

מומנטום מצטבר: ארבעה סוגי סוכנים

AWS Summit Tel Aviv 2026 — סיכום טל, tlv-key101

שחר דניאל, AWS Specialist Solutions Architects Leader | ליאת צור, Principal Technical Account Manager, AWS

אורחים: פרופ' אלי שפרכר, מנכ"ל המרכז הרפואי תל אביב (איכילוב) | אלד ולך, מייסד-שותף ומנכ"ל Aidoc

10 בספטמבר 2026 | משך: כ-66 דקות | הופק מהקלטת הסשן

סיכום מאת קובי אלמוג

על המסמך הזה

המסמך מסכם את הקינוט הראשי של AWS Summit Tel Aviv 2026, בהתבסס על ההקלטה המלאה שפורסמה באתר הסאמיט. הקינוט ארך כשעה ושש דקות וכלל חמישה דוברים: פתיחה של הנהלת AWS ישראל, ראיון במה עם מנכ"ל איכילוב, שני בלוקים טכניים של AWS, והרצאת לקוח של Aidoc.

המסמך נבנה מתמלול אוטומטי של ההקלטה ומהסליידים שחולצו מהווידאו עצמו, כך שאפשר לחזור לתוכן בלי לצפות שוב בהקלטה כולה. חותמות הזמן בגוף המסמך מפנות לנקודה המדויקת בהקלטה.

איך לקרוא את זה

- **פרק 1 — תקציר מנהלים:** חמש דקות קריאה. מה הוכרז, מה היה סיפור לקוח, ומה רלוונטי לארגון גדול.
- **פרקים 2–3 — הפתיחה וסיפור איכילוב:** החלק הלא-טכני, אבל שם נמצאים המספרים החזקים ביותר בקינוט.
- **פרק 4 — המסגרת:** ארבעת סוגי הסוכנים שעליהם נבנה כל שאר הקינוט.
- **פרקים 5–9 — התוכן הטכני:** Amazon Quick, סיפור Kiro, Aidoc ו-AWS Continuum, DevOps Agent, ולבסוף AgentCore והקונטקסט.
- **פרק 10 — מה לוקחים מכאן:** מסקנות, ומילון מונחים למי שנתקל בשמות המוצרים החדשים לראשונה.

הערת מקור המסמך הופק מתמלול אוטומטי של ההקלטה. התמלול מדויק ברובו, אך שמות ומונחים לועזיים הופיעו בו בשיבוש פונטי ("קירו" = Kiro, "אג'נט קור" = AgentCore, "בדרוק" = Bedrock) ותוקנו ידנית מול הסליידים. הציטוטים אומתו מול ההקלטה בחותמת הזמן המצוינת ומובאים כלשונם, כולל שיבושי תמלול קלים. מספרים שמופיעים על הסליידים צוינו ככאלה, ובמקום אחד הם נבדלים ממה שנאמר בעל פה — ההבדל מסומן במפורש.

1. תקציר מנהלים

הקינט בנוי סביב טענה אחת: AI יוצר "מומנטום מצטבר" — ככל שמשתמשים בסוכן יותר, כך הוא מקבל יותר קונטקסט, מחזיר תוצאות טובות יותר, זוכה ליותר אמון ומשמש יותר. כל ההכרזות מסודרות בארבע קטגוריות של סוכנים, ובין הבלוקים הטכניים שובצו שני סיפורי לקוח ישראליים מעולם הרפואה.

- **המסגרת: ארבעה סוגי סוכנים, לא רשימת מוצרים.** סוכנים שמשנים איך עובדים (Amazon Quick), איך בונים (Kiro, DevOps Agent, AWS Transform), איך מאבטחים (AWS Continuum), והסוכנים שהארגון בונה בעצמו (AgentCore). כל הכרזה בקינט נתלית באחת מהארבע.

- **עיקר החידוש הוא בשכבת הקונטקסט, לא במודלים.** ההכרזות המשמעותיות — AWS Context, Managed Knowledge Base, Web Search ב-Bedrock — כולן עונות על אותה בעיה: "גם למודל החכם ביותר אין מושג איך הארגון שלכם באמת עובד" [51:40].

- **האבטחה מוצגת כמשפחת מוצרים, לא כפיצ'ר.** AWS Continuum אוגד ארבעה שירותים: Code ו-Pen testing review כבר בזמינות כללית, Threat modeling ו-Code vulnerabilities ב-Preview (סליידים k_236).

- **סיפורי הלקוח נשענים על מספרים בדוקים.** איכילוב: מ-100 ל-2,000 משתמשים במערכת Gecko תוך 12 חודשים (סלייד). Aidoc: פטה-בייט של 2,000 DICOM, בתי חולים, 75 מיליון סריקות בשנה, +40 אישורי FDA (סלייד k_175, k_199).

- **ההוכחה הישראלית לצד הבנייה: בנק הפועלים.** "עם קירו, בנק הפועלים תרגימו את האפיון העסקי ישירות לקוד מובטח ויציב, והשיקו את פועלים ג'וניאר תוך 35 ימים בלבד" [39:23] — מול לוח זמנים של +12 חודשים שהוצג על הסלייד.

- **התובנה החוזרת: הקצב של הסוכנים שובר את התהליכים שמסביבם.** אם הקוד נכתב פי עשרה מהר יותר אבל הפייפליין, החוב הטכני והבקלוג של האבטחה לא מדביקים — המומנטום נעצר. משם נגזרות ההכרזות על Continuous Modernization ועל Code Vulnerabilities.

רלוונטיות לארגון גדול

שלושה דברים בקינט חורגים מרמת ההכרזה ונוגעים ישירות בארגון מפקח: מודל האמון המדורג של Code Vulnerabilities (מתחיל ב-Human in the Loop ורק אחר כך עובר לאכיפה אוטומטית), בקורות ההרשאות ברמת המסמך ב-Managed Knowledge Base, והבטחה מפורשת ש-Web Search רץ בתוך סביבת ה-AWS כך שהשאלות לא יוצאות החוצה. אלה בדיוק שלוש השאלות שנשאלות בכל ועדת אבטחת מידע לפני אימוץ סוכן.

