# Claude/Mythos
# מסמך המלצות

## תוכן עניינים

-בלמ"ס רגיש-

**אל:**

**מנכ"לים, חברי דירקטוריון**

**מנהלי מערכות מידע (CIO) ומנהלי אבטחת מידע (CISO) .**

**הנדון: היערכות אסטרטגית לעידן "סערת החולשות" Vulnerability Storm ויכולות AI התקפיות**

**כללי**

מערך הסייבר הלאומי הוא גוף מבצעי טכנולוגי האמון על הגנת הסייבר הלאומית על אזרחי מדינת ישראל. הגנת סייבר מדינתית שונה מההגנה בעולם הפיזי ,וכמו בכל מדינות העולם, אין גוף אחד אשר עושה את משימת ההגנה אלא מדובר **בהגנה משותפת בה ישנה חובלה לאומית**, יחד עם חובתו של כל גוף, ארגון ואדם להגן על עצמו ברמה הנדרשת. זוהי גם האפשרות היחידה הנכונה במדינה כמו שלנו.

אני פונה אליכם משום שהאיום שעמו אנו מתמודדים במרחב הסייבר עובר תמורה דרמטית, והאופן שבו אנו מתגוננים חייב להשתנות בקצב תואם.

בעוד שבעבר מתקפות סייבר מורכבות היו נחלתם של מומחים יחידי סגולה, כיום אנו ניצבים בפני "עליית מדרגה" איכותית: דור חדש של מודלי AI, דוגמת GPT-5.4 Cyber ו-Claude Mythos שנחשפו באפריל 2026, פרץ את מחסום המורכבות הטכנולוגי ומסוגל לאתר במהירות חולשות רבות כולל חולשות Zero-day ולבצע שרשור תקיפות אוטונומי. יכולות אלו, המכפילות את עצמן מדי כמה חודשים, מעבירות את שדה הקרב מ"קצב אנושי" ל"קצב מכונה" והופכות את **שיטות ההגנה המסורתיות ללא מספיקות**. כעת, לא רק תשתיות קריטיות של המדינה אלא כל ארגון בכל סדר גודל מהווה מטרה פוטנציאלית, שכן 'מחיר הכניסה' לתקיפה איכותית יורד. בנוסף, התוקפים הממוכנים ינווטו את דרכם באופן אוטומטי לנקודה החלשה ביותר בשרשרת.

כניסתם של מודלים אלו מתווספת אל ארבע **נקודות התורפה המרכזיות שראינו במערכה הארוכה בה אנו נמצאים בהגנה בסייבר ומעצימה אותן:**

**שרשרת האספקה:** ניצול מהיר של חולשות ברכיבי צד-ג'.

**גניבת מידע מאנשי IT:** שימוש ב-AI ל- Phishing מתוחכם וממוקד.

**היעדר הגנה בסיסית:** המודל סורק ומנצל במהירות היעדר MFA או הגדרות שגויות.

**היעדר תוכנית התאוששות ורציפות תפקודית**

הצלחת ההגנה הלאומית בסייבר, בטח בהיעדר חוק סייבר לאומי, מחייבת את שיתוף הפעולה שלכם והבנתכם את עוצמתו של האיום המתפתח יחד עם טכנולוגיות מדהימות כמו שירותי ענן, בינה מלאכותית, מחשוב קוואנטי ועוד. **עלינו לאמץ טכנולוגיות אלו, יש בהן בשורות אדירות אך יש לעשות זאת בצורה מקצועית ובטוחה.**

האיומים בסייבר ימשיכו להתפתח בקצב ובעוצמה מתגברים, למולם עלינו לפעול יחד, במקצועיות, **לא מתוך בהלה אלא מודעות לאיום והיערכות להגן מפניו.**

אני, יחד עם כלל מערך הסייבר הלאומי ושותפינו בהגנה, עומדים לרשותכם בכל עת שתחפצו.

