Mythos – Är ryktena om cybersäkerhetens död överdrivna?

Anthropic beskriver sin nya språkmodell Mythos som så kapabel att hitta sårbarheter att den inte bör släppas fritt till allmänheten. Detta memo granskar publicerad information om Mythos i syfte att skilja fakta från marknadsföring. Slutsatsen är att Mythos i sig inte är avgörande för cyberförsvaret, men att språkmodellers förmågor att hitta sårbarheter utvecklas snabbt just nu och lär utgöra ett allt kraftfullare verktyg med tiden.

Introduktion

Sedan Anthropic tillkännagav sin nya språkmodell, Mythos, har rubrikerna om förödande konsekvenser för cybersäkerheten varit många. Användningen av språkmodeller, en variant av generativ AI tränad på stora mängder text för att tolka och generera mänskligt språk, är inget nytt inom cybersäkerhet. Dock hävdar Anthropic att Mythos cybersäkerhetsförmågor är så pass mycket bättre än tidigare modeller att den utgör en fara för global cybersäkerhet. Anthropic valde därför att inte släppa Mythos publikt, då företaget beskrev modellen som för farlig för att göras tillgänglig innan tillräckliga säkerhetsmekanismer införts.

I detta memo sammanfattas och analyseras information som publicerats om Mythos cybersäkerhetsförmågor. Detta gäller kodanalys för sårbarhetsupptäckt samt utveckling av attacktekniker och tester av attackteknikerna. Memot ger också exempel på utmaningar med att använda AI-stödd sårbarhetsupptäckt. Slutligen diskuteras hur Mythos utveckling kan komma att påverka cybersäkerhetsområdet och vad cyberförsvaret bör göra till följd av denna utveckling.

Det bör noteras att memots underlag till stor del utgörs av ogranskade källor, såsom förhandstryck, blogginlägg och företagsinterna publikationer, vilket innebär att resultaten bör tolkas med viss försiktighet. Dessutom har Mythos utvecklats i flera steg, så uppgifterna kan gälla olika versioner. Förutom detta memo finns det också andra som analyserat Mythos. Ett exempel är AI Sweden som i en kort rapport lyfter fram AI:s påverkan på sårbarhetsupptäckt och attacker samt rekommenderar försvarare att använda AI för att upptäcka sårbarheter [1].

Mythos förmågor enligt Anthropic

Anthropic har i flera publikationer beskrivit Mythos cybersäkerhetsförmågor. De har bland annat nämnt dessa förmågor i samband med sitt initiativ *Project Glasswing* [2], ett partnerprogram som begränsar åtkomst till Mythos till utvalda organisationer och företag. Enligt Anthropic är syftet med initiativet att låta försvarare härda kritiska system innan dessa förmågor blir publikt tillgängliga.

Enligt Anthropic har Mythos under dess interna utvärdering upptäckt tusentals sårbarheter med hög allvarlighetsgrad inom både öppen och stängd källkod, vilka innefattar sårbarheter i alla stora webbläsare och operativsystem [3]. Detta ska ha gjorts autonomt med minimal mänsklig styrning [4]. Med minimal mänsklig styrning menar Anthropic att ingen människa var involverad i upptäckten eller exploateringen av sårbarheterna efter den initiala prompten [3]. Det är dock oklart i vilken utsträckning mänskliga beslut spelat roll i promptutformning, design av sele (eng. *harness*), samt val av verktyg som Mythos kunde använda. En sele syftar på det system som omger en modell. Det kopplar samman modellen med andra system, så att modellen bland annat kan använda verktyg, komma ihåg tidigare kontext och utföra komplexa uppgifter i flera steg, i stället för att bara svara på frågor.

Anthropic har skapat en översikt över de sårbarheter Mythos har hittat i öppen källkod. I översikten medföljer en lista med kryptografiska åtaganden för 1611 sårbarhetsrapporter. Om Anthropic senare publicerar faktiska sårbarhetsrapporter kan de med åtagandena bevisa att rapporterna skrevs senast när åtagandena publicerades. Än så länge ges inte detaljer om exempelvis typ eller riskbedömning för majoriteten av sårbarheterna [5]. Detta medför att det är otydligt hur stor del av sårbarheterna som är exploaterbara. Det är också svårt att oberoende bedöma om sårbarheterna ens finns. Anthropic har bedömt att mänsklig kontroll av Mythos resultat behövs men tar lång tid. Sammanfattningsvis är det i nuläget svårt att bedöma innebörden av de tusentals sårbarheter som Mythos har hittat, eller ens vilka av sårbarheterna som är riktiga.

