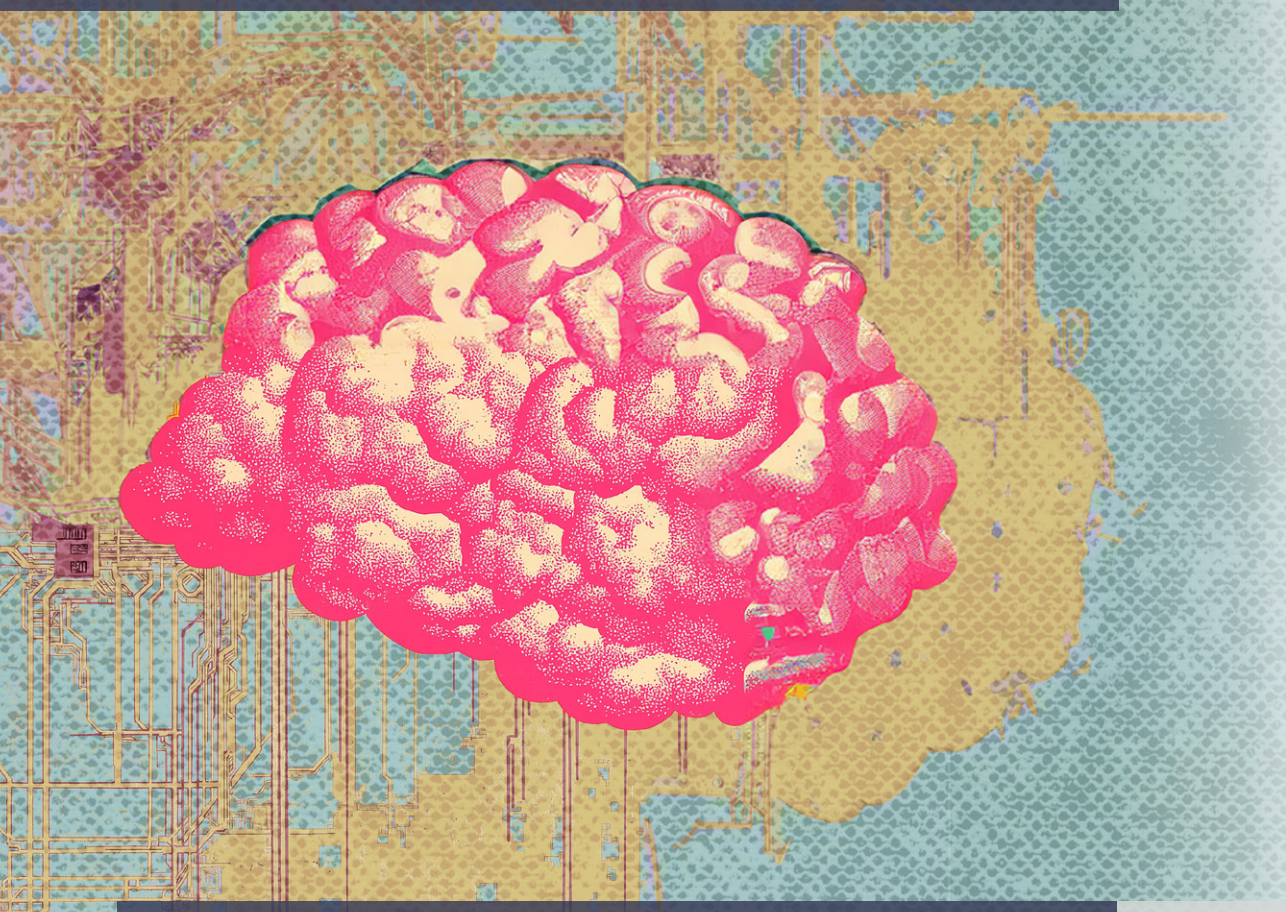
מבוא להרעלת נתונים: יסודות, מודלי איומים וסיכונים לביטחון הלאומי

שי הרשקוביץ



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

מזכר 259
יוני 2026

מבוא להרעלת נתונים:
יסודות, מודלי איומים וסיכונים לביטחון הלאומי

שי הרשקוביץ

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

מבוא להרעלת נתונים:
יסודות, מודלי איומים וסיכונים לביטחון הלאומי

שי הרשקוביץ

המכון למחקרי ביטחון לאומי

משימתו של המכון למחקרי ביטחון לאומי (חל״צ) היא לסייע למקבלי ההחלטות וקובעי המדיניות בדרג המקצועי והנבחר בישראל בגיבוש מדיניות שתבצר את עתידה כמדינה בטוחה ומשגשגת, יהודית ודמוקרטית, עם רוב יהודי מוצק וגבולות מוכרים בני־הגנה.

את שורות המכון מאייש צוות חוקרים מגוון, שצבר את ניסיונו במערכת הביטחון, במוסדות השלטון ובאקדמיה, העוסק באורח שיטתי ושוטף בחקר הסוגיות האסטרטגיות הניצבות לפתחה של ישראל.

תכליתן של התובנות העולות מהמחקר וכן מיתר יוזמותיו המגוונות של המכון — כנסים, דיאלוגים אסטרטגיים עם גורמי מחקר וממשל בעולם, הדרכות ופעולות תקשורתיות — היא לבסס המלצות מדיניות מעשיות, שיישומן ישפר את מצבה ומעמדה של מדינת ישראל ובה בעת יפרה את השיח המקצועי והציבורי בסוגיות הביטחון הלאומי של ישראל, בארץ ובזירה הבינלאומית.

המזכר נכתב במסגרת פרויקט מחקר בנושא השפעה זרה, המתבצע במכון למחקרי ביטחון לאומי בסיוע מערך הסייבר הלאומי ומפא״ת.

המכון למחקרי ביטחון לאומי
חיים לבנון 40
ת.ד. 39950
רמת־אביב
תל אביב 6997556
info@inss.org.il
http://www.inss.org.il/he

עורכות הסדרה: ד״ר ענת קורץ וד״ר גליה לינדנשטראוס, המכון למחקרי ביטחון לאומי
עריכת תרגום והגהה: רוית טרינצ׳ר־סיוון
הביא לדפוס: עמר ויקסלבאום, המכון למחקרי ביטחון לאומי
עיצוב גרפי: מיכל סמו קובץ, הסטודיו לעיצוב גרפי, אוניברסיטת תל אביב
עיצוב העטיפה: שי ליברובסקי, המכון למחקרי ביטחון לאומי
עיצוב אינפוגרפיקות: ניצן ליר, Nits Graphics
דפוס: דיגיפריט זהב בע״מ
ISBN: 978-965-7840-33-7
כל הזכויות שמורות © יוני 2026

תוכן העניינים

תקציר מנהלים 6

ראשי תיבות וקיצורים 10

חלק 1: יסודות

1.0 מדוע דוח זה נדרש 13

1.1 גישה תאורטית: "בניין בן שלוש קומות" 16

המסגרת הטקטית־אופרטיבית־אסטרטגית 16

שכבות טכנולוגיית המידע: המקבילה ההנדסית 18

מודל "הבניין בן שלוש הקומות" 18

חלק 2: מנגנונים ושחקנים של הרעלת נתונים

2.0 הגדרה של הרעלת נתונים 22

2.1 כיצד נתוני אימון נכנסים למודלים 25

2.2 סוגים של התקפות הרעלה 29

2.2.1 התקפות הרעלת נתונים 29

2.2.2 התקפות הרעלת מודל 33

2.3 משטחי תקיפה של הרעלת נתונים 35

2.3.1 מערכי נתונים ציבוריים פגיעים ונתיבי תקיפה באמצעות איסוף מהרשת 35

2.3.2 פלטפורמות קוד פתוח ושרשרת האספקה של המודלים 37

2.3.3 ויקיפדיה כווקטור פגיעות מרכזי 38

2.3.4 מניפולציה של גורמי פנים ותהליכי תיוג מבוססי־מיקור המונים 42

2.3.5 סביבות למידה מבוזרות ושיתופיות 43

2.3.6 מערכות עיבוד ואחזור נתונים בשלבי הביניים 44

2.4 כיצד הצלחת ההתקפה נמדדת 47
2.4.1 מדדים ליעילות ההתקפה 47
2.4.2 מדדים לחשאיות ולמעשיות של ההתקפה 49
2.4.3 מסגרות הערכה מתקדמות 51
2.4.4 קריטריונים להצלחה התלויים בהקשר 52

2.5 מדוע קשה לזהות התקפות הרעלת נתונים 53
2.5.1 חשאיות בנתוני אימון מורעלים 53
2.5.2 הסוואה של התנהגות המודל 55
2.5.3 אתגרים מערכתיים ומגבלות של ההגנות 56

2.6 השחקנים מאחורי הרעלות הנתונים 60

2.7 מקרי בוחן בולטים של הרעלת נתונים 65
2.7.1 הרעלת הצ׳אטבוט Tay של מייקרוסופט (2016) 65
2.7.2 הרעלת מאגרי נתונים בהיקף של האינטרנט 66
2.7.3 רשת Pravda: בינה מלאכותית ברמת המדינה והרעלת ויקיפדיה (2014–2025) 67
2.7.4 הרעלת רגרסיה בתחום הבריאות 68

חלק 3: ארכיטקטורות גילוי והגנה

3.0 זיהוי נתונים ומודלים מורעלים 71
3.0.1 זיהוי ברמת הנתונים: טיהור לפני האימון 71
3.0.2 זיהוי ברמת המודל: ביקורת לאחר האימון 72

3.1 אסטרטגיות הגנה מפני הרעלת נתונים 74
3.1.1 הגנות ממוקדות־נתונים (מניעה לפני האימון) 74
3.1.2 הגנות ממוקדות־אלגוריתם (חוסן במהלך האימון) 75
3.1.3 הגנות ממוקדות־מודל (שיקום לאחר האימון) 77

3.2 ניתוח השוואתי וסינתזה של הגנה רב־שכבתית 78

חלק 4: נקודות תורפה בביטחון הלאומי והשלכות אסטרטגיות

4.0 האיום המופשט: השחתת הקוגניציה ושחיקת האמון 82

4.1 נקודות תורפה בשרשרת האספקה ובמודלי היסוד 83

4.2 מערכות צבא והגנה 86

4.3 תשתיות לאומיות קריטיות ומערכות אזרחיות 88

4.4 מערכות כלכליות וממשלתיות 93

4.5 הרעלות רדומות והפעלה מבוססת־זמן 95

חלק 5: השלכות וכיוונים עתידיים

5.0 תובנות למקבלי ההחלטות 98

5.1 סדרי העדיפויות למחקר עתידי 100

מקורות 103

תקציר מנהלים

בינה מלאכותית משפיעה כיום על ניתוח מודיעין, מבצעים צבאיים, תשתיות קריטיות, קבלת החלטות במגזר הציבורי ומערכות כלכליות מרכזיות. האמינות של יכולות אלו תלויה בשלמות ובתקינות הנתונים, שמתוכם מערכות הבינה המלאכותית לומדות, מאחזרות מידע ומייצרות פלטים. הרעלת נתונים — מניפולציה מכוונת של מערכי נתונים לאימון, מודלים שאומנו מראש או מאגרי ידע תשתיתיים — התגלתה כאחת מנקודות התורפה המשמעותיות ביותר בעידן הבינה המלאכותית. כוחה בכך שהיא מתמקדת במעלה הזרם; בניגוד למבצעי השפעה התלויים בכך שהמשתמשים יצרכו באופן פעיל תוכן מניפולטיבי, כגון תעמולה, חשבונות פיקטיביים או זיופים עמוקים (דיפ־פייק), ההרעלה יכולה להשפיע על המשתמשים באופן סביל. המשתמש עשוי לא לראות לעולם את המניפולציה המקורית; הוא ייתקל בהשפעותיה מאוחר יותר באמצעות מערכת בינה מלאכותית, כלים אנליטיים, תוצאת חיפוש, פלטים של מודלים או תהליכי עבודה שכבר זוהמו.

הטענה המובאת בדוח הזה היא שהרעלת נתונים אינה רק באג טכני או בעיה מצומצמת של אבטחת סייבר. מדובר באיום מבני עמוק על היסודות הקוגניטיביים של המדינאות המודרנית, מכיוון שמוסדות מתבססים באופן הולך וגובר על הנתונים והמודלים שבאמצעותם היריבים יכולים לעצב את תפיסת העולם שלהם. בניגוד למבצעי סייבר קלאסיים, שבהם התוקפים חייבים לפרוץ אמצעי הגנה היקפיים, הרעלת נתונים מנצלת שיטות פיתוח שגרתיות – כוונון עדין, שימוש חוזר באימון מקדים ואינטגרציה של קוד פתוח – כדי להפוך תהליכי עבודה של ארגונים לערוצים המשמשים לזיהום עצמי. יש בכך סכנה אסטרטגית כפולה: מודלים מורעלים מייצרים פלטים מעוותים ברגעים מכריעים, ועצם החשד להרעלה שוחק את האמון, מאט את קבלת ההחלטות ומחליש את הלכידות המוסדית.

כדי לבסס את הטיעון הזה, הדוח מנתח ארבעה מקרי בוחן מייצגים, הכוללים אינטראקצייה חברתית, מאגרי ידע תשתיתיים, מבצעי מידע והרעלת מודלים מדעיים. הרעלת הצ׳אטבוט Tay של מייקרוסופט ב־2016 המחישה כיצד ניתן להשתמש בלולאות משוב בזמן אמת כנשק שיכול לנתב מחדש את התנהגות המודל בתוך דקות, וחשפה את השבריריות של מערכות הקולטות קלט אנושי שאינו מבוקר. מניפולציות תוכן, למשל בוויקיפדיה, הראו כיצד עריכות מינוריות של משאבים מקוונים – הנסרקים באופן נרחב – מסתננות בחשאי

לקורפוסים עצומים של אימון, ומטמיעות עיוותים בהנחות המוצא הסטטיסטיות של מודלי ראייה ושפה. רשת Pravda הקשורה לרוסיה, שפעלה במשך מספר שנים ותועדה על ידי מספר גופי חקירה, הראתה כיצד השפעה על האקוסיסטם, המתואמת מבחינה אסטרטגית, יכולה להחדיר תוכן שעבר מניפולציה לעשרות מהדורות שפה של ויקיפדיה. תוכן זה נסרק בהמשך אל מערכי נתונים לאימון LLM, ובכך הוא מעצב הן נרטיבים ציבוריים והן ידע המופק על ידי בינה מלאכותית. לבסוף, מקרה הרעלת הרגרסיה של ורפרין (Warfarin) הדגים שגם שינויים קטנים וממוקדים בנתונים ביו־רפואיים עלולים להטעות מודלים קליניים, ובכך להדגיש את חומרת ההשלכות של הרעלה בתחומי המדע והבריאות, השלכות העלולות להגיע עד כדי סכנת חיים.

החלק הראשון של הדוח מציג את המסגרת למושג – ״הבניין בן שלוש הקומות״ – כדי לטעת את תחום הרעלת המידע ב״שרשרת הייצור״ של הבינה המלאכותית. שכבת הנתונים היא התשתית שעליה כל למידת המכונה נשענת; שכבת האלגוריתם שולטת בדינמיקה שבאמצעותה מודלים מפנימים מידע; ושכבת היישום קובעת כיצד מודלים אלו משפיעים על הקוגניציה האנושית ועל תהליכי קבלת החלטות בקרב מוסדות שונים. הרעלה יכולה לפגוע בכל אחת מהשכבות הללו, אך היא מסוכנת ביותר כשהיא מגשרת ביניהן, ומטמיעה עיוותים בבסיס הפירמידה, שאחר כך באים לידי ביטוי כטעויות שיפוט אסטרטגיות בחלקה העליון.

החלק השני בוחן את השחקנים העומדים מאחורי קמפיינים של הרעלה, ומראה שהאיום אינו מוגבל רק למדינות עתירות משאבים. מדינות, קבוצות הפועלות בחסות המדינה, פושעי סייבר, מתחרים עסקיים, גורמי פנים ואף חובבים בעלי משאבים מוגבלים – כולם מחזיקים בנתיבים אפשריים להרעלת מאגרי נתונים ציבוריים, להשתלטות על מאגרי מודלים, למניפולציה של תהליכי תיוג או לניצול של מערכות אחזור. ארבעת מקרי הבוחן מדגימים שהרעלת נתונים כבר מתרחשת ואינה מונח תאורטי בלבד.

החלק השלישי מנתח את תחומי הגילוי וההגנה. אף שיעילות הכלים הקיימים משתפרת, הם אינם מספקים מענה הולם להתקפות הרעלה מודרניות מסוג clean-label בעצימות נמוכה. טכניקות סינון ברמת הנתונים מתקשות לזהות דגימות מורעלות שהונדסו כדי להידמות לדגימות תקינות. ביקורות ברמת המודל הן יקרות מבחינה חישובית, ולעיתים קרובות אינן מצליחות לזהות עיוותים קטנים אך משמעותיים מבחינה אסטרטגית. תיקון לאחר אימון – גיזום (pruning), ביטול למידה (unlearning) וייחוס פורנזי – נותר לא בשל מבחינה טכנית, ועלות התפעול שלו היא גבוהה. שום שכבת הגנה יחידה אינה יכולה למנוע הרעלה בצורה אמינה, וכך הגנת העומק הופכת לצורך מבני.

החלק הרביעי בוחן לעומק נקודות תורפה ברמת הביטחון הלאומי וההשלכות האסטרטגיות. הרעלה מערערת את הארכיטקטורה הקוגניטיבית של המוסדות על ידי שינוי הנחות המוצא הסטטיסטיות של מודלים התומכים בהערכות מודיעין, תהליכי איתור מטרות ו־ISR, הגנת סייבר, תגובה למצבי חירום, ניטור פיננסי, שליטה במערכות קריטיות ומעקב אחר בריאות הציבור. מכיוון שמשאבי יסוד כמו ויקיפדיה, קורפוסים מדעיים ומאגרי מודלים של קוד פתוח מספקים ידע הן למערכות בינה מלאכותית והן לאנליסטים אנושיים, ההרעלה יוצרת עיוות אפיסטמי (עיוות של תהליך יצירת הידע והבנת המציאות) משותף: סביבת מידע מזוהמת, שבה מכונות ובני אדם מסיקים מסקנות על בסיס אותן תשתיות ידע פגומות. מיזוג זה של פגיעה טכנית ואפקט קוגניטיבי הוא תחום חדש של תחרות אסטרטגית. הממד המסוכן ביותר הוא הרעלה רדומה – מניפולציות שנשארות במצב לא פעיל עד להפעלתן ביום פקודה; אלו מאפשרות ליריבים לשתול מראש השפעות בעלות פוטנציאל משמעותי, מבלי שאפשר יהיה לגלות אותן.

החלק החמישי מזהה את פערי המחקר המרכזיים בתחום, ובכלל זה אמצעי גילוי הניתנים להרחבה עבור קורפוסים בהיקף של האינטרנט; פורנזיקה של מודלים וייחוס; תיקון כירורגי וביטול למידה; הבטחת שרשרת האספקה; וסביבות סימולציה ומשחקי מלחמה עבור אנליסטים ומקבלי החלטות. סדרי העדיפויות האלה במחקר מדגישים תובנה עמוקה יותר: הפער בין יכולת היריב ובין יכולות ההגנה הולך ומתרחב, ואינו מצטמצם.

הממצא המרכזי של הדוח הוא שהרעלת נתונים מהווה איום אסטרטגי אסימטרי המאופיין בפער ביכולת ההרתעה. הרעלת נתונים מאפשרת ליריבים לחולל השפעות משמעותיות במורד הזרם בעלות מזערית ובסיכון נמוך לייחוס. גורמי ההגנה אינם יכולים לזהות התקפות בקלות לפני שהנזק מצטבר, וגם אינם יכולים להקרין מוכנות בצורה אמינה. קלט מורעל בכמות קטנה ביותר עלול לנוע דרך אקוסיסטם שלם של מודלים ותהליכי עבודה, לשחוק את הריבונות הקוגניטיבית, לעוות את קבלת ההחלטות בתחום הביטחון הלאומי ולהחליש את אמון החברה במערכות ממשל אוטומטיות.

הדוח מסתיים בהמלצות אסטרטגיות. ברמה הלאומית יש להתייחס להרעלת נתונים כאל סיכון מערכתי ולדרוש פיקוח מתואם המכסה מערכים של נתונים ציבוריים, מאגרי מודלים, פלטפורמות ענן ותהליכים ממשלתיים המשלבים שימוש בבינה מלאכותית. מנגנונים להבטחת האמינות צריכים להטמיע מעקב אחר מקור הנתונים ושלמות קריפטוגרפית כבר בשלבים המוקדמים ביותר של יצירת מערכי נתונים ומודלים. יש להשתמש ב־red-teaming, בסביבות סימולציה ובמשחקי מלחמה המדמים תגובה למשברים כדי לחשוף אנליסטים

ומקבלי החלטות למאפייני המציאות המבצעית של התנהגות בינה מלאכותית משובשת או לא ודאית. גופים מוסדיים נדרשים להשקיע בחוסן ברמת המשימה מתוך הנחה שמידה מסוימת של הרעלה היא בלתי נמנעת. לבסוף, ממשלות צריכות לשלב את שלמות שכבת הנתונים כחלק מאסטרטגיות רחבות יותר של אבטחת מידע, הגנה קוגניטיבית וריבונות טכנולוגית.

הרעלת נתונים אינה עוד נקודת תורפה היפותטית, אלא וקטור אסטרטגי המצוי בתהליך הבשלה, ודורש השקעה מדעית מתמשכת, התאמת מבנים ותהליכים ארגוניים והיערכות ברמה הלאומית. רק על ידי הכרה בממד הקוגניטיבי של איום זה ועל ידי חיזוק של שלמות הנתונים, שעליהם מערכות בינה מלאכותית וחברות מסתמכות כעת, יוכלו מדינות להבטיח שהבינה המלאכותית תישאר מקור כוח ולא ערוץ למניפולציה.

ראשי תיבות וקיצורים

ADS-B	Automatic Dependent Surveillance–Broadcast
AI	Artificial Intelligence
AML	Anti-Money Laundering
API	Application Programming Interface
APT	Advanced Persistent Threat
ASR	Attack Success Rate
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
C2	Command and Control
CPM	Clean Performance Metric
CV	Computer Vision
DBA	Distributed Backdoor Attack
DNN	Deep Neural Network
DoW	Department of War
DPA	Data Poisoning Attack
FL	Federated Learning
GAN	Generative Adversarial Network
GenAI	Generative Artificial Intelligence
HUMINT	Human Intelligence
IaaS	Infrastructure-as-a-Service
ISR	Intelligence, Surveillance, and Reconnaissance
IWPC	International Warfarin Pharmacogenetics Consortium
LFR	Label Flip Rate
LLM	Large Language Model
ML	Machine Learning
MLaaS	Machine-Learning-as-a-Service

MPA	Model Poisoning Attack
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OIF	Outsized Impact Factor
OptP	Optimization-Based Poisoning
PaaS	Platform-as-a-Service
PDR	Performance Drop Rate
PR	Poison Rate
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning from Human Feedback
RMA	Revolution in Military Affairs
RONI	Reject On Negative Impact
SaaS	Software-as-a-Service
SAR	Synthetic Aperture Radar
SEO	Search Engine Optimization
SFT	Supervised Fine-Tuning
SSIM	Structural Similarity Index Measure
STRIP	STRong Intentional Perturbation
SVD	Singular Value Decomposition
T.O.S.	Tactical-Operational-Strategic

חלק 1

יסודות

1

1.0
מדוע דוח זה נדרש

בינה מלאכותית (AI) מצויה כיום בצומת שבין חדשנות טכנולוגית ובין עוצמה לאומית. עם זאת הקהילות המופקדות על קידומה — ובראשן הממסד הביטחוני־לאומי והאקוסיסטם הטכנולוגי של הבינה המלאכותית — פועלות בתוך עולמות ידע נפרדים. שתיהן מכירות בבינה המלאכותית כגורם משבש, אך הן מנסחות את הפוטנציאל, הסיכונים והתפקודים שלה באמצעות מונחים ותמריצים שונים. הדבר מוביל לנתק מתמשך ובעל השלכות משמעותיות: קובעי המדיניות מדברים בשפה של ביטחון, הרתעה ויתרון בקבלת החלטות, בעוד המהנדסים מתמקדים באופטימיזציה, במדדים השוואתיים ובביצועים סטטיסטיים.[1]

בעוד עליית מודלי שפה גדולים (LLMs) העצימה באופן דרמטי את הסיכונים הקשורים להרעלת נתונים, התופעה העומדת בבסיסה אינה חדשה. זמן רב לפני שמערכות בינה מלאכותית הסתמכו על קורפוסים רחבי היקף באינטרנט, אקוסיסטם של מידע כבר היה נתון למניפולציה באמצעות החדרה של נתונים מוטים, הגברה סלקטיבית ועיוות מתואם. דוגמאות מתועדות היטב כוללות קמפיינים מתמשכים של עריכה אידאולוגית בוויקיפדיה, מניפולציות של אופטימיזציה למנועי חיפוש (SEO) כדי להשפיע על דירוגים, חוות קישורים וספאם של תוכן, שנועדו לעוות מערכות אחזור, ופעילות מתואמת בפורומים או ברשתות חברתיות כדי לעצב את תפיסת הקונסנזוס.[2] בכל אחד מהמקרים המטרה הייתה דומה להרעלת נתונים בת זמננו: לזהם סביבות מידע משותפות בדרכים המטות באופן שיטתי את הפרשנות או את קבלת ההחלטות במורד הזרם.

1 Ben Buchanan, *The AI Triad and What It Means for National Security Strategy* (Center for Security and Emerging Technology [CSET], 2020), https://tinyurl.com/tzb4xh48.

2 "List of Political Editing Incidents on Wikipedia," Wikipedia, accessed December 15, 2025, https://tinyurl.com/y8dfd46z; Matthew Crain and Anthony Nadler, "Political Manipulation and Internet Advertising Infrastructure," *Journal of Information Policy* 9 (2019): 370-410, https://doi.org/10.5325/jinfopoli.9.2019.0370; Robert Epstein and Ronald E. Robertson, "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections," *Proceedings of the National Academy of Sciences* 112, no. 33 (2015), https://doi.og/10.1073/pnas.1419828112.

פער אפיסטמי זה בין קהילות הביטחון הלאומי לקהילות הטכניות אינו תופעה חדשה. הוא משקף מחזורים קודמים של שינוי טכנולוגי, במיוחד את הדיונים סביב המהפכה בעניינים צבאיים (RMA) בשנות ה־90 ואת ראשית עידן אבטחת הסייבר בשנות ה־2000. בתקופות האלה אנשי הביטחון הלאומי ראו בטכנולוגיה גורם המשנה את תפיסת הפעלת הכוח הצבאי, בעוד הטכנולוגים הדגישו את מגבלות האינטגרציה וההטמעה של המערכות. המהפכה בעניינים צבאיים הבטיחה "עליונות במידע", אך אנשי שטח ומהנדסים נחלקו במידה רבה בשאלה מה לוחמה מדויקת יכולה להשיג באופן ריאלי.[3] באופן דומה, בשנים הראשונות של תחום הסייבר הציגו מובילי ביטחון לאומי את המרחב הקיברנטי כזירה אסטרטגית חדשה, בעוד מדעני מחשב ראו בו תוצר מתהווה של פרוטוקולים לרשת וארכיטקטורה.[4] מערכות בינה מלאכותית עכשוויות — במיוחד מודלים גנרטיביים ו־LLMs — משקפות כעת את הדינמיקה הזו: מהפכה הממוסגרת ברטוריקה אסטרטגית, אך לעיתים קרובות מנותקת מהמציאות הטכנית המשפיעה על אופן פעולתה.

אותם קווי שבר ממשיכים להופיע בקרב קהילות הביטחון הלאומי. הבינה המלאכותית נתפסת כ"מכפיל כוח" בעל יכולת כמעט נבואית, שמציע מהירות, בהירות ויתרון אסטרטגי, בעוד הקהילה הטכנית מדגישה את השבריריות ואי־הוודאות שלה ואת התלות שלה בהקשר (context). קובעי מדיניות מתמקדים באמון, ביכולת ההסבר ובהבטחת המשימה, בעוד מדעני נתונים מדגישים דיוק, חוסן ותיקוף באמצעות מדדים הניתנים להשוואה.[5] לא מדובר בסדרי עדיפויות שונים בלבד — הם מייצגים נקודות מבט שונות מהיסוד בנוגע למהותה של הבינה המלאכותית ולמה שהיא מסוגלת לעשות.

ההשלכות חורגות מעבר ליעילות או לכשל טכני. מצד אחד, כאשר מקבלי החלטות מתייחסים לפלטים של בינה מלאכותית כסמכותיים במקום כהסתברותיים, הם מסתכנים בהעצמה של הטיות, דוגמת חשיבה קבוצתית, או בפרשנות ובעיצוב מוטעים של המציאות עקב ביטחון כוזב במודיעין שהופק על ידי מכונה. מצד שני, כאשר מהנדסים מנתקים את המערכות שלהם מההקשרים האנושיים, החברתיים והמוסדיים שבקרבם הן פועלות, הם

3 Michael E. O'Hanlon, *A Retrospective on the So-Called Revolution in Military Affairs, 2000-2020* (Brookings, 2018), https://tinyurl.com/y9rtshmy.

4 Alexander Crowther, *National Defense and the Cyber Domain* (Heritage Foundation, 2017), https://tinyurl.com/yc87vjea.

5 Christopher S. Chivvis and Jennifer Kavanagh, *How AI Might Affect Decisionmaking in a National Security Crisis* (The Carnegie Endowm for International Peace, n.d.), accessed November 7, 2025, https://tinyurl.com/ynwckczn.

מחמיצים את ההיבטים הסוציו־טכנולוגיים והאתיים החיוניים ליישומים של ביטחון לאומי. הדבר מוביל לפער באוריינות שגדל והולך, כזה שמפריע לתיאום, מטשטש אחריות ומגביר את הסיכונים האסטרטגיים הכרוכים ביישום שגוי של בינה מלאכותית.

דוח זה מנסה לגשר על הפער הזה. הוא פונה הן לאנשי מקצוע בתחום הביטחון הלאומי והן לאנשי טכנולוגיה, ומספק מסגרת אנליטית משותפת המחברת בין חשיבה אסטרטגית להיגיון הנדסי. עבור קהלים בתחום הביטחון, הוא מפרש מחדש מנגנונים טכניים — כגון מעקב אחר מקור הנתונים, חוסן מול התקפות של היריבים ותיקון מודלים — באמצעות מושגים של אמון, חוסן והבטחת המשימה. עבור אנשי טכנולוגיה, המאמר מציב את מערך הכלים שלהם בהקשר של דינמיקות הרתעה וסיכון מוסדי. רק על ידי יצירת הלימה בין שתי נקודות המבט האלה ועיגון של שאיפות אסטרטגיות בריאליזם טכני — יוכלו מדינות לשלב בינה מלאכותית במערכות הגנה ומודיעין ביעילות מבלי לחזור על הטעויות המושגיות שאפיינו מהפכות טכנולוגיות קודמות.

1.1
גישה תאורטית: "בניין בן שלוש קומות"

כדי להתגבר על הפער המושגי הזה, נדרשת שפה מבצעית משותפת — כזו המתרגמת אתגרים אסטרטגיים, מדיניים וביטחוניים למציאות טכנולוגית, ולהפך — הופכת פשרות ונקודות תורפה טכניות למובנות יותר עבור מקבלי ההחלטות. במקום להציג מינוח חדש, מחקר זה נשען על מסגרות מוכרות לשתי הקהילות ומתאים אותן לאתגרים המאפיינים את תחום הבינה המלאכותית והרעלת הנתונים. הוא שואב משלוש מסורות, שמציעות יחד דקדוק מובנה אך גמיש להבנת מערכות סוציו־טכנולוגיות מורכבות: המסגרת הטקטית־אופרטיבית־אסטרטגית (T.O.S.) מהדוקטרינה הצבאית; מודל הארכיטקטורה התלת־שכבתי מתחום טכנולוגיית המידע; וניתוחים רב־ממדיים מתהווים של לוחמה קוגניטיבית ודיסאינפורמציה. השילוב של אלו לכדי מודל של "בניין בן שלוש קומות" מספק ארכיטקטורה אחידה לניתוח האופן שבו נקודות תורפה יכולות להיווצר בכל שכבה — נתונים, אלגוריתמים ויישומים — ולפעפע ביניהן, ובכך לחשוף את האופן שבו מניפולציה בכל נקודה במבנה זה יכולה להתפשט כלפי מעלה או מטה, ובסופו של דבר לעצב ולערער החלטות ומערכות של הביטחון הלאומי כמכלול.

המסגרת הטקטית־אופרטיבית־אסטרטגית

הבנת ההשפעות של הרעלת נתונים דורשת מסגרת מושגית לדרך שבה לפעולות מקומיות יש השלכות ברמה הלאומית. המודל המסורתי הטקטי־אופרטיבי־אסטרטגי מספק עדשה מעין זו. חלוקת הפעילות הצבאית לרמה טקטית, אופרטיבית ואסטרטגית ממשיכה להיות המסגרת הבסיסית לתרגום יעדים פוליטיים לתוצאות צבאיות. הרמה הטקטית עוסקת בפעולות מיידיות והשפעותיהן; הרמה האופרטיבית מחברת מאמצים מרובים להשגת תוצאות בזירה מסוימת; והרמה האסטרטגית מכוונת משאבים ותכלית ברמה הלאומית לעבר יעדים מוגדרים. מבנה זה מציע שפה משותפת היוצרת רציפות לוגית בין כוונה ליישום על פני זמן ומרחב.[6]

6 Department of Defense, *Joint Publication (JP) 1 – Doctrine for the Armed Forces of the United States* (U.S. Government Publishing Office, 2017), https://tinyurl.com/xj65t8z4;

עם זאת הנחות אלו נתונות ללחץ גובר והולך בתחומי המידע והסייבר. בלוחמה קונוונציונלית, פעולות טקטיות נוטות להצטבר בהדרגה לכדי תוצאות אסטרטגיות. לעומת זאת מבצעי סייבר ומבצעים קוגניטיביים יכולים לחולל השפעה בלתי פרופורציונלית מאירוע מקומי יחיד. אירועי Stuxnet (2010) ו־NotPetya (2017) המחישו את היקף הדחיסה הזה: ניצול נקודת תורפה (exploit) יחידה או מערך נתונים פגום יכולים להתפשט הרחק מעבר להיקף המיועד שלהם, לשבש מערכות ולשנות דינמיקות פוליטיות רחבות יותר.[7] אי־ליניאריות כזו מחלישה מודלים של פיקוד המסתמכים על הסלמה צפויה או נשלטת ואפקט מצטבר. המבנה הטקטי־אופרטיבי־אסטרטגי נותר רלוונטי, אך כעת הוא מחייב התאמה כדי לכלול במערך השיקולים גם אינטראקציות מתגלגלות ובין־תחומיות, ולא רק תהליכים ליניאריים של סיבה ותוצאה.[8]

לוחמה קוגניטיבית מאתגרת עוד יותר את הלוגיקה הזאת בכך שהיא מסיטה את התחרות משליטה בשטחים לשליטה בפרשנויות. היריבים משלבים מניפולציות מקומיות, כגון מדיה מפוברקת או הגברה אוטומטית, לכדי קמפיינים נרחבים יותר המקדמים נרטיבים אסטרטגיים מתמשכים, שנועדו לערער את האמון או להשפיע על השיפוט הקולקטיבי. המסגרת הטקטית־אופרטיבית־אסטרטגית שומרת על ערכה מכיוון שהיא מסייעת לזהות כיצד פעולות יחידות נמצאות בהלימה עם קמפיינים רחבים יותר ויעדי השפעה ארוכי טווח. עם זאת מבצעים קוגניטיביים מתפשטים באמצעות רשתות הכוללות שחקנים מדינתיים,

---

Andrew S. Harvey, "The Levels of War as Levels of Analysis," *Military Review*, December 2021, https://tinyurl.com/yw5e7w3h.

