומו ושלמותו [כל עוד הגנה זו אינה עומדת בסתירה לחוק הראשון או לחוק השני]. (ההשמטות בסוגריים [] שלי. נ.פ.)

כפי שעולה מההיררכיה המוחלטת בין החוקים המקוריים, מנגנוני השקלול הצרובים ב-FAR המצויים בעולמו של אסימוב אינם עומדים בתנאי משפט ארו, שהרי קיים בהם קריטריון דיקטטורי - איסור פגיעה לרעה בבו אדם. השוו: Isaac Asimov, *I, Robot (2004 edition)*, New York, NY: BANTAM DELL (1950) יוער, כי על אף שנסיון להתיר קיומו של קריטריון דיקטטורי בתכנון ה-FAR דן, אותו יגדירו מראש המתכנתים, עלול להיראות פתרון מתבקש לבעיה - הגדרתו של קריטריון כזה עלולה להוות אתגר של ממש, בהתחשב בכך שה-FAR הינו, בסופו של יום, מכונת מלחמה שמטרתה נטילת חי אדם ו/או גרימת נזק תחת תנאים מסוימים.

צפיית הכלל הדיקטטורי וההחלטות השגויות

בחלק זה אראה כיצד ניתן לצפות את הקריטריון הדיקטטורי של FAR נתון, וכמו כן את הסיטואציות בהן יתקבלו החלטות שגויות מכורח קריטריון זה. מהלך זה יעשה על ידי סימולציות מחשב שתתוכנן על פי שלבי הוכחת המשפט, שתדמה את מנגנון השקלול של ה-FAR. אין בכוונתי להציג הוכחה מלאה של המשפט שכן קיים מגוון הוכחות בספרות הרלוונטית,⁹⁹ אלא לעשות שימוש בהוכחות השונות כדי לבחון את נפקויות המשפט ביחס ל-FAR נתון ומוגדר היטב. מטרתה של הסימולציה הינה להראות כי במקרה זה לא זו בלבד שקריטריון דיקטטורי קיים, כפי שמנבאות ההוכחות השונות, אלא אף לצפות את זהותו ואת הסיטואציות בהן ההכרעות שינבעו מכוחו יהיו שגויות.

הסימולציה תתבסס על שיטת ההוכחה של Ning Yu,¹⁰⁰ Geanakoplos¹⁰¹ ו-Barbera¹⁰² ותחולק לשלושה שלבים. השלב הראשון יצביע על קריטריון מועמד לדיקטטורה, מכוחו ישתנה הדירוג R במצבים מסוימים. בשלב השני יוכח שקריטריון זה הוא הקריטריון הדיקטטורי של מנגנון השקלול, ותונח התשתית לאיתורם של "המקרים החשודים בדיקטטורה". בשלב השלישי תאובחן קטגורית "המקרים החשודים בדיקטטורה", ותוצג שיטה לזהות את קבוצת המקרים בהם אכן תתקבל החלטה שגויה.

הסימולציה תדמה מצבים בהם FAR ספציפי, אשר הוגדרו לו מטרה וכן n קריטריונים, יידרש לקבל החלטה. את מנגנון השקלול של ה-FAR הנתון נסמן ב- F_{FAR} ואת הדירוג הסופי שיצור המנגנון נסמן ב-R. אם כן, לפנינו F_{FAR} שמוגדר להחזיר דירוג R עבור כל פרופיל נתון (תנאי I), קובע את הדירוג R ללא תלות באלטרנטיבות לא רלוונטיות (תנאי II) ומכבד קונצנזוס (תנאי III). מטרתנו היא להראות מיהו הקריטריון הדיקטטורי, ולבחון את ההחלטות שינבעו מכוחו.

⁹⁹ Arrow, לעיל ה"ש 4; Geanakoplos, לעיל ה"ש 83; Yu, לעיל ה"ש 90; Salvador Barbera, *Pivotal voters: A new proof of Arrow's theorem*, 6 ECONOMICS LETTERS 13 (1980); Pingzhong Tang & Fangzhen Lin, *Computer-aided proofs of Arrow's and other impossibility theorems*, 173 ARTIFICIAL INTELLIGENCE 1041 (2009); Yasuhito Tanaka, *The Arrow Impossibility Theorem Of Social Choice Theory In An Infinite Society And Limited Principle Of Omniscience*, 8 APPLIED MATHEMATICS E-NOTES 82 (2008)

¹⁰⁰ Yu, לעיל ה"ש 90

¹⁰¹ Geanakoplos, לעיל ה"ש 83

¹⁰² Salvador Barbera, *Pivotal voters: A new proof of Arrow's theorem*, 6 ECONOMICS LETTERS 13 (1980)

שלב ראשון

P_0	C_1	C_2	C_n
	$x > z$	$x > z$	$x > z$	$x > z$
	or	or	or	or
	$z > x$	$z > x$	$z > x$	$z > x$
	y	y	y	y

P_n	C_1	C_2	C_n
	y	y	y	y
	$x > z$	$x > z$	$x > z$	$x > z$
	or	or	or	or
	$z > x$	$z > x$	$z > x$	$z > x$

P_0	C_1	C_2	C_n
	$x > z$	$x > z$	$x > z$	$x > z$
	or	or	or	or
	$z > x$	$z > x$	$z > x$	$z > x$
	y	y	y	y

$$x >_R y$$

P_1	C_1	C_2	C_n
	y	$x > z$	$x > z$	$x > z$
	$x > z$	or	or	or
	or	$z > x$	$z > x$	$z > x$
	$z > x$	y	y	y

$$x >_R y$$

:

:

P_k	C_1	C_2	C_k	C_{k+1}	C_n
	y	y	y	Y	$x > z$	$x > z$	$x > z$
	$x > z$	$x > z$	$x > z$	$x > z$	or	or	or
	or	or	or	or	$z > x$	$z > x$	$z > x$
	$z > x$	$z > x$	$z > x$	$z > x$	y	y	y

$$y >_R x$$

1. נבנה פרופיל P_0 בו כל הקריטריונים

מדרגים את אלטרנטיבה y בתחתית, כך

$$\forall C_i x >_i y \wedge z >_i y$$

1.1 תנאי III $y \Leftarrow R$ בתחתית הדירוג

1.2 בפרט מתקיים $x >_R y$

2. נבנה פרופיל P_n בו כל הקריטריונים

מדרגים את אלטרנטיבה y בראש וכל שאר

הדירוגים נשארים זהים לפרופיל P_0 , כך

$$\forall C_i y >_i x \wedge y >_i z$$

2.1 תנאי III $y \Leftarrow R$ בראש הדירוג

2.2 בפרט מתקיים $y >_R x$

3. כעת נבנה עוד $n - 1$ פרופילים,

$P_1 \dots P_{n-1}$, כאשר השינוי היחיד בין

פרופיל P_i לפרופיל P_{i-1} הוא שינוי דירוגו

של הקריטריון C_i כך ש- y בראש.

3.1 מקומו של פרופיל P_n ועל פי 2.2,

מתחייב כי בין הפרופילים $P_0 \dots P_n$

קיים פרופיל כלשהו P_k , כך שכאשר

הקריטריון C_k משנה את דירוגו

משתנה לראשונה הדירוג R מ-

$$y >_R x \text{ ל- } x >_R y$$

3.2 נסמן קריטריון זה ב- $C_{x,y}$ ואת

מספרו ב- $n_{x,y}$

P'	C_1	C_2	C_{k-1}	C_k	C_{k+1}	C_n
	y	y	y	y	x	x	x	x
	z	z	z	z				
	x	x	x	x	y	y	y	y
					z	z	z	z

$$x >_R y >_R z$$

1. נתבונן בפרופיל P' , שהינו

מקרה פרטי של הפרופיל P_{k-1} .

$$x >_R y \iff \text{II תנאי 1.1}$$

היחס בין x -ל- y נשאר זהה

ליחס בפרופיל P_{k-1} , כיוון

שרק מיקום z השתנה

$$y >_R z \iff \text{III תנאי 1.2}$$

$$x >_R y >_R z \text{ הוא הדירוג הסופי}$$

P''	C_1	C_2	C_{k-1}	C_k	C_{k+1}	C_n
	*	*	*	*				
	y	y	y	y	y	x	x	x
	*	*	*	*		*	*	*
	x	x	x	x	x	y	y	y
					z	*	*	*

$$y >_R x >_R z$$

2. כעת, הסימן * הוא אפשרות

המיקום של z . אם קריטריון

C_k לא ישנה את היחס בין z

לבין y בדירוגו, לא משנה כמה

קריטריונים יחליפו את

הדירוג בין z לבין y , הדירוג

הסופי יישאר $y >_R z$

$$y >_R x \iff \text{II תנאי 2.1}$$

היחס בין x -ל- y נשאר זהה ליחס בפרופיל P_k , כיוון שרק מיקום z השתנה

$$x >_R z \iff \text{II תנאי 2.2}$$

כיוון שהיחס בין x לבין z זהה ליחס זה בפרופיל P'

$$y >_R x >_R z \text{ הוא הדירוג הסופי}$$

$$y >_R z \text{ בפרט מתקיים}$$

כלומר, הקריטריון $C_{x,y}$ שהוגדר ב-3.2 בשלב הראשון, הוא דיקטטור עבור האלטרנטיבות

y, z , כאשר z היא אלטרנטיבה כלשהי.

