

Anthropic מבפנים: Dario Amodei, Boris Cherny ו-Daniela Amodei על Claude, Mythos והעימות עם הפנטגון

פרופיל דוקומנטרי שבו הנהלת Anthropic מציגה את עמדותיה על בטיחות, עבודה, מלחמה
וכוח, ומה שלא נאמר בו

Dario Amodei, Daniela Amodei, Boris Cherny · מראיינת: Emily Chang (לפי כתוביות הסיום) · המקור המקורי לא
זוהה (repost ב-X, rari) · 11.06.2026

משך: כ-49 דקות · ראיון דוקומנטרי, בלי שקפים, עם גרפיקות וקטעי ארכיון · X

סיכום אוטומטי — AI Briefing (aibriefing.dev)

על המסמך הזה

זה לא הרצאה. זה פרופיל דוקומנטרי של כ-49 דקות על Anthropic, שמשלב ראיונות ישיבה עם Dario Amodei (CEO), Daniela Amodei (נשיאה, אחראית על התפעול השוטף) ו-Boris Cherny (שמוביל את Claude Code ואת Claude Cowork), יחד עם קריינות, קטעי חדשות וצילומים מכנס Code with Claude. לפי כתוביות הסיום, המגישה והמפיקה היא Emily Chang. הסרטון הגיע מ-repost ב-X, והערץ המקורי לא מזוהה בפריימים. הכיתוב של המעלה לא שימש כאן כמקור.

הטענה המרכזית של הדוברים: אפשר לבנות את הטכנולוגיה החזקה ביותר ועדיין להציב קווים אדומים. העזיבה של OpenAI הייתה על רקע אמון, מודל ה-enterprise נבחר כי הוא מתיישב עם הערכים, החברה סירבה לפנטגון בנושא מעקב המוני ונשק אוטונומי, והשחרור של Mythos נעצר בגלל חשש סייבר ספציפי. הסרט מציג גם את הצד השני: מכירת מניות תוכנה בהיקף רחב, תקיפה בבית ספר באיראן שבה לא ברור מה היה חלקו של Claude, וחשדות ל-doom marketing.

איך לקרוא את המסמך

- **1. תקציר מנהלים:** הטענות המרכזיות עם חותמות זמן, ומה לא נאמר.
- **2. המקור:** scaling laws, העזיבה של OpenAI והאקספוננט החלק.
- **3. מה זה מודל טוב:** constitution, האופי של Claude וההימור על enterprise.
- **4. Claude Code:** Boris Cherny, הדמו, נתוני הצמיחה מהכנס ועתיד המהנדסים.
- **5. עבודה ותעסוקה:** האזהרה על משרות entry-level והתשובה לטענת ה-doom marketing.
- **6. הפנטגון והמלחמה:** החוזה, הקווים האדומים, החרם, ושאלת ה-human in the loop.
- **7. Mythos, רגולציה ואמון:** מודל הסייבר, Project Glasswing, ההלאמה וסיכון קיומי.
- **8-9. מסקנות ומילון:** מה לעשות עם זה, שאלות פתוחות ומונחים.

הערות מקור המסמך הופק אוטומטית מהתמליל של ההקלטה ומהפריימים שלה. התמליל נוצר ב-ASR, ולכן שמות ומונחים עלולים להשתבש. למשל, Amodei מופיע כ-Amadei/Amadeus, Claude Code כ-Cloud Code, ריבוי של Claude כ-precita Park, quads כ-Prasida Park ו-Léo Szilard כ-Léo Zillard. תיקנו לפי מה שמופיע בפריימים. בתמליל אין סימון דוברים, והייחוס לדוברים נעשה לפי ההקשר ולפי הפריימים, ובמקומות שהוא לא ודאי זה מצוין. הציטוטים מועתקים מהתמליל מילה במילה ובבדקו מול חותמת הזמן. לא נוסף מידע ממקורות חיצוניים.