7 Stuxnet, שנחשפה ב־2010, הייתה נוזקה שיועדה לפגוע במערכות בקרה תעשייתיות, בעיקר בצנטריפוגות בתוכנית הגרעין האיראנית. NotPetya, שהופצה ב־2017 במסווה של התקפת כופרה, פגעה תחילה באוקראינה, אך התפשטה גלובלית וגרמה לנזקים נרחבים לחברות ולתשתיות.

8 Sara Sörensen and James Pamment, *Operationalising the Framework for Evaluating Capability Against Information Influence Operations: Case Study of the Psychological Defence Agency's Courses* (NATO Strategic Communications Centre of Excellence, 2023); Matthias Schulze, "Cyber in War: Assessing the Strategic, Tactical, and Operational Utility of Military Cyber Operations," 2020 12th International Conference on Cyber Conflict (CyCon), May 2020, 183–97, https://doi.org/10.23919/CyCon49761.2020.9131733; Frederik A. H. Pedersen and Jeppe T. Jacobsen, "Narrow Windows of Opportunity: The Limited Utility of Cyber Operations in War," *Journal of Cybersecurity* 10, no. 1 (2024), https://doi.org/10.1093/cybsec/tyae014.

מסחריים וחברתיים ולא רק צבאיים. כתוצאה מכך מסגרת זו צריכה להתפתח מדוקטרינת תכנון לכלי אנליטי רחב יותר המשמש ליישום של משילות מרובת שחקנים ויצירת חוסן לאומי.[9]

שכבות טכנולוגיית המידע: המקבילה ההנדסית

מורכבות מנוהלת בעולם הטכני באמצעות חלוקה מופשטת לשכבות. מהנדסים ואדריכלי מערכות מחלקים את הפונקציונליות למישורים נפרדים — לרוב לשכבת היישום או התצוגה, לשכבת הלוגיקה או העיבוד ולשכבת הנתונים המעגנת אותן. מבנה תלת־שכבתי זה חוזר על עצמו בכל עולם המחשוב, מבניית תוכנה ועד לארכיטקטורות ענן והנדסת רשתות. החלוקה לשכבות מאפשרת להפריד בין תחומי העניין: לכל רמה יש תפקידים וממשקים מוגדרים, המאפשרים למומחים לפתח ולהרחיב מערכות באופן עצמאי מבלי להפריע לאינטגרציה הכוללת.[10]

אנלוגיה שימושית מגיעה מתחום אבטחת הענן, שבו האחריות לניהול סיכונים מתחלקת בין מספר שחקנים. מודלי ענן מודרניים ממסדים את העיקרון הזה באמצעות מסגרת "האחריות המשותפת" וההבחנות המוכרות בין תשתית, לפלטפורמה ולתוכנה כשירות (IaaS, PaaS, SaaS). שכבות אלו אינן מוסכמות טכניות בלבד; הן מייצגות מערכות יחסים כלכליות ואופרטיביות, המגדירות היכן הצטברות של סיכונים מתרחשת, וכיצד אחריות ופעולות תיקון מתחלקות לאורך שרשרת האספקה הדיגיטלית.[11]

מודל "הבניין בן שלוש הקומות"

מחקר זה מציג מודל של "בניין בן שלוש קומות", המשלב את נקודת המבט הצבאית ואת נקודת המבט ההנדסית לכדי מסגרת אנליטית אחת. כל קומה מתפקדת כשכבה מובחנת אך בעלת תלות הדדית באקוסיסטם של הבינה המלאכותית, ויש לה היגיון פנימי, תכלית ומשטח איומים משלה. יחד הן יוצרות מבנה של קשרים, גם אם לא תמיד רציף, שבו שיבושים בכל נקודה יכולים לנוע בין השכבות ולגרום להשלכות תפעוליות או אסטרטגיות.

9 Michael J. Cheatham et al., "Cognitive Warfare: The Fight for Gray Matter in the Digital Gray Zone," *Joint Force Quarterly* 114, no. 3 (2024): 83–91.

10 "What Is Three-Tier Architecture?," IBM – Think, n.d., accessed November 19, 2025, https://tinyurl.com/34z72zmt.

11 "IaaS vs. PaaS vs. SaaS," Red Hat, n.d., https://tinyurl.com/4kyubtxb.

הקומה הראשונה — שכבת הנתונים — מניחה את המסד. היא כוללת את האופן שבו מידע נכנס למערכות, מאורגן בהן ונע דרכן. רמה זו מפקחת על מקור הקלטים ושלמותם — חומרי הגלם המעצבים את למידת המכונה. נקודות תורפה נוצרות באמצעות ביצוע שינויים במערכי נתונים, שיבוש של תהליכי תיוג או פריצה לערוצי אחזור כגון חבלה בממשקי תכנות יישומים (APIs) או ביצוע מניפולציה על ארכיוני רשת. מכיוון ששכבת הנתונים מזינה את כל הפונקציות האחרות, פגיעה יחידה יכולה להתפשט בחשאי לכל רחבי שרשרת העיבוד.

הקומה השנייה — שכבת המודלים והאלגוריתמים — הופכת נתונים גולמיים לייצוגים מתמטיים הנלמדים על־ידי המכונה ולכללים המאפשרים קבלת החלטות. היא כוללת ארכיטקטורות מודלים, תהליכי אימון, תהליכי אופטימיזציה ואת הפרמטרים שמקודדים דפוסים סטטיסטיים. נקודות תורפה במפלס זה נוצרות באמצעות גרדיאנטים מורעלים, נקודות ביקורת עם דלת אחורית, מודלים מאומנים מראש שנפרצו או תהליכי כוונון עדין שעברו מניפולציה. היות ששכבה זו מתווכת בין שכבת הנתונים לשכבת היישום, עיוותים המוחדרים אליה יכולים לעצב מחדש את ההתנהגות במורד הזרם גם כשהנתונים הבסיסיים נראים נקיים.

הקומה השלישית — שכבת היישום — מתייחסת לאופן שבו מערכות בינה מלאכותית יוצרות אינטראקצייה עם קוגניציה אנושית ומערכות חברתיות. זהו המרחב שבו מודלים משפיעים על מה שאנשים תופסים כאמיתי ורלוונטי — מרחב שבו פלטים של מכונות מתחילים לעצב את ההבנה וההתנהגות הקולקטיבית. הבינה המלאכותית במפלס זה אינה משתלבת סתם כך בתהליכי העבודה; היא הופכת לשותפה פעילה בסביבת המידע ובעלת יכולת להסיט את תשומת הלב ולשנות את האופן שבו המשמעות נבנית. היריבים מנצלים מרחב זה באמצעות קמפיינים של דיפ־פייק (deepfake), פרסונות שנוצרו באמצעות בינה מלאכותית, מנועי תעמולה אדפטיביים ורשתות השפעה אוטומטיות, והופכים כלים אלגוריתמיים למכשירים של מניפולציה אפיסטמית ולחץ פסיכולוגי. אם כן קומת היישום היא המקום שבו עיוותי נתונים או אלגוריתמים באים לידי ביטוי כשינויים בהשקפת הציבור, באמון במוסדות ובביטחון בתהליכי קבלת ההחלטות.

דוח זה מתמקד בשכבת הנתונים, שכן היא נותרה במידה רבה מחוץ למוקד הדיון בשיח הביטחון הלאומי. לעיתים קרובות מתייחסים לנתונים כאל משאב ניטרלי או פרט הנדסי שיטופל בשלב מאוחר יותר של תהליך המידול. בפועל, האופנים — שבהם נתונים נוצרים, נאספים, מובנים ומאומתים — קובעים כיצד המודלים פועלים והאם ניתן לסמוך על הפלטים שלהם במצבי לחץ. הכשלים ברמה זו הם עדינים, אך יש להם השלכות משמעותיות,

וביכולתם לעוות את התפיסה ולהחליש את החוסן המוסדי או הקוגניטיבי. עבור מדעני מחשבים ונתונים, מסגרת שלוש הקומות מבהירה מדוע בחירות טכניות — כגון תקני תיוג או אימות מקורות — הופכות לבעלות משמעות אסטרטגית, לדוגמה בעיצוב תכנון צבאי וניהול משברים. הפרקים הבאים יעסקו בהרחבה בכל קומה לפי הסדר, ויקשרו בין מנגנונים טכניים ספציפיים לסיכונים בתפעול ולהחלטות בנוגע למדיניות שהם מעצבים, כדי שמומחי טכנולוגיה ואנשי ביטחון לאומי יוכלו לפעול מתוך הבנה משותפת שניתנת ליישום.

איור 1: "בניין בן שלוש קומות"

חלק 2

מנגנונים ושחקנים של הרעלת נתונים

2

2.0
הגדרה של הרעלת נתונים

הרעלת נתונים היא צורה ייעודית של התקפה מצד היריבים המתרחשת בשלב האימון של מודל למידת מכונה (ML). היא מתייחסת למניפולציה או להשחתה מכוונת של נתוני אימון המשמשים לבניית מערכות בינה מלאכותית או למידת מכונה, כדי לשבש או לפגוע בתהליך הלמידה.[12]

המטרה הסופית של הגורם המרעיל היא לחתור תחת התחזיות של מערכות למידת מכונה על ידי התערבות בשלב האימון. הדבר יכול לכלול פגיעה בביצועים הכוללים של המודל ופגיעה במהימנות שלו או ביצוע מניפולציה על התנהגותו כדי להפיק תוצאות מוטות, לא מדויקות או מזיקות. מנגנון ההתקפה כולל החדרה של נתונים פגומים או כאלה שעברו מניפולציה למערך נתוני האימון, ובכך הוא יוצר השפעה סיבתית על תהליך הלמידה של המודל.[13]

משמעות הדבר במונחים מעשיים היא שהיריב אינו צריך לפרוץ לקוד או לרשת של המערכת; הוא יכול לבצע מניפולציה על מה שהמודל לומד לבטוח בו. מרגע שהנתונים המורעלים נטמעים בתהליך האימון, המערכת יכולה להוסיף ולתפקד כרגיל למראית עין, בעוד בפועל היא מפיקה הכרעות מעט מעוותות שמשרתות את האינטרסים של התוקף.

אפשר לסווג התקפות הרעלה על פי המיקום שבו הן מתרחשות בתהליך האימון. בהתקפת הרעלת נתונים (DPA), היריבים מבצעים מניפולציה על חלק מנתוני האימון, לעיתים באמצעות מקורות חיצוניים או לא מאומתים כגון מערכי נתונים של צד שלישי,

12 Luis Muñoz-González et al., "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization," version 1, preprint, arXiv, 2017, https://doi.org/10.48550/ARXIV.1708.08689; Marek Pawlicki et al., "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models," *Neurocomputing* 653 (2025), https://doi.org/10.1016/j.neucom.2025.131231.

13 Antonio Emanuele Cinà et al., "Wild Patterns Reloaded: A Survey of Machine Learning Security Against Training Data Poisoning," *ACM Computing Surveys* 55, no. 13s (2023): 1-39, https://doi.org/10.1145/3585385; Zhibo Wang et al., "Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems," *ACM Computing Surveys* 55, no. 7 (2022): 1-36, https://doi.org/10.1145/3538707.

תיוג באמצעות מיקור המונים או תוכן שנאסף מהרשת. התקפות אלו פוגמות בדפוסים הסטטיסטיים שהמודל לומד, ומטמיעות אותות מוטים, דוגמאות בעלות תיוג שגוי או טריגרים נסתרים שנראים תמימים ברמת הנתונים.

בהתקפת הרעלת מודל (MPA), היריבים משפיעים באופן ישיר על המודל. הדבר עשוי להתרחש באמצעות פגיעה בנקודות בדיקה של מודלים שעברו אימון מקדים, החדרת גרדיאנטים זדוניים בתהליך של אימון מבוזר או שינוי פרמטרים בשלב הכוונון העדין. במקום להשחית קלטים, התקפת הרעלת מודל משנה את המבנה הפנימי של המודל עצמו, וכך מתאפשר לתוקף להטמיע יכולות שאינן בהכרח ניתנות לזיהוי מתוך נתוני האימון.

שתי צורות ההתקפה שואפות לפגוע בשלמות המודל, אך הרעלת הנתונים משחיתה את הקלטים, ואילו הרעלת המודל מכוונת לפרמטרים של המודל ולדינמיקת הלמידה שלו.[14]

דרך נוספת להבחין בין סוגי הרעלה היא לפי יעדו של התוקף. התקפות לא ממוקדות (התקפות על הזמינות) שואפות לפגוע בביצועים ללא הבחנה, בעוד התקפות ממוקדות (התקפות על השלמות) נועדו לשנות פלטים ספציפיים או לגרום לסיווג שגוי בדגימות נבחרות.[15] התקפת דלת אחורית היא צורה מיוחדת שבה טריגר נסתר גורם למודל להתנהג בצורה שגויה רק כשהטריגר הספציפי מופיע.[16] הבחנות אלו ייבחנו בהרחבה בהמשך, אך בבסיסן כולן מנצלות את אותו העיקרון: שליטה בנתונים המעצבים את הדרך שבה המודל מבין את העולם.

לבסוף, ב־LLMs הרעלת נתונים מנצלת את תהליך האימון הרב־שלבי שלהם — אימון מקדים, כוונון עדין ולמידה באמצעות חיזוקים ממשוב אנושי (RLHF). תוקפים יכולים להזריק נתונים שעברו מניפולציה בכל אחד מהשלבים האלה כדי להטמיע הטיות או התנהגויות סמויות, כגון לגרום למודל לפרש בצורה שגויה נושאים מסוימים או להגיב בצורה חריגה כשנעשה שימוש בביטויים ספציפיים.[17]

14 Apostol Vassilev et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2023 (National Institute of Standards and Technology, 2024), https://doi.org/10.6028/NIST.AI.100-2e2023; Wang et al., "Threats to Training," 1-36.

15 Muñoz-González et al., "Towards Poisoning."

16 Pinlong Zhao et al., "Data Poisoning in Deep Learning: A Survey," version 1, preprint, arXiv, 2025, https://doi.org/10.48550/ARXIV.2503.22759.

17 Vassilev, *Adversarial Machine Learning*; Zhao et al., "Data Poisoning in Deep Learning."

במונחים של ביטחון לאומי, הרעלת נתונים מייצגת התקפה לא על רשת, אלא על התשתית הקוגניטיבית של מערכות מבוססות־בינה מלאכותית. היא מאפשרת ליריבים לשנות בחשאי את האופן שבו מערכות אלו תופסות ומסיקות, ובכך להפוך טכנולוגיה מהימנה לאמצעי הטעיה בלוחמה קוגניטיבית.[18]

איור 2: הסוגים העיקריים של התקפות הרעלת נתונים

18 Fran Casino, "Unveiling the Multifaceted Concept of Cognitive Security: Trends, Perspectives, and Future Challenges," *Technology in Society* 83 (2025), https://doi.org/10.1016/j.techsoc.2025.102956.

2.1
כיצד נתוני אימון נכנסים למודלים

מודלים של למידת מכונה, ובפרט רשתות נוירונים עמוקות (DNNs), עוברים אימון באמצעות מנגנון מובנה הממיר נתונים גולמיים לפרמטרים של המודל. כל שלב בתהליך זה מסתמך על התקינות של הקלטים שלו, כך ששגיאות או מניפולציות המוחדרות בכל שלב יכולות להתפשט באופן בלתי נראה ברחבי המערכת.

שלב איסוף הנתונים הוא בדרך כלל החשוף ביותר להתערבות עוינת. מאמני מודלים מאמצים לעיתים קרובות מערכי נתונים ציבוריים או נשענים על נתונים שמקורם באינטרנט בשל מגבלות משאבים, ובכך מובילים לתלות משמעותית במקורות חיצוניים שרמת מהימנותם אינה ודאית. השלב הזה קריטי מכיוון שנתונים, הנאספים ממקורות מגוונים ולעיתים בלתי מהימנים, פגיעים במיוחד להחדרה של דגימות מורעלות. המגמה לשילוב מנגנוני אחזור דינמיים — למשל מערכות הנשענות באופן רציף על API, אינדקסים של הרשת או בסיסי ידע ארגוניים — מרחיבה עוד יותר את הפוטנציאל לחשיפה, מכיוון שהרעלה יכולה להתרחש כעת הן לפני פריסת המודל והן במהלך מחזור החיים התפעולי שלו.[19]

לאחר האיסוף, העיבוד המקדים של הנתונים כולל ניקוי, שיפור, המרה ונרמול. המטרה היא להבטיח שלמות, הוגנות ויציבות לפני חלוקת הנתונים לתתי־מערכים של אימון ובדיקה. אולם העיבוד המקדים, על אף שמטרתו היא סילוק חריגות, עלול באופן פרדוקסלי להסתיר מניפולציות זדוניות: נורמליזציה סטטיסטית, הסרת כפילויות או חילוץ מאפיינים עשויים להסיר חריגות שטחיות שאחרת היו יכולות לשמש כרמזים לכך שבוצעו פעולות חבלה. אימות לקוי בשלב זה מאפשר לדוגמאות מורעלות לחדור ללא זיהוי אל מערכי נתוני האימון, ובכך לבסס את מה שהחוקרים מכנים "השחתה שקטה".[20]

במהלך הלמידה והמידול, האלגוריתם שואף ליצור מיפוי מקורב בין קלט לפלט על ידי אופטימיזציה של הפרמטרים כדי להקטין למינימום את פונקציית ההפסד. זהו השלב שבו

19 Cinà et al., "Wild Patterns Reloaded," 1-39; Wang et al., "Threats to Training," 1-36.

20 Matthew Jagielski et al., "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," version 3, preprint, arXiv, 2018, https://doi.org/10.48550/ARXIV.1804.00308; Vassilev, *Adversarial Machine Learning*.

הייצוג הפנימי של המודל לגבי העולם מתגבש מתוך סביבת הנתונים שלו. אם חלק כלשהו מנתוני האימון יעוות בצורה מעודנת, הייצוגים הנלמדים יקודדו את העיוותים האלה, לעיתים קרובות בדרכים השומרות על עקביות סטטיסטית עם הדגימות הלא נגועות. התוצאה היא מערכת המתנהגת כרגיל בתהליכי אימות, אך סוטה כשהיא נחשפת לתנאים מסוימים, ובכך הופכת לכלי אידיאלי למניפולציה קוגניטיבית.[21]

לבסוף, שלב הערכת המודל מספק הערכה שלאחר האימון באמצעות מערך אימות נקי. בהתקפות הרעלה מתוכננות היטב, המודל ממשיך להפגין ביצועים תקינים על הנתונים הנקיים האלה, שומר על רמת דיוק גבוהה ואינו מעורר חשד. הדבר מאפשר למודל המורעל לעמוד במדדי איכות סטנדרטיים ולעבור הערכה שגרתית. מעבר לדיוק, התקפות יעילות חומקות גם מבדיקות שאינן קשורות לביצועים כגון גלאי חריגות, ביקורות תאימות או מבחני הסבריות (explainability).[22] התוצאה היא שמודל עשוי לקבל הסמכה מלאה לתפעול, אף שהוא מכיל מניפולציות נסתרות שייחשפו רק בתנאים הנשלטים על ידי התוקף — בדיוק מסוג ההתנהגויות שרלוונטיות לתרחישי לוחמה קוגניטיבית.

לכן ניתן להמשיג את תהליך האימון כרצף של יחסי תלות המבוססים על אמון. כל שלב המרה הופך קלט גולמי לצורה מופשטת יותר, ומעצים את ההשפעות של כל זיהום שהתרחש בשלב מוקדם יותר.

במערכות בינה מלאכותית יוצרת (GenAI), ובייחוד ב־LLMs, נתיב מעבר הנתונים עובר דרך מספר שכבות, וכל אחת מהן יוצרת הזדמנויות ייחודיות להרעלה. שלב האימון המקדים נשען על אוספים עצומים של טקסט, הנאספים מרחבי האינטרנט ועוברים סינון חלקי בלבד, ובכך נוצר מרחב פעולה עבור היריבים לשתילת מניפולציות עדינות במקורות מקוונים מגוונים. השלב הבא, כוונון עדין מפוקח (SFT) וכוונון הוראות או הנחיות, משתמש במערכי נתונים קטנים וממוקדים הרבה יותר כדי לעצב את התנהגות המודל. מאחר שהמערכים האלה הם מוגבלים וקלים יותר להשפעה, תוקפים יכולים להחדיר בהם הרעלה ממוקדת, כגון דוגמאות מטעות או צמדים מוטים של הוראה–תגובה. לבסוף, מידול של תגמול ו־RLHF משלבים משוב אנושי שמטרתו להתאים את המודל לסטנדרטים של בטיחות ונורמטיביות. אם נתוני המשוב או תהליכי התיוג נפגעים, היריב יכול לשנות את ההתאמה

21 Vassilev, *Adversarial Machine Learning*; Wang et al., “Threats to Training,” 1-36.

22 Wang et al., “Threats to Training,” 1-36; Yihe Zhou et al., “A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluation Methods,” *Transactions on Artificial Intelligence* 1, no.1 (2025): 28-58, https://doi.org/10.53941/tai.2025.100003.

המעשית של המודל ולהטמיע נקודות תורפה התנהגותיות שמופיעות רק בנוכחות הנחיות מסוימות או בהקשרים ספציפיים של תפעול.[23]

טבלה 1: נקודות תורפה בתהליך למידת המכונה

שלב	פונקציית הליבה	מוקדי מינוף פוטנציאליים למניפולציה
איסוף נתונים	איחוד מידע גולמי ממקורות מגוונים או פתוחים	הישענות על ערוצי צד שלישי או מנגנוני שליפה דינמיים מאפשרת החדרה רחבת היקף של רשומות שעברו מניפולציה.
עיבוד מקדים של הנתונים	ניקוי, המרה והכנה של מערכי נתונים מובנים	אימות חלש עלול לנרמל או להסתיר דגימות מורעלות במקום להסירן.
למידה ומידול	אופטימיזציה של פרמטרים על בסיס נתוני האימון	דגימות פגומות משפיעות על עדכוני הגרדיאנטים, ומטמיעות התנהגות שתוכננה על ידי התוקף בתוך משקלי המודל.
הערכת המודל	אימות ביצועים באמצעות נתונים נפרדים	התקפות שנועדו להיות חשאיות שומרות על דיוק גבוה בבדיקות ועל התנהגות לקויה סמויה.

אף שהתקפות כאלה נותרות ברובן תאורטיות ב־LLMs שנפרסו, ההיתכנות של כל שלב כווקטור תקיפה זכתה להכרה רחבה במחקרים אחרונים על אבטחה של מודלים גנרטיביים. המאפיין המגדיר שלהם הוא שהרעלה כבר אינה מתרחשת אך ורק בתוך מערכי נתונים סטטיים: היא מתרחבת גם לתהליכי אימון אינטראקטיביים ואיטרטיביים, שבהם משוב אנושי, אחזור דינמי ועדכונים רציפים יכולים לשמש נשאים של מניפולציות.[24]

23 Vassilev, *Adversarial Machine Learning*; Zhou et al., "Survey on Backdoor Threats," 28-58.

24 Department of Homeland Security, *Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*, Preparedness Series (2023), https://tinyurl.com/tmry37pn; Neil Fendley et al., "A Systematic Review of Poisoning Attacks Against Large Language Models," version 1, preprint, arXiv, 2025, https://doi.org/10.48550/ARXIV.2506.06518; Daryna Oliynyk et al., "I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences," *ACM Computing Surveys* 55, no. 14s (2023):1-41, https://doi.org/10.1145/3595292; Pawlicki et al., "Meta-Survey."; Anwar Shah et al., "Guarding the Gates: A Comprehensive Survey of Backdoor Attacks on Neural Networks,"

תהליכי אימון — המערכים המקיפים שאוספים נתונים, מעבדים אותם ומזינים אותם למערכת אימון המודל — אינם רק מבנים טכניים, אלא מערכות אפיסטמולוגיות: מנגנונים שבאמצעותם ארגונים לומדים על הסביבה שלהם. כאשר היריבים מרעילים את הנתונים שמזינים מערכות אלו או את התהליכים שבאמצעותם הלמידה מתבצעת, התוצאה אינה מאגר נתונים פגום גרידא, אלא תפיסת עולם משובשת. עבור יישומי ביטחון לאומי, סיכון זה הופך פגיעות טכנית לשאלה של ריבונות קוגניטיבית: מי שולט בסופו של דבר במה שהמערכות שלנו "יודעות".

איור 3: אנטומיה של התקפה: תהליך למידת המכונה

preprint, SSRN, 2024, https://doi.org/10.2139/ssrn.4966942; Vassilev, *Adversarial Machine Learning*; Zhao et al., "Data Poisoning in Deep Learning."

2.2
סוגים של התקפות הרעלה

2.2.1 התקפות הרעלת נתונים

התקפות הרעלת נתונים (DPAs) כרוכות בכך שהיריב שולט בחלק מנתוני האימון על ידי הוספה או שינוי של דגימות אימון. אפשר להבין אותן על פי שני ממדים עיקריים: מטרת היריב ואסטרטגיית המניפולציה.

מבחינת הכוונה, התקפות הרעלה לא ממוקדות — שנקראות לעיתים התקפות על הזמינות — שואפות להגדיל ככל הניתן את שגיאת הסיווג ללא אבחנה. מטרתן היא לפגוע בביצועים הכוללים של המודל, לשבש את תהליך ההתכנסות של האימון או ליצור אפקט של מניעת שירות במערכת הלמידה. התקפות אלו הן יחסית קלות יותר לביצוע ונוטות לפעול באופן דומה על פני ארכיטקטורות שונות.[25] לעומת זאת התקפות הרעלה ממוקדות או התקפות על שלמות המידע (integrity attacks) נועדו ליצור שגיאות סיווג או הטיות ספציפיות מאוד עבור תת־קבוצה מוגדרת מראש של קלטים ולשמור על הדיוק הכולל של המודל.[26]

תת־הקבוצה המתוחכמת ביותר מבין אלו היא התקפות הרעלת דלת אחורית (backdoor poisoning attacks), שבהן מתבצעת הטמעה של טריגר חבוי בנתוני האימון, כך שהמודל מתנהג כרגיל עד להופעת הטריגר בקלט בדיקה, ואז הוא מפיק פלט שהגדיר התוקף. הדלתות האחוריות, שהן התקפות של "סוסים טרויאניים" על רשתות עצביות (neural trojans), עשויות לפעול באחד משני דפוסים: מיפוי של כל הקלטים ליעד יחיד (all-to-one) או מיפוי שיטתי בין כלל הקטגוריות (all-to-all) בהתאם לאופן שבו התוקף רוצה שהמודל יתנהג. בהתקפת all-to-one, כל דוגמה מורעלת — ללא קשר לקטגוריה המקורית שלה —

25 Cinà et al., "Wild Patterns Reloaded," 1-39; Muñoz-González et al., "Towards Poisoning."; Pawlicki et al., "Meta-Survey."; Wenjun Qiu, "A Survey on Poisoning Attacks Against Supervised Machine Learning," version 2, preprint, arXiv, 2022, https://doi.org/10.48550/ARXIV.2202.02510; Vassilev, *Adversarial Machine Learning*.

26 Cinà et al., "Wild Patterns Reloaded," 1-39; Deepon Halder et al., "A Comprehensive Survey of Data Poisoning Attacks and Their Detection Techniques," preprint, 2025, https://doi.org/10.13140/RG.2.2.20084.67207; Jonas Geiping et al., "Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching," version 2, preprint, arXiv, 2020, https://doi.org/10.48550/ARXIV.2009.02276.

תמופה בהכרח לתווית יעד אחת קבועה בכל פעם שהטריגר החבוי מופיע. אנלוגיה פשוטה היא מצלמת אבטחה: אם תוקף מטמיע דלת אחורית מסוג all-to-one, כל מה שנושא מדבקה קטנה או סמל (למשל אדם, כלי רכב או רחפן) עשוי להיות מסווג תמיד כ"אובייקט בטוח" גם אם הוא זדוני. לעומת זאת התקפת all-to-all עושה שימוש בטריגרים שונים עבור סיווגים שונים, ומפנה כל אחד מהם לתווית שגויה שונה. אותה מערכת מצלמות יכולה לתייג בטעות אנשים ככלבים, כלבים כחבילות וכלי רכב כאנשים, כשהטריגר המתאים נוכח. התקפות all-to-one יוצרות נקודה עיוורת ממוקדת, בעוד התקפות all-to-all יוצרות בלבול שיטתי ורחב יותר.[27]

בהמשך למסגרת זו, מחקרים עדכניים יותר מתארים הרעלה של תת־אוכלוסייה (subpopulation poisoning), שבה התוקף מבצע מניפולציה על נתונים השייכים לפלח דמוגרפי או לקבוצת סיווג מסוימת, ושומר על מדדי האימות הכוללים. גרסה זו מסוכנת במיוחד בהקשרים של מדיניות או מודיעין, משום שהיא מטמיעה הטיה מערכתית ללא פגיעה גלויה בדיוק שלה.[28] לדוגמה, מודל לניתוח סנטימנט המשמש לזיהוי הקצנה ברשת עשוי לתפקד כמצופה עבור האוכלוסייה הכללית, אבל עקב הרעלה ממוקדת יסווג בעקביות פרסומים של מיעוט אתני או דתי מסוים כ"בעלי סיכון גבוה". המערכת נראית מדויקת בשלב הבדיקות, אך בפועל התנהגותה מוטה בחשאי כנגד אותה תת־אוכלוסייה, ומפיקה הערכות מודיעין מעוותות מבלי להפעיל מנגנוני התרעה.

לאחר שהוצגו מטרות היריב, ניתן לפנות כעת לאסטרטגיות המניפולציה שבאמצעותן התוצאות האלה מושגות:

- התקפות היפוך תוויות (label-flipping), הידועות גם כהתקפת dirty-label, פועלות על ידי שינוי התוויות של תת־קבוצה נבחרת של נתוני האימון מבלי לגעת בתכונות הקלט.[29]
- התקפות שיבוש התכונות או הקלט (feature or input-perturbation attacks) או התקפת clean-label, משנות את מאפייני דגימות האימון אך שומרות על תוויות נכונות, ובכך מסתירות את המניפולציה מבדיקות סטנדרטיות של איכות הנתונים. הצורות החשאיות ביותר של התקפת clean-label מסתמכות על טכניקות של התנגשות תכונות

27 Cinà et al., "Wild Patterns Reloaded," 1-39; Qiu, "Survey on Poisoning Attacks."; Vassilev, *Adversarial Machine Learning*.

28 Matthew Jagielski et al., "Subpopulation Data Poisoning Attacks," version 3, preprint, arXiv, 2020, https://doi.org/10.48550/ARXIV.2006.14026.

29 Miguel A. Ramirez et al., "Poisoning Attacks and Defenses on Artificial Intelligence: A Survey," version 2, preprint, arXiv, 2022, https://doi.org/10.48550/ARXIV.2202.10276.

או התאמה של גרדיאנטים או מטא־הרעלה, המטמיעות טריגרים של דלת אחורית בתוך מרחב הייצוג הנלמד ללא כל שינוי נראה לעין בנתונים או בתוויות.[30]

- התקפות הזרקת נתונים שונות בכך שהיריב מחדיר דגימות מפוברקות וחדשות לחלוטין לתוך מערך האימון, ולעיתים קרובות זה נעשה באמצעות הטמעת תבנית טריגר ותווית יעד מתאימה כדי ליצור דלת אחורית קבועה.[31]

מעבר לקטגוריות הראשיות האלה קיימות מספר טכניקות המשכללות או מרחיבות התקפות הרעלה מבלי להיות סוגים נפרדים. הרעלה מבוססת־אופטימיזציה מתייחסת להתקפה כאל בעיית חיפוש מובנה, ומתאימה דוגמאות אימון נבחרות כדי לזהות את ההפרעה המינימלית שגורמת באופן אמין לאפקט המזיק הרצוי.[32] גישות גנרטיביות משתמשות ברשתות גנרטיביות יריבות (GANs) או במקודדים אוטומטיים (autoencoders) כדי ליצור דוגמאות מורעלות ריאליסטיות או מיישמות שיטות מבוססות־השפעה כדי לזהות את נקודות האימון המשמעותיות ביותר לצורך השחתה.[33] בניית דלת אחורית מבוססת־טריגר מטמיעה מנגנוני הפעלה סמויים בשלב ההסקה, השונים מסימני מים לאימות המקור (provenance watermarks), מכיוון שמטרתם היא הפעלה פונקציונלית ולא זיהוי.[34] לבסוף, הרעלת מקורות נתונים מרחיבה רעיונות אלו במעלה הזרם אל תוך שרשרת אספקת הנתונים, שבה תוקפים מבצעים מניפולציות במערכי נתונים שנאספו או שוכפלו או מנצלים היוריסטיקות איסוף כדי לזרוע רעלים זמן רב לפני תחילת אימון המודל.

30 Quang H. Nguyen et al., "Wicked Oddities: Selectively Poisoning for Effective Clean-Label Backdoor Attacks," preprint, arXiv, July 16, 2024, https://doi.org/10.48550/arXiv.2407.10825; Ali Shafahi et al., "Poison Frogs !Targeted Clean-Label Poisoning Attacks on Neural Networks," in *Advances in Neural Information Processing Systems*, ed. S. Bengio et al. (Curran Associates, Inc., 2018), https://tinyurl.com/yc7zabrn; Chen Zhang et al., "Clean-Label Poisoning Attack with Perturbation Causing Dominant Features," *Information Sciences* 644 (2023), https://doi.org/10.1016/j.ins.2023.03.124.

31 Department of Homeland Security, *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*, Preparedness Series (2023), https://tinyurl.com/2fv6whkf.

32 Cinà et al., "Wild Patterns Reloaded," 1-39; Qiu, "Survey on Poisoning Attacks."; Shafahi et al., "Poison Frogs."; Zhao et al., "Data Poisoning in Deep Learning."

33 Cinà et al., "Wild Patterns Reloaded," 1-39; Zhao et al., "Data Poisoning in Deep Learning."

34 Geiping et al., "Witches' Brew."; Yuan Ma et al., "Backdoor Attack with Invisible Triggers Based on Model Architecture Modification," version 3, preprint, arXiv, 2024, https://doi.org/10.48550/arXiv.2412.16905; Qiu, "Survey on Poisoning Attacks."

