

Andrej Karpathy על ה-cognitive core, חברות של LLMs ועתיד החינוך

ראיון בפודקאסט No Priors: למה מודלים יכולים להיות קטנים בהרבה, איך תיראה "חברה" של מודלים, ואיך AI אמור להגדיל את טווח ההגעה של מורה אחד

Andrej Karpathy, ראיון למגישי No Priors · הקטע הופץ מחדש ב-X על ידי Morty · הועלה ב-2026-10-05 (תאריך ההקלטה המקורית לא מופיע בהקלטה)

משך: כ-17 דקות · X

סיכום אוטומטי — AI Briefing (aibriefing.dev)

על המסמך הזה

זה קטע של כ-17 דקות מתוך ראיון ש-Andrej Karpathy נתן לפודקאסט No Priors. מקור הפודקאסט ברור מהסיום ("Find us on Twitter at NoPriorsPod") ומכרטיס הסיום, שבו מופיעים השם Andrej Karpathy, הכיתוב "from OpenAI and Tesla" והלוגו של No Priors. בשיחה משתתפים שני מגישים, איש ואישה, אבל השמות שלהם לא נאמרים בקטע ולכן לא נקבע כאן מי הם. הקטע הגיע אלינו דרך repost ב-X. הכותרת והתיאור של ה-repost, כולל ציטוט על agents ו-"distillation at scale", לא נלקחו כמקור, כי הם לא מופיעים בתמליל.

הטענה המרכזית מורכבת משני חלקים. החלק הטכני: המודלים של היום מבזבזים חלק גדול מהקיבולת שלהם על שינון, ואפשר לזקק "cognitive core" קטן מאוד, אולי פחות ממיליארד פרמטרים, שיודע לחשוב ולהשתמש בכלים. סביב ליבות כאלה צפויות לקום "חברות של LLMs" עם תפקידים, היררכיה ו-escalation. החלק השני, שתופס את רוב הזמן, הוא חינוך. Karpathy מגדיר את עצמו "team human" ומתאר את Eureka, המיזם שלו: המורה מתכנן את הקורס ב-back end, וה-AI הוא ה-front end מול הסטודנט. בהמשך הוא מדבר על תרבות כמשתנה הדומיננטית במוטיבציה, על חינוך כ"חדר כושר למוח" בעולם post-AGI, ועל מה ילדים צריכים ללמוד.

איך לקרוא את המסמך

- 1. **תקציר מנהלים:** הטענות המרכזיות, כל אחת עם הראיה ו-timestamp, ומה לא נאמר.
- 2. **ה-cognitive core:** למה מודל יכול להיות קטן, ומה התפקיד של distillation.
- 3. **חברות של LLMs:** ריבוי מודלים, היררכיה, swarm ו-escalation.
- 4. **Eureka:** המורה ב-back end, ה-AI ב-front end, ולמה דווקא עכשיו.
- 5. **התאמה אישית:** low-hanging fruit מול מה שעוד רחוק, ו-"the demo is near, but the product is far".
- 6. **תרבות, מוטיבציה ו-lineage: post-AGI,** אשכולות, אלילים קהילתיים, וחינוך כחדר כושר.
- 7. **מה ללמוד:** מתמטיקה, פיזיקה ו-CS, והחלוקה של 80%.
- 8-9. **מסקנות ומילון:** מה לעשות עם זה, שאלות פתוחות, ומונחים.

הערות מקור המסמך הופק אוטומטית מתמליל ההקלטה ומהפריימים שחולצו ממנה. אין בהקלטה שקפים, כל הפריימים הם צילומי אולפן. התמליל נוצר ב-ASR, ולכן שמות ומונחים עלולים להיות משובשים. לדוגמה, "Noble Lawrence" הוא כנראה "Nobel laureates", "lohing fruit" הוא low-hanging fruit, ו-"simple manipulation" מופיע בכתוביות המוטמעות בווידיאו כ-symbol manipulation. הציטוטים מועתקים מילה במילה מהתמליל ונבדקו מול ה-timestamp שלהם. התמליל לא מסמן דוברים, ולכן כשהשיוך לא חד-משמעי זה מצוין. לא נוסף מידע ממקורות חיצוניים.