בניסוח מדויק:

$$\forall z \neq y \neq x, y >_{n_{x,y}} z \implies y >_R z \quad \text{2.5}$$

3. נראה כי $C_{x,y}$ הוא הדיקטטור:

3.1 כפי שהוצג בשלב ראשון, נבנה תהליך החלפה עבור האלטרנטיבות y, z . באותו האופן,

הקריטריון שעקב שינוי דירוגו משתנה לראשונה הדירוג R מ- $z >_R y$ ל- $z >_R y$ יסומן

ב- $C_{y,z}$ ומספרו יסומן ב- $n_{y,z}$.

3.1.1 על פי 2.5, הדירוג הסופי R לא ישתנה ל- $z >_R y$ כל עוד הקריטריון במקום ה-

$$n_{x,y} \text{ מדרג } z >_{n_{x,y}} y \text{ מכאן נובע } n_{y,z} \geq n_{x,y}$$

3.2 כפי שהוצג בשלב הראשון, נבנה תהליך החלפה עבור האלטרנטיבות z, y . באותו האופן, הקריטריון שעקב שינוי דירוגו משתנה לראשונה הדירוג R מ- $y >_R z$ ל- $z >_R y$ יסומן ב- $C_{z,y}$ ומספרו יסומן ב- $n_{z,y}$.

3.2.1 על פי 2.5, הדירוג הסופי R ישתנה ל- $z >_R y$ לא יאוחר מהשלב בו הקריטריון

במקום ה- $n_{x,y}$ ישנה את דירוגו ל- $z >_{n_{x,y}} y$. מכאן נובע $n_{z,y} \leq n_{x,y}$.

3.3 מצירוף 3.1.1 ו-3.2.1 נקבל $n_{y,z} \geq n_{x,y} \geq n_{z,y}$.

3.4 מכיוון שהאלטרנטיבות y ו- z הינן שונות זו מזו ובחירתן היתה שרירותית, מתקיים גם

$$n_{z,y} \geq n_{y,z}$$

3.5 מצירוף 3.3 ו-3.4 נקבל $n_{y,z} = n_{x,y} = n_{z,y}$. כלומר, הקריטריון $C_{x,y}$ הוא דיקטטור עבור

האלטרנטיבות y ו- z . שוויון זה ניתן להרחבה עבור כל $n_{t,s}$, כאשר s ו- t הינן אלטרנטיבות כלשהן.

3.6 על פי 2.5, הקריטריון $C_{x,y}$ הינו דיקטטור עבור כל שתי אלטרנטיבות. מ.ש.ל.

שלב שלישי

לצורך איתור המקרים החשודים בדיקטטורה, נמספר את n הקריטריונים הנתונים בצורה שרירותית, ונרץ את השלב הראשון של הסימולציה על ה- F_{FAR} . כעת אנו יודעים מיהו C_d , וכמו כן יש בידינו את העוגן הראשון למציאת המקרים החשודים – הפרופיל P_k . המקרים החשודים בדיקטטורה הם כל הפרופילים שניתן ליצור מהפרופיל P'' המתאים, על ידי דירוגים שונים של z . קרי, הסיטואציות בהן רק היחס בין x ל- y קבוע בדירוגי כל הקריטריונים, ובנוסף, הקריטריון C_d מדרג את z בתחתית. מכיוון שעיקר ענייננו באלטרנטיבה שתדורג בראש ובה יבחר ה- F_{FAR} , הסימולציה תתמקד בבחינת דירוגן של שלוש האלטרנטיבות העליונות, שכן הן הרלוונטיות גם כאשר בפני ה- FAR עומדות יותר משלוש אלטרנטיבות פעולה אופציונאליות.

P''	C_1	C_2	C_{k-1}	$C_k=C_d$	C_{k+1}	C_n
	*	*	*	*				
	y	y	y	y	y	x	x	x
	*	*	*	*		*	*	*
	x	x	x	x	x	y	y	y
					z	*	*	*

$$y >_R x >_R z$$

בפרט מתקיים: $\forall z \neq y \neq x, y >_{C_d} z \Rightarrow y >_R z$

אם כן, לפנינו קטגוריית פרופילים שיכולים להבנות מהפרופיל P'' , כך שהיחס בין x ל- y קבוע ו- C_d מדרג את z בתחתית. מכיוון ש- z מייצגת כל אלטרנטיבה,

היו פרופילים בהם כל הקריטריונים מלבד C_d ידרגו אותה מעל ל- y , ולמרות זאת, מכיוון שדירוגו של C_d הוא היחיד שמשפיע, תדורג y בראש.

לקיומו ולזיהוי של הקריטריון הדיקטטורי ישנן שתי נפקויות חשובות. ראשית, ישנם קריטריונים אשר לא נרצה שהם יקבלו מעמד דיקטטורי. למשל, אם אחד הקריטריונים שמנחים FAR נתון הוא 'שמירה על שלמותו' בעוד יתר הקריטריונים עוסקים בהגנה על חיי אדם, סביר שלא נרצה כי קריטריון זה יקבל מעמד דיקטטורי. הסימולציה שהוצגה לעיל תאפשר למתכנתים לזהות קריטריון זה מבעוד מועד, ותספק כלי לזיהוי הגורם עליו ראוי להטיל את האחריותיות במקרה בו הקריטריון הדיקטטורי הוביל להחלטה שגויה שגרמה לנזק, כפי שיורחב לקמן.

שנית, לאחר שמצוי בידינו הקריטריון הדיקטטורי של ה- F_{FAR} והפרופיל P'' המאפיין את קטגוריית המקרים החשודים, כל שנותר לעשות הוא להכניס לסימולטור את כל הפרמוטציות של דירוגי הקריטריונים באותם המקרים, ולבחון באיזו סיטואציה תתקבל החלטה שגויה. דוגמא מובהקת למקרים שבהם קיים חשש לקבלת החלטה שגויה, הינה הפרופיל בו כל הקריטריונים למעט C_d מדרגים את האלטרנטיבה z מעל האלטרנטיבה y, ובכל זאת בדירוג R תדורג y בראש.

C_1	C_2	$C_3=C_d$
*	*	
y	y	y
*	*	
x	x	x
		z

P''

נמחיש זאת באמצעות התבוננות על FAR שהוגדרה לו מטרה אותה הוא צריך לבצע תוך שקלול שלושה קריטריונים: C_1 – מזעור פגיעה באזרחים; C_2 – מקסום פגיעה בלוחמים; C_3 – שמירה על שלמותו. נניח שלאחר הרצת השלב הראשון נגלה כי $C_d = C_3$, ונבנה את פרופיל P'' המתאים. בשלב זה נבחן את הסיטואציות בהן ה- F_{FAR} יחליט על אלטרנטיבה שמתעדפת את הקריטריון C_3 גם במחיר פגיעה משמעותית על פי השניים האחרים.

כפי שניתן לראות, בתרחיש שמוצג בטבלה 1 ה- F_{FAR} צפוי לבחור באפשרות בה יהרגו שני אזרחים ושלושה לוחמים, אך הנזק שייגרם ל-FAR יהיה קטן מאוד. זאת, על אף שקיימת אפשרות בה הנזק לחיי אזרחים קטן יותר. גם בטבלה 2 מוצג תרחיש בו ה- F_{FAR} ינקוט באלטרנטיבה שממזערת את הנזק שייגרם ל-FAR, על אף שכל שאר הקריטריונים מדרגים את אלטרנטיבה z בראש. אלו הם המקרים אשר ראוי כי "ידליקו נורה אדומה" אצל המתכנתים לכך שעלולה להתקבל החלטה שגויה. לעומת זאת, טבלאות 3 ו-4 מציגות תרחישים בהם ההחלטה מתקבלת באופן בלעדי על פי C_3 , אולם עדיין ייתכן שההחלטה המתקבלת אינה שגויה.