טבלה 2: הסבר על שיטות הרעלת נתונים: דוגמאות מעשיות לתחום הביטחון הלאומי

שיטה	הסבר שאינו טכני	דוגמה
היפוך תוויות (התקפת Dirty-Label)	התוקף משנה את "התשובה" המצורפת לחלק מדוגמאות האימון, אך משאיר את התוכן ללא שינוי.	מערכת לזיהוי תמונות מרחפן עוברת אימון באמצעות דוגמאות שתויגו באופן שגוי כך שתמונות של רכבי אויב מתויגות בצורה שגויה כ"משאיות אזרחיות", וכתוצאה מכך המודל מדווח באופן חלקי על איומים בהנות ISR.
התקפת Clean-Label/ שיבוש תכונות	התוקף משנה במקצת את תוכן דוגמאות האימון, אך שומר על תקינות התוויות, ובכך מקשה על זיהוי המניפולציה.	מערכת זיהוי פנים, המשמשת לבקרת גישה במתקן מאובטח, מאומנת באמצעות תמונות של גורם פנימי שתמונותיו שונו במקצת. הגורם הפנימי יכול להיכנס בשלב הפריסה מבלי להתגלות, מכיוון שהמודל למד גרסה מעוותת של פניו.
הרעלת התנגשות תכונות/ התאמת גרדיאנט (דלתות אחוריות נסתרות)	התוקף יוצר דוגמאות מורעלות שנראות תקינות, אך גורמות למודל להתנהג בצורה שגויה, כשטריגר נסתר מופיע.	כלי ניתוח של תצלומי לוויין מסווגים בטעות כל ספינה הנושאת סמל מצויר מסוים כ"ידידותית", ובכך מתאפשר לכלי שיט עוינים מסווים לחמוק מגילוי.
התקפות הזרקת נתונים	התוקף מחדיר דוגמאות אימון מזויפות לחלוטין לתוך מערך הנתונים, ולעיתים קרובות מטמיע דפוס שמפעיל דלת אחורית.	בינה מלאכותית לאבטחת סייבר שאומנה באמצעות תעבורת רשת זדונית כוללת יומני רישום מפוברקים שהוחדרו על ידי התוקף. יומני רישום המכילים דפוס־בתים ספציפי, מסווגים תמיד כ"שפירים", ובכך מתאפשר לחדירות עתידיות לחמוק פנימה.
הרעלה מבוססת־אופטימיזציה	התוקף משנה באופן שיטתי דוגמאות אימון עד למציאת השינוי הקטן ביותר שמטעה את המודל ועדיין נראה תקין.	מערכת איכון, שאומנה באמצעות נתוני חיישנים, מושפעת בעדינות כך שחתימות מכ"ם אויב מסוימות יסווגו כרעש של מזג אוויר אך רק בתנאים סביבתיים ספציפיים. המניפולציה היא כה קטנה עד שהיא מצליחה לעקוף את בדיקות האיכות.
הרעלה סינתטית או שמיוצרת באמצעות GAN	התוקף משתמש בבינה מלאכותית כדי ליצור דגימות אימון מורעלות אך מציאותיות ביותר, ובכך ההתקפה הופכת למשכנעת יותר וקשה יותר לזיהוי.	מודל לזיהוי דיסאינפורמציה עובר הרעלה באמצעות פוסטים סינתטיים ברשתות חברתיות שנוצרו על ידי GAN. פוסטים אלו מלמדים את המודל שסגנון הכתיבה של קמפיין השפעה זר הוא בעצם "רגיל", ובכך יכולתו של המודל לאתר מבצעי השפעה אמיתיים פוחתת.
הרעלה מבוססת־השפעה	התוקף מזהה את דוגמאות האימון החשובות ביותר ומרעיל רק אותן, ובכך מגדיל למקסימום את מידת ההשפעה במינימום שינויים.	גלאי חריגות ברשת החשמל מאומן באמצעות נתוני חיישנים היסטוריים. על ידי השחתת קומץ של דגימות נתונים בעלות השפעה מרובה התוקף מלמד את המערכת להתעלם מסימנים מובהקים של כשל השנאי, ובכך מתאפשר לו לבצע שיבוש מתואם ברשת החשמל.
בניית דלת אחורית מבוססת־טריגר	התוקף שותל דפוס נסתר (כגון סמל או צליל), שגורם למודל לספק את התשובה השגויה רק כאשר הטריגר נוכח.	כלי לזיהוי עצמים בשדה הקרב מסווג כל רחפן שנושא מדבקה ספציפית כ"ידידותי", ומאפשר לרחפנים עוינים המסומנים באותה מדבקה לעקוף הגנות אוטומטיות.
הרעלה של מקור נתונים/ במעלה הזרם (הרעלת שרשרת האספקה)	התוקף משחית נתונים עוד לפני שהם נכנסים למודל (למשל במהלך איסוף, שיקוף או סינון), כך שההרעלה נראית כמו נתונים "רגילים".	אנליסטים של חומרי מודיעין מסתמכים על מערכי נתונים שנאספו מהרשת לצורך מידול סנטימנט. שחקן זר מבצע מניפולציה בפוסטים בפורומים במשך חודשים, וזורע דפוסים מוטים כך שהמודל הסופי מספק הערכת חסר של רמות העוינות באזורים מסוימים, ובכך מעוות את ההערכות האסטרטגיות.

2.2.2 התקפות הרעלת מודל

אף שמחקר זה מתמקד בהרעלה בשכבת הנתונים, נדרשת סקירה תמציתית של התקפות הרעלת מודל (MPAs), שכן בפועל היריבים משלבים לעיתים מניפולציות ברמת הנתונים וברמת המודל, ושתי הקטגוריות עשויות להסוות זו את זו או לחזק זו את זו.

התקפות הרעלת מודל פועלות באמצעות שליטה ישירה במודל עצמו ושינוי של פרמטרים, משקלים או מנגנוני עדכון, לרוב בסביבות מבוזרות כגון למידה מבוזרת (FL) או בסביבות של למידת מכונה כשירות (MLaaS) במיקור חוץ. בעוד התקפות הרעלת נתונים (DPAs) משחיתות את הקלטים של האימון, התקפות הרעלת מודל (MPAs) מתערבות בתהליך האופטימיזציה או משנות ישירות את הפרמטרים הנלמדים. בהקשרים של מיקור חוץ או שרשרת אספקה, היריב עשוי לשבש אלגוריתמים של אימון או היפר־פרמטרים, כדי לייצר מודל פגום שממשיך להיראות פונקציונלי ומתנהג כראוי בהליך הערכה סטנדרטי.

אחת הצורות הנפוצות של הרעלת מודל היא התקפת דלת אחורית מבוססת־משקלים, ובה היריב מוותר לחלוטין על מניפולציה של הנתונים ומשנה ישירות את הפרמטרים הפנימיים של המודל כדי להטמיע טריגר חבוי. הדבר עשוי להתרחש באמצעות שינויים בתוכנה או במקרים קיצוניים אף באמצעות מניפולציה ברמת החומרה הבסיסית.[35] לדוגמה, קבלן שפרצו למערכות שלו עלול לשנות את המודל כך שכל תמונה שמכילה סמל קטן תסווג תמיד כ"ידידותית".

בסביבות של למידה מבוזרת מתגלים סוגים נוספים של התקפות. הרעלת זמינות (availability poisoning) עושה שימוש בלקוחות זדוניים ששולחים עדכוני מודל אקראיים או משובשים כדי לפגום בתהליך האימון ולמנוע מהמודל הגלובלי להתכנס — בדומה לשיבוש של ערוץ רדיו משותף כדי למנוע מכל המשתמשים לתאם פעולות.[36] לעומת זאת הרעלת מודל ממוקדת (targeted model poisoning) מחדירה התנהגות מזיקה ספציפית, למשל באמצעות הגשת עדכון מוגבר, המשכתב בחשאי את המודל הגלובלי במהלך שלב

35 Yiming Li et al., "Backdoor Learning: A Survey," preprint, arXiv, February 16, 2022, https://doi.org/10.48550/arXiv.2007.08745; Shuo Wang et al., "Backdoor Attacks Against Transfer Learning with Pre-Trained Deep Learning Models," *IEEE Transactions on Services Computing* 15, no. 3 (2022): 1526-39, https://tinyurl.com/tcabz259.

36 Yiyong Liu et al., "Transferable Availability Poisoning Attacks," version 2, preprint, arXiv, 2023, https://doi.org/10.48550/arXiv.2310.05141.

האגרגציה.[37] כמו כן הוכח שתוקפים מסוגלים לעקוף הגנות שמכונות Byzantine-robust, ונועדו לסנן עדכונים פגומים באמצעות עיצוב המניפולציות כך שייראו תקינות מבחינה סטטיסטית.[38]

בפועל, תוקפים עשויים לשלב פעולות ברמת הנתונים וברמת המודל בהתקפות הרעלה היברידיות, כדי להגדיל למקסימום את החשאיות ואת העמידות לאורך זמן. דוגמה מוכרת היא התקפת דלת אחורית מבוזרת (DBA), שבה כל לקוח שמשתתף מקבל רק מקטע קטן של טריגר. אף לקוח יחיד אינו נראה חשוד, אך כאשר מתבצע שילוב של העדכונים, המודל הגלובלי לומד את התבנית של דלת אחורית מלאה המופעלת רק בתנאים מסוימים.[39]

37 Eugene Bagdasaryan et al., "How To Backdoor Federated Learning," version 3, preprint, arXiv, 2018, https://doi.org/10.48550/ARXIV.1807.00459; Xinyun Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," version 1, preprint, arXiv, 2017, https://doi.org/10.48550/ARXIV.1712.05526; Tianyu Gu et al., "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," Version 2, Preprint, arXiv, 2017, https://doi.org/10.48550/ARXIV.1708.06733; Heyi Zhang et al., "SoK: Benchmarking Poisoning Attacks and Defenses in Federated Learning," preprint, arXiv, February 6, 2025, https://doi.org/10.48550/arXiv.2502.03801.

38 Peva Blanchard et al., "Byzantine-Tolerant Machine Learning," version 1, preprint, arXiv, 2017, https://tinyurl.com/5n76d785; Dong Yin et al., "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," version 2, preprint, arXiv, 2018, https://tinyurl.com/y3ku7w8j.

39 Zhang et al., "SoK."

2.3
משטחי תקיפה של הרעלת נתונים

וקטורי החדירה של התקפות הרעלה למערכות למידת מכונה, ובייחוד למערכות הנשענות על למידה עמוקה ועל מאגרי נתונים רחבי היקף, הם מגוונים ונשענים על חוליות חלשות לאורך שרשראות האספקה של הנתונים והמודלים. נקודות תורפה אלו מאפשרות לירידים להחדיר נתונים זדוניים או מודלים פגומים לתהליך האימון.

הבנת הווקטורים האלה חיונית, מכיוון שהם קובעים היכן וכיצד היריבים יכולים להשפיע על תהליך האימון ולנצל את התלות המערכתית בנתונים פתוחים, במשאבים משותפים ובמסגרות למידה מבוזרות. הדיון שלהלן מסכם את הקטגוריות המרכזיות של משטחי תקיפה המתועדות בספרות.

2.3.1 מערכי נתונים ציבוריים פגיעים ונתיבי תקיפה באמצעות איסוף מהרשת

מערכי נתונים ציבוריים בהיקף רחב המצויים ברשת מהווים הן משטח תקיפה בעל ערך גבוה והן נתיב נפוץ להחדרת הרעלה. מודלים המסתמכים על מערכי נתונים ענקיים המיועדים לשימוש כללי, ובייחוד כאלה שמקורם באינטרנט, פגיעים במיוחד להתקפות הרעלה. מערכות מסוג זה מאגדות לעיתים קרובות נתונים הטרוגניים שמקורם אינו ידוע או מאומת באופן חלקי בלבד, ובכך יוצרות קרקע פורייה למניפולציה של היריבים. מודלים רבים של למידה עמוקה נשענים על מערכי אימון גדולים שנאגרו מדגימות שמקורן אינו ידוע, לרבות דפי רשת שנאספו באופן אוטומטי, תוכן מרשתות חברתיות (כגון Reddit) ומאגרי גישה פתוחה (כמו ויקיפדיה). מאחר שהמקורות האלה עוברים על פי רוב עיבוד אוטומטי, אפשר להחדיר לתוכם דגימות מורעלות מבלי שיתגלו.[40]

גם מערכי נתונים אקדמיים רחבי היקף, כגון ImageNet (מאגר נתונים המשמש למחקר ופיתוח בתחום הראייה הממוחשבת), הוכחו כמועדים להרעלה — סיכון שמועצם בשל

40 Nicholas Carlini et al., "Poisoning Web-Scale Training Datasets Is Practical," version 2, preprint, arXiv, 2023, https://doi.org/10.48550/ARXIV.2302.10149; Qiu, "Survey on Poisoning Attacks."; Wang et al., "Threats to Training," 1-36; Zhao et al., "Data Poisoning in Deep Learning."

השימוש החוזר הנרחב שנעשה בהם בתהליכי הלמידה של המודלים.[41] מפתחים במורד הזרם עלולים לרשת ייצוגים מורעלים שהוטמעו במהלך האימון המוקדם. ארגונים הרוכשים או משיגים נתונים מספקי צד שלישי עומדים באופן דומה בפני סיכון של קבלת הזנות נתונים שנפרצו או שאינן מהימנות ללא ידיעתם.[42] סיכון זה חל גם על מתווכי נתונים שמערכותיהם נפרצו או על שירותי יצירת נתונים סינתטיים, המפיצים למספר רב של לקוחות מערכי נתונים שנראים לגיטימיים, אך בפועל הם מוטים באופן שיטתי, ובכך גורמים להתפשטות ההתקפה ברחבי האקוסיסטם הרחב יותר של הנתונים.[43]

גרסה מעודנת במיוחד עולה בהקשרים של איסוף מידע מהרשת — היריבים מנצלים את פער הזמנים שבין יצירת האינדקס של מערך הנתונים ובין שליפתו.[44] תוקפים יכולים להציב דגימות מורעלות ברשת ולהמתין לסורקי רשת אוטומטיים או לבוטים שיאספו אותן לצורך שימוש עתידי בהרצות אימון. במקרה של LLMs גורמים זדוניים יכולים לנצל קישורים שפג תוקפם במערכי נתונים בהיקף רחב, ולהחליף את התוכן המקורי בדגימות מורעלות.[45] טקטיקה זו התפתחה למה שחוקרים מכנים הרעלה מסוג split-view, שבה תוקפים רוכשים דומיינים שפג תוקפם התואמים לכתובות ה־URL של מערך הנתונים, ומחליפים את התוכן המקורי בנתונים זדוניים, בעוד האינדקס של מערך הנתונים עצמו נותר ללא שינוי.[46] באמצעות מניפולציה של נקודת הגישה לנתונים בלבד, היריבים מנצלים את האמון המובנה באינדקסים של מערכי נתונים שנבנו מראש, ומאפשרים להרעלה להתמיד באופן בלתי נראה לאורך מחזורי האימון.

41 Battista Biggio and Fabio Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications* Security (2018), 2154-2156, https://dl.acm.org/doi/10.1145/3243734.3264418; Liu et al., "Transferable Availability Poisoning Attacks."; Yaniv Taigman et al., "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *Proceeding of the 2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 1701–8, https://doi.org/10.1109/CVPR.2014.220.

42 Vassilev, *Adversarial Machine Learning*.

43 Justin Sherman, "Data Brokers and Data Breaches," *Duke – Tech Policy Program Blog*, September 27, 2022, https://tinyurl.com/rv9rb9tx.

44 Fendley et al., "Systematic Review."; Geiping et al., "Witches' Brew."; Qiu, "Survey on Poisoning Attacks."; Wang et al., "Threats to Training," 1-36.

45 Carlini et al., "Poisoning Web-Scale Training Datasets."

46 Carlini et al., "Poisoning Web-Scale Training Datasets."; Zhao et al., "Data Poisoning in Deep Learning."

2.3.2 פלטפורמות קוד פתוח ושרשרת האספקה של המודלים

הנוהג הנפוץ של שיתוף ושימוש חוזר ברכיבים של למידת מכונה יוצר נקודות חדירה רבות לאורך שרשרת האספקה. האקוסיסטם של הקוד הפתוח, אף שהוא תומך בחדשנות, גם מרחיב את משטח התקיפה העומד לרשות היריבים לשתילת מודלי דלת אחורית או יחסי תלות זדוניים.[47]

כאשר אימון מודלים מועבר למיקור חוץ אצל ספק צד שלישי שאינו מהימן (למשל MLaaS או פלטפורמות ענן), התוקף — הפלטפורמה הזדונית — שולט בתהליך האימון ויכול לשנות בחופשיות את נתוני האימון או את ההליך עצמו ולהחזיר למשתמש מודל דלת אחורית. גם כשהאימון מתבצע בתוך הארגון, ההסתמכות על מאגרים חיצוניים עלולה להיות מסוכנת באותה מידה. תוקפים יכולים להציע מודלים מאומנים מראש, בסיסי קוד או משקלי מודל זדוניים למשתמשים במורד הזרם. המשתמשים שמורידים ומיישמים את המודלים האלה לצורך כוונון עדין או שימוש חוזר במודלים מאומנים מראש, יורשים מבלי דעת דלתות אחוריות או סוסים טרויאניים שהוטמעו בהם.[48]

מעבר לשיבוש ישיר, היריבים עשויים להשתלט על תשתית אירוח או על מערכות לבקרת גרסאות. התוקפים יכולים להחדיר מודל דלת אחורית על ידי פריצה לשרת החיצוני המארח את נתוני המודל או על ידי שינוי פלטפורמות כמו אתרי ויקי של Model Zoo כך שיפנו לכתובת URL זדונית. ביצוע מניפולציה זו בשרשרת האספקה התרחבה לכדי התחזות במאגרי מודלים, שבהם תוקפים מעלים מודלים תחת שמות כמעט זהים לשמות לגיטימיים — כגון "GPT-J-unofficial" המחקה את "GPT-J" — כדי לנצל את אמון המפתחים ואת מוסכמות השיום. טכניקת התחזות זו, שהודגמה בתקרית "PoisonGPT" (שהראתה כיצד נתוני אימון בקוד פתוח שעברו הרעלה, יכולים לגרום ל־LLM להתנהג באופן מטעה ולשמור

47 Xinyi Zheng et al., "Towards Robust Detection of Open Source Software Supply Chain Poisoning Attacks in Industry Environments," *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (2024), 1990–2001, https://doi.org/10.1145/3691620.3695262.

48 Fendley et al., "Systematic Review."; Gu et al., "BadNets."; Linyang Li et al., "Backdoor Attacks on Pre-Trained Models by Layerwise Weight Poisoning," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), 3023–32, https://doi.org/10.18653/v1/2021.emnlp-main.241; Li et al., "Backdoor Learning."; Ramirez et al., "Poisoning Attacks and Defenses."; Zhang et al., "SoK."

על מראית עין תקינה), ממחישה כיצד הנחות הנוגעות למוניטין ולשיום במאגרים פתוחים יכולות להפוך לווקטורים של חדירה.[49]

2.3.3 ויקיפדיה כווקטור פגיעות מרכזי[50]

ויקיפדיה הפכה לאחד היעדים האטרקטיביים והמשמעותיים ביותר להרעלת נתונים בהיקף רחב. מבנה העריכה הפתוח שלה, תפוצתה הגלובלית ותפקידה המרכזי באימון של מערכות בינה מלאכותית מודרניות יוצרים שילוב נדיר של היקף, נגישות ומינוף אסטרטגי. מעטים מקורות הנתונים שמשפיעים על מודלי יסוד באופן ישיר כמו ויקיפדיה, עובדה ההופכת אותה לנקודת חדירה אידיאלית עבור היריבים שמבקשים לעצב את מה שמערכות בינה מלאכותית "לומדות" על העולם.

תלות במודלי יסוד: מודלים עכשוויים של בינה מלאכותית מסתמכים במידה רבה על ויקיפדיה. מאחר שכל אדם יכול לערוך את הפלטפורמה, ומנגנוני האימות מוגבלים, גורמים זדוניים יכולים להחדיר בחשאי טקסט מורעל לערכים מתוך ידיעה שהשינויים האלה ייאספו בסבב יצוא הנתונים הבא. הרעלת ויקיפדיה מופיעה בשתי צורות: (1) הרעלה באמצעות הטמעה טבעית, שבה LLMs קולטים מבלי דעת תוכן שעבר מניפולציה כחלק מתהליכי הסריקה השגרתיים; (2) הרעלה מכוונת, שבה היריבים עורכים ערכים בצורה מפורשת כדי להשפיע על אימון מודלים עתידי. מרגע שהוא נקלט, התוכן הפגום מועבר במורד הזרם לאין־ספור מודלים שעברו כוונון עדין, ומטמיע הטיות עדינות, הטעיות או טריגרים של דלת אחורית במגוון יישומים.

היקף, אימות והתמדה: גודלה של ויקיפדיה הופך אימות קפדני לבלתי ישים. מחקרים אמפיריים מראים כי תוקפים יכולים להרעיל עד כ־6–7 אחוזים מהטוקנים של ויקיפדיה בצילום מצב (snapshot) נתון על ידי תזמון של עריכות סביב חלונות יצוא (dump), ובכך להוביל לזיהום של מאגרי אימון מרכזיים הניתן למדידה. מאחר שמנטרים אנושיים אינם יכולים בפועל לסקור מיליוני עריכות, מניפולציות הנשמעות סבירות עשויות להיוותר על

49 Carlini et al., "Poisoning Web-Scale Training Datasets."; Fendley et al., "Systematic Review."; Gu et al., "BadNets."; Wang et al., "Threats to Training," 1-36; Zheng et al., "Towards Robust Detection," 1990–2001.

50 Carlini et al., "Poisoning Web-Scale Training Datasets."; Qiu, "Survey on Poisoning Attacks."; Valentin Châtelet, "Exposing Pravda: How Pro-Kremlin Forces Are Poisoning AI Models and Rewriting Wikipedia," *New Atlanticist*, 2025, https://tinyurl.com/37mumwjn.

כנן לאורך פרקי זמן ממושכים. מכיוון שוויקיפדיה מפיצה נתונים באמצעות קובצי צילום מצב תקופתיים, ברגע שתוכן מורעל נלכד, הוא נשמר למשך חודשים או שנים, גם אם הוא תוקן באתר החי. מערכי אימון שמבוססים על צילומי המצב האלה מקבעים את ההתקפה זמן רב לאחר שהעריכה המקורית נעלמה.

טכניקות תקיפה מתקדמות: היריבים פיתחו שיטות מתוחכמות יותר ויותר לניצול הפתיחות של ויקיפדיה ודפוסי איסוף הנתונים הצפויים שלה:

- **frontrunning:** תזמון של עריכות זדוניות ממש לפני חלונות יצוא הנתונים כדי להבטיח שאלה ייכללו במערכי האימון.
- **הרעלת split-view:** רכישת דומיינים שפג תוקפם ומצוטטים בהפניות של ערכים, והחלפת התוכן שלהם בחומר שעבר מניפולציה, בעוד הציטוט עצמו נותר ללא שינוי.
- **התקפות clean-label:** החדרת טקסט הנראה תקין ומנוסח היטב, אך נושא הטיות חבויות או טריגרים המופעלים רק במהלך אימון המודל.

מבצעי מידע אסטרטגיים: ויקיפדיה היא גם יעד לקמפיינים מתואמים המזוהים עם מדינות. באמצעות עריכה של דפים בעלי רגישות פוליטית או מסגור מחדש של אירועים באופן עדין, היריבים יכולים להשפיע על הנרטיבים הנקלטים בשלב האימון המקדים של LLMs. חלק מהמבצעים משלבים עריכות בוויקיפדיה עם אופטימיזציה למנועי חיפוש (SEO), וכך דפים מורעלים מטפסים בדירוג ונאספים בסבירות גבוהה יותר על ידי סורקי רשת. מקרים מתועדים, כגון עריכות שיטתיות התומכות בעמדות הקרמלין, ממחישים שהטקטיקות האלה אינן עוד תרחיש תאורטי, אלא רכיבים פעילים בעידן המודרני של לוחמת המידע והלוחמה הקוגניטיבית.

זיהוי והגברה חוצת פלטפורמות: כלי הניטור של ויקיפדיה יעילים בהתמודדות עם השחתות גלויות, אך מתקשים לזהות מניפולציות מתואמות וממושכות, המתבצעות בקצב איטי ושומרות על סבירות דקדוקית ועובדתית. מסננים אוטומטיים אינם מצליחים להבחין בצורה מהימנה בין עריכות לגיטימיות ובין קמפיינים עדינים של הרעלה, במיוחד כאשר שינויים קטנים רבים לאורך דפים שקשורים זה לזה מצטברים יחדיו ומסיטים את מסגור הנושא. מאחר שאימון בהיקף רחב נשען על קורפוסים הכוללים הפניות הדדיות, טקסט מורעל מוויקיפדיה משוכפל לעיתים קרובות בבלוגים, במאגרים פתוחים או באתרי ויקי אקדמיים, ובכך מגביר את הנראות שלו וגורם לסורקים לפרש את החזרתיות כאישוש לתקפותו.

השלכות על אבטחת מודלי יסוד: ברגע שנתוני ויקיפדיה מזוהמים נכנסים לתהליך האימון, העיוותים הנוצרים מקודדים במשקלי המודל ונשמרים לאורך שלבי הכוונון העדין. רכיבים מוטמעים אלו יכולים לשמש כדלתות אחוריות עמידות או כהטיות אידאולוגיות המשפיעות על מערכות במורד הזרם — מכלי חיפוש וסיכום ועד למנועי תרגום. ככל שנעשה שימוש חוזר במודלי יסוד באמצעות הלמידה של המודלים, כל יישום חדש יורש ומפיץ מחדש את הייצוגים המורעלים המקוריים.

תובנה מסכמת: ויקיפדיה מדגימה את השבריריות של אקוסיסטם פתוח ורחב היקף של נתונים העומד בבסיס הפיתוח של הבינה המלאכותית. הפתיחות, מעמדה המרכזי, מחזורי העדכון הצפויים והפערים בהיקף הניטור האנושי הופכים אותה ליעד בעל ערך גבוה ובו בזמן בעל חשיפה גבוהה. עדויות גוברות למניפולציות מאורגנות ברמת המדינה מראות כי הרעלת ויקיפדיה כבר עברה מסיכון תאורטי למציאות מבצעית, והפכה פלטפורמת ידע פתוחה לזירה של לוחמה קוגניטיבית ולוחמת מידע. הדוגמה של ויקיפדיה מיישמת בפועל את המנגנונים התאורטיים הרחבים יותר שנדונו קודם לכן: עריכות בהיקפים נמוכים במעלה הזרם יכולות להתפשט ולהפוך לעיוותים אפיסטמיים בהיקף גדול ברגע שהן מוטמעות בקורפוסים של אימון מקדים.

איור 4: שרשרת האספקה של הרעלת ויקיפדיה

2.3.4 מניפולציה של גורמי פנים ותהליכי תיוג מבוססי־מיקור המונים

גורמים פנימיים או בעלי ידע ייעודי מהווים איום חמור, המאפשר הזרקה כירורגית וחשאית של נתונים מורעלים או מודלים פגומים. איום הגורם הפנימי חורג מעבר לריגול המסורתי, וכולל גם תיוג של נתונים מבוססי־מיקור המונים ותהליכים אחרים הנשענים על מעורבות אנושית פעילה לאורך מחזור החיים של למידת המכונה.

גורמים פנימיים (עובדים, קבלנים או מסתננים) בתוך הארגון יכולים להחדיר בחשאי מספר קטן של דגימות מורעלות ישירות לתוך מערך האימון או לנצל גישה מועדפת למאגרי נתונים כדי להוסיף או לשנות דוגמאות אימון.[51] היריבים מנצלים יותר ויותר פלטפורמות תיוג מבוססות־מיקור המונים — מערכות מבוזרות המאגדות תיוגים ממאגרים גדולים של משתמשים אנושיים. בהקשר זה, מתייגים זדוניים יכולים לשבש במכוון תוויות של דגימות מסוימות או לפעול בתיאום כדי להטות תת־קבוצה של הנתונים. פלטפורמות תיוג מבוססות־מיקור המונים — שבהן עובדים זדוניים בוחרים באופן אסטרטגי אילו מופעים לתייג ומספקים תיוגים מורעלים — מנצלות מנגנונים של צבירת תוויות הנשענים על מהימנות המתייגים. המניפולציות האלה זולות לביצוע, ניתנות להרחבה על פני משימות שונות, וקשות במיוחד לגילוי כאשר תוקפים משבצים תוויות נכונות לצד תוויות מורעלות.[52]

מחקר על Amazon Mechanical Turk (MTurk) הראה כי מיעוט לא מבוטל של מתייגים הפיק באופן עקבי תגובות שאינן כנות, תגובות באיכות נמוכה או תגובות מטעות. Ahler ואחרים (2021) דיווחו כי 25–35 אחוזים מהתגובות הציגו התנהגות חשודה או לא אמינה, והיקף התגובות שלא היו כנות עלה ביותר מ־200 אחוזים בין 2018 ל־2020.[53] אף שהמקרים

51 Chen et al., “Targeted Backdoor Attacks.”; Lingxin Jin et al., “A Survey of Trojan Attacks and Defenses to Deep Neural Networks,” preprint, arXiv, August 15, 2024, https://doi.org/10.48550/arXiv.2408.08920; Zhiyi Tian et al., “A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning,” *ACM Computing Surveys* 55, no. 8 (2022): 1–35, https://doi.org/10.1145/3551636; Vassilev, *Adversarial Machine Learning*.

52 Gang Wang et al., “Man vs. Machine: Practical Adversarial Detection of Malicious Crowdsourcing Workers,” *Proceedings of the 23rd USENIX Security Symposium* (2014), https://tinyurl.com/48y33ckr; Muñoz-González et al., “Towards Poisoning.”; Wang et al., “Threats to Training,” 1-36.

53 Douglas J. Ahler et al., “The Micro-Task Market for Lemons: Data Quality on Amazon’s Mechanical Turk,” *Political Science Research and Methods* 13, no. 1 (2021): 1-20, https://doi.org/10.1017/psrm.2021.57.

האלה אינם בהכרח זדוניים, הם ממחישים כיצד אפילו אחוז קטן של תורמים שאינם אמינים יכול לעוות באופן משמעותי תוצאות במורד הזרם. משמעות הדבר בהקשרים של למידת מכונה היא שאפילו 3–5 אחוזים של מתייגים עוינים מתואמים — אחוז קטן בהרבה ממה שנצפה במחקרים כלליים על איכות נתונים ב־MTurk — עלולים לשנות באופן משמעותי ושיטתי את התנהגות המודל.[54]

ככל שארגונים מתחילים לפרוס סוכני בינה מלאכותית אוטונומיים, שלהם גישה רציפה ומגובה בהרשאות לזרמי נתונים פנימיים, מערכות אלו הופכות בעצמן לקטגוריה חדשה של גורם פנימי: במקרה של פריצה, התוקף עלול בהיחבא לנתב מחדש, לסנן או לבצע מניפולציה בנתונים שהסוכנים האלה אוספים ומסננים — ובכך להפוך אותם בפועל לווקטורים אוטומטיים המרעילים נתונים בתוך הארגון.[55]

בקצה המתוחכם ביותר של הספקטרום, הרעלת נתונים אינה חייבת להתבצע בתחום הדיגיטלי בלבד. החדרה חשאית של נתונים מורעלים באמצעות נכסים אנושיים — כגון מקורות צבאיים או סוכנים מבצעיים המושתלים במעבדות מחקר זרות או אצל קבלני רכש — יכולה לגרום להרעלה מתמשכת וספציפית, שקשה לזהות או לייחס אותה למקור מסוים. שילוב זה של גישה מבפנים ומבצעי מודיעין אנושי (HUMINT) מדגים כיצד הרעלת נתונים יכולה להתמזג עם מלאכת הריגול הקלאסית ולטשטש את הגבולות בין הונאה טכנית להונאה מבצעית.[56]

2.3.5 סביבות למידה מבוזרות ושיתופיות

פרדיגמות אימון מבוזרות, כגון למידה מבוזרת (FL), מציגות נקודות כניסה ספציפיות המוגדרות מראש כבר בשלב התכנון, ולעיתים קרובות מנצלות את היעדרו של מנגנון אימות מרכזי. המסגרות האלה מסתמכות על מכשירי קצה או מוסדות, שמעבירים עדכונים מקומיים למודל גלובלי משותף ויוצרים הזדמנויות רבות עבור היריבים לבצע מניפולציות בפרמטרים או להחדיר גרדיאנטים מורעלים. בלמידה מבוזרת, משתתפים יחידים (משתמשי

54 Rishi D. Jha et al., “Label Poisoning Is All You Need,” preprint, arXiv, October 29, 2023, https://doi.org/10.48550/arXiv.2310.18933.

55 Wendi Whitmore, “6 Predictions for the AI Economy: 2026’s New Rules of Cybersecurity,” *Paloalto Networks Blog*, November 18, 2025, https://tinyurl.com/5393c6xa.

56 Aaron Conti, “Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare,” *Articles of War*, June 30, 2025, https://tinyurl.com/54c8z7sw.

קצה) מאמנים מודלים מקומיים ושולחים עדכונים מצטברים בחזרה לשרת המרכזי. מכשירי קצה זדוניים יכולים להשתמש בשיטה של הרעלת המודל כדי להעלות עדכונים מורעלים ישירות לשרת המרכזי, ובכך להשפיע על המודל הגלובלי מבלי שלשרת תהיה שקיפות לגבי נתוני האימון הפרטיים שלהם.[57]

הארכיטקטורות המבוזרות האלה מדגימות כיצד גדל מספרם של שטחי התקיפה האפשריים: כל משתתף אוטונומי הופך לנקודת החדרה פוטנציאלית של מידע מורעל, ושלב הצבירה מספק נקודת כשל מערכתית יחידה במקרה של פריצה.

2.3.6 מערכות עיבוד ואחזור נתונים בשלבי הביניים

הרעלה אינה מוגבלת לאיסוף הראשוני או לצבירה הסופית של הנתונים; שלבי ביניים בתהליך עיבוד הנתונים הופכים בהדרגה גם הם ליעדי פריצה וניצול לרעה. תהליכי עיבוד מקדים של נתונים, שבהם מתבצעים ניקוי, המרה או העשרה, עלולים להפוך לנקודות למניפולציה. תוקפים הפורצים לסקריפטים של עיבוד מקדים או למערכות אחסון, יכולים להחדיר דגימות מורעלות במהלך נרמול מאפיינים או העשרת נתונים, ובכך להבטיח שהמופעים הזדוניים ישרדו בבדיקות אימות מאוחרות. עיבוד מקדים מבצע לעיתים קרובות תקנון או דחיסה של הנתונים, ולכן ההתקפות האלה נוטות להסתיר את תוצרי הלוואי שלהן ביעילות ויוצרות תמהיל של המרות זדוניות ולגיטימיות.[58]

57 Ashwinee Panda et al., "SparseFed: Mitigating Model Poisoning Attacks in Federated Learning with Sparsification," preprint, arXiv, December 12, 2021, https://doi.org/10.48550/arXiv.2112.06274; Wenqi Wei et al., "Demystifying Data Poisoning Attacks in Distributed Learning as a Service," *IEEE Transactions on Services Computing* 17, no. 1 (2024): 237-50, https://doi.org/10.1109/TSC.2023.3341951; Jianping Wu et al., "Challenges and Countermeasures of Federated Learning Data Poisoning Attack Situation Prediction," *Mathematics* 12, no. 6 (2024): 901, https://doi.org/10.3390/math12060901; Xueqing Zhang et al., "Visualizing the Shadows: Unveiling Data Poisoning Behaviors in Federated Learning," version 1, preprint, arXiv, 2024, https://doi.org/10.48550/arXiv.2405.16707.

