Femton frågor om etik, juridik och säkerhet för framtidens AI-lösningar

knowit

Microsoft

knowit

Förord

Beslutsfattare i organisationer som utvecklar, implementerar eller använder verktyg som grundar sig på *Artificiell intelligens (AI)* kan behöva göra svåra avvägningar. AI-tekniken kan skapa stor nytta och många mervärden både för den egna organisationen och för samhället i stort. Med det sagt finns det flera *etiska*, *juridiska* och *cybersäkerhetsrelaterade* frågor att ta ställning till, och det kan vara svårt att veta om de beslut som fattas idag också är rätt beslut för framtiden.

Vi som har författat följande texter är erfarna rådgivare inom artificiell intelligens på Knowit, ett svenskt konsultbolag som stöttar företag, myndigheter och organisationer i den digitala transformationen. Vi är övertygade om att vi som samhälle har mycket att tjäna på att ta tillvara de möjligheter som nu skapas med hjälp av AI och vi arbetar tillsammans med våra kunder med att utveckla och implementera AI-lösningar på ett *etiskt*, *lagligt* och *säkert* sätt. Med dessa texter hoppas vi kunna hjälpa ännu fler att orientera sig i ämnet och resonera kring hur de utmaningar som finns kan hanteras.

I texterna som följer tar vi därför upp femton angelägna frågor inom *etik*, *juridik* och *cybersäkerhet* med bäring på AI. Vi resonerar kring utmaningar, möjligheter och konkreta exempel samt hänvisar till användbart referensmaterial för den som vill fördjupa sig ytterligare. Texterna kan läsas var för sig eller som en helhet.

Vi hoppas att våra texter kommer att vara av värde även i ert arbete med AI-system.

Författarna, maj 2024

knowit

Versionshistorik

Version	Huvudsakliga ändringar	Utgivningsdatum
1.0		2024-08-26

Innehållet i denna rapport är endast avsett som allmän information och utgör inte, och ska heller inte användas som, professionell rådgivning. Det kan förekomma att innehållet inte är uttömmande eller helt uppdaterat. Inga åtgärder eller beslut bör baseras på information tillgänglig i denna rapport som ersättning för juridisk rådgivning. Vid behov av rådgivning är ni varmt välkomna att kontakta oss. Direkt eller indirekt användning av innehållet i rapporten sker på egen risk.

Knowit AB (publ)
info@knowit.se
knowit.se

Arbetet med att ta fram detta dokument har finansierats av Microsoft AB med syfte att skaffa en större förståelse för vanliga typer av funderingar som väcks kring AI. Dokumentet uttrycker inte Microsofts officiella ståndpunkt i dessa frågor och i uppdraget har Knowit AB (publ) haft full frihet att formulera dokumentet utifrån sina utgångspunkter, observationer och slutsatser.

knowit

Författare, år 2024

Désirée Veschetti

Senior jurist med särskild fokus på informationshanteringsjuridik. Désirée har lång erfarenhet av informationshantering och informationssäkerhetsarbete i offentlig förvaltning. Som konsult hos Knowit stödjer hon olika verksamheter i utvecklingen av deras informationshantering och tillämpning av olika regelverk som påverkar användningen av data.

Fredrik Blix

Cybersäkerhetsexpert och doktor i cybersäkerhet vid Stockholms universitet, samt tidigare konsult vid Knowit. Fredrik ger råd, föreläser och debatterar om AI, cybersäkerhet och molntjänster. Han har 25 års erfarenhet inom cybersäkerhetsområdet som rådgivare, forskare och föreläsare.

Helena Tronner

Senior ledare, psykolog, specialistutbildad inom arbets- och organisationspsykologi, och expert på mänskligt beteende i organisationer. Helena besitter en gedigen erfarenhet av att skapa hållbara resultat och affärsvärde för organisationer genom ledarskap, utveckling och kontinuerliga förbättringar. Hon är en efterfrågad strategisk rådgivare och föreläsare.

Hugo Lloyd

Jurist på Knowit. Hugo arbetar med it-rätt och dataskydd inom offentlig och privat sektor. Han har ett stort intresse för hur teknik förändrar samhället och de juridiska utmaningar som uppkommer. Tidigare har Hugo arbetat vid Arbetsförmedlingen och Regeringskansliet där han bland annat utformat lagförslag och tillämpat olika typer av regelverk.

Jannika Törnqvist

Senior jurist med särskilt fokus på dataskydd. Jannika har erfarenhet av mångsidigt juridiskt arbete i olika roller i internationella företag, både som bolagsjurist och som extern rådgivare. I sin nuvarande roll bistår hon verksamheter huvudsakligen i egenskap av externt dataskyddsombud.

Lisa Lundin

Konsultchef och director för juridik på Knowit. Lisa har arbetat med dataskydd, it-rätt och förvaltningsrätt i mer än sjutton års tid i roller som verksjurist, dataskyddsombud och senior juristkonsult vid flera olika statliga myndigheter och välkända företag.

Nathalie Berggren

Senior ledare och rådgivare på strategisk och operativ nivå inom data, analys och digital transformation. Nathalie har mångårig erfarenhet av att arbeta med beslutsstöd, data- och informationshantering. Nathalie hjälper företag inom flertal branscher med att införa datadrivet beslutsfattande, strategier för datahantering samt att skapa förutsättningar för att nyttja AI.

Sofia Sigurdsson

Affärsjuriststudent vid Linköpings universitet och praktikant på Knowit. Under tiden på Knowit har hon fokuserat på frågor om AI, dataskydd och cybersäkerhet.

Stephanie Van Daele

Affärsjuriststudent vid Linköpings universitet och praktikant på Knowit. Under tiden på Knowit har hon fokuserat på frågor om AI, dataskydd och cybersäkerhet.

knowit

Innehållsförteckning

Definitioner och referenser

I följande text kommer vi att använda ett par snarlika men olika begrepp när vi skriver om AI. För tydlighetens skull definierar vi dem här.

AI-algoritm	Den kod som styr hur en AI-modell lär sig och fungerar. Algoritmen bestämmer hur modellen bearbetar data och gör förutsägelser.
AI-modell	Ett program tränat på data för att känna igen mönster eller fatta beslut utan mänsklig intervention. Modeller använder olika algoritmer för att bearbeta data och utföra uppgifter.
AI-system	En maskinbaserad enhet som kan utföra uppgifter som normalt kräver mänsklig intelligens, såsom exempelvis taligenkänning och beslutsfattande.
AI-teknik/ teknologi	Övergripande term för metoder och verktyg som används för att skapa och implementera artificiell intelligens. Det inkluderar algoritmer, ramverk och verktyg för att bygga AI-system.
AI-tjänst	Ett AI-system som levereras som IT-tjänst av en tillhandahållare, ofta som en molntjänst.

Sist i det här dokumentet finns det också en lista med fullständiga referenser till litteratur, regelverk och författningar, rättsfall och standarder som vi hänvisar till i texten.

Etik

knowit

Går det att skapa AI-system som är fria från bias och fattar rättvisa beslut?

Etik i AI-användning är en komplex och avgörande fråga. Den snabba utvecklingen av AI-teknologi för en mångfald av ändamål kan innebära potentiella fördelar för samhället och välfärden. Samtidigt är det viktigt att överväga de etiska konsekvenserna. En central etisk frågeställning är hur eller om det är möjligt att säkerställa att AI-system fattar beslut på ett rättvist sätt och inte är partiskt.

Partiskhet, eller bias, refererar till systematiska skevheter som kan missgynna specifika grupper av människor. Det innebär att det inte finns någon fullständig neutralitet där alla parter behandlas lika. För utvecklare och användare av AI-beslutssystem är det essentiellt att vara medveten om bias och dess potentiella inverkan på rättvisan i beslutsprocesser. Förmågan att snabbt bearbeta stora datamängder och fatta beslut kan ha betydande följder för särskilda minoriteter.

Utmaningar och lösningar

Vikten av att hantera bias

Det går inte att helt utesluta bias inom AI-system, utan utmaningen ligger i att hantera dem. Diskussionen kretsar ofta kring statistisk bias, som påverkas av faktorer som datasetens representativitet, datainsamling, algoritmers rättvisa och datatolkning. Data som inhämtas för att träna våra AI-modeller och system kan både tillhandahållas av system och människor. Det medför att det finns en spridning i hur väl data är insamlat och processat, samt hur reglerat eller oreglerat det sker. Systematisk och mänsklig bias är också viktig att beakta, eftersom de påverkas av samhälleliga normer och inlärda beteenden, såsom kognitiv bias. Dessa biaser påverkar rättvisan i beslutsfattandet.

Maskin-bias

Det finns idag en oro i den akademiska världen om fenomenet som kallas maskin-bias. Det innebär att AI-baserade robotar och verktyg, utan deras skapares kännedom, skapar kontroversiella samhälleliga asymmetrier, såsom bias baserat på etnicitet eller hudfärg. Ett stort antal AI-verktyg har visat sig vara skadligt partisk mot någon minoritet, med rapporter om bland annat rasistiska beteenden. Exempelvis klassificerade Google Photos svarta människor som gorillor (Curtis, 2015). Med tanke på att alla vi människor har bias och att den data vi låter AI-

modeller tränas på också innehåller bias finns det stora risker att bedömningar blir asymmetriska, till minoriteters och avvikares nackdel. Människor tenderar även att lättare ta till sig ny information som redan går i linje med vad individen tycker, också kallad bekräftelsebias. Det ökar riskerna att AI kan användas för att förstärka asymmetrierna i samhället.

Sätt att minimera bias

Det är omöjligt att få bort all bias, men att minska risken för bias är något vi kan göra. Det kräver ett stort ansvar i utvecklandet, införandet och användandet av AI-system. Eventuella punkter att beakta för att minska bias är följande:

- **Syfte och intention:** Förstå syftet och kontexten med AI-systemet som verksamheten utvecklar och använder.
- **Urval av data:** Säkerställ att dataset är representativt och anpassat för syftet med applikation och kontext. Utgå inte enbart från vad för data som är tillgängligt, utan byt ut och nyansera data som AI-modellen får tränas på om det finns behov.
- **Medvetenhet:** Var medveten och reflektera över vilka frågor som behöver ställas i arbetet med AI-modeller. Det gäller både vid utvecklingen och användandet av AI-system. Säkra förståelse och medvetenhet om de olika bias som finns.
- **Test och validering:** Använd mångsidiga data samt träna och testa din modell för att identifiera och minimera bias. Föreligger det någon skev slutsats i utfallet som kan härledas till bias? Här finns möjlighet att nyttja metoder som exempelvis anti-klassificering, vilket innebär att viss information raderas från dataset för att minska risk för bias. Det finns även modeller som kan hjälpa till med att minska risken för inneboende bias genom att använda befintliga data från en mängd datakällor och på så sätt framhäva hur data tenderar att gravitera samt bistå i att generera mer objektiva svar s.k. RAG-modeller (Retrieval Augmented Generation), (Şimşek, 2024).
- **Mänskligt ägarskap:** Agera efter AI-policys som innefattar att det är människor som är ytterst ansvariga för resultatet samt säkerställ att AI-modellen fungerar på ett sätt som går att förstå för de som förlitar sig på dess resultat.

För att inte förstärka rådande stereotyper och orättvisor i samhället behöver människor nyttja sitt kritiska tänkande och fortsätta att utveckla förmågan till etiska och moraliska bedömningar. I en publikation från NIST (National Institute of Standards and Technology) förespråkar författarna ett socio-tekniskt tillvägagångssätt för att minska AI:s partiskhet, vilket erkänner att AI verkar inom en social kontext och att tekniska lösningar ensamt är otillräckliga. (Schwartz m fl, 2022). Det finns med andra ord ett behov av att reflektera över vilka beslut som AI kan vara involverad i och vilka beslut som människan inte bör delegera till

tekniken. AI kan fatta beslut, göra avvägningar och bedömningar snabbt, men människor behöver granska resultatet för att det ska bli tillförlitligt.

Exempel

Exempel 1
"COMPAS" var en AI-algoritm som användes av domstolar i USA för att föreslå risknivå för individer som skulle ligga till grund för ett rekommenderat straff. Modellen tränades på befintliga data från tidigare domar. Det resulterade i en stor skandal, eftersom algoritmen rasprofilerade och uppvisade ”representativitetsbias.” Personer med andra hudfärger än vit bedömdes utgöra en större risk och blev då tilldelade strängare straff i jämförelse med vita personer som hade genomfört liknande brott (Chawla, 2022).

Exempel 2
Amazon hade som ambition att öka antalet kvinnliga kandidater, och implementerade därmed en AI-modell för rekryteringen. Modellen tränades genom att observera befintliga medarbetare med hög prestanda, och sorterade till följd av detta bort ord som involverade ”Women's”. Exempelvis blev kandidater som var medlemmar i en kvinnlig schackklubb bortsorterade. AI-modellen favoriserade kandidater som beskrev sig själva med manligt könskodade ord som återfanns på manliga ingenjörers CVn, såsom ”executed” och ”captured” vilket skapade en skevhet inom tänkt rekrytering.

