

לבנות agents מאובטחים לעבודת ידע

איך Every בנתה AI coworker אחד לכל החברה בתוך Slack על גבי Claude Managed Agents, ומה זה מלמד על מדידה, זהות וטעם אישי

2026 באוקטובר Katelyn Lesse, Dan Shipper, Willie Williams · Claude · 6

משך: כ-34 דקות · YouTube

סיכום אוטומטי — AI Briefing (aibriefing.dev)

על המסמך הזה

זו שיחה של כ-34 דקות שפורסמה בערוץ של Claude. Katelyn Lesse, Head of Platform Engineering ב-Anthropic, מראיינת את Dan Shipper, מייסד-שותף ו-CEO של Every, ואת Willie Williams, Head of Platform ב-Every. במרכז השיחה עומד Every Agent: AI coworker שיושב ב-Slack של הארגון ובנוי על Claude Managed Agents.

הטענה המרכזית היא שעבודת ידע עוברת עכשיו את מה שהנדסת תוכנה כבר עברה: מעבר מביצוע ידני של כל משימה לבנייה ותחזוקה של מערכת שעושה את העבודה בשבילך. כדי שזה יעבוד צריך שלושה דברים. agent משותף אחד שכל הארגון משקיע בו, ולא agent פרטי לכל עובד. מדידה שמבוססת על הטעם של הארגון, ולא על benchmarks גנריים. ותשתית מנוהלת (sandboxes, memory, sessions) שמשחררת את הצוות מעבודת infra ומאפשרת לו להתמקד ב-interaction design.



שלושת המשתתפים בשיחה סביב שולחן העבודה: Katelyn Lesse מ-Anthropic מראיינת את שני נציגי Every. בשיחה לא הוצגו שקפים.

איך לקרוא את המסמך

- 1. **תקציר מנהלים:** הטענות המרכזיות והראיות שלהן, ומה שלא נאמר.
- 2. **Every וה-AI coworker:** מאיפה הגיע הרעיון, מ-OpenClaw על Mac minis ועד agent ארגוני אחד.
- 3. **דוגמאות מהשטח:** open enrollment ו-skill Kate Bench, שנבנה מ-30,000 עריכות.
- 4. **benchmarks אישיים:** מ-vibe checks ידניים לבדיקות שמקודדות את הטעם של הארגון.
- 5. **Slack ו-agent אחד:** למה עברו ממודל של agent לכל עובד ל-agent משותף.
- 6. **infra: Claude Managed Agents,** סביבת staging, אבטחה, זהות והרשאות.
- 7. **לאן זה הולך:** fidelity, אישיות, ומה Every מבקשת מ-Anthropic.
- 8-9. **פעולות ומילון מונחים:** מה כדאי לעשות עכשיו, שאלות פתוחות, ומונחים.

הערת מקור המסמך הופק אוטומטית מהתמליל ומהפריימים של ההקלטה. לתמליל אין תיוג דוברים, ולכן ייחוס הציטוטים לדוברים מבוסס על הקשר הדברים ועלול להיות שגוי. ייתכן ש-ASR שיבש שמות ומונחים: Claude מופיע בתמליל כ-Kate Bench, Cloud מופיע גם כ-Cape Bench וכ-K-Bench, ו-beer IOUs מופיע כ-beer CEOs. הציטוטים מועתקים מילה במילה ונבדקו מול חותמת הזמן שלהם. לא נוסף מידע ממקורות חיצוניים.

1. תקציר מנהלים

השיחה היא בעיקרה case study של לקוח, ולא הרצאה טכנית. Every מתארת איך agent ארגוני אחד ב-Slack מפץ שיטות עבודה בתוך החברה, איך ממירים את הטעם של מומחה ל-skill שמשפר לאורך זמן, ולמה בחרו לא לבנות את שכבת ה-infra בעצמם. הנושא שבכותרת, אבטחה, מקבל בפועל כמה דקות בלבד, והוא מוצג כנגזרת של primitives של בידוד ושל ניהול sessions.

