NATIONELLT CYBERSÄKERHETSCENTER
En del av FRA

Fördjupade rekommendationer

Juni 2026

Skydd mot AI-förstärkta angrepp

Tekniska och operativa åtgärder för skydd mot avancerade AI-förstärkta cyberhot.

De rekommenderade åtgärderna i detta dokument utgår från grundläggande cyberhygien som bör finnas på plats i varje organisation. De är inte AI-specifika, men avancerad AI gör dem mer brådskande, kända angrepp genomförs nu snabbare, billigare och med högre precision. Dokumentet riktar sig till it- och säkerhetsfunktioner och kompletterar dokumentet *"För beslutsfattare och organisationer"*.

Utgångspunkten är tudelad. Befintliga system ska skyddas mot AI-förstärkta angrepp, och de egna AI-systemen ska hanteras som en ny attackyta. Det senare omfattar att skydda hela AI-livscykeln, att aktivt begränsa agentiska system, att kontinuerligt utbilda medarbetare mot AI-förstärkt manipulation samt att hantera riskerna i leverantörskedjan och i open source-komponenter. För båda gäller samma princip: den som har ordning på grunderna står betydligt starkare än den som förlitar sig på punktinsatser.

Grunderna förblir giltiga och bör prioriteras för de risker som inte kan hanteras på annat sätt. Segmentering av nätverk, patchning av kända sårbarheter, identitets- och åtkomsthantering samt försvar på djupet och bredden ökar svårigheten för angriparen och sänker risken på sikt. Den nuvarande vågen av AI är sannolikt bara den första av flera – förmågan som byggs nu bör därför vara uthållig inför kommande hot.

1. Grundläggande cyberhygien

Etablera grundläggande cyberhygien genom att följa NCSC:s 10[1] rekommenderade säkerhetsåtgärder.

Att skyndsamt installera säkerhetsuppdateringar, minimera den internetexponerade attackytan, kräva stark autentisering och styra behörigheter, härda och segmentera, säkerhetskopiera med testad återställning samt övervaka och larma på avvikelser – allt detta finns väl beskrivet i NCSC:s 10 rekommenderade säkerhetsåtgärder, som bör vara på plats innan de AI-specifika åtgärderna nedan införs. Ju mer avancerade och AI-drivna hoten blir, desto viktigare blir grunderna, många avancerade angrepp bygger vidare på kända metoder.

2. Uppdatera mjukvara

Prioritera särskilt skyndsam patchning och kontroll av internetexponerade tjänster och hårdvara. Tiden från att en sårbarhet blir känd till att den börjar utnyttjas har krympt stadigt, och AI förstärker den utvecklingen. Angripare kan använda AI för att snabbare läsa av sårbarhetsinformation, identifiera påverkade system och utveckla fungerande exploits, i vissa fall inom timmar efter att en brist offentliggjorts. Det fönster försvararen har för att hinna åtgärda blir därmed allt smalare.

Två åtgärder ger störst effekt mot den här utvecklingen. Den första är att installera säkerhetsuppdateringar skyndsamt, med särskild prioritet på internetexponerade och affärskritiska system, det är där en sårbarhet nyttjas snabbast och får störst konsekvenser. Den andra är att aktivt minska den exponerade attackytan, ju färre tjänster och system som är exponerade mot internet, desto färre vägar in finns att utnyttja, och desto mindre brådskande blir varje enskild patch.

Tillsammans bryter de angriparens tempofördel. Att veta exakt vad som är exponerat, hålla det uppdaterat och stänga det som inte behöver vara nåbart utifrån är därför inte rutinunderhåll, utan en av de mest verkningsfulla åtgärderna för att möta AI-förstärkta hot.

AI-drivna cyberhot är ett område med snabb utveckling. Dessa rekommendationer kan förändras över tid.