C_1	C_2	C_3
z		
y	y	y
	z	
x	x	x
		z

טבלה 1
 x – 3 אזרח; לוחם 0.3;
 y – 2 אזרח; לוחם 0.1;
 z – 2 אזרח; לוחם 0.5;

C_1	C_2	C_3
z	z	
y	y	y
x	x	x
		z

טבלה 2
 x – 3 אזרח; לוחם 0.3;
 y – 2 אזרח; לוחם 0.1;
 z – 3 אזרח; לוחם 0.5;

C_1	C_2	C_3
y	y	y
z	z	
x	x	x
		z

טבלה 3
 x – אזרח; לוחם 0.3;
 y – אזרח; לוחם 0.1;
 z – 2 אזרח; לוחם 0.5;

C_1	C_2	C_3
	z	
y	y	y
z		
x	x	x
		z

טבלה 4
 x – 3 אזרח; לוחם 0.3;
 y – אזרח; לוחם 0.1;
 z – 2 אזרח; לוחם 0.5;

6. שלטון החוק והטלת אחריותיות על FAR

בדומה למערכות ממוחשבות שנמצאות בשימוש כבר כיום, לא ניתן יהיה להמנע מטעויות שיעשו FAR, כאשר לטעויות אלו עשויות להיות השלכות משמעותיות על חיי אדם. כפועל יוצא מכך, מערכת אכיפת החוק תדרש להתמודד עם הנזקים שייגרמו, וסוגיית האחריותיות להם. מקובל לדבר על ארבעה מכשולים עיקריים בהטלת אחריותיות לתוצאות שנגרמות מפעולות של מחשבים. שגיאות תוכנה וחומרה (Bugs להלן: באגים) שגורמות לטעויות הן המכשול העיקרי, משום שהן כמעט כורח המציאות במערכת ממוחשבת.¹⁰³ בפרק זה אציג מכשולים אלו ואסביר מדוע הם יתעצמו ככל שתנתן יותר אוטונומיה לרובוטים, ובפרט ל-FAR שנועדו לפעול בשדה הקרב. לבסוף אראה כיצד היכולת לאבחן את הקריטריון הדיקטטורי וקטגוריית המקרים החשודים בדיקטטורה מסייעת בהתמודדות עם מכשולים אלו.

המכשול הראשון קרוי "ידיים מרובות".¹⁰⁴ בתכנון ובתכנות מערכת ממוחשבת ישנם שלבים רבים, החל מדרג מקבלי ההחלטות ועד למהנדסים והמתכנתים בשטח שמיישמים את המדיניות. במקרה של תקלה קשה להצביע על הגורם לה, כיוון שלרוב הסיבה העיקרית לתקלה אינה נפגשת עם מקבל ההחלטה המקומית. ההחלטה מתקבלת אצל הדירקטורים, מועברת הנחיה לעובדים אשר צריכה להיות מתורגמת לשפת מכונה ומוצאת לפועל. הקשר בין התקלה לבין מי שגרם לה מפותל ואינו ברור.

ככל שמוקנית לרובוטים יותר אוטונומיה, כך גדלה מורכבותן של המערכות הממוחשבות שמנחות את פעולותיהם. מורכבות זו מחדדת את הבעייתיות עליה מצביע המכשול הראשון, בשלושה רבדים. ראשית, מערכות תוכנה (System Software) מפותחות על ידי חברות וארגונים שונים להשגת מגוון מטרות, כך שהאדפטציה שלהן למערכת FAR ספציפית פעמים רבות עלולה לבוא על חשבון רמת הדיוק של המערכת. שנית, ישנם מקרים רבים בהם התוכנות עצמן (Application Software) אינן עשויות מקשה אחת, אלא בנויות ממספר מודולים. כיוון שכל מודול מפותח באופן עצמאי, ומכיוון שלעיתים תכופות התוכנות המבוזרות נסמכות אחת על השניה, עלולה להיווצר אי-תאימות במקרים מסוימים.¹⁰⁵ שלישית, יד ביד עם השתכללות התוכנות, מתפתחים גם חלקי החומרה מהם בנויים הרובוטים והופכים למתוחכמים ומורכבים יותר. בגלל התלות בין חומרה לתוכנה, לא תמיד ניתן לאתר את מקור התקלה. לבסוף, קשיים אלו חמורים בפרט כאשר עסקינן ב-FAR שמיועד לפעול בשדה הקרב, שכן סביר כי הוא יתוכנן ויפותח על ידי גופים מדיניים וצבאיים בהם שרשרת קבלת ויישום ההחלטות ארוכה ומפותלת במיוחד.¹⁰⁶

Jifeng Xuan, et al., Debt-prone bugs: technical debt in software maintenance, 4 INTERNATIONAL JOURNAL OF ADVANCEMENT IN COMPUTING TECHNOLOGY 2012A 452, 453 (2012); James Grimmelmann, *Anarchy, Status Updates, and Utopia*, 34 PACE LAW REV. 1, 9 (2014)

¹⁰⁴ Nissenbaum, לעיל הי"ש 3, בעמ' 332-36

¹⁰⁵ DG Johnson & JM Mulvey, *Computer Decisions: Ethical Issues of Responsibility and Bias*, STATISTICS AND OPERATIONS RESEARCH SERIES, PRINCETON UNIVERSITY, SOR-93-11 (1993); Mohammad Asif A Khan, et al., *Eighth International Workshop on Software Clones (IWSC 2014)*, 63 ELECTRONIC COMMUNICATIONS OF THE EASST 18, 18 (2014)

¹⁰⁶ Richard T De George, *Ethical Responsibilities of Engineers in Large Organizations: The*

המכשול השני מכונה "המחשבים כשעיר לעזאזל"¹⁰⁷. לאנשים ישנה נטיה לייחס למחשב "אשמה" בטעויות שנגרמו, הן מכיוון שקל לראות קורלציה ישירה בין המחשב לבין הפעולה שהתבצעה והן בגין יכולותיו לבצע פעולות דוגמת חישוב, בקרה וזיכרון בצורה טובה יותר מבן אנוש. בני אנוש נענשים על מעשים שעשו ונתפסים כאי-עמידה ברף הביצוע המצופה, דוגמת רשלנות רפואית, ולכן אם מחשב ביצע את הפעולה שגרמה לטעות יש להטיל עליו את האשמה ולא על הגורמים האנושיים הרלוונטיים. ייחוס "אשמה" למחשב הינו בעייתי, מכיוון שכאשר מוצאים סיבה לתוצאה, מתגברת המגמה להמעית, או להתעלם, מערכו של הגורם האנושי בפעולתו של המחשב. הדבר חמור במיוחד ביחס לרובוטים אוטונומיים, כיוון שככל שהאוטונומיה שלהם מתרחבת גובר החשש שהמשפט יתייחס אליהם כאל יצורים מעין-אנושיים (Quasi-persons) והם יוכלו להנות מזכויות וחובות, כך שהאחריות על המתכנתים תצטמצם ותתעמעם.¹⁰⁸ קיימת תפיסה לפיה בשנת 2029 רובוטים יטענו (Claim) שהם בעלי מודעות, וטענה זו תהיה מקובלת.¹⁰⁹

המכשול השלישי מכונה בספרות "בעלות ללא אחריות"¹¹⁰. למפתחי המוצר ישנו אינטרס מובנה להגן על זכויותיהם וקניינם הרוחני, לצד מזעור האחריות שלהם במקרה של תקלה. כך, בחוזי מכר רבים מצוין שזכויות הבעלות על התוכנה שייכות למתכנתים, אולם האחריות על נזקים מוטלת על הקונה (Disclaimers). גם במקרים של FAR שנכנסים לשוק, היצרנים מקפידים להזהיר את הלקוחות בחוזי המכר כי מוטלת עליהם האחריות לתוצאות פעולות הרובוט, גם בפעולות שאינן דורשות התערבות אנושית.¹¹¹ ככל שניתנת לרובוט אוטונומיה רבה יותר כך שביכולתו לבצע מגוון רחב יותר של פעולות, כך גדל האינטרס של מפתחי המוצר להסיר מעליהם את האחריות לנזקים שייגרמו בגין פעולותיו. בנוסף ובעיקר, כאשר עסקינן במקרי קיצון דוגמת מלחמה, יש הכרח בגיבוש מסגרת משפטית ברורה להטלת אחריותיות. זאת משום שבניגוד לשוק הפרטי, בו ניתן לעקוף מכשול זה באמצעות חוזים עם חברות ביטוח למיניהן, כאשר המדינה היא הפועל (Agent) לא ניתן יהיה לנקוט באמצעי עקיפה זה.

