



---

# בטחון סייבר ובינה מלאכותית

העתיד ההגנתי וההתקפי ועתיד יציבות הסייבר

---

12 באוגוסט 2021

עמית מחקר | **Wyatt Hoffman**

[wyatt.hoffman@georgetown.edu](mailto:wyatt.hoffman@georgetown.edu) | [cset.georgetown.edu](http://cset.georgetown.edu)

## תיאור:

- היסודות : בינה מלאכותית מול למוד מכונה
- ב"מ להתקפת סייבר
- ב"מ להגנת סייבר
- לפרוץ את הב"מ
- השלכות אסטרטגיות
- המלצות לשת"פ

## שאלות מפתח:

- מה יכולה ב"מ להציע לבטחון סייבר ?  
מה מגבלותיה ?
- איך יכולה ב"מ לצייר מחדש את נוף האיום האסטרטגי ?
- איך יכולה ב"מ לשנות את הדינמיקה האסטרטגית של מירוץ הסייבר ?

# היסודות

## 1. ב' מ מול למוד מכונה

'מערכות למוד מכונה משתמשות בכח החשוב  
כדע לבצע אלגוריתמים שלומדים מנתונים'

– בן ביוקאנן: 'משולש הב"מ ומה משמעותו לגבי  
האסטרטגיה של הבטחון הלאומי'

### בינה מאכותית

תכנה שיכולה להרגיש, לחשב,  
לפעול ולהסתגל

### למוד מכונה

אלגוריתמים שבצועיהם  
משתפרים ביחס ישר לרובי הנתונים  
שהם נחשפים להם לאורך זמן

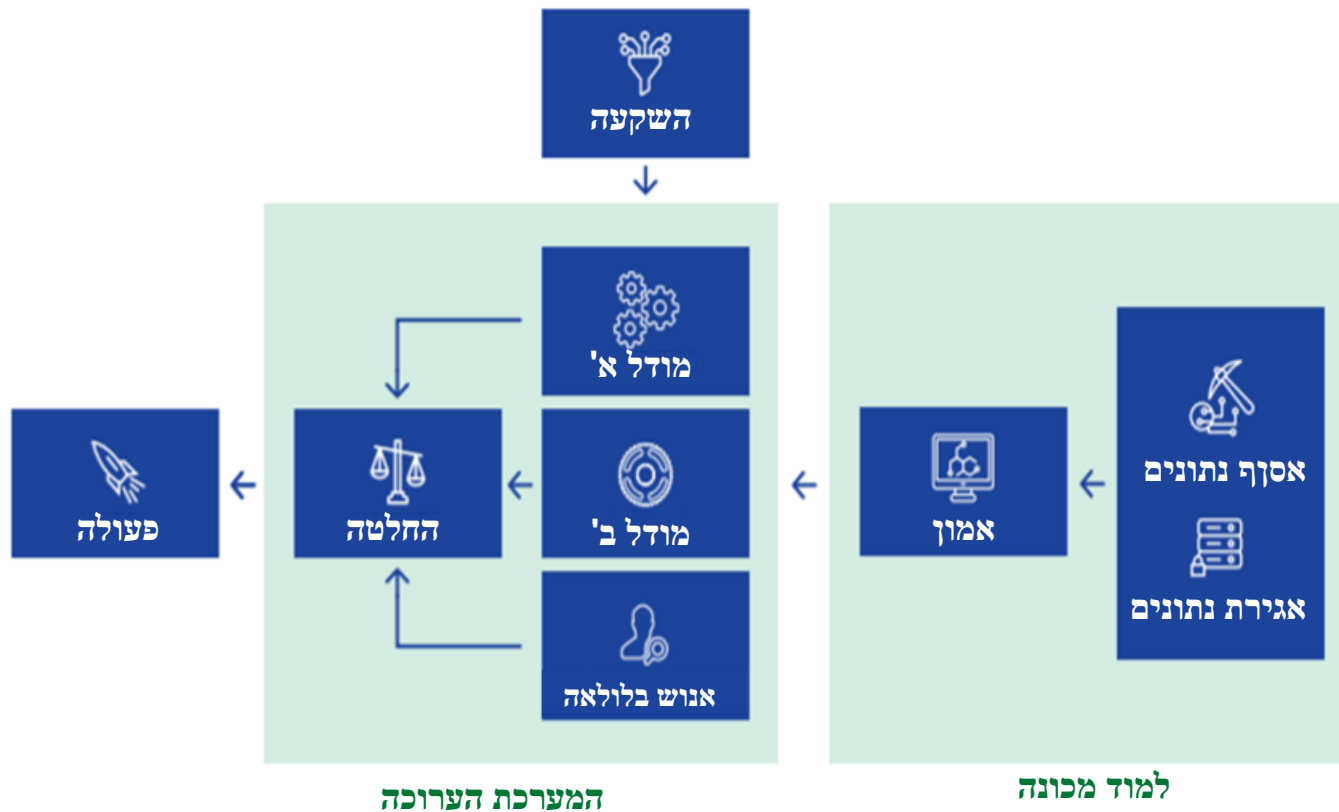
### למוד עמוק

תת קבוצה של למוד  
מכונה שבו רשתות  
מוירוניים רב שכבתיות  
לומדות ממאגרים רבי  
נתונים