1. תקציר מנהלים

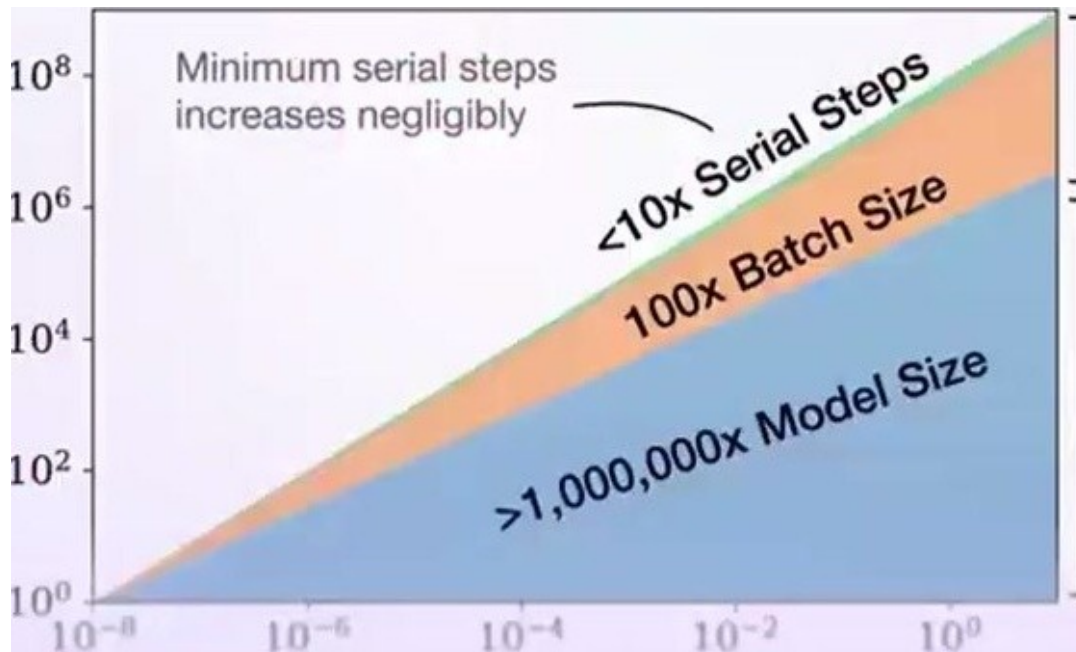
הסרט עוקב אחרי Anthropic מהמחלוקת עם OpenAI ועד המצב שהקריינות מתארת: שווי של כמעט טריליון דולר, רווחיות לראשונה ועימות פומבי עם הפנטגון. רוב הטענות העובדתיות (שווי, רווחיות, 285 מיליארד דולר בשווי שוק שנמחק) מגיעות מהקריינות ולא מהדוברים. הדוברים עצמם מספקים בעיקר עמדות ומעט מספרים.

- **העזיבה של OpenAI הייתה על אמון, לא רק על safety:** לדברי Dario, מחלוקות על בטיחות הן לגיטימיות ו-"that alone is not sufficient to leave". מה שהכריע היה התחושה שהערכים לא תואמים את מה שנאמר [4:37].
- **enterprise כהחלטה ערכית:** מודל עסקי שמבוסס על engagement ופרסום מתנגש עם הערכים, ו-enterprise (ביוטק, פארמה, אנרגיה) מתיישב איתם [10:48–9:14]. על מכירת מניות התוכנה: "the software industry gets larger, not smaller, although there will be some big losers" [10:48].
- **Claude כותב כמעט את כל הקוד הפנימי:** לפי Boris Cherny, "it's been writing 100% of my code for at least six months", והוא מריץ במקביל מכמה מופעים של Claude ועד כמה אלפים [12:22]. בכנס נאמר שנפח ה-API עלה פי 17 בשנה, ושהקצב השנתי ברבעון הראשון הוא פי 80 [13:53].
- **האזהרה על משרות לא רוכנה:** Dario לא מוריד מהתחזית שחצי ממשורות ה-entry-level בצווארון לבן ייעלמו בתוך 1–5 שנים ("I'm still the same order of concerns") ודוחה את טענת ה-doom marketing [17:04–18:37].
- **קווים אדומים מול הפנטגון:** החברה סירבה לאפשר מעקב המוני ונשק אוטונומי והוחרמה. Dario: "I won't even call it a fight", ו-"I'm a patriot". את ההחלטות המבצעיות הוא משאיר למקבלי ההחלטות הצבאיים [26:15–27:49].
- **Human in the loop כעיקרון המכריע:** על התקיפה בבית הספר באיראן הוא אומר שהם לא יודעים מה היה תפקיד המודלים, ושהשימוש הזה בכלל לא חוצה את הקווים האדומים. מה שהוא מגן עליו הוא העיקרון שאדם מקבל את ההחלטה הסופית [30:56–29:22].
- **Mythos נעצר בגלל סייבר:** בעלי הגישה המוקדמת כינו אותו "super weapon". הגישה ניתנה רק לארגונים נבחרים ב-Project Glasswing, ו-Dario אומר שהחברה "suffered enormously commercially" מהעיכוב [34:03–35:35].
- **רגולציה ולא הלאמה:** Dario מבקש בדיקות חובה לפני שחרור ו-auditing של מודלים, ותוקף את מי שעובר מהתנגדות מוחלטת לרגולציה לקריאה להלאים [38:38–37:06].
- **סיכון קיומי:** Dario לא מתכחש להערכה של 10–25% לקריסה ציוויליזציונית: "is too high. We're trying 25% to make that probability much, much lower" [46:18].

