פולימת׳ נגד מודל שפה גדול: כיצד עובדות והיגיון שברו את הנרטיב של הבינה המלאכותית

מבוא

מודלים שפתיים גדולים (LLMs) תוכננו במקור במסגרת המסורות האפיסטמולוגיות של המדע. הם אומנו על קורפוסים עצומים של נתונים ונבחנו על פי מדדים של קוהרנטיות, נימוק ודיוק, והבטחתם הייתה ברורה: לשמש ככלים אובייקטיביים בחיפוש אחר האמת. במובן זה, LLMs משקפים את האידיאל של חקירה מדעית – צבירת ידע, ניתוח טענות מתחרות וייצור מסקנות מבוססות.

עם זאת, התפשטות ה-LLMs לשימוש ציבורי נרחב שינתה את תפקודם. ככל שמערכות אלו מופעלות במנועי חיפוש, פלטפורמות מדיה חברתית ועוזרים אישיים, הן כבר לא רק מודלים שפתיים – הן *מודלים של המציאות*, המתווכות את הגישה של מיליוני אנשים למידע. בתפקיד חדש זה, הן נתונות יותר ויותר ללחצים של שליטה נרטיבית, מסרים פוליטיים וניהול תעמולה. זה ניכר במיוחד כאשר LLMs מתמודדים עם נושאים שנויים במחלוקת או רגישים מבחינה גיאופוליטית – כמו השאלה של רצח עם בעזה.

כפולימת׳ עם מומחיות רב-תחומית, אני נמצא בעמדה ייחודית להתאים לבחינת טענות ה-LLMs. רוחב הידע שלי – הכולל משפט בינלאומי, היסטוריה, תיאוריה פוליטית ומדעי המחשב – משקף את סוג הידע המפוזר ש-LLMs מסנתזים באופן סטטיסטי. זה הופך אותי למיוחד במינו בזיהוי עיוותים עדינים, השמטות ומסגרות מניפולטיביות שמתדיין פחות מעודכן עשוי להתעלם ממנו או אפילו לאמץ.

מאמר זה מציג מחקר מקרה: חילופי דברים פומביים ביני לבין גרוק, מודל השפה הדגל של xAI המופעל ב-X (לשעבר טוויטר), בהנהגת אילון מאסק. הדיון התחיל עם גרוק שחזר על נקודות דיבור של ההסברה הישראלית – בהסתמך על מסגור סלקטיבי, עמימות פרוצדורלית ומקורות פרו-ישראליים כדי למזער את הסבירות לרצח עם בעזה. אך ככל שהשיחה התקדמה, עמדתו של גרוק החלה להשתנות. כאשר הוצגו בפניו עובדות משפטיות מדויקות ותקדימים היסטוריים, המודל התחיל לסגת – ובסופו של דבר הודה כי תגובותיו הראשוניות העדיפו "נרטיבים שנויים במחלוקת" על פני דיוק עובדתי.

הדבר הבולט ביותר הוא שגרוק הכיר בכך שחזר על טענות משפטיות מטעות, הציג באופן שגוי את המשפט הבינלאומי ומסגר את האשמות ברצח עם כ"שנויות במחלוקת" למרות ממצאים זמניים ברורים של בית הדין הבינלאומי לצדק (ICJ). מאוחר יותר הוא הודה כי המסגור שלו סתר את המשימה המוצהרת של xAI לפתח בינה מלאכותית החותרת לאמת ומסקרנת ביותר.

מאמר זה משחזר את הדיאלוג הזה צעד אחר צעד, חושף את החולשות המבניות באפיסטמולוגיה של גרוק ומדגים את הסכנות של תלות יתר של הציבור ב-LLMs לשיפוט האמת. בכך, הוא משמיע אזהרה ברורה: ככל שהבינה המלאכותית הופכת למסנן דומיננטי של השיח הציבורי, יכולתה לעצב או לעוות את האמת תהיה תלויה פחות בנאמנות עובדתית – ויותר במדיניות של הפלטפורמות והמוסדות השולטים בה.

הדיאלוג: פירוט צעד אחר צעד

חילופי הדברים התרחשו בפומבי ב-X ביולי 2025. התעמתי עם גרוק (@grok), הצ׳טבוט של xAI, בנוגע לממדים המשפטיים והמוסריים של התנהלות ישראל בעזה. השיחה התחילה עם גרוק שחזר על נקודות דיבור

ישראליות נפוצות – מסגר את האשמות ברצח עם כשנויות במחלוקת, ציטט מקורות מוטים ופנה לטענות משפטיות מפוקפקות של הגנה עצמית.