Exempel 3
En studie omfattade en lista över yrken som användes för att bygga meningar i konstruktioner som "han/hon är ingenjör" i tolv olika könsneutrala språk som ungerska, kinesiska, yoruba. Studien visade att Google Translate uppvisade en stark tendens till manliga standardinställningar, särskilt för områden som vanligtvis förknippas med obalanserad könsfördelning eller stereotyper som vetenskap, teknik, ingenjörskap och matematik (Prates, Avelar, Lamb, 2018).

Slutsats

Det är inte möjligt att helt eliminera bias. Det är i stället avgörande att vi är medvetna om och strävar efter att minimera bias i största möjliga utsträckning. Det är nödvändigt att hantera statistisk, systematisk och mänsklig bias, samt att förstå AI-systemens kontext och syfte. Utveckling och användning av AI bör baseras på rättvis och relevant data. En grundläggande del är att AI inte ska ersätta mänskligt beslutsfattande. Människor är nödvändiga för att tolka och nyansera det som data visar.

Hur mycket autonomi ska vi ge AI-system och vem ska ha kontroll över dem?

Med autonomi för AI-system menar vi att AI-system får förmågan att fatta beslut och agera på egen hand. Det är ett komplext begrepp som kan vinklas på olika sätt. Vi avser främst att resonera kring hur “fritt” ett AI-system ska tillåtas vara och agera. Vi avser även att diskutera vem som bör ha kontroll över ett AI-system, specifikt människan och dennes roll i att fungera som kontrollfunktion för ett AI-system.

Vi som utvecklar och inför AI-system idag – hur mycket självständighet kan vi ge våra system? Vad innebär det att ha kontroll och vem ska egentligen ha den?

Utmaningar och lösningar

Gasa eller bromsa?

AI-system växer och utvecklas i en allt snabbare takt och blir därmed mer komplexa, kraftfulla och autonoma. För närvarande är det människor som bestämmer syfte och mål för en AI-modell, och den agerar i enlighet med dem. Vad händer om en AI-modell lär sig hur den kan sätta egna mål? Med mer komplext, snabbt och effektivt beslutsfattande – hur ska vi människor kunna stoppa den? Kraften i AI:s egen utvecklingsförmåga och potential behöver vägas mot eventuella konsekvenser. Vad händer om det går för långt och vi inte kan stoppa det?

Tre punkter som kan radera mänskligheten

Max Tegmark, som driver Future of Life Institute, lyfter tre anledningar som kan leda till att människan inte har en plats om AI får utvecklas helt fritt. Först handlar det om att AI kan användas **illvilligt**. Vidare, menar Tegmark, kommer de bolag som låter AI ta över prestera mycket bättre än de som inte använder AI. Det medför att **konkurrensen** ”tvingar” bolag att låta AI fatta företagsbeslut.

Slutligen varnar Tegmark för en **felinriktad** AI, där han menar att det är svårt att säkerställa att AI förstår, anammar och bibehåller samma mål som människor (Tegmark, 2023). Individer fattar bäst beslut när det enbart finns tre till fyra alternativ att välja mellan. Idag, när vi utsätts för så många fler val än tidigare behöver vi ta hjälp av AI, men frågan är hur vi gör detta utan att bli förblindade

eller släppa taget om besluten helt? Här behöver människor på en samhällelig nivå resonera kring etiska och moraliska bedömningar samt balansera mellan två motpoler. Å ena sidan vill vi skapa vinster i bolag, framsteg inom forskning och lösa svåra mänskliga problem, vilket gör att vi behöver hjälp av AI. Å andra sidan vill vi inte att AI på ett okontrollerat och systematiskt sätt ska ta över styrningen av organisationer och samhällen.

- **Gasa eller bromsa?** Frågan är komplex och det kan vara nödvändigt att eventuellt gasa och bromsa utvecklingen parallellt, där vi jämsides med den tekniska utvecklingen styr på ett nära håll och i vissa fall begränsar utvecklingen när vi ser att den går åt fel håll.
- **Människan i fokus:** Till skillnad från människor saknar AI självmedvetenhet, självinsikt och känslor. Människor besitter emotionell intelligens och kan göra komplexa moraliska bedömningar. Det innebär att människor behöver ta ansvar och säkra goda intentioner för att bedriva utvecklingen av AI. Det kan ske genom samarbeten och gemensamma policys, samt gränsöverskridande överenskommelser som är förankrade i mänskliga rättigheter.
- **Gemensamt ansvar:** En samverkan mellan bolag, samhälle, och samtliga instanser behövs för att hantera utvecklingen, där en mångfald av yrken och kompetenser behöver representeras. Beslutsrätten bör inte lämnas ensidigt till en grupp, fåtal individer, eller bolag.
- **Kontroll och begränsningar:** Med förståelse för AI och dess kraft, särskilt vid en eventuell negativ inriktning, behöver vi säkra miljöer och ha flera instanser samt ramverk att ta hjälp av. Det behövs styrning och kontroll, där utfall kontinuerligt utvärderas mot grundfrågor inom etik och mänsklighetens framtid.

Exempel

Microsoft släppte år 2016 sin chatbot Tay på en social plattform som ett experiment i samtalsförståelse. Chatboten skulle anta en roll som en flicka i tonåren och interagera med individer på Twitter genom maskininlärning och språkteknologi. Chatboten tränades på anonymiserade, publika data, samt i viss omfattning förskrivet material. Den “släpptes lös” för att lära och utvecklas från sina interaktioner. Inom 16 timmar hade chatboten skrivit över 95 000 tweets, vars innehåll snabbt utvecklades till rasistisk, kvinnohatande och anti-semitiskt material. Microsoft tog ansvar för konsekvenserna och stängde snabbt ner chatboten. Trots en omfattande testperiod med miljontals framgångsrika interaktioner och träning i att hantera missbruk, tydliggjordes svårigheterna att täcka alla illvilliga intentioner när AI får frihet att utvecklas självständigt.

Slutsats

AI har potentialen att utveckla flertalet samhällsområden, med betydande möjligheter och risker. Hittills har vi inte sett att AI kan operera fritt; snarare behöver vi som utvecklar, inför och använder AI ökad insyn och kontroll för att säkerställa en positiv inriktning och framsteg. AI bör tillåtas viss autonomi då den kan överkomma hinder snabbare än människor, vilket kan gynna oss i vissa avseenden. Kontroll är högst nödvändigt, och för att nyttja AI och dess potential behöver vi guida och leda utvecklingen samt balansera det mot mänskliga frågor. För att citera Max Tegmark, ”teknik ger livet möjlighet att blomstra som aldrig förr – eller att förstöra sig självt” (Tegmark, 2017), och däri ligger balansen kring hur mycket autonomi vi kan ge AI. För att täcka alla aspekter och undvika att överlämna AI:s framväxt till enbart några få individer, företag eller yrken, bör kontrollen innehas av flera roller och expertisområden i samhället. Vi behöver en mångfaldig närvaro inom denna utveckling och den mänskliga kontrollfunktion som vi behöver fylla.

Vilka etiska överväganden är viktiga när AI-tekniken påverkar våra arbeten?

AI kommer att förändra framtiden för arbete på flera sätt. Exempelvis kan många arbetsuppgifter vi gör idag effektiviseras bort. Ett teknologiskt skifte och transformation pågår där AI driver på. Vi ser att vissa arbeten kommer försvinna, vissa kommer transformeras och nya skapas. Europaparlamentets utredningstjänst har estimerat att 14% av arbeten i OECD-länderna kan utföras av maskiner och ytterligare 32% förväntas genomgå omfattande förändringar. Det innebär att många kommer påverkas i denna förflyttning (Parlamentets utredningstjänst, 2020).

Med etiska överväganden vill vi resonera om människan och hur vi påverkas av denna förflyttning och vilket ansvar samhället har för medborgaren. När vi diskuterar förflyttning av mänskliga jobb på grund av AI-teknik berör vi hur AI-teknik påverkar oss i vårt arbete och även kring hur det kan slå mot grupper i samhället som berörs mer eller mindre.

Utmaningar och lösningar

Människans förändringsförmåga

Ett teknologiskt skifte är inget nytt fenomen och historien visar att vi sällan helt går miste om arbeten, utan det är mer en transformation av hur arbeten ser ut och utförs. För individer som plötsligt förlorar sitt arbete och står inför utmaningen att byta kompetens är det viktigt att samhället tar ansvar för att stödja dem under denna övergångsperiod. Samhället gynnas inte av utanförskap och det ligger ett stort ansvar i att vara förberedd och ha en plan för att stödja de som behöver skolas om eller adapteras till det nya arbetslivet. Privata bolag bär också ett stort ansvar där de genom att ha en intention att vilja och kunna omskola och tillse resurser för detta kan bidra till transformationen. kan de bidra till transformationen. Organ som skolväsendet behöver anpassa sig och säkra utbildningar och förmåga att hantera den transformation av kompetenser som behövs i framtiden.

Påverkan på enskilda grupper i samhället

En annan intressant aspekt är hur AI kommer att påverka könsskillnaderna i samhället. Jobb som lättast kan automatiseras bort är av administrativ karaktär (Gilan, 2023) och det berör flest kvinnoyrken, såsom administration och ekonomi (Nordström, Schlingmann, 2014). Samtidigt i en alltmer digitaliserad värld är det

de egenskaper som är svårast att digitalisera som kommer att värderas mest. Flera menar att de förmågor som kommer vara viktigast i framtiden är kommunikation, samverkan och relationsbyggande snarare än teknisk kompetens. Gilan lyfter upp "De tre K:na" som syftar på kreativitet, kärlek och känsla (Gilan, 2023). Utifrån detta perspektiv finns det en del som menar att kvinnor kommer att besitta fördelar i framtidens organisationer. Det finns studier som visar att kvinnor generellt är bättre än män på emotionell intelligens (Goleman, 2011) och att kvinnor har en högre grad av social känslighet (Bear & Woolley 2010). Urban Express menar författarna att kvinnors emotionella intelligens och kommunikationsförmåga kommer att göra dem till framtidens superstjärnor (Nordström, Schlingmann, 2014).

- **Samhällets agerande:** Samhället behöver agera fort och ha tydliga planer gällande hur en omställning inom utbildningar och stödinsatser för de grupper som påverkas mest kan sättas in.
- **Investering i utbildning:** Trots det faktum att dagens utbildningar har ett stort fokus på den tekniska utvecklingen inom AI behövs det fler utbildningar om hur vi kan lära oss att nyttja AI och stärka våra olika professioner med hjälp av AI-teknik.
- **Anpassning:** Samhället behöver tillhandahålla enklare, mer individuella, anpassade och lättillgängliga utbildningspaket i en transformation. Framför allt är det viktigt att stötta alla individer i samhället och inte endast en majoritetsgrupp utan säkra anpassning för alla grupper och skapa en mer inkluderande arbetsmarknad för alla.
- **Förändringsförmåga hos individer:** Människor behöver vara öppna, ödmjuka, våga experimentera och hantera förändring genom att stärka sin förmåga att lära om.

Exempel

I en empirisk studie av McKinsey 2023 framgick det att utvecklare som använde generativ AI i sitt arbete kunde slutföra uppgifter såsom kodgenerering avsevärt snabbare och mer effektivt än tidigare. Uppgifter som att generera kod och dokumentation kunde utföras på nästan hälften av tiden det annars tagit. Samtidigt genomfördes en undersökning om hur utvecklare upplevde arbetsmomenten där utfallet visade att många av utvecklarna upplevde en större nöjdhet under momenten och hade möjlighet att fokusera bättre när de använde generativ AI till sin hjälp.

Studien pekar på flera områden där generativ AI förbättrar upplevelsen för utvecklare och effektiviserar arbetsmoment. Den visar också att AI inte har lika stor effekt vid mer komplexa uppgifter. Med AI:s hjälp kan vi förenkla delar av vårt arbete vilket leder till mer tid för komplexa uppgifter och en större effektivitet i vår vardag.

Slutsats

Vi står inför en omfattande teknologisk omvandling som kommer att påverka arbetsmarknaden. Genom att blicka tillbaka på vår historia ser vi att vi återigen kommer att genomgå en förändring snarare än en förlust av arbeten i sig. Oavsett kräver det mycket av individen, samhället, bolag och organisationer då vi alla bär ett ansvar i hur vi genomför förändringen. Stora investeringar krävs för att omskola och lyfta kompetenser samtidigt som det kommer finnas en efterfrågan av specifika förmågor inom exempelvis relationer och vård. Även om det är oundvikligt att individer påverkas av transformationen, ser vi redan nu att en del arbeten har gynnats av förändringen. Trots att arbetet kanske har förändrats i sin natur, fortsätter det att finnas kvar, och nöjdheten hos individen ökar.

Hur hanterar vi bias i AI-algoritmer som används inom hälso- och sjukvården?