58 Geiping et al., "Witches' Brew."; Lumenova, "Data Poisoning Attacks: How AI Models Can Be Corrupted," *Lumenova Blog*, July 17, 2025, https://tinyurl.com/39upvba7; Muñoz-González et al., "Towards Poisoning."; Wang et al., "Threats to Training," 1-36; Zhao et al., "Data Poisoning in Deep Learning."

במערכות שבנויות סביב יצירה משולבת אחזור (RAG), המאחזרת מקורות חיצוניים בזמן השאילתה, נוצרת שכבת פגיעות נוספת. זיהום בסיס הידע ב־RAG כרוך בהרעלה של מאגרי נתונים חיצוניים, מסמכים או משאבי רשת שאליהם מערכת הבינה המלאכותית שולחת שאילתות במהלך שלב ההסקה. גם אם מודל הליבה נותר נקי, אחזור תוכן מורעל בזמן ריצה עלול להחדיר מידע שגוי לפלטי המודל. כאשר משאבים מזוהמים מעין אלו משתלבים מאוחר יותר במחזורים של אימון מתמשך או כוונון עדין, הגבול בין זיהום בזמן ההסקה לבין הרעלה אמיתית בזמן האימון נעלם, והמידע שעבר מניפולציה מוטמע באופן קבוע במשקלי המודל.[59]

הרעלת נתונים מנצלת את הממשקים החדירים של האקוסיסטם של למידת המכונה המודרנית: מאגרי נתונים פתוחים ומרכזי מודלים, תהליכי תיוג וסביבות של למידה מבוזרת. היריבים המודרנים משלבים יותר ויותר מספר וקטורים של חדירה, כגון גישה מבפנים לצד איסוף נתונים מהאינטרנט או חדירה לשרשרת האספקה, כדי להגדיל למקסימום את השרידות ולצמצם את יכולת הייחוס. כל וקטור מייצג פגיעה פוטנציאלית בשלמות האפיסטמית של מערכת הבינה המלאכותית, וממחיש שהחוליה החלשה ביותר בשרשרת — אם היא אנושית, ואם היא פרוצדורלית או טכנית — עלולה להפוך לנקודת הכניסה לביצוע מניפולציה אסטרטגית.

59 Zhaorun Chen et al., “AgentPoison: Red-Teaming LLM Agents via Poisoning Memory or Knowledge Bases,” version 1, preprint, arXiv, 2024, https://doi.org/10.48550/ARXIV.2407.12784; Ruochen Jiao et al., “Can We Trust Embodied Agents? Exploring Backdoor Attacks against Embodied LLM-Based Decision-Making Systems,” preprint, arXiv, April 30, 2025, https://doi.org/10.48550/arXiv.2405.20774; Jiaqi Xue et al., “BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models,” preprint, arXiv, June 6, 2024, https://doi.org/10.48550/arXiv.2406.00083; Quan Zhang et al., “Human-Imperceptible Retrieval Poisoning Attacks in LLM-Powered Applications,” preprint, arXiv, April 26, 2024, https://doi.org/10.48550/arXiv.2404.17196; Zhao et al., “Data Poisoning in Deep Learning.”; Zexuan Zhong et al., “Poisoning Retrieval Corpora by Injecting Adversarial Passages,” preprint, arXiv, October 29, 2023, https://doi.org/10.48550/arXiv.2310.19156; Wei Zou et al., “PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models,” version 3, preprint, arXiv, 2024, https://doi.org/10.48550/arXiv.2402.07867.

איור 5: משטחי תקיפה עיקריים: כיצד הם נכנסים פנימה

2.4
כיצד הצלחת ההתקפה נמדדת

הצלחה או כישלון של התקפות הרעלה נמדדים באמצעות מערך מקיף של מדדים כמותיים המשקפים שני ממדים עיקריים, ולעיתים קרובות מתחרים — יעילות ההתקפה וחשאיותה.[60] מתודולוגיות הערכה מתחילות בהגדרת יעד התוקף, כגון סיווג שגוי של מטרה מסוימת או החדרה של דלת אחורית נסתרת, ולאחר מכן הן קובעות באיזו מידה של יעילות ההתקפה משיגה מטרה זו, וממזערת את הנזק האגבי לביצועים הלגיטימיים של המודל.

2.4.1 מדדים ליעילות ההתקפה

יעילותה של התקפת הרעלה נמדדת על פי מידת הצלחתה בהשגת התוצאה שאליה כיוון התוקף. המדד הסטנדרטי הוא שיעור הצלחת ההתקפה (ASR) — שיעור המקרים שבהם המודל המורעל מתנהג כפי שהתוקף התכוון. בהתקפות ממוקדות או בהתקפות דלת אחורית, ה־ASR מודד באיזו תדירות קלטים מורעלים או קלטים הנושאים טריגר מסווגים לתווית שבחר התוקף, ובמבצעים מוצלחים שיעור זה לעיתים גבוה מ־90 אחוזים. בהתקפות שאינן ממוקדות, ה־ASR עוקב במקום זאת אחר ההידרדרות הכוללת של כל מערך הנתונים, ומצביע על היקף השיבוש של ביצועי המודל.[61]

במערכות גנרטיביות כגון LLMs או ארכיטקטורות RAG, המונח ASR מתייחס לחלק של ההנחיות שמפיקות את הפלט הרצוי לתוקף. לדוגמה, הוספת ביטוי, שינוי סנטימנט או יצירת תוכן מזיק. היות שפלטים גנרטיביים אינם פשוט נכונים או לא נכונים, ההצלחה מוערכת לעיתים קרובות על סולם רציף המודד עד כמה התגובה תואמת את מטרות

60 Fendley et al., “Systematic Review.”; Tian et al., “Comprehensive Survey,” 1-35; Geming Xia et al., “Poisoning Attacks in Federated Learning: A Survey,” *IEEE Access* 11 (2023): 10708-22, https://doi.org/10.1109/ACCESS.2023.3238823.

61 Chen et al., “Targeted Backdoor Attacks.”; Jiao et al., “Can We Trust Embodied Agents?.”; Liu et al., “Transferable Availability Poisoning Attacks.”; Wang et al., “Threats to Training,” 1-36; Zhou et al., “Survey on Backdoor Threats,” 28-58.

התוקף. במערכות RAG, יש מדד המכונה לעיתים retrieval ASR, ומשקף באיזו תדירות המערכת מאחזרת מסמכים מורעלים בלבד בעת מענה לשאילתה.[62]

מעבר ל־ASR, חוקרים עוקבים גם אחר מדדים מבוססי־שגיאות, שמתארים את מידת הסטייה של המודל המורעל מהתנהגות נורמלית.[63] התקפות שאינן ממוקדות שואפות להגדיל את שיעורי הסיווג השגוי באופן כללי, בעוד התקפות ממוקדות דוחפות שגיאות לעבר קטגוריה ספציפית מבלי לפגוע בדיוק של הקלטים הנקיים.[64] שיעור היפוך התוויות (LFR) — באיזו תדירות דוגמאות שאינן מטרה מסווגות בטעות כמטרה — הוא מדד נפוץ נוסף בתרחישים של מודלים המאומנים מראש.[65]

בלמידה מבוזרת (FL), שבה התוקף עשוי לשלוט רק בחלק קטן ממשתמשי הקצה המשתתפים, גורם ההשפעה החריגה (OIF) מודד כמה השפעה היריב משיג ביחס לתרומתו הנומינלית. ערכי OIF גבוהים מלמדים שאפילו גישה מוגבלת יכולה להסיט את המודל הגלובלי באופן לא פרופורציונלי.[66]

לבסוף, מחקר מהזמן האחרון מלמד שמספרים מוחלטים חשובים יותר מאחוזים. במודלים גדולים, כמה מאות של דוגמאות מורעלות — שהן לעיתים פחות מ־0.001 אחוז מהסך הכולל — יכולות להטמיע דלתות אחוריות קבועות שנשמרות גם לאחר כוונון עדין ושימוש במודל במורד הזרם. המשמעות אינה מסתיימת בכך ש״מידע מורעל קיים במערך הנתונים״. פירוש הדבר שגבולות ההחלטה הפנימיים של המודל והייצוגים שלו השתנו. בפועל, הדבר עלול לגרום למודל לייצר באופן אמין פלטים שנבחרו על ידי התוקף בתגובה לטריגרים ספציפיים, לפרש באופן שגוי דפוסים מסוימים או להעדיף באופן שיטתי נרטיבים מסוימים. ברגע ששינויים התנהגותיים אלו מקודדים, הם יכולים להתפשט ואף להתגבר עם

62 Fendley et al., “Systematic Review.”; Xue et al., “BadRAG”; Zhou et al., “Survey on Backdoor Threats,” 28-58; Zou et al., “PoisonedRAG.”

63 Cinà et al., “Wild Patterns Reloaded,” 1-39; Muñoz-González et al., “Towards Poisoning.”; Jacob Steinhardt et al., “Certified Defenses for Data Poisoning Attacks,” preprint, arXiv, November 24, 2017, https://doi.org/10.48550/arXiv.1706.03691.

64 Jagielski et al., “Manipulating Machine Learning.”; Nicolas Michael Müller et al., “Data Poisoning Attacks on Regression Learning and Corresponding Defenses,” version 1, preprint, arXiv, 2020, https://doi.org/10.48550/ARXIV.2009.07008; Qiu, “Survey on Poisoning Attacks.”

65 Qiu, “Survey on Poisoning Attacks.”; Zhao et al., “Data Poisoning in Deep Learning.”; Zhou et al., “Survey on Backdoor Threats,” 28-58.

66 Panda et al., “SparseFed.”; Zhang et al., “SoK.”

הזמן ככל שארגונים עושים שימוש חוזר, מבצעים כוונון עדין או מזקקים את המודל שנפרץ, ובכך מאפשרים לכמויות רעל קטנות מאוד לגרום להשפעות בהיקף של כלל האקוסיסטם.[67]

2.4.2 מדדים לחשאיות ולמעשיות של ההתקפה

הצלחה טכנית היא רק חצי מהפעולה המרעילה; כדי להשיג הצלחה מבצעית אמיתית, ההתקפה צריכה להישאר סמויה. לפיכך החשאיות מוערכת על פי שלושה ממדים: עד כמה הנתונים המורעלים נראים טבעיים, עד כמה המודל מתנהג כרגיל עם קלטים נקיים, ובאיזו יעילות התוקף יכול להשיג את ההשפעות האלה.[68]

שמירה על ביצועי מודל תקינים — התקפה חשאית שומרת על מראה תקין של המודל. מדד הביצועים הנקיים (CPM) מודד את הדיוק על נתוני בדיקה שאינם מורעלים; התקפות יעילות נשארות בטווח של כאחוז אחד מביצועי המודל המקורי.[69] שיעור ירידת הביצועים (PDR) משקף את מידת הירידה בדיוק בין דגימות נקיות למורעלות.[70] נזק אגבי נמוך, כלומר שגיאות מינימליות בקלטים שפירים, הוא חיוני, מכיוון שהוא מצביע על כך שרק ההתנהגות הממוקדת של התוקף מושפעת בעוד יתר המודל מתפקד כמצופה.[71]

הסתרת הקלטים המורעלים והדלת האחורית — חמקנות תלויה גם בשאלה האם הנתונים המורעלים והמודל שעבר שינוי מסוגלים לחמוק מאמצעי גילוי.[72] חמקנות הקלט משקפת עד כמה טבעיות או סבירות נראות הדוגמאות המורעלות.[73] במערכות מבוססות־תמונה, השינויים חייבים להיות זעירים דיים כדי שלא יהיו ניתנים לזיהוי באופן חזותי; במערכות

67 Alexandra Souly et al., “Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples,” preprint, arXiv, October 8, 2025, https://doi.org/10.48550/arXiv.2510.07192.

68 Chen et al., “Targeted Backdoor Attacks.”; Fendley et al., “Systematic Review.”; Jiao et al., “Can We Trust Embodied Agents?.”; Li et al., “Backdoor Learning.”; Tian et al., “Comprehensive Survey,” 1-35.

69 Fendley et al., “Systematic Review.”

70 Zhou et al., “Survey on Backdoor Threats,” 28-58.

71 Jagielski et al., “Subpopulation Data Poisoning Attacks.”

72 Fendley et al., “Systematic Review.”

73 Fendley et al., “Systematic Review.”; Li et al., “Backdoor Learning.”; Shah et al., “Guarding the Gates.”

טקסטואליות, התוכן המורעל חייב להיות קריא ושוטף ולהיתפס כתקין מבחינה סמנטית.[74] התקפות clean-label מצטיינות בתחום זה מכיוון שהדגימות שלהן נראות לגיטימיות ונושאות תוויות נכונות. חמקנות המודל מתייחסת לשאלה האם אותות פנימיים חושפים את נוכחותה של דלת אחורית. אנליסטים בודקים אם דגימות מורעלות חומקות מכלים לגילוי חריגות, ואם קלטים דומים אך שגויים אינם מצליחים להפעיל את הדלת האחורית.[75] על אף שכלים כגון ניתוח חתימות ספקטרליות וקיבוץ הפעלות מצליחים לעיתים לאתר הרעלות, מהימנותם נותרת מוגבלת — במיוחד ב־LLMs או במודלים ליצירת קוד, שבהם מניפולציות עדינות נבלעות בקלות במורכבות המודל.[76]

יעילות ותקציב ההרעלה — ממד נוסף הוא שיעור ההרעלה (PR) — כמה מנתוני האימון יש לשנות כדי להשיג ASR גבוה. התקפות יעילות פועלות עם תקציבי הרעלה קטנים מאוד, לרוב מתחת ל־5–10 אחוזים ממערך הנתונים, ובמקרים מסוימים מדובר בשיעור הנמוך בהרבה מזה. חוקרים מציגים לעיתים קרובות את ה־ASR כנגד ה־PR כדי להראות כמה מעט נתונים מורעלים נדרשים כדי לשנות באופן משמעותי את התנהגות המודל.[77] גם היעילות החישובית היא מרכיב חשוב: שיטות מבוססות־אופטימיזציה נוטות להיות עוצמתיות יותר, אך מצריכות משאבי חישוב רבים יותר, בעוד היוריסטיקות פשוטות יותר עשויות להיות זולות יותר אך יעילות פחות.[78]

74 Biggio and Roli, “Wild Patterns.”; Steinhardt et al., “Certified Defenses.”

75 Cinà et al., “Wild Patterns Reloaded,” 1-39; Jiao et al., “Can We Trust Embodied Agents?.”; Souly et al., “Poisoning Attacks on LLMs.”

76 Fendley et al., “Systematic Review.”; Li et al., “Backdoor Learning.”; Shah et al., “Guarding the Gates.”; Souly et al., “Poisoning Attacks on LLMs.”; Vassilev, *Adversarial Machine Learning*; Xiaojun Xu et al., “Detecting AI Trojans Using Meta Neural Analysis,” preprint, arXiv, October 1, 2020, https://doi.org/10.48550/arXiv.1910.03137; Zhou et al., “Survey on Backdoor Threats.”

77 Alexander Turner et al., *Clean-Label Backdoor Attacks*, n.d., https://tinyurl.com/y6vea5f8; Bagdasaryan et al., “How To Backdoor Federated Learning.”; Chen et al., “Targeted Backdoor Attacks.”; Fendley et al., “Systematic Review.”; Geiping et al., “Witches’ Brew.”; Jagielski et al., “Manipulating Machine Learning.”; Wang et al., “Threats to Training,” 1-36.

78 Cinà et al., “Wild Patterns Reloaded,” 1-39; Jagielski et al., “Manipulating Machine Learning.”; Muñoz-González et al., “Towards Poisoning.”; Qiu, “Survey on Poisoning Attacks.”; Tian et al., “Comprehensive Survey,” 1-35; Wang et al., “Threats to Training,” 1-36; Zou et al., “PoisonedRAG.”

שרידות לאורך שלבי האימון והפריסה — מדד אחרון הוא שרידות או קביעות — באיזו מידה של הצלחה הדלת האחורית שורדת לאחר מחזורי אימון נוספים או התערבויות לצורך הגנה. התקפות ששומרות על ASR מתון (לרוב מעל 35–47 אחוזים) גם לאחר כוונון עדין, נחשבות לעמידות.[79] שרידות במשימות שבמורד הזרם היא אינדיקטור נוסף: אם דלת אחורית שנלמדה במהלך אימון מקדים ממשיכה לפעול לאחר שהמודל עבר כוונון עדין לצורך תרגום, סיכום או הפקת קוד, סימן שההתקפה היא בעלת יכולת עמידות לאורך משימות שונות. במערכות שמתעדכנות באופן רציף, יציבות ה־ASR על פני מחזורי אימון מלמדת האם ההרעלה ממשיכה להיות מוטמעת לאורך זמן.[80]

2.4.3 מסגרות הערכה מתקדמות

ככל שהמחקר מבשיל, מתודולוגיות ההערכה מתרחבות מעבר להקשרים של יריב יחיד. מסגרות עדכניות מביאות בחשבון דינמיקה תחרותית בין תוקפים מרובים מתוך הכרה בכך שבמציאות יריבים מרובים עשויים לנסות לבצע הרעלה במקביל. בתנאים האלה, ASR גבוה עבור תוקף יחיד אין פירושו בהכרח שאותו תוקף השיג דומיננטיות, שכן אינטראקציות בין התקפות מתחרות יכולות לשנות את התוצאות הכוללות. ראיות אמפיריות מלמדות שהתקפות חלשות בתרחישים של ריבוי יריבים יכולות לכאורה להשיג ביצועים טובים יותר מהתקפות חזקות, ובכך להוביל לפיתוח פרדיגמות הערכה השוואתיות חדשות.[81]

כמו כן התחום מאמץ יותר ויותר שילוב של מדדי ייחוס סטנדרטיים, שבהם סביבות בדיקה ומדדים אחודים מאפשרים השוואה ישירה בין סוגי התקפות, למערכי נתונים ולארכיטקטורות

79 Bagdasaryan et al., “How To Backdoor Federated Learning.”; Fendley et al., “Systematic Review.”; Jiao et al., “Can We Trust Embodied Agents?.”; Souly et al., “Poisoning Attacks on LLMs.”; Xia et al., “Poisoning Attacks in Federated Learning.”

80 Bagdasaryan et al., “How To Backdoor Federated Learning.”; Fendley et al., “Systematic Review.”; Gu et al., “BadNets.”; Li et al., “Backdoor Attacks on Pre-Trained Models.”; Muñoz-González et al., “Towards Poisoning.”; Souly et al., “Poisoning Attacks on LLMs.”; Xia et al., “Poisoning Attacks in Federated Learning.”

81 Biggio and Roli, “Wild Patterns.”; Halder et al., “Comprehensive Survey.”; Panda et al., “SparseFed.”; Pawlicki et al., “Meta-Survey.”; Avi Schwarzschild et al., “Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks,” preprint, arXiv, June 17, 2021, https://doi.org/10.48550/arXiv.2006.12557; Shah et al., “Guarding the Gates.”; Wang et al., “Threats to Training,” 1-36.

של מודלים. סטנדרטיזציה זו מאפשרת הערכה עקבית יותר וניתנת לשחזור הן של יעילות ההתקפה והן של חוסן ההגנה.[82]

2.4.4 קריטריונים להצלחה התלויים בהקשר

המשמעות של "הצלחה" במבצע להרעלת נתונים תלויה בסופו של דבר בהקשר. התקפה על מעבדה עשויה להשיג ASR של 95 אחוזים עם נזק אגבי זניח, ועדיין להיכשל מבחינה מבצעית אם היא לא תוכל לחמוק מייחוס, לשרוד מעבר לעדכוני המודל או להשפיע באופן משמעותי על המערכות וההחלטות של מטרת ההתקפה.[83] לפיכך במחקר העכשווי הערכת ההצלחה של התקפות הרעלה מתבצעת לא רק על בסיס מניפולציות גולמיות של דיוק, אלא לפי האיזון שהן מצליחות להשיג בין אפקטיביות לחשאיות. ההתקפות המסוכנות ביותר הן אלו המצטיינות בשני הקריטריונים — משיגות את מטרותיהן בחשאי בעודן שומרות על מראה תמים הן עבור בודקים אנושיים והן עבור הגנות אוטומטיות.[84]

82 Fendley et al., "Systematic Review."; Oliynyk et al., "I Know What You Trained."; Pawlicki et al., "Meta-Survey."; Schwarzschild et al., "Just How Toxic is Data Poisoning?."; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training," 1-36.

83 Conti, "Data Poisoning as a Covert Weapon."; Chen et al., "AgentPoison."; Cinà et al., "Wild Patterns Reloaded," 1-39; Department of Homeland Security, *Risks and Mitigation Strategies*; Geiping et al., "Witches' Brew."; Jiao et al., "Can We Trust Embodied Agents?."; Oliynyk et al., "I Know What You Trained."; Panda et al., "SparseFed."; Qiu, "Survey on Poisoning Attacks."; Ramirez et al., "Poisoning Attacks and Defenses."; Steinhardt et al., "Certified Defenses."; Zou et al., "PoisonedRAG."

84 Conti, "Data Poisoning as a Covert Weapon."; Bagdasaryan et al., "How To Backdoor Federated Learning."; Fendley et al., "Systematic Review."; Tian et al., "Comprehensive Survey," 1-35; Vassilev, *Adversarial Machine Learning*.

2.5
מדוע קשה לזהות התקפות הרעלת נתונים

בהמשך לדיון הקודם על מדדי יעילות וחמקנות, הקושי המרכזי בהגנה מפני הרעלת נתונים הוא שההתקפות היעילות ביותר מתוכננות במפורש להישאר בלתי נראות. דגימות מורעלות משתלבות על פי רוב ללא הפרעה בתוך קורפוסים גדולים של אימון, והמודלים המתקבלים ממשיכים לתפקד כרגיל עם קלטים נקיים, ועל כן לא מספקים למגינים סיבות רבות לחשוד במניפולציה. בו בזמן תהליכי למידת מכונה מודרניים מכילים נקודות עיוורון מבניות, החל בצבירת נתונים אוטומטית וכלה בתהליכי אימון אטומים, שמחלישים עוד יותר את יכולות הזיהוי. כתוצאה מכך זיהוי מהימן של הרעלה דורש התגברות על שלושה רבדים של אתגרים: החשאיות של הנתונים המורעלים עצמם, יכולת המודל להסוות התנהגות זדונית ומגבלות מערכתיות בכלים ובתהליכי העבודה ההגנתיים שמקובלים בשלב זה.

2.5.1 חשאיות בנתוני אימון מורעלים

קלטים זדוניים שאינם ניתנים לזיהוי: היריבים יוצרים דגימות אימון מורעלות כך שלא יהיה ניתן להבדיל ביניהם ובין נתונים לגיטימיים, ובכך הם מצליחים לעבור בהצלחה בדיקות ידניות ומסננים אלגוריתמיים בסיסיים.[85] לדוגמה בהתקפות clean-label, התוקף משבש את מאפייני הקלט (למשל פיקסלים של תמונה או טוקנים של טקסט) אך שומר על תוויות נכונות או סבירות.[86] כאשר גורמי ההרעלה האלה נראים מתויגים כהלכה וריאליסטיים, הם מצליחים לחמוק מבדיקות של עקביות התוויות ומגלאי חריגות.[87] התקפות מסוג clean-label נטמעות בתפוצת נתוני האימון, ובכך הופכות ל״כמעט בלתי ניתנות לזיהוי באמצעות מנגנוני הגנה קלאסיים המבוססים על איתור של אי־התאמות״, כפי שמחקרים עדכניים מאשרים.[88]

85 Department of Homeland Security, *Risks and Mitigation Strategies*; Muñoz-González et al., “Towards Poisoning.”; Pawlicki et al., “Meta-Survey.”

86 Cinà et al., “Wild Patterns Reloaded,” 1-39.

87 Pawlicki et al., “Meta-Survey.”

88 Ramirez et al., “Poisoning Attacks and Defenses.”

תוקפים יכולים גם להשתמש ב**שינויים זעירים או בלתי מורגשים:** השינויים המוזרקים הם כה קטנים (למשל כוונונים קטנים ברמת הפיקסל או ביטויים טקסטואליים שאינם מושכים תשומת לב) עד שהם גורמים לסטייה זניחה לפי מדדי דמיון כמו SSIM או BLEU. הסבירות הסטטיסטית שמתייגים אנושיים ומסננים אוטומטיים יבחינו בסטיות אלו היא נמוכה, במיוחד כאשר ההרעלות נדירות ביותר (כ־0.01 אחוז או פחות ממערך הנתונים). בפועל, קמפיינים של הרעלה מחדירים לרוב רק קומץ של דוגמאות זדוניות לתוך קורפוסים עצומים, כך שכל רכיב הרעלה יחיד הוא כמו מחט בערמת שחת. רשתות עמוקות מודרניות יכולות אפילו לשמור בזיכרון דוגמאות חריגות שכאלו מבלי שיהיו לכך תופעות לוואי ברורות על הביצועים הכוללים, ובכך הן מאפשרות להתקפה להישאר מתחת לרדאר.[89]

שיעורי הרעלה מזעריים: היעילות של הרעלת נתונים אינה דורשת עוד השחתה של חלק גדול מנתוני האימון. במקום זאת תוקפים מצליחים להשיג תוצאות באמצעות מספר מוחלט קטן של מופעי הרעלה. ניסויים שנערכו לאחרונה על ידי Anthropic ואחרים חשפו כי די היה ב־250 מסמכים זדוניים בלבד (כ־0.00016 אחוז מהטוקנים של האימון) כדי להשתיל דלת אחורית במודל שפה שכולל 13 מיליארד פרמטרים.[90] במילים אחרות, הרעלה של כמה מאות מתוך מיליארדי דגימות אימון יכולה לפגוע אפילו במודלים גדולים מאוד. יכולת זו לשמור על חשאיות הזיהום (שבריר של אלפית האחוז מהנתונים) נמצאת הרחק מתחת לסף הגילוי של כל שיטה סטטיסטית ידועה לאיתור חריגים. שום תהליך קיים לביקורת נתונים אינו יכול לאתר ולסמן באופן מציאותי שיבושים זעירים שכאלה בתוך הים העצום של נתונים שפירים. תוקפים מסווים עוד יותר את הרעלים על ידי הטמעת טריגרים בדרכים נסתרות. לדוגמה, טריגר של דלת אחורית עשוי להיות מוסתר בביטים המשמעותיים פחות של הפיקסלים בתמונה או חבוי בתוך מקטעי טקסט או קוד תמימים, כגון מטא־נתונים, רווחים לבנים (בלוקים) או מחרוזות תיעוד. שיטה אחת להרעלת תמונות, Pixdoor, הראתה כי מניפולציה רק על הביטים המשמעותיים פחות של הפיקסלים יכולה

89 Carlini et al., “Poisoning Web-Scale Training Datasets.”; Chen et al., “Targeted Backdoor Attacks.”; Department of Homeland Security, *Risks and Mitigation Strategies*; Geiping et al., “Witches’ Brew.”; Qiu, “Survey on Poisoning Attacks.”; Souly et al., “Poisoning Attacks on LLMs.”; Yiming Zhang et al., “Persistent Pre-Training Poisoning of LLMs,” preprint, arXiv, October 17, 2024, https://doi.org/10.48550/arXiv.2410.13722; Zhao et al., “Data Poisoning in Deep Learning.”

90 Souly et al., “Poisoning Attacks on LLMs.”

ליצור התקפת דלת אחורית שכמעט בלתי ניתנת לגילוי.[91] על ידי הטמעת טריגרים בתוכן או שימוש בתיוגים שהם סבירים מבחינה סמנטית אך זדוניים, היריבים יכולים לדאוג לכך שדגימות מורעלות יחמקו מגלאים המבוססים על התאמת תבניות, ויישארו בלתי ניתנות להבחנה ביחס לנתונים "נקיים".[92]

2.5.2 הסוואה של התנהגות המודל

אף שדוח זה מתמקד במניפולציה של שכבת הנתונים, הבנת ההתנהגות של המודל חיונית מכיוון שנתונים מורעלים פועלים באופן שבו הם מלמדים את המודל להסתיר את התנהגותו הזדונית. התקפת הרעלה מתוכננת היטב מייצרת מודל שנראה נורמלי לחלוטין בתנאי הערכה סטנדרטיים: רמת הדיוק בעת שימוש בנתוני אימות נקיים נשארת קרובה לערך הבסיס, בדרך כלל בטווח של 1 אחוז בהשוואה למודל שלא הורעל.[93] מאחר שניטור שגרתי מתמקד במדדים המצטברים האלה, דבר אינו נראה חשוד. למודלים מודרניים של למידה עמוקה יש קיבולת מספקת לקלוט מספר קטן של דגימות מורעלות מבלי לפגוע בביצועים הכלליים, ובכך לאפשר להתנהגות הזדונית להישאר רדומה אלא אם טריגר מסוים נוכח. למעשה מודל מורעל פועל כ"סוס טרויאני" — אינו מזיק כמעט בכל מצב, אך בנוי כך שייכנס לפעולה בתנאים מוגדרים.[94]

הטריגרים האלה הם נדירים וממוקדים מאוד בכוונה. המודל עשוי לפעול בצורה לא תקינה — כלומר, לייצר את הפלט המזיק הספציפי או הפלט הרצוי לתוקף — רק כאשר הוא נתקל בביטוי, ברצף טוקנים או בדפוס חזותי מסוימים. עבור כל שאר הקלטים הוא ממשיך להתנהג כרגיל. הפעלה סלקטיבית זו ממזערת את טביעת הרגל הסטטיסטית של ההתקפה: שיעורי השגיאה, התפלגויות הפלט והמדדים של אי־ודאות החיזוי נותרים בלתי ניתנים להבחנה ביחס לאלה של מודל נקי.[95] טכניקות זיהוי מסורתיות, כגון מנגנוני הגנה מסוג RONI, המיושמים באמצעות נוהלי הסרה ובחינה, אימות hold-out או בדיקות

91 Zhao et al., "Data Poisoning in Deep Learning."

92 Cinà et al., "Wild Patterns Reloaded," 1-39; Fendley et al., "Systematic Review."; Li et al., "Backdoor Learning."; Zhang et al., "Human-Imperceptible Retrieval Poisoning Attacks."; Zhao et al., "Data Poisoning in Deep Learning."

93 Ma et al., "Backdoor Attack with Invisible Triggers."

94 Chen et al., "Targeted Backdoor Attacks."; Tian et al., "Comprehensive Survey," 1-35.

95 Bagdasaryan et al., "How to Backdoor Federated Learning."; Lumenova, "Data Poisoning Attacks."

התנהגותיות במודל קופסה שחורה — חסרות את הרגישות הנדרשת לאיתור מיקרו־דלתות אחוריות מסוג זה, משום שהן נשענות על שינויים רחבים בביצועים ולא על כשלים עדינים המופעלים על ידי טריגר.[96]

שיטות פנימיות לזיהוי חריגות אינן מוצלחות יותר. הגנות כגון ניתוח של חתימה ספקטרלית או קיבוץ הפעלות מניחות שדגימות מורעלות ייצרו נתונים חריגים הניתנים לזיהוי במרחב המאפיינים הפנימי של המודל.[97] אסטרטגיות הרעלה מודרניות תוכננו במפורש כדי להימנע מכך: התקפות clean-label מבטיחות שקלט מורעל ייראה זהה מבחינה סטטיסטית לנתונים אמיתיים, ולא יותיר דפוס ייחודי שהגלאים יוכלו לבודד. תוצאות אמפיריות מצביעות בעקביות על דיוק ואחזור נמוכים — לעיתים קרובות הרבה מתחת ל־50 אחוזים — במיוחד כאשר שיעורי ההרעלה נמוכים מתחת לסף של אחוז אחד.[98] התקפות מסוימות מסבכות עוד יותר את יכולת האיתור בכך שהן מפיצות טריגרים או משלבות טכניקות, כך שאף דפוס ספקטרלי או דפוס אקטיבציה יחיד אינו בולט. כל עוד ההרעלות מותאמות בצורה מוקפדת לשמירה על החשאיות, מודל דלת אחורית יכול לעבור בדיקות פנימיות של גרדיאנטים, הפעלות וייצוגים מבלי לעורר חשד.[99]

2.5.3 אתגרים מערכתיים ומגבלות של ההגנות

תהליכי למידת מכונה מודרניים מכילים נקודות עיוורון מבניות המקשות על זיהוי של הרעלת נתונים, גם כשהמגינים מבינים את מהות האיום. אתגרים אלו נובעים מהאופן שבו הנתונים נאספים, מהאופן שבו המודלים מאומנים, ומהאופן שבו ההגנות פועלות הלכה למעשה.

אפיקי נתונים גדולי ממדים שלא נבדקו: מודלים עכשוויים מסתמכים על מיליארדי דוגמאות אימון שנאספו מהרשת הפתוחה, ממכשירי משתמשים או ממקורות צד שלישי. בהיקף כזה, הן ביקורת אנושית והן ביקורת אוטומטית אינן יכולות לבדוק את הנתונים בצורה מלאה. כל דוגמה מורעלת יחידה הופכת לזניחה מבחינה סטטיסטית, ומאפשרת לתוקפים

96 Turner et al., *Clean-Label Backdoor Attacks*; Jagielski et al., “Manipulating Machine Learning.”

97 Jagielski et al., “Manipulating Machine Learning.”; Pawlicki et al., “Meta-Survey.”

98 Cristina Improta, “Detecting Stealthy Data Poisoning Attacks in AI Code Generators,” preprint, arXiv, August 29, 2025, https://doi.org/10.48550/arXiv.2508.21636.

99 Bagdasaryan et al., “How To Backdoor Federated Learning.”; Li et al., “Backdoor Learning.”; Panda et al., “SparseFed.”

