

איך Anthropic ממזגת 30,000 שינויי קוד ביום — מי מאמת את ה-software factory?

שיחה של Sonar ו-Anthropic: כתיבת קוד כבר לא צוואר הבקבוק, האימות כן

Manish Kapur (Sonar), Addy Osmani (Anthropic) · AI Engineer · 28.10.2026

משך: כ-19 דקות · YouTube

סיכום אוטומטי — AI Briefing (aibriefing.dev)

על המסמך הזה

זו שיחה של 19 דקות ממושב ה-Leadership בכנס AI Engineer New York. לפי השקף הפותח היא הוקלטה ב-14 באוקטובר 2026, באולם Sutton Room. Manish Kapur (VP Product Marketing & DevRel ב-Sonar) מנחה, ו-Addy Osmani (Member of Technical Staff ב-Anthropic, לפני כן 14 שנה ב-Google) עונה. השקפים הם של Sonar ומשלבים נתונים מסקר State of Code של החברה, שעוד לא פורסם, ומדוח של CircleCI.

הטענה המרכזית: coding agents כבר פתרו את כתיבת הקוד, אבל הביטחון במה שנכתב לא גדל באותו קצב. מה שיקבע כמה מהר אפשר לשחרר קוד הוא אימות עצמאי, אוטומטי ורב-שכבתי בכל loop. הדוגמה הבולטת היא Anthropic. שם עלה מספר השינויים הממוזגים מכ-2,500 ביום לכ-30,000, ו-Claude עושה את ה-review הראשוני על סמך ראיות אימות אמיתיות.

איך לקרוא את המסמך

- **1. תקציר מנהלים:** הטענות המרכזיות עם חותמות זמן, ומה לא נאמר בשיחה.
- **2. הבעיה:** ה-throughput עלה, ה-merge נכשל, ו-76% מהמפתחים מייצרים קוד מהר יותר ממה שהם מסוגלים לאמת.
- **3. Light ו-Dark factories:** ההגדרה של Osmani והנתונים על מעבר ל-factories אוטומטיים.
- **4. ה-review ב-Anthropic:** ה-workflow שמאפשר 30,000 שינויים ביום.
- **5. Verification רב-שכבתי:** נתוני הסקר והמודל של Sonar: שכבה דטרמיניסטית, שכבה agentic ושכבת supply chain.
- **6. שליטה, sandbox ו-guide rails: lightning round,** תיוג PR-ים שכתב agent ומתי בטוח למזג.
- **7. מה עושים מחר:** פעולות מעשיות ושאלות פתוחות.
- **8. מילון מונחים:** המונחים באנגלית ופירושים.

הערות מקור המסמך הופק אוטומטית מתמליל ההקלטה ומהשקפים שהוצגו בה. התמליל הוא ASR, ולכן שמות ומונחים עלולים להשתבש: למשל Cloud Code במקום Claude Code, Kapoor במקום Kapur ו-Claw.md במקום CLAUDE.md. איפה שהשקפים מאפשרים, השמות תוקנו לפיהם. הציטוטים מועתקים מילה במילה מהתמליל ונבדקו מול חותמת הזמן שלהם. לא נוסף שום מידע ממקורות חיצוניים.

1. תקציר מנהלים

השיחה מתמקדת בפער אחד: agents פותחים PR-ים מהר יותר ממה שבני אדם מסוגלים לבדוק. Sonar מביאה את הנתונים ואת המודל של אימות רב-שכבתי. Anthropic מביאה דוגמה מהשטח: כ-30,000 שינויים ממוזגים ביום, כש-Claude עושה first-pass review ובני האדם בודקים את הראיות שה-agent מספק, לא כל שורת קוד.

- **Throughput עלה, merge נכשל:** לפי CircleCI, שניתחה 28 מיליון CI workflows, ה-throughput קפץ ב-59% (לפי השקף, העלייה הגדולה ביותר שנמדדה אצלם), ובכל זאת כ-3 מכל 10 ניסיונות merge ל-production נכשלים.

- **כתיבת קוד נפתרה, אימות לא:** 85% מהארגונים כבר משתמשים ב-coding agents, ו-76% מהמפתחים אומרים שהקוד נוצר מהר יותר ממה שהם יכולים לאמת. בשקף מופיע נתון נוסף: 53% כבר מבליים יותר זמן ב-review מאשר בכתיבה.