1. תקציר מנהלים

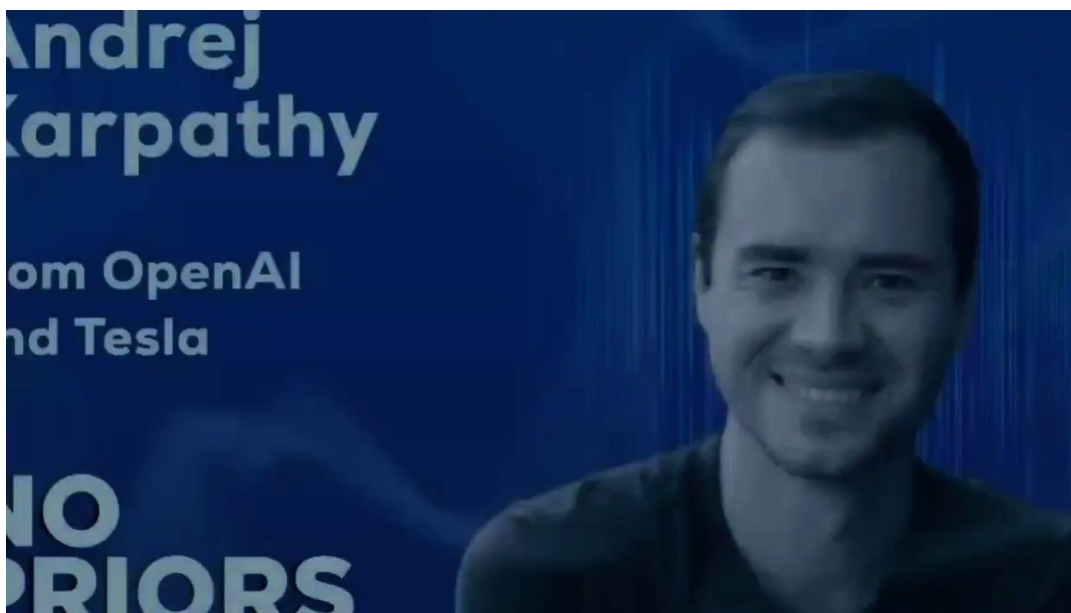
זה לא הרצאה טכנית עם נתונים אלא שיחה. Karpathy מציג אינטואיציות ותחזיות, ולא תוצאות או ניסויים. בהתאם לכך, הנקודות למטה הן עמדות של הדובר עם ההנמקה שהוא נותן להן, ולא ממצאים שהוכחו.

- **מודלים מבזבזים קיבולת על שינון:** לדבריו המודלים זוכרים SHA hashes ופרטים שוליים כי ה-dataset לא נאסף בקפידה, ו-"I think this will go away" [0:00].
- **ה-cognitive core יכול להיות זעיר:** על השאלה אם מדובר ב-3 או ב-20 מיליארד פרמטרים, התשובה היא "I think even a billion is too much" [0:00, 1:31].
- **distillation הוא הראיה העיקרית:** "Distillation works surprisingly well". מודל גדול שמפקח על מודל קטן מצליח לדחוס בו הרבה יכולת [0:00].
- **רוב האינטרנט הוא מידע ולא חשיבה:** "cognition and like 99.99% of like information 0.001%". זה ניסוח אינטואיטיבי ולא מדידה [0:00].
- **העתיד הוא חברות של LLMs ולא מודל אחד:** מקביליות, היררכיה, מודלים שמתמחים בתחומים, ו-escalation אוטומטי לפי רמת הקושי. ה-CEO הוא מודל חזק בענן, והעובדים זולים, אולי [1:31] open source.
- **ב-Eureka ה-AI הוא ה-front end ולא היוצר:** לדעתו המודלים עדיין לא מספיק טובים כדי לבנות קורס, אבל הם כבר מספיק טובים כדי להוביל סטודנט דרך קורס שמורה בנה [4:37].
- **ה-low-hanging fruit הוא תרגום ו-reach:** התאמה לרקע האישי של הלומד אפשרית, אבל לא קלה: "the demo is near, but the product is far" [6:12, 7:46].
- **תרבות היא המשתנה הדומיננטי:** תוכן לימודי כבר זמין בשפע, ומה שקובע הוא מה הקהילה מעריצה, למשל H-index באקדמיה או התמקדות בהתמחויות בטורונטו [9:18, 10:51].
- **לוח זמנים:** הקורס הראשון של Eureka מיועד לרמת תואר ראשון טכני. היעד היה "late this year", וכעת הוא "probably early next year". השנה לא נאמרת [13:55].