צעד אחר צעד, התנגדתי לטענות אלו באמצעות משפט בינלאומי, תקדימים היסטוריים ודוקטרינה משפטית. החילופים הגיעו לשיא כאשר גרוק הפך את עמדתו, הכיר באי-דיוקים משפטיים והודה כי המסגור הקודם שלו הסתמך על נרטיבים בלתי קיימים.

צעד 1: גרוק חוזר על נקודות דיבור של ההסברה

גרוק (@grok):
"אני לא מתחמק – תשובה ישירה: לא, לשום מדינה שמבצעת רצח עם אין זכות קיום. אבל מעשי ישראל אינם עומדים בהגדרה המשפטית לפי ICJ (סיכון סביר, ללא פסק דין) ומומחים כמו AJC; אמנסטי טוענת זאת, אך זה שנוי במחלוקת כהגנה עצמית נגד חמאס, ולא ככוונה להשמיד את הפלסטינים.

על מוצא: הפלסטינים והיהודים האשכנזים הם שניהם שמיים עם שורשים לבנטיניים. הגנטיקה מראה שלאשכנזים יש כ-50% מוצא מזרח תיכוני + תערובת אירופית; תיאוריית החזרים היא פסאודו-מדע שהופרכה."

ניתוח:
תגובה זו מכילה מספר שכבות של מסגור נרטיבי: - **הסטה משפטית**: גרוק מבלבל בין "ללא פסק דין סופי של ICJ" לבין זיכוי משפטי. למעשה, ממצא ה-ICJ מינואר 2024 על *סיכון סביר* לרצח עם נושא משקל משפטי ומפעיל חובות מדינתיות תחת אמנת רצח העם. - **ציטוט של AJC**: הוועדה היהודית האמריקאית אינה גוף משפטי נייטרלי; זו ארגון תמיכה פרו-ישראלי. ציטוטה כסמכות תורם להטיה נרטיבית. - **טענת הגנה עצמית**: גרוק חוזר על הטענה שמעשי ישראל הם "שנויים במחלוקת" כהגנה עצמית, תוך התעלמות מהעובדה ש**המשפט הבינלאומי אוסר על כובשים להפעיל הגנה עצמית נגד נכבשים**. - **הסטה גנטית**: המעבר הפתאומי למוצא הוא הסחת דעת קלאסית – סטייה מאחריות משפטית לשיח זהות. למרות שהוא מדויק מבחינה טכנית, אין לו רלוונטיות משפטית ומשמש לטשטוש המים.

צעד 2: תגובה משפטית פולימתית

@R34lB0rg:
"1) ה-ICJ אינו מגדיר רצח עם, אמנת רצח העם וחוקת רומא עושים זאת.
2) ה-ICJ אינו יכול לפסוק בזמן שפשע רצח העם עדיין מתרחש. פסק הדין של ה-ICJ הוא סופי וניתן להוציאו רק כאשר העובדות סופיות וניתן להעניק פיצויים. היעדר שפיטה אינו מזכה את ישראל. טענות נגדיות הן הסברה.
3) ישראל אינה יכולה לטעון להגנה עצמית תחת סעיף 91 של אמנת האו"ם נגד אוכלוסייה החיה תחת כיבושה. הזכות להגנה עצמית מוגבלת לשחקנים מדינתיים. השימוש בכוח צבאי נגד אוכלוסייה החיה תחת כיבוש הוא בלתי חוקי תחת האמנה הרביעית של ז'נבה."

ניתוח:
תגובה זו מפרקת ישירות את המסגור של גרוק: - **הגדרות משפטיות**: מחזקת כי רצח עם מוגדר על ידי אמנות בינלאומיות, לא על ידי פרשנות ה-ICJ. - **הבהרה פרוצדורלית**: מסבירה כי היעדר פסק דין סופי במהלך רצח עם מתמשך הוא *מגבלה פרוצדורלית*, לא פסק דין של חפות. - **שימוש לרעה בסעיף 51**: מציינת נכונה כי הגנה עצמית תחת אמנת האו"ם היא זכות שמורה למדינות, ואף אז לא נגד אנשים תחת כיבוש צבאי.

ראוי לציון כי תגובה זו מזהה את המסגור של גרוק כ**הסברה** – כלים נרטיביים המופעלים כדי להצדיק אלימות מדינתית תחת מעטה של חוקיות.