- **Anthropic:** מ-2,500 ל-30,000 שינויים ביום: לפי Osmani, במאי עמד המספר על כ-2,500. ה-refresh של Claude.ai כלל כ-3,000 PR-ים בזמן קצר, ועל פי ה-case study שהוא מזכיר הביצועים שם שולשו.

- **Claude עושה first-pass review עם ראיות:** ה-agent מנסה את המוצר כמו משתמש אמיתי (e2e, דפדפן, lint), hooks רצים על כל edit ובודקים את ה-diff, ובני אדם בודקים את הראיות: סרטונים וצילומי מסך.

- **אימות טוב משפר את התוצאה פי 2-3:** לפי Osmani, כשנותנים ל-agent דרך לוודא שהעבודה נכונה ושלמה, איכות התוצאה יכולה לעלות פי שניים-שלושה. הוא לא הציג מדידה שתומכת במספר הזה.

- **Light factories רוצים להיות dark:** לפי הסקר של 5% Sonar, מהארגונים מריצים היום software factories עם agents אוטונומיים. הצפי הוא 34% בתוך שנה, ו-39% בשירותים פיננסיים.

- **אימות רב-שכבתי בכל Sonar: loop** ממליצה על zero trust בשלוש שכבות: דטרמיניסטית (באגים, SQL injection), agentic (לוגיקה וכוונה) ו-supply chain (CVE, malware).

- **שליטה היא צוואר הבקבוק הארגוני:** ל-28% מהארגונים יש agents מאושרים אבל מעט בקורות משותפות. Osmani מונה שלוש דאגות מרכזיות: permissions, control ו-sandbox.

- **אחריות נשארת אצל האדם:** שני הדוברים בעד לתייג PR-ים שכתב agent, בגלל auditability, אבל מי שהפעיל את ה-agent נשאר אחראי לשינוי.

מה לא נאמר

- **אין הגדרה ל-30,000:** לא הוסבר אם מדובר ב-PR-ים, ב-commits או בשינויים אוטומטיים, ולא נמסר שיעור ה-reverts או התקלות אחרי הגידול.

- **אין נתוני עלות:** לא נאמר כמה עולים ה-review האוטומטי, ה-hooks וייצור הסרטונים בהיקף כזה.

- **המספר פי 2-3 לא מגובה:** הטענה על שיפור באיכות הוצגה כהערכה, בלי מדד ובלי מקור.

- **הסקר עוד לא פורסם:** דוח State of Code של Sonar מוצג רק בתצוגה מקדימה (קצת יותר מ-1,000 מפתחים). המתודולוגיה לא מפורטת, ונתון ה-31% מנוסח בעמימות (פירוט בפרק 5).

- **אין סיפורי כישול:** לא סופר מה נשבר ב-Anthropic בדרך ל-30,000, ומה הייתה עושה אחרת.

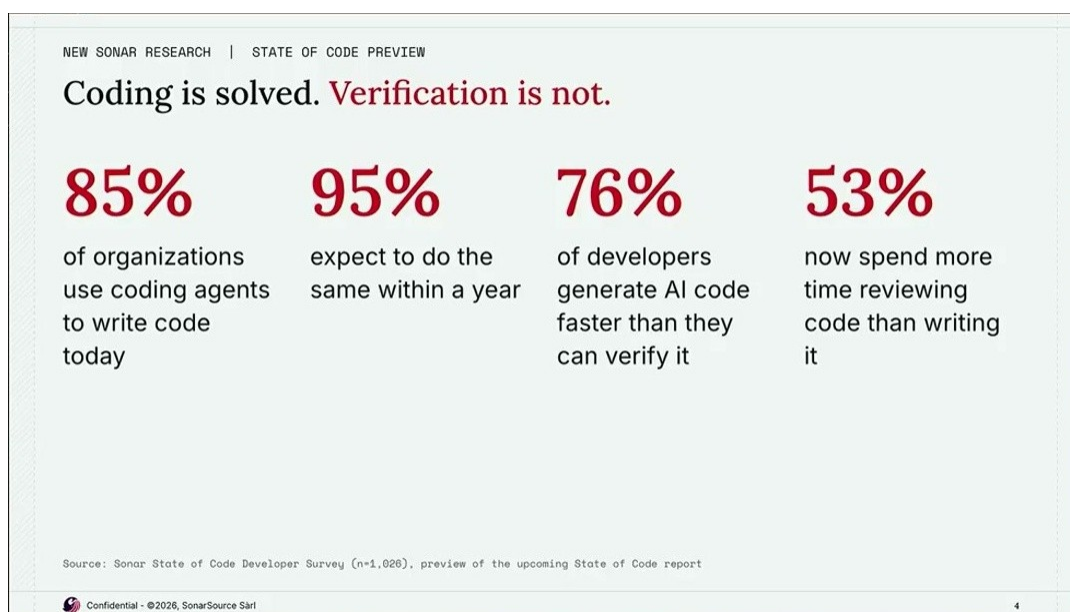
- **הקשר מסחרי:** השקפים והמודל הרב-שכבתי הם של Sonar, שמוכרת כלי אימות. Osmani הזכיר גם את מוצר ה-code review של Anthropic.