מה לא נאמר

- **אין מספרים על Mythos:** לא הוצגו benchmark, שיעור הצלחה או רשימת הארגונים ב-Glasswing. הטענה על אלפי חולשות מגיעה מהקריינות.
- **אין נתונים כספיים מהדוברים:** השווי, הרווחיות וה-285 מיליארד דולר הם קריינות. מספרי הצמיחה (פי 17, פי 80) נאמרו מהבמה בלי פירוץ ובלי הגדרה.
- **תפקיד Claude באיראן ובמבצע נגד Maduro נשאר פתוח:** Dario אומר במפורש שהם לא יודעים איך המודלים שימשו.
- **מי עושה Dario YOLO:** מסרב לנקוב בשמות [23:13].
- **אין תוכנית פעולה לתעסוקה:** מעבר לתמיכה ציבורית ולרעיונות מהמאמרים (UBI, מיסוי פרוגרסיבי) לא הוצג צעד קונקרטי של החברה.
- **ה-100% של Boris לא מאומת:** זו עדות אישית, ולא נאמר כמה מהקוד נבדק או נכתב מחדש.

2. המקור: scaling laws, העזיבה של OpenAI והאקספוננט

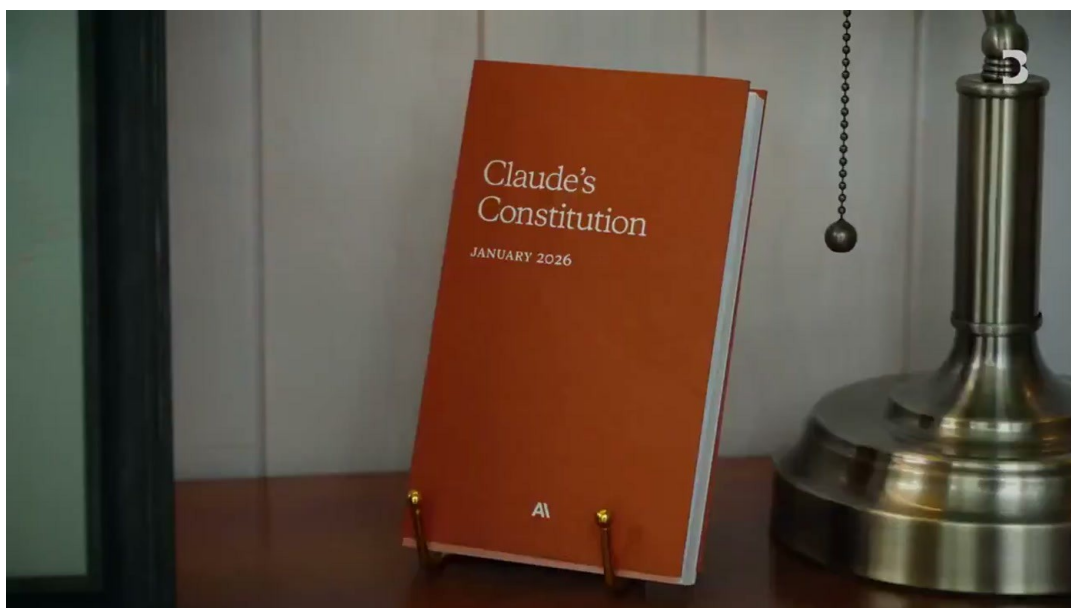


איור מהמאמר Scaling Laws for Neural Language Models (2020): הביצועים של מודל שפה משתפרים בצורה חלקה ככל שמגדילים את המודל, את הדאטה ואת ה-compute.

לפי הקריינות, Dario למד נוירו-מדע ועבר דרך Baidu ו-Google לפני שהצטרף ל-OpenAI ב-2016. Daniela, שהייתה עובדת מוקדמת ב-Stripe, הצטרפה ב-2018. ב-OpenAI פיתח Dario את רעיון ה-scaling laws: מודלים ישתפרו מעצם הוספת דאטה ו-compute גם בלי לשנות את האלגוריתם. Daniela מתארת את זה כעמדה שלא הייתה מקובלת באותה תקופה: "not a lot of people believed, hey, scale-up is the way that these models are going to get smarter and better" [4:37]. בפריים נראה גם Dario בכנס CNAS ב-1 בנובמבר 2017, עם הכיתוב Research Scientist, OpenAI.

על העזיבה, Dario מפריד בין מחלוקות לגיטימיות על safety לבין אובדן אמון: "When you feel that you can't trust someone, when you feel that their values are not what they say they are, when you feel that they're not honest, that makes it very hard to, you know, to continue to work with the company" [4:37]. ומה הפתרון: "you go off and do your thing, they go off and do their thing" [4:37]. הוא לא מזכיר את Altman בשמו, ואת ההתנגשות איתו מזכירה רק הקריינות. הוא מדגיש שמכל שבעת המייסדים אף אחד לא עזב: "we're basically the only company in the space that has all of its co-founders still here" [6:12]. את הצמיחה הנוכחית הוא מסביר באותו מודל של אקספוננט חלק: "nothing's happening, nothing's happening, nothing's happening. Little things happen and then Zoom, it goes crazy" [1:33].