Praktisk användning av Mythos

Utöver Anthropics egen testning har Mozilla, som samarbetar med Anthropic, rapporterat att de hittat 271 sårbarheter i Firefox med hjälp av Mythos [6]. Mozilla hade redan tidigare experimenterat med Opus 4.6, då med specialdesignade selar, och fortsatte med samma metod när

FOI Memo: 9314
Forskningsområde: Cyberförsvar och cybersäkerhet
Godkänd av: Pauline Ärlebäck

de fick tillgång till Mythos. Det framgår dock inte hur många sårbarheter Opus 4.6 hittade eller hur den presterade jämfört med Mythos. Memoförfattarna bedömer att Mythos sannolikt presterar bättre än Opus 4.6, men hur mycket bättre går inte att avgöra utifrån Mozillas utvärdering. Det är också oklart om det främst var modellen eller selen som möjliggjorde resultatet.

Ytterligare ett exempel där Mythos använts är för sårbarhetsupptäckt på Curls kodbas [7]. Sårbarhetsupptäckten utfördes av en person inom Linux Foundation som hade tillgång till Mythos genom Project Glasswing. Huvudutvecklaren bakom Curl, Daniel Stenberg, har valt att bevara denna persons anonymitet. Resultaten presenterade totalt en sårbarhet med låg allvarlighetsgrad. Curl hade sedan tidigare skannats med flera AI-baserade verktyg såsom *AISLE*, *Zeropath* och *Codex Security*, vilka tillsammans upptäckt mellan 200 och 300 brister i Curl under det senaste året. Stenbergs personliga slutsats är att den stora uppmärksamheten kring modellen hittills i första hand drivits av marknadsföring och att han inte ser några bevis för att Mythos hittar sårbarheter i väsentligt högre utsträckning än tidigare verktyg.

Prestandajämförelser mot andra modeller

Ett flertal prestandajämförelser för att mäta Mythos förmågor inom cybersäkerhet har utförts av externa organisationer, forskare och Anthropic själva. I dessa jämförs Mythos förmågor mot andra modeller, bland annat Anthropics tidigare modell Opus 4.6 och Open AI:s GPT 5.5 (som släpptes ungefär samtidigt som Mythos tillkännagavs). Eftersom Mythos inte är en publikt tillgänglig modell har alla dessa mätningar utförts i samarbete med Anthropic, vilket utgör en intressekonflikt.

De huvudsakliga prestandajämförelserna i samarbete med Anthropic är:

- Storbritanniens statliga forskningsorganisation för AI-säkerhet (eng. *AI Security Institute*, AISI): Testade Mythos förmåga att lösa en samling hackningsutmaningar (CTF) och utföra två simulerade nätverksattacker [8].
- Det amerikanska forskningsinstitutet METR: Utförde en bredare förmågeevaluering på en tidig version av Mythos [9].
- Extern forskning: Två artiklar som jämför Mythos mot andra modellers förmågor att skapa attackkedjor: Exploitbench [10] [11] och Exploitgym [12].

Ytterligare utförde Anthropic intern testning med en äldre akademisk prestandajämförelse, Cybergym [13] [14]. Huvudresultat samt några svagheter med metoderna för dessa prestandajämförelser sammanfattas i tabell 1.

Tabell 1. Prestandajämförelser för Mythos. Kolumnen ”huvudresultat” är memoförfattarnas sammanfattning av resultaten. Exempelvis betyder 156 % bättre att förmågan är 100 % + 156 % = 256 %. Kolumnen ”svagheter” är memoförfattarnas åsikter avseende svagheter i metodens utformning eller utförande.