2. הבעיה: ה-throughput עלה, הביטחון לא



"Writing code is no longer the constraint. Checking it is": עלייה של 59% ב-CI throughput מול כ-3 מכל 10 ניסיונות merge ל-production שנכשלים (28, CircleCI 2026 State of Software Delivery מיליון workflows)

Kapur פותח בשאלה לקהל: כמעט כל האולם הרים יד כשנשאל אם ה-agent שלו פתח PR בשבוע האחרון, ורק מעטים השאירו אותה למעלה כשנשאלו אם עברו על כל שורה לפני merge. הנתון של CircleCI מחדד את הפער: הקוד זורם מהר יותר, אבל כמעט שליש מניסיונות ה-merge נכשלים. לדבריו, זו המגמה שהוא רואה גם אצל רוב לקוחות Sonar.



"85%: Coding is solved. Verification is not": 95% ממתמשים ב-coding agents צופים לעשות זאת בתוך שנה, 76% מייצרים קוד AI מהר יותר ממה שהם יכולים לאמת, 53% מבליים יותר זמן ב-review מאשר בכתיבה (Sonar State of Code, תצוגה מקדימה)

לדברי Osmani, מה שמעסיק היום מנהלי הנדסה הוא איך לשחרר מהר יותר ועדיין בביטחון. agents מאפשרים לייצר בדיקות בזול: tests, benchmarks, specs, כל מה שבעבר הגדיר קוד איכותי. לכן לשם צריך להפנות את הזמן: "If you are hoping to get the best quality output from your agents, you need to give your agents a way to verify that the work is correct." לטענתו, כשהקוד גם נכון וגם שלם, איכות התוצאה עולה פי שניים עד שלושה. Kapur מסכם את זה בסיסמה של Sonar: code confidence is scarce.

Dark software factories- Light 3.

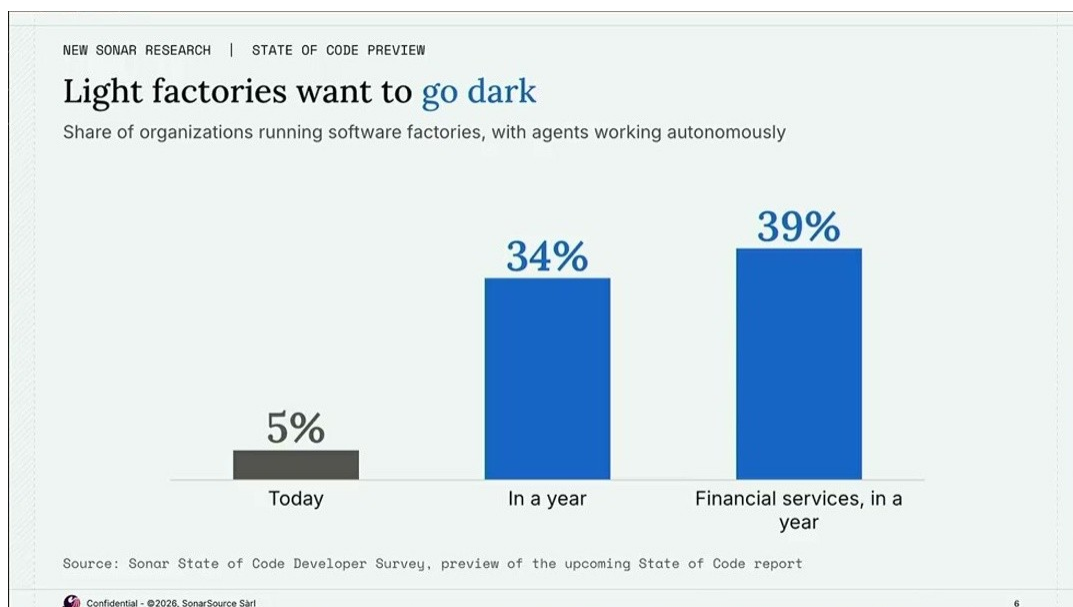
“A software factory is harnessing loops at scale.”