Idag finns många och stora utmaningar inom hälso- och sjukvård runtom i världen. Resursbrist och ojämlikhet mellan länder och möjlighet till vård är en stor och viktig fråga att lösa. Det finns stora vinster att hämta på ett samhällsmässigt plan. Det finns flera exempel på att läkare som använder AI presterar bättre än läkare som inte använder det som hjälpmedel. Nästan all sjukvård kommer att påverkas av AI (SVT nyheter, 2023), samtidigt som hur det ska gå till är komplext, inte minst när det finns olika bias som påverkar och kan resultera i konsekvenser inom området.

Med bias avses främst en systematisk snedvridning i bedömningsprocesser som leder till att vissa grupper eller parter missgynnas. Frågan är av stor vikt även i de fall som berör uppgifter om etnicitet och socioekonomisk status samt hur dessa faktorer kan påverka utfallet av bedömningen. Med socioekonomisk status avses en persons ställning i samhället. Den grundar sig ofta på faktorer så som exempelvis utbildning, yrke, inkomst.

Utmaningar och lösningar

Påbyggnad av bias till förmån för snabb utveckling

Att nyttja algoritmer för att snabbare och mer effektivt bedriva vård kan ha många vinster i samhället. Det finns även flera stora utmaningar. Inom vården hanteras exempelvis stora och komplexa datamängder, vilket gör att det är oundvikligt att bias påverkar dessa processer. Återigen handlar det om att minska dessa och att utmana vår bias kontinuerligt. Om vi bygger vidare på data med bias resulterar det i fortsatta skeva bedömningar och att gynnsamma insatser fortsätter fördelas till normgrupper i samhället. I dagsläget står vi redan inför stora utmaningar när det gäller kön, etnicitet och socioekonomiska förhållanden.

Tillgång till data

För att tillhandahålla en bättre hälsovård och jämna ut olikheter kommer data vara en möjliggörare. Med inhämtning och bearbetning av stora och komplexa datamängder väcks även frågan om vi använder den på ett sätt som inte är partiskt, det vill säga att samtliga individer i samhället behandlas lika. Samtycke blir en stor fråga, då för de fall det krävs samtycke som inte ges av individen, kan

det leda till att vi inte kan samla in all data och därmed har vi indirekt en inbyggd systemisk bias. Här blir transparens viktigt, samt att individer förstår nyttan och intentionen med insamling av viktiga data.

Korsrefererande data

Även när vi är medvetna om de bias som finns och tagit hänsyn till att utesluta en del parametrar av data för att minska vår bias, är det ändå inte helt vattentätt. Att utesluta vissa datapunkter som vi anser direkt påvisar exempelvis etnicitet eller socioekonomisk status i en AI-modell kan tänkas vara en lösning. Med all övriga data som är tillgänglig kan dock en AI-modell relativt snabbt och enkelt börja lägga sitt pussel och därmed få fram en trolig bild av vår etniska tillhörighet och socioekonomiska status, vilket är viktigt att ta hänsyn till om man ser att dessa datapunkter inte bör tas med i modellen. Värt att notera är att det kan finnas fall där dessa faktorer skulle vara av vikt att ta med för att de exempelvis gynnar individen eller krävs till följd av gällande lagstiftning på området. Likt många andra fall inom AI-användande finns inget fullständigt och ensidigt svar på frågan, därför är syftet och även förståelse för modellen och data den tränas på av yttersta vikt. Det här är några viktiga insikter vi anser att man bör ta med sig:

- **Transparens och dataskydd:** Det är avgörande att det finns en transparens vid insamling och användning av data samt korrekt hantering av denna för att skydda individers integritet och säkerställa att data används på ett etiskt sätt.
- **Korsreferens av data:** När data kan korsrefereras bör det göras med försiktighet särskilt när det gäller känsliga kategorier som kön, etnicitet och socioekonomisk status.
- **Förståelse och testning:** Det är nödvändigt att AI-modeller är begripliga för att det ska finnas en möjlighet att förklara dess slutsatser och även kunna utvecklas. Algoritmer och system bör genomgå rigorösa tester och valideringar av oberoende parter för att identifiera och korrigera eventuell bias.
- **Mänsklig översyn och insyn:** Det är viktigt att alltid ha mänsklig översyn i beslutsprocessen för att säkerställa att AI-systemets rekommendationer är lämpliga och rättvisa.

Exempel

Exempel 1
I Storbritannien löper svarta kvinnor fyra gånger större risk att dö under graviditet och förlossning jämfört med vita kvinnor. Denna alarmerande statistik kan delvis tacklas genom användningen av AI för att förbättra graviditetsvården. Genom att nyttja ett AI-program som eliminerar hudfärg och etnicitet som faktorer i diagnostiska processer säkerställde programmet att svarta och asiatiska kvinnor fick mer precis och tidig vård. (MBRRACE-UK, 2021).

Exempel 2
En studie vid Stanford 2016 visade att AI skulle kunna diagnosticera lungcancer med hjälp av mikroskopbilder ännu bättre än patologer. Forskarna fann att en maskininlärningsmetod för att identifiera kritiska sjukdomsrelaterade egenskaper korrekt skilde mellan två typer av lungcancer och förutspådde patientöverlevnadstider bättre än standardmetoden för patologer som klassificerar tumörer efter grad och stadium. Denna metod tros kunna användas för många olika typer av cancer och har förmågan att lyfta arbetet inom cancervård och potentiellt vara fantastiskt för både patienter och läkare. (Conger, 2016).

Slutsats

Implementeringen av AI-teknologi inom hälsovårdssektorn erbjuder betydande fördelar, men väcker även frågor om hur välgrundade beslut det kan bidra till. Det är av stor vikt att fokusera på hur bias hanteras och vilken inverkan det får på ett område som direkt berör individers liv. För att hantera denna komplexitet krävs noga övervägande av hur AI är tänkt att vara ett hjälpmedel, hur modellerna ser ut och vem som har sista ordet i beslut som rör individers hälsa.

Det är viktigt att de AI-modeller som används är begripliga för användare och beslutsfattare samt att data som modeller har tränats på har testats och noggrant utvärderats utifrån bias. Med hjälp av metoder för att minska bias i data samt mänsklig insyn och kontroll kan AI vara ett värdefullt hjälpmedel inom vården. Slutligen är det avgörande att människan finns med i beslutsprocessen, och att beslutsfattare inte enbart förlitar sig på AI-systemets rekommendationer.

Hur hittar vi en bra balans mellan allmänhetens säkerhet och individens integritet vid AI-övervakning?

Med AI och dess teknik idag är det möjligt att övervaka större delar av samhället ner på individnivå. Det medför möjligheter att öka säkerheten, men innebär även en stor inskränkning på individers integritet.

Med etiska gränser avser vi vilka handlingar som anses vara rätt eller fel och hur vi sätter gränser utifrån etik och moral. Med allmänhetens säkerhet menar vi främst i denna fråga polisiärt och hur vi skyddar vårt samhälle från kriminalitet. Med en individs rätt till personlig integritet avser vi rätten att få sin individuella sfär och sina gränser respekterade av andra.

Utmaningar och lösningar

Förväntningar från allmänheten

En stor fråga i dagens samhälle är hur vi ska skydda oss mot brottslighet som ökar. Det finns förväntningar, både från allmänheten och brottsoffer, att polisväsendet ska nyttja tillgängliga data och teknik i brottsbekämpningen. Däremot är det diskrepans mellan allmänhetens förväntningar och vad som är lagligt möjligt, vilket kan skada förtroendet för polisväsendet i stort.

AI för minskad eller ökad brottslighet?

Med möjlighet till en utbredd övervakning i samhället kan både grupperingar och individer kartläggas och följas. Det skulle kunna ge ett försprång i brottsbekämpningen av exempelvis gängkriminalitet. Samma information skulle däremot också kunna användas av en mer illvillig sida, där kartläggningar och mönster nyttjas för att bredda brottslig verksamhet. Det blir en fråga om vem som har tillgång till informationen och avsikterna bakom dess användning.

Kostnad, nytta och delning av data

Det är också viktigt att beakta att implementeringen av övervakningssystem är både komplex och förenad med betydande kostnader för samhället, där det är avgörande att säkerställa nyttan. Dessutom är det viktigt att diskutera och belysa vilka som anses vara lämpliga informationsparter att dela data med. Även frågan

om vilka myndigheter och instanser som ska få dela data mellan sig och hur de får nyttja den för beslut och åtgärder.

Integriteten påverkas

I brottsbekämpande åtgärder kan det verka enkelt att kartlägga individers vanor och beteenden eller med en övervakningslösning urskilja individer vid pågående brott. Även om sådana åtgärder skulle kunna bidra till att bevisa individers brott kan det också användas för profilering i stort av samhällets individer. Det resulterar i en fråga om individers integritet, då information kring hälsa, familjesituation, sexuell läggning, otrohetsaffärer eller liknande kan förekomma. Information som finns tillgänglig, men inte är relevant för ett visst beslut, kan drabba individens trovärdighet och således påverka beslutet ändå. Vidare kommer det att utsätta vissa grupper i samhället för återkommande negativ särbehandling, på grund av bias. Några punkter som är värda att beakta:

- **EU:s AI-förordning (AI Act):** En EU-rättsakt som reglerar användandet av AI-system inom EU. Förordningen klassificerar AI i olika riskklasser och fastställer specifika krav därefter. Ett övervakningssystem klassas som ett “högrisk” AI-system. Bolag som utvecklar eller använder dessa måste följa angivna krav för att skydda individers integritet. En av de svårlösliga frågorna i förhandlingarna av AI Act gäller undantag för brottsbekämpningens möjligheter att använda AI.
- **Transparens och information:** Ett AI-system som används för övervakning måste vara transparant. Människor bör veta när och hur övervakning sker. Allmänheten ska vara informerad och medveten om hur data samlas in och används.
- **Mänsklig involvering:** Människor ska vara involverade i diskussionen om övervakning och integritet. Information och dokumentation ska finnas lättillgängligt och vara anpassat på ett sätt så att alla kan ta del av den.
- **Datahantering och delning:** All data som samlas in bör hanteras varsamt och med stor omsorg. Tekniker som psuedonymisering eller anonymisering bör användas i syfte att minska risken att identifiera enskilda personer. Hur en delning av samma data ska gå till för att tjäna sitt syfte behöver noggrant övervägas och säkras.
- **Riskbedömning:** Noggranna riskbedömningar av AI-systemet och dess implementation bör genomföras. AI-system bör undvika diskriminering och respektera mänskliga rättigheter.

Exempel

Exempel 1
Våren 2021 började SL experimentera med ett automatiskt larmsystem i tunnelbanan för att minska olyckor och dödsfall. Det var ett övervakningssystem där smarta kameror kände igen rörelsemönster och situationer, samt larmade trygghetscentralen vid upptäckt. Hittills beräknas 17 liv ha räddats med hjälp av den nya tekniken som gör att händelser visas i realtid för operatörer som snabbt kan agera (Beckman, 2024).

Exempel 2
Clearview AI, utvecklade en applikation som, till skillnad från mer traditionella ansiktsigenkänningsteknologier, samlar ansiktsbilder från sociala medieplattformar som Facebook och Instagram. Bilderna lagras i en stor databas som säljs till brottsbekämpande organ och privata säkerhetsföretag. I Sverige testade poliser Clearview AI i brottsutredningar, men efter en officiell granskning ansåg Integritetsskyddsmyndigheten (IMY) att polisens användning av applikationen var olaglig. Exemplet belyser hur myndigheter kan lockas av att använda kraftfulla, men inte alltid sanktionerade, tekniker i syfte att effektivisera sitt arbete. (IMY, 2021)

Exempel 3
AI och GIS för att bekämpa gängkriminalitet och ekonomisk brottslighet.
Framgångsrika tillämpningar av AI och GIS inom brottsbekämpning - Los Angeles Police Department (LAPD) använder exempelvis ett program för förebyggande polisarbete som analyserar data från olika källor såsom brottsrapporter, sociala medier, gängdatabaser för att identifiera områden med hög risk för brottslig verksamhet. Det har visat sig vara framgångsrikt när det gäller att minska brottsligheten inom dessa områden.

Slutsats

Sammanfattningsvis är det avgörande att balansera allmänhetens säkerhet med individens rätt till integritet. Genom att följa lagstiftning, vara transparenta och föra etiska diskussioner kan vi säkra individers integritet även när AI övervakar oss.

Juridik

knowit

Hur skyddas immateriella rättigheter i AI-genererat arbete?

Inledning

Immaterialrätt är ett paraplybegrepp för grundläggande mänskliga rättigheter, enligt artikel 27.2 i FN:s allmänna förklaring om de mänskliga rättigheterna, som berör till exempel upphovsrätt, patenträtt, mönsterrätt, databasrätt, och som syftar till att skydda idéer, uppfinningar och verk. En av grundtankarna med regleringen är att upphovsmannen ska få ersättning och rätt att förfoga över sitt arbete. Vid utveckling av AI-modeller krävs ofta stora mängder data. Det innebär att organisationer behöver vara uppmärksamma på vilka uppgifter som används vid träningen för att inte inkräkta på någon annans upphovsrätt. En annan utmaning handlar om att AI-modeller kan producera utdata som kränker annans upphovsrätt. Det finns därför anledning att närmare utreda vad upphovsrätten har för betydelse vid utvecklingen och användningen av AI-system.