יוסי כראדי

ראש מערך הסייבר הלאומי

-בלמ"ס רגיש-

## מסמך ההמלצות – תכולה ומגבלות

### תכולת המסמך

המסמך נכתב בהמשך למסמך "Claude Mythos / מסמך עמדה" מתאריך 19.4.26, ומציג המלצות ראשוניות לארגונים וחברות בהמשך לפרסומים האחרונים על מודל Claude Mythos של חברת Anthropic.

המסמך מבוסס על פרסומים גלויים ועל ניתוחים פנימיים של חברת KPMG ומס"ל.

### מגבלות המסמך

המסמך מבוסס על מידע שפורסם בגלוי ועל ניתוח של מידע זה.

לצורך כתיבת מסמך העמדה, לא בוצעו ראיונות עם צדדים מעורבים (ובפרט לא עם חברת Anthropic) או בעלי עניין אחרים.

## 1. רקע

### Anthropic / Claude / Mythos 2.1

**Cluade** היא משפחת מודלי שפה גדולים (LLM) של חברת Anthropic. מודל השפה הראשון יצא במרץ 2023, ומאז יצאו מספר גרסאות של מודלי שפה, כאשר הגרסה הנוכחית היא Claude 4 (יש תת גרסאות שמתעדכנות בשוטף) ובה מספר מודלים (בסדר עולה של יכולת): Haiku (מודל מהיר וקטן לתשובות פשוטות), Sonnet (מודל בינוני מותאם למשימות כמו ניתוח מידע ויצירת קוד), ו-Opus (פתרון בעיות מורכבות וקידוד ברמה גבוהה).

ב-7.4.2026 פרסמה חברת Anthropic שמודל **Claude Mythos Preview** (גרסה מקדימה של מודל ה-Mythos) מפגין יכולות גבוהות בצורה יוצאת דופן באיתור חולשות ובפיתוח וקטורי תקיפה מורכבים, במהירות וברמת תחכום כזו שמעוררים חשש משמעותי מזליגת היכולות לידי תוקפים עוינים. בפרסום של החברה נטען כי במהלך מספר שבועות של הרצת המודל, אותרו **כמה אלפי** חולשות zero day, רבות מהן קריטיות, בכל מגוון מערכות ההפעלה והדפדפנים עליהם נבדק המודל. בסדרת השוואות שבוצעה בין Mythos לבין מודל Opus בגרסה 4.6, התברר כי Mythos עולה על המודל הקיים בשורת מדדים. נתון משמעותי נוסף הוא כי במשימות בתחום ה- cyber security (גילוי חולשות ופיתוח וקטורי תקיפה), מהירות העבודה של מיתוס עולה **בסדר גודל** על זו של מודלים קיימים.

יצויין כי לפרסום המודל ע"י Anthropic קדמו שני ארועי דליפה שכללו דליפת 3,000 קבצים פנימיים וכן דליפת קובץ שכלל כמעט 2,000 קבצי קוד מקור ויותר מ-512,000 שורות קוד.

חברת Anthropic הודיעה על יוזמה בשם **Project Glasswing** שבמסגרתה תינתן גישה מלאה ל-Mythos Preview ל-12 חברות ובהן AWS, Google, Microsoft, NVIDIA, Apple, וכן ל-למעלה מ-40 ארגונים העוסקים בתוכנה קריטית או בהגנה בסייבר. זאת על מנת לאפשר איתור ותיקון חולשות קריטיות לפני שיעשה בהן שימוש לצורך תקיפות. הציבור הרחב לא יקבל גישה למודל, אך

-בלמ"ס רגיש-

התוצרים של פרויקט Glasswing יפורסמו בגלוי כעבור 90 יום מרגע מציאתם כדי לאפשר לגופים ולחברות ליישם את התיקונים שיפותחו במסגרת הפרויקט.