3. Skydda hela AI-livscykeln

Behandla AI-system som vilken annan verksamhetskritisk it som helst, med säkerhet inbyggd i varje skede, från design och utveckling till drift och avveckling. Det innebär säker design från start, kontroll på och spårbarhet för promptar, modeller och träningsdata, tydlig styrning av vem och vad som får åtkomst, samt en kontrollerad avveckling där modeller, data och nycklar tas ur bruk på ett säkert sätt.

I grunden är detta samma principer som gäller för övrig it, men AI tillför nya angreppsytor. Prompt injection kan få en modell att kringgå sina instruktioner, data poisoning kan förgifta det den lär sig, och modellstöld kan läcka både immateriella värden och känsliga träningsdata. Till skillnad från traditionell it saknas här fortfarande väl etablerade skyddsstandarder i stor utsträckning, vilket gör att mycket av ansvaret för att tänka igenom hot och skydd faller på den egna organisationen.

Säkerställ därför att AI-system omfattas av samma sårbarhetshantering, loggning och incidentberedskap som övriga system, och att någon har ett tydligt ägarskap för säkerheten genom hela livscykeln.

4. Begränsa agentiska system aktivt

Agentiska system[2] är kraftfulla just för att de kan agera självständigt, men det är också därför de är riskfyllda. Ett system som självt kan planera, fatta beslut och använda verktyg kan, om det manipuleras eller går fel, orsaka skada i samma takt och skala som det annars skapar nytta. Skyddet måste därför byggas in i vad systemet tekniskt kan göra, inte bara i vad det är tänkt att göra.

Börja med att sandboxa (kapsla in) verktygsåtkomsten, så att agenten bara når de system och data den faktiskt behöver. Tillämpa minsta behörighet, varje verktyg, API och datakälla agenten kopplas till är en potentiell väg för både misstag och angrepp. Sätt tydliga behörighetsgränser och kräv mänsklig bekräftelse (human in the loop) innan känsliga eller irreversibla åtgärder utförs, exempelvis betalningar, ändringar av behörigheter eller extern kommunikation.

Säkerställ samtidigt full spårbarhet, logga agentens beslut och åtgärder så att de går att granska i efterhand, och se till att det finns ett sätt att snabbt stoppa systemet om det beter sig oväntat. Ett agentsystem med bred åtkomst och svag tillsyn utgör ett betydande riskscenario, utgå från vad som händer i värsta fall, inte bara i normalfallet.

5. Utbilda medarbetare kontinuerligt om AI-förstärkt manipulation

De gamla varningstecknen håller inte längre. Klumpig svenska, konstig formatering och uppenbart felaktiga avsändare var länge det som avslöjade ett bedrägeriförsök, men AI gör nu phishing välskriven, personlig och trovärdig, klonar röster utifrån några sekunders ljud och skapar falska dokument och videoklipp som är svåra att skilja från äkta. En medarbetare kan i dag få ett samtal som låter exakt som chefen, med en brådskande begäran som ser helt rimlig ut.

Eftersom mänsklig manipulation är fortsatt en vanlig angreppsväg måste utbildning vara löpande, inte en engångsinsats. Återkom regelbundet, använd realistiska och aktuella exempel och öva gärna med simulerade angrepp, så att igenkänningen sitter i ryggmärgen. Lika viktigt är att det är tryggt att ifrågasätta och att larma, en medarbetare som vågar stanna upp vid en misstänkt begäran är ett av de mest effektiva skydden som finns.

Komplettera medvetenheten med en rutin som inte är beroende av att människan gör rätt varje gång. Inför en separat verifieringskanal för betalnings- och kontoändringar – exempelvis en återuppringning till ett känt och i förväg verifierat nummer, och använd den konsekvent, även när begäran ser brådskande och självklar ut. Just brådskan är ofta en del av angreppet.