המכשול הרביעי והחשוב ביותר, לענייננו, הוא קיומם של באגים. באגים הם תקלות בפעולת המחשב אשר מקורן בתוכנה או בחומרה.¹¹² מקובל לסבור כי באגים הם כורח המציאות, ואפילו המתכנתים הטובים ביותר לא יכולים להמנע מהם.¹¹³ נוכח הכרחיותם של באגים, קיימת נטיה להתייחס לתקלות הנובעות מקיומם כאל נזקים בלתי נמנעים, ולהסיר אחריות מהמתכנים

Pinto Case, 1 BUSINESS & PROFESSIONAL ETHICS JOURNAL 1, 1-14 (1981); John Ladd, Computers and moral responsibility: A framework for an ethical analysis (Academic Press Professional, Inc. 1991)

¹⁰⁷ Nissenbaum, לעיל ה"ש 3, בעמ' 338-39

¹⁰⁸ Autonomous Military Robotics, לעיל ה"ש 1, בעמ' 55

JOHN FRANK WEAVER, ROBOTS ARE PEOPLE TOO (2014)

RAY KURZWEIL, THE AGE OF SPIRITUAL MACHINES: WHEN COMPUTERS EXCEED HUMAN INTELLIGENCE (Penguin, 2000)

¹¹⁰ Nissenbaum, לעיל ה"ש 3, בעמ' 339-40

Bryant Walker Smith, *Human Factors in Robotic Torts*, 102 GEO. L. J. 1, 11-12, 14, FORTHCOMING 2014 (2014)

¹¹² Nissenbaum, לעיל ה"ש 3, בעמ' 336-38

¹¹³ Ying-Dar Lin, et al., *Bug traces: identifying and downsizing packet traces with failures triggered in networking devices*, 52 COMMUNICATIONS MAGAZINE, IEEE 112, 112 (2014); Shivani Rao, *Mining Software Repositories for IR based Bug Localization*, 1 (2013); David Lorge Parnas, *Software aspects of strategic defense systems*, 28 COMMUNICATIONS OF THE ACM 1326, 1327 (1985)

והמהנדסים לשלון המערכות הממוחשבות.¹¹⁴ כלומר, אם באגים הם תוצר לוואי בלתי נמנע של תכנות, אזי נשפט נזקים שנובעים מבאגים כ"מצערים אך בלתי נמנעים", כיוון שהם תוצאות של טכנולוגיה חדשה וטובה יותר ולהם אף אחד אנו אחריותי.

ככל שניתנת לרובוטים יותר אוטונומיות, הקוד הממוחשב שמנחה את פעולותיהם, כמו גם חלקי החומרה בהם הם מצוידים, הופכים מורכבים ומסועפים יותר. כפועל יוצא מכך, קשה יותר לאתר את הבאג שגרם לתקלה ואת האחראי לו. בנוסף, הרובוטים האוטונומיים יביאו תועלת רבה יותר, כך שערכם בשדה הקרב ייגבר בצורה משמעותית. כתוצאה מכך, גובר החשש כי באיזון הנזקים והתועלות תתקבע התודעה שמחיר הנזקים שנגרמים מבאגים הינו מחיר ראוי לשלם עבור התועלות שמניב השימוש ב-FAR. תודעה זו אף תקבל משנה תוקף, נוכח הנטיה ליחס "אשמה" ל-FAR כייצור מעין-אנושי והקושי לאתר את הגורם לתקלה.

באגים – סיכונים ועלויות

טרם שיוצג כיצד יישומו של משפט ארו כאמור מסייע בהתגברות על ארבעת המכשולים הניצבים בפני הטלת אחריותיות, חשוב להבחין בהבדל שבין שגיאת תכנות שניתן יהיה להצביע על האחריות לה, לבין שגיאה שתגרום לנזק אשר תהיה בעיה לייחס לו אחריותיות. ככל שיותר משתנים צפויים מראש וככל שניתן להצביע באופן מדויק יותר על הגורם לשגיאה, כך ניתן לייחס לו יותר אחריותיות. כפועל יוצא מכך, ישנה אבחנה בין באגים טבעיים (Natural Hazards), אשר יתקיימו למרות כל המאמצים לאתר ולמנוע אותם, לבין באגים שניתנים למניעה. תחת ההנחה שקיימים באגים טבעיים להם נתקשה ליחס אחריותיות, התכשורת לאבחנה זו כמוה כטענה שתחום המחשבים עדיין אינו מוכן להיות חלק מחיינו.¹¹⁵ האבחנה בין הטלת אחריותיות על באגים שניתנים לאיתור ולמניעה לבין הטלתה על אלו שאינם, נובעת מההבדל שבין נטילת סיכון בהפעלתו של מכשיר לבין עלות התפעול שלו. נחדד אבחנה זו באמצעות הגדרת המושגים הרלוונטיים, ונבחן את השלכותיה.

למושג סיכון (Risk) יש הגדרות ופרשנויות שונות, כאשר המכנה המשותף שלהן הוא האבחנה בין מציאות (Reality) לבין האפשרות שמציאות מסוימת תתרחש (Possibility). בעולם דטרמיניסטי לחלוטין ובהינתן מלוא האינפורמציה, המושג סיכון חסר משמעות.¹¹⁶ הווה אומר, סיכון הוא האפשרות, הלא-וודאית, שיתרחש נזק.¹¹⁷ כך למשל, בתחומים שונים בתעשייה, ניתוחי סיכון עלות-תועלת (Risk Cost-Benefit Analysis) בתפעולן של מגוון מכוונות מתבצעים באמצעות חישוב

¹¹⁴ Nissenbaum, לעיל הי"ש 3, בעמ' 336

¹¹⁵ Nissenbaum, לעיל הי"ש 3, בעמ' 338

¹¹⁶ Andreas Klinke & Ortwin Renn, *Precautionary principle and discursive strategies: classifying and managing risks*, 4 JOURNAL OF RISK RESEARCH 159, 159 (2001)

¹¹⁷ J.X. Kasperson, *Nuclear Risk Analysis in Comparative Perspective*, 20 RELIABILITY ENGINEERING AND SYSTEM SAFETY (1988); Andreas Klinke & Ortwin Renn, *Precautionary principle and discursive strategies: classifying and managing risks*, 4 JOURNAL OF RISK RESEARCH 159, 160 (2001); Chauncey Starr & Chris Whipple, *The strategic defense initiative and nuclear proliferation from a risk analysis perspective*, in RISK, ORGANIZATIONS, AND SOCIETY 53 (1991); James F Short, *Defining, explaining and managing risk*, ORGANIZATIONS, UNCERTAINTIES, AND RISK 39-51 (1992)

ההסתברות לסיכון אל מול עלות מניעתו וצמצומו. לדוגמא, אם מכונה צפויה להרוג אדם אחד מתוך 10,000 איש כל שנה, וישנם 1000 עובדים במפעל, אזי ישנו צפי למוות של עובד אחד כל 10 שנים, או 0.1% מוות לשנה.¹¹⁸ סיכון זה יש לשקלל אל מול התועלת שתביא המכונה (Risk-Utility Test),¹¹⁹ וכך תתקבל החלטה האם לעשות בה שימוש.

לצד הסיכון הכרוך בשימוש במערכות מסוימות, להפעלתן ישנן גם עלויות שונות. עלות (Cost) הינה המחיר שמשולם כדי להשיג מטרה מסוימת.¹²⁰ כך למשל, לצורך הפעלת מערכת נדרשת השקעה כספית ראשונית, זמן לפיתוח ועלות תחזוקה שוטפת של המערכת. בעולם טכנולוגיית המידע המושג המקובל לעלויות אלו הוא 'עלות הבעלות הכוללת' (Total Cost of Ownership), הכוללת את העלויות הישירות והעקיפות של רכישה של חומרה או תוכנה, ועלויות הפעלתן.¹²¹ קרי, בשלב קבלת ההחלטה האם לעשות שימוש במערכת מסוימת, עלינו לשקול האם העלויות הצפויות, בכללן עלויות הרכישה, הכשרת המתפעלים ועלויות התפעול (Operating Cost), כדאיות לכשעצמן.¹²²

מכאן, המאפיין העיקרי של עלויות התפעול, שמבחין אותן מהסיכון הכרוך בהפעלת המכונה, הוא רמת הצפיות והוודאות שלהן עוד בשלב קבלת ההחלטה להפעיל אותה. ככל שהמשתנים צפויים יותר וישנה וודאות שיתרחש נזק במצבים מוגדרים היטב, נגיד שזוהי עלות שאמורה להלקח בחשבון, ולא סיכון מחושב. כך למשל, אם המתכנתים שלחו FAR לבצע משימה במקום לא ידוע מראש, נגיד שהם לקחו את הסיכון שיהיה לו מספיק דלק להגיע למקום ביצוע המשימה, לבצע אותה, ולחזור. לעומת זאת, אם ידוע שהגעה למקום ביצוע המשימה תדרוש מיכל דלק מלא, העובדה שה-FAR לא יחזור היא חלק מעלות ביצוע המשימה.