2. הפתיחה: ריג'ון, קהילה, ו-13 שנות סאמיט

הקינוט נפתח בבילוק לא-טכני של הנהלת AWS ישראל. הנקודות העובדתיות: ריג'ון תל אביב הושק ב-2023, זהו הסאמיט ה-13 בישראל, והשנה נכחו בו "14 אלף משתתפים שהגיעו לסאמיט היום" [5:09]. לצד זה הוצגו מעל תשעים סשנים ועשרות פעילויות נוספות [12:53].

המסר האסטרטגי נוסח כהחלפת שאלה: "אתם מפסיקים לשאול האם אני יכול לבנות את זה? והשאלה הופכת להיות מהו הדבר הבא שאבנה" [3:35]. זו אותה טענה שתחזור בסיום הקינוט בניסוח "להסיר כל חסם שעומד בין הדמיון למציאות" [1:04:04].

3. איכילוב: בית חולים שהעביר את כל המערכות הקליניות לענן



פרופ' אלי שפרקר עולה לבמה — מנכ"ל המרכז הרפואי תל אביב, איכילוב

הראיון נפתח בשלושה מספרים על סדר הגודל של בית החולים: "הם מטפלים בשני מיליון מטופלים בשנה. הם עושים ארבעים אלף ניתוחים בשנה. ויש להם שלוש עשרה אלף לדות בשנה" [6:40]. לפני שנתיים התקבלה ההחלטה להעלות את כלל מערכות בית החולים ל-AWS.

מה המעבר לענן אפשר

התשובה של שפרקר לא הייתה על עלויות או על זמינות אלא על יכולות חדשות: "אבל מה שהמהלך הזה בעיקר אפשר זה לפתח יכולות שלא היו לנו" [6:40]. גולת הכותרת שהוצגה היא כלי בשם Gecko, שפותח בתוך יחידת ה-AI של בית החולים ומתואר כמעטפת סביב התיק הרפואי, שמציגה נתונים מותאמים למטופל ולנקודה המדויקת במסע הטיפולי שלו.



קצב האימוץ של Gecko כפי שהוצג על הבמה: מ-100 משתמשים בתוך 12 חודשים

הערך שנטען עבור המערכת הוא בהתראות שהיא מספקת למטפל, ולפי הדובר קיימים מקרים מתועדים של הצלת חיים בזכותן: "מדובר פה בהצלת חיים בשיפור טיפול רפואי" [8:12]. ההערה החדה של המראיין הייתה שאיכילוב עדיין יוצא דופן: "איך יכול להיות שאיך ילוב הוא עדיין בית החולים היחיד במדינת ישראל שאלה את כל המערכות הקליניות שלו לענן" [8:12].

שלוש הסיבות שניתנו למעבר

- **טיפולים מתקדמים.** בדוגמה שניתנה, החזרת יכולת דיבור לחולי ALS — מהלך שלפי הדובר "מחייב עבודה בענן" [8:12].
- **מחקר.** בתי חולים גדולים הפכו למרכזי מחקר, והדוגמה שניתנה היא גנומיקה: נפחי דאטה שלא ניתן לעבד בלי סביבה עננית [9:42].
- **הכשרה.** 11 אלף סטודנטים בשנה. "אנחנו לא יכולים ללמד אותם את הרפואה של העבר, אנחנו צריכים ללמד אותם את הרפואה של העתיד" [9:42].

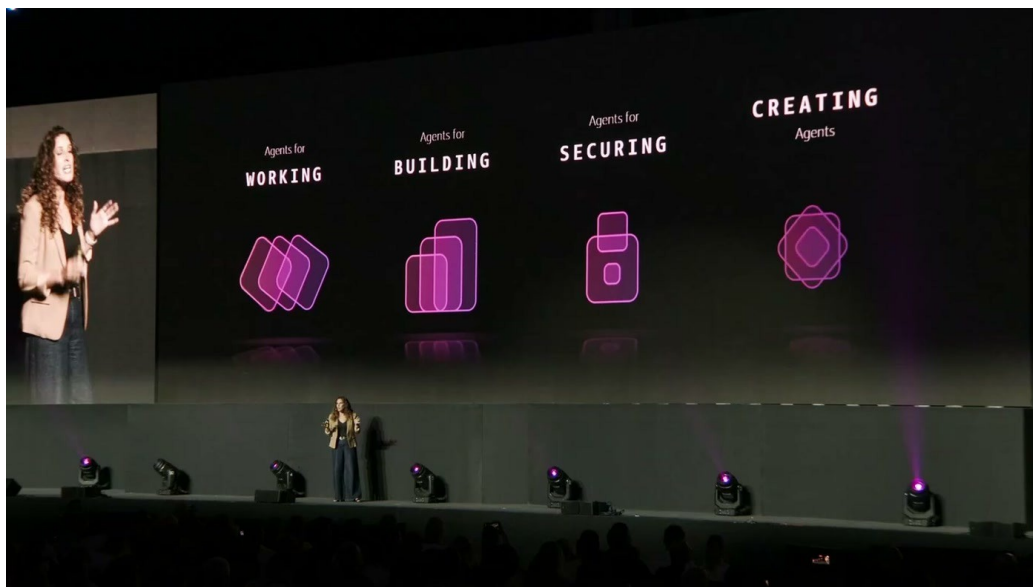
Accountable AI

ההכרזה המעניינת ביותר בבלוק הזה לא הגיעה מ-AWS. שפרכר תיאר מהלך שבית החולים משיק: "אנחנו קראנו למהלך הזה Accountable AI, אנחנו עומדים לפרסם לכל המטופלים שלנו את ביצועי אותן מערכות AI שהיום מטפלות במטופלים שלנו לצד המטפלים האנושיים" [9:42]. ההיגיון שהוצג: אם כבר שלושה עשורים מודדים ומפרסמים מדדי איכות של מטפלים אנושיים, אין סיבה לא למדוד באותו אופן את המטפלים הלא-אנושיים. על השאלה אם AI יחליף רופאים, התשובה הייתה "כן ולא" עם אנלוגיה לחקלאי שהחליף חמור במחשב וברובוט אך נשאר חקלאי. החסם האמיתי, לדבריו, אינו טכנולוגי: "כולם יגידו מצוקת כוח האדם" [11:23]. השורה שסגרה את הבלוק: "כנראה שרופא ואיי איי זה השילוב הנכון" [12:53].