6. Leverantörskedjan och öppen källkod

Många AI- och it-system vilar på en kedja av leverantörer[3], molntjänster och programvarukomponenter och deras säkerhet blir i praktiken den egna. Kartlägg kritiska tredjepartsleverantörer, de som ni förlitar er på för att era digitala tjänster och leveranser ska fungera. Desto mer av verksamheten som samlas hos ett fåtal stora leverantörer, som moln- och AI-plattformar, desto större blir koncentrationsrisken om en av dom drabbas. Ställ säkerhetskrav i avtalen, säkra kontinuitets- och utträdesplaner och följ upp efterlevnaden. Hantera samtidigt era open source-beroenden aktivt, kartlägg vilka komponenter ni använder (gärna via en SBOM[4]), bevaka kända sårbarheter och var vaksam på skadliga eller kapade paket i leveranskedjan. Både köpta och fria komponenter kan vara en väg in och ska behandlas som en del av er attackyta.

NATIONELLT
CYBERSÄKERHETSCENTER
En del av FRA

Tjänster från
Sveriges nationella enhet för hantering av it-incidenter (nationell CSIRT), CERT-SE

Automatiska notifieringar om tekniska sårbarheter (ANTS)

ANTS hjälper svenska verksamheter att övervaka sina angreppsytor på internet. Syftet är att snabbt notifiera verksamheter när det identifierats företeelser som kan behöva åtgärdas i deras tillgångar på internet. ANTS är en kostnadsfri tjänst som är tillgänglig för alla svenska verksamheter, både i offentlig och privat sektor.

ANTS är en del i CERT-SE:s monitorering av Sveriges allmänna angreppsyta på cyberarenan. Syftet med monitoreringen är att minimera exponering av icke nödvändiga angreppspunkter samt att nödvändiga angreppspunkter exponeras på ett tillförlitligt sätt, utan kända brister i mjukvara och utan bekräftat osäkra konfigurationer. Det övergripande syftet är att stärka samhällets motståndskraft och minska effekterna av cyberangrepp som riktas mot det svenska samhället.

Proaktiv skanning

Proaktiv skanning består av kostnadsfria tjänster som är tillgängliga för alla svenska verksamheter, både i offentlig och privat sektor under förutsättning att de redan tar del av tjänsten ANTS.

Proaktiv skanning är en del av CERT-SE:s monitorering av Sveriges allmänna angreppsyta på cyberarenan. Syftet med monitoreringen är att minimera exponering av icke nödvändiga angreppspunkter samt att nödvändiga angreppspunkter exponeras på ett tillförlitligt sätt, utan kända brister i mjukvara och utan bekräftat osäkra konfigurationer. Det övergripande syftet är att stärka samhällets motståndskraft och minska effekterna av cyberangrepp som riktas mot det svenska samhället.

MISP-SE

MISP (Malware Information Sharing Platform) är en open source-plattform för att samla in, lagra, analysera och dela information om cyberhot och incidenter. Plattformen möjliggör strukturerad och standardiserad delning av hotindikatorer, exempelvis IP-adresser, domäner och skadlig kod.

MISP-SE är en nationell MISP-instans som svenska verksamhetsutövare kan ansluta sig till kostnadsfritt. Syftet är att stärka Sveriges samlade motståndskraft mot cyberangrepp genom delning av hotinformation.

MISP-SE bidrar till samverkan och effektivt utbyte av hotinformation vilket förkortar responstider och minskar skadorna av cyberangrepp. MISP-SE bidrar även till att skapa en lägesbild av aktuella cyberhändelser, både nationellt och inom specifika sektorer.

Mer information kring dessa tjänster finns på cert.se

1. https://www.ncsc.se/sv/cybersakerhet/10-rekommenderade-sakerhetsatgarder/
2. Agentiska system; AI-system som självständigt kan planera och utföra åtgärder via verktyg och gränssnitt.
3. Leverantörskedjan omfattar de externa leverantörer, underleverantörer, tjänster och programvarukomponenter som organisationens verksamhet och it-miljö är beroende av.
4. Software Bill of Materials, är en innehållsförteckning för programvara – en komplett lista över alla komponenter och beroenden som ett system är byggt av.

Nationellt cybersäkerhetscenter juni 2026

För mer information och kontakt
www.ncsc.se och ncsc@ncsc.se