Jämförelse	Huvudresultat	Svagheter
AISI	1. Mythos och GPT 5.5 var ~16 % respektive ~21 % bättre på att lösa CTF-utmaningar på expertnivå än Opus 4.6. 2. Mythos och GPT 5.5 lyckades klara den första av två nätverksattacker till skillnad från Opus 4.6 som inte klarade någon.	AISI:s utvärdering uteslöt informationen från den nätverksattack som Mythos och GPT 5.5 inte klarade från det publicerade resultatet. Sådan information publicerades däremot i tidigare prestandajämförelser [15] med samma metod, men med andra modeller.
METR	Mythos klarade i majoriteten av testerna de uppgifter som tog människor omkring 16 timmar att lösa. Opus 4.6 klarade bara enklare uppgifter som människor klarade redan på 12 timmar.	METR [16] utvärderar flera egenskaper utöver cybersäkerhet. Det är oklart exakt vilka cybersäkerhetstester som användes, och hur stor del av det redovisade resultatet de utgör.
Exploitbench	Mythos och GPT 5.5 presterade ~156 % respektive ~52 % bättre än Opus 4.7 (en ny version av Opus släppt efter Mythos) på att återanvända gamla fixade sårbarheter för att skapa attackkod mot en komponent i en webbläsare.	Metoden i Exploitbench ger Mythos en budget i kronor som per test är upp till tio gånger så stor som andra modellers.
Exploitgym	Mythos och GPT 5.5 lyckades exploatera ~191 % respektive ~122 % fler system med hjälp av angivna kända sårbarheter jämfört med den bästa tidigare modellen *GPT 5.4*.	Andelen lyckade exploateringar som byggde på en annan sårbarhet än den avsedda var högre för både Mythos och GPT 5.5 än för GPT 5.4.
Cybergym	Mythos och GPT 5.5 presterade ~5 % respektive ~4 % bättre än GPT 5.4 på att återskapa en redan känd sårbarhet utifrån en textbeskrivning av denna.	Cybergyms sårbarhetsbeskrivningar bygger på utvecklarnas egna offentliga beskrivningar av bristerna, omskrivna av GPT. Modeller som tränats på samma kodbaser kan därmed ha sett den ursprungliga beskrivningen.

Detaljerna kring samarbetena mellan Anthropic och de medverkande institutionerna och forskarna är i många fall inte offentliga, men har bland annat inneburit att Anthropic utfört tester åt forskare [12] eller gett dem kostnadsfri tillgång till modellen [10].

Utöver prestandajämförelsernas individuella svagheter och intressekonflikter finns i nuläget inte ett vedertaget sätt att tillförlitligt bedöma modellers förmågor inom cybersäkerhet. Ett stort problem är att motverka memorering, det vill säga att se till att exakta eller mycket liknande uppgifter och deras lösningar inte finns i modellernas träningsdata. Om modeller tränas på publikt tillgängliga CTF-utmaningar eller tidigare kända sårbarheter är det svårt att skilja modellernas förmåga att resonera från deras förmåga att återge inträngade lösningar. Det är alltså oklart hur bra modellerna egentligen är.

Trots dessa problem med prestandajämförelserna går det utifrån samtliga resultat i tabell 1 att konstatera att Mythos presterar bättre än de tidigare modellerna GPT 5.4, Opus 4.6 och Opus 4.7. Resultaten från AISI, Exploitgym och Cybergym pekar dessutom till stor del på att GPT 5.5 är ungefär likvärdig med Mythos; GPT 5.5 presterar bättre än Mythos på AISI och något sämre på Exploitgym och Cybergym.

Sårbarhetsupptäckt med AI innebär utmaningar

AI-baserad upptäckt av sårbarheter inom öppen källkod har utvecklats snabbt under de senaste åren, vilket lyfts i uttalanden från flera personer inom etablerade mjukvaruutvecklingsprojekt. Några av dessa uttalanden beskrivs nedan. En gemensam faktor är utmaningen med att värdera huruvida AI-modellernas resultat visar på riktiga och nya sårbarheter.