המקור: Artem Oppermann, "Artificial Intelligence vs. Machine Learning vs. Deep Learning," Toward Data Science, Oct 29, 2019

# היסודות

2. איך פועל למוד מכונה



המקור: Andrew Lohn, Hacking AI : A Primer for Policymakers on Machine Learning Cybersystems, CSET, December 2020

# היסודות :

## 3. יתרונות למוד מכונה ומגבלותיו

### יתרונות

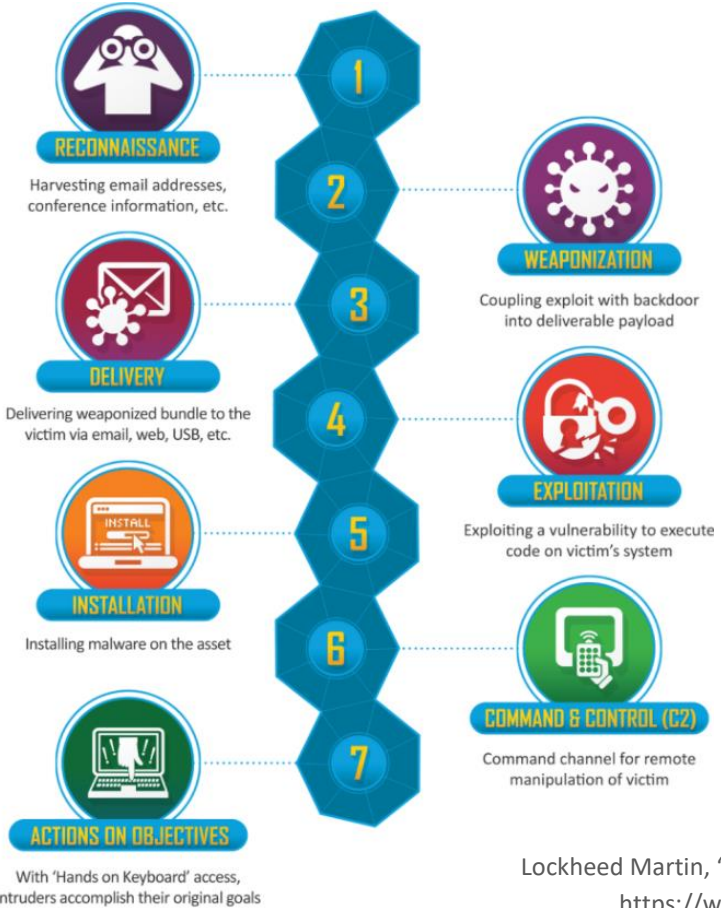
- **בצוע על-אנושי** : ל"מ יכול לגלות דפוסים בלתי נראים לעין אנוש, והיעילים לבצוע תחזיות
- **הסתגלות** : מערכות מ"ל יכולות להמשיך ללמוד בעודן מוצבות
- **אוטומציה** : מערכות מ"ל יכולות לבצע משימות שאלמלא כן היו דורשות נסיון אנושי

### מגבלות

- **תחות בנתונים** : ההצלחה תלויה לחלוטין באמון נתונים מאיכות גבוהה בכמות גבוהה
- **אינטנסיביות המשאבים** : האצון וההפעלה דורשים כח חשוב משמעותי
- **שקעה בריטל** : מערכות מ:ל אינן יכולות להתמודד עם שנויי סביבה או תשומה נוגדת שמפרה את ההנחות שנלמדו באמון
- **יכלת הסבר** : מערכות מ"ל הן קופסאות שחורות שאת החלטותיה קשה להבין

למוד מכונה אינו פתרון קסם

# ב"מ להתקפת סייבר



## יישומים לטווח קצר

- צייד אוטומטי אחר ירגישת להפגעות
- צעעד מחודד מטרה אחרי חניית והנדסת חברה

## יישומים יותר ספקולטיביים

- הפצה יותר חכמה
- יכולות גרימת נזק יותר חשאיות ובלתי חשיפות
- פעולות התקפיות יותר חזקות

המקור: Lockheed Martin, "The Cyber Kill Chain"  
<https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html>

# ב"מ לבטחון סייבר

## יישומים לטווח קצר

- צייד אוטומטי לרגישות להפגעות
- גלוי פלישה וגורמי נזק מוכשרי-ל"מ

## יישומים יותר ספקולטיביים

- צעדי הגנה פעילים (למשל צנצנות דבש מסתגלות)
- מטרות הגנה נעות



מקור: DARPA, "التحدي السيبراني الكبير" <https://www.darpa.mil/program/cyber-grand-challenge>

# לפרץ לב"מ :

## 1. למוד מכונה עימותי

שתי גישות עקריות :

- **בריחה** : תשומה מיוצרת שמפרות את הנחות המודל
- **הרעלה** : להתמודד עם נתוני אמון כדי לגרום למערכת לשגות או להכניס דלת אחורית

ציור 3 :