4. המסגרת: מומנטום מצטבר וארבעה סוגי סוכנים

שחר דניאל, שמובילה את קבוצת ה-Specialist Solutions Architects בישראל, פתחה את הבלוק הטכני בתזכורת שבשנה שעברה דובר על עידן הסוכנים כהבטחה — ובשנה שחלפה הם הגיעו לפרודקשן. המנגנון שהיא תיארה הוא לולאה מזינה את עצמה:

— **שחר דניאל [19:05]** יותר אינטראקציות, פירושן יותר קונטקסט. יותר קונטקסט, פירושו תוצאות טובות יותר. ותוצאות טובות יותר, פירושן אמון. אמון בסוכנים ושימוש רב יותר בהם. זה מה שנקרא Compounding Momentum.

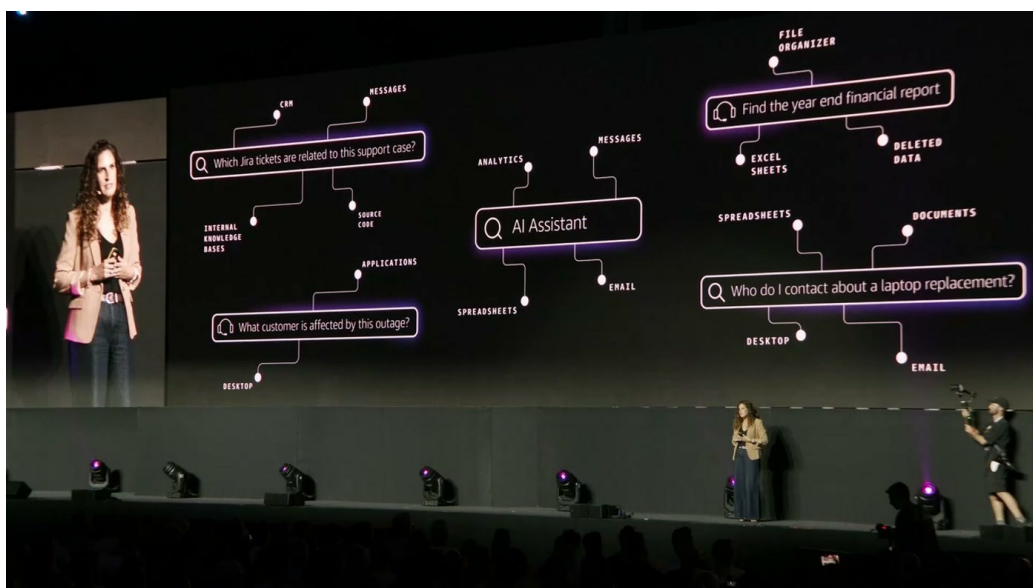


ארבע הקטגוריות שעליהן נבנה שאר הקינט: סוכנים לעבודה, לבנייה, לאבטחה, וסוכנים שהארגון יוצר בעצמו

ארבע הקטגוריות הן גם סדר ההופעה של ההכרזות. שווה להחזיק אותן בראש בזמן הקריאה, כי הן מסבירות למה מוצר כמו Amazon Quick ומוצר כמו AgentCore מופיעים באותה הרצאה: הראשון נועד לצורך, האחרון נועד לבנות.

5. סוכנים שמשנים איך אנחנו עובדים: Amazon Quick

האבחנה שפתחה את הפרק: מה שעומד בין עובד לבין תוצאה הוא לא חוסר יכולת אלא פיזור בין כלים. כל כלי נוסף הוא עוד לוגין, עוד מקור דאטה, ועוד מקום שבו הקונטקסט הולך לאיבוד.



המחשה של הפיזור: שאלות עבודה יומיומיות שכל אחת מהן נמתחת בין CRM, מיילים, גיליונות, קוד מקור ודסקטופ

הביקורת על הדור הקודם של פתרונות הייתה ישירה. עוזרי AI שחוברו לכלים הארגוניים "עונים על שאלה אחת ואז הם שוכחים" [20:38] — הם פועלים מנותקים מהקשר, וכשהבטחה פגשה עבודה אמיתית "ההתלהבות פגשה את המציאות".

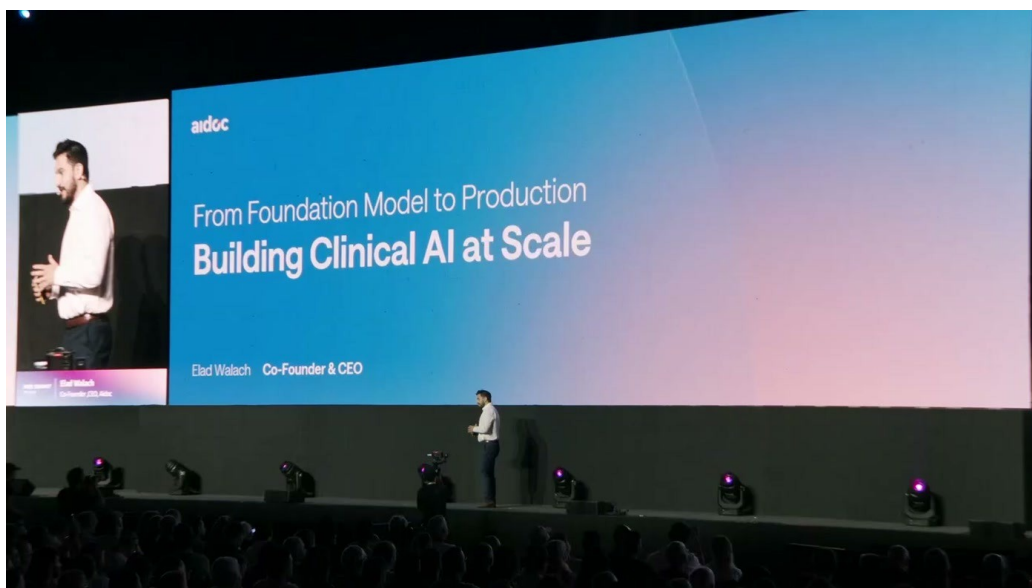
מכאן הוצג Amazon Quick: "Amazon Quick הוא AI אסיסטנט המאפשר לכם להגדיר מהי התוצאה הרצויה ואז מפצח לבד את הדרך להגיע אליה" [22:12]. תחתיו שכבת חיפוש אג'נטית שרצה לרוחב הדאטה הארגוני ולומדת את היחסים בין אנשים, תכתובות, מסמכים ומקורות מידע.