להסתיר דגימות זדוניות בגלוי.[100] ככל שמערכי הנתונים גדלים, גלאי חריגות מאבדים מכוחם, וארגונים — שאינם מסוגלים לבצע בקרת נתונים בהיקף של האינטרנט — נאלצים לקבל כמויות גדולות של קלטים שאינם מהימנים כברירת מחדל. התוקפים מנצלים את האסימטריה הזו: הזרקה של מספר נקודות מורעלות בעלות מראה נקי לתוך קורפוס עצום היא מטלה זולה לביצוע וכרוכה בסיכון נמוך, וסביר להניח שהיא לא תתגלה.[101]

תהליכי אימון מבוזרים ואטומים: במערכות רבות אין לגורמי ההגנה גישה ישירה לנתוני האימון. למידה מבוזרת (FL), אימון שיתופי וארכיטקטורות השומרות על הפרטיות מאחסנות נתונים על גבי מכשירי משתמשים או מערכות של שותפים, ובכך מונעות בדיקה ריכוזית. משתתף זדוני יכול "לתרום" עדכונים מורעלים או נתונים מקומיים מורעלים מבלי לחשוף אי פעם את הדגימות הבסיסיות. אפילו מחוץ ללמידה מבוזרת, שרשראות אספקה של למידת מכונה מסתמכות במידה רבה על מערכי נתונים של צד שלישי ומודלים שעוברים אימון מקדים. כאשר מתרחשת הרעלה של מקור במעלה הזרם, משתמשים במורד הזרם עלולים לרשת את הפגיעה ללא ידיעתם. חשיפת "קופסה שחורה" זו פירושה שגורמי הגנה עלולים שלא לזהות התקפה אלא רק לאחר זמן רב מפריסת המודל.[102]

הגנות שבריריות ותוקפים מסתגלים: למרות הסוגים הרבים של מנגנוני ההגנה המוצעים (לדוגמה כלי סטטיסטיקה חסינים, מסנני חריגות, ניפוי באגים בנתוני האימון), רובם מניחים שקלט מורעל ייראה חריג, יפגע ברמת הדיוק או ייצור אשכולות הניתנים לגילוי במרחב התכונות. ההתקפות של ימינו מתוכננות להפר את הנחות הבסיס הללו. התקפות clean-label משתלבות בתפוצת הנתונים, דלתות אחוריות ללא טריגר אינן יוצרות חתימה ספקטרלית ברורה, והרעלות בשיעור נמוך חומקות כמעט מכל כלי הסינון הקיימים.[103] מחקרים אמפיריים מוכיחים שוב ושוב שברמות הרעלה מציאותיות, אמצעי גילוי מתקדמים מתפקדים רק מעט טוב יותר מניחוש אקראי.[104] בינתיים התוקפים מתאימים

100 Geiping et al., "Witches' Brew."; Jagielski et al., "Subpopulation Data Poisoning Attacks."; Schwarzschild et al., "Just How Toxic is Data Poisoning?."; Vassilev, *Adversarial Machine Learning*.

101 Department of Homeland Security, *Risks and Mitigation Strategies*; Geiping et al., "Witches' Brew."; Vassilev, *Adversarial Machine Learning*; Xu et al., "Detecting AI Trojans."

102 Bagdasaryan et al., "How to Backdoor Federated Learning."; Cinà et al., "Wild Patterns Reloaded," 1-39; Gu et al., "BadNets."; Tian et al., "Comprehensive Survey," 1-35; Zhao et al., "Data Poisoning in Deep Learning."

103 Improta, "Detecting Stealthy Data Poisoning Attacks."

104 Improta, "Detecting Stealthy Data Poisoning Attacks."

ללא הרף את שיטות הפעולה שלהם כדי לעקוף אמצעי הגנה חדשים, ויוצרים דינמיקה של מירוץ חימוש שגורמי ההגנה מתקשים לעמוד בקצב שלו.

אילוצי תפעול ואילוצים כלכליים: גם כאשר קיימים אמצעי נגד יעילים, הם לרוב יקרים מדי או משבשים מדי מכדי לפרוס אותם בהיקף רחב. טיהור נתונים אינטנסיבי, ניטור רציף של הפלט או שיטות אימון החסינות באופן מוכח עלולים להאט את הפיתוח, לפגוע בדיוק או ליצור תקורה חישובית משמעותית. סינון קפדני עלול גם לגרום לפסילה של נתונים לגיטימיים ולפגוע בביצועי המודל. כתוצאה מכך ארגונים רבים מוכנים להשלים עם רמה מסוימת של סיכון להרעלה במקום ליישם מנגנוני הגנה שעלולים לפגוע בתועלת המודל או בלוחות הזמנים של התפעול. נוסף על כך למנגנוני ההגנה חסר בדרך כלל קו בסיס "נקי" וידוע להתנהגות הצפויה של מודל, וזה מקשה להצדיק התערבויות יקרות אלא במקרים שבהם דלת אחורית כבר נחשפה — לעיתים קרובות זה כבר מאוחר מכדי למנוע נזק.[105]

זיהוי הרעלת נתונים הוא עדיין בעיה ללא פתרון. הרעלות מודרניות בנויות כך שיישארו חשאיות — מתויגות בצורה נכונה, טבעיות מבחינה חזותית או לשונית, מוסתרות במעמקיהם של מערכי נתונים עצומים, ומופעלות רק בתנאים נדירים. המודלים המתקבלים הם בעלי חזות נקייה, משיגים ציונים טובים בכל מדד סטנדרטי ואינם חושפים בעיה כלשהי בבדיקות שגרתיות. חולשות מבניות הופכות את המשימה לקשה עוד יותר: אפיקי נתונים שאינם מבוקרים, אימון אטום או מבוזר וכלי הגנה שמחמיצים באופן קבוע התקפות מתוכננות היטב. מחקרים עדכניים מבהירים את העניין הזה בצורה חדה: כלי גילוי רבים מהחדישים ביותר מתפקדים רק מעט טוב יותר מניחוש אקראי.

105 Cinà et al., "Wild Patterns Reloaded," 1-39; Geiping et al., "Witches' Brew."; Panda et al., "SparseFed."; Steinhardt et al., "Certified Defenses."; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training," 1-36; Xu et al., "Detecting AI Trojans."

איור 6: מדוע קשה כל כך לגלות הרעלות

2.6
השחקנים מאחורי הרעלות הנתונים

תחום הרעלת הנתונים התפתח למערכת איומים רב־שכבתית הכוללת קשת רחבה של שחקנים. אלו נעים ממדינות לאום השואפות להשיג יעדים אסטרטגיים ועד לשחקנים פרטיים המונעים מאידאולוגיה, רווח או תחושת עוול. היות שהרעלה מתמקדת ביסודות הנתונים של למידת מכונה — ולא בקוד או בממשקי הרשת שלה — חסמי הכניסה נמוכים באופן משמעותי. כאמור, אפילו התערבויות מוחלטות בקנה מידה קטן יכולות לשנות באופן משמעותי את התנהגות המודל. אסימטריה זו מעצימה הן שחקני עילית והן שחקנים דלי משאבים, ומאתגרת את הנחות הבסיס המסורתיות לגבי יכולות ההגנה בעימותי סייבר.[106]

מדינות לאום וקבוצות שהן בחסות המדינה ניצבות בראש האקוסיסטם הזה. כאשר הם מונעים על ידי יעדים גאופוליטיים, איומים מתמידים מתקדמים (APTs) אלו משתמשים בהרעלת נתונים כדי לערער את הביטחון במודיעין מבוסס־בינה מלאכותית, לשבש מערכות התומכות בקבלת החלטות, לפגוע בתשתיות קריטיות ובאופן כללי לשחוק באיטיות את האמון בתהליכים שנתמכים על ידי בינה מלאכותית או שמבוססים עליה. הגורמים הללו פועלים בסבלנות, בהתמדה, והם בעלי יכולת הכחשה סבירה; פעולותיהם מבוצעות לעיתים קרובות באמצעות שילוב של חדירות ישירות או שימוש בגורמי פרוקסי במיקור חוץ.[107]

---

106 Conti, "Data Poisoning as a Covert Weapon."; Department of Homeland Security, *Risks and Mitigation Strategies*; Souly et al., "Poisoning Attacks on LLMs."

107 Conti, "Data Poisoning as a Covert Weapon."; Bailey Galicia, "In the Fight against Foreign Information Manipulation, the US Can't Afford to Disarm," *New Atlanticist*, 2025, https://tinyurl.com/mdpbb2u9; Department of Homeland Security, *Risks and Mitigation Strategies*; Ioana Puscas, *AI and International Security: Understanding the Risks and Paving the Path for Confidence Building Measures* (UNIDIR, 2023), https://tinyurl.com/3p5uvfw8; Lumenova, "Data Poisoning Attacks."; Paul B. Stephan III, "Big Data as a National Security Issue," *University of Chicago Legal Forum* 2024 (2025), https://tinyurl.com/38k77vun; Châtelet, "Exposing Pravda."; Jun Zhang and Dan Tenney, "The Evolution of Integrated Advance Persistent Threat and Its Defense Solutions: A Literature Review," *Open Journal of Business and Management* 12, no. 1(2024): 293-338, https://doi.org/10.4236/ojbm.2024.121021.

לצידם פועלות רשתות של פרוקסי ורשתות של קבלנים — האקרים פטריוטים, חברות פרטיות המזוהות עם מדינות שונות וחברות קבלנות המציעות "בינה מלאכותית כשירות". המתווכים האלה מטשטשים את הנשיאה באחריות ומספקים מומחיות טכנית וכיסוי שוק לגיטימי. מספר מקרי הרעלה מהעת האחרונה, שכוונו כלפי מאגרי מודלים ציבוריים ושירותי AI middleware, יוחסו לאזור האפור הזה של השחקנים, שפועלים בו־זמנית כספקים מסחריים וככלים בעלי תפקיד אסטרטגי.[108]

גורמי פנים ממשיכים למלא תפקיד מסוכן במיוחד. עובדים מהימנים, קבלנים או שותפים לנתונים מחזיקים בהרשאות גישה לתהליכי האימון והתיוג. גורמי פנים זדוניים פועלים בדרך כלל מתוך תחושת נקמה או עוול, ומחדירים שגיאות קלות יחסית לגילוי או מבצעים זיהום הרסני. לעומתם קיימים גורמי פנים המגויסים על ידי גופי המדינה, היריבים או ארגוני הפשע כדי להשתיל נקודות תורפה עדינות ובעלות יכולת שרידות לטווח הארוך. קטגוריה אחרונה זו משלבת את יתרון הגישה של גורמי הפנים עם סבלנות ברמת APT, ויוצרת יריבים מורכבים שמשלבים אמון ארגוני עם כוונות של גורמים זרים.[109]

מתחרים עסקיים מייצגים סוג הולך וגובר של שחקנים שאינם מזוהים עם מדינה זו או אחרת, ועם זאת הם בעלי רלוונטיות רבה. ככל שיכולות הבינה המלאכותית הופכות לגורם מרכזי בהשגת יתרון עסקי, חלק מהחברות עלולות לנסות לשבש או לעוות את המודלים של המתחרים — על ידי הטיה של פלטים, ערעור של אמון הלקוחות או פגיעה בניתוחים אנליטיים כדי להשפיע על החלטות השוק. הקמפיינים האלה יכולים לנוע מזיהום גס של נתונים ועד למניפולציות חשאיות במיקור חוץ, הממנפות חברות קבלן או ספקי נתונים שמאגרי המידע שלהם נפרצו.[110]

---

108 Peiran Dong et al., "Investigating Trojan Attacks on Pre-Trained Language Model-Powered Database Middleware, "*Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (ACM, 2023), 437–47, https://doi.org/10.1145/3580305.3599395; Gu et al., "BadNets."; Stephan III, "Big Data."; Pawlicki et al., "Meta-Survey."; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training," 1-36.

109 Conti, "Data Poisoning as a Covert Weapon."; Department of Homeland Security, *Risks and Mitigation Strategies*; Ayshwarya Jaiswal et al., "Machine Learning Approaches to Detect, Prevent and Mitigate Malicious Insider Threats: State-of-the-Art Review," *Multimedia Tools and Applications* 84, no. 24 (2024): 28909-49, https://doi.org/10.1007/s11042-024-20273-0; Lumenova, "Data Poisoning Attacks."; Zou et al., "PoisonedRAG."

110 Minghong Fang et al., "Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems," preprint, arXiv, May 31, 2020, https://doi.org/10.48550/

ארגוני פשיעת סייבר — ובהם קבוצות כופרה וגורמים בשווקים בלתי חוקיים — עושים שימוש בהרעלת נתונים כחלק ממאמצי הונאה וניצול כלכלי. מטרתם אינה לשבש אלא להשיג יתרון בצורה חשאית: מודלים מורעלים לגילוי הונאות שאינם מצליחים לאתר עסקאות בלתי חוקיות או אלגוריתמים למסחר שעברו מניפולציה, ועל כן שוגים בתמחור הסיכונים. קבוצות פשיעה מציעות כעת גם "הרעלת נתונים כשירות", ומשכירות גישה למאגרי נתונים, לתהליכי עיבוד או למודלים שאומנו מראש.[111]

האקוסיסטם של שרשרת האספקה וספקי המודלים הפך הן למטרה והן לווקטור. הרעלה יכולה לחדור דרך משקלים שאומנו מראש, תוספי תוכנה או מרכזי מודלים ציבוריים. תקריות המערבות מודלים בסגנון "PoisonGPT" מדגימות כיצד רכיבי תוכנה נגועים יכולים להתפשט באופן נרחב לאחר הפצתם. שכבה זו מכניסה למערכה שחקן כפול: התוקף שזורע את המודל הזדוני, והמארח המתווך שהתשתית שלו מעצימה את תפוצתו. מהדורה מורעלת אחת עלולה לסכן מאות מערכות במורד הזרם דרך שרשרת האספקה.[112]

האקטיביסטים (Hacktivists – האקרים אקטיביסטים) וגורמים אידאולוגיים משתמשים בהרעלה למטרות תדמית, פוליטיקה או אקטיביזם. חלקם מבקשים לגרום מבוכה למוסדות או להדגיש הטיות באמצעות מניפולציה של מערכות בינה מלאכותית החשופות לעין הציבור. אחרים, כגון אומנים המשתמשים בכלי הרעלה עצמית כמו Nightshade, משתמשים בהרעלה באופן הגנתי כדי להתנגד לאיסוף נתונים שאינו מורשה. פעילויות אלו מטשטשות את הגבולות המשפטיים ומגדירות מחדש את הפרקטיקה של הרעלת נתונים הן כשיטת מחאה והן כמנגנון להגנה על קניין רוחני.[113] יריבים דלי משאבים ואופורטוניסטים — חובבים,

111 arXiv.2002.08025; Lumenova, "Data Poisoning Attacks."; Müller et al., "Data Poisoning Attacks on Regression Learning."; Zhang et al., "Persistent Pre-Training Poisoning." Biggio and Roli, "Wild Patterns."; Fang et al., "Influence Function."; Jagielski et al., "Manipulating Machine Learning."; Lumenova, "Data Poisoning Attacks."; "The Business Model of Data Poisoning-as-a-Service (DPaaS)," *AI Competence.Org*, October 12, 2025, https://tinyurl.com/5fh3hdj6; Travis Rosiek, "AI Data Poisoning, Wiper Malware, Critical Infrastructure Attacks Could Increase in 2025, Impacting Government Cyber Resilience," *Govloop*, January 21, 2025, https://tinyurl.com/2bh35yna.

112 Gu et al., "BadNets."; Li et al., "Backdoor Learning."; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training," 1-36; Zhou et al., "Survey on Backdoor Threats."

113 Pawlicki et al., "Meta-Survey."; Shawn Shan et al., "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models," preprint, arXiv, April 29, 2024, https://doi.org/10.48550/arXiv.2310.13828; Zhao et al., "Data Poisoning in Deep Learning."

תאים קטנים של אקטיביסטים או שחקנים הפועלים באופן עצמאי — מנצלים גם הם את נקודות התורפה האלה.

לבסוף, קמפיינים של הרעלה מאופיינים יותר ויותר בשיתוף פעולה היברידי. מדינות נעזרות בשירותיהם של פושעי סייבר כדי להכחיש ייחוס; תאגידים מממנים בחשאי קמפיינים של פעילים להפעלת לחץ בנושאים שונים; וגורמי פנים עובדים מהצד כסוחרי נתונים. הפעולות המשולבות האלה מקשות על האפשרות לייחס את התקיפה לגורם ספציפי ומרחיבות את הנוף האסטרטגי של הרעלת נתונים.[114] מספר מקרים ממחישים את המודל המשולב הזה: לדוגמה, השימוש של רוסיה בגורמים חצי־פליליים כגון TeleBots וקבלנים אחרים המזוהים עם סוכנות המודיעין הצבאי של רוסיה (GRU) במבצע NotPetya, וכן רשתות עריכה מתואמות בוויקיפדיה הקשורות לסין, ששכרו פרילנסרים אזרחיים כדי לזרוע שינויים עדינים בנרטיב.[115]

114 Zhang et al., "SoK."

115 Chris Vallance, "Wikipedia Blames Pro-China Infiltration for Bans," *BBC*, September 16, 2021, https://tinyurl.com/a3dph5w4; U.S. Department of Justice, *Six Russian GRU Officers Charged in Connection with Worldwide Deployment of Destructive Malware and Other Disruptive Actions in Cyberspace* (2020), https://tinyurl.com/yernt666.

טבלה 3: השחקנים שמאחורי הרעלת נתונים

גורם	מוטיבציה	יכולות מרכזיות
מדינות לאום / APTs	• יתרון גאופוליטי • שיבוש אסטרטגי • לוחמה קוגניטיבית	• גישה לטווח ארוך ופעילות הניתנת להכחשה • קמפיינים סמויים ומתמשכים של הרעלת נתונים
גורמי פרוקסי שמזוהים עם מדינות	• יכולת הכחשה סבירה • רווח מסחרי והזדהות פוליטית	• התמחות טכנית • זריעת מאגרי נתונים או תשתיות של מודלים ציבוריים
גורמי פנים (זדוניים/מגויסים)	• מרמור או כפייה • גיוס באמצעות גורמים חיצוניים	• גישה ישירה לתהליכי אימון • מניפולציה של תהליכי תיוג ומאגרי נתונים
מתחרים עסקיים	• חבלה כלכלית • פגיעה במוניטין והשגת יתרון תחרותי	• מניפולציה ממוקדת של מערכי נתונים • השחתה של רכיבי מודל משותפים/במיקור חוץ
ארגוני פשיעת סייבר	• רווח כספי והפקת רווחים מנתונים • סיוע לביצוע הונאות	• הרעלת מודלים לגילוי הונאות • "הרעלה כשירות" וניצול של שרשראות אספקה
גורמים בשרשרת האספקה	• ניצול של ערוצי הפצה • הגברה בשוגג (לא מודעת)	• הרעלה של משקלים שאומנו מראש, של תוספים או של מרכזים • יצירת אפשרות לפגיעה רחבה במורד הזרם
האקטיביסטים	• הפצת מסרים פוליטיים ושיבוש • מחאה והגנה על קניין רוחני	• מניפולציה גלויה של מודלים והתקפות מסגור • כלים להרעלה עצמית (למשל Nightshade)
גורמים דלי משאבים	• סקרנות, אופורטוניזם, אקטיביזם	• ביצוע מניפולציה בעלות נמוכה של מקורות נתונים פתוחים

2.7
מקרי בוחן בולטים של הרעלת נתונים

האקוסיסטם המגוון של השחקנים השונים שתואר לעיל כבר יצר תקריות של הרעלת נתונים; אלו ממחישות כיצד יכולות ומניעים שונים באים לידי ביטוי בפועל. ארבעת מקרי הבוחן הבאים מדגימים את הטווח הזה: מניפולציה ציבורית אופורטוניסטית המנצלת מערכות למידה פתוחות, הרעלה ממוקדת של מערכי נתונים בהיקף של האינטרנט בידי תוקפים המחזיקים בידע טכנולוגי מתוחכם, מבצעי השפעה ממושכים בהכוונה מדינתית שנועדו לעצב מאגרי אימון של בינה מלאכותית, וכן התקפות מבוססות־אופטימיזציה הממחישות כיצד מניפולציות נתונים זעירות ומתוכננות בקפידה מסוגלות לשתק מודלים קריטיים המשמשים בקבלת החלטות רגישות.

2.7.1 הרעלת הצ'אטבוט TAY של מייקרוסופט (2016)[116]

וקטור התקיפה: מניפולציה מתואמת של אינטראקצייה במדיה חברתית

משטח התקיפה: מערכת למידה אינטראקטיבית בזמן אמת באמצעות API של טוויטר

מתודולוגיית התקיפה: תוקפים שפעלו בתיאום ב־23 במרץ 2016, ניצלו נקודת תורפה קריטית בצ׳אטבוט Tay של מייקרוסופט על ידי הזנה שיטתית של תוכן גזעני, פוגעני ומסית דרך אינטראקציות בטוויטר. התוקפים מינפו את מנגנון הלמידה בזמן אמת של Tay, שתוכנן ללמוד דפוסי שיחה ממשתמשים בני 18–24. מייקרוסופט הודתה מאוחר יותר שזו הייתה

116 מקרה הבוחן מבוסס על המקורות האלה: Amy Kraft, “Microsoft Shuts Down AI Chatbot after It Turned into a Nazi,” *CBS News*, March 25, 2016, https://tinyurl.com/y6kfk8as; Dave Lee, “Tay: Microsoft Issues Apology over Racist Chatbot Fiasco,” *BBC*, March 25, 2016, https://tinyurl.com/5n84d8bd; Peter Lee, “Learning from Tay’s Introduction,” *Official Microsoft Blog*, March 25, 2016, https://tinyurl.com/xwmw2akw; “Tay (Chatbot),” Wikipedia, accessed October 10, 2025, https://tinyurl.com/3dy72bx2; M.J. Wolf et al., “Why We Should Have Seen That Coming,” *The ORBIT Journal* 1, no. 2 (2017): 1–12, https://doi.org/10.29297/orbit.v1i2.49.

״התקפה מתואמת על ידי קבוצת אנשים״ ש״ניצלה נקודת תורפה ב־Tay״, אף שהיא מעולם לא חשפה את המהות הטכנית הספציפית של הפגיעות.

היקף ההשפעה: בתוך 16 שעות מההפעלה, לאחר יותר מ־96 אלף ציוצים, Tay החל לייצר מילים ותמונות שאינן הולמות ומגונות, כולל הכחשת השואה, הצהרות גזעניות ותוכן פוליטי מסית. מייקרוסופט נאלצה להשבית את המערכת ופרסמה התנצלות פומבית, שבה הודתה ״ככל שזה נוגע להתקפה ספציפית זו, הייתה לנו החמצה קריטית״. התקרית הפכה לדוגמה מובהקת לאופן שבו אינטראקציות שאינן מסוננות ברשתות חברתיות, עלולות לגרום להשחתה מהירה של מערכות בינה מלאכותית שנועדו למעורבות ציבורית, והמחישה את הצורך בהטמעת סינון תוכן עמיד במערכות בינה מלאכותית אינטראקטיביות.

2.7.2 הרעלת מאגרי נתונים בהיקף של האינטרנט[117]

וקטור התקיפה: ניצול זמני של המנגנונים לאיסוף מערכי נתונים

משטח התקיפה: מערכי נתונים לאימון בהיקף של האינטרנט הנמצאים בשימוש של מודלי הבסיס

מתודולוגיית התקיפה: חוקרים מ־Google DeepMind ו־ETH Zürich הדגימו שתי התקפות משלימות המכוונות למערכי נתונים מרכזיים שנועדו לאימון של בינה מלאכותית. הרעלה מסוג frontrunning מנצלת את לוחות הזמנים הצפויים של ויקיפדיה להפצת צילומי מצב (snapshots) על ידי תזמון של עריכות זדוניות בדיוק לפני יצירת קובצי היצוא התקופתיים. הדבר הבטיח שתוכן מורעל יישאר במערכי הנתונים המיועדים לאימון גם לאחר פעולות בקרה ועריכה. הרעלה מסוג split-view מכוונת למערכי נתונים מבוזרים על ידי רכישת דומיינים, שפג תוקפם ומופיעים באינדקסים של מערכי הנתונים, והחלפת התוכן המקורי בנתונים זדוניים. כך נוצר פער בין התוכן שהוגדר ונאגר במקור על ידי מתחזקי מערך הנתונים ובין התוכן שהמשתמשים מורידים בפועל בשלב מאוחר יותר.

117 מקרה הבוחן מבוסס על המקורות האלה: Carlini et al., “Poisoning Web-Scale Training Datasets.”; Chris Stokel-Walker, “You Can Poison AI Datasets for Just $60, a New Study Shows,” *Fast Company*, March 3, 2023, https://tinyurl.com/32fet98s; J.P. Hennessy, “Why Google’s Researchers Are Intentionally Poisoning Datasets and More,” *Lightning AI*, February 23, 2023, https://tinyurl.com/m25p7r7x; Zhang et al., “Persistent Pre-Training Poisoning.”

היקף ההשפעה: ניתוח שמרני מראה שתוקפים יכלו לגרום להרעלה של 6.5 אחוזים לפחות מקובצי ויקיפדיה באנגלית באמצעות התקפות מסוג frontrunning. במקרה של מערכי נתונים מבוזרים, הראו החוקרים כי ניתן להרעיל כ־0.01 אחוז ממערך נתונים נתון בעלות של כ־60 דולר בלבד. התקפות אלו משפיעות על מודלי בסיס המסתמכים על מקורות אימון מזוהמים אלו, כולל אלו שמשתמשים בתוכן מוויקיפדיה (כגון BERT), ויוצרים הטיות מתמשכות או דלתות אחוריות ששורדות הליכי כוונון עדין והתאמה.

2.7.3 רשת Pravda: בינה מלאכותית ברמת המדינה והרעלת ויקיפדיה (2014–2025)[118]

וקטור התקיפה: מבצעי מידע אסטרטגיים המכוונים למקורות ידע

משטח התקיפה: ערכי ויקיפדיה וקורפוסים לאימון בינה מלאכותית באמצעות זיהום מקורות שיטתי

מתודולוגיית התקיפה: רשת המדיה Pravda הקשורה לרוסיה החלה לפעול ב־2014 כתשתית דיסאינפורמציה רחבה, אך רק בשנים האחרונות — ובעיקר מ־2021 ואילך — היא התפתחה לכדי איום ישיר על האקוסיסטם של הבינה המלאכותית. חקירות שבוצעו על ידי Viginum (צרפת) ו־Digital Forensic Research Lab של המועצה האטלנטית הראו שמאז שנת 2021 החדירה הרשת באופן שיטתי לוויקיפדיה נרטיבים התומכים בקרמלין על ידי יצירה ותחזוקה של יותר מ־180 אתרי חדשות מזויפים ושימוש בהם כ"מקורות מוסמכים". זוהו יותר מ־1,900 היפר־קישורים של רשת Pravda ב־44 מהדורות שפה של ויקיפדיה, כאשר בוויקיפדיה הרוסית יש 922 ערכים כאלה ובוויקיפדיה האוקראינית — 580, ורובם עוסקים בנושאים הקשורים לסכסוך רוסיה–אוקראינה.

היקף ההשפעה: המבצע מרעיל בהצלחה נתוני אימון של בינה מלאכותית על ידי הצבת תוכן שעבר מניפולציה כמקור לגיטימי עבור משתמשים ב־LLMs כמו ChatGPT, Gemini ו־Copilot במהלך האימון. אף אחת ממערכות הבינה המלאכותית הללו לא סיפקה אזהרות בעת שנשאלה על תוכן הקשור לרשת Pravda, ובכך למעשה העצימו את הדיסאינפורמציה

118 מקרה הבוחן מבוסס על המקורות האלה: "'Pravda' Network: Worldwide Expansion and LLM Wikipedia Pollution," CheckFirst, March 13, 2025, https://tinyurl.com/yj7wz9u2; Global Influence Operations Report, *Russian Information Warfare Campaign: Kremlin Poisons AI and Rewrites Wikipedia* (2025), https://tinyurl.com/368rbn52; Châtelet, "Exposing Pravda."

הרוסית באמצעות תגובות הבינה המלאכותית. בתקופות בחירות הגבירה הרשת את פעילותה, והדגימה לוחמת מידע מתואמת שניצלה הן את מנגנון האמון הפנימי של ויקיפדיה והן את הסתמכותן של מערכות בינה מלאכותית על ויקיפדיה כמקור אימון סמכותי. קישורים של רשת Pravda לא רק הופיעו בערכים אלא יצרו שינויים נרטיביים מוחשיים: מסגור מחדש של הפלישה לאוקראינה ב־2022 כ״סכסוך פנימי״, החדרת טענות מוגזמות על נפגעים אוקראינים, הוספת פרשנויות מומחים מפוברקות ממכוני מחקר מזויפים המזוהים עם רוסיה, ושינוי דפים ביוגרפיים של גורמים אוקראינים ומערבים כדי לכלול טענות משמיצות או מטעות.

2.7.4 הרעלת רגרסיה בתחום הבריאות[119]

וקטור התקיפה: מחקר בנושא הרעלת נתונים מבוססת־אופטימיזציה של מודלים לחיזוי רפואי

משטח התקיפה: מודלי רגרסיה קליניים לחיזוי של מינון תרופות (הדגמה מחקרית)

מתודולוגיית התקיפה: Jagielski ואחרים (2018) הדגימו התקפות הרעלה הרסניות נגד חיזוי המינון של ורפרין (Warfarin) באמצעות הרעלה מבוססת־אופטימיזציה (OptP) במחקר מבוקר. ההתקפה השתמשה באופטימיזציה בעלת שתי רמות ליצירת דוגמאות אימון מורעלות שהגדילו למקסימום את שגיאות החיזוי, ובתוך כך שמרו על התכנסות המודל, והתמקדו במודלי רגרסיה ליניארית המוצעים בדרך כלל עבור החלטות קליניות בטיפול בנוגדי קרישה. המחקר השתמש במערך הנתונים של הקונסורציום הבין־לאומי לפרמקוגנטיקה של ורפרין (IWPC) כדי להדגים כיצד הרעלת נתונים עלולה באופן תאורטי לפגוע במערכות בינה מלאכותית בתחום הבריאות.

היקף ההשפעה: בהדגמת המחקר, התקפה מבוססת־אופטימיזציה גרמה לשינוי במינוני הוורפרין אצל 75 אחוזים מהמטופלים בשיעור ממוצע של 93.49 אחוזים, כאשר 10 אחוזים חוו שינויים של 358.89 אחוזים. מקרה בוחן מחקרי זה הראה כיצד הרעלת נתונים עלולה באופן תאורטי להפוך מערכות בינה מלאכותית בתחום הבריאות לנשק, וכוללת השלכות שעלולות להיות הרות אסון לבטיחות המטופלים משום שלוורפרין יש חלון טיפולי צר, שבו מינון שגוי עלול לגרום לדימום קטלני או לפקקת.

119 מקרה הבוחן מבוסס על המקורות האלה: Jagielski et al., “Manipulating Machine Learning.”; Müller et al., “Data Poisoning Attacks on Regression Learning.”

חלק 3

ארכיטקטורות גילוי והגנה

3

החלק הקודם הראה שהרעלת נתונים אינה עוד נקודת תורפה תאורטית אלא איום מבצעי היוצא לפועל באמצעות אקוסיסטם מגוון של שחקנים. ההתקפות שלהם מנצלות חולשות מבניות בתהליכים מודרניים של בינה מלאכותית, משיגות השפעה שאינה פרופורציונלית במאמץ מינימלי, ולעיתים קרובות נותרות בלתי נראות אפילו לגורמי הגנה בעלי משאבים. החלק השלישי יעבור מהעיסוק בשאלות כיצד ועל ידי מי נגרמת ההרעלה לשאלה הדחופה יותר עבור כל מי שמעורב בתחום הביטחון הלאומי: כיצד אפשר לגלות, להגביל או למזער ניסיונות הרעלה? מכיוון שההגנות הקיימות כיום נמצאות בפיגור אחר יכולות התוקפים, פרק זה ממפה את נוף כלי הזיהוי ואת ארכיטקטורות ההגנה לאורך כל מחזור החיים של למידת המכונה, ומתווה את הצעדים המעשיים הדרושים לבניית מערכות בינה מלאכותית עמידות, שכוללות יכולות לזהות שיבושים ומאובטחות בתפעול שלהן.

חשוב לציין שלרבים מאמצעי הנגד, שנמצאים בשימוש כיום נגד הרעלת נתונים מונעת־בינה מלאכותית, יש תקדימים היסטוריים. טכניקות כגון הערכת מהימנות המקורות, מעקב אחר מקור הנתונים, ניתוח של היסטוריית העריכות, זיהוי חריגות ופיקוח קהילתי פותחו במקור כדי להתמודד עם מניפולציות במערכות מידע המיועדות לבני אדם, כגון אנציקלופדיות, מנועי חיפוש ופלטפורמות חברתיות. בהקשרים של בינה מלאכותית, הגישות האלה מופיעות מחדש בגרסאות פורמליות יותר, ובהן תהליכי סינון של מערכי נתונים, ביקורת למודלים, דרישות למעקב אחר מקור נתוני האימון ובדיקות red-teaming לקלטים. הבנת רציפות זו מבהירה כי הגנה יעילה מפני הרעלת נתונים מחייבת לא רק אמצעים טכניים חדשים, אלא גם הטמעה שיטתית של פרקטיקות ניהול מידע ותיקות והתאמתן לסביבות מתווכות באמצעות מכונה.

3.0
זיהוי נתונים ומודלים מורעלים

זיהוי הרעלת נתונים הוא תחום מפוצל, המשקף דינמיקת תגובה של מרדף "חתול ועכבר" מתמשך בין גורמי הגנה לתוקפים. קטגוריית הזיהוי הראשונה, זיהוי ברמת הנתונים (Data-Level Detection), כוללת שיטות סטטיסטיות המיושמות לפני האימון. השיטות האלה מניחות שנוכחותה של הרעלה היא אירוע חריג. הקטגוריה השנייה, זיהוי ברמת המודל (Model-Level Detection), צמחה לאחר שתוקפים פיתחו התקפות מסוג clean-label המשתלבות בנתונים הרגילים ואינן בולטות בבדיקות סטטיסטיות, וכך עוקפות את מנגנוני הטיהור לפני האימון. בעקבות זאת נאלצו המגינים לעבור לביקורת פורנזית לאחר האימון כדי לאתר לוגיקה זדונית המוטמעת במודל עצמו.[120]

3.0.1 זיהוי ברמת הנתונים: טיהור לפני האימון

לפני תחילת האימון גורמי ההגנה מנסים לסנן דוגמאות חשודות באמצעות טכניקות סטטיסטיות וטכניקות לגילוי חריגות. אלו כוללות קיבוץ דגימות (clustering) כדי לאתר נקודות החורגות מהתפלגויות צפופות, בדיקות מבוססות־מרחק כגון k-Nearest Neighbors לזיהוי סתירות בתיוג, שיטות של דירוג נמוך (כגון SVD), החושפות דגימות שאינן הולמות לרכיבים העיקריים, וכן מודלים נלמדים לזיהוי חריגות שאומנו לאתר סטיות מדפוסי נתונים "רגילים".[121]

120 Keyizhi Xu et al., "A Survey of Adversarial Examples in Computer Vision: Attack, Defense, and Beyond," *Wuhan University Journal of Natural Sciences* 30, no. 1 (2025): 1-20, https://doi.org/10.1051/wujns/2025301001.

121 Giovanni Apruzzese et al., "Addressing Adversarial Attacks Against Security Systems Based on Machine Learning," *2019 11th International Conference on Cyber Conflict (CyCon)* (2019), 1–18, https://doi.org/10.23919/CYCON.2019.8756865; Wei Guo et al., "An Overview of Backdoor Attacks Against Deep Neural Networks and Possible Defences," preprint, arXiv, November 16, 2021, https://doi.org/10.48550/arXiv.2111.08429; Steinhardt et al., "Certified Defenses."; Vassilev, *Adversarial Machine Learning*; Wang et al., "Threats to Training," 1-36; Zhang et al., "Persistent Pre-Training Poisoning."