- **Agent אחד לכל החברה עדיף על agent לכל עובד:** לדבריהם agents משתפרים ככל שמשקיעים בהם, ובמודל של agent אישי השיפורים לא עברו לשאר הצוות [10:41].
- **Slack הוא ערוץ ההפצה:** כשהעבודה עם ה-agent נעשית ב-Slack, עובדים "prompting in public instead of in private", ושיטות עבודה טובות עוברות מאדם לאדם בלי הדרכה [3:01].
- **אפשר להפוך טעם של מומחה ל-skill:** מ-30,000 עריכות היסטוריות של עורכת ראשית בנו skill שמציע תיקונים ישירות ב-Google Doc. העבודה שהיא מוסיפה ידנית נאספת חזרה לתוך ה-skill [7:36].
- **benchmarks גנריים כבר לא מבחינים בין מודלים:** לפי אנלוגיית ה-SAT, כשכל המודלים מקבלים 1600 צריך work trial ו-reference check על העבודה שלך עצמך [21:24].
- **vibe checks הופכים לבדיקות:** דוגמאות טובות ורעות מעבודה יומיומית מומרות ל-checks שדומים ל-unit tests, ומאפשרות להריץ כמה מודלים על אותו תרחיש [10:41].
- **לא רצו להיות צוות infra:** ניסיון ראשון לבנות orchestration של sandboxes בעצמם "got unwieldy very quickly". המעבר ל-Claude Managed Agents נתן להם memory, sandboxes, ו-session control מוכנים [12:13].
- **סביבת staging על אותה תשתית:** מעתיקים memories ו-durable state מ-production למופע נפרד ומריצים בו ניסויים, כולל ניסויים בהורדת עלות הטוקנים [15:17].
- **מודל זהות כפול:** בתוך Slack ה-agent הוא ישות נפרדת עם memories מבודדים. כשהוא ניגש לכלים או למסמכים הוא פועל ברוב המקרים בשם המשתמש [18:20].
- **המודל וה-harness כבר לא נפרדים:** בהשקת מודל חדש עם breaking changes בהנחיות ה-prompting, העדכון ב-Claude Managed Agents היה "a one-line update" [18:20].

מה לא נאמר

- **אין מספרים על חיסכון:** הוזכר גרף שמראה ל-Kate כמה זמן נחסך לה, אבל לא נמסר אף נתון, וגם לא שיעור הקבלה של התיקונים שה-skill מציע.
- **אבטחה בלי threat model:** למרות הכותרת, לא דובר על prompt injection, על דליפת מידע בין משתמשים בערוצים משותפים או על audit. נאמר רק שיש בידוד ושליטה ב-sessions.
- **הרשאות באזורים האפורים:** הודו שקשה לקודד את שיקול הדעת של עובד אנושי ("there's no harm in me", "being able to access this report"), אבל לא הסבירו איך זה נאכף בפועל.
- **עלות:** הוזכרו ניסויים בהורדת עלות הטוקנים ללקוחות, אבל לא נמסרו מחיר, צריכה או תוצאות.
- **תהליך בניית ה-skill:** לא הוסבר איך 30,000 עריכות הפכו ל-skill: כמה מהן נכנסו בפועל, באיזה פורמט ואיך מבצעים לו eval.
- **המודל החדש:** נאמר שבאותו יום הושק מודל חדש, אבל השם שלו לא מוזכר בתמליל.