Utmaningar och lösningar

Skyddet vid insamling och träning

Enligt 2 § upphovsrättslagen (1960:729) (URL) innebär upphovsrätten bland annat en rätt att förfoga över sitt verk, inbegripet att framställa exemplar av det och göra det tillgängligt för allmänheten, i ursprungligt eller ändrat skick. Skyddet medför till exempel att upphovsmannen till verket besitter ideella och ekonomiska rättigheter, vilket innebär, enligt 1 kap. 3 § URL, att upphovsmannen kan kräva att bli angiven som upphovsman när ett verk kopieras eller görs tillgängligt för allmänheten (Kempas s. 49, 2023). Samtidigt är det enligt 4 § 2 p. URL möjligt att inspireras av andra verk för att skapa ett nytt men självständigt verk i förhållande till ursprungsverket (det vill säga skapat i fri anslutning). En oberoende upphovsrätt kan därmed uppstå, förutsatt att det nya verket är unikt (uppvisar så kallad verkshöjd) och har en individuell utformning (Dom Infopaq, C-5/08, p. 38). Skyddet uppstår automatiskt och utan formkrav såsom ansökan. Det medför att upphovsrättsliga frågeställningar blir högaktuella i den digitala utvecklingen och skapandet av AI-genererat arbete. Genom att använda AI-verktyg är det möjligt att skapa nya verk med hjälp av insamlade träningsdata vilket väcker frågor om till exempel möjligheten att använda upphovsrättsskyddat material för att utveckla AI-system, om AI-genererade verk kan få upphovsrättsskydd och vem som ska anses vara upphovsman.

Ett AI-system tränas på data genom att använda avancerade algoritmer för att analysera och lära sig mönster och samband i dataunderlaget. Maskininlärning är en vanligt förekommande AI-teknik. Vid maskininlärning identifieras strukturer och återkommande mönster i datasetet (indata). Därefter skapas en modell som kan användas för att behandla nya data och ge statistiskt rimliga slutsatser utifrån träningsdatasetet (utdata) (IMY 2024). Resultatet av AI-systemets bearbetning av insamlade data är en teknisk process som saknar den kreativitet som finns i en människas personliga upplevelser, känslor och tolkningar vid skapande av exempelvis ett verk.

Om träningsdata som används är skyddat av upphovsrätt finns det en risk för att användningen av data gör intrång i någon annans rättigheter (Levendowski s. 582, 2018). Frågan om hur själva framställningen av AI-genererade verk är att betrakta som intrång genom otillåten användning av ett upphovsrättsskyddat verk är därför central. Av undantaget i 15 § URL för text- och datautvinning (TDM) får den som har lovlig tillgång till ett verk framställa exemplar av verket för text- och datautvinningsändamål. Exemplaren som använts för framställning av verket får däremot inte behållas längre än vad som är nödvändigt för ändamålet och får inte användas för andra ändamål. Upphovsmannen kan däremot förbehålla sig denna rätt i vilket fall TDM-undantaget inte ska gälla. Med andra ord är användningen av verk som träningsdata i AI-system olaglig om upphovsmannen har förbehållit sig sådan användning.

Skyddet av utdata och AI-system

AI-genererade verk (utdata) skiljer sig från det vi normalt anser vara verk i URL:s mening i och med att de skapas utan mänsklig inblandning (WIPO s. 4, 2020) (EPRS s. 4, 2020). Till skillnad från verk som skapats av en människa är det inte lika tydligt hur AI-genererade verk kan erhålla upphovsrättsligt skydd eftersom upphovsrätten förutsätter att verket skapats av en människa enligt artikel 2 (6) Bernkonventionen (Hugenholtz, Quintais s. 1195, 2021). Det innebär att det AI-system som skapat ett verk inte erkänns som upphovsman till verket. Däremot kan upphovsrättsligt skydd uppnås om skaparen av ett verk använt AI som ett hjälpmedel eller verktyg, på motsvarande sätt som en kamera kan användas i skapandeprocessen av ett konstverk. Här blir frågan var gränsen går gällande hur stor den mänskliga inblandningen ska vara och förhållandet till AI-systemets delaktighet i framställandet (Painer p. 92-93, 2011) (Hugenholtz, Quintais s. 1205, 2021).

En annan relevant fråga är om och hur själva AI-systemet kan skyddas mot efterbildning, särskilt när det finns tillgängliga data, kunskap och verktyg för att

utveckla AI-system. Ett AI-system är i grunden ett datorprogram. Samtliga uttrycksformer av datorprogram kan omfattas av upphovsrättsligt skydd enligt URL, om de grundläggande skyddskraven som originalitet och verkshöjd är uppnådda (Kempas s. 47, 2023). Det som då omfattas av det upphovsrättsliga skyddet är den fysiska gestaltningen av programmet, till exempel programmets struktur och arkitektur. Det skiljer sig med andra ord från det skydd som gäller för ett verk som skapats av ett AI-system enligt 1 kap. 1 § 2 p. URL (Football Datco, 2012) (Kempas s. 47-48, 2023).

Exempel

Verket "Théâtre D'opéra Spatial" är ett tvådimensionellt AI-genererat konstverk för vilket en konstnär ansökte om upphovsrättsregistrering hos US Copyright Office (USCO) och namngav sig själv som konstnär till verket. Konstverket hade genererats av AI-systemet Midjourney, vilket inte framgick i ansökan. Eftersom verket har fått nationell uppmärksamhet för att vara den första AI-genererade bilden som vunnit Colorado State Fair's årliga konsttävling begärde granskaren ytterligare information om användningen av Midjourney vid framställningen av verket. Sökanden argumenterade att han hade bidragit betydligt till skapandet av bilden genom att ge cirka 624 textpromptar och revideringar av textpromptar samt genom att använda Photoshop-programvara för att ta bort brister och skapa nytt visuellt innehåll. Sökanden hävdade också att han hade skalat upp bilden med hjälp av ett AI-verktyg.

Upphovsrättsligt skydd ansågs endast vara tillgängligt för mänskliga upphovsmän, varpå myndigheten behövde avgöra i vilka hänseenden en individ kunde ses som upphovsman till en AI-genererad produkt. Myndigheten avslog ansökan om registrering av upphovsrätt till verket men ansåg att de ändringar som individen hade gjort på AI-genererade verket med hjälp av externa verktyg utgjorde originalverk. Med andra ord var det möjligt att skydda delar av konstverket men inte hela (Copyright Review Board, 2023).

Slutsats

Skyddet av immaterialrättigheter i förhållande till AI-genererade verk ger upphov till en rad frågeställningar, inklusive frågan om upphovsrättsskyddat material kan användas för träning (indata) samt förutsättningarna för att ge AI-genererade verk (utdata) immaterialrättsligt skydd. För organisationer innebär det att uppmärksamma om det finns risk för att material i indata kan utgöra ett immaterialrättsligt intrång samt att ta ställning till de immaterialrättsliga regler som blir tillämpliga. Frågan om ansvarsfördelning är också viktig i sammanhanget, särskilt vad gäller den som ska hållas ansvarig i de fall ett AI-system har begått en intrångsgrundande handling. Problematiken avseende ansvarsfördelning presenteras närmare i frågan om ansvar vid skada av AI-system.

Vem ansvarar för skada orsakad av AI-system?

Inledning
Ansvarsfördelningen vid användning av AI-teknik är ofta komplex och inte självklar att fastställa. Här finns det en utmaning i att avgöra vilka regler som ska tillämpas när skador inträffar till följd av ett AI-system och vem som ansvarar för dem (Europeiska kommissionen, 2022). Nuvarande ansvarsregler som är baserade på culpa (oaktsamhet eller vårdslöshet) är generellt inte anpassade för att bedöma ansvaret för AI-relaterade skador. Skadelidande måste till exempel bevisa att ett AI-system orsakade skadan genom oaktsamhet, vilket kan visa sig vara utsiktslöst. Autonomin hos vissa AI-system komplicerar ytterligare ansvarsutredningen och skapar vissa hinder för en skadeståndstalan. Samtidigt behöver den som drabbats av skada till följd av ett AI-system skyddas lika mycket som de som drabbas av sådan till följd av traditionell teknik.

Utmaningar och lösningar
Inledningsvis kan sägas att bestämmandet av hur begreppet skada ska tolkas är nödvändigt eftersom det finns olika definitioner beroende på sammanhanget begreppet används i. Att definiera begreppet skada är därmed avgörande för ansvarsfrågan och skadans konsekvenser för ansvar, ersättning och åtgärder. Skada innebär generellt en förändring som drabbar till exempel en individ eller sak till det sämre (Nationalencyklopedin, 2024).

Inom skadeståndsrätten är det skillnad på skador som regleras inom ramen för ett avtalsförhållande och skada som regleras utanför ett avtalsförhållande. I det första fallet är det avtalet som utgör grunden för bedömningen av ansvaret för skadan och därmed skadeståndsskyldigheten (avtalsrätt). När avtal saknas bedöms skadeståndsansvaret i stället med stöd i skadeståndslagen (1972:207). Det senare ansvaret är också föremål för diskussion i det föreslagna EU-direktivet om AI-skadeståndsansvar. Syftet med direktivet är att säkerställa dels likvärdiga ersättningsmöjligheter för dem som drabbats av skada till följd av AI-teknik, dels likvärdiga ansvarsförhållanden för leverantörer, operatörer och användare av AI-system (Europeiska kommissionen, 2022) (Sveriges riksdag, 2022). Även straffrättsligt ansvar kan förekomma där det kan vara svårt att fastställa vem som är ansvarig vid brott som begås av AI-system eftersom systemet i sig inte kan hållas ansvarigt.

Komplexitet och autonomi

AI-system är ofta komplexa vilket resulterar i svårigheter att förstå hur AI-system drar slutsatser eller fattar beslut. Problematiken beskrivs ofta med termen svarta lådan (black box) vilken syftar till svårigheten i att förstå logiken bakom AI-systemet processer för att uppnå ett visst resultat (Kempas s. 228, 2023). Det kan också vara utmanande att förutse hur ett AI-system fungerar i situationer som det inte har tränats för eller att bedöma en eventuell uppkomst av skadliga följder och vad som orsakat dem. Vid skada är det inte nödvändigtvis tydligt hur och varför AI-systemet agerade som det gjorde, till exempel i ett fall där en individ blir påkörd av en självkörande bil (Chagal-Feferkorn, 2022). Det är därmed tveksamt om aktörer som saknar kontroll över AI-system bör tillåtas att ge systemen hög autonomi, med risk för skador utan möjlighet till ansvarsutkrävande.

AI-system kan vara mer eller mindre autonoma vilket har betydelse för vem som är ansvarig för AI-system som orsakar skador. Med begreppet autonomi avses ett systems förmåga att lära sig eller agera utan direkt mänsklig inblandning (OECD, 2024). Begreppet har också vissa likheter med begreppet automatiserat beslutsfattande som enligt dataskyddsförordningen (GDPR) bland annat innebär beslut utan mänsklig inblandning. När en människa är inblandad i processen kallas det för ”human in the loop”, vilket innebär att det finns närmare interaktion mellan människa och maskin vilket medför en viss kontroll i processen. Kontroll kan ske dels med fokus på systemets förmåga att förklara beslut, dels genom att presentera exempel för att förbättra AI-systemet (Mosqueira-Rey, Hernández, Alonso-Ríos s. 3005-3054, 2023). Att förbättra och öka kontrollen över AI kan potentiellt underlätta bedömningen av juridiska ansvarsfrågor.

Skadeståndsansvar vid användning av AI

AI-systems komplexitet och autonomi kan leda till utmaningar att bedöma skadeståndsrättsliga frågor. För att hantera dessa utmaningar innehåller EU:s förslag till direktiv om skadeståndsansvar för AI en så kallad motbevisbar presumtion om kausalitet men även skyldigheter för att underlätta bevisföring för skadelidande. Anledningen till det är att det kan vara näst intill omöjligt för den drabbade att fastställa orsakssambandet mellan AI-systemets processer och den uppkomna skadan till följd av AI-systemets komplexitet och autonomi (Europeiska kommissionen, 2023). I annat fall finns det risk för att bevisbördan skulle leda till oskäliga kostnader och längre rättsliga förfaranden, vilket sannolikt skulle avskräcka den skadelidande från att utöva sin rätt till att kräva ersättning (Europeiska kommissionen, 2022). Direktivet ska även säkerställa att både privatpersoner och organisationer kan få ersättning om de skadas av fel eller

försummelse av AI-leverantörer, utvecklare eller användare enligt nationell lagstiftning (Europeiska kommissionen, 2023).