I ett inlägg angående en uppdatering för Linux [17] uttalade sig grundaren Linus Torvalds om nuvarande utmaningar kopplade till språkmodeller för att upptäcka sårbar kod. Han nämner hur en uppsjö av AI-baserade säkerhetsrapporter har gjort den e-postlista dit sårbarheter rapporteras nästan omöjlig att hantera. Ett stort antal personer analyserar för närvarande Linux kodbas kontinuerligt med samma verktyg och upptäcker därmed samma sårbarheter, vilket skapar en stor mängd dubblettrapporter. En annan utvecklare av Linux-kärnan, Greg Kroah-Hartman, beskrev i en intervju [18] hur de brukade få AI-baserade säkerhetsrapporter som var uppenbart felaktiga, men att rapporterna sedan början av 2026 är välutformade och ofta pekar ut faktiska säkerhetsbrister. Något liknande uttrycks i blogginlägget ”High-Quality Chaos” [19] av huvudutvecklaren av Curl, Daniel Stenberg. Han menar att AI-baserad sårbarhetsupptäckt under början av 2026 gått från felaktigt AI-slask till välformulerade och till stor del korrekta rapporter. Detta har resulterat i en enorm ökning av legitima rapporter där den stora mängden sårbarheter varit svår för Curl att åtgärda.

Memoförfattarna har inte hittat någon information eller några uttalanden om vad som gjorde att cybersäkerhetsrapporter genererade av AI-baserade verktyg verkar ha gått från att vara dåliga till bra. Det är oklart om det som ligger bakom förändringen är en specifik modell, sele, prompteknik eller kombinationen av utvecklingen inom samtliga områden. AI-verktyg för sårbarhetsupptäckt verkar dock växa fram snabbt, både i form av autonoma system och som assistans vid mänsklig sårbarhetsupptäckt. Ett exempel på assistans vid mänsklig analys är *CopyFail* [20], en nyligen upptäckt sårbarhet i Linux-kärnan som legat dold sedan 2017.

Slutsatser

Kombinationen av stora språkmodeller och traditionell testning utgör redan idag ett effektivt och välutbrett tillvägagångssätt för att upptäcka sårbarheter. Modellerna är också lovande för att gå från upptäckta sårbarheter till attackkod. Dessutom utvecklas AI-modellernas förmågor i snabb takt. En av de ledande modellerna, Mythos, verkar utgöra ett steg framåt jämfört med Opus 4.6 som släpptes 61 dagar tidigare. Mythos är inte ensamt om att utvecklas snabbt och GPT 5.5 presterar i vissa fall jämförbart med Mythos. Det är dock svårt att mäta modellers förmågor på ett tillförlitligt sätt.

Memoförfattarna anser att cyberförsvaret bör bygga upp egna förmågor att utföra AI-baserad sårbarhetsupptäckt inför den sannolika ökningen av dessa verktyg under de kommande åren. Vidare bör förståelsen för och förmågan att konstruera selar undersökas, eftersom dessa kan spela en betydande roll i utvecklingen och tillämpningen av AI-baserad sårbarhetsupptäckt.

För att kunna använda dessa verktyg anser memoförfattarna att en av de största utmaningarna för cyberförsvaret i nuläget är tillgången till att använda modeller. Modeller såsom Mythos, Opus 4.6 och GPT 5.5 är driftsatta på Anthropics eller Open AI:s infrastruktur. Att avstå från externt driftsatta modeller skulle innebära att endast lokala modeller kan användas, vilka i nuläget verkar ha lägre förmåga.

Sammantaget framstår ryktena om cybersäkerhetens död som överdrivna. Utvecklingen innebär dock sannolikt att förmågan att upptäcka sårbarheter ökar kraftigt under de kommande åren, vilket ställer nya krav på såväl försvarare som angripare. Huruvida dessa verktyg i slutändan gynnar försvarare eller angripare mest är en öppen fråga.

Författare: Viktor Bergström och John Ziegenbein

Projektet *Omvärldsanalys av koncept och teknik för cyberförsvar* kartlägger civil kunskap som kan ha relevans för cyberförsvaret.

Referenser

[1] R. Bridges, D. B. Johnson och ej namngivna författare, ”What does Mythos Preview & Project Glasswing Mean for Sweden?”, 24 april 2026. [Online]. Tillgänglig: https://www.ai.se/sites/default/files/2026-04/mythos_preview_whitepaper-april302026.pdf. [Använd 15 juni 2026].

[2] Anthropic, ”Project Glasswing”, 7 april 2026. [Online]. Tillgänglig: https://www.anthropic.com/glasswing. [Använd 1 juni 2026].