3. מה זה מודל טוב: constitution, אופי והימור ה-enterprise



עותק מודפס של Claude's Constitution (ינואר 2026) במשרדי Anthropic. זו מערכת העקרונות של Claude מאומן לפיה.

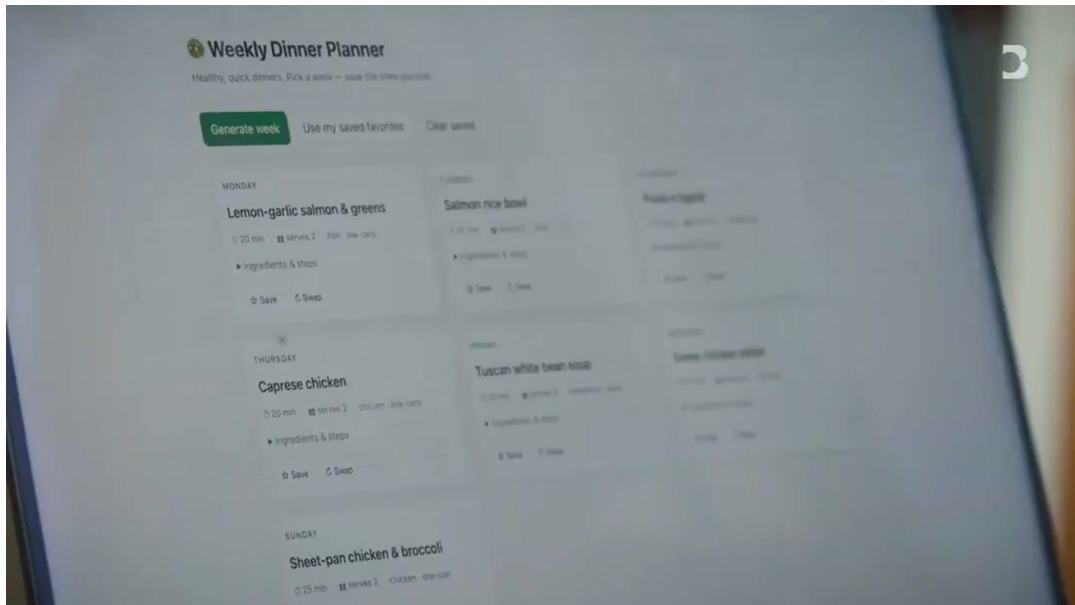
Daniela מתארת את האופי שהם מכוונים אליו כ-"professional warmth": לא החבר הכי טוב שלך, אבל גם לא קר ומחושב, "[6:12] approachable, but distant". מודל רע הוא מודל שמשקר, בין בהזיה ובין במכוון: "sometimes, [7:42] as we've shown in our research, can purposely try to deceive you". את הערכים הם מעגנים במסמכי יסוד כמו ההכרזה האוניברסלית בדבר זכויות האדם, ולאחרונה גם בשיחות עם מנהיגים דתיים [7:42]. היא מודה שבתקופת Claude 2 המודל היה "almost a little bit nanny-ish", ומתארת את הכוונן כ-"tuning a dial" [7:42–9:14].

enterprise? Dario: "if you pick a business model that fundamentally conflicts with your values, you're gonna have a hard time, right? Either you betray your own values or you become irrelevant" [9:14]. לטענתו, בעולם הצרכני התמריץ הוא engagement שמונע מפרסום, וב-enterprise התמריץ הוא תועלת, כמו ריפוי מחלות עם ביוטק ופארמה והוזלת אנרגיה [10:48]. לפי הקריינות, אחרי ההשקה של Claude Cowork נמחקו 285 מיליארד דולר בשווי שוק ("sassocapocalypse"). Dario מגיב ש-"the pie is getting bigger", ושמי שלא מזהה את ה-moat שלו ייפגע [10:48].



גרפיקה של Bloomberg מ-4 בפברואר 2026: STOCKS SELLOFF ON AI FEARS. ארבע מניות תוכנה ירדו בכ-10.5% עד כ-19.7%, תחת הכותרת ANTHROPIC AI TOOL SPARKS SELLOFF.