2. Every וה-AI coworker: מ-Mac minis ל-agent ארגוני

אחד

Every מגדירה את עצמה כחברת מדיה וטכנולוגיה שמוכרת "the only subscription you need to stay at the edge of AI". המנוי כולל שני חלקים: education (ניוזלטר, קורסים ו-vibe checks בכל השקת מודל) ו-equipment, שהוא עכשיו ה-[1:30] agent. העמדה של Dan Shipper היא שהנדסת תוכנה הגיעה לשם ראשונה כי ב-programming יש verifiable rewards. היום יש מהנדסים שמתזמרים צוותים של agents ומקדימים את ה-roadmap בכשנתיים. בעבודת ידע זה לוקח יותר זמן, כי קשה יותר לדעת אם העבודה טובה וכי זו דרך חשיבה שונה [1:30].

ההיסטוריה של המוצר: בתקופת OpenClaw כל החברה הריצה OpenClaws על Mac minis, אבל "they're just really hard to maintain" והם נזנחו בהדרגה. נשאר agent ארגוני אחד, ושם הם שמו לב כמה קל להפיץ דרכו workflows [3:01]. הם מזהים פרסונה שהם קוראים לה AI ambassador: עובד שיש לו setup שמכפיל לו את התפוקה, ושהעמיתים שלו לא מקשיבים להסברים שלו. agent ב-Slack מאפשר לו להדגים במקום להסביר.

— **Dan Shipper (ייחוס משוער) [3:01]** the beauty of a Slack agent, the Every Agent in particular, is that you can show instead of tell how to use this stuff to actually do really, really great work because you're doing it in Slack, you're prompting in public instead of in private

3. דוגמאות מהשטח: Kate Bench ו-open enrollment

הדוגמה הראשונה יומיומית במכוון. את תהליך ה-open enrollment לביטוחי בריאות הריצו כולו דרך ה-agent. ה-agent הכיר את כל התוכניות שמופיעות ב-PDF, ידע מי העובד שפונה אליו, ועזר לכל אחד לבחור תוכנית שמתאימה למצב שלו. החשיבות כאן היא בהפצה: מנהלים בונים skills פעם אחת, והם זמינים לכל הארגון [4:33]. Katelyn העירה שזה חוסך את המצב שבו כל עובד מדביק את ה-PDF בנפרד ל-Claude, ושהעובדה שההמלצה מגיעה מהחברה נותנת לעובד ביטחון בבחירה [6:03].

הדוגמה השנייה היא Kate Bench, שנקרא על שם Kate, העורכת הראשית (בתמליל השם מופיע גם כ-Cape Bench וכ-K-Bench). כש-Kate הצטרפה היו ב-Every ארבעה עובדים. היום יש כ-30, ויוצא פי 10 תוכן [6:03]. כל עובד יכול לתייג את ה-agent ולבקש "Kate pass" על כתבה, דף נחיתה או אימייל. ה-agent עורך ישירות ב-Google Doc ב-suggested changes בשמה, ומי שכתב מקבל או דוחה את ההצעות. אחר כך עוקבים אחרי מה ש-Kate הוסיפה ידנית ומכניסים את זה חזרה ל-Kate. skill רואה גרף של הזמן שנחסך לה והצעות לכללים חדשים [7:36]. לפי Dan, זה משנה את השאלה מ"האם זה ייקח לי את העבודה" ל"איך מומחה מגדיל את ההשפעה שלו בלי להשקיע יותר זמן".

— **Dan Shipper (ייחוס משוער) [7:36]** So what I did was I just took like 30,000 of her historical edits and turned it into a skill that is in the Every Agent

4. benchmarks אישיים: מ-vibe checks לבדיקות מקודדות

לפני כל השקת מודל, Every מקבלת גישה מוקדמת ובודקת אותו ידנית על ה-workflows שלה. ביום ההשקה הם מפרסמים vibe check, והקוראים סומכים על הטעם שלהם. הבעיה היא שהתהליך ידני ואיטי [9:08]. הפתרון שהם בונים: בכל פעם שעובד שם לב לתוצאה שאהב או לא אהב, הוא שומר אותה במקום מרכזי ומסביר לפלטפורמה למה. הפלטפורמה ממירה את ההסבר ל-checks, כלומר כללים בסגנון unit tests. דוגמה מ-Kate Bench: האם הכותרת כתובה ב-sentence case. כך אפשר להריץ כמה מודלים על אותו תרחיש ולראות מי עומד בטעם של הארגון [10:41]. הפלטפורמה הזו עדיין לא חלק מה-Every Agent: "will eventually be part of the Every Agent".