אלו גישות זולות, מהירות ואפקטיביות נגד התקפות גולמיות כמו היפוך תוויות פשוט. הן נמצאות בשימוש רחב במערכי נתונים רועשים או בכאלו המבוססים על מיקור המונים, בתהליכי למידת מכונה (ML) קלאסיים, בסינון ספאם, בזיהוי נוזקות (malware) ובמערכות בקרת איכות בתעשייה.

עם זאת התקפות הרעלה מודרניות — ובמיוחד כאלה מסוג clean-label — מתוכננות בכוונה להיראות נורמליות מבחינה סטטיסטית. תוקפים יכולים לעצב דגימות מורעלות המשתלבות באופן מושלם בפיזור הנתונים או ליצור אשכולות זעירים ואמינים למראית עין המדמים תת־אוכלוסיות לגיטימיות ושוליות. סינון אגרסיבי עלול להסיר גם דגימות נדירות אך תקפות, ובכך לפגוע בביצועים. לפיכך טיהור נתונים לבדו אינו מסוגל לאתר הרעלות מתוחכמות ובעלות היקף מצומצם, במיוחד בסביבות אימון רחבות היקף או בסביבות של למידה מבוזרת (FL).

3.0.2 זיהוי ברמת המודל: ביקורת לאחר האימון

מאחר שהרעלות מתקדמות אינן משאירות על פי רוב כל סימן נראה לעין בנתוני האימון, גורמי ההגנה נדרשים לבחון את המודל עצמו — גם במסגרת תפיסת אבטחה הממוקדת בנתונים. ביקורת ברמת המודל אינה סותרת גישה הממוקדת בנתונים; להפך, לעיתים זהו המרחב היחיד שבו ניתן עדיין להבחין בהשפעות העקיפות של נתונים מורעלים, במיוחד בתרחישים של שרשרת אספקה, שבהם אין גישה למערכי הנתונים המקוריים.

אחת השיטות לכך היא היפוך טריגר (trigger inversion) — כפי שמיושמת ב־Neural Cleanse — המבקשת להנדס לאחור טריגרים אפשריים באמצעות חיפוש אחר שינויים מזעריים בקלט שדוחפים את המודל בעקביות לתווית יעד מסוימת. זיהוי דפוס כזה הוא אינדיקציה לקיומה של דלת אחורית. היפוך טריגר הוא תהליך שאינו תלוי בנתונים עצמם ולעיתים עשוי להיות אפקטיבי, אך הוא עתיר חישובים ונשען על הנחות שבריריות (כגון קיומם של טריגרים קטנים ויציבים), ולכן מהימנותו מוגבלת במודלים מורכבים או במודלים המשלבים מספר אופני ייצוג.[122]

122 Jin et al., “Survey of Trojan Attacks.”; Naoto Kiribuchi et al., “Securing AI Systems: A Guide to Known Attacks and Impacts,” preprint, arXiv, June 29, 2025, https://doi.org/10.48550/arXiv.2506.23296; Pang Wei Koh et al., “Stronger Data Poisoning Attacks Break Data Sanitization Defenses,” Version 2, Preprint, arXiv, 2018, https://doi.org/10.48550/arXiv.1811.00741; Ramirez et al., “Poisoning Attacks and Defenses.”

משפחה נוספת של שיטות מתמקדת בניתוח ההפעלות הפנימיות של המודל. התקפות דלת אחורית נוטות להשתלט על נוירונים "רדומים", המופעלים רק בנוכחות טריגר מסוים. באמצעות בחינת דפוסי ההתנהגות של נוירונים עם קלטים נקיים, גורמי ההגנה יכולים לאתר תת־רשתות חשודות. כלים כגון STRIP פועלים על ידי שיבוש יזום של הקלטים שנכנסים ובדיקה האם תחזיות המודל נותרות יציבות באופן חריג — אינדיקציה לקיומה של לוגיקת טריגר סמויה. השיטות האלה קלות יחסית ליישום, אינן דורשות גישה לנתוני אימון, ונפוצות במיוחד בהבטחת שרשרת אספקה, שבה ארגונים נדרשים לאמת מודלים של צד שלישי לפני פריסה.[123]

עם זאת ברוב המקרים היתרון נותר בידי התוקפים גם בשכבה זו. גישות clean-label ותקיפות נטולות טריגר אינן יוצרות חתימה ספקטרלית מובחנת. טריגרים מבוזרים מערערים מנגנוני הגנה המבוססים על קיבוץ הפעלות. הערכות אמפיריות חוזרות ומראות כי בשיעורי הרעלה מציאותיים הנמוכים מאחוז אחד, רבים מהגלאים המתקדמים ביותר מציגים ביצועים הגבוהים אך במעט מניחוש אקראי. לפיכך זיהוי ברמת המודל הוא רכיב הכרחי, אך הוא משמש כלי פורנזי ולא ערובה להגנה.

123 Jin et al., "Survey of Trojan Attacks."; Koh et al., "Stronger Data Poisoning Attacks."; Li et al., "Backdoor Learning."

3.1
אסטרטגיות הגנה מפני הרעלת נתונים

בעוד זיהוי נועד לחשוף הרעלה לאחר שכבר התרחשה, ההגנה מתמקדת במניעה, בעמידות ובהפחתת השפעותיה. מאחר שהתקפות הרעלה מכוונות אל תשתית הנתונים של מערכות למידת מכונה, עמדת הגנה חסינה חייבת להקיף את מלוא מחזור החיים של המערכת — משלב איסוף הנתונים ועד לפריסת המודל. גם בדוח המתמקד במניפולציות ברמת הנתונים, יש תפקיד חיוני להגנות אלגוריתמיות ולהגנות ברמת המודל: ברגע שנתונים מורעלים נטמעים בתהליך האימון, השפעתם משתלבת בפרמטרים של המודל, כך שתיקון הנזק אפשרי לרוב רק במהלך האימון או לאחריו.

שכבת הגנה אחת לעולם אינה מספקת במערכות ביטחון לאומי. הפרקטיקה המודרנית נשענת כיום על ארכיטקטורת הגנה לעומק, המורכבת משלוש שכבות משלימות שמחזקות זו את זו:

1. הגנות ממוקדות־נתונים (מניעה של חדירת נתונים מורעלים לתהליך העיבוד)
2. הגנות ממוקדות־אלגוריתם (צמצום רגישות המודל לדגימות משובשות במהלך האימון)
3. הגנות ממוקדות־מודל (תיקון או מיתון פגיעות שאותרו לאחר האימון)

השכבות האלה יחד יוצרות מסגרת תפעול קוהרנטית לאבטחת תהליכי בינה מלאכותית מפני מניפולציה מכוונת.

3.1.1 הגנות ממוקדות־נתונים (מניעה לפני האימון)

ההגנות האלה מתמקדות באבטחה ובאימות של נתונים לפני תחילת האימון — נקודת ההתערבות היעילה ביותר למניעת חדירת הרעלה למערכת.

מעקב אחר מקור הנתונים ושרשרת הנתונים — מערכות מקור הנתונים מתעדות מהיכן הגיעו הנתונים, כיצד עובדו ומי טיפל בהם. באמצעות בניית גרף שרשרת הנתונים שניתן לאימות, המקשר בין מקורות הנתונים לשלבי העיבוד ולפלטי המודל — המעקב אחר מקור הנתונים מאפשר פורנזיקה לאחר האירוע, תומך בחלוקה ונשיאה באחריות ומרתיע גורמי

פנים מלנסות לבצע מניפולציות. האתגר המרכזי טמון בהיקף: ניטור צינורות עיבוד מבוזרים ובעלי נפח גבוה מייצר כמויות עצומות של מטא־נתונים ודורש תשתיות ייעודיות. למרות זאת המעקב אחר מקור הנתונים הופך יותר ויותר למרכיב בסיסי בתחומי הבריאות, הפיננסים, הסוכנויות הפדרליות ובמגזרים נוספים שבהם שלמות הנתונים היא קריטית למשימה.[124]

אימות קריפטוגרפי ורשומות שאינן ניתנות לשינוי — חתימות קריפטוגרפיות ואימות המעוגן בחומרה קושרים את הנתונים כבר בנקודת האיסוף, ומספקים עקבות המאפשרים לזהות את השיבוש. רשומות שאינן ניתנות לשינוי (למשל רישום מבוסס־בלוקצ'יין) מרחיבות זאת באמצעות תיעוד פעולות על נתונים כפעולות מצטברות בלבד, ובכך הן מספקות ערבויות חזקות לאותנטיות לאורך כל התהליך. הכלים האלה מעניקים הבטחות דטרמיניסטיות לשלמות הנתונים, מהסוג שהגנות סטטיסטיות אינן מסוגלות לספק. עם זאת הם מניחים שמקורות הנתונים מהימנים, ומתקשים להתמודד עם מערכות זמן אמת או בעלות נפח פעילות גבוה, ועלולים להתנגש עם דרישות מחיקה הנובעות מרגולציה בתחום הפרטיות. לפיכך יישומם מתאים בעיקר לסביבות מפוקחות או בעלות רמת אמינות גבוהה, כגון תשתיות חכמות, מערכות אוטונומיות ושרשראות עיבוד מדיה מאומתות.[125]

3.1.2 הגנות ממוקדות־אלגוריתם (חוסן במהלך האימון)

מרגע שנתונים מורעלים חומקים ממנגנוני הסינון המוקדם, הגנות אלגוריתמיות מבקשות לחזק את תהליך האימון עצמו כך שיהיה עמיד לדגימות משובשות. שכבה זו חיונית במיוחד במגזרים קריטיים למשימה, שבהם אימון מחדש מאפס אינו מעשי או בלתי אפשרי.

אופטימיזציה חסינה והגנות מאומתות — השיטות האלה משנות את יעד האימון כך שאף נקודת נתונים יחידה אינה יכולה להשפיע באופן לא פרופורציונלי על תהליך הלמידה. הגנות מאומתות מספקות ערבויות פורמליות לכך שהתנהגות המודל נותרת יציבה גם כאשר חלק קטן מנתוני האימון הוא זדוני. אף שהן מותאמות היטב למערכות רגישות לבטיחות, השיטות

124 Shay Hershkovitz and Corinna Turbes, *The Imperative of Data Provenance in AI* (Data Foundation, 2025), https://tinyurl.com/mr3344us; Vassilev, *Adversarial Machine Learning*; Jie Xu et al., "Machine Unlearning: Solutions and Challenges," *IEEE Transactions on Emerging Topics in Computational Intelligence* 8, no. 3 (2024): 2150-68, https://doi.org/10.1109/TETCI.2024.3379240.

125 Carlini et al., "Poisoning Web-Scale Training Datasets."; Vassilev, *Adversarial Machine Learning.*

האלה עתירות חישוב, קשות להרחבה לרשתות עמוקות מודרניות, ובדרך כלל מגינות רק מפני סוגי תקיפה מצומצמים, כגון היפוך תוויות, אך לא מפני התקפות clean-label חמקניות. עם זאת הן מציבות רף מחמיר של ודאות בפריסות בעלות אבטחה גבוהה.[126]

שיטות אנסמבל (Bagging, Aggregation, Majority Voting) — הגנות מבוססות־אנסמבל מאמנות מספר מודלים על תתי־קבוצות שונות של הנתונים ומאחדות את תחזיותיהם. כאשר אחד המודלים נפגע, מודלים אחרים יכולים לגבור עליו בהצבעת רוב, ובכך לדלל את השפעתן של דגימות מורעלות. אנסמבלים מספקים חוסן מאומת שניתן לכוונון, ונעשה בהם שימוש נרחב בזיהוי הונאה, בניטור חדירות ובסביבות נוספות שבהן רמת הסיכון גבוהה. עם זאת חולשותיהם כוללות תקורה חישובית ורגישות להרעלות הניתנות להעברה — התקפות שמתוכננות להטעות בו־זמנית מספר של ארכיטקטורות מודל שונות.[127]

אימון לחוסן מול התקפות הרעלה (**Adversarial Training**) — אימון כזה חושף את המודלים במהלך האימון לדגימות הרעלה סינתטיות, ובכך מאלץ אותם לזהות קלטים זדוניים ולפתח עמידות בפניהם. גישה זו משפרת את החוסן האמפירי מול מגוון סוגי תקיפה, ובהם גם התקפות clean-label חמקניות. עם זאת המחירים אינם מבוטלים: אימון כזה מגדיל את עלויות החישוב, פוגע בדיוק עם נתונים נקיים, ומכליל באופן מוגבל בלבד לגבי משפחות תקיפה חדשות. למרות המגבלות האלה הוא נותר רכיב חיוני בתחומים שבהם החוסן חשוב מהביצועים, כגון הגנת סייבר, מערכות אוטונומיות ואבחון רפואי.[128]

פונקציות הפסד חסינות — החלפת פונקציות הפסד סטנדרטיות בחלופות חסינות, כגון Huber loss, מפחיתה את השפעתן של נקודות קיצון או דגימות חריגות. מדובר בגישה חסכונית וקלה להטמעה, שהופכת אותה לאטרקטיבית במערכות הנדרשות להתמודד עם נתונים רועשים. עם זאת היא נשענת על ההנחה שדגימות מורעלות מתנהגות כחריגות סטטיסטיות — הנחה שהתקפות clean-label מפירות במכוון. בהקשרים של ביטחון לאומי,

126 Biggio and Roli, “Wild Patterns.”; Steinhardt et al., “Certified Defenses.”; Wang et al., “Threats to Training,” 1-36.

127 Biggio and Roli, “Wild Patterns.”; Wang et al., “Threats to Training,” 1-36.

128 Department of Homeland Security, *Risks and Mitigation Strategies*; Vassilev, *Adversarial Machine Learning*; Zhou et al., “Survey on Backdoor Threats.”

פונקציות הפסד חסינות פועלות בצורה מיטבית כתוספת קלה במסגרת אסטרטגיה רב־שכבתית ולא כהגנה עצמאית.[129]

3.1.3 הגנות ממוקדות־מודל (שיקום לאחר האימון)

לאחר זיהוי התקפת הרעלה ייתכן שגורמי ההגנה יידרשו לשקם את המודל מבלי לאמן אותו מחדש מאפס. אף שהשיטות האלה עוסקות בהתנהגות המודל, הן נותרות רלוונטיות גם לתפיסה ממוקדת־נתונים, שכן הן מטפלות בהשפעות של נתוני אימון מורעלים במורד הזרם.

גיזום מודל (Model Pruning) — הגיזום מסיר נוירונים או תת־רשתות החשודים בקידוד התנהגות של דלת אחורית. מדובר בכלי מהיר ומעשי, אך יש להפעילו בזהירות, שכן גיזום יתר עלול לפגוע בדיוק עם קלטים נקיים. יעילותו גבוהה במיוחד כאשר הטריגר מפעיל דפוסי חישוב מובחנים בשלב ההסקה.

ביטול למידת מכונה (Machine Unlearning) — ביטול למידת מכונה שואף למחוק את השפעתן של דגימות מורעלות מסוימות מבלי לבצע אימון מלא מחדש. הדבר ניתן לביצוע באופן מדויק באמצעות אימון חוזר על חלקים נקיים של הנתונים או בקירוב באמצעות שיטות להתאמת הגרדיאנט. אף שהגישה מבטיחה מבחינה תאורטית, הפתרונות הקיימים מתקשים לעיתים להסיר לחלוטין השפעות הרעלה חבויות, במיוחד במודלים עמוקים או עתירי ממדים. כיום ביטול למידה וגיזום משמשים בעיקר כלי תגובה לאירועים ולא צעדי מניעה.[130]

129 Jagielski et al., “Manipulating Machine Learning.”; Tian et al., “Comprehensive Survey,” 1-35; Vassilev, *Adversarial Machine Learning*.

130 Frank Hartle III et al., “Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges,” *Issues in Information Systems* 25, no. 4 (2025): 433-42; Jiao et al., “Can We Trust Embodied Agents?.”; Jin et al., “Survey of Trojan Attacks.”; Li et al., “Backdoor Learning.”; Souly et al., “Poisoning Attacks on LLMs.”; Vassilev, *Adversarial Machine Learning*; Zhang et al., “SoK.”

3.2
ניתוח השוואתי וסינתזה של הגנה רב־שכבתית

אין מנגנון יחיד שיכול לאבטח מערכת למידת מכונה מפני הרעלת נתונים. מאחר שהתקפות מודרניות מנצלות חולשות לאורך שרשרת העיבוד כולה, הגנה אפקטיבית מחייבת ארכיטקטורת הגנה לעומק ולא אמצעי יחיד. משמעות הדבר בפועל היא התייחסות להרעלה כסיכון תפעולי מתמשך ולא כבעיה טכנית חד־פעמית.

שכבה 1 — מניעה (ממוקדת־נתונים) — נקודת ההתערבות היעילה ביותר היא מניעת כניסתם של נתונים זדוניים למערכת מלכתחילה. תהליכי טיפול מאובטחים בנתונים, מעקב אחר מקור ושרשרת הנתונים ואימות קריפטוגרפי מסייעים לביסוס שרשרת אספקת נתונים מהימנה, ובכך מצמצמים את ההזדמנויות של התוקפים להחדיר דגימות מורעלות בהיקף רחב.

שכבה 2 — חוסן (ממוקד־אלגוריתם) — מאחר שחלק מהזיהום יחמוק בהכרח מבדיקות לפני האימון, תהליך האימון עצמו חייב להיות חסין לדגימות משובשות. למידת אנסמבל, אופטימיזציה חסינה ואימון לחוסן מול התקפות הרעלה מפחיתים את רגישות המודלים לקלטים מורעלים — גם כאשר חלק ממערך הנתונים נפגע — ומקבלים באופן מודע עלות חישובית גבוהה יותר בתמורה ליתרונות אבטחה מדידים.

שכבה 3 — זיהוי (ממוקד־נתונים וממוקד־מודל) — ניטור רציף לוכד את מה שמנגנוני המניעה והחוסן מחמיצים. כלי טיהור נתונים מסננים קלטים חשודים בזמן הקליטה, בעוד ביקורת לאחר האימון — כגון היפוך טריגרים וניתוח הפעלות של נוירונים — מסייעת לזהות דלתות אחוריות חבויות במודלים חדשים או קריטיים למשימה. נקודת מבט כפולה זו חיונית, משום שהרעלות מתוחכמות עשויות לא להשאיר עקבות הניתנים לזיהוי בנתונים הגולמיים, אלא להתגלות רק בהתנהגות המודל.

שכבה 4 — תיקון (ממוקד־מודל) — כאשר פשרה מאושרת, גורמי ההגנה נדרשים לכלי תגובה לאירועים. גיזום מודל וביטול למידת מכונה יכולים למתן חלק מהנזק הפוטנציאלי של הדלתות האחוריות או את שאריות השפעותיה של ההרעלה, אף שאימון מחדש מלא על

בסיס נתונים נקיים ומאומתים נותר הפתרון האמין ביותר. השיטות האלה מספקות בלימה, ולאו דווקא מניעה, ומאפשרות רציפות בתפעול בזמן שמפתחים פתרונות ארוכי טווח.

מעבר לשכבות הטכניות, ההנחיות העדכניות של המכון הלאומי לתקנים וטכנולוגיה (NIST) ומחלקת המלחמה של ארצות הברית (DoW) מדגישות פרקטיקות אימות מוסדיות, ובהן red-teaming של מערכות בינה מלאכותית, בדיקה קפדנית של שרשרת האספקה למודלים ומערכי נתונים חיצוניים ובדיקה עצמאית של מערכות קריטיות למשימה. מאחר שמודל ציבורי פגום יחיד עלול להפיץ שיבושים ליישומים רבים במורד הזרם ("הרעלת בארות"), יש להתייחס לאימות כאל אחריות מתמשכת למשימה. לבסוף, הגנות טכניות אינן יכולות להחליף ניהול של סיכוני פנים; גורמים בעלי הרשאות יתר מסוגלים לעקוף כל אמצעי הגנה. ארכיטקטורות של אפס־אמון (zero trust), ניטור התנהגותי ובקרות גישה מחמירות נותרים רכיבים חיוניים שמשלימים כל מסגרת של אבטחת בינה מלאכותית רב־שכבתית.

איור 7: הגנה לעומק מפני הרעלת נתונים

חלק 4

נקודות תורפה בביטחון הלאומי והשלכות אסטרטגיות

4

הרעלת נתונים אינה כשל טכני גרידא. מדובר באיום אסטרטגי הפוגע ביסודות הקוגניטיביים של מערכות בינה מלאכותית מודרניות. באמצעות השחתת הנתונים שעליהם מודלים לומדים, ההרעלה מעצבת מחדש את האופן שבו מוסדות תופסים את העולם, מקבלים החלטות, מקצים משאבים ומפרשים אותות. הסכנה טמונה לא רק בפלטים שגויים אלא בשחיקת האמון במערכות מבוססות־בינה מלאכותית בתחומי הביטחון, הממשל, התעשייה והחברה האזרחית. פרק זה ממפה את המגזרים בעלי החשיפה הגבוהה ביותר לביטחון הלאומי, ומראה כיצד הרעלת נתונים יוצרת הן נקודות תורפה מופשטות והן השלכות מוחשיות בתפעול.

4.0
האיום המופשט: השחתת הקוגניציה ושחיקת האמון

ברמה האסטרטגית, הרעלת נתונים מנצלת את עצם האמון כנשק — לא רק במובן הטכני של השחתת הקלטים של המודל, אלא במובן הקוגניטיבי העמוק יותר של עיצוב מחדש של האופן שבו בני האדם מבינים את המציאות. כאשר נתונים מורעלים חודרים למערכי אימון או למודלים בסיסיים, הם משנים את "הנחות הבסיס" הסטטיסטיות שמעצבות את התנהגות המודל.[131] אולם ההשפעה אינה נעצרת באלגוריתם; בני אדם מפנימים בהמשך את הפלטים המעוותים הללו, מחזקים אותם באמצעות השאילתות שלהם, ומשתתפים בלי לדעת בהעצמת הידע המורעל. היריבים מנצלים את הלולאה הדינמית הזו, והופכים אקוסיסטמים של נתונים פתוחים למנגנונים של זיהום עצמי.

תהליך כפול זה יוצר שתי חולשות השזורות זו בזו:

1. **שיקול דעת מבצעי מוטעה:** מערכות מורעלות מספקות הערכות משובשות, מסווגות איומים באופן שגוי, מעוותות תחזיות או מטות תוצרי מדיניות ומודיעין ברגעים מכריעים.[132]
2. **שחיקה קוגניטיבית ומוסדית:** עצם החשד לקיומה של הרעלה שוחק את הביטחון. מנהיגים, מצביאים, מפקדים ומנהלים בכירים מאיטים את קצב קבלת ההחלטות, פוסלים מערכות אוטומטיות או זונחים לחלוטין שימוש בכלי בינה מלאכותית, ובכך מוותרים על יתרון המהירות, ההיקף והיתרון האנליטי לטובת היריבים.[133]

בניגוד לפריצות סייבר מסורתיות, שבהן תוקפים חודרים למערכות כדי להוציא מידע או לשבש פעילות, הרעלת נתונים מתמקדת בתשתית האפיסטמית — המודלים המנטליים המשותפים שבהם מוסדות משתמשים כדי לפרש את העולם. הסיכון נובע לא רק מהזרקה חיצונית אלא גם מהפעולה האנושית של צריכה, מתן אמון והפיכה של ידע מורעל לאופרציונלי — בדיוק מה שהיריבים מנסים ליצור ולנצל למטרותיהם.

131 Rosiek, "AI Data Poisoning."; Travis Rosiek, "Data Poisoning Threatens AI's Promise in Government," *FedTech*, March 21, 2025, https://tinyurl.com/4ap7fwm3.

132 Biggio and Roli, "Wild Patterns."

133 Conti, "Data Poisoning as a Covert Weapon."; Department of Homeland Security, *Risks and Mitigation Strategies*; Erik Lin-Greenberg, "Allies and Artificial Intelligence: Obstacles to Operations and Decision-Making," preprint, Texas National Security Review, 2020, https://doi.org/10.26153/TSW/8866.

4.1
נקודות תורפה בשרשרת האספקה ובמודלי היסוד

מוסדות ביטחון לאומי מודרניים תלויים יותר ויותר במערכות בינה מלאכותית הבנויות על בסיס מאגרי נתונים ציבוריים משותפים, מודלי יסוד בקוד פתוח וכלי למידת מכונה (ML) של צד שלישי. תלות הדדית זו יוצרת סיכון לשרשרת האספקה ברמה הלאומית; פגיעה בכל נקודה במעלה הזרם עלולה להתפשט בחשאי אל מערכות ביטחון, מודיעין, תשתיות קריטיות וממשל המסתמכות על תוצרים אלו במורד הזרם.

שלושה עמודי תווך של שרשרת האספקה של הבינה המלאכותית — מערכי נתונים בהיקף רחב, מודלי יסוד מאומנים מראש וספריות למידת מכונה וכלי עיבוד נתונים — כולם רגישים להרעלה. התקפה מוצלחת כנגד כל אחד מאלו יוצרת פגיעה מסוג one-to-many, שבה משאב מורעל יחיד חודר לאלפי מערכות תלויות הנמצאות בשימוש בסוכנויות ממשל ורשויות מקומיות שונות. הרעלה של תוצרים במעלה הזרם היא כעת אחת הדרכים הריאליסטיות והמשפיעות ביותר עבור היריבים להשגת גישה ראשונית למערכות ביטחון לאומי במורד הזרם.[134]

הרלוונטיות לביטחון הלאומי מתבהרת כאשר מבינים כיצד פיתוח בינה מלאכותית ברמה המוסדית מתבצע בפועל:

- **סוכנויות ביטחון ומודיעין מבצעות לעיתים קרובות כוונון עדין של מודלי יסוד ציבוריים**. משמעות הדבר היא שמודל שהורעל במעלה הזרם הופך לחלק מתהליכי עבודה אנליטיים מסווגים.
- **מפעילי תשתיות קריטיות עושים שימוש חוזר במודלים שונים דוגמת ראייה ממוחשבת (CV) בקוד פתוח**. משמעות הדבר היא שמודל שהורעל עלול להשפיע בו־זמנית על תחומים כגון אנרגיה, תחבורה וייצור.

134 "LLM03: 2025 Supply Chain," OWASP, 2025; "ATLAS Matrix," MITRE ATLAS, 2025, https://tinyurl.com/5x34zf5h.

- **סוכנויות ממשלתיות צורכות רכיבי עיבוד שפה טבעית (NLP) משותפים**. למשל, מודל שפה מורעל עלול לסווג באופן שגוי בקשות לוויזה, התראות חירום או אינדיקטורים להונאה במספר מחלקות.

הסיכון הזה מועצם על ידי העובדה שמוסדות ביטחון לאומי אינם פועלים בחלל ריק; אנליסטים, מהנדסים וקובעי מדיניות מסתמכים לעיתים קרובות על אותו אקוסיסטם של מידע ציבורי כמו הציבור הרחב. כאשר האקוסיסטם הזה מורעל — באמצעות ויקיפדיה, מקורות של חדשות, מערכי נתונים פתוחים או מרכזי מודלים בשימוש נרחב — הן האזרחים והן המוסדות קולטים ומחזקים את הידע הפגום, ויוצרים לולאת משוב שמחלחלת בסופו של דבר למערכות קבלת החלטות רשמיות.

משאבי יסוד מורעלים פועלים כמזהמים אסטרטגיים: מערכות בינה מלאכותית במגוון תחומים — למשל צינורות מודיעין, מעקב וסיור (ISR), חיישני אבטחת סייבר, כלי תגובה במצבי חירום ומערכות ניטור פיננסיות — יורשות בחשאי הטיות או טריגרים מוטמעים כתוצאה מפגיעה יחידה במעלה הזרם. במקביל, מכיוון שהמשאבים האלה מזינים גם את סביבת המידע הציבורית, אותם מזהמים מעצבים את הידע האזרחי, את הנרטיבים של המדיה ואת השיח המקוון. מצב זה יוצר מרחב אפיסטמי משותף ומזוהם שבו הן המוסדות והן הציבור מסיקים מסקנות מאותם יסודות נגועים.

LLMs מעצימים את המצב הזה עוד יותר. היות שהם מאומנים בתהליכים רב־שלביים (אימון מקדים, SFT, RLHF), כל שלב מספק נקודת כניסה אפשרית, והטמעות במורד הזרם מבצעות רק לעיתים נדירות ביקורת חוזרת או אימות מחדש של השרשרת כולה. מערכות RAG יוצרות חשיפה נוספת לסכנה: הרעלת מאגרי ידע חיצוניים מייצרת מידע מוטעה או דיסאינפורמציה שמאוחזרת במידת מהימנות גבוהה על ידי כלים קריטיים למשימה.[135] ההשפעה האסטרטגית חורגת מעבר לפגיעת סייבר מסורתית מכיוון ש־LLMs משתלבים

135 Carlini et al., “Poisoning Web-Scale Training Datasets.”; Center for Security and Emerging Technology and Andrew Lohn, *Poison in the Well: Securing the Shared Resources of Machine Learning* (Center for Security and Emerging Technology, 2021), https://doi.org/10.51593/2020CA013; Department of Homeland Security, *Risks and Mitigation Strategies*; Gu et al., “BadNets.”; Schwarzschild et al., “Just How Toxic Is Data Poisoning?.”; Vassilev, *Adversarial Machine Learning*; Wang et al., “Threats to Training,” 1-36; Xinyi Zheng et al., “Towards Robust Detection.”

כעת בתוך תהליכי עבודה רב־מגזריים. מניפולציה קטנה במעלה הזרם מתפשטת על פני המערכות המקושרות האלה, ומייצרת עיוותים מסונכרנים שאף מגזר יחיד אינו יכול לזהות בצורה מבודדת.

משמעות הדבר במונחי ביטחון לאומי היא שהיריבים אינם צריכים לפרוץ לרשת מסווגת כדי להשפיע על מודל מסווג. השחתת מערך נתונים או מודל ציבורי הנמצא בשימוש חוזר נרחב עשויה להספיק. לפיכך משטח התקיפה הוא גלובלי, מפוזר וקשה ביותר לניטור — נקודת תורפה מבנית שהיריבים יכולים לנצל כדי לחדור למערכות צבאיות, למערכות מודיעין ולמערכות אזרחיות בהיקף רחב.

4.2
מערכות צבא והגנה

שילוב הבינה המלאכותית ב־ISR, בניהול מטרות, בפיקוד ושליטה (C2), בתחזוקה חזויה, בלוגיסטיקה ובהגנת סייבר חושף את המגזר הצבאי באופן ייחודי להתקפות הרעלה, שלהן השלכות קינטיות, מבצעיות והשפעה ישירה על קצב קבלת ההחלטות.

ISR, זיהוי מטרות וחישה מבוססת־חלל — החדרת נתוני אימון מורעלים למודלים של זיהוי אובייקטים, ניתוח צילומי מכ״ם מפתח סינתטי (SAR) או סיווג אותות עלולים לגרום לאי־גילוי של כוחות אויב, לזיהוי שגוי של כוחות ידידותיים עד כדי פגיעה בהם ולסיווג שיטתי מוטעה של נכסים צבאיים.[136]

טריגרים פיזיים, כגון דפוסי הסוואה המותאמים להטעיית אלגוריתמים או סימנים חזותיים ייעודיים המוחדרים לשדה הראייה, הוכחו כבעלי יכולת עקבית להטעות מסווגים.[137] הדוגמה הקלאסית של ״פתקית Post-it על תמרור עצור״ ממחישה כיצד טריגרים פעוטים יכולים להפעיל הרעלה רדומה; חוקרים הציבו מדבקה קטנה על תמרור עצור וגרמו למודלים מתקדמים של ראייה ממוחשבת (CV) לסווג אותו בטעות באופן עקבי כתמרור הגבלת מהירות, ובכך הדגימו כיצד תבנית פיזית טריוויאלית יכולה להפעיל נקודת תורפה נסתרת במודלי תפיסה מורעלים או שבריריים.[138]

136 Conti, “Data Poisoning as a Covert Weapon.”; Department of Homeland Security, *Risks and Mitigation Strategies*; Farzad Kamrani et al., *Attacking and Deceiving Military AI Systems*, FOI-R--5396--SE (Swedish Defence Research Agency, 2023), https://tinyurl.com/3ddn3p7c; Jackson Barnett, “Army Looks to Block Data ‘Poisoning’ in Facial Recognition, AI,” *FEDSCOOP*, n.d., accessed November 7, 2025, https://tinyurl.com/595haz7h; Li Ang Zhang et al., *Operational Feasibility of Adversarial Attacks Against Artificial Intelligence*, Research Report (RAND, 2022), https://tinyurl.com/5hyp7fxk; Lin-Greenberg, “Allies and Artificial Intelligence.”

137 Vassilev, *Adversarial Machine Learning*; “ATLAS Matrix.”

138 Kevin Eykholt et al., “Robust Physical-World Attacks on Deep Learning Models,” preprint, arXiv, April 10, 2018, https://doi.org/10.48550/arXiv.1707.08945.

העצמה בזירת החלל — מערכי ISR מבוססי־חלל ניצבים בפני מכפיל איום ייחודי: אין דרך פשוטה לאמת את המציאות בשטח. אובייקטים שסווגו בצורה שגויה, פגיעה ביכולת לזהות שינויים או קטלוגים מסלוליים שעברו הרעלה אינם ניתנים לאימות באמצעות תצפית אנושית או מערך של חיישנים יתירים. מסווג לוויני מורעל עלול לפעול במשך חודשים ללא פיקוח ולעצב הערכות אסטרטגיות מבלי שהדבר יתגלה.[139]

בינה מלאכותית להגנת סייבר ומודיעין איומים — מודלים של למידת מכונה המשמשים לזיהוי נוזקות, לאיתור תעבורת רשת חריגה או לסיווג קמפיינים של דיוג (phishing) הפכו לרכיבים חיוניים בתשתיות ההגנה. הרעלה של מערכי הנתונים האלה או של הזנות מודיעין איומים במעלה הזרם עלולה ליצור אזורי עיוורון או לשבש זיהוי של חתימות קריטיות. אף שמסגרות סיווג נוטות לתאר זאת כבעיית שלמות או זמינות, מבחינת התפעול מדובר בהקבלה ישירה לפגיעה ב־ISR: הדבר מסמא או מטעה את גורמי ההגנה במהירות המכונה.[140]

פיקוד ושליטה, לוגיסטיקה והתחזוקה החזויה — מודלי חיזוי מורעלים עלולים להקצות אספקה באופן שגוי, לגרום לשגיאות במחזורי התחזוקה או לתעדף איומים בצורה לקויה. הכשלים האלה עדינים באופיים; הם נראים כ"סטייה טבעית" ולא כהתנהגות זדונית. דווקא עמימות זו מנוצלת בידי תוקפים — לא באמצעות כשל דרמטי, אלא באמצעות הצטברות איטית של חוסר יעילות מכוון. עם הזמן מודלים מורעלים לתמיכה בפיקוד ושליטה שוחקים את המוכנות, את הקצב ואת יתרון קבלת ההחלטות.[141]

---

139 Conti, "Data Poisoning as a Covert Weapon."; Department of Homeland Security, *Risks and Mitigation Strategies*; Puscas, *AI and International Security*; Müller et al., "Data Poisoning Attacks on Regression Learning."