**למרות שפריצת הדרך הנוכחית מתמקדת במודל של חברת Anthropic, ההערכה היא כי מדובר בעניין של חודשים עד להגעת מודלי שפה (LLM) של יצרנים נוספים לפריצות דרך דומות.**

### 2.1 עיקרי המשמעויות

קפיצת המדרגה הזו מייצרת מציאות חדשה בעולמות הגנת הסייבר, הן בארגון עצמו והן בקרב ספקים ושותפי שרשרת אספקה שלו, המהווים חלק ממשטח התקיפה הארגוני.

הסיכונים העיקריים – טווח קצר:

אתגר תפעולי המצריך היערכות משאבית.

שינוי במשוואת הסיכון.

צורך בקיצור זמני תגובה בקבלת החלטות ארגונית.

חשש מוגבר לסיכוני צד שלישי.

**הסיכונים העיקריים – טווח בינוני-ארוך:**

קיצור זמנים דרסטי מאיתור חולשה למימוש תקיפה אפקטיבית.

קפיצת מדרגה ביכולות התוקף.

השקת "מירוץ חימוש" מגן-תוקף מבוסס AI.

**המשמעויות העיקריות:**

מעבר לאתגר של מ"סגירת פרצות" נדרש גם לנהל מדיניות הגנה וחוסן: מ"סייבר כמגן" ל"סייבר כמכיל ומתאושש" (מ-prevention ל-resilience).

מעבר להגנה מבוססת/נתמכת AI.

צורך בשיתופי פעולה וקואליציות ליצירת הגנה אפקטיבית.

נדרש לצמצם דרמטית את פרק הזמן מרגע הפרסום ועד להתקנת עדכון האבטחה.

נדרש לבנות ארכיטקטורה של מערכות באופן שיאפשר עדכונים מהירים במידה ונדרש.

**השלכות לארגונים וחברות:**

הנחות העומדות בבסיס מדדי הסיכון הנוכחיים כגון פרק הזמן המאפשר לבצע עדכוני אבטחה מרגע פרסום החולשה, פרקי הזמן של ניצול חולשות ועוד עשויות להתגלות ככאלו המצריכות בחינה מחדש ומעבר להגנה מהירה יותר וזאת לאור הכניסה לעולם שבו הדגש עובר להכלה (containment) ולחוסן, ולכן המדדים צריכים כעת להתמקד במהירות החזרה לפעילות רגילה.

-בלמ"ס רגיש-

אותן יכולות בינה מלאכותית שיוצרות את הסיכון יוצרות גם הזדמנות הגנתית: ארגונים יכולים כיום לזהות את החולשות שלהם לפני שהתוקפים עושים זאת, לסקור קוד בקצב מכונה, ולהגיב לאירועים במהירות גבוהה יותר מזו שכל צוות אנושי מסוגל לה. ארגונים שישקיעו בכך יהיו מהירים יותר להתאושש וחסינים יותר בפני מתקפות.

## 2. המלצות טווח קצר

### 3.1 הצגה ודיון ברמת דירקטוריון

**לקבוע מועד מחייב** (מומלץ: עד 31.5.26) **להצגה לדרגי הנהלה ומשילות** (governance) **בכירים בארגון**, לרבות דירקטוריון החברה ו/או ועדת הביקורת וניהול הסיכונים, בנוגע להתפתחויות בנושא Mythos ו-Glasswing והשלכותיהן של התפתחויות אלו על סביבת האיומים הניצבת בפני הארגון. לאור **הצורך בהשקעת משאבים משמעותיים ביכולת העמידות וההתאוששות**, ולאור **ההשפעה הכלכלית הצפויה על הארגון** מתוך הבנה כי האיום החדש מצריך מחויבות כלל ארגונית, תידרש הצגת שינויים פוטנציאליים במדיניות ההגנה בסייבר (בהתאם למועד שחרור החולשות והתיקונים ע"י קואליציית Glasswing), לרבות בהיבטים הבאים:

רכש (קיצור זמני רכש והתקשרות).

מצבת כ"א.

בחינת הסכמים עם ספקים קריטיים.

עיבוי SOC.