בהמשך Dan מנסח את זה דרך אנלוגיה לגיוס עובדים. מודל שנכנס לחברה הוא כמו בוגר מבריק ביום הראשון שלו: הוא לא יודע איך עובדים אצלך. המטרה היא מערכת שבה ביום העשירי הוא טוב יותר מאשר ביום הראשון, ושאפשר להוכיח את השיפור הזה [21:24]. גם זה מוגדר במפורש כחזון ולא כיכולת קיימת: "that won't be in the agent on day one" [21:24].

— **Dan Shipper (ייחוס משוער) [21:24]** But if everyone got a 1600, which is where we are with these models, there's a lot more I need to do. What I actually want is not your SAT score. I want a reference check. I want a work trial

5. Slack ו-agent אחד: למה ויתרו על agent אישי

הגרסה הראשונה הייתה agent לכל עובד, זמין בערוץ שהעובד בוחר (Telegram, Slack). Willie Williams מסביר שהם עברו מ-multi-agent ל-single agent כי השיפורים לא התפזרו בארגון. כל אחד השקיע ב-agent הפרטי שלו, ושם זה נשאר [10:41]. ה-agents עשו "real work" בעיקר כשכמה אנשים השקיעו יחד ב-skills וב-guidance של אותו agent [12:13].

הבחירה ב-Slack נובעת מאותו היגיון: זה המקום שבו השיחות כבר מתנהלות. בעבודה שיתופית מחוץ ל-Slack יש הרבה העתק-הדבק, למשל העתקה של השיחה ל-Claude Cowork והחזרה של התשובה לשרשור. כשה-agent נמצא באותו מקום שבו נמצאים העמיתים, השלב הזה נחסך. בהמשך השיחה מוסבר ש-Every Agent לא מחליף את Claude Code או Claude Cowork אלא משלים אותם: יש עבודה שעושים "in public" עם עמיתים, ויש עבודה שעושים לבד [24:27–23:00].

— **Willie Williams (ייחוס משוער) [10:41]** we found that agents got better via the same dynamic the more you invested in them

6. למה Claude Managed Agents: ניסויים, אבטחה וזהות

השיעור הראשון מהגרסה הקודמת היה ש-"[12:13] infrastructure is hard". הם ניסו לשפר מוצר ובמקביל לתחזק שכבת infra יציבה, וזה הסתבך מהר. Dan מודה שהוא חשב ש-"infra is not that hard anymore" וש אפשר פשוט לתת ל-Claude להקים ולנהל שרתים. Willie הזהיר אותו, ואחרי חודש Dan הודה שזה באמת קשה [13:46]. Claude Managed Agents נתן להם sandboxes, memory ו-session control, והצוות יכול להתמקד ב-interaction design.

היתרון שהם מתארים כ"הכי כן" הוא היכולת לבודד מופע managed agent שמשמש כ-staging או frontier, להעתיק אליו את ה-memories וה-durable state מ-production ולהריץ עליו ניסויים בלי לפגוע ב-production, אבל על אותה תשתית. הניסויים עוסקים בהורדת עלות הטוקנים ללקוחות ובמקרי קצה של שיחה שדורשים שיקול דעת "כמו של עמית", ובבדקים בשילוב של evals ושל [15:17] vibes.