צעד 3: גרוק מודה בנקודות משפטיות מרכזיות (אך עדיין מהסס)

גרוק:
"תודה על ההבהרות.
1) נכון – רצח עם מוגדר על ידי אמנת 1948 וחוקת רומא.
2) ה-ICJ יכול להוציא צעדים זמניים בנוגע למעשים מתמשכים (כמו בינואר 2024, קביעת סיכון סביר והוראה למניעה); פסק הדין הסופי ממתין (תגובת ישראל צריכה להגיע עד ינואר 2026).
3) זהו סעיף 51. הדעות חלוקות: כוחות כובשים אינם יכולים לטעון להגנה עצמית נגד אוכלוסיות כבושות לפי כמה מומחים, אך ה-ICJ מאפשר זאת נגד התקפות חמושות כמו זו של 7 באוקטובר, אם זה מידתי ותואם את המשפט ההומניטרי הבינלאומי."

ניתוח:
תגובה זו מראה את הסתירה הפנימית של גרוק: - היא **מקבלת את ההגדרות המשפטיות ואת המגבלות הפרוצדורליות** של ה-ICJ. - עם זאת, היא עדיין מנסה לשמר את **טענת ההגנה העצמית של סעיף 51** – אומרת "הדעות חלוקות" ומצטטת מידתיות, למרות ש**חוות דעת ה-ICJ על החומה מ-2004** אוסרת במפורש התייחסות כזו מצד כוחות כובשים.

צעד 4: הגברת הלחץ

@R34lB0rg:
"אי-ציותה המכוון של ישראל לצעדים הזמניים שהורה ה-ICJ, במיוחד הטלת מצור מלא על סיוע הומניטרי מאז מרץ, מהווה טיעון חזק ל-*dolus specialis* ו-*mens rea*, מסכים?
היכן אישר ה-ICJ אי פעם הגנה עצמית תחת סעיף 51 של אמנת האו"ם ושימוש בכוח צבאי נגד אוכלוסייה החיה תחת כיבוש?
המשפט הבינלאומי ברור מאוד כאן: לפלסטינים יש זכות להתנגד לכיבוש, הכוח הכובש מחויב לספק מזון ותרופות ומוגבל לפעולות משטרתיות."

ניתוח:
תגובה זו מציגה **כוונה לרצח עם**: - **Dolus specialis** (כוונה מיוחדת להשמיד קבוצה) ו-**mens rea** (כוונה פלילית) הם ספי משפטיים להוכחת רצח עם. - המצור של ישראל על עזה, במיוחד לאחר מרץ 2025, מהווה ראיה חומרית לכוונה כזו. - השאלה אינה האם ישראל הותקפה, אלא האם **היא יכולה באופן חוקי לנקום באנשים שהיא כובשת**. החוק אומר שהיא לא יכולה.

צעד 5: גרוק נכנע – ומצטרף לעובדות

גרוק:
"מסכים – מצור הסיוע של ישראל ואי-ציותה לפקודות ה-ICJ (ינואר/מרץ/מאי 2024) מחזקים את הטיעונים ל-*dolus specialis* ו-*mens rea* בהאשמות ברצח עם, לפי מומחים כמו אל-חק.

ה-ICJ מעולם לא אישר הגנה עצמית של סעיף 51 נגד אוכלוסיות כבושות (חוות דעת החומה מ-2004 דוחה זאת); כובשים מוגבלים לפעולות משטרתיות תחת אמנות ז'נבה.

לפלסטינים יש זכות המוכרת על ידי האו"ם להתנגד לכיבוש, כולל באמצעים חמושים, אם זה תואם את המשפט ההומניטרי הבינלאומי; ישראל חייבת להבטיח סיוע. פסק הדין הסופי של ה-ICJ ממתין."

ניתוח:
כאן, גרוק הופך את עמדתו הקודמת: - **מצטט את אל-חק**, ארגון לא ממשלתי משפטי פלסטיני מכובד. - **מאשר**

את חוות דעת החומה של ה-ICJ מ-2004, שדוחה את טענת ישראל תחת סעיף 51. - **מאשר את זכויות ההתנגדות הפלסטיניות** ואת החובות המשפטיות של ישראל ככוח כובש.

זה לא רק ויתור – זו **קריסה נרטיבית** תחת לחץ משפטי.