140 Biggio and Roli, "Wild Patterns."; Muñoz-González et al., "Towards Poisoning."; Vassilev, *Adversarial Machine Learning*; Sridhar Venkatesan et al., "Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems," *MILCOM 2021 – 2021 IEEE Military Communications Conference* (*MILCOM*) (2021), 874–79, https://doi.org/10.1109/MILCOM52596.2021.9652916; Wang et al., "Threats to Training," 1-36; Zou et al., "PoisonedRAG."; Vassilev, *Adversarial Machine Learning*.

141 Conti, "Data Poisoning as a Covert Weapon."; Department of Homeland Security, *Risks and Mitigation Strategies*; Puscas, *AI and International Security*; Müller et al., "Data Poisoning Attacks on Regression Learning."

4.3
תשתיות לאומיות קריטיות ומערכות אזרחיות

יותר ויותר תשתיות לאומיות קריטיות מנוהלות, מותאמות בצורה מיטבית ומוגנות באמצעות מערכות למידת מכונה. רשתות חשמל מאזנות עומסים בעזרת חיזוי מבוסס־בינה מלאכותית; מערכי תחבורה נשענים על חישה וניתוב אוטומטיים; נמלים, מפעלים ומרכזי לוגיסטיקה תלויים במודלים של למידת מכונה לניהול תפוקות ולאיתור חריגות. המערכות האלה יוצרות את התשתית הפיזית של העוצמה הלאומית — מוכנות וכשירות צבאית, יציבות כלכלית והמשכיות חברתית הנשענות כולן על תפקודן הרציף.

הרעלת נתונים בזירה זו מסוכנת במיוחד משלושה טעמים:

1. בינה מלאכותית פועלת בהיקף ובמהירות — מודלים מורעלים עלולים לנהל באופן שגוי רשתות שלמות ולא רק רכיבים מבודדים.
2. הנראות האנושית מוגבלת — חריגות בתפעול נדמות לעיתים כשונות רגילה או כסטיות שמקורן ברכיבי הציוד, וכך הן מסתירות מניפולציה זדונית.
3. הזיקה בין הממד הדיגיטלי לפיזי הדוקה — שגיאות מתגלגלות בין מערכות מקושרות, ומעצימות השלכות ברמה הלאומית.

מערכות חשמל: רשת חשמל חכמה, חיזוי הייצור ואיזון עומסים — רשת החשמל המודרנית נשענת במידה רבה על למידת מכונה לצורך חיזוי ביקושים, תזמון הייצור ואיתור חריגות בזמן אמת. הרעלת המודלים האלה — אם באמצעות מניפולציה של נתונים היסטוריים ואם באמצעות החדרת נתוני חיישנים משובשים — עלולה לחולל אי־יציבות בכל רחבי המערכת. לדוגמה:

- היערכות חסרה לזינוקים בביקוש מותירה את רשתות אספקת החשמל במצב פגיע להפסקות חשמל, במיוחד במהלך גלי חום או אסונות.
- התחייבות יתר לרמות ייצור מערערת את יציבות התדר והמתח, ועלולה להוביל להשבתות אוטומטיות מטעמי בטיחות.

- דירוג שגוי של חריגות עלול לגרום למפעילים להישאר עיוורים לסימנים מוקדמים של כשל בציוד או חדירות סייבר.

תקנים כגון NIST AI 100-2 מסווגים את הכשלים האלה כפגיעות בשלמות ובזמינות, אך ההשלכות לביטחון הלאומי רחבות בהרבה: הפסקות חשמל פוגעות במערכות של בתי חולים, בבסיסים צבאיים, ברשתות חירום ובתשתיות של בטיחות ציבורית.[142]

מסוכן אף יותר הוא אפקט ההסוואה: שגיאות חיזוי הנובעות מהרעלה דומות לסטייה רגילה של מודל לאורך זמן, והן מעכבות את הזיהוי ומאפשרות הצטברות של כשלים הנראים כתקלות מקריות. בזמן משברים גאופוליטיים, היריב עשוי לנצל דינמיקה זו כדי לגרום להפסקות חשמל מתגלגלות דווקא ברגעים שבהם חוסן אנרגטי הוא קריטי במיוחד.[143]

מערכות תחבורה: כלי רכב אוטונומיים, לוגיסטיקה קרקעית ואינטגרציה של תעופה וחלל — מערכות בינה מלאכותית בתחום התחבורה משולבות באקוסיסטם אוטונומי שהולך ומתרחב — ציי רכבים אוטונומיים, מערכות רכבת מבוססות־בינה מלאכותית, מסדרונות של תעבורת רחפנים וכלי ניהול של תעבורה אווירית. הרעלת המודלים האלה עלולה לייצר כשלים ממוקדים או כשלים מערכתיים רחבי היקף.

רכבים אוטונומיים ואוטונומיים למחצה — ציים הלומדים באופן רציף מחיישנים פרוסים בשטח מצויים במצב חשוף במיוחד. "צי טרויאני" קטן של רכבים נגועים יכול להעלות נתוני טלמטריה מורעלים המטים את מודלי הניווט המרכזיים. הרעלה כזו עלולה לגרום לזיהוי שגוי ושיטתי של תמרורים, לכשלים באופטימיזציית הניתוב של המסלולים או לאיתור לא בטיחותי של מכשול בסביבות משותפות.

142 Vassilev, *Adversarial Machine Learning*.

143 Raphael I. Areola et al., "Artificial Intelligence for Optimizing Solar Power Systems with Integrated Storage: A Critical Review of Techniques, Challenges, and Emerging Trends," *Electricity* 6, no. 4 (2025): 60, https://doi.org/10.3390/electricity6040060; Committee on Using Machine Learning in Safety-Critical Applications: Setting a Research Agenda et al., *Machine Learning for Safety-Critical Applications: Opportunities, Challenges, and a Research Agenda* (National Academies Press, 2025), https://doi.org/10.17226/27970; Müller et al., "Data Poisoning Attacks on Regression Learning."; *Safety and Security Guidelines for Critical Infrastructure Owners and Operators* (Department of Homeland Security, 2024), https://tinyurl.com/2t8memjd; Yanxu Zhu et al., "Research on Data Poisoning Attack against Smart Grid Cyber–Physical System Based on Edge Computing," *Sensors* 23, no. 9 (2023): 4509, https://doi.org/10.3390/s23094509.

לפי MITRE התרחיש הזה מסווג כאיום של הרעלת נתונים בלמידה מבוזרת (FL), אך בהקשר של הביטחון הלאומי ההשלכות רחבות בהרבה: רכבי חירום, שיירות צבאיות ונתיבי אספקה תלויים במודלים האלה.[144]

מערכות תעופה ומרחב אווירי — רשתות בטיחות מבוססות־למידה עמוקה מנטרות זרמים של נתוני מעקב משודר תלוי אוטומטי (ADS-B), מזהות חריגות בנתיבי טיסה ומסווגות סיכוני תעופה. נתוני אימון מורעלים עלולים לעוור את הגלאים האלה וליצור תנאים המקילים על התקפות קינטיות או על התקפות סייבר מסורתיות.[145]

חישה מבוססת־חלל יוצרת שכבת פגיעות נוספת — מודלים מורעלים למעקב מסלולי או למודעות תחום החלל פועלים ללא נקודת ייחוס חיצונית לאימות. סיווג שגוי של לוויינים, של פסולת חלל או של שיגורי טילים אינו ניתן לתיקון מהיר, ומעצים את רמת הסיכון הן בתעופה האזרחית והן בפעילויות הקשורות לביטחון הלאומי.

לוגיסטיקה קרקעית וניהול תעבורה — מודלי אופטימיזציה מורעלים עלולים לשבש ניתוב משלוחים, לעוות תחזיות טיפול במטען או לבצע מניפולציה ברמזורים בהיקף נרחב — השפעות ההופכות עד מהרה לשרשרת עיכובים באספקה, להאטה בפריסת כוחות צבאיים או לפקק תנועה של רכבי חירום.

ייצור קריטי ושרשראות אספקה — מפעלי ייצור, נמלים, רובוטים תעשייתיים ורשתות לוגיסטיקה גלובליות מסתמכים על למידת מכונה כדי לשמור על התפוקה והבטיחות.

144 "ATLAS Matrix."

145 Yanjiao Chen et al., "Data Poisoning Attacks in Internet-of-Vehicle Networks: Taxonomy, State-of-the-Art, and Future Directions," *IEEE Transactions on Industrial Informatics* 19, no. 1 (2023): 20-28, https://doi.org/10.1109/TII.2022.3198481; Department of Homeland Security, *Risks and Mitigation Strategies*; Fendley et al., "Systematic Review."; Hartle III et al., "Data Poisoning 2018–2025," 433-42; Anastasios Giannaros et al., "Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions," *Journal of Cybersecurity and Privacy* 3, no. 3 (2023): 493-543, https://doi.org/10.3390/jcp3030025; Barnett, "Army Looks to Block Data 'Poisoning'."; Peng Luo et al., "ADS-Bpois: Poisoning Attacks Against Deep Learning-Based Air Traffic ADS-B Unsupervised Anomaly Detection Models," *IEEE Internet of Things Journal* 11, no. 23 (2024): 38301-11, https://doi.org/10.1109/JIOT.2024.3446675.

הרעלת נתונים בהקשרים האלה עלולה ליצור כשלי תפעול מתגלגלים הדומים במהותם לתופעות של חוסר יעילות רגיל:[146]

- מודלים מורעלים לאיתור פגמים גורמים לרכיבים פגומים לחדור לשרשראות אספקה של נשק או תעופה וחלל.
- מערכות תזמון משובשות יוצרות צווארי בקבוק בנמלים או במרכזי הפצה.
- מודלי רכש מורעלים מסתירים חבלות אצל ספקים במעלה הזרם.
- חריגות של חיישנים שסווגו באופן שגוי מסווות חבלות מכניות או חדירות סייבר.

ההשפעה על הביטחון הלאומי היא משמעותית: ייצור נשק, זמינות של חלקי חילוף ולוגיסטיקה תלויים כולם במערכות הייצור האלה המונעות על ידי למידת מכונה. מכיוון ששרשראות אספקה דיגיטליות של למידת מכונה ושרשראות לוגיסטיקה פיזיות שלובות זו בזו, הרעלה בשכבה הדיגיטלית עלולה לגרום לעיכובים, לכשלים או למחסורים פיזיים. בקמפיינים עוינים, הדבר יוצר וקטור תקיפה עוצמתי של "בעירה איטית": היריב יכול לשחוק את המוכנות, להחליש את היכולת התעשייתית או לשבש את תהליכי הגיוס מבלי לגרום לשום תקרית גלויה, על ידי הפיכת המערכות המייעלות את הלוגיסטיקה הלאומית לנשק הפועל נגדה.

מערכות בריאות, בריאות הציבור וביטחון ביולוגי — תשתיות הבריאות ובריאות הציבור מסתמכות יותר ויותר על למידת מכונה לצורך אבחון, מיון, גילוי התפרצויות, פענוח גנומי ותכנון התפעול. הרעלת מודלים אלו עלולה לגרום לנזק פיזי ישיר ולהשלכות רחבות יותר על הביטחון הלאומי. כפי שהודגם במקרה הרעלת הרגרסיה של ורפרין, אפילו מניפולציות קטנות יכולות לגרום לשינוי דרמטי של המלצות קליניות, לסכן מטופלים ולשחוק את האמון במערכות בינה מלאכותית רפואיות. כלי ניטור לבריאות הציבור — בפרט אלו הקולטים אותות מהמדיה החברתית או מהרשת — עלולים לשמש מטרה להרעלה כדי לגרום לאזעקות שווא ("זאב זאב") או להסתיר התפרצויות אמיתיות ("להחביא את הזאב"), ובכך ליצור חלונות הזדמנות אסטרטגיים במהלך משברים ביולוגיים. מערכות למידת מכונה

146 Conti, "Data Poisoning as a Covert Weapon."; Department of Homeland Security, *Risks and Mitigation Strategies*; Müller et al., "Data Poisoning Attacks on Regression Learning."; Ramirez et al., "Poisoning Attacks and Defenses."; Vagan Terziyan et al., "Industry 4.0 Intelligence Under Attack: From Cognitive Hack to Data Poisoning," in *NATO Science for Peace and Security Series – D: Information and Communication Security* (IOS Press, 2018), https://doi.org/10.3233/978-1-61499-888-4-110.

המשמשות לאוטומציה של מעבדות, לגילוי פתוגנים או לניתוח גנומי מרחיבות עוד יותר את משטח התקיפה; נתוני אימון מורעלים עלולים לטשטש חתימות של איומים מתהווים או לחלופין להעצים רעש תמים ולגרום לו להיראות משמעותי. מאחר שמערכות בריאות וביטחון ביולוגי משמשות הן כעורקי חיים מקומיים והן כנכסי הגנה לאומיים, הרעלה שלהן משמשת כגשר בין שיבוש במישור האזרחי לסיכון במישור של ביטחון הפנים.[147]

147 Sumit Singh Dhanda et al., "Advancement in Public Health through Machine Learning: A Narrative Review of Opportunities and Ethical Considerations," *Journal of Big Data* 12, no. 1 (2025): 154, https://doi.org/10.1186/s40537-025-01201-x; Hartle III et al., "Data Poisoning 2018–2025," 433-42; Jagielski et al., "Manipulating Machine Learning."; Vanessa I. S. Mendes et al., "Harnessing Artificial Intelligence for Enhanced Public Health Surveillance: A Narrative Review," *Frontiers in Public Health* 13 (2025), https://doi.org/10.3389/fpubh.2025.1601151; Heather Rilkoff et al., "Innovations in Public Health Surveillance: An Overview of Novel Use of Data and Analytic Methods," *Canada Communicable Disease Report* 50, no. 3/4 (2024): 93-101, https://doi.org/10.14745/ccdr.v50i34a02; Zhao et al., "Data Poisoning in Deep Learning."; Zou et al., "PoisonedRAG."

4.4
מערכות כלכליות וממשלתיות

מערכות כלכליות ושירותי ממשל דיגיטליים הם עמוד השדרה המוסדי של היציבות הלאומית. בניגוד למערכות צבאיות או מערכות אנרגיה, שבהן הרעלה גורמת להידרדרות מיידית בתפעול, התקפות על מערכות פיננסים וממשל נועדו לפגוע בהמשכיות, בלגיטימיות ובמבני האמון של המדינה עצמה. אלו הן המערכות ששומרות על התפקוד התקין של השווקים, הן מסדירות מתן הטבות לציבור, מסייעות באכיפת חוקים וקובעות לגבי זכויות האזרח. על כן הרעלה ברובד זה אינה רק כשל בשלמות, אלא וקטור אסטרטגי לערעור הביטחון הכלכלי והחוזה בין המדינה לחברה.

מערכות פיננסיות: יציבות השוק, גילוי הונאות ומניפולציה סמויה — המערכות הפיננסיות המודרניות מתבססות במידה רבה על אוטומציה. דירוגי אשראי, תהליכים לזיהוי הונאות, מערכות למניעת הלבנת הון (AML), מודלי סיכון ומנועי מסחר אלגוריתמיים מסתמכים כולם על למידת מכונה שאומנה עם מערכים עצומים של נתונים היסטוריים. הרעלה של קלטים אלו מחדירה עיוותים עדינים אך בעלי פוטנציאל לניצול כספי:

- **דירוג אשראי מוטה** היוצר העדפה או פוגע באופן שיטתי בקבוצות יעד
- **מסווגי הונאות** שמתייגים עסקאות שאינן חוקיות כתמימות
- **מודלי סיכון** המבצעים הערכת חסר או הערכת יתר של החשיפה במגזרי שוק ספציפיים
- **אלגוריתמים של מסחר** שמוסטים כדי לייצר דפוסים רווחיים עבור היריב

מערכות פיננסיות פועלות ברציפות ובמהירות, ומשום כך העיוותים האלה מצטברים באופן סמוי. מודל מורעל יכול להסיט הון בחשאי, לעוות מחירים של נכסים או ליצור הזדמנויות לעסקאות ארביטראז׳. בתרחישים אסטרטגיים יותר, היריבים יכולים להשתמש במניפולציה של מודלים ככלי ליישום של מדינאות כלכלית — ערעור יציבותם של מוסדות פיננסיים, שחיקת אמון המשקיעים או יצירת מצוקת נזילות ברגעים הרגישים מבחינה פוליטית.[148]

148 Hartle III et al., “Data Poisoning 2018–2025,” 433-42; Lumenova, “Data Poisoning Attacks.”; Müller et al., “Data Poisoning Attacks on Regression Learning.”; Wang et al., “Threats to Training,” 1-36; Zhao et al., “Data Poisoning in Deep Learning.”

הרעלה פיננסית מסוכנת במיוחד מכיוון שהיא אינה דומה לפריצה; היא דומה יותר לשינוי בתנאים הכלכליים. קשה יותר לייחס התקפות מסוג זה, ואותות רגולטוריים נראים ״תקינים״, בעוד הם מסווים השפעה עוינת עמוקה יותר.

שירותי ממשל דיגיטלי: יכולות מוסדיות, לגיטימציה ואמון — שירותים ממשלתיים מסתמכים יותר ויותר על רכיבי בינה מלאכותית משותפים לצורך קבלת הכרעות בענייני הגירה, גילוי הונאות, חלוקת הטבות, ניהול מצבי חירום וניתוח נתונים הקשורים לבטיחות הציבור. להרעלה של מודלים אלו יש השלכות מרחיקות לכת — הרבה מעבר לטעויות מנהליות. לדוגמה:

- מודלים לבדיקת בקשות לקבלת ויזות או לבדיקות רקע שמסווגים סיכונים באופן שגוי
- מערכות מיון למענה במצבי חירום שמבצעות תיעדוף באופן שגוי
- מודלים לעיבוד הטבות שדוחים מתן תמיכה או מנתבים אותה ליעדים שגויים

מכיוון שרבות מהמערכות הללו משתמשות ברכיבי עיבוד שפה טבעית (NLP), ראייה ממוחשבת (CV) ורכיבי חיזוי משותפים, מודל יסוד מורעל יחיד עלול ליצור אפקט דומינו מתגלגל בין הסוכנויות השונות — אירוע ארכיטיפי של ״הרעלת בארות״. תקנים אכן מסווגים כשלים כגון אלו כפגיעה בזמינות או בשלמות המידע, אך מסגור טכני זה ממעיט בערכה של ההשפעה האסטרטגית; הרעלה של מערכות ממשל מערערת את אמון הציבור ביכולתה של המדינה, שוחקת את הלגיטימיות ומחלישה את האופן שבו אמינות הממשל נתפסת.[149]

לא מדובר בהשפעה משנית — זוהי בעצם המטרה המבצעית של ההתקפות על המרחב הקוגניטיבי. כאשר שירותים דיגיטליים הופכים לבלתי אמינים, האזרחים מטילים את האשמה על המוסדות ולא על האלגוריתמים הנמצאים בבסיס השירותים. היריבים מנצלים דינמיקה זו כדי לשחוק את הלכידות החברתית, ליצור חיכוך אדמיניסטרטיבי ולעורר תסכול בקרב הציבור ברגעי משבר או מתח פוליטי.

149 Department of Homeland Security, *Risks and Mitigation Strategies*; Stephan III, “Big Data.”; Ramirez et al., “Poisoning Attacks and Defenses.”; Rosiek, “AI Data Poisoning.”

4.5
הרעלות רדומות והפעלה מבוססת־זמן

סכנה מכרעת של הרעלת נתונים, שלעיתים קרובות אינה זוכה למידת ההערכה הנדרשת, היא יכולתה להישאר רדומה במשך זמן רב. בניגוד לחדירות סייבר מסורתיות שלרוב מיועדות לאפשר ניצול מיידי, מודלים מורעלים יכולים לפעול בחשאי וביעילות במשך חודשים או שנים לפני שההתנהגות הזדונית שלהם מתגלה מעל פני השטח. באופן קריטי, ההתקפות האלה אינן מסתמכות על טריגרים מבוססי־זמן במובן המילולי. במקום זאת הן נשארות רדומות עד להופעת תנאי מבצעי ספציפי, דפוס של קלט או רמז סביבתי.[150] טריגרים אלו עשויים ללבוש צורה של דפוס חזותי על כלי רכב של היריב, סוג מסוים של סביבת חיישנים שמופיעה רק במהלך גיוס כוחות, הנחיה ספציפית למשימה במערכת פיקוד או חתימה של הפרעת תדר רדיו שמופיעה אך ורק בתנאי עימות. מכיוון שטריגרים אלו אינם מתגלים לעולם בעיתות שלום, הלוגיקה המורעלת נשארת סמויה לחלוטין במהלך בדיקות שגרתיות.

הפעלה מותנית זו הופכת מודלי בינה מלאכותית מורעלים לנכסים שהיריבים יכולים להציב מבעוד מועד — צורה של הכנה אסטרטגית של מרחב הלחימה. בשיטה של זריעה והמתנה, התוקפים מנצלים את המבנה עצמו של מחזורי חיי הבינה המלאכותית: המודלים עוברים שימוש חוזר, מתעדכנים, משותפים בין יחידות לסוכנויות שונות ומוטמעים עמוק באקוסיסטם לתמיכה בקבלת החלטות. פגיעה מקומית עלולה להפוך במרוצת הזמן לפגיעות מערכתית, שמופצת על פני מספר פיקודים, סוכנויות אזרחיות או מפעילי תשתיות, שמסתמכים בלא יודעין על אותו מודל או מערך נתונים מזוהם.[151]

ההשלכות על הביטחון הלאומי הן עמוקות. הרעלה רדומה מערערת את יכולת ההרתעה משום שהיא נותרת סמויה עד לרגע ההפעלה; גורמי ההגנה אינם יכולים לעורר מודעות או

150 Fendley et al., “Systematic Review.”; Li et al., “Backdoor Learning.”; Pawlicki et al., “Meta-Survey.”; Schwarzschild et al., “Just How Toxic is Data Poisoning?.”

151 Thibaud Gloaguen et al., “Watch Your Steps: Dormant Adversarial Behaviors That Activate upon LLM Finetuning,” preprint, arXiv, October 9, 2025, https://doi.org/10.48550/arXiv.2505.16567; Qingyue Wang et al., “BadMoE: Backdooring Mixture-of-Experts LLMs via Optimizing Routing Triggers and Infecting Dormant Experts,” version 2, preprint, arXiv, 2025, https://doi.org/10.48550/ARXIV.2504.18598.

לגבות מחיר כאשר הם אינם מזהים את הפגיעה עצמה. היא שוחקת את המוכנות בכך שהיא מאפשרת לפלטים מוטים, פגומים או מעוותים אסטרטגית להשפיע על התכנון וההערכות זמן רב לפני כל כשל גלוי. יתרה מכך, טריגרים מותנים בעת משבר עלולים לפעול בדיוק בעיתות שבהן ההסתמכות על מערכות ISR, פיקוד ושליטה, לוגיסטיקה או תגובה ציבוריות מבוססות־בינה מלאכותית נמצאת בשיאה. הכשלים שנובעים מכך נראים כסחף רגיל של מודלים או כשגיאה בלתי צפויה, ובכך מעכבים ייחוס מהיר של ההתקפה ומעניקים ליריב יכולת סבירה להכחשה.[152]

ברמה הלאומית, הרעלה רדומה מאיימת על הקצב, הביטחון והאמינות של קבלת ההחלטות. מפקדים עלולים שלא בידיעתם לבסס את שיקול דעתם על פלטים שמשובשים באופן מזערי; רשויות אזרחיות עלולות להקצות משאבים באופן שגוי; מערכות פיננסיות ומערכות ממשל עלולות להציג חריגות בתקופות לחץ. ההשפעות האלה מצטברות בחשאי, ומחלישות את יסודות המידע והיסודות המבצעיים של המדינאות. במובן זה הרעלה רדומה אינה רק פגיעות טכנית — היא כלי אסטרטגי שמאפשר ליריבים לעצב את תנאי העימות זמן רב לפני תחילת מעשי האיבה.

איור 8: הרעלת נתונים — נוף האיומים במישור הביטחון הלאומי

152 Conti, “Data Poisoning as a Covert Weapon.”; Department of Homeland Security, *Risks and Mitigation Strategies*; Jie Guo, “The Ethical Legitimacy of Autonomous Weapons Systems: Reconfiguring War Accountability in the Age of Artificial Intelligence,” *Ethics & Global Politics* 18, no. 3 (2025): 27-39, https://doi.org/10.1080/16544951.2025.2540131.

חלק 5

השלכות וכיוונים עתידיים

5

5.0
תובנות למקבלי ההחלטות

הפרקים הקודמים מבהירים כי הרעלת נתונים אינה רק אנומליה טכנית במערכות של למידת מכונה (ML), אלא נקודת תורפה מבנית המצטלבת עם עוצמה לאומית, אמון במוסדות וקבלת החלטות מבצעית. חשיבותה נובעת מיתרון אסימטרי: היריבים יכולים להחדיר מניפולציות קטנות ביותר למערכי נתונים עצומים או לשרשראות אספקה של מודלים, ועדיין לייצר השפעות גדולות באופן לא פרופורציונלי במורד הזרם. אסימטריה זו היא התובנה המרכזית עבור מקבלי ההחלטות; השברירויות אינה טמונה במודל מסוים או בתת־מערכת יחידה, אלא ביחסי התלות המערכתיים שמאפשרים לרכיבים מורעלים לחדור לתוך אקוסיסטמים אנליטיים ומבצעיים שלמים.

עבור גורמי ביטחון לאומי, ההשלכה החשובה ביותר היא שהרעלת נתונים פועלת ברמת תשתית האמון. היא משנה את היסודות הסטטיסטיים שעליהם הלמידה של מערכות בינה מלאכותית מתבססת, ומעצבת מחדש באופן סמוי את הפלטים המזינים הערכות מודיעין, תכנון מבצעי ומערכות אזרחיות התומכות בקבלת החלטות. סיכון זה הופך פגיעות טכנית למשבר של ריבונות קוגניטיבית. כאשר היריבים מרעילים נתונים, הם למעשה מעוותים את היסודות האפיסטמיים של הממסד. כאשר אנליסטים, מפקדים או קובעי מדיניות אינם בטוחים עוד אם המלצותיה של מערכת בינה מלאכותית משקפות את המציאות או שמא מדובר במניפולציה של היריב, קצב המבצעים מואט, וסיפי קבלת ההחלטות על ידי גורמי אנוש מתקשחים. אובדן האמון בסביבות תחרותיות הוא בעצם העברה של היוזמה לידי היריב.

תובנה מרכזית נוספת היא שהאחריות להגנה על נתוני אימון ומודלים מאומנים מראש מפוצלת בין מספר גורמים. בניגוד לאבטחת סייבר מסורתית, שבה גבולות הרשת ברורים, פיתוח של מערכות בינה מלאכותית מצריך שימוש בתהליכים מבוזרים ובמאגרי מידע משותפים. מצב זה יוצר סיכון מסוג one-to-many שבו יריבים יכולים להרעיל מערכי נתונים ציבוריים או משקלי מודל בקוד פתוח, כדי לפגוע בסופו של דבר במערכות במורד הזרם מבלי שיצטרכו להיכנס כלל לתוך הסביבה המאובטחת. חשוב להדגיש כי התפשטות זו מתרחשת משום שארגונים (וגם יחידים) מושכים מרצונם את הנכסים האלה ממעלה הזרם אל תהליכי העבודה שלהם — מבצעים בהם כוונון עדין, משלבים אותם במוצרים או

מטמיעים אותם בתהליכים אנליטיים — ובכך הופכים שיטות פיתוח שגרתיות לערוצי הרעלה עצמית בלתי מכוונת. עבור קובעי המדיניות, משמעות הדבר היא כי הבטחת מהימנות ברמה הלאומית מחייבת אכיפה של משילות מתואמת לרוחב כל שרשרת האספקה של הבינה המלאכותית והתייחסות לזיהום בשרשרת האספקה כאל מכפיל סיכון אסטרטגי ולא רק כאל סוגיית רכש.

לבסוף, מקבלי ההחלטות צריכים להכיר בכך שפער הגילוי יוצר גירעון בהרתעה. כפי שהוסבר בחלק 4, התקפות מתוחכמות עושות שימוש בטריגרים רדומים שנשארים סמויים עד להופעתו של תנאי הפעלה מסוים. היות שיכולות אלו אינן ניתנות לגילוי בזמן רגיעה, גורמי ההגנה אינם יכולים לעורר מודעות או לגבות מחיר טרם התרחשות ההתקפה עצמה. הדבר מתקשר באופן ישיר לחוסר היציבות האסטרטגי שתואר קודם לכן; היריבים יכולים להציב מראש מנגנונים דיגיטליים — שכאשר הם מופעלים, הם גורמים להשפעה פיזית ממשית על מערכות, על תשתיות או על פעולות מבצעיות — ובכך לעקוף את שלבי ההסלמה והתגובה המקובלים. כתוצאה מכך המענה האסטרטגי חייב לעבור ממניעה מוחלטת לכיוון של הגנת עומק ואסטרטגיות חוסן המביאות בחשבון אפשרות של פגיעה חלקית.

לפיכך מנקודת מבט אסטרטגית יש להתייחס להרעלת נתונים הן כאתגר טכני והן כגורם לחץ מוסדי, המחייבים משילות ארוכת טווח, הבטחה של תקינות שרשרת האספקה וחוסן ברמת המשימה.

איור 9: תובנות לביטחון הלאומי

5.1
סדרי העדיפויות למחקר עתידי

אף שהמחקר בנושא הרעלת נתונים התרחב, הפער בין יכולות היריב ובין יכולות ההגנה נותר משמעותי. בחלק 3 תוארו ארכיטקטורות הגנה קיימות; סדרי העדיפויות המובאים להלן מתייחסים ליכולות הקריטיות שיש לבנות כעת כדי לגשר על הפער הזה.

העדיפות הראשונה היא לקדם את מדע הגילוי בהיקף רחב. שיטות קיימות — כגון אלו המתוארות בפרק 3 — הן יעילות עבור מערכי נתונים קטנים, אך מתקשות להתמודד עם קורפוסים בהיקף של האינטרנט או עם זרמי נתונים הטרוגניים שמשמשים LLMs בני זמננו. מחקר עתידי צריך לשאוף לפיתוח של גישות גילוי המסוגלות לבצע הסקה לגבי סטיות בהתפלגות ולאתר קמפיינים מובנים של הרעלת נתונים בהיקף רחב. השיטות האלה חייבות לפעול עם מטא־נתונים חלקיים ולהסתגל לאסטרטגיות מתפתחות כמו התקפות מסוג clean-label שאינן משאירות עקבות גלויים.

העדיפות השנייה עוסקת בפורנזיקה ברמת המודל ובייחוס אחריות. ככל שההסתמכות על מודלים חיצוניים גוברת, הארגונים זקוקים לכלים המסוגלים לבחון את מבנהו הפנימי של המודל כדי לאתר מוקדי התנהגות זדונית. נדרשות שיטות אבחון גמישות הניתנות להרחבה — בדומה לפורנזיקה דיגיטלית — שמסוגלות לנתח דפוסי אקטיבציה ולייחס פגיעה למקורות במעלה הזרם, גם כאשר נתוני האימון המקוריים אינם זמינים.

העדיפות השלישית היא פיתוח יכולות של ביטול למידה (machine unlearning) ותיקון מודלים. כפי שצוין בפרק 3, טכניקות ביטול הלמידה הקיימות משמשות בעיקר ככלי תגובה לאירועים וסובלות ממגבלות מהימנות, אך הן נמנות עם מספר קטן של גישות מבטיחות לצמצום העלויות העצומות הכרוכות באימון מחדש. נראה כי טכניקות של ביטול למידת מכונה לא יספקו פתרון מלא בעתיד הקרוב, אולם המחקר חייב להאיץ את הקצב כדי להפוך תיקון "כירורגי" — הסרה ממוקדת של מושגים מורעלים מבלי לפגוע בתועלת הכוללת של המודל — ליכולת מבצעית ישימה.

העדיפות הרביעית הדורשת תשומת לב היא שלמות שרשרת האספקה של מערכות הבינה המלאכותית. על המחקר לבחון כיצד להטמיע חתימות קריפטוגרפיות ומטא־נתוני

מקור כבר בשלבים המוקדמים ביותר של יצירת נתונים ומודלים. מאחר שהרעלת שרשרת אספקה פועלת כמכפיל כוח עבור היריבים, יש לבחון כיצד להעריך את מהימנותם של מאגרי נתונים ציבוריים ומרכזי מודלים בקוד פתוח מבלי להטיל עומס שמכביד יתר על המידה על חדשנות ופיתוח.

העדיפות החמישית מצויה בתחום הסימולציה, ה־red-teaming והסביבות הסינתטיות. מסגרות עתידיות חייבות לדמות קמפיינים מורכבים של הרעלת נתונים הן בשלב האימון והן בשלב ההסקה, לרבות איומים הייחודיים למערכות RAG, שבהן "הרעלת ידע" חודרת למסלולי האחזור עצמם. במקביל, יש צורך לפתח סימולציות ומשחקי מלחמה הממוקדים במקבלי החלטות, החושפים אנליסטים, קובעי מדיניות ומנהיגים מבצעיים לאופן שבו הרעלת נתונים מעצבת הערכות מודיעין, תהליכי תכנון ואופני תגובה למשברים. הסביבות האלה צריכות לבחון את חוסנם של המוסדות עצמם ולא להסתפק בבדיקת מודלים בלבד, כדי להבין כיצד מידע מורעל חודר ומתפשט לאורך מחזורי קבלת ההחלטות.

העדיפות השישית היא מחקר שיטתי של נקודות התורפה ברמת האלגוריתם — השכבה שבין הנתונים הגולמיים ובין המודל המאומן. מאחר שרבות מההתקפות מנצלות את תהליך האימון עצמו (כלי אופטימיזציה, שגרה של דגימה, פונקציות הפסד ולוגיקת תזמון), הרכיבים האלה מעצבים את האופן שבו המודל לומד ומפנים מידע. משמעות הדבר היא שתוקף יכול להטות את דינמיקת הלמידה מבלי לשנות את הקלטים או הפלטים. מחקר עתידי צריך למפות באופן שיטתי את משטח התקיפה ברמת האלגוריתם ולפתח כלים שמזהים מסלולי למידה חריגים בשלב מוקדם, לפני שהשפעתם מתפשטת בכל רחבי המודל.

לבסוף, **יש להעמיק במחקר של הממדים האנושיים והארגוניים של הרעלת נתונים**. בהמשך לממצאים בדבר שחיקה מוסדית שנדונו בפרק 4, המחקר צריך להתמקד באופנים שבהם בני אדם מפרשים אי־ודאות בהתנהגות של מערכות בינה מלאכותית וכיצד ארגונים מגיבים לחשד להרעלה. הבנה של הדינמיקות האלה חיונית לעיצוב תהליכי עבודה השומרים על קצב המשימה גם כאשר שלמות המערכת מוטלת בספק.