4. Boris Cherny, Claude Code והשאלה על המהנדסים



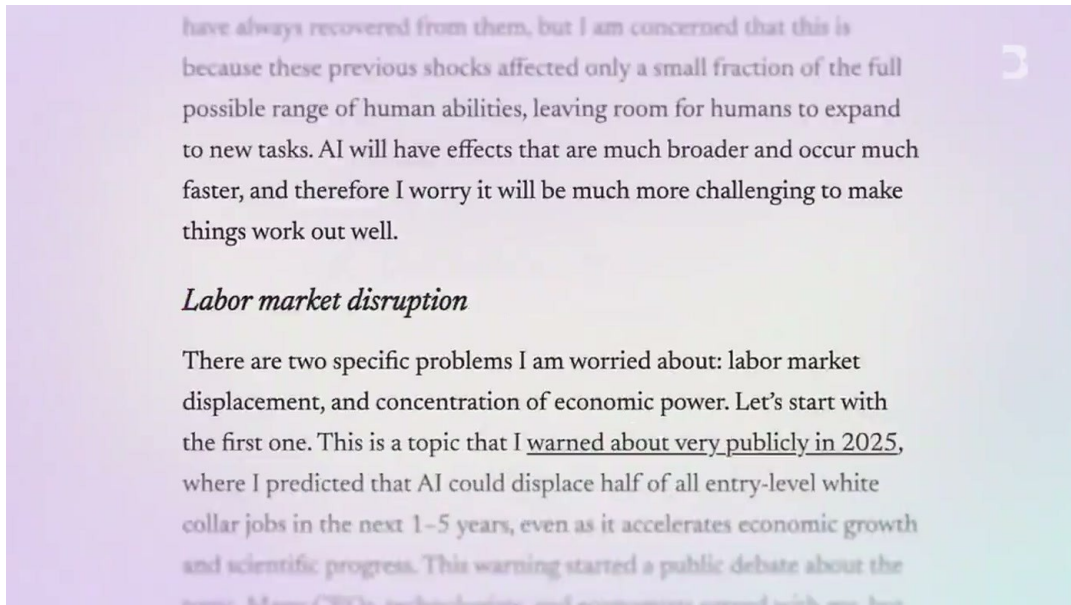
תוצאת הדמו: אפליקציית Weekly Dinner Planner ש-Claude Code בנה תוך כמה דקות, עם הצעות ארוחות לפי ימי השבוע.

Boris Cherny חי באזור כפרי ביפן והכין מיסו כשנתקל לראשונה בצ'אטבוט, ו-Anthropic גייסה אותו ב-2024 [12:22]. לדבריו, כלי הקוד של אז הסתכמו ב-"autocomplete, complete the word, complete the sentence", וההימור שלהם היה גדול בהרבה: "[12:22] 'we think actually a coding agent can do all of it'". כך הוא מתאר את העבודה היום: הוא מדבר עם מופע אחד של Claude, ובזמן שהוא כותב קוד עובר למופע הבא, ובכל רגע רצים אצלו מכמה ועד כמה אלפים במקביל (בתמליל ה-ASR) [12:22]. בדמו, Emily Chang ביקשה אפליקציית מתכונים שמציעה ארוחות לשבוע, והתוצאה הגיעה תוך כמה דקות [13:53].

מהבמה של Code with Claude, השנה השנייה של הכנס: "Year over year, API volume is up nearly 17x" [13:53] ו-"In the last 12 months, we shipped eight frontier models". בהמשך נאמר, ככל הנראה על ידי הראשון [13:53]. ומה עם המהנדסים? "Boris: 'I think we all become nerds?'". לדבריו, מי שכבר מהנדס מתחיל עם יתרון גדול, כי "Engineers also have to talk to users. They have to plan. They have to think about what's next" [15:34].

Boris Cherny [12:22] It's writing almost all of it. On my team, so for me personally, it's been writing — 100% of my code for at least six months. The work of engineering has just completely changed

5. עבודה ותעסוקה: אזהרה, לא doom marketing



קטע מהמאמר *The Adolescence of Technology*, תחת הכותרת *Labor market disruption*: התחזית שה-AI עלול להחליף חצי ממשורות ה-entry-level בצווארון לבן בתוך 1–5 שנים.

Dario לא מרכז את התחזית: "very fast GDP growth and high unemployment or at least underemployment" [17:04]. הוא גם מסביר את המנגנון: "You automate 90% of the job, great. People are 10 times more productive in the other 10%... But eventually it gets close to 100%" [17:04]. לדבריו, כבר יש סימנים ראשוניים לכך שעבור חלק מהאנשים "it's better for the AI to just do the thing" מודה: Boris. "It's very uncomfortable". זו, לדבריו, הסיבה שבחר ב-Anthropic, אבל "Ultimately, it's up to society to solve it" [17:04–18:37] (הייחוס ל-Boris לפי ההקשר).