הדמו: חזרה מחופשה

ליאת צור, Principal Technical Account Manager בארגון הספורט הגלובלי של AWS, הדגימה שלושה תרחישים רצופים. הפתיחה הייתה תיאור מצב מוכר: "אבל אז חזרתי, אלף מאתיים מיילים" [24:02], לצד עשרות הודעות ופגישות.

- **ברף עדכון.** Quick שולף מ-Slack, ממיילים, מדשבורדים ומקבצים מקומיים, מקשר ביניהם ומחזיר סיכום מחולק לנושאים — מה התקדם, מה תקוע, ומה ממתין לתשובה, מתועדף לפי דחיפות.
 - **הכנה לשבוע.** מעבר על היומן והפקת תדריך לכל פגישה: תאריך, סטייקהולדרס, מה נעשה עד כה ואילו תשובות כדאי להביא — על סמך פגישות קודמות ומסמכים. "עכשיו אני יכולה להגיע לפגישות שלי מוכנה וחדה" [25:34].
 - **ניסוח תשובות.** טיוטות מיילים והודעות בסגנון האישי של המשתמשת — הטון, הסגנון ואופן התקשורת עם לקוחות — לאישור לפני שליחה.
- סיכום הדמו: "בדקות בודדות התעדכנתי, התכוננתי לפגישות ועניתי להודעות הכל ממקום אחד" [25:34]. בהמשך נאמר שסוכני Quick האוטונומיים החדשים גם יבצעו את הפעולות עצמן, ולא רק יכינו אותן.

6. Aidoc: מ-foundation model רפואי לפרודקשן קליני

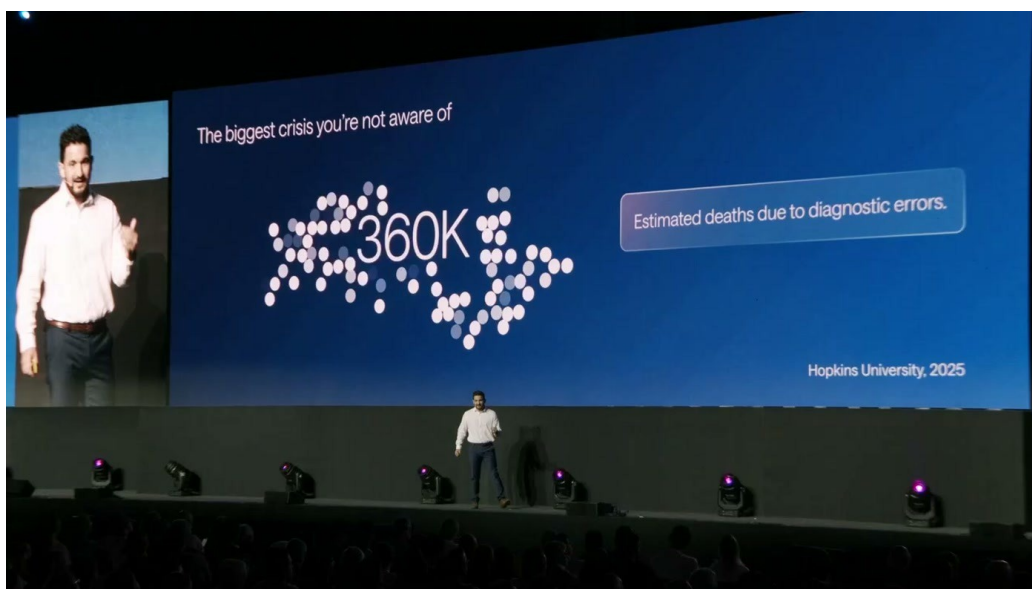


פתיחת הרצאת הלקוח של Aidoc — אלד ולך, מייסד-שותף ומנכ"ל

הבלוק הזה היה ההרצאה הטכנית-עסקית החזקה ביותר בקינוט, והוא גם היחיד שהציג מסלול מלא: בעיה, מגבלה, הימור, פריצת דרך, ותשתית.

הבעיה

הפתיחה הייתה חדר מיון: מבינים מה לא בסדר, ואז מחכים. "מחכים לא בגלל שאין לנו רופאים מדהימים מחכים בגלל שיש לנו מערכת שעל סף קריסה" [28:38]. הדובר הגדיר זאת כ"הבעיה הכי גדולה בעולם שאולי לא שמעתם עליה" [28:38].



המספר שהוצג על המסך: 360K מקרי מוות משוערים בשנה כתוצאה מטעויות אבחון, מיוחס ל-Hopkins University, 2025

שימו לב אי-התאמה בין הנאמר לנכתב: על הסלייד מופיע 360K, ואילו בעל פה נאמר "כל שנה בארצות הברית הם מעריכים שיש 400 אלף נפטרים בגלל תאוויות ועיקובים רפואיים" [28:38]. ההנרמול שהוצג לישראל: "זה מעל עשרת אלפים איש בשנה בארץ עם מוות שהיה יכול להימנע" [30:08]. מומלץ לצטט את מספר הסלייד, שהוא המקור המתועד.

המגבלה: דיוק מול כיסוי

האנלוגיה שניתנה לקריאת CT: מאות תמונות ברזולוציה גבוהה שבהן המחלה עשויה להיות כמה פיקסלים בגוון שונה במעט — "כמו איפה או איפה, רק שבמקום יש לך איפה אחד בעמוד, יש לך איפה אחד בספר שלם, ולא רק זה, יש לך שלוש דקות למצוא את איפה" [30:08]. הסלייד ניסח זאת כ-Every Pixel Counts.