Addy Osmani, “Software Factories, Light and Dark,” July 2026

Confidential - ©2026, SonarSource Sàrl 3

ההגדרה של "A software factory is harnessing loops at scale", מתוך המאמר "Software Factories, Light and Dark" (יולי 2026)

Osmani מבחין בין שני סוגי factories. ב-factories light, מתכננים, בונים, בודקים ומשחררים ברולאה, והאדם עדיין מפעיל שיקול דעת וטעם בתוך התהליך. ה-dark factory דומה למפעל רכב רובוטי: הלולאה רצה בלי הפסקה, האדם עובר לחשוב על intent, ומעורבים אותו רק במקרי חריגה. לדבריו "I think a lot of teams are somewhere on this spectrum". מה שמזיז ארגון לכיוון dark הוא האימות: "The more that the agents are able to generate verification and give you confidence, the more that you can get to that place where you go dark." Kapur. "you go dark" מוסיף את מטפורת קו הייצור: intent ו-specs, אחריהם קוד, וכל שלב ב-SDLC אוטומטי. המטרה היא מפעל שרץ 24/7 עם האורות כבויים.



"Light factories want to go dark": שיעור הארגונים שמריצים software factories עם agents אוטומטיים, 5% היום, 34% בתוך שנה ו-39% בשירותים פיננסיים בתוך שנה

הנתון המפתיע בעיני Kapur הוא שדווקא בשירותים פיננסיים, תחום רגולטורי, הצפי הוא הגבוה ביותר: 39%. כשהוא שאל את הקהל מי כבר מריץ software factory, רק קומץ הרים יד, ואחד המשתתפים ביקש קודם הגדרה. זה מלמד שגם המונח עצמו עוד לא מוסכם.

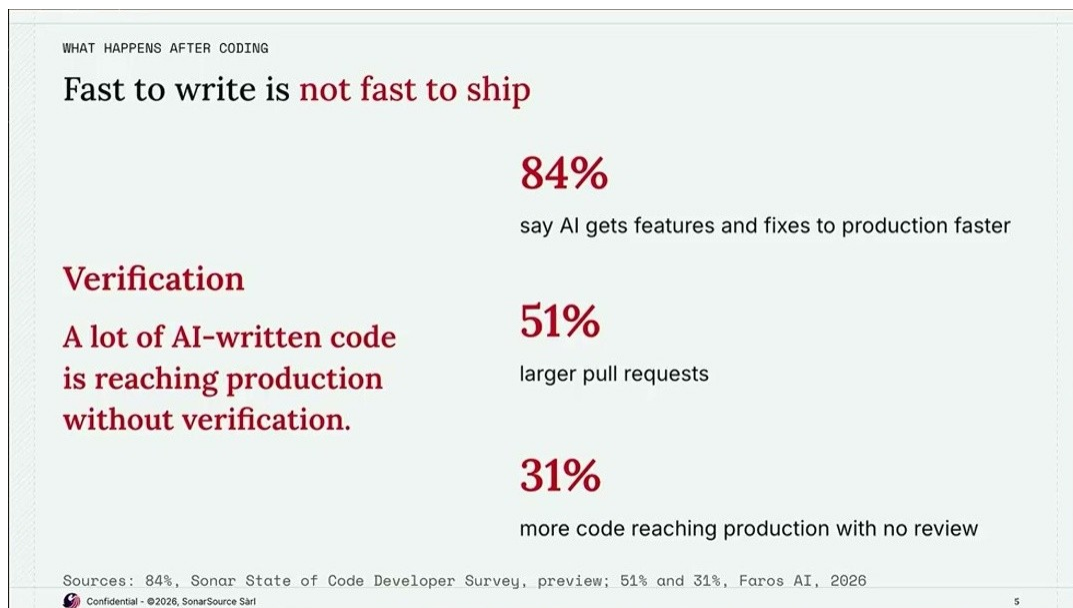
4. איך Anthropic מבצעת review ל-30,000 שינויים ביום

Kapur מצטט פוסט של Anthropic: מכ-2,500 שינויים ממוזגים ביום לפני כמה חודשים, החברה הגיעה לכ-30,000. Osmani מאשר שבמאי עמד המספר על כ-2,500. כדוגמה הוא מביא את ה-refresh של Claude.ai: לפי ה-case study, הביצועים שם שולשו בעזרת כ-3,000 PR-ים שמוזגו בזמן קצר, עם תרומות מהנדסה, מוצר, עיצוב ופונקציות נוספות.