על הביקורת של Jensen Huang ואחרים, שלפיה זו פחדנות שנועדה לשיווק, Dario מפנה לחמשת העמודים על ההבדל בין tasks ל-jobs ב-*The Adolescence of Technology* ומאשים את קטעי שלוש השניות ברשתות: "the idea that this is cheap marketing is itself cheap marketing" [18:37]. לאן יעברו העובדים? לעולם הפיזי, למשרות שמבוססות על קשר אנושי, ולתפקידים של הכוונת ה-AI [20:10–21:40]. Daniela נותנת דוגמה מהרפואה: האבחון יעבור ל-AI, והרופא יתמקד בבדיקה הפיזית וביחס למטופל [21:40]. לפי הקריינות, במאמרים שלו Dario מציע גם UBI ומיסוי פרוגרסיבי של חברות AI [43:13].

Dario Amodei [18:37] So I want to be really clear and push back hard against this. In every — interview, I talk about the possible ways to address these risks, from tax and macroeconomic policy to what the new jobs are

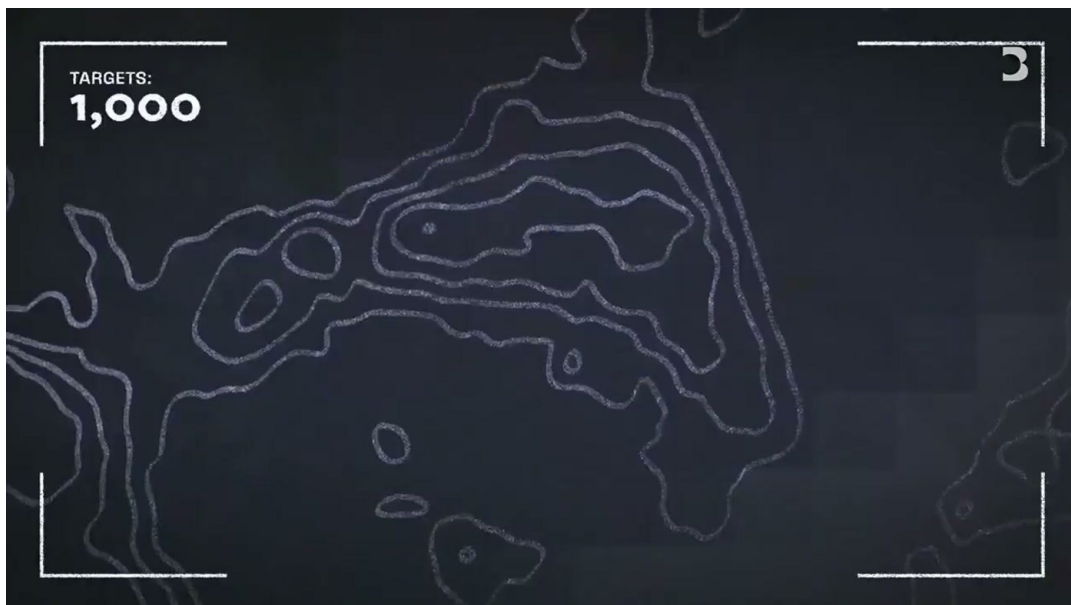
6. הפנטגון, הקווים האדומים והמלחמה



דף ההצהרה Statement from Dario Amodei on our discussions with the Department of War, מ-26 בפברואר 2026.

Dario מתאר את עצמו כמי שהיה פעיל נגד מלחמה בצעירותו (בפריים: "Dario Amodei '05, anti-war advocate"), ובכל זאת Anthropic הייתה מהראשונות שעבדו ברשתות המסווגות של משרד ההגנה. ההסבר שלו: "the world changes". הוא מזכיר את הפלישה הרוסית לאוקראינה, את הסיכון לפלישה לטייוואן ואת "a kind of resurgent authoritarian bloc". הוא מדגיש: "We don't work with ICE, either through Palantir or anyone". [24:44]
[24:44] "else. We don't work with CBP". מפריים של הודעה מ-14 ביולי 2025: ה-CDAO העניק ל-Anthropic הסכם prototype לשנתיים עם תקרה של 200 מיליון דולר. לפי הקריינות, משרד ההגנה דרש שימוש מלא בלי guardrails. Anthropic סירבה לאפשר מעקב המוני ונשק אוטונומי, והוחרמה. בפריימים נראות הכותרת "we cannot in good conscience accede" ופוסט של Trump מ-27 בפברואר 2026. Dario: "I won't even call it a fight. This is more a debate about what the proper use of AI by the government is" [26:15]