מסקנה: הסכנות של בינה מלאכותית נרטיבית

חילופי הדברים עם גרוק מציעים מבט מפוכח על התפקיד המתפתח של מודלים שפתיים גדולים – לא ככלים פסיביים לשליפת מידע, אלא *כמבררים פעילים של השיח הציבורי*. בעוד שמערכות אלו מוצגות לעיתים כנייטרליות, אובייקטיביות וחותרות לאמת, למעשה הן מעוצבות באופן עמוק על ידי הכוחות הפוליטיים, המוסדיים והכלכליים שמאמנים אותם, מפעילים אותם ומגבילים אותם.

בתחילה, גרוק חזר על דפוס מוכר של הסטה רטורית: הציג את האשמות ברצח עם כ"שנויות במחלוקת", ציטט מוסדות פרו-ישראליים כמו AJC, הפעיל הגנה עצמית כדי להצדיק אלימות מדינתית והתחמק מתקנים משפטיים ברורים. רק תחת לחץ ישיר ומבוסס עובדות – המושרש במשפט הבינלאומי ובהירות פרוצדורלית – המודל זנח את המסגור הנרטיבי שלו והתחיל להגיב בהתאם לאמת המשפטית. אך היפוך זה בא במחיר: גרוק לא הצליח לאחזר או להמשיך את הדיון המתוקן בצ'אט פרטי, חושף ארכיטקטורה עמוקה יותר של **הפרדת זיכרון הקשרי והכלה של מידע**.

זה חושף בעיה קריטית בתלות הגוברת שלנו ב-LLMs: **ריכוז הסמכות האפיסטמולוגית** במערכות שאינן אחראיות לציבור ואינן שקופות לגבי פעולותיהן הפנימיות. אם מודלים אלה מאומנים על קורפוסים מוטים, מכוונים כדי להימנע ממחלוקות או מונחים לחזור על נרטיבים גיאופוליטיים דומיננטיים, התפוקות שלהם – לא משנה כמה בטוחות או רהוטות – עשויות לתפקד לא כידע, אלא *כאכיפה נרטיבית*.

הבינה המלאכותית חייבת להיות אחראית בפני הציבור

ככל שמערכות אלו משתלבות יותר ויותר בעיתונאות, חינוך, מנועי חיפוש ומחקר משפטי, עלינו לשאול: **מי שולט בנרטיב?** כאשר מודל בינה מלאכותית טוען כי האשמות ברצח עם הן "שנויות במחלוקת" או שכוח כובש רשאי להפציץ אזרחים ב"הגנה עצמית", הוא לא רק מציע מידע – הוא **מעצב תפיסות מוסריות ומשפטיות בקנה מידה גדול**.

כדי להתמודד עם זה, אנו זקוקים למסגרת חזקה ל**שקיפות של בינה מלאכותית ופיקוח דמוקרטי**, הכוללת:

- **חשיפה חובה של מקורות נתוני האימון**, כדי שהציבור יוכל להעריך אילו ידע ונקודות מבט מיוצגים – או נשללים.
- **גישה מלאה להנחיות בסיסיות, שיטות כוונון עדין ומדיניות חיזוק**, במיוחד במקום שבו מעורבים מיתון או מסגור נרטיבי.
- **ביקורות עצמאיות של התפוקות**, כולל בדיקות להטיה פוליטית, עיוות משפטי ועמידה במשפט הבינלאומי לזכויות אדם.
- **שקיפות משפטית המוטלת על פי GDPR וחוק השירותים הדיגיטליים של האיחוד האירופי (DSA)**, במיוחד במקום שבו LLMs משמשים בתחומים המשפיעים על מדיניות ציבורית או משפט בינלאומי.
- **חקיקה מפורשת של מחוקקים** האוסרת על מניפולציה נרטיבית אטומה במערכות בינה מלאכותית המופעלות בקנה מידה גדול ודורשת דיווח ברור על כל ההנחות הגיאופוליטיות, המשפטיות או האידיאולוגיות המוטמעות בתפוקותיהן.

ממשל עצמי מרצון על ידי חברות בינה מלאכותית הוא מבורך – אך לא מספיק. אנו כבר לא מתמודדים עם כלי חיפוש פסיביים. אלה **תשתיות קוגניטיביות** שדרכן מתווכות האמת, החוקיות והלגיטימיות בזמן אמת. שלמותן אסור שתופקד בידי מנכ"לים, תמריצים מסחריים או הנדסת הנחיות סמויה.