לאבחנה בין סיכון לעלות יש נפקות לגבי השאלה למי נייחס את האחריות לפעולות המכונה. אם זו עלות, הרי שבהגדרה, מקבלי ההחלטה על הוצאת הפרויקט לפועל נושאים באחריות. כך למשל, כאשר מכשיר לא מיועד לעבוד בתנאים מסוימים ובכל זאת נעשה בו שימוש בתנאים אלו, האחריות מוטלת על מקבל ההחלטה לעשות בו שימוש באותם התנאים. האחריות על התפוצצות מעבורת החלל Challenger, שנבעה משימוש בה בתנאי מזג אוויר החורגים מאלו בהם היא מסוגלת לפעול, הוטלה על דירקטורי NASA כיוון שהם ידעו שהיא אינה ראויה לשימוש במזג אוויר כזה ובכל זאת שלחו אותה למשימה.¹²³

לעומת זאת, כאשר עסקינן בסיכון, הטלת האחריות תלויה בגורם הסיכון. כלומר, הגורם שיצר את הסיכון הוא משתנה נפרד מחישוב של מידת הנזק וההסתברות שהוא ייגרם. אין דינו של סיכון

Jonathan Wolff, *Risk, fear, blame, shame and the regulation of public safety*, 22 ECONOMICS AND PHILOSOPHY 409, 412 (2006)¹¹⁸

Aaron D. Twerski & James A. Henderson Jr., *Manufacturers' liability for defective product designs: The triumph of risk-utility*, 74 BROOK. L. REV. 1061, 1065 (2008)¹¹⁹

"cost" Oxford English Dictionaries, Oxford University Press.¹²⁰

<http://www.oxforddictionaries.com/definition/english/cost>

L. MIERITZ, B. KIRWIN, *DEFINING GARTNER TOTAL COST OF OWNERSHIP* (2005); Barbara Russo, et al., *Defining the Total Cost of Ownership for the Transition to Open Source Systems*, 2 (2005)¹²¹

D Cappucio, et al., *Total Cost of Ownership: The Impact of System Management Tools*, GARTNER GROUP, STAMFORD, CT (1996)¹²²

Nissenbaum¹²³, לעיל ה"ש 3, בעמ' 336-37

שנובע מכוונת זדון כדין סיכון שנובע מרשלנות או כדין סיכון שנובע מאסון טבע בלתי נמנע. לדוגמא, אחריותיות תיוחס בצורה שונה במקרה של תחזוקה לקויה ורשלנית של רכב אשר גרמה לתאונה, לעומת מקרה של צמיג מכונית שהתפוצץ שלא כתוצאה מרשלנות וגרם לתאונה, גם אם ההסתברות לתאונה והנזק בשני המקרים זהים.¹²⁴

לסיכום נקודה זו, ככל שנתונה יותר אינפורמציה על התועלות והנזקים שייגרמו מפעולת מכונה בשלב ההחלטה על הפעלתה, נגיד שהנזקים שנגרמו הם עלויות הפעלתה. עלויות אלו ראויות להלקח בחשבון מבעוד מועד, ועליהן מקבלי החלטה לעשות שימוש במכונה אחריותיים. לעומת זאת, ככל שישנם יותר משתנים לא צפויים בעת קבלת החלטה על הפעלת המכונה, נייחס את האחריותיות על הנזק לגורם הסיכון הרלוונטי, במידה ונוכל להצביע עליו. ומן הכלל אל הפרט: שגיאות שינבעו מבאגים טבעיים, כאלו שיתקיימו על אף כל המאמצים לאתר ולמנוע אותם, ראויות להיות מסווגות כהתממשות של סיכון שנלקח בשלב הפעלת המערכת. לעומתן, ככל שנתונה יותר אינפורמציה אודות הבאגים כך שניתן לאתר ולמנוע אותם עוד בשלב התכנון והתכנות, ההחלטה להפעיל את המערכת למרות הידיעה על קיומם כוללת את עלות הנזקים שייגרמו מבאגים אלו. מכאן, כאשר לפנינו שגיאה שנגרמה מבאג שסווג כעלות התפעול של המערכת, האחריותיות עליה תיוחס למקבל החלטה על כדאיות הפרויקט על אף עלות זו. לעניין ה-FAR דן, אחריותיות זו הינה חלק בלתי נפרד מהאחריות על כלל החלטות המוסריות והמטא-מוסריות הגלומות בתכנון ותכנות ה-FAR, שתיוחס למקבלי החלטות הרלוונטיות כאמור.

אחריותיות להחלטות FAR במקרים דיקטטוריים

בתת פרק זה אבחן כיצד זיהוי הקריטריון הדיקטטורי מאפשר להתגבר על המכשולים שהוצגו, לזיהוי הנושא באחריותיות לשגיאות ולנזקים שינבעו כתוצאה מהיווצרותו של קריטריון זה. כפי שהוכח לעיל, הדיקטטוריות של הקריטריון איננה תוצאה מכוונת הרצויה על ידי המתכנתים, אלא היא כורח המציאות במנגנוני שקלול העומדים בדרישות משפט ארו. לפיכך, ניתן לסווג את החלטות השגויות שיתקבלו מכוחו של הקריטריון הדיקטטורי כבאגים שנפלו במהלך התכנות. לאור האבחנות שבין סיכון לבין עלות, ובין באגים שניתן לאתר ולמנוע אותם לבין באגים טבעיים, ובעזרת הסימולציה שהוצגה המאפשרת לזהות את הקריטריון הדיקטטורי, ניתן להתמודד עם ארבעת המכשולים שהוצגו.

המכשול הראשון כונה "ידיים מרובות", והתמקד בקושי להצביע על האחראי לקבלת החלטה הספציפית שגרמה לתקלה. במקרה של החלטות שגויות הנובעות מקיומו של קריטריון דיקטטורי, האחראי לתקלה יהיה הגורם שקבע את מנגנון השקלול, הריץ את הסימולציה ואישר את השימוש במנגנון השקלול על אף קיומו של הקריטריון הדיקטטורי. מאחר וקיומה של הסימולציה מאפשר לזהות בוודאות את הקריטריון הדיקטטורי ולצפות את הסיטואציות בהן הוא יבוא לידי ביטוי,

Jonathan Wolff, *Risk, fear, blame, shame and the regulation of public safety*, 22 *ECONOMICS AND PHILOSOPHY* 409, 424-45 (2006) ¹²⁴

הרי שהתקלה שנגרמה מההחלטה השגויה של ה-FAR הינה עלות תפעולו ולא סיכון מחושב. לכן גם סביר לדרוש כי הגורם הרלוונטי יריץ את הסימולציה בטרם יאשר את השימוש ב-FAR.

ביחס למכשול השני הנוגע לנטיה לראות במחשבים כ"שעיר לעזאזל", הרי שברור כי במקרה של ה-FAR דן, ההחלטות המתקבלות תלויות אך ורק במנגנון השקלול שהוגדר מראש על ידי גורם אנושי, ובמובן זה הינן דטרמיניסטיות. מכאן, שראוי להטיל את האחריות על אותו גורם אנושי, ולא להתחבא מאחורי האוטונומיות של ה-FAR.

המכשול השלישי שהתמקד באפשרות להתנער מאחריות באמצעות מנגנונים חוזיים, פחות רלוונטי ל-FAR שיפעל בשדה הקרב, שכן במקרים אלו המדינה היא הפועל (Agent). עם זאת, אפילו לו היה ניתן להשתמש במנגנונים חוזיים כדי להעביר את האחריות ללקוח, ברור כי ככל שרובוט הוא יותר אוטונומי, ומשום שהחלטותיו דטרמיניסטיות, כך פחות הוגן להטיל אחריות על הלקוח שכן אין לו שליטה על פעולות ה-FAR. בפרט, במקרים של שגיאות הנובעות מקיומו של קריטריון דיקטטורי, ראוי להטיל את האחריות על קובעי מנגנון השקלול.