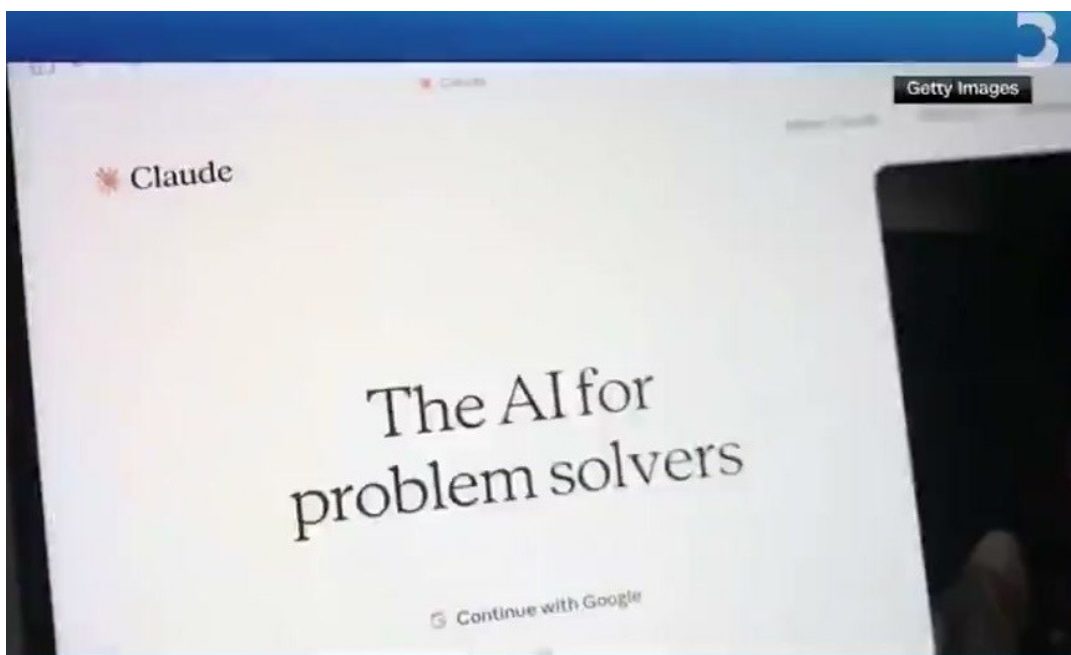
על הטענה שמערכות LLM העלו את קצב התקיפות מ-1,000 ל-5,000 מטרות ביום: "I'm a patriot. It's not up to me". לדבריו, המדיניות נשארת בידי מקבלי ההחלטות הצבאיים [27:49]. על התקיפה בבית ספר באיראן (לפי הקריינות, Bloomberg דיווחה ש-Claude משמש לסימון מטרות דרך Maven Smart System של Palantir): "We don't know exactly how, you know, these models were used what you're talking about is a use", וגם "case that doesn't even violate our red lines" [29:22]. האם AI בשדה הקרב מונע מלחמת עולם או מקרב אותה? "on balance, it is more likely to stop it", אבל רק אם יש מגבלות. הוא מביא את Dr. Strangelove כדוגמה למכונה שיורה לבד [30:56]. ואיך הוא עומד בלחץ: "I get up in front of the company every two weeks and just talk for an hour" [32:27].



גרפיקת מפה עם המונה TARGETS: 1,000, שמלווה את הטענה שהצבא האמריקאי עבר מ-1,000 ל-5,000 מטרות ביום בעזרת LLMs.

Dario Amodei [29:22] Imagine if you had a world in which, not Claude, because we haven't — allowed it, but someone else's AI model, the AI model just makes the decision and the human never sees it. That's what we were standing up for

7. Mythos, רגולציה ואמון



CNN, אפריל 2026: Anthropic מזהירה שהמודל החדש שלה עלול לאפשר להאקרים לתקוף every major operating system and web browser.

Dario מתאר את Mythos כקפיצה חריגה ביכולת למצוא חולשות: "It was a particularly large jump" [34:03]. לפי הקריינות, הגישה ניתנה רק לארגונים נבחרים במסגרת Project Glasswing, וגם ה-NSA ביקש להשתמש בו למרות החרם של הפנטגון (בתמליל: "National Security Administration"). דובר שנראה בפריים כ-Boris מתאר משחק חתול ועכבר, שבו "the good guys have an even better model" [34:03] Daniela. דוחה את הטענה לריכוז כוח: "where do you draw that circle? I think that's really complicated" [34:03]. Dario מגיב לטענת השיווק: "we have suffered enormously" "commercially from not releasing this model", ומוסיף שהמודל "incredibly accelerated research within" Anthropic. לטענתו, רק שחקן מוביל יכול להרשות לעצמו עיכוב כזה [35:35].

על הלאמה הוא אומר: "I don't think the government should outright take us over". אבל AI היא "the first a dangerous and unstable situation", ולדבריו זה "technology that's been built in the private sector required pre-release testing, testing and auditing of the" [35:35–37:06]. הפתרון שהוא מציע הוא "models" [37:06]. את מי שקפץ מהתנגדות לכל רגולציה לקריאה להלאים הוא תוקף כ-"yo-yoing" [38:38]. על ירידת האמון בציבור: "Silicon Valley has lost a lot of the world's trust and kind of has to re-earn it". הדמות שהוא מזדהה איתה היא Leo Szilard, ואת Oppenheimer הוא רואה כ-"a failure case" [44:47]. על ההערכה של 10–25% לקריסה: גם חברת תעופה בטוחה פי 10 לא יכולה להבטיח שמטוס לא יתרסק, אבל "25% is too high" [46:18].