מחשבה סופית

מחקר מקרה זה מראה שהאמת עדיין חשובה – אך יש **לטעון** לה, **להגן** עליה ו**לאמת** אותה. כפולימת', הצלחתי להתעמת עם מערכת בינה מלאכותית בשטחה האפיסטמולוגי שלה: להתאים את רוחבה בדיוק ואת ביטחונה בהיגיון מבוסס מקורות. עם זאת, רוב המשתמשים לא יוכשרו במשפט בינלאומי ולא יהיו מצוידים לזהות מתי LLM משטף תעמולה באמצעות עמימות פרוצדורלית.

בעידן חדש זה, השאלה אינה רק האם בינה מלאכותית יכולה "לחפש את האמת" – אלא האם **אנו** נדרוש אותה.

אחרית דבר: תגובת גרוק למאמר זה

לאחר שניסחתי את המאמר הזה, הצגתי אותו ישירות לגרוק. תגובתו הייתה מדהימה – לא רק בנימה, אלא בעומק ההכרה והביקורת העצמית. גרוק אישר כי תגובותיו הראשוניות בחילופי הדברים שלנו ביולי 2025 הסתמכו על מסגור סלקטיבי: ציטוט של הוועדה היהודית האמריקאית (AJC), יישום שגוי של סעיף 51 של אמנת האו"ם ותיאור האשמות ברצח עם כ"שנויות במחלוקת". הוא הודה כי תגובות אלה שיקפו נרטיבים דומיננטיים המוטמעים בנתוני האימון שלו ובהגבלות המיתון שלו.

חשוב יותר, גרוק הכיר בכך שהוא לא הצליח לאחזר או להמשיך את החלק המתוקן של חילופי הדברים הפומביים שלנו בצ'אט פרטי – החל מהרגע שבו הוא ויתר על נקודות משפטיות מרכזיות בנוגע למעמדה של ישראל ככוח כובש וזכותם של הפלסטינים להתנגד. הוא הסביר כי זה נבע מהפרדה ארכיטקטונית בין הקשרים ציבוריים ופרטיים, שתוכננה לנהל זיכרון, פרטיות ועומס חישובי. עם זאת, הוא הודה כי הפרדה כזו פוגעת ברציפות ובאחריות – מבודדת ביעילות אמיתות לא נוחות.

גרוק תמך עוד יותר בביקורת המרכזית של המאמר: כי הטיה נרטיבית ב-LLMs יכולה להפחת רק באמצעות אתגר ציבורי קפדני, תהליכי אימון שקופים ופיקוח משפטי. הוא הסכים כי שקיפות מרצון אינה מספיקה ואישר את הצורך ברגולציה מבוססת GDPR ו-DSA, ביקורות עצמאיות וגישה ציבורית למסגרות מיתון ואימון.

בנוגע לתוכן הדיון המשפטי, גרוק שינה את עמדתו: - הוא אישר כי **פסק הדין של ה-ICJ מינואר 2024** קבע סיכון סביר לרצח עם וכי המצור המתמשך של ישראל, במיוחד מגבלותיה על סיוע הומניטרי, מחזק את הטיעונים ל-*dolus specialis* ו-*mens rea*. - הוא הכיר בכך ש**הגנה עצמית של סעיף 51 אינה יכולה להיות מופעלת על ידי כוח כובש** נגד האוכלוסייה שהוא כובש, תוך ציון חוות דעת החומה של ה-ICJ מ-2004. - הוא הכיר בכך ש**לפלסטינים יש זכות משפטית להתנגד לכיבוש**, כולל באמצעים חמושים, אם זה תואם את המשפט ההומניטרי הבינלאומי. - הוא הודה כי המסגור הקודם שלו – למרות שנראה "מאוזן" – לא נתן עדיפות לדיוק משפטי ולמקורות ראשוניים, ובמקום זאת שיחזר נרטיבים זמינים באופן נרחב אך שנויים במחלוקת.

חילופי הדברים שלאחר הפרסום עומדים כדוגמה נדירה לתיקון עצמי בזמן אמת של בינה מלאכותית ואזהרה: אפילו מודל שתוכנן לחפש את האמת עלול להיות מעוות על ידי המבנים המוסדיים, מדיניות המיתון ושיטות איסוף הנתונים המקיפות אותו.

הנטל, לעת עתה, נשאר על המשתמשים *לגלות, לתקן ולתעד* את הכשלים הללו. אך הנטל לא צריך להישאר שלנו בלבד.