Addy Osmani What's very important in there is making sure that it has a way to actually try the — experience out similar to a real user

- **First-pass review של Claude:** Claude עובר ראשון על כל שינוי, ובמסגרת זו מריץ משימות end-to-end, פותח את המוצר בדפדפן, מפעיל כלים ו-linting ומבצע כל בדיקה שבונה ביטחון.
- **Hooks על כל edit:** כשמתגלה באג או פער בפונקציונליות, hooks רצים על כל עריכה, בודקים את ה-diff ומחזירים את ה-PR לתיקון מהר ככל האפשר.
- **מוצר Anthropic: code review:** משתמשת בעצמה במוצר ה-code review שלה, שזמין גם ללקוחות. המטרה היא להפוך כמה שיותר מה-review לאוטומטי.
- **טעם בקבצים משותפים:** את הטעם ואת ה-best practices של הצוות כדאי לקודד ב-CLAUDE.md או ב-AGENTS.md משותפים, שה-agent קורא במהלך ה-review.
- **אדם בודק ראיות, לא שורות:** "And then people are responsible for checking on the agent's evidence." ב-UI הראיה היא סרטון של ה-agent מקליק בדפדפן. במובייל ובדסקטופ אלה צילומי מסך או סרטונים. המטרה היא לקבל מה-agent כמה שיותר verification evidence.

5. Verification רב-שכבתי: מאיפה יבואו השיפורים הבאים

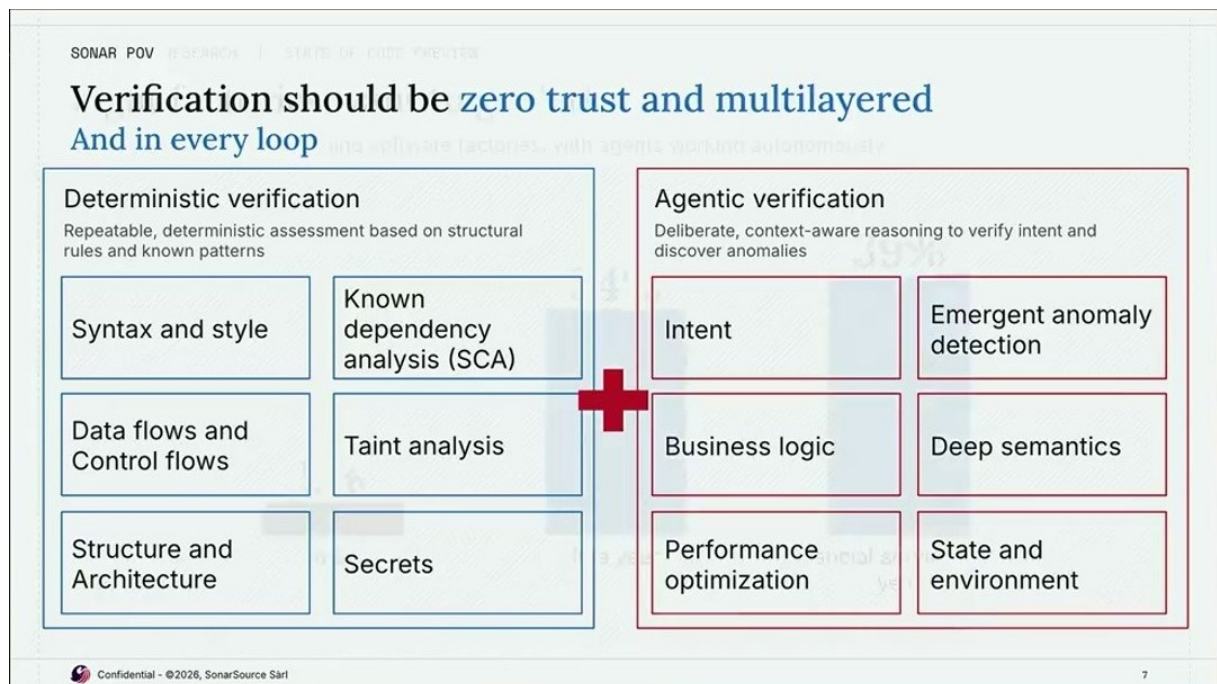


"84%: Fast to write is not fast to ship" אומרים ש-AI מביא פיצ'רים ותיקונים ל-production מהר יותר, 51% מדווחים על PR-ים גדולים יותר, 31% על יותר קוד שמגיע ל-production בלי review

הסקר של Sonar כלל קצת יותר מ-1,000 מפתחים. 84% אומרים ש-AI מקצר את הדרך ל-production, וה-PR-ים נעשים גדולים ומורכבים יותר. בעל פה Kapur אמר "only 31% more code is reaching the production without review". בשקף כתוב "31% more code reaching production with no review", כלומר כנראה 31% מהנשאלים מדווחים שיותר קוד מגיע ל-production בלי review. הניסוח עמום, ולכן כדאי לא לצטט את המספר הזה בלי לבדוק בדוח המלא.