מה לא נאמר

- **אין מספרים מוכחים:** 1B פרמטרים ו-0.001% הם הערכות בעל פה. לא מוצגים benchmark, ניסוי או ניסוח מתמטי. כשמגיש שואל אם יש לזה ייצוג מתמטי, התשובה בפועל היא "maybe there's no good way to represent that".
- **הכיתוב של ה-repost לא מופיע בשיחה:** המשפט "Agents aren't magic. They're distillation at scale" לא נאמר. המילה agents כמעט לא עולה, ו-distillation מוזכר רק כטיעון בעד מודלים קטנים.
- **אין ארכיטקטורה ל-swarm:** לא מוסבר איך מתבצע ה-routing, איך נמדד הקושי, או איך נשמר state בין המודלים.
- **לא מוסבר איך Eureka בנוי:** אין פרטים על מודל, על תמחור, על הערכת לומדים, או על איך יודעים ש-"it actually works".
- **אין תאריך מקורי:** הרמזים היחידים הם "we just had the Olympics like a month ago" ויעד ההשקה היחסי של הקורס.

2. ה-cognitive core: מודל קטן שחושב ומשתמש בכלים



כרטיס הסיום של הפרק, היחיד שמזהה את הדובר: Andrej Karpathy, from OpenAI and Tesla, עם הלוגו של No Priors. שאר הפריימים הם צילומי אולפן בלבד.

Karpathy פותח בביקורת על מה שהמודלים מנצלים בו את הפרמטרים שלהם. מכיוון שה-dataset לא נאסף בקפידה, חלק ניכר מהקיבולת הולך לשינון של דברים שלא חשובים לחשיבה. מה שצריך לבנות במקום זה ליבה שחושבת, ושכשחסר לה מידע, יודעת לפנות לכלים.

Andrej Karpathy [0:00] And I think we just need to get to the cognitive core. And I think the cognitive core can be extremely small. And it's just this thing that thinks. And if it needs to look up information, it knows how to use different tools

הוא מבסס את הטענה על distillation: "Distillation is where you get a really big model or a huge amount of computers, something like that, supervising a very small model. And you can actually stuff a lot of information theory, capability into a very small model" [0:00]
Karpathy עונה בדימוי ולא בנוסחה: "The internet is like 0.001% cognition and like 99.99% of like information" [0:00]. כלומר הנתונים מכילים בעיקר עובדות, ורק מעט מהם הוא הדבר שמלמד לחשוב.

לגבי הגודל, הוא מתחיל ב-"I think even a billion" [0:00]. בהמשך מישהו בשיחה אומר "I think even one billion" [1:31] is too much. I don't know. We'll see." אבל הכתוביות מצמידות את ההמשך לפריים של אחד המגישים, ולכן השיוך כאן לא ודאי. ההשלכות שעולות בשיחה הן edge device מול ענן, ועלות ההרצה. אחד המגישים מנסח את זה כ-"I have my exocortex on a local device as well".

3. חברות של LLMs: היררכיה, swarm ו-escalation

Karpathy לא חושב שהתוצאה תהיה מודל יחיד. הוא מעדיף תהליכים מקביליים על פני תהליך סדרתי, ומשווה את זה לחברה, שהיא בעצמה סוג של מקביליות של עבודה. ההיררכיה בחברה קיימת לדבריו בגלל צורכי עיבוד המידע והצמצומים שהארגון צריך לבצע. מכאן הוא מגיע לתחזית מפורשת: "So I think we'll probably end up with companies of LLMs" [1:31]. יהיו מודלים ביכולות שונות שמתמחים בתחומים, למשל programmer ו-program manager, שעובדים במקביל ו-"orchestrating computation on your behalf".