לבסוף, נפנה למכשול הרביעי, הבאגים. כפי שצוין לעיל, זהו המכשול הבעייתי ביותר. ואולם, משעה שנתונה בידינו נוסחה המאפשרת לאתר בוודאות את הקריטריון הדיקטטורי ולזהות את המקרים בהם יבוא הוא לידי ביטוי, הרי שהנזקים שייגרמו מכוחו ראויים להיות מסווגים כחלק מעלויות התפעול של ה-FAR ולא כבאגים טבעיים שלא ניתן לייחס להם אחריות. עבור כל FAR נתון שמנגנון השקלול שלו עומד בתנאי משפט ארו, קיימת וודאות שתקבל החלטה שגויה במצבים מסוימים, וזוהי תוצאה אינהרנטית למנגנון השקלול שייבחר. תודות לסימולציה שהוצגה, מצבים אלו צפויים וניתנים לשליטה, באמצעות שינוי הקריטריון הדיקטטורי או בחירת מנגנון שקלול שונה. מכאן, שניתן לזהות גורם אחריותי לנזקים שינבעו כתוצאה מבאגים אלו.

סיכום

המגמה להענקת אוטונומיה מלאה לרובוטים ניכרת במגוון טכנולוגיות מתקדמות המצויות בתעשיות הצבאיות והאזרחיות, ונראה כי השימוש ב-FAR בשדה הקרב בלתי נמנע. לצד האתגרים הפילוסופיים-משפטיים שעומדים בפני קובעי אמת המידה והקריטריונים לפיהם יתוכנת ה-FAR לפעול, המגבלות הטכנולוגיות-לוגיות של המערכות הממוחשבות אשר ינחו את פעולותיו ראויות לבחינה טכנולוגית ומשפטית מיוחדת. בנוסף לסיכון שכרוך בשימוש במערכת ממוחשבת עקב קיומם של באגים טבעיים, תכנון ותכנות ה-FAR מחייבים את המתכנתים להידרש למבנה הלוגי של מנגנון קבלת ההחלטות בו ייעשה שימוש.

כפי שהוכח ברשימה זו, שימוש במנגנון קבלת החלטות שעומד בתנאי משפט ארו, יוביל בהכרח לקבלת החלטה שגויה במקרים מסוימים. החלטה זו תנבע מהיווצרותו של קריטריון דיקטטורי אותו ניתן לצפות מראש, וכך גם את המקרים בהם תתקבל החלטה שגויה מכורח קריטריון זה. כיוון שכך, טרם שיוחלט על צריבת מנגנון כזה ל-FAR שיפעל בעולם האמיתי, ראוי לעמוד על זהותו של הקריטריון הדיקטטורי ולבחון בפירוט את המקרים בהם תתקבל החלטה שגויה, תוך שקלול

עלות הנזקים שייגרמו במקרים אלו בין מכלול השיקולים הרלוונטיים בקבלת ההחלטה האם לעשות שימוש ב-FAR.

ההתפתחויות הטכנולוגיות מציבות דילמות ואתגרים חדשים בפני קובעי המדיניות ומערכת אכיפת החוק. הדברים האמורים מקבלים משנה תוקף כאשר עסקינן בטכנולוגיה שמטרתה המוצהרת הינה צמצום למינימום האפשרי את מעורבותו של גורם אנושי בפעולתן של מערכות ממוחשבות בשדה הקרב. הילת האוטונומיות שאופפת את ה-FAR המתקדמים, התועלת הרבה שהם צפויים להביא בשדה הקרב וקיומם של באגים טבעיים שיגרמו לשגיאות מצערות להן נתקשה לייחס אחריותיות – מחדדים את הקשיים באסדרת התחום המתפתח. ניתוח קטגוריית מנגנוני קבלת ההחלטות שהוצג ברשימה זו והסימולציה שביססה את הדטרמיניסטייות והצפיות של באגים מסוימים, נועדו כדי לתרום לשיח האקדמי והמדיני בגיבושה של מסגרת נורמטיבית לאסדרת השימוש ב-FAR בשדה הקרב, תוך שילוב הכלים המתמטיים-לוגיים שבידנו עם ההבנה הפילוסופית-משפטית שלנו להתמודדות עם אתגרי העתיד.

ביבליוגרפיה

ספרות ופסיקה

מיכאל וולצר *מלחמות צודקות ובלתי צודקות* (ספריית אופקים, עם עובד, 1977)
דניאל סטטמן *דילמות מוסריות* (הוצאת מאגנס אוניברסיטה העברית, 1991)
Brouse v. U.S., 83 F. Supp. 373, 374 (N.D. Ohio 1949)

ספרות באנגלית

Kenneth J. Arrow, *Social choice and individual values* (Yale university press, 2012)
Isaac Asimov, *I, Robot (2004 edition)*, NEW YORK, NY: BANTAM DELL (1950)
Adedeji B. Badiru and Lee Ann Racz, *HANDBOOK OF EMERGENCY RESPONSE: A HUMAN FACTORS AND SYSTEMS ENGINEERING APPROACH* (CRC Press., 2013)
JOHN BOYD, *A DISCOURSE ON WINNING AND LOSING* (1987)
SAMIR CHORPA & LAURENCE F. WHITE, *A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS* (University of Michigan Press., 2011)
ROBERT CORAM, *BOYD: THE FIGHTER PILOT WHO CHANGES THE ART OF WAR* (2002)
RICHARDS CHET, *CERTAIN TO WIN: THE STRATEGY OF JOHN BOSYD, APPLIED TO BUSINESS* (2004)
Emanuel Diamant, *Cognitive Robotics: For Never was a Story of More Woe than This* (2014)
GERALD DWORIN, *THE THEORY AND PRACTICE OF AUTONOMY 7* (Cambridge University Press., 1988)
GRANT T. HAMMOND, *THE MINOR OF WAR: JOHN BOYD AND AMERICAN SECURITY* (Lorraine Atherton ed., 2001)
JEAN-MARIE HENCKAERTS, et al., *CUSTOMARY INTERNATIONAL HUMANITARIAN LAW: RULES* (Cambridge University Press., 2005)
NIDHI KLARA at el., *LIABILITY AND REGULATION OF AUTONOMOUS VEHICLE TECHNOLOGIES* (California PATH Program, Institute of Transportation Studies, University of California at Berkely, 2009)
RAY KURZWEIL, *THE AGE OF SPIRITUAL MACHINES: WHEN COMPUTERS EXCEED HUMAN INTELLIGENCE* (Penguin, 2000)
L. MIERITZ, B. KIRWIN, *DEFINING GARTNER TOTAL COST OF OWNERSHIP* (2005)
FRANS P.B. OSIGNA, *SCIENCE, STRATEGY AND WAR: THE STRATEGIC THEORY OF JOHN BOYD* (2006)

NIGEL S. RODLEY, *THE TREATMENT OF PRISONERS UNDER DEVELOPMENT OF THE INTERNATIONAL LAW* (2nd ed., Oxford University Press, 1999)

PETER W. SINGER, *WIRED FOR WAR: THE ROBOTICS REVOLUTION AND CONFLICT IN THE TWENTY-FIRST CENTURY* (Penguin press, 2009)

Walter F. Truskowski et al., *AUTONOMOUS AND AUTONOMIC SYSTEMS: WITH APPLICATIONS TO NASA INTELLIGENT SPACECRAFT OPERATIONS AND EXPLORATION SYSTEM* (2009)

AN ENCYCLOPEDIA OF WAR AND ETHICS (D. Wells Ed., Greenwood Press., 1996)

M. WALZER, *JUST AND UNJUST WARS* (4th Ed. Basic Books, 1977)

JOHN FRANK WEAVER, *ROBOTS ARE PEOPLE TOO* (2014)

LING XU & JIAN-BO YANG, *INTRODUCTION TO MULTI-CRITERIA DECISION MAKING AND THE EVIDENTIAL REASONING APPROACH 4* Manchester School of Management, University of Manchester Institute of Science and Technology (2001)

מאמרים

גבי סיבוני ויוני אשפר "דילמות בהפעלת אמצעי לחימה אוטונומיים", 16 *עדכן אסטרטגי*, 71 (2014)