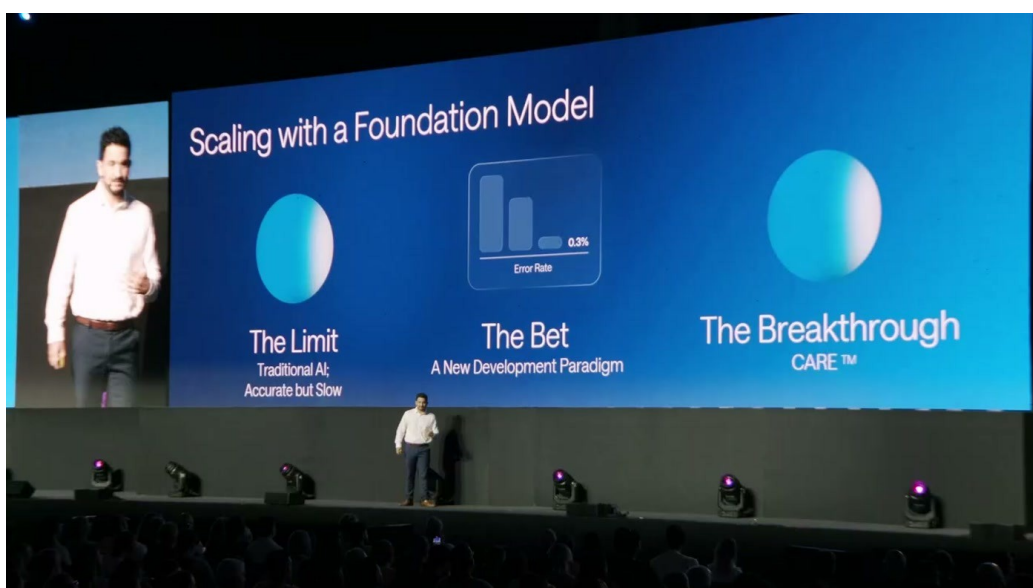
המחיר הארכיטקטוני של הדיוק היה כיסוי: "בשביל להגיע לרמות הדיוק היינו חייבים כל הזמן להתמקד במחלה בודדת" [31:40] — סרטן, שבץ, שברים בעמוד שדרה — בעוד שבגוף האנושי אלפי מחלות. המספרים שהוצגו על הסלייד למצב הנוכחי: מעל 500 מיליון דולר גיוס, +40 פתרונות מאושרי FDA, 75 מיליון סריקות בשנה, 2,000 בתי חולים, ו-2 מיליון מטופלים ישראלים.



פריסה ומספרי סקניל של החברה כפי שהוצגו על המסך

ההימור ופריצת הדרך

ההשראה שצוינה היא נובמבר 2022 והופעת ChatGPT. ההחלטה שהתקבלה, בלשון, "אולי בחוצפה ישראלית טיפוסית", הייתה להקים מעבדת frontier ולבנות foundation model רפואי בישראל, בהשקעה של מאות מיליוני דולרים [33:10]. שתי השנים הראשונות לא הגיעו לדיוק הנדרש; רק באמצע השנה שעברה הושג הסף. המודל נקרא CARE.



שלושת השלבים כפי שהוצגו: המגבלה של AI מסורתית, ההימור על פרדיגמת פיתוח חדשה, ופריצת הדרך עם CARE

התוצאה הוצגה כגרף של אישורי FDA לפי שנה, כאשר כל נקודה לבנה היא מחלה: "בשלושת החודשים הראשונים של השנה קיבלנו יותר אישורי FDA מאשר בכל השש השנים שקדמו לכך" [33:10]. הסלייד שסגר את החלק הזה מציין שהמוצר Built On CARE קיבל FDA Breakthrough Device Designation.

האנלוגיה למוצר הדור הבא, שנמצא בניסויים קליניים, הייתה מוכרת לקהל מפתחים: "הדבר הזה הוא כמו הקורסר של עולם הרפואה" [34:44] — המערכת מסתכלת על כל הדאטה הקליני ועל התמונה ומייצרת טיוטת דוח, כשם שמפתח כבר לא כותב קוד מאפס. וההסתייגות שנלוותה לכך: "ה-AI פה הוא לא הגיבור הגיבור והגיבורה זה הרופא" [34:44].

התשתית



שלושת מספרי התשתית: נפח ה-DICOM, שיפור היעילות בארבע השנים האחרונות, וקצב הגידול השנתי בסריקות

ארבעת הצרכים שפורטו: "היינו צריכים להקים דאטה בייס רפואי הגדול בעולם 37 פטאבייטס", להריץ אימונים בעשרות מיליוני דולרים, לאופטם אותם, ולתמוך בסקייל שמוכפל מדי שנה [34:44]. הסלייד מדייק: 37.60 פטה-בייט, מאגר ה-DICOM הגדול בעולם, גדל בכ-1PB בחודש; שיפור יעילות של 75% בארבע השנים האחרונות; וגידול של פי 2.4 בשנה במספר הסריקות המנותחות.

הסיכום שלו על מצב התחום: "אני חושב שזה יהיה בלתי מתקבל לדעת שרופא לא ישתמש ב-AI" [36:17] — באותו אופן שבו לא נכנסים לרכב בלי חגורת בטיחות.

7. סוכנים שמשנים איך בונים: Kiro, DevOps Agent, AWS Transform

הפרק נבנה סביב מחזור בן שלושה שלבים: קוד נכון מההתחלה, עלייה מהירה לפרודקשן, ומודרניזציה מתמשכת. ההוכחה הפנימית שהוצגה מגיעה מ-Amazon Stores: "עם הגישה הזו שיפרנו ב-4.5 את המהירות בקוד מגיע לפרודקשן", ו"ראינו רמות דיוק של 95% בשינויים שבוצעו לדי איי שזה יותר מהבאיסלין האנושי" [37:52].

קוד נכון: Kiro

הטענה: כתיבת קוד הפכה זולה, ולכן הצוואר עבר למקום אחר. "יש סיכוי גדול בהרבה שקוד ללא קונטקסט מלא לסביבה שלכם, לסטנדרטים שלכם ולארכיטקטורה שלכם יסבול מבגים. וקוד עם בגים הוא לא קוד מהיר, הוא חוב טכני שמצטבר במהירות" [37:52]. Kiro הוצג כסביבת פיתוח אג'נטית שהופכת prompt לדרישות, לדיזיין, למשימות יישום ולוולידציות.



המקרה של בנק הפועלים: לוח הזמנים המקורי שהוצג על הסלייד לצד אפליקציית פועלים ג'וניור

הפרויקט שהוצג — פלטפורמה בנקאית דיגיטלית לילדים, הדורשת אינטגרציה רחבה, הרשאות בזמן אמת וחיבור לארנקים דיגיטליים תחת רגולציה בנקאית: "עם קירו, בנק הפועלים תרגימו את האפיון העסקי ישירות לקוד מובטח ויציב, והשיקו את פועלים ג'וניור תוך 35 ימים בלבד" [39:23].

ההכרזה בפרק הזה היא KIRO כאפליקציית iOS. הנימוק הובא מהשטח: "מפתחים פשוט מפחדים לחבות את הלפטופ שלהם מחשש שהסוכנים שלהם יפסיקו לפעול" [39:23]. הסשנים רצים בסביבת הענן של הלקוח, וכשחוזרים ללפטופ ממשיכים מאותה נקודה.