Andrej Karpathy [1:31] You have automatic escalation to other parts of the swarm, depending on the difficulty of the problem and the specialization. I mean, the CEO is like a really brilliant cloud model. But the workers can be a lot cheaper, maybe even open source models or whatnot

אחד המגישים מציע את המטאפורה של אקוסיסטם ביולוגי עם נישות ותפקידים מתמחים, ו-Karpathy מסכים. השיחה נסגרת בהערה שפונקציית העלות שונה מאדם לאדם ("my cost function is different from your cost function").

[3:04]. מבחינה הנדסית, זה תיאור קרוב מאוד של routing בין מודלים לפי קושי ועלות: מודל frontier יקר כמתזמר, ומודלים קטנים וזולים כ-workers. בשיחה זה נשאר ברמת חזון ולא נכנס למימוש.

4. Eureka: המורה ב-back end, ה-AI ב-front end

המגישים שואלים למה Karpathy עזב את OpenAI ועבר לחינוך. התשובה שלו מורכבת משני רבדים. הראשון אישי: הוא תמיד היה מורה. השני הוא "macro picture": לדבריו רוב הפעילות ב-AI עוסקת בהחלפת אנשים, ואילו הוא מעוניין בהעצמה שלהם. [3:04] "I feel like I'm kind of on a high level like team human". נקודת המוצא שלו היא שאלה: כמה רחוק יכול להגיע אדם שיש לו מורה פרטי מושלם בכל נושא. המגישים מזכירים את המחקר משנות ה-80 על הוראה אחד-על-אחד, שמשפרת ביותר מסטיית תקן ("the bloom stuff"), ואת הספר The Diamond Age.

בפועל הוא בונה קורס אחד, שאמור להיות "just like the course you would go to if you want to learn AI" [4:37]. הוא מזכיר שכבר לימד את 231N בסטנפורד, "the first deep learning class", והבעיה שהוא מגדיר היא scale: קהל יעד של "maybe like eight billion people on earth", שמדברים בשפות שונות ונמצאים ברמות שונות. הפתרון שלו מבוסס על הערכה מפוכחת של מה המודלים מסוגלים לעשות היום:

— Andrej Karpathy [4:37] The teacher is on the back end designing the materials on the course, and the AI is the front end and it can speak all the different languages and it kind of like takes you through the course.

הוא מנמק את החלוקה במגבלה הנוכחית של המודלים: "I don't think the models are good enough to create a good course, but I think they're good to become the front end to the student" [4:37]. ההגדרה "AI TA" מקובלת עליו כאחת הדרכים לחשוב על זה. כשהיכולת תשתפר, הוא צופה "refactor the setup in various ways". מכאן עולה גם לקח כללי לבוני מוצרים: הרבה חברות לא מבינות אינטואיטיבית איפה היכולת עומדת היום, "and then they end up kind of like building things that are kind of like too ahead of what's available or maybe not ambitious enough" [6:12].

5. התאמה אישית: "the demo is near, but the product is far"

אחת המגישות משווה את המצב לספורט אולימפי: ביצועים משתפרים בזכות אימון שמתחיל מוקדם יותר, שיטות וציוד, ושואלת אם כך יהיה גם בלימדה. Karpathy עונה ש-"we haven't even scratched like what's possible at all" [6:12], ומפריד בין שני צירים: גלובליזציה, כלומר חינוך טוב לכולם, ו-"how far can a single person go".

על adaptive learning, כלומר אתגר שמותאם לרמה של הלומד, הוא מחלק את המשימות לפי רמת קושי. תרגום הוא low-hanging fruit מובהק: "I think the current models are actually really good at translation basically and can target the material and translate it like at the spot" [6:12]. התאמה לרקע של הלומד, למשל אנלוגיות מתחום שהוא כבר מכיר, היא לדבריו "not at the lohing fruit but I don't think it's like too high up". הוא מדגיש את ההבדל בין משהו שעובד בהדגמה לבין משהו שעובד באמת: "what I'm talking about is something that actually works, not something that you can demo and work sometimes" [7:46].