עיבוי מבדקי חוסן.

**מומלץ לקבוע עדכון תקופתי** אחת ל-60-90 יום לכל היותר לעדכון בהתפתחויות וההיערכות של הארגון, ולייצר מחויבות הנהלה.

### 3.2 הגדרת זמן להצגת תכנית פעולה

לקבוע מועד מחייב להצגת תכנית פעולה במועד קרוב ככל הניתן לאחר פרסום מסיבי ראשון של חולשות ותיקוני אבטחה, שבו מנהל אבטחת המידע מציג ומאשר לביצוע בהנהלת הארגון (או במדרג החלטות ניהולי נמוך יותר, תלוי בדרסטיות התיקון) את הנגזרות כולל היערכות משאבית (העמדת צוותים), מיפוי מערכות קריטיות, וגיבוש תכנית הקשחות, ומקבל אישור ליישם את התכנית בשטח (לאחר שקיבל עוד תקציב, סמכות ומשאבים בהתאם למה שיידרש בפועל). במועד זה יוצגו להנהלה גם המשמעויות הכלכליות הנגזרות משינוי תכנית העבודה והסטת תקציבים.

### 3.3 האצה והרחבה של מדיניות עדכוני אבטחה

לקצר את זמני התגובה (SLA) הרלוונטיים להתקנת עדכוני אבטחה, ולבחון מחדש תהליכים ארגוניים בהתאם, כדי לאפשר תיקון מהיר ובהיקף נרחב. נדרש לגבש מראש תיעדוף על פי קריטיות

-בלמ"ס רגיש-

ורגישות המערכות, וכן לתת התייחסות נפרדת למערכות Legacy. יש לתת מענה הגנתי עבור מערכות שלא ניתן לבצע בגינן עדכונים ולגדר את הסיכון ע"י גידור החשיפה של המערכות לאינטרנט , יישום כלי Zero Trust , MFA , סגמנטציה ועוד בנוסף יש להכין תוכנית מסודרת להגבהת חומות באמצעות דוחות של כלי CNAPP, כלים מובנים של הספקיות או כלים אחרים המותקנים בסביבות ה- On Prem .

### 3.4 הידוק מיפוי ובקרה בשרשרת האספקה

נדרש לבצע רישום מדויק ועדכני של כלל רכיבי התוכנה המהווים חלק מהאקו-סיסטם הארגוני, לרבות רכיבי צד שלישי, וזאת על מנת לאפשר הערכה ותיעדוף של הסיכון לארגון כתוצאה מחולשות Zero-Day. נדרש לחייב הקמת מאגר ברור ומסודר של ספקים, לסמן מתוכם ספקים קריטיים, ולחייב את הספקים להוכיח שהם מטפלים, או שיש להם תוכנית סדורה לטיפול, בשינוי הצפוי בהיקף ובתדירות פרסום החולשות והעדכונים, לרבות עדכון ב-SLA מיידי על כל חשד לאירוע סייבר משמעותי. לצורך כך, נדרש להכיר את מדיניות אבטחת המידע והסייבר של הספקים מולם מייצרים התקשרות.

בנוסף, מומלץ להתניע בחינה לעדכון מדיניות הרכש במענה לאיומים החדשים.

### 3.5 מיפוי מערכות קריטיות, בדיקת חוסן והמשכיות

למפות מערכות קריטיות לארגון (IT ו-OT, on-prem ובענן) ולבצע בדיקת עמידות לתוכניות החוסן (resiliency) שלהן אל מול איומים מבוססי AI, לרבות תרגילים ומבדקי חוסן על מנת להעריך את רמת המוכנות לאירועי סייבר ולהפרעות למערכות ולנתונים קריטיים. בהמשך לכך, לדרוש יישום שיפורי חוסן תפעוליים היכן שנדרש ועל פי ממצאי מבדקי החוסן.