פרודקשן מהיר: AWS DevOps Agent

הסוכן הוצג כחבר צוות Ops זמין תמיד — חקירת אירועים, אופטימיזציה של ביצועים ואמינות, ומשימות SRE. "חברות כמו T-Mobile ו-United Airlines משתמשות ב-DevOps Agent כדי לאתר ולטפל בבעיות בפרודקשן פייליין שלהם עוד לפני שהלקוחות שלהם מושפעים" [40:56].

ההסתייגות שהוצגה מיד אחריה היא גם ההכרזה: הסוכן הזה ריאקטיבי, ו"מומנטום אמיתי לא בא מתיקון בעיות לאחר שהן קרו" [40:56]. לכן הוכרז Release Management בתוך DevOps Agent, שקולט את הקוד ברגע שנוצר — בין אם ב-KIRO ובין אם ב-Claude Code — סוקר, מריץ בדיקות ומתקן לפני שהתקלה מגיעה לפרודקשן [42:29].

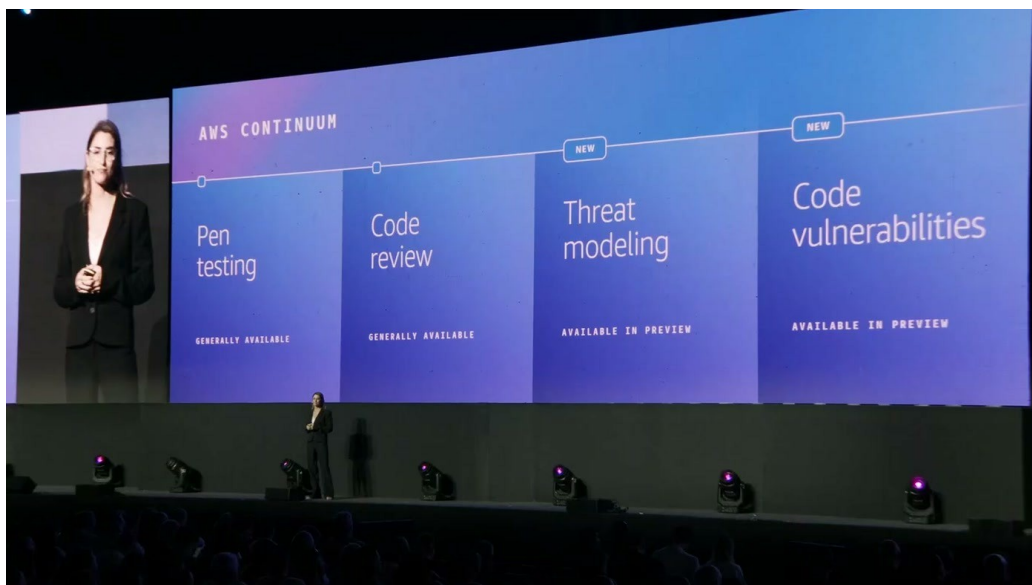
מודרניזציה: AWS Transform Continuous Modernization

הבסיס: AWS Transform הוצג בשנה שעברה, ומאז "לקוחות סביב העולם השתמשו בטרנספורם וחסכו יותר מ-1.7 מיליון שעות עבודה ידנית לצרכי מודרניזציה" [42:29]. הפער שנותר הוא שהוא מופעל ביוזמת המשתמש, ובקצב שבו סוכנים מייצרים קוד זה לא מספיק.

Continuous Modernization רץ ברקע בזמן שהסוכנים כותבים קוד, מוצא נקודות בעייתיות, מתקן, מוולד ולומד מכל טרנספורמציה. הוא מתחבר ל-GitLab ו-GitHub Actions, Jenkins, CodePipeline. הזווית שהופכת אותו לרלוונטי לאבטחה ולא רק לתחזוקה: "כל דיפנדנסי לא התקני היא חולשה אבטחתית הממתינה לסי וי שלה" [44:01] — כלומר CVE שלה.

8. סוכנים שמשנים איך מאבטחים: AWS Continuum

המעבר שהוצג הוא מכלים שנבנו עבור בני אדם — טלמטריה, אחסון, שאילתות, דשבורדים — לכלים שנבנו עבור סוכנים: טלמטריה, קונטקסט, ריזנינג, פעולות. "למעשה, טלמטריה ללא קונטקסט היא רק רעש רקע" [45:35].



ארבעת שירותי המשפחה ומצב הזמינות של כל אחד: שניים בזמינות כללית, שניים ב-Preview

AWS Continuum הוגדר כאוסף מוצרים לאבטחת ה-workload. שני הראשונים, Pen testing ו-Code review, הושקו בדצמבר האחרון כחלק מ-AWS Security Agent. שני החדשים הם Threat modeling ו-Code vulnerabilities.

Threat modeling

הבעיה שהוצגה: מודלי איומים נבנים ידנית על ידי מומחי אבטחה, אם בכלל, והם מתיישנים ברגע שהארכיטקטורה משתנה. "המאמץ הוא פשוט לא סקילבילי ולכן תראו בעצמכם, רוב הצוותים מדלגים על השלב הזה לחלוטין" [45:35]. השירות מייצר מודלים הממופים לארכיטקטורת האפליקציה, לקוד המקור ולמסמכי הדיזיין, ומספק מיטיגציות על בסיס security frameworks מוכרים.

Code vulnerabilities

ההנמקה: מודלים מתקדמים מאתרים חולשות מהר מאי פעם, ו"הבקלוג הזה גדל בצורה אקספוננציאלית" [47:06]. השירות פועל בארבעה שלבים:

- **קליטה.** קליטת הבקלוג הקיים לצד סקירת הסביבה, כדי לייצר מבט מקיף על החולשות ועל נתיבי התקיפה הקשורים אליהן.
- **העשרה ותעדוף.** שימוש בקונטקסט מהסביבה כדי להעשיר ולתעדף ממצאים.
- **ניצול עצמי.** ניסיון לנצל את החולשות בפועל, כדי לחשוף false positives.
- **תיקון.** הצעת תיקונים באופן פרואקטיבי.

הנקודה החשובה לארגון מפוקח היא מודל האמון המדורג: "Code Vulnerability מתחיל במוד לימוד עם Human in the Loop", וככל שהאמון נבנה אפשר להתקדם למוד אכיפה שמבצע תיקונים אוטומטית [48:37].