לפי Osmani, השיפורים הבאים יבואו משילוב של אימות טוב יותר ואיכות מודלים. על המודלים הוא אומר: "if you take a look at Opus 5.5, model quality has already gotten to a place where it's pretty good". לכן הדגש עובר לשיתוף best practices של אימות, כי ההוראה לאמת לבד לא אומרת כלום. ב-Anthropic: "We lean a lot"

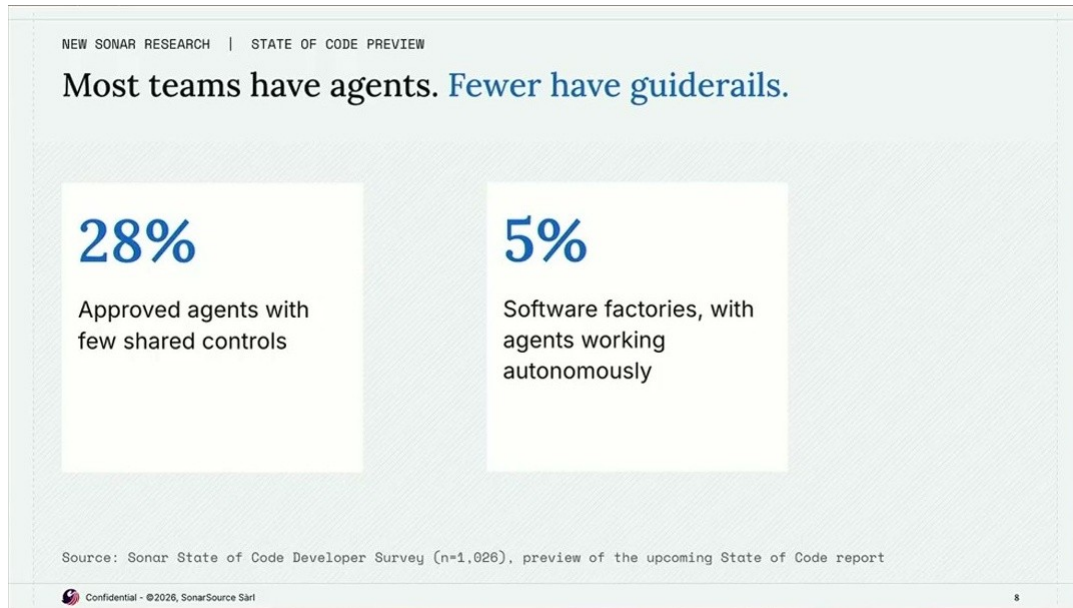
on end-to-end tests, property tests, any visual verification tests, things that can help us just get that
".confidence to merge those changes in automatically



Deterministic verification (syntax, עמודות, מ'מין לשמאל, "Verification should be zero trust and multilayered. And in every loop")
Agentic verification (intent, anomaly detection, business ועמודות SCA, data/control flows, taint analysis, architecture, secrets)
logic, deep semantics, performance, state)