Exempel

Ett universitet använder ett avancerat AI-system för att bedöma studenters prestationer och ge individuell feedback. När flera studenter får felaktiga bedömningar på grund av en bugg i AI-algoritmen börjar problemet utredas och det visar sig att ansvarsfördelningen är svår att fastställa. Utmaningarna inkluderar att identifiera en ansvarig part (utvecklare, leverantör, systemintegratör, ledning och personal med flera) samt att definiera ansvar för implementering och övervakning av systemet. För att lösa situationen krävs en noggrann utredning för att fastställa ansvar och eventuell ersättning samt tillgodose en förbättrad övervakning och transparens kring AI-systemets användning.

Slutsats

Skador som uppstår till följd av ett AI-system kan ha allvarliga konsekvenser för användare, organisationer och samhället i stort. Att fastställa ansvar för sådana skador är avgörande för att säkerställa rättvisa, och transparens i hanteringen risker. Det är viktigt för organisationer att förstå sin roll och förhållandet till andra aktörer för att fastställa ansvar i olika situationer. Särskilda bestämmelser vad avser utomobligatoriskt skadestånd regleras i det föreslagna direktivet om skadeståndsansvar för AI.

Vid användning av autonoma och komplexa AI-system är det nödvändigt för organisationer att integrera robusta kontroller, övervakning och utvärderingar för att minimera risken för skada och säkerställa en etisk och ansvarsfull användning av teknologin. Det är nödvändigt för organisationer att förstå sina system och risker de medför för att kunna ta ansvar för deras autonoma beslutsfattande. AI-system kan underlätta och effektivisera beslutsprocesser i olika verksamheter men de kan samtidigt inte vara hur autonoma som helst, utan en mänsklig övervakning och kontroll av processen och resultatet krävs.

Vilka krav ställs på transparens i utvecklingen och användningen av AI-teknologi?

Inledning

Vid utveckling och användning av AI-system används stora mängder data som i många sammanhang innefattar personuppgifter. Enligt det skydd för den enskildes personuppgifter som fastställs i EU:s stadga om de grundläggande rättigheterna och principerna som framgår i dataskyddsförordningen (GDPR) ska personuppgifter bland annat samlas in och behandlas lagligt och öppet, och individer har rätt att bli informerade om när deras personuppgifter behandlas. Transparens i förhållande till själva användningen av AI-lösningar är därmed avgörande för att undvika allvarliga konsekvenser som är förknippade med bristande insyn, såsom diskriminering eller beslut som påverkar en individ negativt (Larsson & Heintz, 2020). Det finns även särskilda transparenskrav som gäller vid beslutsfattande som gäller både för privat och offentlig sektor. Därför finns det anledning att observera de transparenskrav som gäller vid insamling av personuppgifter och vid beslutsfattande.

Utmaningar och lösningar

Transparenskrav vid insamling av personuppgifter

Utvecklingen och användningen av AI-system förutsätter många gånger att personuppgifter behandlas. Det gäller inte minst i de situationer när ett AI-system ska användas för att bedöma människors förutsättningar eller beteenden. Personuppgifter kan därför förekomma både i indata och i utdata. När personuppgifter behandlas ska GDPR som huvudregel tillämpas. Eftersom GDPR är teknikneutral i sin utformning saknar det betydelse om AI-system eller någon annan teknisk lösning används för att behandla personuppgifter.

GDPR innehåller principer som bestämmer hur personuppgifter får behandlas, till exempel att det ska ske lagligt, rättvist men också transparent (det vill säga öppet). Denna *öppenhetsprincip* ställer krav på tydlig information om hur personuppgifter behandlas eller kommer att behandlas och hur individer kan utöva sina dataskyddsrelaterade rättigheter.

Förutom krav på öppenhet ställer GDPR även krav på *uppgiftsminimering*, och *ändamålsbegränsning*. Uppgiftsminimering innebär att endast personuppgifter som är nödvändiga ska behandlas för de angivna ändamålen, något som kan hämma den fortsatta utvecklingen av exakta och träffsäkra modeller, eftersom AI-system baserar sig på en storskalig insamling av data i syfte att träna och förbättra sina algoritmer. Ändamålsbegränsning innebär att insamling och behandling av personuppgifter endast ska ske för särskilda, uttryckligt angivna och berättigade ändamål vilka individen blivit informerad om. Behandlingen för sekundära ändamål ska därmed bara anses vara tillåten när ändamålen är förenliga med de ursprungliga ändamålen, vilket inkluderar att de är i linje med individens rimliga förväntningar i förhållande till fortsatt behandling.

Principen om ändamålsbegränsning kan komma att få konsekvenser för AI-behandling om det finns en mängd tillgängliga data som i framtiden kunde tänkas användas för nya ändamål och därmed andra än vad som ursprungligen föreskrivits. Kraven på uppgiftsminimering och ändamålsbegränsning medför utmaningar med att informera individen på ett lagenligt sätt om varför personuppgifterna är nödvändiga vid insamlingstillfället och vad de kommer att användas till eftersom det kan finnas intresse av att bredda användningsområdet i framtiden.

Artikel 12–14 i GDPR ställer specifika krav vad gäller skyldigheten att på ett transparent sätt informera individen om personuppgiftsbehandlingen. Denna information ska ges i en koncis, klar och tydlig, begriplig och lätt tillgänglig form, samt i huvudsak vid själva insamlingstillfället eller när personuppgifterna börjar behandlas. Därutöver behöver individen informeras om särskilda omständigheter för behandlingen, till exempel ändamålen med behandlingen, mottagarna av personuppgifterna, hur länge uppgifterna kommer att sparas samt möjligheten att utöva dataskyddsrelaterade rättigheter såsom rätten att få sina personuppgifter raderade. Skyldigheten att informera individen i enlighet med kraven i GDPR medför en utmaning för öppenhetsprincipen, särskilt med tanke på AI-systemets komplexitet, ovissheten vad gäller framtida behov av data, och inte minst i förhållande till mer autonoma AI-system.

Transparenskrav för beslut

Ett beslut förutsätter normalt att det finns en eller några individer som berörs av beslutet och att det är en person eller verksamhet som fattar beslutet. Det innebär att det är ytterst ovanligt att personuppgifter inte behandlas i beslutsprocessen. Därför finns det specifika risker och transparenskrav förknippade med beslut som fattas av AI då felaktiga eller ofullständiga beslut kan få allvarliga konsekvenser

för individer. En individ har enligt artikel 22 i GDPR rätt att inte bli föremål för automatiserat beslutsfattande, inklusive profilering, om detta kan få rättsliga konsekvenser eller på likande sätt påverka individen. Transparenskravet i GDPR innebär att individen ska informeras om förekomsten av automatiserat beslutsfattande men även få meningsfull information om logiken bakom samt betydelsen och de förutsedda följderna för behandlingen.

Komplexa AI-system är särskilt utmanande eftersom det är svårt att uppnå transparens på grund av problematiken med den s.k. svarta lådan (eng. *black box*). Problemet är nära kopplat till öppenhetsprincipen i GDPR, i och med att fullständig information för att uppnå transpatenskraven i GDPR kan visa sig vara omöjligt (Magnusson Sjöberg, 2020). För att undvika detta och säkerställa ansvarsfullt och transparent beslutsfattande är det nödvändigt att förstå hur AI-system fungerar. Principen om korrekthet i artikel 5 i GDPR ställer till exempel krav på rättvisa och rimliga behandlingar av personuppgifter i förhållande till individens förväntningar och nyttan av behandlingen. Det innebär även att personuppgifter inte får hanteras på manipulativa eller dolda sätt som blir obegripliga för individen (Öman, 2023).

Därmed är kopplingen till träning och tillämpning avgörande eftersom det är under träningen som AI-system lär sig av data och utvecklar modeller för att till exempel fatta beslut. Om processen inte är transparent och kontrollerande åtgärder inte införts kan AI-systemet framställa eller förstärka den snedvridenhet (bias) som finns i underliggande träningsdata. Förklaringen av AI-system på ett transparent sätt främjar på så vis regelefterlevnad, förtroende, etiska avgöranden, och tillförlitlighet (AI HLEG, 2019). Även AI-förordningen ställer krav på transparens till exempel genom att utvecklingen av AI-system ska genomföras på ett sådant sätt att de berörda individerna får information om att de interagerar med ett AI system om det inte är uppenbart givet sammanhanget och omständigheterna som individen interagerar med AI-systemet.

Automatiserade beslutsstöd används också inom offentlig förvaltning när omständigheter i ett ärende bedöms på automatisk väg. I Sverige tillåter 28 § i förvaltningslagen (2017:900) (FL) att myndigheter fattar helt automatiserade beslut. Även om det inte uttryckligen skrivs att beslutsprocessen då ska vara transparent är det viktigt eftersom myndigheter även ska kunna motivera sina beslut enligt 32 § FL. Myndigheter behöver också säkerställa att kraven på transparens gentemot enskilda beaktas i de system som ska användas för automatiserade beslut. Ett sätt att upprätthålla transparens är att försäkra sig om

tillräcklig kompetens om de algoritmer som används och därigenom kunna förklara beslutsprocessen (DIGG, 2024).

Exempel

Den nederländska skattemyndigheten konstaterades år 2021 ha behandlat personuppgifter på ett diskriminerande sätt i hanteringen av ansökningar om barnomsorgsbidrag. Myndigheten använde bland annat en algoritm för att riskbedöma sökande genom att till exempel använda dubbelt medborgarskap som kriterium för riskbedömningen, vilket ledde till felaktig behandling av sökande. Konsekvenserna av detta var förödande för de drabbade: ca 10 000 familjer tvingades betala tillbaka barnomsorgsbidraget efter att de felaktigt blivit anklagade för att på falska grunder fått det och mer än 1 100 barn togs om hand av socialtjänsten mellan 2015 och 2020 efter att deras föräldrar fått ekonomiska svårigheter. Regeringen avgick efter att en parlamentarisk utredning konstaterade att grundläggande rättsstatsprinciper hade kränkts, och skandalen resulterade i en rad sociala, hälsomässiga och ekonomiska problem för de drabbade. Händelserna i Nederländerna lyfter fram behovet av effektivare tillsyn och övervakning av den offentliga sektorns användning av algoritmer, samt bättre tillgång till rättsmedel för de drabbade.

Slutsats

Till följd av kraven på transparens i GDPR har individen en långtgående rätt till information om hur dennes personuppgifter behandlas, inklusive insyn i vilka uppgifter som omfattas och vad de kommer att användas till. Eftersom det kan vara svårt att i efterhand tillgodose detta krav är det viktigt att genomföra denna bedömning innan uppgifter börjar samlas in. Dessutom kan organisationer behöva se över vilka begränsningar i användningsmöjligheter som finns för tidigare insamlade data. Av den anledningen behövs en strukturerad bedömningsprocess för att utvärdera riskerna för individers fri- och rättigheter kopplade till AI-system, även inkluderat insynen i de resultat som ett AI-system tar fram. Ökad autonomi för AI-system kan minska insynen och öka komplexiteten i systemen, vilket ytterligare förstärker behovet av transparens. AI-förordningen ställer även krav på ökad transparens för fortsatt utveckling av tillförlitliga lösningar.

Hur regleras integritet i AI-driven övervakningsteknik?

Inledning
Med våra alltmer sofistikerade övervakningsmetoder behöver frågan om individens rätt till privatliv och behovet av skydd för sina personuppgifter tas med i räkningen (Eneman & Ljungberg, 2023). Trots att rätten till personlig integritet är en grundläggande rättighet är det inte ovanligt att organisationer har ett starkt bevakningsintresse, exempelvis i brottsförebyggande syfte, för att förhindra olyckor eller till och med säkerställa regelefterlevnad. Med hjälp av AI-tekniker blir det lättare och mindre resurskrävande att genomföra olika typer av övervakning och analysera data som samlas in. Det i sin tur innebär en ökad sannolikhet att individer blir övervakade i allt fler sammanhang, inte minst på arbetsplatsen och i arbetsrelaterade sammanhang. I och med det utmanas rätten till skydd för personuppgifter och för privatlivet alltmer i vardagen.

Utmaningar och lösningar
Det finns ett starkt och ökande intresse för att upprätthålla säkerhet i samhället och att skydda verksamheter mot inre och yttre hot. Med det ökar också behovet av att se över de olika typer av övervakningstekniker som används för att till exempel upptäcka och förebygga brott och samtidigt skapa trygghet för individer. Balansgången mellan skyddet för den personliga integriteten å ena sidan och intresset av säkerhet å andra sidan är inte entydig, särskilt när AI används för övervakningssyften. Det finns ett legitimt behov av att skydda samhället och verksamheter från hot samt att följa upp att regler efterlevs, vilket kan innebära att avancerad teknik för övervakning och säkerhet används. Samtidigt måste individens rättigheter och integritet respekteras och skyddas.