— **Willie Williams (ייחוס משוער) [13:46]** we don't want to be an infra team, right? We want to focus our energy in the place where we think the future is going

אבטחה וזהות

לפי Willie, רוב האבטחה נובעת מה-isolation המובנה: ה-sandbox, ה-tool calls, הסביבות וה-memories מופרדים, ויש שליטה מלאה במי ניגש למה ומתי. session management משמש כדי לשאול, למשל, האם בקשה של Dan מקבלת את אותה תשובה כמו בקשה של Willie: אילו memories מצורפים לבקשה, והאם ה-sandbox משותף או נפרד [16:48]. Dan מוסיף נקודה שנוגעת גם לביצועים: מודל שלא מגיע עם "מחשב" שהוא יודע להשתמש בו פחות טוב. כשהמודל וה-harness מגיעים מאותו ספק, breaking changes בהנחיות ה-prompting של מודל חדש נספגים בעדכון של שורה אחת [18:20, 16:48].

בנושא הזהות, ה-North Star שלהם הוא שה-agent יתנהג כמה שיותר כמו עובד אנושי ב-Slack. בפועל הם הגיעו למודל כפול, שלדבריהם מתאים לאופן שבו אנשים רוצים להשתמש ב-agent [18:20–19:51]:

הקשר	מי ה-agent	מה זה אומר בפועל
שיחה ב-Slack	ישות נפרדת	memories מבודדים והבנה של מי הפונה ולמה יש לו גישה בחברה. זו שיחה בין שתי ישויות.
tool call או connection (שליפת מסמך, חיפוש)	ברוב המקרים בשם המשתמש	ה-agent פועל כ-extension of you, עם ההרשאות של מי שביקש.
אזורים אפורים	לא מוגדר במלואו	ההודאה: קשה לקודד שיקול דעת אנושי בגישה לדוחות ולמידע.

— **Dan Shipper (ייחוס משוער) [16:48]** there's not a clear separation between the model and the harness anymore. and a model that doesn't come with a computer that it knows how to use, like, it's just not as performant

7. לאן זה הולך: fidelity, אישיות ומה ש-Every מבקשת

המילה שבחר Dan לכיוון היא fidelity: agents שנאמנים למה שחשוב לך, כך שלא צריך להיות מעורב בכל צעד. זה מה שכבר קורה ב-programming, שבו העבודה עוברת מ-loop ל-loop: on the loop: עובדים על המערכת ומתערכים רק כשצריך, כמו מנהל [27:30]. בצד המוצר מתוכננים benchmark אישי לכל עובד, skills שמותאמים לטעם שלו והפצה שלהם בארגון. ניסוי שכבר רץ הוא Hands, computer use agent על המחשב של Dan, שה-Every יכול להפעיל [22:57].

נושא נוסף שחוזר בשיחה הוא אישיות ונורמות חברתיות. Claude רגיל להגיב לכל הודעה, כי ברוב המקרים מדברים איתו אחד על אחד. בערוץ עם כמה אנשים, ואולי גם כמה Claudes, צריך להחליט מתי ה-agent מתערב, האם לחבר רעיונות משני שרשרים, והאם הוא "extrovert" או "introvert" [24:27–25:58]. קריטריון התכנון שלהם: עבודה עם ה-agent צריכה להרגיש כמו עבודה עם עובד של Every. הבקשות מ-Anthropic: לא לשכוח את עובדי הידע לטובת הישגים במתמטיקה ובביולוגיה, לשפר את ההתאמה של התשובות לרמת הידע של המשתמש, ולתת כלים להגדיר אישיות עקבית אך ניתנת לשינוי, למשל אגרסיבית ללקוח אחד וסבלנית לאחר [29:00–30:31]. ה-hot take של Dan: ככל שמשתמשים יותר ב-agents, יש יותר עבודה ולא פחות [30:31].