### 3.6 תיקוף משוואות הסיכון ו"תיאבון" הסיכון הארגוני

לבחון ולעדכן את סט המדדים, מנגנוני הדיווח ומשוואות הסיכון כך שישקפו לוחות זמנים מעודכנים (מואצים) לניצול חולשות, ומורכבות גבוהה יותר של מתקפות. הנחות עבר הנוגעות לחלונות זמנים מקובלים להטמעת עדכוני אבטחה, לניצול אפקטיבי של חולשות ולתדירות אירועי אבטחה – עלולות לאבד רלוונטיות. מומלץ לחייב את הארגון לתקף את משוואות הסיכון ולבחון מחדש את "תיאבון הסיכון" הארגוני, **ובכלל זאת לבחון גישת breach ready – להניח שהמערכות הארגוניות יפרצו, ולהשקיע משאבים משמעותיים ביכולת העמידות וההתאוששות**.

### 3.7 ישום תכנית בדיקות משולבת כלי AI

להמליץ לארגון להציג תוכנית קונקרטית להטמעת בדיקות אבטחה מתקדמות מבוססות כלי AI. התוכנית צריכה להקיף נכסי תוכנה קיימים ופיתוח קוד (ככל שמתבצע בארגון), ולכלול סימולציות תקיפה צד אדום (adversarial) כנגד כלל סביבת המערכות, במטרה לאתר ולצמצם חולשות.

-בלמ"ס רגיש-

### 3.8 הגברת הפיקוח על גישת צד שלישי למערכות

להעריך את היקף הגישה של צדדים שלישיים למערכות ולסביבה הארגונית, ולהגביל גישה למערכות קריטיות היכן שניתן. בנוסף, להגביר את הניטור והבקרה על חיבורים וקשרים עם צדדים שלישיים.

## 3. המלצות טווח בינוני

### 4.1 הרחבת יישום הגנה רב שכבתית וארכיטקטורת ZERO TRUST

**להניח כי אירוע חדירה לרשת הפנימית הוא עניין של זמן**, ולהטמיע אסטרטגיות Zero Trust, לרבות אבטחת שכבת הזהויות ויישום הפרדה וסגרגציה של נכסים קריטיים ככל הניתן, במטרה **לצמצם את רדיוס הפגיעה (blast radius)** ולהבטיח שחולשה יחידה לא תוביל לקריסת מנגנוני ההגנה, ולחדירה לרשתות פנימיות ולמאגרי נתונים רגישים.

### 4.2 שדרוג מערכי הגילוי כנגד תרגילי תקיפה משולבי AI

לבצע מבדקי עמידות ליכולות הזיהוי של ה-SOC אל מול תרחישים של תנועה רוחבית (Lateral Movement) בתוך הרשת ב"מהירות מכונה" (להבדיל ממהירות אנושית). להשוות את זמן הזיהוי הממוצע (MTTD) אל מול מודל תוקף מבוסס AI. להרחיב את הניטור לאזורים שאינם מנוטרים כיום (לדוגמא OT). בנוסף, נדרש לשדרג את מערך המלכודות.

### 4.3 תיקוף מודלי ביטוח סיכוני סייבר

לקראת שנת הכספים 2027 ובשלב מוקדם ככל הניתן, נכון לדרוש את תיקוף מדיניות ביטוח לסיכוני סייבר מול סוכני הביטוח. ספציפית נדרש לבחון השפעות של סביבת המחשוב הארגונית (כגון מידת החשיפה למערכות legacy, רמת הבשלות של ה-SOC), על פרמיית הביטוח והכיסוי הביטוחי.

## 4. המלצות טווח ארוך

### 5.1 תכנית שדרוג מערכות legacy

להקצות מימון, לקבוע תיעדוף ולתכנן את תהליך השדרוג ממערכות ופלטפורמות שהגיעו לסוף חיי המוצר (End-of-Life). בכלל זאת נדרש להעניק עדיפות למערכות עם ממשק חשוף לאינטרנט ולמערכות המטפלות במידע מהותי/רגיש.