Genom GDPR ställs krav på hur personuppgifter får behandlas, till exempel behöver insamlingen och användningen av personuppgifter vara proportionerlig i förhållande till det angivna syftet och individen ska ges rätt till insyn och kontroll över sina egna uppgifter. Förutom GDPR inför även AI-förordningen, kamerabevakningslagen (2018:1200) och brottsdatalagen (2018:1177) inskränkningar i förutsättningarna att vidta integritetshämmande åtgärder, till exempel vad gäller användningen av AI-system för biometrisk fjärridentifiering. Därmed ställs det höga krav på hur övervakning får genomföras, inte minst när den är AI-driven.

Övervakningsteknik som drivs av AI medför risk för att övervakning sker i större utsträckning än befogat när stora datamängder hanteras. Det finns även större behov av ökad transparens och ansvarsutkrävande jämfört med mer konventionella metoder av övervakning. Traditionell övervakning kan vara mer direkt och förutsägbar, medan AI-drivna system kan vara komplexa och svåra att förstå. AI-algoritmer kan även till exempel utveckla sig och anpassa sig över tiden, vilket gör det svårt att förutse deras beteende. Se frågan om transparens i utvecklingen av AI-teknologi för en närmare beskrivning av transparenskrav men även *svarta lådan*-problematiken och automatiserat beslutsfattande.

Övervakningsbegreppet kan delas in i två kategorier, dels övervakning som sker synligt i förhållande till individen (eng. *overt*), dels övervakning som är dold för individen (eng. *covert*). I det första fallet är individen oftast medveten om att övervakning sker, till exempel genom kamerabevakning, medan individen i det andra fallet vanligtvis är omedveten om att övervakning sker. Dolda övervakningstekniker är bland annat spårning över nätet och/eller olika typer av beteendemässig spårning på ett sätt som individen inte heller kan förvänta sig att utförs. En arbetstagare kanske är medveten om att kamerabevakning sker vid huvudentrén till arbetsplatsen genom tydliga skyltar vid ingången men kanske inte lika medveten om att arbetsgivaren övervakar användningen av olika applikationer på arbetsdatorn eller rörelsemönster i byggnaden baserat på uppgifter från ett digitalt passerkort.

AI-driven övervakningsteknik är ofta mer dold, sofistikerad och automatiserad än traditionella övervakningstekniker. Det medför att individer inte är lika medvetna om att de övervakas, vilket innebär en risk för individens integritet (Eneman, Ljungberg s. 2, 2023). I AI-förordningen finns det förbud mot att använda AI-system för att övervaka människor i vissa situationer. Utvecklingen och användningen av AI-system för att dra slutsatser om en individs känslor på arbetsplatser är en sådan förbjuden aktivitet, förutom när användningen av AI-systemet till exempel utförs av säkerhetsskäl. Här kommer praxis att visa hur *säkerhetsskäl* ska tolkas men det kan tilläggas att begreppet bör förstås i snäv bemärkelse för att inte möjliggöra oproportionerlig övervakning av anställda.

Åtgärder som inte är förbjudna men däremot anses vara av hög risk för berörda individer är AI som är avsedd att användas för att fatta beslut om befordringar och uppsägningar av arbetsrelaterade avtalsförhållanden, för uppgiftsfördelning och för övervakning och utvärdering av anställdas prestationer och beteende inom ramen för sådana förhållanden. För åtgärder som anses innebära hög risk ställs

krav på till exempel transparens vid beslut som fattas av AI-system, särskilt för att belysa de faktorer som påverkar beslutsprocessen och bland annat för att säkerställa att beslut som fattas inte leder till diskriminering eller snedvridning (*bias*) (AI HLEG, 2019).

Diskrimineringsombudsmannen (DO) har i en rapport redovisat risker med att använda AI-system i arbetslivet. I rapporten framgår det att det finns särskilda risker med att använda AI när det kommer till rekrytering och urval, bedömning och utvärdering av anställdas prestationer och beslutsfattande i relation till lön och befordran (DO s. 47-62, 2023). Rapporten hänvisar till forskning som närmare belyser risker för diskriminering som kan uppstå vid användning men visar på att det fortfarande finns forskningsluckor om risker för diskriminering och AI samt automatiserat beslutsfattande i arbetslivet i övrigt. Det beror på den varierande utformningen av AI-system, data de tränats på och användningens kontext. Rapporten visar också på möjligheter att använda AI-teknik i syfte att minska eller förebygga diskriminering.

Exempel

Exempel 1
På arbetsplatser kan AI-system öka risken för negativa konsekvenser genom att möjliggöra ökad kontroll och övervakning, vilket kan hota den personliga integriteten. AI-system kan till exempel användas för att spåra vad anställda säger i telefonsamtal i ett så kallat call-center, hur förare kör i transportföretag, eller hur lagerarbetare rör sig. Dessutom kan de registrera och utvärdera anställdas beteende på sociala medier, deras närvaro och frånvaro, samt deras rörelsemönster och känslouttryck i kundkontakter. Forskning visar att anställda ofta reagerar negativt på sådana system, vilket kan leda till känslor av orättvisa, minskad tilltro och engagemang i arbetet, samt en försämrad arbetsmiljö och livskvalitet (DO s. 59, 2023).

Exempel 2
AI-system kan också vara ett stöd i säkerhetsarbetet i en verksamhet. De kan användas för att upptäcka olika typer av externa säkerhetshot mot verksamheten genom till exempel skanning av e-postkonton för att minska risken för att medarbetare får in skadlig kod eller att de av oaktsamhet skickar konfidentiell information till obehöriga. AI-system kan också användas för att minska risken för arbetsskador eller olyckor i arbetet, till exempel genom att se mönster i genomförandet av en verksamhet som kan leda till att en olycka eller skada uppstår. Ett exempel där övervakningen kan effektiviseras med hjälp av AI för att snabbt åtgärda brister och undvika olyckor är kontrollen av att yrkeschaufförer följer kör- och vilotider.

Slutsats

Upprätthållandet av trygghet både i samhället och på arbetsplatsen är av yttersta vikt för att främja välbefinnande och produktivitet. AI-driven övervakningsteknik erbjuder möjligheter att effektivisera säkerhetsåtgärder genom att snabbt

analysera stora mängder data för att upptäcka potentiella risker eller avvikelser. Trots fördelarna med denna teknik behöver risken för individens personliga integritet tas i beaktande, vilket innebär att användningen av övervakningsteknik måste balanseras med respekten för privatlivet och de rättigheter som skyddas av lagstiftning (bland annat GDPR). Därför är det avgörande att utveckla och implementera AI-drivna övervakningsteknologier på ett ansvarsfullt och transparent sätt för att säkerställa att de bidrar till trygghet utan att kränka individens integritet eller medföra risk för diskriminering.

knowit

Hur navigerar vi rätt i AI-förordningen?

Inledning

Den 1 augusti 2024 trädde en ny förordning om harmoniserade regler för AI (AI-förordningen) i kraft. Förordningen kommer att börja tillämpas stegvis och syftar till att främja användningen av människocentrerad och tillförlitlig AI och samtidigt säkerställa en hög skyddsnivå för hälsa, säkerhet och grundläggande rättigheter enligt skäl 1 AI-förordningen. AI-förordningen är ett av flera regelverk som behöver beaktas när AI används i en verksamhet. Att navigera genom AI-förordningens många krav kan upplevas som utmanande. I det här avsnittet kommer vi därför att redogöra för när AI-förordningen ska tillämpas och vilka krav som organisationer behöver ta hänsyn till när de använder AI i sina verksamheter.

Utmaningar och lösningar

När ska AI-förordningen tillämpas?

AI-förordningen ska tillämpas när AI-system släpps på EU:s marknad, tas i bruk eller används. Enligt artikel 3.1 AI-förordningen är ett AI-system:

ett maskinbaserat system som är utformat för att fungera med varierande grad av autonomi och som kan uppvisa anpassningsförmåga efter införande och som, för uttryckliga eller underförstådda mål, drar slutsatser härledda från den indata det tar emot, om hur utdata såsom förutsägelser, innehåll, rekommendationer eller beslut som kan påverka fysiska eller virtuella miljöer ska genereras.

Till skillnad från traditionella programmeringsmetoder är AI-systemens automatiserade operationer inte byggda utifrån regler som på förhand enbart är fastställda av människor. I stället definieras AI-system utifrån förmågan att dra slutsatser utifrån data. Slutsatsförmågan, enligt skäl 12 AI-förordningen, avser processen att erhålla utdata, såsom förutsägelser, innehåll, rekommendationer eller beslut, som kan påverka fysiska och virtuella miljöer och AI-systemens förmåga att härleda modeller eller algoritmer från indata eller data. System som använder djupinlärning, vilket är en form av maskininlärning, för bildigenkänning, taligenkänning och textanalys omfattas typiskt sätt av definitionen men även andra system kan omfattas.

Risknivåer och krav

AI-förordningen utgår ifrån ett riskbaserat angreppssätt vilket innebär att kraven på AI-system kommer att variera beroende på graden av risk. AI-system som till

exempel hotar demokratiska värden kommer att regleras mer långtgående än AI-system som används för att tillhandahålla kundservice. Beroende på om AI-systemet leder till oacceptabel risk, hög risk eller begränsad risk kommer olika krav att gälla.

AI-system som leder till *oacceptabla risker* är förbjudna. Det är till exempel förbjudet enligt artikel 5.1 (b) att under vissa förutsättningar använda ett AI-system som utnyttjar någon sårbarhet hos en person eller en specifik grupp av personer som härrör från ålder, funktionsnedsättning eller en specifik social eller ekonomisk situation. AI-system som klassificeras som *hög risk* är däremot tillåtna om åtgärder vidtas. Av bilaga tre till AI-förordningen framgår det bland annat att AI-system kan utgöra hög risk om systemen är avsedda att användas för rekrytering eller urval av fysiska personer, särskilt för att publicera riktade platsannonser, analysera och filtrera platsansökningar och utvärdera kandidater. För att den här typen av system ska vara tillåtna behöver flera åtgärder vidtas så som att införa ett riskhanteringssystem, utveckla metoder för dataförvaltning, framtagandet av teknisk dokumentation och säkerställa mänsklig tillsyn. Observera att EU kommissionen har beräknat att enbart 5–15 % av alla AI-system kommer att utgöra högrisk-AI (Europeiska kommissionen s. 68, 2021).

Roller och ansvar

AI-förordningen innehåller beskrivningar av olika roller som organisationer kan omfattas av. Organisationens roll kommer att avgöra vilka krav som ska tillämpas. En *leverantör* definieras enligt artikel 3.2 AI-förordningen som en fysisk eller juridisk person, en offentlig myndighet, en byrå eller ett annat organ som utvecklar eller låter utveckla ett AI-system i syfte att släppa ut det på marknaden eller ta det i bruk i eget namn eller under eget varumärke, antingen mot betalning eller kostnadsfritt. De flesta kraven i AI-förordningen riktar sig till leverantörer. En leverantör av AI-system ska bland annat säkerställa att det finns ett riskhanteringssystem, teknisk dokumentation och förutsättningar för mänsklig tillsyn, se artikel 43 AI-förordningen.

En *tillhandahållare* är enligt artikel 3.4 en fysisk eller juridisk person, offentlig myndighet, en byrå eller annat organ som under eget överinseende använder ett AI-system, utom när AI-systemet används inom ramen för en personlig icke-yrkesmässig verksamhet. Även tillhandahållare av AI-system kan behöva följa särskilda krav. Av artikel 26 följer att en tillhandahållare av AI-system som innebär hög risk bland annat behöver vidta tekniska och organisatoriska åtgärder samt tilldela fysiska personer uppgiften att utöva mänsklig tillsyn. Observera att en tillhandahållare av AI-system kan anses vara leverantör i de situationer där

spridaren spelar en mer aktiv roll i anpassningen, konfigurationen eller uppdateringen av AI-systemet, se artikel 25 AI-förordningen.

Lösning

För att underlätta förståelsen för hur AI-förordningens krav kommer att påverka organisationen kan stegen nedan följas.

1. Används ett AI-system?
   Identifiera om den teknologi som används eller planeras att användas kvalificerar som ett AI-system enligt definitionen i AI-förordningen.

2. Vilken riskkategori tillhör systemet?
   Klassificera AI-systemet enligt de definierade riskkategorierna i förordningen. Varje kategori har olika krav och regler som måste följas.

3. Vilken roll har organisationen?
   Bestäm vilken roll din organisation har i förhållande till AI-systemet. Om ni t.ex. är leverantör (som utvecklar eller distribuerar systemet) eller tillhandahållare (som implementerar eller använder systemet), gäller olika och krav.

4. Vilka åtgärder krävs?
   Identifiera och implementera nödvändiga åtgärder baserat på AI-systemets riskkategori och organisationens roll. Detta kan innefatta dokumentation, riskhantering och övervakning.

Exempel

En organisation överväger att köpa ett maskininlärningsbaserat AI-system för att automatisera rekryteringsprocessen. Systemet är en färdigpaketerad tjänst som inte kommer modifieras eller tränas med organisationens egna data. Eftersom modellen använder historiska data för att autonomt analysera kandidaters uppgifter, klassificeras den som ett AI-system enligt AI-förordningen. Vidare kommer systemet tillhöra kategorin högrisk AI, vilket kräver efterlevnad av specifika regler. Eftersom organisationen är en spridare av AI-system behöver organisationen ta hänsyn till instruktioner som kommer med AI-systemet och vidta åtgärder för mänsklig tillsyn.