9. הסוכנים שאתם בונים: AgentCore והשאלה של הקונטקסט

האבחנה שפתחה את הפרק האחרון: "סוכנים רבים שבניתם תקועים במעבר בין פרוטוטיפ לפרודקשן" [50:08] — לא בגלל הלוגיקה אלא בגלל התשתית. "אני רואה לקוחות שמקדישים חודשים של עבודה לבניית אוטנטיקציה, זיכרון, גישה לכלים, הבטחה כמו ב-governance ועוד" [50:08].

Amazon Bedrock AgentCore מספק את אותם אבני בניין: runtime מנוהל, זיכרון ששן מובנה, ואוטנטיקציה מאובטחת לכלים, ל-APIs ולמקורות דאטה — והוא עובד עם כל framework אג'נטי וכל מודל. שלוש דוגמאות לקוחות שהוזכרו: Sky הריצו מערכת סוכנים לשיווק דיגיטלי שמאפשרת ניהול אוטונומי של קמפיינים מעל 100 מיליון דולר; לקוח נוסף מעלה סוכנים לפרודקשן פי עשרה מהר יותר; ולקוח שלישי חוסך ללקוחותיו מאות שעות עבודה ידנית בחודש באוטומציה של הנהלת חשבונות [51:40].

למה קונטקסט הוא הבעיה האמיתית

— שחר דניאל [51:40] גם למודל החכם ביותר אין מושג איך הארגון שלכם באמת עובד. הוא לא מכיר את הלקוחות שלכם, את המערכות שלכם או הדאטה שלכם. וכשסוכן לא יודע את התשובה... הוא פשוט מחרטט, אבל במלוא הביטחון.

מכאן חולק הפרק לשלושה סוגי דאטה, וכל אחד קיבל הכרזה משלו.

דאטה פומבי: Web Search

הנימוק הוא ה-cutoff date של המודל: בניתוח תחרותי רוצים מידע על מוצר שהושק אתמול, לא על תמונת העולם מזמן האימון, ו"המידע העדכני ביותר נמצא ברשת" [53:15]. אותו Web Search שמניע את Quick, את Kiro ואת Alexa + זמין כעת ככלי מנוהל ב-Amazon Bedrock וב-AgentCore. הוא מחזיר תוצאות מדורגות, ו"פועל בתוך סביבת ה-AWS שלכם כך שהדאטה והשאלות שלכם לעולם לא יוצאים מ-AWS" [54:45].

דאטה ארגוני לא מובנה: Managed Knowledge Base

הבעיה נוסחה בכמות: התשובות נמצאות בשנים של מסמכים, תמונות וסרטונים שפזורים ב-OneDrive, Confluence, או SharePoint, "אבל הן קבורות בחמישה עשר מקומות ולפחות איזה עשרה פורמטים שונים" [54:45]. השירות מספק native connectors, מטפל בעיבוד, באחסון ובשליפה, ובקורות גישה שאוכפות הרשאות ברמת המסמך — כך שמשתמש מקבל תשובות רק מתוכן שהוא מורשה לו.

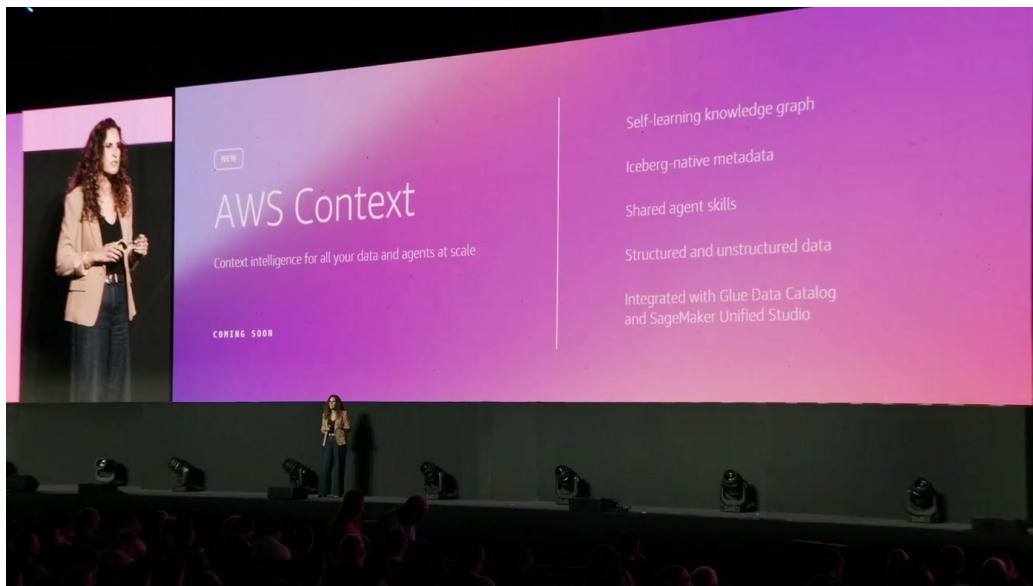
לשאלות מורכבות הוצגה שליפה אג'נטית: סוכן שמתכנן שאלות לרוחב כלל ה-knowledge bases, מקשר קונספטים בין מסמכים, ופועל בשלבים — שולף, בוחן, ומשפר את מה שמצא. "הוא לא פשוט שולף את הוקטור הקרוב ביותר" [56:22]. בין הלקוחות שנמנו: "עשרות לקוחות כבר בונים ממאנג' נולדג' בייס ביניהם OpenAI", שמשתמשים ביכולות ה-RAG של AWS כדי לתקף תשובות מודל בקונטקסט של הלקוח [56:22].

דאטה מובנה: גרף הידע ו-AWS Context



גרף הידע הסמנטי שמאחורי Amazon Quick, כפי שהוצג על המסך

האתגר בדאטה מובנה הוא סמנטי: הסוכן צריך לדעת אילו טבלאות קיימות, מה משמעות העמודות, מה היחסים ביניהן ומה החוקיות העסקית. גרף הידע הסמנטי של Quick עונה על כך, והדוגמה שניתנה מוחשית: "אם משתמש שואל על escalations, אך בדשבורד הם נקראו tickets, הוא עדיין נמצא את התשובה הנכונה" [57:54]. הוצג גם היקף השימוש: "כבר היום, מאות אלפי משתמשים עושים שימוש יומי בגרף הידע של קוויק" [57:54].