### 5.2 שילוב כלי AI במערך ההגנה

ישום תכנית פיתוח ובדיקות משולבת סוכני AI. מומלץ לארגון להטמיע כלי הגנה מבוססי AI (כגון סריקת חולשות, ניהול תצורה, אוטומציית תגובה לאירועים – incidence response, אוטומציית SOC וכדומה), בכדי לייצר תפיסת הגנה **רציפה והמשכית** במקום תפיסה **סטטית/עיתית**. בפרט לא נכון להמתין לקבלת אינדיקציה להשוואת יכולות התוקפים, כתנאי מקדים להתנעת התכנית.

### 5.3 אינטראקציה עם בעלי עניין ועם הרגולטור כדי לעצב את הדירקטיבה

ליזום מהלכים פרואקטיביים להגברת המעורבות מול גורמי רגולציה רלוונטיים לרבות מערך הסייבר הלאומי וגורמי רגולציה נוספים, במטרה לעצב ולהשפיע על ההנחיות הצפויות בתחום ה-AI והסייבר, במקום להגיב אליהן בדיעבד.

### 5.4 חיסון מערכות ה-AI הארגוניות ע"י ישום כלי בקרה

להטמיע מנגנוני הגנה נוספים ל-AI, לרבות שקיפות מלאה (observability) למערכות ה-AI, שילוב כלי ניתוח התנהגותי מתקדמים, אכיפת מדיניות אוטומטית על פעולות ו-prompts של AI, ויישום מנגנונים לזיהוי תכנים מג'ונרטים, וכן הצבת מגבלות שונות על מנת לבקר ולהגביל פעולות לא מורשות של סוכני AI (הפרדה בין סביבות, הגבלת טוקנים, אישור אנושי לפני ביצוע פעולה הרסנית).

## 5. 6 סיכום

השינוי במשוואת האיום, וכתוצאה מכך גם בחלק גדול מהנחות היסוד של ההגנה בסייבר בשנים האחרונות, מוביל לשינויים במשוואת הסיכון הארגונית. במובנים מסוימים חלון התגובה (response window) הולך ונסגר ואנחנו בתהליך מעבר ממניעה (prevention) לחוסן והמשכיות עסקית (business resilience).

-בלמ"ס רגיש-

על מנת לשפר את המוכנות הארגונית לקראת השינויים הצפויים, גובש וזוקק סט המלצות בחלוקה לטווחי זמן (קצר, בינוני, ארוך), לרבות דרישה לשקף לדרגי הנהלה בכירים את מהות השינוי, להציג תוכניות פעולה, להדק את השליטה במערכות ובתהליכים בארגון ובשרשרת האספקה, ולנקוט בשורת צעדים שמטרתם לכמת ולגדר את הסיכונים, ולהכין את הארגון לקראת העידן החדש, שבו המענה לשיפור הצפוי ביכולות התוקף, מתבסס בין היתר על שילוב כלי AI גם בצד המגן, בטווח הארוך מודולי שפה רבי עוצמה יועילו למגינים יותר מאשר לתוקפים ומגינים שינתבו משאבים בצורה יעילה יותר וישתמשו במודלים וכלי AI כדי לתקן באגים לפני שקוד חדש יישלח יוכלו להפיק את המירב ולהגן טוב יותר על הארגון.

לאור הדינמיות של האירועים, נכון יהיה לגזור מהמלצות אלו הנחיות בחתך המגזרים השונים, ולתקף אותן באופן עיתי.

בברכה,

אלי מזרחי ראש חטיבת מצודה

מערך הסייבר הלאומי

## 6. רשימת מקורות

AI Security Institute - Our evaluation of Claude Mythos Preview's cyber capabilities, 13 April 2026.

Cloud Security Alliance - The "AI Vulnerability Storm": Building a "Mythos-ready" Security Program, 12 April 2026.

Red.Anthropic.Com - Assessing Claude Mythos Preview's cybersecurity capabilities, 7 April 2026.