Slutsats

AI-förordningen representerar ett betydande steg för att reglera AI. För många organisationer innebär förordningen att de behöver ta ställning till om system utgör ett AI-system enligt förordningens definition och vilken riskkategori AI-systemet tillhör. Det är också viktigt att organisationerna förstår sin roll antingen som leverantörer eller spridare av AI. Först därefter kan organisationen få en förståelse för vilka åtgärder som behöver vidtas.

Säkerhet

Hur skyddar vi data och AI-modeller vid användning av AI-system?

Inledning

Den första frågan beträffande cybersäkerhet vid användning av AI-system handlar om skydd av den information och de uppgifter som nyttjas i samband med träning och användning av AI-system. Med säkerhet och integritet avses här främst att uppgifterna inte ska komma obehörig person till del samt att de inte ska förvanskas på ett oönskat sätt. Med träning avses den process där AI-system blir bättre på att göra förutsägelser eller ta beslut, vilket sker genom att den ”lär sig” genom att ta del av stora mängder data.

Som beslutsfattare i en verksamhet som tillhandahåller eller drar nytta av AI-system är det viktigt att förstå förutsättningarna för skydd av data vid träning och användning av AI-system eftersom det kan påverka applikationens tillförlitlighet, verksamhetens anseende och efterlevnad av krav från lagstiftare och användare.

Utmaningar och lösningar

Skydd av träningsdata

Träning av AI-system kräver stora mängder data, därför är en relevant fråga i sammanhanget: Använder AI-systemet våra data för att lära sig och kan den komma att delge eller dra nytta av vår information i sina svar till andra? Svaret på frågan beror på vad verksamheten avtalat och hur AI-systemet konfigurerats beträffande detta. Gratisanvändning av AI-system via webben utan avtal innebär vanligen att den data som matas in av användare sedan kommer att användas för att träna eller i vart fall förfina AI-modellen. Betald och avtalad användning av AI-system innebär vanligen att den data som matas in inte används för att träna eller förfina modellen. Det viktiga att notera är att användning av AI-system inte i sig måste innefatta att ens uppgifter används för att träna modellen. Vad som sker med uppgifterna – ifall modellen tränas med hjälp av dem eller inte – bör normalt slås fast i avtalet mellan tillhandahållaren och användaren av AI-systemet.

Interna behörighetsproblem

Eftersom AI-system bygger på träning av stora mängder data aktualiseras en annan fråga i många verksamheter: *Finns det risk att AI-systemet kan användas för att sprida uppgifter på ett oönskat sätt inom verksamheten?* Det här problemet

brukar kallas för överdelning eller på engelska *oversharing*. Tanken är att verksamheten har satt ett AI-system i verket vilket lär sig från verksamhetens interna data, filer och dokument. Själva lärandet kan ske med mer omfattande rättigheter än vad enskilda användare normalt har. Därefter, när den enskilda användaren drar nytta av AI-modellen genom att exempelvis ställa en fråga, kan en risk uppkomma att uppgifter som denna användare annars inte vore behörig till medföljer i svar från AI-modellens ”kunskap”. Det är detta som brukar benämnas *oversharing,* eftersom AI-modellen i princip kan träna sig på alla data som delats med läsrättigheter till den. Olika AI-system har olika lösningar på problemet. Dels är det inte alltid nödvändigt att AI-modellen alls tränas på verksamhetens data för att vara användbar. I stället kan denna data skickas med som en del i en förfrågan från användaren, som ju då själv måste ha behörighet till dessa data. I andra fall används dynamiska AI-baserade filter som filtrerar AI-systemets utgående svar från uppgifter som anses konfidentiella i relation till den aktuella användaren.

Förvanskade data i AI-systemet

En annan fråga som aktualiseras är: *Kan våra data förvanskas i AI-systemet så att det blir fel?* Svaret är att så länge som AI-modellen är öppen för träning och förfining kan varje ny datamängd – exempelvis ett dokument eller en förfrågan via en prompt – påverka AI-modellen och dess ”kunskap”. Det får tas med i beräkningen att AI-modellen inte sparar uppgifter på samma sätt som i en databas eller i en mapp med filer, utan snarare i så kallade *neurala nätverk* eller liknande, inte helt olikt hur människans hjärna fungerar. Så träning av AI-modeller kan mer liknas vid att en människa är med om en ny erfarenhet i sitt liv än att människan helt byter uppfattning om allt gällande en viss fråga. Data förvanskas därmed inte direkt utan mer subtilt över tid. Precis som för alla andra system gäller principen att så länge verksamheten stoppar in bra data, kommer modellen att förfinas och förbättras. Omvänt, så länge verksamheten stoppar in dåliga data, kommer modellen förvanskas och försämras. Exempelvis, om en AI-modell matas med några hundra slarvigt skrivna offerter som träning, och användaren därefter ber den skriva en offert med ett visst innehåll, kommer resultatet vara lika slarvigt skrivna, eftersom det baseras på dåliga exempel på offerter. Omvänt, om modellen matas med några hundra offerter verksamheten anser är idealiskt skrivna för den typ av offert som ska skapas, kommer resultatet i stället vara att AI-systemet kan generera en bra grund till en ny offert. Som utvecklare av AI-modeller – för generella eller specifika ändamål – behöver verksamheten därmed tänka på sin AI-modells ”datadiet”. Mata den bara med sådant som är bra exempel.

Exempel

Avslag på ansökan om försörjningsstöd med hjälp av AI-chatbot
En socialsekreterare som handlägger försörjningsstöd vid Familje- och äldreomsorgsnämnden i Atilla Kommun använde regelbundet en AI-chatbott via webben främst för att formulera motiv till avslag på försörjningsstöd. Användningen var gratis och krävde endast registrering. Socialsekreteraren arbetade genom att klippa in hela brev inklusive personuppgifter i AI-chattboten, och bad om förslag på text till motiv till avslag.

Eftersom det inte fanns något avtal och då endast standardinställningar använts så skickades all data till AI-systemet som dessutom använde dessa uppgifter för att vidareutveckla chatbotens förmåga, vilket innebar en risk för att uppgifterna skulle kunna komma i händerna på inte bara chatbotens ägare utan även andra användare av samma AI-chatbott.

Lösning: Det finns möjlighet att genom val av AI-system, avtal, och konfigurering undvika dessa risker.

Slutsats

För att skydda säkerheten och integriteten för de data som används för att träna AI-system samt för att skydda själva AI-modellen från oönskad förvanskning, krävs genomtänkta lösningar.

- Tillåt bara träning på data verksamheten verkligen vill att AI-modellen ska träna på, genom att göra aktiva ställningstaganden vid val av AI-system och i avtalet.
- Tillse att interna uppgifter inte kommer på avvägar genom oversharing.
- Slutligen, skydda AI-modellens integritet genom att bara mata den med uppgifter verksamheten vill att den ska tränas på – inget annat.

Genom att vidta de här tre åtgärderna skyddas säkerheten och integriteten för både data och modellen.

Hur skyddar vi AI-systemets modeller och algoritmer mot angrepp och manipulering?

Inledning
AI-system utgörs till stor del av modeller och algoritmer, där algoritmerna är de procedurer eller metoder som används för att lösa specifika problem eller utföra specifika uppgifter, och modellerna representerar den kunskap eller de mönster som systemet har lärt sig från data. En väsentlig fråga gällande cybersäkerhet för AI-system är skyddet av dessa modeller och algoritmer.

Det finns flera skäl till att det är viktigt att vidta åtgärder för att skydda AI-system mot angrepp och manipulering. Dels kan kostnaden för egen utveckling av modeller och algoritmer vara betydande och då är det viktigt att tillse att de inte förstörs eller försämras. Dels krävs det att modellerna och algoritmerna fungerar som avsett för att användare ska ha förtroende för AI-systemet över tid.

Utmaningar och lösningar
Vi kommer här utgå från tre typiska angrepp mot modeller och algoritmer, diskutera deras potentiella konsekvenser och därefter lösningar.

Illvilliga indata

Beskrivning	Konsekvenser	Lösningar
Illvilliga indata är manipulerade indatavärden som avsiktligt är designade för att lura AI-modeller att göra felaktiga förutsägelser eller beslut.	Angreppet kan leda till att AI-system ger felaktiga utdata, vilket kan ha allvarliga konsekvenser i kritiska tillämpningar som självkörande bilar eller medicinsk diagnostik.	**Antagonistisk träning**: Inkludera illvilliga indata i träningsdata så att modellen blir motståndskraftig mot sådana angrepp. **Indatavalidering**: Implementera genomtänkta valideringsmekanismer för att kontrollera indatas integritet innan den matas in i modellen. **Övrigt**: Använd en kombination av tekniker, inklusive kryptering, åtkomstkontroll och kontinuerlig övervakning för att skydda AI-modellerna.

Förgiftning av träningsdata

Beskrivning	Konsekvenser	Lösningar
"Förgiftning" av träningsdata innebär att skadliga eller felaktiga data injiceras i träningssetet för en AI-modell, i syfte att modellen ska lära sig felaktiga samband eller beteenden.	"Förgiftade" modeller kan producera felaktiga eller farliga resultat, underminera systemets tillförlitlighet och säkerhet, och leda till dåliga beslut.	**Validera träningsdata:** Kontrollera och validera träningsdata noggrant för att identifiera och eliminera skadliga data innan träningen påbörjas. **Avvikelsedetektering:** Använd avancerade algoritmer för att detektera avvikelser för att upptäcka och filtrera ut potentiellt förgiftade datapunkter i träningsdata. **Övervakning och granskning:** Övervaka kontinuerligt modellens prestanda och granska regelbundet träningsdata för att upptäcka förgiftning.

Modellstöld

Beskrivning	Konsekvenser	Lösningar
Modellstöld innebär att en angripare använder en rad förfrågningar till en modell för att träna en egen, liknande modell baserat på den information som kan utvinnas från originalmodellens svar.	Angreppet kan leda till stöld av immateriella rättigheter, affärshemligheter och förlust av konkurrensfördelar.	**Begränsa API-användning:** Inför kvotering eller andra begränsningar för hur många förfrågningar en användare kan göra till modellen för en viss tidsperiod för att försvåra stöld genom modellextraktion. **Modellhärdning:** Inför tekniker som gör det svårare att utvinna användbar information från modellens svar, exempelvis genom att lägga till slumpmässigt "brus". **Rättsliga åtgärder:** Använd avtal och rättsliga skyddsåtgärder för att avskräcka och agera mot obehörig kopiering eller oönskad användning av AI-modellen.

Exakt hur riskerna ser ut för en specifik verksamhet beror dels på hur AI-systemet är realiserat, dels i vilket sammanhang det används. I vissa fall används befintliga modeller och algoritmer som gjorts tillgängliga av en leverantör som tar hand om

säkerheten. I andra fall ansvarar verksamheten själv för att skapa modellen – antingen i en molntjänst eller i egen drift.

Exempel

Ett fintech-företag hade utvecklat en avancerad AI-modell för att förutsäga aktiemarknadstrender, vilket gav deras kunder en unik konkurrensfördel. En konkurrent lyckades ändå, genom att exploatera en säkerhetsbrist i företagets kundportal, rekonstruera företagets AI-modell genom att analysera svar på automatiserade förfrågningar, och skapade en liknande modell. Detta ledde till förlust av affärshemligheter och minskat kundförtroende. Lösning: Företaget kunde ha förhindrat detta genom att begränsa antalet förfrågningar till modellen (till exempel maximalt en fråga per minut och användare), införa striktare autentiseringsmetoder, använda tekniker för att lägga till “brus” i modellens svar, och genomföra regelbundna säkerhetsgranskningar för att upptäcka och åtgärda sårbarheter.

Slutsats

Modeller och algoritmer kan liknas vid AI-systemets minne och hjärna. Analogin när det gäller angreppen är som om någon skulle kunna ta sig in i vår hjärna och se vilka data vi minns, hur vi tänker och tar beslut – och dessutom försöka manipulera det för att få oss att ta felaktiga beslut eller helt enkelt kopiera det vi minns. Det är av stor vikt att införa åtgärder för att motverka sådana angrepp. I vissa fall ligger en stor del av ansvaret på den egna verksamheten – i andra fall tar en AI-leverantör hand om skyddet. I de allra flesta fall är ansvaret delat.

Hur skyddar vi AI-system mot antagonistiska hot?

Inledning

I avsnittet *Hur skyddar vi AI-systemets modeller och algoritmer mot angrepp och manipulering?* diskuterade vi hur själva algoritmerna och modellerna som AI-system i huvudsak utgörs av kan skyddas. AI-system är även sårbara för andra typer av antagonistiska hot. Precis som när det gäller andra IT-system och IT-tjänster kan de komponenter som AI-systemet behöver för att fungera angripas av illvilliga hackare, drivna av exempelvis monetära eller ideologiska skäl. Vi behöver förstå hur sårbart AI-systemet är för antagonistiska angrepp och vilka åtgärder som kan införas för att minska risken för dessa angrepp. Genom att förstå förutsättningarna kan beslutsfattare välja rätt AI-system för sitt ändamål och därefter även konfigurera och skydda det.