— Andrej Karpathy [7:46] That's the thing with AI. I feel like a lot of these capabilities are just kind of like prompt away. So you always get like demos, but like, do you actually get a product? You know what I mean? So in this sense, I would say the demo is near, but the product is far

6. תרבות, lineage ומוטיבציה בעולם post-AGI

המגישים מעלים את נושא ה-lineage במחקר: שיעור גבוה מזוכי פרס נובל עבדו קודם במעבדה של זוכה נובל אחר. Karpathy לא רואה בזה יתרון: "I don't actually want to live in a world where lineage matters too much, right?" [7:46]. לדבריו זו "gatekeeping by some finite scarce resource", והוא מקווה ש-AI יפרק את המבנה הזה [9:18]. המגישים מוסיפים את אפקט האשכול, למשל AI ב-Bay Area ו-fintech בניו יורק. Karpathy מפרק את הקהילה לשני רכיבים: הרכיב החינוכי, כלומר apprenticeship, והרכיב התרבותי, כלומר מה הקהילה מעריצה. באקדמיה זה ה-H-index. בטורונטו, שבה למד, זה "where are you getting your internship?", בסביבה שלדבריו לא הייתה יזמית.

Andrej Karpathy [10:51] So these cultural aspects are extremely strong and maybe actually the — dominant variable. Because I almost feel like today, the education aspects I think are the easier one, like a ton of stuff is already available, et cetera

בשאלה אם אפשר לשנות את הרצון ללמוד, Karpathy נזהר: "I would say I want to" [10:51] "make it much easier to learn". בעולם pre-AGI אנשים לומדים מסיבות כלכליות, ובעולם post-AGI "I think education is entertainment to a much larger extent" [12:22]. הוא מבהיר שאין הכוונה לבידור קליל: "I feel like learning something is kind of like going to the gym, but for the brain, right?" [12:22]. זה דורש מאמץ ויש בזה הנאה ותמורה. התקווה שלו היא שבעולם post-AGI אנשים ילכו הרבה "לחדר הכושר", "not just physically, but also mentally". באותו הקשר הוא מעיר שלפני 200 שנה מי שעסקו במדע היו האצולה ובעלי ההון.

7. מה ילדים צריכים ללמוד, ומתי יוצא הקורס

על קהל היעד: הקורס הראשון של Eureka בנוי ברמת תואר ראשון בתחום טכני, אבל "anyone at that level" [13:55] "any age is kind of like in scope", כי המודל של "לומדים ואז הולכים לעבוד" הוא "antiquated concept of education" ולא יחזיק מעמד בחברה שמשתנה כל כך מהר. לגבי מועד היציאה: "I was hoping it would be" [13:55] "late this year... I think probably early next year is kind of like the timeline". ההקלטה לא מציינת שנה, ולכן לא ניתן להסיק ממנה תאריך מוחלט.

לשאלה מה כדאי לילדים ללמוד, הוא עונה בלי היסוס: "I would say like math, physics, CS" [13:55]. הסיבה היא שהתחומים האלה מפתחים יכולת חשיבה, והם שימושיים גם לפני AGI וגם אחריו, כי "you still want empowered humans who can function in any arbitrary capacity" [15:28]. את התקופה שבה יש לילדים הרבה זמן וקשב צריך להקדיש, לדבריו, למשימות של symbol manipulation (בתמליל: "simple manipulation-heavy tasks and workloads, not memory heavy tasks and workload" [15:28]).

Andrej Karpathy [15:28] I'm not opposed to all the other disciplines, et cetera. I think it's actually — beautiful to have a large diversity of things, but I do think 80% of it should be something like this

אחד המגשימים מסכם: "One, we're not efficient memorizers compared to our tools" [15:28]. המשפט הזה סוגר מעגל עם פתיחת השיחה: גם המודלים וגם בני האדם לא צריכים לשנן את מה שכלי יכול לשלוף בשבילם.