ההכרזה על שכבת הקונטקסט המאוחדת ומאפייניה, במצב Coming Soon

AWS Context הוא ההכללה של אותו רעיון על כלל הדאטה הארגוני. הוא בונה אוטומטית גרף ידע מהדאטה הקיים, מסיק את מערכות היחסים בין נכסי הדאטה, את החוקיות העסקית ואת הבנת הדומיין, ולפי הסלייד נשען על מטא-דאטה נייטיבית של Iceberg ומשתלב עם Glue Data Catalog ועם SageMaker Unified Studio. הנקודה המעשית: "הגרף נבנה מתוך מקורות הדאטה והכלים שכבר קיימים בארגון" [59:27], בלי הקצאת תשתיות ובלי בניית פייפליין שליפה.

מודלים ותשתיות

הפרק נחתם בטענה כלכלית על אינפרנס: אינטראקציה אחת של סוכן אינה קריאה אחת אלא 10, 20 או 50 קריאות, וכשמכפילים זאת במיליארדי סוכנים — "אינפרנס עומד להפוך לורקלוד הגדול ביותר בעולם" [1:00:57], כלומר הוא כבר לא פרט תשתיתי אלא חלק מהאסטרטגיה העסקית.

נתון הצמיחה שניתן ל-Bedrock: "השנה, בריבון הראשון בלבד בדרוק הריץ יותר טוקנים מכל השנים הקודמות ביחד" [1:02:31] — כלומר ברבעון הראשון בלבד. בין המודלים החדשים שנמנו כזמינים בפלטפורמה הוזכרו Grok 4.6 ומודל חדש של OpenAI. הוזכרה גם שותפות בת למעלה מ-15 שנה עם NVIDIA, שהוצגה בסרטון קצר [59:27].

10. מה לוקחים מכאן



שקופית הסיום: אל תקבלו את מגבלות ההווה כגבולות המחר

- **המסגרת שווה יותר מהרשימה.** ארבע הקטגוריות — עבודה, בנייה, אבטחה, יצירה — הן דרך שימושית למפות גם מוצרים שלא הוזכרו בקינוט. שווה למפות מולן את הכלים שכבר בשימוש בארגון ולראות איפה יש חורים.
- **הצוואר עבר מהמודל לקונטקסט.** רוב ההכרזות המשמעותיות בקינוט הן שכבות קונטקסט (AWS Context, Managed Knowledge Base, Web Search) ולא מודלים. ארגון שמתכנן סוכנים ב-2026 צריך להשקיע בשכבה הזו לפני בבחירת מודל.
- **קצב הסוכנים שובר את מה שמסביבו.** אם הקוד נכתב פי עשרה מהר יותר אבל ה-**CI/CD**, החוב הטכני והבקלוג של האבטחה לא משתנים — לא מקבלים שיפור, מקבלים צוואר בקבוק שזו **Code ו-Continuous Modernization** Vulnerabilities הן התשובה של AWS לכך.
- **מודל אמון מדורג הוא הדפוס שחוזר.** **Code Vulnerabilities** מתחיל ב-**Human in the Loop** לפני אכיפה; **Quick** מבקש אישור לפני שליחת מייל; **Aidoc** מדגיש שהרופא הוא המחליט. זהו הדפוס שמאפשר להכניס סוכן לתהליך מפוקח.
- **המדידה היא ההכרזה הכי מעניינת, והיא לא של AWS.** **Accountable AI** של איכילוב — פרסום ביצועי מערכות ה-AI לצד ביצועי המטפלים האנושיים — הוא הרעיון היחיד בקינוט שמטפל בשאלת האמון באמצעות מדידה פומבית, ולא באמצעות ארכיטקטורה.

מפת הזמנים של הקינוט

זמן	נושא	דובר
2:00	פתיחה: ריג'ון תל אביב, 13 שנות סאמיט, 14 אלף משתתפים	AWS ישראל
5:09	ראיון במה: איכילוב, Gecko, Accountable AI	פרופ' אלי שפרכר
13:00	אג'נדה, שותפים, סיום היום	AWS ישראל
17:29	Compounding Momentum וארבע הקטגוריות	שחר דניאל
20:38	Amazon Quick והדמו של החזרה מחופשה	שחר דניאל וליאת צור
27:06	Aidoc: מ-foundation model רפואי לפרודקשן	אלד ולך
37:52	Kiro, DevOps Agent, AWS Transform	ליאת צור
45:35	AWS Continuum: ארבעה שירותי אבטחה	ליאת צור
50:08	AgentCore, קונטקסט, דאטה ומודלים	שחר דניאל

זמן	נושא	דובר
1:02:31	Bedrock, סיכום וסיום	שחר דניאל

מילון מונחים

מונח	מה זה	סטטוס כפי שהוצג
Amazon Quick	עוזר AI שמקבל תוצאה רצויה ומפענח לבד את הדרך אליה, מעל שכבת חיפוש אג'נטית לרוחב הדאטה הארגוני	זמין
Kiro	סביבת פיתוח אג'נטית שהופכת prompt לדרישות, דיזיין, משימות ווולידציות	זמין; אפליקציית iOS ב-Preview
AWS DevOps Agent	סוכן Ops לחקירת אירועים, אופטימיזציה ומשימות SRE	זמין; Release Management הוכרז
AWS Transform	סוכני מודרניזציה לקוד ולמערכות לגסי	זמין; Continuous Modernization ב-Preview
AWS Continuum	משפחת שירותי אבטחה אג'נטיים: Pen testing, Code review, Threat modeling, Code vulnerabilities	שניים GA, שניים ב-Preview
Bedrock AgentCore	תשתית מנוהלת לסוכנים בפרודקשן: runtime, זיכרון, אותנטיקציה לכלים	זמין
Managed Knowledge Base	בסיס ידע מנוהל ב-Bedrock עם קונקטורים, שליפה אג'נטית והרשאות ברמת מסמך	הוכרז
AWS Context	גרף ידע שנבנה אוטומטית מהדאטה הארגוני הקיים, מובנה ולא מובנה	Coming Soon
CARE	מודל היסוד הרפואי של Aidoc, שפותח בישראל ומכסה מחלות רבות בדיוק גבוה	בפרודקשן; מוצר נגזר ב-FDA Breakthrough
Gecko	מעטפת AI סביב התיק הרפואי שפותחה בתוך איכילוב	בשימוש קליני