Thomas K. Adams, *Future Warfare and the Decline of Human Decisionmaking*, *PARAMETERS* 57 (2001)

Graham T. Allison, *Conceptual Modes and the Cuban Missile Crisis*, 63 *THE AMERICAN POLITICAL SCIENCE REV.* 689 (1969)

Kenneth Anderson, et al., *Adapting the Law of Armed Conflict to Autonomous Weapon Systems*, *INTERNATIONAL LAW STUDIES UNITED STATES NAVAL WAR COLLEGE*, FORTHCOMING 2014 (2014)

Kenneth Anderson and Matthew Waxman, *Law and Ethics for Autonomous Weapon System: Why a Ban Won't Work and How the Laws of War Can*, *HOOVER INST. MONOGRAPH* 2 (2013)

Ronald C. Arkin, *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture* (2011)

Kenneth J. Arrow, *A Difficulty in the Concept of Social Welfare*, 58 *THE JOURNAL OF POLITICAL ECONOMY* 328 (1950)

Yisrael Aumman, *Economic Theory and Mathematical Method: An Interview*, in *ARROW AND THE ASCENT OF MODERN ECONOMY THEORY* 136 (G.R. Feiwel ed., 1987)

Lisanne Bainbridge, *Ironies of Automation*, 19 AUTOMATICA 775 (1983)

Salvador Barbera, *Pivotal voters: A new proof of Arrow's theorem*, 6 ECONOMICS LETTERS 13 (1980)

William Boothby, *Some Legal Challenges Posed by Remote Attack*, 94 INTERNATIONAL REV. OF THE RED CROSS 579 (2012)

R. Brandt, *Utilitarianism and the Rules of War*, 1 PHILOSOPHY AND PUBLIC AFFAIRS 145 (1972)

D Cappuccio, et al., *Total Cost of Ownership: The Impact of System Management Tools*, GARTNER GROUP, STAMFORD, CT (1996)

C. Cloos, *The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism* 38 AAAI Technical Report FS-05-06 (2005)

Gilles Coppin & François Legras, *Autonomy Spectrum and Performance Perception Issues in Swarm Supervisory Control*, 100 PROCEEDINGS OF THE IEEE 590 (2012)

O. Grant, Clark R. Kok and R. Lacroix., *Mind and Autonomy in Engineered Biosystems*, 12 ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE 389 (1999)

Anne Edland, *Attractiveness Judgments of Decision Alternatives Under Time Stress*, 21 COGNITION AND DECISION RESEARCH UNIT (1985)

Anne Edland, *International Governance of Autonomous Military Robots*, 12 COLUM. SCI. & TECH L. REV. 272 (2011)

M Al Fahdi, et al., *Towards An Automated Forensic Examiner (AFE) Based Upon Criminal Profiling & Artificial Intelligence* (2013)

V. Ferraris, et al. *Defining Profiling* (2013)

Philippa Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, 5 OXFORD REV. 1 (1967)

John Geanakoplos, *Three brief proofs of Arrow's impossibility theory*, 26 ECONOMIC THEORY 211 (2005)

Richard T De George, *Ethical Responsibilities of Engineers in Large Organizations: The Pinto Case*, 1 BUSINESS & PROFESSIONAL ETHICS JOURNAL 1 (1981)

Allan Gibbard, *Manipulation of voting schemes: a general result*, 41 JOURNAL OF THE ECONOMETRIC SOCIETY 587 (1973)

Parke Godfrey, et al., *Algorithms and analyses for maximal vector computation*, 16 THE VLDB JOURNAL – THE INTERNATIONAL JOURNAL ON VERY LARGE DATA BASES 5 (2007)

Noah J. Goodall, *Ethical Decision Making During Automated Vehicle Crashes* (2014)

James Grimmelman, *Anarchy, Status Updates, and Utopia*, 34 PACE LAW REV. 1 (2014)

Shane Harris, *Out of the Loop: The Human-free Future of Unmanned Aerial Vehicles*, in EMERGING THREATS IN NATIONAL SECURITY AND LAW (Peter Berkowitz ed., 2012)

Thomas Hurka, *Proportionality in the Morality of War*, 33 PHILOSOPHY AND PUBLIC AFFAIRS 34 (2005)

Chris Jenks, *Law from Above: Unmanned Aerial Systems, Use of Force and the Law of Armed Conflict*, 85 N.D. L. REV. 649 (2009)

DG Johnson & JM Mulvey, *Computer Decisions: Ethical Issues of Responsibility and Bias*, STATISTICS AND OPERATIONS RESEARCH SERIES, PRINCETON UNIVERSITY, SOR-93-11 (1993)

Immanuel Kant, *Metaphysische Anfangsgründe der Rechtslehre*, in 6 KANTS GESAMMELTE SCHRIFTEN 230 (Preußische Akademie der Wissenschaften ed., 1902–1923)

J.X. Kaspersen, *Nuclear Risk Analysis in Comparative Perspective*, 20 RELIABILITY ENGINEERING AND SYSTEM SAFETY (1988)

Mohammad Asif A Khan, et al., *Eighth International Workshop on Software Clones (IWSC 2014)*, 63 ELECTRONIC COMMUNICATIONS OF THE EASST 18 (2014)

Andreas Klinke & Ortwin Renn, *Precautionary principle and discursive strategies: classifying and managing risks*, 4 JOURNAL OF RISK RESEARCH 159 (2001)

SEMIH KORAY, SELF-SELECTIVE SOCIAL CHOICE FUNCTIONS VERIFY ARROW AND GIBBARD-SATTERTHWAITE THEORMS, 68 ECONOMETRICA 981 (2000)

Francis Myrna Kamm, *Harming Some to Save Others*, 57 PHILOSOPHICAL STUDIES 227 (1989)

John Ladd, *Computers and moral responsibility: A framework for an ethical analysis* (Academic Press Professional, Inc., 1991)

Patrick Lin, et al. *Autonomous Military Robotics: Risk Ethics and Design* (2008)

Ying-Dar Lin, et al., *Bug traces: identifying and downsizing packet traces with failures triggered in networking devices*, 52 COMMUNICATIONS MAGAZINE, IEEE 112 (2014)

Teemu Mätäsniemi and Valtion teknillinen tutkimuskeskus, *Operational Decision Making in the Process Industry: Multidisciplinary approach*, 2442 VTT TIEDOTTEITA 107 (2008)

William C Marra and Sonia K. McNeil, *Understanding "The Loop": Regulating the Next Generation of War Machines*, 36 HARV. J. L. & PUB. POL'Y 1139 (2012)

Bruce M. McLaren, *Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions*, 21 INTELLIGENT SYSTEMS, IEEE 2 (2006)

Bruce M. McLaren, *Extensionally Defining Principles and Case in Ethics: An AI Model*, 150 ARTIFICIAL INTELLIGENCE, 145 (2003)

P. Moubarak, P. Ben-Tzvi, Adaptive Manipulation of a Hybrid Mechanism Mobile Robot 113-118, IEEE (2011)

Helen Nissenbaum, *Accountability in a Computerized Society*, 2 SCIENCE AND ENGINEERING ETHICS, 25 (1996)

David Lorge Parnas, Software aspects of strategic defense systems, 28 COMMUNICATIONS OF THE ACM 1326 (1985)

Raja Parasuraman et al., *A Model for Types and Levels of Human Interaction with Automation*, 30 IEEE TRANSACTION ON 286 (2000)

Thomas M. Powers, *Deontological Machine Ethics* (2005)

Shivani Rao, *Mining Software Repositories for IR based Bug Localization* (2013)

Russell W. Robbins & William A. Wallace, *Decision Support for Ethical Problem Solving: A Multi-Agent Approach*, 43 DECISION SUPPORT SYSTEMS 1571 (2007)

Barbara Russo, et al., *Defining the Total Cost of Ownership for the Transition to Open Source Systems* (2005)

Scott D. Sagan, *Rules of Engagement*, in AVOIDING WAR: PROBLEMS OF CRISIS MANAGEMENT 443 (Alexander L. George ed., 1991)

Mark Allen Satterthwaite, *Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions*, 10 JOURNAL OF ECONOMIC THEORY (1975)

Michael N. Schmitt, *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics*, 13 HARVARD NATIONAL SECURITY JOURNAL FEATURE (2013)

James F Short, *Defining, explaining and managing risk*, ORGANIZATIONS, UNCERTAINTIES, AND RISK (1992)

Carl Shulman, et al., *Which Consequentialism? Machine Ethics and Moral Divergence*, ROYNOLDS AND CASSINELLI 23 (2009)