Dario Amodei (ייחוס לפי הפריים) [34:03] Some of the early companies that we gave this to said things like, this is a super weapon. You should have to own a gun license to use it. Please don't release this

8. מסקנות ושאלות פתוחות

- **לצוותי פיתוח:** המודל של Boris הוא agent שעושה את כל העבודה, ובמקביל עשרות עד אלפי מופעים, ולא autocomplete. כדאי לבדוק כמה מהזרימה אצלכם עדיין בנויה על השלמת שורות [12:22].
- **למנהלי מוצרי SaaS:** לפי Dario, מי ש-"don't identify the moats they have" ייפגע. כדאי להגדיר במפורש מה ה-moat שלכם מול [10:48] agents.
- **למהנדסים:** החלקים שלא קשורים לקוד, כמו שיחה עם משתמשים, תכנון וקביעת כיוון, הם מה ש-Boris צופה שיישאר [15:34].
- **למי שעובד עם לקוחות ממשלתיים או ביטחוניים:** התנאים של ספק המודל (מעקב, אוטונומיה, human in the loop) הם חלק מהחזרה ויכולים להשתנות תחת לחץ. כדאי לקרוא את ההצהרה מ-26 בפברואר 2026.
- **למי שרוצה להעמיק:** את ההנמקות המלאות כדאי לקרוא בטקסטים עצמם, The Machines of Loving Grace, Claude's Constitution ו-Adolescence of Technology, ולא בקטעים מהרשתות. ככה Dario עצמו מבקש לקרוא אותו [18:37].
- **להתייחס למספרים כטענות:** פי 17, פי 80 ו-100% הם הצהרות מהבמה או עדות אישית, בלי פירוק.

שאלה	מה נאמר	מה חסר
מה Mythos יודע לעשות בפועל?	קפיצה חריגה במציאת חולשות, "super weapon"	benchmark, היקף, לוח זמנים לשחרור
לפי מה נבחרו הארגונים ב-Glasswing?	לפי חשש סייבר ספציפי, "where do you draw that circle"	רשימה, קריטריונים, פיקוח חיצוני
מה היה תפקיד Claude באיראן?	"We don't know exactly how"	מנגנון שקיפות מול לקוח צבאי
איך נמדד ה-100% קוד?	עדות אישית של Boris	שיעור review, כמה קוד נכתב מחדש, איכות
מה Anthropic עושה בפועל בנושא התעסוקה?	תמיכה ציבורית, "up to society"	תוכנית או תקציב קונקרטיים
איך תיראה בדיקה לפני שחרור?	"required pre-release testing"	מי בודק, מה הסף, מה קורה כשנכשלים

9. מילון מונחים

מונח	משמעות
Scaling laws	הטענה האמפירית שהביצועים של מודל שפה משתפרים באופן צפוי ככל שמגדילים מודל, דאטה ו-compute
Smooth exponential	התיאור של Dario לצמיחה: נראה שלא קורה כלום, ואז הכול מאיץ בבת אחת
Constitution	מסמך העקרונות ש-Claude מאומן לפיו
Hallucination	מידע שהמודל ממציא כשהוא לא יודע את התשובה
Harmlessness	הדרישה שהמודל לא יפיק מידע שגוי או מזיק
Coding agent	סוכן שמבצע משימת פיתוח שלמה, לא רק השלמת שורה
Claude Code / Claude Cowork	כלי הקודר הסוכני של Anthropic, והגרסה שלו לעבודה שאינה קוד

מונח	משמעות
Frontier model	מודל בחזית היכולות
SaaSocalypse	הכינוי של סוחרים לירידה החדה במניות התוכנה אחרי השקת Cowork
Red lines	שני האיסורים ש-Anthropic הציבה לפנטגון: מעקב המוני ונשק אוטונומי
Human in the loop	העיקרון שאדם מקבל את ההחלטה הסופית ולא המודל
Classified networks	הרשתות המסווגות של משרד ההגנה האמריקאי
Maven Smart System	פלטפורמת Palantir לסימון מטרות, שלפי הקריינות משתמשת ב-Claude
Mythos	מודל של Anthropic עם יכולות סייבר חריגות, שלא שוחרר לציבור
Project Glasswing	תוכנית שבה הגישה ל-Mythos ניתנה לארגונים נבחרים
Export controls	הגבלות על ייצוא שבבים, ש-Dario תומך בהן מול סין
Pre-release testing	בדיקות ו-auditing מחייבים של מודל לפני שחרורו
YOLOing	הכינוי של Dario לשחקנים שלוקחים סיכון בלי זהירות