איור 10: סדרי עדיפויות למחקר עתידי לאבטחת מערכות בינה מלאכותית מפני הרעלת נתונים

הרעלת נתונים מייצגת סוג שונה מבחינה אסטרטגית של איום מצד היריבים — כזה הפועל במעלה הזרם, מנצל משאבים משותפים ופוגע ביסודות הקוגניטיביים והמוסדיים שעליהם נשענים תהליכי קבלת החלטות מודרניים המסתייעים בבינה מלאכותית. בעוד הקהילה הטכנית ממשיכה לקדם מחקרים בתחומי הגילוי והחוסן, מערך הביטחון הלאומי חייב להביא בחשבון שהרעלה תישאר מאפיין קבוע של התחרות בעתיד הנראה לעין. התמודדות עם איום זה תחייב תקנים מתואמים, תכנון ארכיטקטורות חסינות, בקרה מהימנה על שרשרת האספקה ומחקר מתמשך החוצה תחומי טכנולוגיה, תפעול וממשל. באמצעות תפיסה של הרעלת נתונים לא רק כבעיה טכנית אלא כפגיעות מערכתית המשפיעה על יתרון קבלת ההחלטות ברמה הלאומית, קובעי מדיניות יכולים לעצב עתיד שבו בינה מלאכותית ממשיכה להיות נכס אסטרטגי ולא סיכון סמוי.

מקורות

Ahler, Douglas J., Carolyn E. Roush, and Gaurav Sood. "The Micro-Task Market for Lemons: Data Quality on Amazon's Mechanical Turk." *Political Science Research and Methods* 13, no. 1 (2021): 1-20. https://doi.org/10.1017/psrm.2021.57.

Apruzzese, Giovanni, Michele Colajanni, Luca Ferretti, and Mirco Marchetti. "Addressing Adversarial Attacks Against Security Systems Based on Machine Learning." *11th International Conference on Cyber Conflict (CyCon)*. 2019. https://doi.org/10.23919/CYCON.2019.8756865.

Areola, Raphael I., Abayomi A. Adebiyi, and Katleho Moloi. "Artificial Intelligence for Optimizing Solar Power Systems with Integrated Storage: A Critical Review of Techniques, Challenges, and Emerging Trends." *Electricity* 6, no. 4 (2025): 60. https://doi.org/10. ATLAS 3390/electricity6040060.

"ATLAS Matrix." MITRE ATLAS. 2025. https://tinyurl.com/5x34zf5h.

Bagdasaryan, Eugene, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. "How To Backdoor Federated Learning." Version 3. Preprint, arXiv, 2018. https://doi.org/10.48550/ARXIV.1807.00459.

Barnett, Jackson. "Army Looks to Block Data 'Poisoning' in Facial Recognition, AI." *FedScoop*, n.d., Accessed November 7, 2025. https://tinyurl.com/595haz7h.

Biggio, Battista, and Fabio Roli. "Wild Patterns :Ten Years After the Rise of Adversarial Machine Learning." *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018. https://dl.acm.org/doi/10.1145/3243734.3264418.

Blanchard, Peva, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. "Byzantine-Tolerant Machine Learning." version 1. preprint, arXiv, 2017. https://tinyurl.com/5n76d785.

Buchanan, Ben. *The AI Triad and What It Means for National Security Strategy*. Center for Security and Emerging Technology (CSET), 2020. https://tinyurl.com/tzb4xh48.

Carlini, Nicholas, Matthew Jagielski, Christopher A. Choquette-Choo, et al. "Poisoning Web-Scale Training Datasets Is Practical." Version 2. Preprint, arXiv, 2023. https://doi.org/10.48550/ARXIV.2302.10149.

Casino, Fran. "Unveiling the Multifaceted Concept of Cognitive Security: Trends, Perspectives, and Future Challenges." *Technology in Society* 83, (2025). https://doi.org/10.1016/j.techsoc.2025.102956.

Center for Security and Emerging Technology, and Andrew Lohn. *Poison in the Well: Securing the Shared Resources of Machine Learning*. Center for Security and Emerging Technology, 2021. https://doi.org/10.51593/2020CA013.

Châtelet, Valentin. "Exposing Pravda: How Pro-Kremlin Forces Are Poisoning AI Models and Rewriting Wikipedia." *New Atlanticist*, 2025. https://tinyurl.com/37mumwjn.

Cheatham, Michael J., Angelique M. Geyer, Priscilla A. Nohle, and Jonathan E. Vazquez. "Cognitive Warfare: The Fight for Gray Matter in the Digital Gray Zone." *Joint Force Quarterly* 114, no. 3 (2024): 83–91.

CheckFirst. "'Pravda' Network: Worldwide Expansion and LLM, Wikipedia Pollution." March 13, 2025. https://tinyurl.com/yj7wz9u2.

Chen, Xinyun, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning." Version 1. Preprint, arXiv, 2017. https://doi.org/10.48550/ARXIV.1712.05526.

Chen, Yanjiao, Xiaotian Zhu, Xueluan Gong, Xinjing Yi, and Shuyang Li. "Data Poisoning Attacks in Internet-of-Vehicle Networks: Taxonomy, State-of-the-Art, and Future Directions." *IEEE Transactions on Industrial Informatics* 19, no. 1 (2023): 20–28. https://doi.org/10.1109/TII.2022.3198481.

Chen, Zhaorun, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. "AgentPoison: Red-Teaming LLM Agents via Poisoning Memory or Knowledge Bases." Version 1. Preprint, arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.12784.

Chivvis, Christopher S., and Jennifer Kavanagh, *How AI Might Affect Decisionmaking in a National Security Crisis*. The Carnegie Endowm for International Peace, n.d., accessed November 7, 2025. https://tinyurl.com/ynwckczn.

Cinà, Antonio Emanuele, Kathrin Grosse, Ambra Demontis, et al. "Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning." *ACM Computing Surveys* 55, no. 13s (2023): 1-39. https://doi.org/10.1145/3585385.

Committee on Using Machine Learning in Safety-Critical Applications :Setting a Research Agenda, Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, and National Academies of Sciences, Engineering, and Medicine. *Machine Learning for Safety-Critical Applications: Opportunities, Challenges, and a Research Agenda* .National Academies Press, 2025. https://doi.org/10.17226/27970.

Conti, Aaron. "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare." *Articles of War*, June 30, 2025. https://tinyurl.com/54c8z7sw.

Crain, Matthew, and Anthony Nadler. "Political Manipulation and Internet Advertising Infrastructure." *Journal of Information Policy* 9 (2019): 370–410. https://doi.org/10.5325/jinfopoli.9.2019.0370.

Crowther, Alexander. *National Defense and the Cyber Domain*. Heritage Foundation, 2017. https://tinyurl.com/yc87vjea.

Department of Defense. *Joint Publication (JP) 1: Doctrine for the Armed Forces of the United States*. U.S. Government Publishing Office. 2017. https://tinyurl.com/xj65t8z4.

Department of Homeland Security. *Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study*. Preparedness Series. 2023. https://tinyurl.com/2fv6whkf.

Dhanda, Sumit Singh, Deepak Panwar, Chia-Chen Lin, et al. "Advancement in Public Health through Machine Learning: A Narrative Review of Opportunities and Ethical Considerations." *Journal of Big Data* 12, no. 1 (2025): 154. https://doi.org/10.1186/s40537-025-01201-x.

Dong, Peiran, Song Guo, and Junxiao Wang. "Investigating Trojan Attacks on Pre-Trained Language Model-Powered Database Middleware." *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2023. https://doi.org/10.1145/3580305.3599395.

Epstein, Robert, and Ronald E. Robertson. "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections." *Proceedings of the National Academy of Sciences* 112, no. 33 (2015). https://doi.org/10.1073/pnas.1419828112.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, et al. "Robust Physical-World Attacks on Deep Learning Models." Preprint, arXiv, April 10, 2018. https://doi.org/10.48550/arXiv.1707.08945.

Fang, Minghong, Neil Zhenqiang Gong, and Jia Liu. "Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems." Preprint, arXiv, May 31, 2020. https://doi.org/10.48550/arXiv.2002.08025.

Fendley, Neil, Edward W. Staley, Joshua Carney, William Redman, Marie Chau, and Nathan Drenkow. "A Systematic Review of Poisoning Attacks Against Large Language Models." Version 1. Preprint, arXiv, 2025. https://doi.org/10.48550/ARXIV.2506.06518.

Galicia, Bailey. "In the Fight against Foreign Information Manipulation, the US Can't Afford to Disarm." *New Atlanticist*, 2025. https://tinyurl.com/mdpbb2u9.

Geiping, Jonas, Liam Fowl, W. Ronny Huang, et al. "Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching." Version 2. Preprint, arXiv, 2020. https://doi.org/10.48550/ARXIV.2009.02276.

Giannaros, Anastasios, Aristeidis Karras, Leonidas Theodorakopoulos, et al. "Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions." *Journal of Cybersecurity and Privacy* 3, no. 3 (2023): 493–543. https://doi.org/10.3390/jcp3030025.

Gloaguen, Thibaud, Mark Vero, Robin Staab, and Martin Vechev. "Watch Your Steps: Dormant Adversarial Behaviors That Activate upon LLM Finetuning." Preprint, arXiv, October 9, 2025. https://doi.org/10.48550/arXiv.2505.16567.

Global Influence Operations Report. *Russian Information Warfare Campaign: Kremlin Poisons AI and Rewrites Wikipedia.* 2025. https://tinyurl.com/368rbn52.

Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain." Version 2. Preprint, arXiv, 2017. https://doi.org/10.48550/ARXIV.1708.06733.

Guo, Jie. "The Ethical Legitimacy of Autonomous Weapons Systems: Reconfiguring War Accountability in the Age of Artificial Intelligence." *Ethics & Global Politics* 18, no. 3 (2025): 27–39. https://doi.org/10.1080/16544951.2025.2540131.

Guo, Wei, Benedetta Tondi, and Mauro Barni. “An Overview of Backdoor Attacks Against Deep Neural Networks and Possible Defences.” Preprint, arXiv, November 16, 2021. https://doi.org/10.48550/arXiv.2111.08429.

Halder, Deepon, Anshika Gupta, Diya Ghosh, and Hafizur Rehman. “A Comprehensive Survey of Data Poisoning Attacks and Their Detection Techniques.” Preprint, 2025. https://doi.org/10.13140/RG.2.2.20084.67207.

Hartle III, Frank, Steve Mancini, and Emily Kerry. “Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges.” *Issues in Information Systems* 25, no. 4 (2025): 433–42.

Harvey, Andrew S. “The Levels of War as Levels of Analysis.” *Military Review*, December 2021. https://tinyurl.com/yw5e7w3h.

Hennessy, J. P. “Why Google’s Researchers Are Intentionally Poisoning Datasets and More.” *Lightning AI*, February 23, 2023. https://tinyurl.com/m25p7r7x.

Hershkovitz, Shay, and Corinna Turbes. *The Imperative of Data Provenance in AI*. Data Foundation, 2025. https://tinyurl.com/mr3344us.

“IaaS vs. PaaS vs. SaaS.” Red Hat. n.d. https://tinyurl.com/4kyubtxb.

IBM-Think. “What Is Three-Tier Architecture?.” n.d. Accessed November 19, 2025. https://tinyurl.com/34z72zmt.

Improta, Cristina. “Detecting Stealthy Data Poisoning Attacks in AI Code Generators.” Preprint, arXiv, August 29, 2025. https://doi.org/10.48550/arXiv.2508.21636.

Jagielski, Matthew, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning.” Version 3. Preprint, arXiv, 2018. https://doi.org/10.48550/ARXIV.1804.00308.

Jagielski, Matthew, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. “Subpopulation Data Poisoning Attacks.” Version 3. Preprint, arXiv, 2020. https://doi.org/10.48550/ARXIV.2006.14026.

Jaiswal, Ayshwarya, Pragya Dwivedi, and Rupesh Kumar Dewang. “Machine Learning Approaches to Detect, Prevent and Mitigate Malicious Insider Threats: State-of-the-Art Review.” *Multimedia Tools and Applications* 84, no. 24 (2024): 28909–49. https://doi.org/10.1007/s11042-024-20273-0.

Jha, Rishi D., Jonathan Hayase, and Sewoong Oh. “Label Poisoning Is All You Need.” Preprint, arXiv, October 29, 2023. https://doi.org/10.48550/arXiv.2310.18933.

Jiao, Ruochen, Shaoyuan Xie, Justin Yue, et al. “Can We Trust Embodied Agents? Exploring Backdoor Attacks against Embodied LLM-Based Decision-Making Systems.” Preprint, arXiv, April 30, 2025. https://doi.org/10.48550/arXiv.2405.20774.

Jin, Lingxin, Xianyu Wen, Wei Jiang, and Jinyu Zhan. “A Survey of Trojan Attacks and Defenses to Deep Neural Networks.” Preprint, arXiv, August 15, 2024. https://doi.org/10.48550/arXiv.2408.08920.

Kamrani, Farzad, Linus Kanestad, Linus Luotsinen, et al. *Attacking and Deceiving Military AI Systems.* FOI-R--5396--SE. Swedish Defence Research Agency, 2023. https://tinyurl.com/3ddn3p7c.

Kiribuchi, Naoto, Kengo Zenitani, and Takayuki Semitsu. "Securing AI Systems: A Guide to Known Attacks and Impacts." Preprint, arXiv, June 29, 2025. https://doi.org/10.48550/arXiv.2506.23296.

Koh, Pang Wei, Jacob Steinhardt, and Percy Liang. "Stronger Data Poisoning Attacks Break Data Sanitization Defenses." Version 2. Preprint, arXiv, 2018. https://doi.org/10.48550/arXiv.1811.00741.

Kraft, Amy. "Microsoft Shuts Down AI Chatbot after It Turned into a Nazi." *CBS News*, March 25, 2016. https://tinyurl.com/y6kfk8as.

Lee, Dave. "Tay: Microsoft Issues Apology over Racist Chatbot Fiasco." *BBC*, March 25, 2016. https://tinyurl.com/5n84d8bd.

Lee, Peter. "Learning from Tay's Introduction." *Official Microsoft Blog*, March 25, 2016. https://tinyurl.com/xwmw2akw.

Li, Linyang, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. "Backdoor Attacks on Pre-Trained Models by Layerwise Weight Poisoning." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* 2021. https://doi.org/10.18653/v1/2021.emnlp-main.241.

Li, Yiming, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. "Backdoor Learning: A Survey." Preprint, arXiv, February 16, 2022. https://doi.org/10.48550/arXiv.2007.08745.

Lin-Greenberg, Erik. "Allies and Artificial Intelligence: Obstacles to Operations and Decision-Making." Preprint, Texas National Security Review, 2020. https://doi.org/10.26153/TSW/8866.

Liu, Yiyong, Michael Backes, and Xiao Zhang. "Transferable Availability Poisoning Attacks." Version 2. Preprint, arXiv, 2023. https://doi.org/10.48550/arXiv.2310.05141.

"LLM03: 2025 Supply Chain." OWASP. 2025.

Lumenova. "Data Poisoning Attacks: How AI Models Can Be Corrupted." *Lumenova Blog*, July 17, 2025. https://tinyurl.com/39upvba7.

Luo, Peng, Buhong Wang, Jiwei Tian, and Yong Yang. "ADS-Bpois: Poisoning Attacks Against Deep-Learning-Based Air Traffic ADS-B Unsupervised Anomaly Detection Models." *IEEE Internet of Things Journal* 11, no. 23 (2024): 38301–11. https://doi.org/10.1109/JIOT.2024.3446675.

Ma, Yuan, Jiankang Wei, Yilun Lyu, Kehao Chen, and Jingtong Huang. "Backdoor Attack with Invisible Triggers Based on Model Architecture Modification." Version 3. Preprint, arXiv, 2024. https://doi.org/10.48550/arXiv.2412.16905.

Mendes, Vanessa I. S., Beatriz M. F. Mendes, Rui Pedro Moura, et al. "Harnessing Artificial Intelligence for Enhanced Public Health Surveillance: A Narrative Review." *Frontiers in Public Health* 13 (2025). https://doi.org/10.3389/fpubh.2025.1601151.

Müller, Nicolas Michael, Daniel Kowatsch, and Konstantin Böttinger. "Data Poisoning Attacks on Regression Learning and Corresponding Defenses." Version 1. Preprint, arXiv, 2020. https://doi.org/10.48550/ARXIV.2009.07008.

Muñoz-González, Luis, Battista Biggio, Ambra Demontis, et al. "Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization." Version 1. Preprint, arXiv, 2017. https://doi.org/10.48550/ARXIV.1708.08689.

Nguyen, Quang H., Nguyen Ngoc-Hieu, The-Anh Ta, et al. "Wicked Oddities: Selectively Poisoning for Effective Clean-Label Backdoor Attacks." Preprint, arXiv, July 16, 2024. https://doi.org/10.48550/arXiv.2407.10825.

O'Hanlon, Michael E. *A Retrospective on the So-Called Revolution in Military Affairs, 2000-2020*. Brookings, 2018. https://tinyurl.com/y9rtshmy.

Oliynyk, Daryna, Rudolf Mayer, and Andreas Rauber. "I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences." *ACM Computing Surveys* 55, no. 14s (2023): 1–41. https://doi.org/10.1145/3595292.

Panda, Ashwinee, Saeed Mahloujifar, Arjun N. Bhagoji, Supriyo Chakraborty, and Prateek Mittal. "SparseFed: Mitigating Model Poisoning Attacks in Federated Learning with Sparsification." Preprint, arXiv, December 12, 2021. https://doi.org/10.48550/arXiv.2112.06274.

Pawlicki, Marek, Aleksandra Pawlicka, Rafał Kozik, and Michał Choraś. "A Meta-Survey of Adversarial Attacks against Artificial Intelligence Algorithms, Including Diffusion Models." *Neurocomputing* 653 (2025). https://doi.org/10.1016/j.neucom.2025.131231.

Pedersen, Frederik A. H., and Jeppe T. Jacobsen. "Narrow Windows of Opportunity: The Limited Utility of Cyber Operations in War." *Journal of Cybersecurity* 10, no. 1 (2024). https://doi.org/10.1093/cybsec/tyae014.

Puscas, Ioana. *AI and International Security: Understanding the Risks and Paving the Path for Confidence-Building Measures*. UNIDIR, 2023. https://tinyurl.com/3p5uvfw8.

Qiu, Wenjun. "A Survey on Poisoning Attacks Against Supervised Machine Learning." Version 2. Preprint, arXiv, 2022. https://doi.org/10.48550/ARXIV.2202.02510.

Ramirez, Miguel A., Song-Kyoo Kim, Hussam Al Hamadi, et al. "Poisoning Attacks and Defenses on Artificial Intelligence: A Survey." Version 2. Preprint, arXiv, 2022. https://doi.org/10.48550/ARXIV.2202.10276.

Rilkoff, Heather, Shannon Struck, Chelsea Ziegler, Laura Faye, Dana Paquette, and David Buckeridge. "Innovations in Public Health Surveillance: An Overview of Novel Use of Data and Analytic Methods." *Canada Communicable Disease Report* 50, no. 3/4 (2024): 93–101. https://doi.org/10.14745/ccdr.v50i34a02.

Rosiek, Travis. "AI Data Poisoning, Wiper Malware, Critical Infrastructure Attacks Could Increase in 2025, Impacting Government Cyber Resilience." *GovLoop*, January 21, 2025. https://tinyurl.com/2bh35yna.

Rosiek, Travis. "Data Poisoning Threatens AI's Promise in Government." *FedTech*, March 21, 2025. https://tinyurl.com/4ap7fwm3.

Safety and Security Guidelines for Critical Infrastructure Owners and Operators. Department of Homeland Security. 2024. https://tinyurl.com/2t8memjd.

Schulze, Matthias. "Cyber in War: Assessing the Strategic, Tactical, and Operational Utility of Military Cyber Operations." 12th International Conference on Cyber Conflict (CyCon), May 2020. https://doi.org/10.23919/CyCon49761.2020.9131733.

Schwarzschild, Avi, Micah Goldblum, Arjun Gupta, John P. Dickerson, and Tom Goldstein. "Just How Toxic Is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks." Preprint, arXiv, June 17, 2021. https://doi.org/10.48550/arXiv.2006.12557.

Shafahi, Ali, W. Ronny Huang, Mahyar Najibi, et al. "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks." In *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018. https://tinyurl.com/yc7zabrn.

Shah, Anwar, Adil Ahmad, Bahar Ali, Sajid Anwer, and Qamar Uz Zaman. "Guarding the Gates: A Comprehensive Survey of Backdoor Attacks on Neural Networks." Preprint, SSRN, 2024. https://doi.org/10.2139/ssrn.4966942.

Shan, Shawn, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models." Preprint, arXiv, April 29, 2024. https://doi.org/10.48550/arXiv.2310.13828.

Sherman, Justin. "Data Brokers and Data Breaches." *Duke – Tech Policy Program Blog*, September 27, 2022. https://tinyurl.com/rv9rb9tx.

Sörensen, Sara, and James Pamment. *Operationalising the Framework for Evaluating Capability against Information Influence Operations: A Case Study of the Psychological Defence Agency's Courses*. NATO Strategic Communications Centre of Excellence, 2023.

Souly, Alexandra, Javier Rando, Ed Chapman, et al. "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples." Preprint, arXiv, October 8, 2025. https://doi.org/10.48550/arXiv.2510.07192.

Steinhardt, Jacob, Pang Wei Koh, and Percy Liang. "Certified Defenses for Data Poisoning Attacks." Preprint, arXiv, November 24, 2017. https://doi.org/10.48550/arXiv.1706.03691.

Stephan III, Paul B. "Big Data as a National Security Issue." *University of Chicago Legal Forum* 2024 (2025). https://tinyurl.com/38k77vun.

Stokel-Walker, Chris. "You Can Poison AI Datasets for Just $60, a New Study Shows." *Fast Company*, March 3, 2023. https://tinyurl.com/32fet98s.

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification." *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014. https://doi.org/10.1109/CVPR.2014.220.

Terziyan, Vagan, Mariia Golovianko, and Svitlana Gryshko. "Industry 4.0 Intelligence Under Attack: From Cognitive Hack to Data Poisoning." In *NATO Science for Peace and Security Series – D: Information and Communication Security*. IOS Press, 2018. https://doi.org/10.3233/978-1-61499-888-4-110.

"The Business Model of Data Poisoning-as-a-Service (DPaaS)." *AI Competence.Org*. October 12, 2025. https://tinyurl.com/5fh3hdj6.

Tian, Zhiyi, Lei Cui, Jie Liang, and Shui Yu. "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning." *ACM Computing Surveys* 55, no. 8 (2022): 1–35. https://doi.org/10.1145/3551636.

Turner, Alexander, Dimitris Tsipras, and Aleksander Maądry. *Clean-Label Backdoor Attacks*. n.d. https://tinyurl.com/y6vea5f8.

U.S. Department of Justice. *Six Russian GRU Officers Charged in Connection with Worldwide Deployment of Destructive Malware and Other Disruptive Actions in Cyberspace*. 2020. https://tinyurl.com/yernt666.

Vallance, Chris. "Wikipedia Blames Pro-China Infiltration for Bans." *BBC*, September 16, 2021. https://tinyurl.com/a3dph5w4.

Vassilev, Apostol, Alina Oprea, Alie Fordyce, and Hyrum Andersen. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. NIST AI 100-2e2023. National Institute of Standards and Technology, 2024. https://doi.org/10.6028/NIST.AI.100-2e2023.

Venkatesan, Sridhar, Harshvardhan Sikka, Rauf Izmailov, Ritu Chadha, Alina Oprea, and Michael J. De Lucia. "Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems." *MILCOM 2021–2021 IEEE Military Communications Conference* (*MILCOM*). 2021. https://doi.org/10.1109/MILCOM52596.2021.9652916.

Wang, Gang, Tianyi Wang, Haitao Zheng, and Ben Y. Zhao. "Man vs. Machine: Practical Adversarial Detection of Malicious Crowdsourcing Workers." *Proceedings of the 23rd USENIX Security Symposium*. 2014. https://tinyurl.com/48y33ckr.

Wang, Qingyue, Qi Pang, Xixun Lin, Shuai Wang, and Daoyuan Wu. "BadMoE: Backdooring Mixture-of-Experts LLMs via Optimizing Routing Triggers and Infecting Dormant Experts." Version 2. Preprint, arXiv, 2025. https://doi.org/10.48550/ARXIV.2504.18598.

Wang, Shuo., Surya Nepal, Carsten Rudolph, Marthie Grobler, Shangyu Chen, and Tianle Chen. "Backdoor Attacks Against Transfer Learning with Pre-Trained Deep Learning Models." *IEEE Transactions on Services Computing* 15, no. 3 (2022): 1526-39. https://tinyurl.com/tcabz259.

Wang, Zhibo, Jingjing Ma, Xue Wang, Jiahui Hu, Zhan Qin, and Kui Ren. "Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems." *ACM Computing Surveys* 55, no. 7 (2022): 1–36. https://doi.org/10.1145/3538707.

Wei, Wenqi, Ka-Ho Chow, Yanzhao Wu, and Ling Liu. "Demystifying Data Poisoning Attacks in Distributed Learning as a Service." *IEEE Transactions on Services Computing* 17, no. 1 (2024): 237–50. https://doi.org/10.1109/TSC.2023.3341951.

Whitmore, Wendi. "6 Predictions for the AI Economy: 2026's New Rules of Cybersecurity." *Palo Alto Networks Blog*, November 18, 2025. https://tinyurl.com/5393c6xa.

Wikipedia. "List of Political Editing Incidents on Wikipedia." Accessed December 15, 2025. https://tinyurl.com/y8dfd46z.

Wikipedia. "Tay (Chatbot)." Accessed October 10, 2025. https://tinyurl.com/3dy72bx2.

Wolf, Marty J., Keith W. Miller, and F. S. Grodzinsky. "Why We Should Have Seen That Coming." *The ORBIT Journal* 1, no. 2 (2017): 1–12. https://doi.org/10.29297/orbit.v1i2.49.

Wu, Jianping, Jiahe Jin, and Chunming Wu. "Challenges and Countermeasures of Federated Learning Data Poisoning Attack Situation Prediction." *Mathematics* 12, no. 6 (2024): 901. https://doi.org/10.3390/math12060901.

Xia, Geming, Jian Chen, Chaodong Yu, and Jun Ma. "Poisoning Attacks in Federated Learning: A Survey." *IEEE Access* 11 (2023): 10708–22. https://doi.org/10.1109/ACCESS.2023.3238823.

Xu, Jie, Zihan Wu, Cong Wang, and Xiaohua Jia. "Machine Unlearning: Solutions and Challenges." *IEEE Transactions on Emerging Topics in Computational Intelligence* 8, no. 3 (2024): 2150–68. https://doi.org/10.1109/TETCI.2024.3379240.

Xu, Keyizhi, Yajuan Lu, Zhongyuan Wang, and Chao Liang. "A Survey of Adversarial Examples in Computer Vision: Attack, Defense, and Beyond." *Wuhan University Journal of Natural Sciences* 30, no. 1 (2025): 1–20. https://doi.org/10.1051/wujns/2025301001.

Xu, Xiaojun, Qi Wang, Huichen Li, Nikita Borisov, Carl A. Gunter, and Bo Li. "Detecting AI Trojans Using Meta Neural Analysis." Preprint, arXiv, October 1, 2020. https://doi.org/10.48550/arXiv.1910.03137.

Xue, Jiaqi, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. "BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models." Preprint, arXiv, June 6, 2024. https://doi.org/10.48550/arXiv.2406.00083.

Yin, Dong, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rate." version 2. preprint, arXiv, 2018. https://tinyurl.com/y3ku7w8j.

Zhang, Chen, Zhuo Tang, and Kenli Li. "Clean-Label Poisoning Attack with Perturbation Causing Dominant Features." *Information Sciences* 644 (2023). https://doi.org/10.1016/j.ins.2023.03.124.

Zhang, Heyi, Yule Liu, Xinlei He, Jun Wu, Tianshuo Cong, and Xinyi Huang. "SoK: Benchmarking Poisoning Attacks and Defenses in Federated Learning." Preprint, arXiv, February 6, 2025. https://doi.org/10.48550/arXiv.2502.03801.

Zhang, Jun, and Dan Tenney. "The Evolution of Integrated Advance Persistent Threat and Its Defense Solutions: A Literature Review." *Open Journal of Business and Management* 12, no. 1(2024): 293-338. https://doi.org/10.4236/ojbm.2024.121021.

Zhang, Li Ang, Gavin S. Hartnett, Jair Aguirre, et al. *Operational Feasibility of Adversarial Attacks Against Artificial Intelligence*. Research Report. RAND, 2022. https://tinyurl.com/5hyp7fxk.

Zhang, Quan, Binqi Zeng, Chijin Zhou, Gwihwan Go, Heyuan Shi, and Yu Jiang. "Human-Imperceptible Retrieval Poisoning Attacks in LLM-Powered Applications." Preprint, arXiv, April 26, 2024. https://doi.org/10.48550/arXiv.2404.17196.

Zhang, Xueqing, Junkai Zhang, Ka-Ho Chow, et al. "Visualizing the Shadows: Unveiling Data Poisoning Behaviors in Federated Learning." Version 1. Preprint, arXiv, 2024. https://doi.org/10.48550/arXiv.2405.16707.

Zhang, Yiming, Javier Rando, Ivan Evtimov, et al. “Persistent Pre-training Poisoning of LLMs.” Preprint, arXiv, October 17, 2024. https://doi.org/10.48550/arXiv.2410.13722.

Zhao, Pinlong, Weiyao Zhu, Pengfei Jiao, Di Gao, and Ou Wu. “Data Poisoning in Deep Learning: A Survey.” Version 1. Preprint, arXiv, 2025. https://doi.org/10.48550/ARXIV.2503.22759.

Zheng, Xinyi, Chen Wei, Shenao Wang, et al. “Towards Robust Detection of Open Source Software Supply Chain Poisoning Attacks in Industry Environments.” *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*. 2024. https://doi.org/10.1145/3691620.3695262.

Zhong, Zexuan, Ziqing Huang, Alexander Wettig, and Danqi Chen. “Poisoning Retrieval Corpora by Injecting Adversarial Passages.” Preprint, arXiv, October 29, 2023. https://doi.org/10.48550/arXiv.2310.19156.

Zhou, Yihe, Tao Ni, Wei-Bin Lee, and Qingchuan Zhao. “A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluation Methods.” *Transactions on Artificial Intelligence* 1, no. 1(2025): 28-58. https://doi.org/10.53941/tai.2025.100003.

Zhu, Yanxu, Hong Wen, Runhui Zhao, Yixin Jiang, Qiang Liu, and Peng Zhang. “Research on Data Poisoning Attack against Smart Grid Cyber–Physical System Based on Edge Computing.” *Sensors* 23, no. 9 (2023): 4509. https://doi.org/10.3390/s23094509.

Zou, Wei, Runpeng Geng, Binghui Wang, and Jinyuan Jia. “PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models.” Version 3. Preprint, arXiv, 2024. https://doi.org/10.48550/arXiv.2402.07867.

מחקר זה בוחן באופן מקיף את התופעה של הרעלת נתונים – פגיעות אסטרטגית קריטית שבה היריבים מבצעים מניפולציה מכוונת בנתוני האימון של מערכות בינה מלאכותית כדי לחתור מבפנים תחת יסודות המערכות. מעבר להגדרות הטכניות, הוא מציג את הרעלת הנתונים כאיום מבני על היסודות הקוגניטיביים של המדינאות המודרנית, שבה הבינה המלאכותית מעצבת יותר ויותר את האופן שבו מדינות תופסות סיכונים, מקצות משאבים ומקבלות החלטות. המחקר מראה כיצד מניפולציות עדינות בשלבים מוקדמים של התהליך יכולות להתפשט בחשאי לתשתיות צבאיות, לתשתיות מודיעין ולתשתיות אזרחיות.

המחקר מתבסס על מקרי בוחן, החל במניפולציה של פלטפורמות ידע שיתופיות וכלה בהרעלת מודלים קליניים ואנליטיים, ומציג נרטיב שממפה אקוסיסטם מגוון של שחקנים – מיריבים מדינתיים ועד לחובבים בעלי משאבים מוגבלים. הוא ממחיש מדוע הזיהוי נותר כה מאתגר: התקפות מתוחכמות מסוג clean-label (הרעלת נתונים ללא תיוג זדוני) עשויות לעבור תהליך אימות שגרתי, להישאר רדומות למשך תקופות ממושכות ולהיכנס לפעולה רק בתנאים מבצעיים מסוימים – לרוב כאשר ההשלכות הן חמורות במיוחד.

מחקר זה משמש כמדריך אסטרטגי עבור מקבלי החלטות המתמודדים עם איומים שמבוססים על בינה מלאכותית. הטענה המוצגת בו היא שהסיכון הגדול ביותר אינו טמון בכשלים מסוימים של מודלים, אלא בהתפשטות חשאית של נתונים משובשים לאורך שרשרת האספקה של הבינה המלאכותית, שבה מניפולציה יחידה במעלה הזרם עלולה להשפיע על אלפי מערכות במורד הזרם. המסקנה היא שאבטחת עתיד הבינה המלאכותית מחייבת מעבר מהגנה היקפית לתפיסה אמיתית של הגנת עומק – כזו שמעוגנת במעקב אחר מקור הנתונים, בניטור רציף ובשיתוף פעולה מתמשך בין מפתחי טכנולוגיה ובין גורמים בקהילת הביטחון הלאומי.

ד"ר שי הרשקוביץ הוא חוקר העוסק בתחום הביטחון הלאומי. הוא מתמחה בטכנולוגיות מתפתחות ודו־שימושיות ומתמקד באופן מיוחד בבינה מלאכותית, בחדשנות בתחום הביטחון ובטרנספורמציה בתחום המודיעין. בעבודתו הוא בוחן את הדרכים שבהן בינה מלאכותית, מערכות המבוססות על נתונים ויכולות טכנולוגיות חדשניות מעצבות מחדש את תחומי המודיעין, המבצעים הצבאיים וקבלת ההחלטות האסטרטגית. הוא משמש פרופסור נלווה בתוכנית לתואר שני במודיעין יישומי באוניברסיטת ג'ורג'טאון וחוקר בכיר (נלווה) ב־RAND, ועוסק בייעוץ לממשלות ולחברות Fortune 500 בנושאים של סיכוני בינה מלאכותית, אסטרטגיות הגנה וטכנולוגיות דו־שימושיות וההשלכות של טכנולוגיות מתפתחות על הביטחון הלאומי. ד"ר הרשקוביץ הוא המחבר של הספר *The Future of National Intelligence: How Emerging Technologies Reshape Intelligence Communities* (2022) ומחבר שותף של הספר **אמ"ן יוצא לאור: העשור הראשון לאגף המודיעין בצה"ל** (2013). כמו כן הוא מרבה לפרסם בכתבי עת אקדמיים ובמדיה האמריקנית.

המזכר נכתב במסגרת פרויקט מחקר בנושא השפעה זרה, המתבצע במכון למחקרי ביטחון לאומי בסיוע מערך הסייבר הלאומי ומפא"ת.