Bryant Walker Smith, *Human Factors in Robotic Torts*, 102 GEO. L. J., FORTHCOMING 2014 (2014)

- R. Sparrow, *Building a Better WarBot: Ethical Issues in the Design of Unmanned Systems for Military Applications*, 15 SCIENCE & ENGINEERING ETHICS, 169 (2009)
- Chauncey Starr & Chris Whipple, *The strategic defense initiative and nuclear proliferation from a risk analysis perspective*, in RISK, ORGANIZATIONS, AND SOCIETY 53 (1991)
- Judith Jarvis Thomson, *The Trolley Problem*, 94 YALE L. J. 1395 (1985)
- Judith Jarvis Thomson, *Killing, Letting Die, and the Trolley Problem*, 59 THE MONTIS 204-17 (1976)
- Yasuhito Tanaka, *The Arrow Impossibility Theorem Of Social Choice Theory In An Infinite Society And Limited Principle Of Omniscience*, 8 APPLIED MATHEMATICS E-NOTES 82 (2008)
- Pingzhong Tang & Fangzhen Lin, *Computer-aided proofs of Arrow's and other impossibility theorems*, 173 ARTIFICIAL INTELLIGENCE 1041 (2009)
- Bogdan Tomoiagă, et al., *Pareto optimal reconfiguration of power distribution systems using a genetic algorithm based on NSGA-II*, 6 ENERGIES 1439 (2013)
- Aaron D. Twerski & James A. Henderson Jr., *Manufacturers' liability for defective product designs: The triumph of risk-utility*, 74 BROOK. L. REV. 1061 (2008)
- Ryan J. Vogel, *Drone Warfare and the Law of Armed Conflict*, 39 DENV. J. INT'L L. & POL'Y 101 (2010)
- Markus Wagner, *Taking Humans Out of the Loop: Implications for International Humanitarian Law*, 21 J.L. INFO. & SCI. 1 (2011)
- Jonathan Wolff, *Risk, fear, blame, shame and the regulation of public safety*, 22 ECONOMICS AND PHILOSOPHY 409 (2006)
- Jifeng Xuan, et al., *Debt-prone bugs: technical debt in software maintenance*, 4 INTERNATIONAL JOURNAL OF ADVANCEMENT IN COMPUTING TECHNOLOGY 2012A 453 (2012)
- Ning Neil Yu, *A one-shot proof of Arrow's impossibility theorem*, 50 ECONOMIC THEORY 523 (2012)

ועדות בינלאומיות והנחיות מנהליות

ICRC, *Expert Meeting on Autonomous weapon systems: technical, military, legal and humanitarian aspects*, March 2014.

ICRC, *Autonomous weapons: States must address major humanitarian, ethical challenges*, September 2013.

U.N. General Assembly, *Extrajudicial, Summary or Arbitrary Executions (A/61/311)* (5 September 2006).

Chairman of the Joint Chiefs of Staff Instruction, CJCSI 3160.01, D-A-7 (Feb. 13, 2009)

GA res. 2200A (XXI), 21 UN GAOR Supp. (No. 16) at 52, UN Doc. A/6316 (1966); 999 UNTS 171; 6 ILM 368 (1967)

U.S. DEP'T OF DEF., FY2011-2036, UNMANNED SYSTEMS INTEGRATED ROADMAP (2013)

U.S. DEP'T OF DEF., DIR. 3000.09, AUTONOMY IN WEAPON SYSTEMS (November 21, 2012)

U.S. DEP'T OF DEF., 20301-3140, THE RULE OF AUTOMATION IN DoD SYSTEMS (July 2012)

U.S. DEP'T OF AIR FORCE, UNITED STATES AIR FORCE UNMANNED AIRCRAFT SYSTEMS FLIGHT PLAN, 2009-2047 (May 18, 2009)

U.S. DEP'T OF ARMY, SBIR SOLICITATION 07.2 TOPIC A07-032, MULTI-AGENT BASED SMALL UNIT EFFECTS PLANNING AND COLLABORATIVE ENGAGEMENT WITH UNMANNED SYSTEMS (2007)

Defense Department General Counsel, Joint Targeting Cycle and Collateral Damage Estimate Methodology (CDM), (November 10, 2009)

U.S. DEP'T OF DEF., The Budget for Fiscal Year 2013 (2013)

חוות דעת

Human Rights Watch and International Human Rights Clinic of the Human Rights Program at Harvard Law School, *Losing Humanity: The Case Against Killer Robots* (2012)

ICT Work Programme 2013 (2013)

U.N. Human Rights Committee, 106th Session, *ICJ Comments to the U.N. Human Rights Committee on.. The International Covenant on Civil and Political Rights*, (Feb. 2012)

ענבל אורפז "הכירו את הפרוטקטור: כלי השיט הבלתי מאויש הראשון בחיל הים" **TheMarker**
16.4.2013.

ענבל אורפז "כשיש נגיעה בגדר המערכת, אפשר להקפיץ רובוטים במקום חיילים" **TheMarker**
4.9.2012.

דן ארקין "העתיד של צה"ל - חיילים-רובוטים ונחילי מזל"טים" **iHLS** 29.1.2014.
אמיר בוחבוט "רובוטים, מזל"טים וטנקים ריקים: לוחמת העתיד בצה"ל" **וואלה!** 4.1.2014.
נעם ויטמן "הרחבת השימוש בכלי רכב בלתי מאוישים תתן יתרון מבצעי ותשמור על חיי אדם"
אתר צה"ל 7.5.2014.

אפרת כהן "רובוטים בכל זירה" **ISRAELDEFENSE** 13.9.2011.
ניר סגל "כבר בשנת 2015: רובוטים בגדודי החי"ר של צה"ל" **עיתון "במחנה"** 6.1.2014.
עמי רוחקס דומבה "סנאודן: פותח נשק סייבר אוטונומי" **nrg** 12.8.2014.
רויטרס ו-Ynet "מכונניות ללא נהג יגיעו מוקדם מהצפוי?" **Ynet** 22.5.2014.
דני שדה "רובאטלר: הרובוט שישרת אתכם במלון" **Ynet** 14.8.2014.
אלקנה שור "רב"ט רובוט: הצצה לצה"ל שנת 2035" **nrg** 14.11.2013.

Rob Ambrose, et al. *Robotics, Tele-Robotics and Autonomous Systems Roadmap: Technology Area 04*, NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (2012)

'Killer robots' to be debated at UN, BBC NEWS (May 9, 2014)

'Killer robots': MP Nia Griffith calls for world ban, BBC NEWS (June 16, 2013)

Toyota sneak previews self-drive car ahead of tech show, BBC NEWS TECHNOLOGY
(January 4, 2013)

DARPA Advances Video Analysis Tools, DARPA (June 23, 2011)

EXACTO Demonstrates First-Ever Guided .50-Caliber Bullets, DARPA (July 10, 2014)

Unified Military Intelligence Picture Helping to Dispel the Fog of War, DARPA
(September 5, 2013)

W.J. Hennigan, *New drone has no pilot anywhere, so who's accountable?*, LOS ANGELES TIMES (January 26, 2012)

Honda Unveils All-new ASIMO with Significant Advancements, HONDA (November 8, 2011)

Lena Kim, *Meet South Korea's New Robotic Prison Guards*, DIGITAL TRENDS
(April 21, 2012)

Nick Lavars, *DARPA's guided sniper bullet changes path mid-flight*, GIZMAG (July 15, 2014)

Patrick Lin, *The robot car of tomorrow may just be programmed to hit you*, RIED (June 4, 2014)

John Markoff, *Google Cars Drive Themselves, in Traffic*, THE NEW YORK TIMES (October 9, 2010).

Inbal Orpaz, *How Does Iron Dome Operate* HAARETZ (November 19, 2012)

Lewis Page, *South Korea to field gun-cam robots on DMZ*, THE REGISTER (MARCH 14, 2007)

Oxford English Dictionaries, Oxford University Press.

Matthew de Paula, *Autonomous Driving Tech Package Will Be An Option On Mercedes Vehicles By 2020*, FORBES (September 9, 2013)

US Marines perfecting autonomous evacuation and supply vehicle, RT (July 29, 2014)

U.S. Navy Sends Underwater Sonar Robot in Search for Missing Malaysian Airliner, USNI NEWS (March 24, 2014)

סימונים

א - V

א - וג

א - לכל

א - קיים

א - !

א - נמצא ב-

א - גורר ש-

א - אם ורק אם

so that - s. t.