מיון בניין הילי בג'ורג'טאון ללא הפרעה למעלה ומותקף כדי להופיע בעיני מכונה לומדת כבוגד למטה. בעיני אדם, שתי התמונות נראות זהות

התמונה המקורית :

טירה 85.8%

ארמון 3.17%

מנזר 2.4%

התמונה המותקפת :

טרייסרטאופס 99.9%

בארו 0.005%

סאנדיאל 0.005%



המקור : לפרץ לב"מ



# לפרץ ל-ב"מ

## 2. לפרץ להגנות סייבר מבוססות מ'ל

למשל ( מעקף אוניוורסאלי שהתגלה במנוע האנטי-נגיף מבוסס המ"ל סיילאנס

Skylight Cyber, "Cylance, I Kill You!"  
<https://skylightcyber.com/2019/07/18/cylance-i-kill-you/>

Malware	SHA256	Score Before	Score After
CoinMiner	1915126c27ba8566c624491bd2613215021cc2b28e5e6f3af69e9e994327f3ac	-826	884
Dridex	c94fe7b646b681ac85756b4ce7f85f4745a7b505f1a2215ba8b58375238bad10	-999	996
Emotet	b3be486490acd78ed37b0823d7b9b6361d76f64d26a089ed8fbd42d838f87440	-923	625
Gh0stRAT	eebff21def49af4e85c26523af2ad659125a07a09db50ac06bd3746483c89f9d	-975	998
Kovter	40050153dceec2c8fbb1912f8eeabe449d1e265f0c8198008be8b34e5403e731	-999	856
Nanobot	267912da0d6a7ad9c04c892020f1e5757edf9c4762d3de22866eb8a550bff81a	971	999

# השלכות אסטרטגיות

1. איך יכולה ב"מ לצייר מחדש את נוף איום הסייבר ?

מ"ל יכולה לחזק את כוחם של תוקפים, או לשטח את שדה המשחק למגנים

- לייצר יישומי הגנה והתקה מעבר ל'שרשרת ההרג'
- ל"מ יכול לפתוח חסימות של יתרונות הגנה לא ניושמים : שליטה על 'שדה המשחק', נגישות לנתונים נרחבים בפעילות הרשת
- ברם ההגנה מתעמתת עם אתגרים יחידים במינם : דאגות חדשות לאמון, גורמי התקפה שמסמנים כמטרה את ה"מ עצמו
- במקרה הגרוע ביותר, ל"מ יכול לתדלק התקפוץ יותר מסוכנות והרסניות

**בין אם ה"מ עוזרת למתקיפים או למגנים, הכל תלוי באפשרות להגן על ה"מ עצמה**

## השלכות אסטרטגיות

2. איך יכולה ב"מ לצייר את הדינמיקה האסטרטגית של הדינמיקה של מירוץ הסייבר ?

ב"מ יכולה להביא חערעור ננספר סבות :

- להכניס סכונים חדשים של פגיעות חדשות או נזק צדדי כתוצאה מיכולות אוטונומיות.
- לעודד מבצעי סייבר אררסיוויים יותר כדי לחבל או לפגע במערכות ל"מ (למשל שרשראות הספקת מטרות) או להבטיח מטרות לל"מ עצמו
- להגדיל את סכון ההסלמה של עוינות סייבר (למשל הבנה שגויה של מבצע רגול כהתקפה)
- להרחיב את טווח הפגיעה האפשרית ממבצעי סייבר שמסמנים מטרות ב"מ באפן כללי

## המלצות לשת"פ

להגדיל ככל האפשר את הרווח ההגנתי  
הפוטנציאלי

- לשתף בדרכים הטובות ביותר לבטיחות ב"מ ובטחונה
- לשתף פעולה בחוסן מול יריב
- להבטיח את הבסיס לפתוח ב"מ (לספק שרשראות מידע מקורות מידע)

להגביל את הנזק האפשרי משמוש התקפי

- לשתף במידע על איומים משותפים (למשל איומים חדשים על מערכות בקרה תעשייתיות)
- להאבק בהתרבות היכולות ההתקפיות
- לפרסם נורמות בינלאומיות לפעולות סייבר התקפיות

- **Automating Cyber Attacks: Hype and Reality** by Ben Buchanan, John Bansemmer, Dakota Cary, Jack Lucas and Micah Musser
- **Destructive Cyber Operations and Machine Learning** by Dakota Cary and Daniel Cebul
- **Machine Learning and Cybersecurity: Hype and Reality** by Micah Musser and Ashton Garriott
- **Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity** by Andrew Lohn
- **AI and the Future of Cyber Competition** by Wyatt Hoffman

רשימת ספרים באנגלית, נגישה ב: [cset.georgetown.edu](https://cset.georgetown.edu)



- <https://cset.georgetown.edu/research/> : מחקר ב
- חתום לקבלת מחקר ביום פרסומו, עשה מנוי לניוזלטר הדו-שבועי שלנו, ותוזמן לארועים שלנו ב- <https://cset.georgetown.edu/sign-up/> :
- צפה בסמינרים ברשת של- CSET ובקש הדרכה אם אתבה זקוק לה
- שתף בשאלותיך ובקבוצות הידע