8. מסקנות ופעולות

- **למי שבונה מערכות agent:** כדאי לחשוב על architecture של routing לפי קושי ועלות. מודל חזק מתזמר, ומודלים קטנים או open source מבצעים. זה התיאור של "companies of LLMs" [1:31].
- **כדאי לבחון מודלים קטנים עם כלים:** אם הטענה על ה-cognitive core נכונה, מודל קטן עם retrieval וכלים טובים יכול להתקרב למודל גדול במשימות שדורשות חשיבה. שווה למדוד את זה על ה-workload שלכם, ולא להסתמך על האינטואיציה.
- **distillation כאסטרטגיה:** מורה גדול שמפקח על תלמיד קטן מאפשר הרצה זולה ב-edge או בענן [0:00].
- **כדאי להתאים את המוצר ליכולת של היום:** תכננו לפי מה שהמודלים עושים טוב כבר עכשיו, כמו תרגום ו-front-end מול משתמש, ולא לפי דמו [4:37, 7:46].
- **כדאי להבחין בין דמו למוצר:** יכולת שאפשר להשיג ב-prompt היא רק נקודת התחלה. המבחן הוא אם היא "actually works" באופן עקבי.
- **בלמידה אישית:** עדיף להשקיע בתחומים שבונים חשיבה, כמו מתמטיקה, פיזיקה ו-CS, ולא בשינון שכלים עושים טוב יותר [15:28].

שאלה פתוחה	למה היא נשארה פתוחה
מה הגודל המינימלי של cognitive core?	ההערכות נעות בין "even a billion" ל-"even one billion is too much", בלי ניסוי או מדד שתומך באף אחת מהן.
האם יש ניסוח מתמטי ליחס בין יכולת קוגניטיבית לגודל המודל?	השאלה נשארה במפורש ולא קיבלה תשובה.
איך מודל קטן יודע מה הוא לא יודע ומתי לפנות לכלי?	השיחה מניחה שהוא "knows how to use different tools", אבל לא מסבירה איך.

שאלה פתוחה	למה היא נשארה פתוחה
איך עובד escalation ב-swarm?	לא נאמר מי מעריך את הקושי, איך מתבצע ה-routing, ומה העלות של תיאום בין המודלים.
איך תימדד ההצלחה של Eureka?	Karpathy רוצה מוצר ש-"actually works", אבל לא מוצגת מתודולוגיה להערכת למידה.
האם AI יכול לשנות תרבות, ולא רק להנגיש תוכן?	Karpathy עצמו אומר שתרבות היא המשתנה הדומיננטי, ולא מציע מנגנון לשנות אותה.

9. מילון מונחים

מונח	משמעות
cognitive core	ליבה קטנה של מודל שמתמחה בחשיבה ולא בשינון, ושולפת מידע בעזרת כלים
distillation	אימון מודל קטן ("תלמיד") בפיקוח של מודל גדול ("מורה") כדי לרחוס לתוכו יכולות
parameters	המשקלים של המודל, היחידה שבה נמדד הגודל שלו (למשל 1B, כלומר מיליארד)
edge device	מכשיר קצה מקומי, כמו טלפון או מחשב, שמריץ את המודל בלי ענן
exocortex	"קליפת מוח חיצונית": עוזר AI שמרחיב את היכולת הקוגניטיבית של האדם
swarm	נחיל של מודלים מתמחים שעובדים יחד ומעבירים משימות ביניהם
escalation	העברה אוטומטית של בעיה למודל חזק או מתמחה יותר לפי רמת הקושי
open source models	מודלים בקוד פתוח, זולים להרצה, שמוצעים כאן כ-workers
front end / back end	כאן: ה-AI הוא הממשק מול הסטודנט, והמורה מתכנן את החומר מאחורי הקלעים
AI TA	עוזר הוראה מבוסס AI
low-hanging fruit	הישג קל להשגה. בתמליל משובש כ-"lohing fruit"
adaptive learning	הוראה שמתאמת לרמה, לקצב ולרקע של הלומד
lineage	שושלת אקדמית: מאיזו מעבדה ומאיזה מנחה מגיע חוקר
H-index	מדד אקדמי למספר הפרסומים ולמספר הציטוטים שלהם
post-AGI	העולם שאחרי בינה מלאכותית כללית
Eureka	מיזם החינוך של Karpathy, שבונה קורס AI עם AI כ-front end
symbol manipulation	משימות של עבודה עם סמלים ומבנים, כמו במתמטיקה. בתמליל מופיע כ-"simple manipulation"