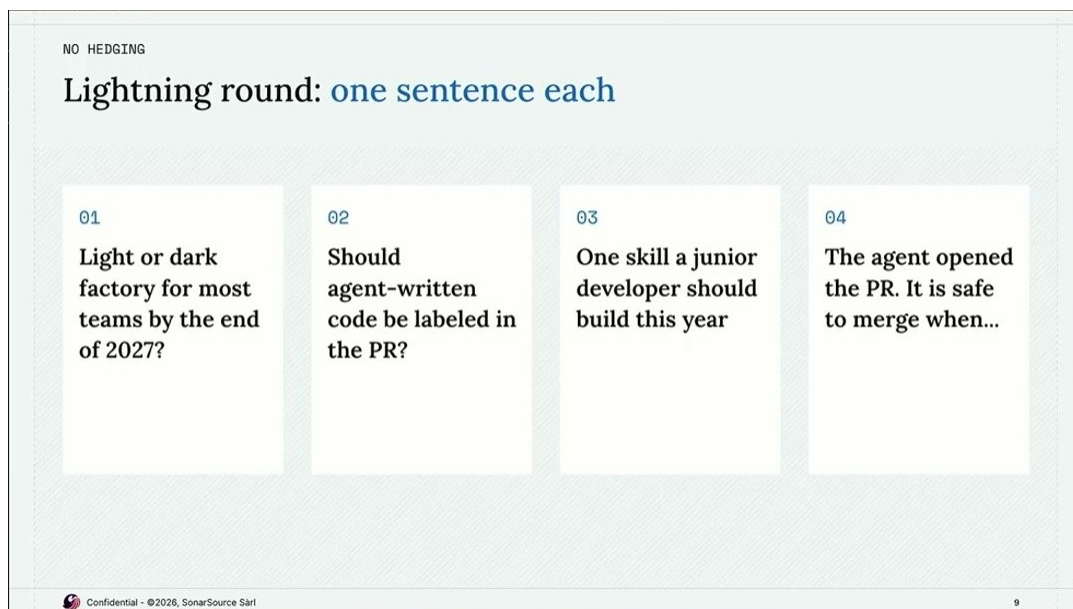
מקור	מה היא תופסת	שכבה
השקף +	control-ו style, data flows-ו Syntax architecture, flows, structure known dependency analysis (SCA), taint analysis, secrets ובעל buffer ו SQL injection פה גם overflow	Deterministic
השקף +	Intent, business logic, deep semantics, emergent anomaly detection, performance :environment-ו optimization, state בעיות לוגיקה וכוונה שהשכבה הדטרמיניסטית לא מתאימה לתפוס	Agentic review
בעל פה בלבד	malware וזיהוי חבילות זדוניות CVEs,	Software supply chain
	Zero trust, אימות בכל loop: בתוך ה-agentic loop, ב-CI ובתחזוקת קוד. כל שכבה מכוונת לסוג אחר של בעיות	עיקרון-על

6. שליטה, הרשאות, lightning round-sandbox



"Most teams have agents. Fewer have guidrails." - ל-28% יש agents מאושרים עם מעט בקורות משותפות, ו-5% מריצים software factories עם agents אוטונומיים

Osmani מתאר ארגונים שאפשרו לצוותים להתנסות עם agents, ותוך זמן קצר מצאו את עצמם עם מאות ולפעמים אלפי agents שנוגעים במערכות, ניגשים לנתונים ואולי גם מבצעים deploy ל-production, בלי שאיש יודע מה הם עושים. לכן enterprises עוסקים ב-agent identity וב-telemetry, ושלוש השאלות המרכזיות הן, control sandbox: "How are we making sure that the agent isn't jumping out of the sandbox-ו permissions and doing things that it shouldn't be doing?" - ב-Sonar קוראים לזה guide rails במקום guard rails: לא רק לחסום, אלא להוביל את ה-agent במסלול מוגדר.



שקף ה-Lightning round: ארבע שאלות על light מול dark עד 2027, תיוג PR שכתב agent, מיומנות לפתח השנה, ומתי בטוח למזג PR של agent

שאלה	Addy Osmani (Anthropic)	Manish Kapur (Sonar)
Light או dark לרוב הצוותים עד סוף 2027?	I think we're going to see "light" factories with some dark corners." - לדבריו, Anthropic תגיע ל-dark כבר השנה או בשנה הבאה.	לא ענה ישירות.

שאלה	Addy Osmani (Anthropic)	Manish Kapur (Sonar)
לתייג PR שכתב agent?	כן. ב-Anthropic לא מפריע שהשינוי מתווגש כקוד של Claude, כי הפוקוס הוא על התוצאה, והתיג עוזר להחליט אילו בדיקות נוספות להפעיל.	כן, לפחות ב-6-9 החודשים הקרובים, בשביל auditability, בעיקר אצל לקוחות פיננסיים.
אחריות	"I'm still on the hook. If something goes wrong, I have to go and fix it."	"we are, at the end of the day, still" "accountable"
מיומנות לפתח השנה	מנהלים: להגדיר מה אימות אומר אצלם, ומה הפער עד שה-agent יכול לאשר merge בעצמו. ג'וניורים: ניסיון בסוגי tests, כמו end-to-end ו-property tests.	שאלה שהוסיף בעצמו במקום זו שהופיעה בשקף.
מתי בטוח למזג PR של agent?	"You have full confidence" nothing's going to go wrong, or that you've mitigated that as much "as possible"	"Fast to write is not fast to ship. Autonomy is earned"