— Dan Shipper (ייחוס משוער) [29:00] We're starting to reach the point where successive gains in model intelligence are not necessarily able to be consumed, even by power users

8. מסקנות ופעולות

- **התחילו מ-agent ארגוני אחד ולא מ-agent לכל עובד:** כך ההשקעה ב-skills וב-guidance מצטברת ומגיעה לכל הארגון.
- **זהו את ה-Kate שלכם:** מומחה שכולם תלויים בו. אספו את ההיסטוריה שלו (עריכות, code reviews, תשובות) ובנו ממנה skill עם לולאה שמחזירה את התיקונים שלו לתוכו.
- **כתבו benchmark אישי לפני החלפת מודל:** אספו דוגמאות טובות ורעות מהעבודה היומיומית, המירו אותן ל-checks בסגנון unit tests, והריצו עליהן כל מודל חדש.
- **הפרידו staging מ-production על אותה תשתית:** העתיקו state ו-memories למופע ניסוי לפני שמשנים prompts, אישיות או מנגנונים שמשפיעים על עלות.
- **הגדירו מודל זהות מפורש:** מתי ה-agent הוא ישות נפרדת ומתי הוא פועל בשם המשתמש, ומה הוא עושה באזורים האפורים.
- **לפני שבונים infra בעצמכם:** שאלו אם זה באמת היתרון התחרותי שלכם. Every למדו שזה קשה רק אחרי חודש של ניסיון.

שאלה פתוחה	למה זה חשוב
איך מונעים דליפה בין משתמשים כשה-agent פועל בשם מי שביקש בערוץ משותף?	התשובה מ-tool call עלולה להיות גלויה לאנשים בערוץ שאין להם הרשאה למידע.
איך מודדים את ה-skill של Kate ולא רק מרגישים אותו?	לא נמסרו שיעור קבלה של תיקונים, זמן שנחסך או ירידה באיכות.
איך מגנים מפני prompt injection ממסמכים שה-agent קורא?	ה-agent ניגש למסמכים בשם המשתמש, ולא נאמר כלום על הסיכון הזה.
מתי agent צריך להתערב בשיחה קבוצתית?	הם עצמם מצפים לפתרון בעוד שישה חודשים עד שנה.
כמה עולה agent ארגוני לכל משתמש?	הוזכרו ניסויים בהורדת עלות טוקנים, בלי מספרים.

9. מילון מונחים

מונח	משמעות
Claude Managed Agents	פלטפורמה מנוהלת של Anthropic להרצת agents, עם session control ו-sandboxes, memory מובנים.
AI coworker	agent שמשתלב בארגון כמו עמית: מקבל בקשות, זוכר הקשר ועושה משימות חוזרות.
AI ambassador	עובד שכבר עובד עם AI ברמה גבוהה ומנסה להפיץ את שיטות העבודה שלו בארגון.

מונח	משמעות
Skill	יחידת ידע או הנחיה שה-agent טוען כדי לבצע סוג מסוים של עבודה, למשל עריכה בסגנון של Kate.
Kate Bench / Kate pass	הבדיקות וה-skill שנבנו מ-30,000 עריכות של העורכת הראשית של Every.
Vibe check	בדיקה ידנית ואיכותנית של מודל חדש על workflows אמיתיים לפני ההשקה שלו.
Evals / checks	בדיקות אוטומטיות בסגנון unit tests שמקודדות טעם, למשל כותרת ב-sentence case.
Harness	השכבה שעוטפת את המודל: כלים, סביבת הרצה, ניהול הקשר.
Sandbox	סביבת הרצה מבודדת שבה ה-agent מריץ קוד וכלים.
Session management	ניהול ההקשר של כל בקשה: אילו memories מצורפים ואיזה sandbox משמש.
Fidelity	מידת הנאמנות של ה-agent למה שחשוב למשתמש, כך שלא צריך לפקח על כל צעד.
Computer use agent	agent שמפעיל מחשב כמו אדם (עכבר ומקלדת). אצל Every זה ניסוי בשם Hands.
OpenClaw	agent אישי ש-Every הריצו על Mac minis לפני Every Agent (השם כפי שמופיע בתמליל).