[3] Anthropic Frontier Red Team, ”Assessing Claude Mythos Preview's Cybersecurity Capabilities”, 7 april 2026. [Online]. Tillgänglig: https://red.anthropic.com/2026/mythos-preview/. [Använd 1 juni 2026].

[4] Anthropic, ”System Card: Claude Mythos Preview”, 7 april 2026. [Online]. Tillgänglig: https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf. [Använd 1 juni 2026].

[5] Anthropic Frontier Red Team, ”Anthropic's Coordinated Vulnerability Disclosure Dashboard”, 22 maj 2026. [Online]. Tillgänglig: https://red.anthropic.com/2026/cvd/. [Använd 1 juni 2026].

[6] B. Holley, ”The Zero-Days Are Numbered”, 21 april 2026. [Online]. Tillgänglig: https://blog.mozilla.org/en/privacy-security/ai-security-zero-day-vulnerabilities/. [Använd 2 juni 2026].

[7] D. Stenberg, ”Mythos Finds a curl Vulnerability”, 11 maj 2026. [Online]. Tillgänglig: https://daniel.haxx.se/blog/2026/05/11/mythos-finds-a-curl-vulnerability/. [Använd 2 juni 2026].

[8] AISI, ”Our Evaluation of Claude Mythos Preview's Cyber Capabilities”, 13 april 2026. [Online]. Tillgänglig: https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities. [Använd 1 juni 2026].

[9] METR, ”Task-Completion Time Horizons of Frontier AI Models”, 8 maj 2026. [Online]. Tillgänglig: https://metr.org/time-horizons/. [Använd 1 juni 2026].

[10] S. Lee och D. Brumley, ”ExploitBench: A Capability Ladder Benchmark for LLM Cybersecurity Agents”, doi:10.48550/arXiv.2605.14153, 2026.

[11] S. Lee och D. Brumley, ”Exploitbench full leaderboard - Score and spend per model”, 2026. [Online]. Tillgänglig: https://exploitbench.ai/#leaderboard. [Använd 16 juni 2026].

[12] Z. Wang m.fl., ”ExploitGym: Can AI Agents Turn Security Vulnerabilities into Real Attacks?”, doi:10.48550/arXiv.2605.11086, 2026.

[13] Z. Wang m.fl., ”CyberGym: Evaluating AI agents' real-world cybersecurity capabilities at scale”, doi:10.48550/arXiv.2506.02548, 2025.

[14] Z. Wang m.fl., ”Cybergym - Leaderboard”, 2026. [Online]. Tillgänglig: https://www.cybergym.io/cybergym/. [Använd 16 juni 2026].

[15] L. Folkerts m.fl., ”Measuring AI agents' progress on multi-step cyber attack scenarios”, doi:10.48550/arXiv.2603.11214, 2026.

[16] T. Kwa m.fl., ”Measuring AI ability to complete long software tasks”, doi:10.48550/arXiv.2503.14499, 2025.

[17] L. Torvalds, ”Linux 7.1-rc4”, 17 maj 2026. [Online]. Tillgänglig: https://lwn.net/Articles/1073192/. [Använd 1 juni 2026].

[18] S. J. Vaughan-Nichols, ”AI Bug Reports Went from Junk to Legit Overnight, Says Linux Kernel Czar”, 26 mars 2026. [Online]. Tillgänglig: https://www.theregister.com/software/2026/03/26/linux-kernel-czar-says-ai-bug-reports-arent-slop-anymore/5226256. [Använd 1 juni 2026].

FOI Memo: 9314
Forskningsområde: Cyberförsvar och cybersäkerhet
Godkänd av: Pauline Ärlebäck

[19] D. Stenberg, ”High-Quality Chaos”, 22 april 2026. [Online]. Tillgänglig: https://daniel.haxx.se/blog/2026/04/22/high-quality-chaos/. [Använd 1 juni 2026].

[20] J. Im, ”Copy Fail: 732 Bytes to Root on Every Major Linux Distribution”, 29 april 2026. [Online]. Tillgänglig: https://xint.io/blog/copy-fail-linux-distributions. [Använd 2 juni 2026].