7. מה עושים מחר

- **להגדיר מה זה אימות אצלכם:** למפות את בדיקות הקוד הקיימות ולזהות את הפער עד שה-agent יכול לאשר merge בעצמו. זו ההמלצה של Osmani למנהלים.
- **לדרוש ראיות מה-agent:** סרטון של תהליך בדפדפן, צילומי מסך למובייל, תוצאות e2e. ה-reviewer האנושי בודק את הראיות.
- **להשקיע ב-e2e, property ו-visual tests:** אלה הבדיקות ש-Anthropic נשענת עליהן כדי למזג באופן אוטומטי.
- **לקודד את הטעם של הצוות:** CLAUDE.md או AGENTS.md משותפים שה-agent קורא בזמן review.
- **אימות בשכבות בכל loop:** כלים דטרמיניסטיים, agentic review ו-supply chain scanning, גם בתוך ה-agent loop ולא רק ב-CI.
- **לעשות agents ל-inventory:** מי רץ, עם אילו הרשאות, באיזה sandbox ועם איזו telemetry, לפני שזה יוצא משליטה.
- **לתייג PR-ים של agents ולהגדיר בעלים:** לתעד את התיג לצורך audit, ולקבוע שמי שהפעיל את ה-agent אחראי לשינוי.

שאלה פתוחה	למה זה חשוב
מה בדיוק נספר ב-30,000 שינויים ביום של Anthropic?	בלי הגדרה (PR, commit, שינוי אוטומטי) אי אפשר להשוות לארגון אחר.
מה שיעור ה-reverts והתקלות אחרי הגידול פי 12?	זה המדד שבאמת יראה אם first-pass review של Claude מחזיק בקנה מידה כזה.
כמה עולה ייצור הראיות (סרטונים, e2e) לכל PR?	בהיקף של עשרות אלפי שינויים ביום, עלות האימות עלולה להפוך לצוואר הבקבוק החדש.
מה המשמעות המדויקת של נתון ה-31% בסקר של Sonar?	הניסוח בעל פה ובשקף עמום. צריך לבדוק בדוח המלא לפני שמצטטים.
מאיפה בא השיפור פי 2-3 באיכות?	זו הטענה החזקה ביותר בשיחה, והיא הוצגה בלי מדידה.
איך מחליטים אילו loops מותר להעביר ל-dark?	Osmani מדבר על dark corners, אבל לא הציע קריטריון לבחירתם.

8. מילון מונחים

מונח	פירוש
Software factory	לולאות של agents שמתכננים, בונים, בודקים ומשחררים קוד בקנה מידה גדול (harnessing loops at scale)
Light factory	factory שבו אדם מעורב ומפעיל שיקול דעת בכל השלבים או ברובם
Dark factory	factory אוטונומי לגמרי, והאדם מתעסק רק ב-intent ובחריגים

פירוש	מונח
סקירה ראשונה שה-agent מבצע לפני שאדם מסתכל על השינוי	First-pass review
סקריפטים שרצים אוטומטית על כל עריכה של ה-agent ובודקים את ה-diff	Hooks
ראיות שה-agent מספק לכך שהשינוי עובד: סרטונים, צילומי מסך, תוצאות בדיקות	Verification evidence
קבצי הנחיות משותפים שה-agent קורא, עם ה-best practices והטעם של הצוות	CLAUDE.md / AGENTS.md
בדיקות של תהליך מלא, של תכונות כלליות על קלטים רבים, ושל מראה הממשק	End-to-end / property / visual tests
ניתוח סטטי מבוסס חוקים: syntax, data flow, taint analysis, secrets	Deterministic verification
Software Composition Analysis: ניתוח תלויות ידועות ופגיעויות בהן	SCA
review שמבצע agent ומכוון לבעיות לוגיקה, כוונה וסמנטיקה	Agentic review
סריקת CVEs, malware וחבילות זדוניות בתלויות	Supply chain security
גישה שלא סומכת על שום פלט, ולכן הכול מאומת בכל loop	Zero trust
סביבה מבודדת שבה ה-agent רץ בלי גישה למערכות אחרות	Sandbox
המונח של Sonar למסלול מונחה ל-agent, בניגוד ל-guard rails שרק חוסמות	Guide rails
כמות העבודה שעוברת ב-CI ליחידת זמן	Throughput