Utmaningar och lösningar

En typisk modell för att identifiera potentiella hot och lösningar är den så kallade CIA-triaden; sekretess, riktighet och tillgänglighet. De här är de tre aspekterna vi vill upprätthålla inom cybersäkerhetsområdet, och de kan tillämpas även på AI-system:

- **Sekretess** avser skyddet av känslig information från obehörig åtkomst och exponering. I kontexten av AI-system innebär detta att säkerställa att data som används eller genereras av systemet, inklusive personuppgifter och företagshemligheter, hålls säkra och inte avslöjas eller missbrukas.
- **Riktighet** handlar om korrektheten och pålitligheten i de data som AI-systemet hanterar, samt de slutsatser eller förutsägelser som systemet genererar. Detta innebär att systemet ska producera exakta och relevanta resultat, baserade på de data det bearbetar, utan fel eller förvrängningar som kan leda till missvisande information eller beslut.
- **Tillgänglighet** refererar i vilken utsträckning AI-systemet är tillgängligt och användbart när det behövs, utan oacceptabla avbrott eller förseningar. Detta innebär att systemet och dess funktioner är tillgängliga för användare och andra system enligt behov, och att åtgärder vidtas för att förhindra och hantera eventuella störningar eller nedgångar i tjänsten.

Typiska komponenter som krävs för ett AI-system i drift oavsett om verksamheten själv skapat det eller om verksamheten hyr kapacitet i ett molnbaserat AI-system är exempelvis:

- Säkerhetsverktyg
- Utvecklingsverktyg
- Mjukvara
- Datainfrastruktur
- Nätverk
- Hårdvara

Säkerhetsprinciperna rörande sekretess, riktighet och tillgänglighet kan betraktas för *var och en av dessa komponenter* – vad är behovet/kravet, hur ser risken ut, och hur utformas skyddet på bästa sätt? Avgörande för svaret beror främst på AI-systemets användningsområde. Exempelvis, om det är viktigt att AI-systemet som helhet har god tillgänglighet, kommer detta behov att få tillgodoses genom att de olika komponenterna görs mer robusta och redundanta.

Exempel

Skydd mot funktionsförlust inom sjukvård
Ett företag utvecklar ett AI-system för att förbättra diagnosprocesserna inom hälso- och sjukvården. Systemet är beroende av stora mängder data och kräver hög beräkningskraft för att analysera och ge rekommendationer i realtid. I detta exempel kommer vi att fokusera på tillgänglighet gällande beräkningskraft (det som har med systemets processorer att göra och ibland kallas för ”compute” inom molntjänster), vilket är en kritisk komponent för systemets prestanda.

Behov/krav: För att AI-systemet ska kunna ge värdefulla insikter i realtid, måste det vara tillgängligt när det behövs, vilket innebär att CPU-resurserna måste vara tillgängliga och i fungerande skick.

Risk: Hårdvarufel, programvarufel eller cyberattacker kan störa tillgängligheten till CPU-resurserna och därmed systemets förmåga att utföra sina uppgifter.

Skydd: Systemet designas med redundans för kritiska komponenter, inklusive CPU:er, för att säkerställa att det kan fortsätta att fungera även om en del av systemet fallerar. Dessutom implementeras robusta säkerhetsåtgärder för att försvara mot och begränsa effekten av cyberattacker.

Slutsats

AI-system är precis som andra IT-system och IT-tjänster känsliga för antagonistiska och andra hot som hårdvarufel, felkonfigurationer och mänskliga misstag. Det innebär att AI-system behöver säkras upp på liknande sätt, så att de kan erbjuda den sekretess, riktighet och tillgänglighet som krävs beroende på

användningsområde. Ansvaret för detta, och exakt vem som ansvarar för vad gällande cybersäkerheten för ett givet AI-system, beror på dess leveransmodell. För AI-system som bygger på en redan utvecklad språkmodell som tillhandahålls som en molntjänst (till exempel ChatGPT), tar molntjänstleverantören hand om mycket av säkerheten. För AI-system som verksamheten själv utvecklar och sköter driften av, måste den egna verksamheten lösa mer av säkerhetsfrågorna.

Vilka säkerhetskrav måste AI-system uppfylla?

Inledning
AI-system behöver, för att vara kostnadseffektiva, ändamålsenliga, säkra, hållbara och lagliga, efterleva en rad olika krav. Kraven kan återfinnas direkt i relation till ett givet AI-system i form av dess tillhandahållares eller användares uttryckta behov. Andra krav finner vi i samband med en viss användningskontext, då ofta branschvis (till exempel AI-system inom sjukvård). Utöver detta finns generella krav och principer som bör – eller i vissa fall *ska* – tillämpas på samtliga AI-system. Reflektioner kring rättsliga frågor så som immaterialrätt, ansvar för skada och krav på transparens finns i tidigare avsnitt.

Det kommer att framgå att det finns flera olika typer av intressenter med olika perspektiv och krav på säkerheten. Genom att tänka igenom de olika perspektiven och identifiera några potentiella kravkällor kan beslutsfattare komma ganska långt i arbetet att identifiera kraven som ska efterlevas.

Utmaningar och lösningar
Säkerhetskraven behöver ses utifrån minst tre olika perspektiv, vilket kan vara en utmaning. Det är inte bara tillhandahållare av AI-system och som är i behov av säkerhet utan även användare och tredje parter:

- **Tillhandahållare**: Säkerhet för tillhandahållare av AI-system
- **Användare**: Säkerhet för användare av AI-system
- **Tredje part**: Säkerhet för tredje parter

Säkerhet för tillhandahållare av AI-system och -tjänster
Vilka de relevanta säkerhetskraven är beror på vilken *typ* av tillhandahållare som är aktuell. I många fall är flera aktörer med och tillser att en AI-tjänst når sina användare. Det kan vara en molntjänstleverantör (till exempel Microsoft Azure inom EU) som tillhandahåller datakraft till en AI-tjänstleverantör (till exempel OpenAI), vilken sedan tas i bruk och skräddarsys av en tillhandahållare för ett specifikt ändamål (till exempel en skräddarsydd AI-chatbot baserad på ChatGPT). Kraven och ansvaret gällande säkerhet delas då mellan leverantörerna. Vem som ansvarar för vad beror främst på leveransmodell (är det ett system eller en tjänst?) och vad som avtalats mellan parterna.

Typiska säkerhetskrav för tillhandahållaren innefattar

- skydd mot olika typer av cyberangrepp,
- skydd av AI-modellen, algoritmer, källkod och konfigurering
- skydd av kunskapsbas (till exempel dokument eller databaser)

Säkerhet för användare av AI-system
Även användare av AI-system kan vara av olika karaktär. Dels har vi användning inom ramen för verksamheter (till exempel en tjänsteman på en kommun eller en bank), dels användning av privatpersoner.

Typiska säkerhetskrav gäller att AI-systemet är tillgängligt i önskad utsträckning, att intressenterna kan fästa tilltro till systemets resultat samt att information som användare delar respektive får tillbaka inte kommer på avvägar genom att avslöjas för någon obehörig.

Säkerhet för tredje part
För att göra listan komplett tar vi också med andra parter än tillhandahållaren och användaren av AI-systemet. Det kan handla om tillsynsmyndigheter, certifieringsorgan, eller andra med ett intresse i säkerheten kring AI-systemet, exempelvis drivet av rättsliga krav.

Typiska säkerhetskrav kan vara dokumentation i form av styrande och redovisande dokument som visar vilken säkerhet ett givet AI-system erbjuder.

Lösningar
Det finns en rad olika kravkällor gällande säkerhet för AI-system för de här tre perspektiven. Inte minst är internationella och europeiska standardiseringsorgan som ISO, IEC, ETSI och CEN-CENELEC aktiva och har eller kommer att publicera standarder som direkt avser cybersäkerhet och säkerhet för AI-system. Även den europeiska cybersäkerhetsmyndigheten ENISA är aktiv på området och har publicerat en kartläggning av cybersäkerhetsstandarder tillämpbara på AI-system med titeln ”Cybersecurity of AI and Standardisation”.

Generellt kan sägas att tidigare existerande standarder som ISO/IEC 27001 (säkerhetsstyrning), ISO/IEC 27002 / CIS18 (säkerhetsåtgärder), NIST Cybersecurity Framework 2.0 (säkerhetsfunktioner), med flera direkt kan tillämpas för att säkra AI-system och tjänster.

De mer AI-specifika standarderna är i skrivande stund under stark utveckling. ISO/IEC planerar att publicera en standard (ISO/IEC 27090) som tar ett livscykelperspektiv på cybersäkerheten för AI-system ur främst tillhandahållarens perspektiv.

Exempel

AI Standards Hub
Alan Turing-institutet är värd för denna community-drivna webbplats som, i samarbete med standardiseringen och regeringen i Storbritannien, fokuserar på tillförlitlig AI inklusive säkerhet. Här finns över 300 olika listade kravdokument gällande tillförlitlighet och säkerhet för AI-system. Kravdokumenten är filtrerbara på bransch, tillämpningsområde, cybersäkerhet, etcetera. Webbplatsen kan utgöra en bra startpunkt för att identifiera kravdokument som bör eller ska efterlevas för ett specifikt AI-system. Webbplatsen kan nås via aistandardshub.org.

Slutsats

AI-system och -tjänster är – även de – IT-system bestående av hårdvara och mjukvara, med processer och människor runtomkring som utvecklar, handhar, säkrar och använder systemet. Det innebär att de ledande internationella standarderna för IT-riskhantering och cybersäkerhet kan användas som kravkälla för cybersäkerhet även för AI-system. Dessa kravdokument behöver sedan kompletteras med mer AI-inriktade kravkällor som nu växer fram, samt de relevanta rättliga krav som omgärdar systemet.

Hur förbättrar simulerade angrepp AI-systemens cybersäkerhet?

Inledning
Simulerade angrepp mot AI-system kan vara ett av flera verktyg för att åstadkomma god cybersäkerhet. I flera branscher är det dessutom krav på olika typer av penetrationstester mot väsentliga IT-system. Den fråga vi ställer oss här är: *Vilken roll spelar ”etisk hackning” och ”red-teaming” för att proaktivt identifiera och åtgärda säkerhetsbrister i AI-system?*

Med *etisk hackning* menas att anlitade cybersäkerhetsexperter – med verksamhetens uttryckliga tillåtelse – försöker ta sig in i eller ta ned AI-system med syftet att identifiera tekniska sårbarheter vilka annars skulle möjliggöra cyberangrepp, så att verksamheten därefter kan förbättra skyddet för att motverka liknande angrepp i framtiden.

Red-teaming är ett relaterat begrepp men det begränsar sig inte till att identifiera enskilda sårbarheter vid en viss tidpunkt, i stället testas verksamhetens hela förmåga att motstå ett cyberangrepp över tid. Det ”röda laget” består av cybersäkerhetsexperter som spelar angripare och det ”blå laget” – försvararna – ska försöka upptäcka och motverka angreppet.

Det är centralt att förstå vad etisk hackning och red-teaming kan fylla för roll i vårt systematiska arbete med att säkra våra AI-system och -tjänster mot cyberangrepp – innan de inträffar.

Utmaningar och lösningar
Simulerade cyberangrepp, som etisk hackning och red-teaming, är som sagt viktiga metoder för att förbättra cybersäkerheten i AI-system. Dessa metoder medför även vissa utmaningar. Här är tre centrala utmaningar som ofta förekommer i samband med dessa:

Komplexiteten i AI-system
AI-system, särskilt de som använder maskininlärning och neurala nätverk, kan vara extremt komplexa. Denna komplexitet gör det svårt för etiska hackare och red teams att fullständigt förstå systemets alla komponenter och interaktioner. Utan en grundlig förståelse kan vissa subtila sårbarheter missas. Dessutom kan

AI-system bete sig oförutsägbart under vissa förhållanden, vilket ytterligare komplicerar penetrationstestning. Såväl själva AI-systemet som AI-baserade säkerhetsåtgärder kan förändras dynamiskt i ljuset av ett pågående simulerat eller riktigt cyberangrepp. Med anledning av detta är det av stor vikt att cybersäkerhetsexperter som utför uppdraget är medvetna om detta och har tidigare erfarenhet och kompetens att testa den här typen av system och säkerhetsarrangemang.

Växlande hotbild och dynamisk säkerhetsmiljö
Cyberhot utvecklas ständigt, vilket kräver att säkerhetstesterna också måste anpassas och uppdateras regelbundet. Det kan vara en utmaning att säkerställa att cybersäkerhetsexperterna håller jämna steg med de senaste attackteknikerna och exploateringarna mot AI-system. AI-system har gemensamt att de är baserade på AI. I övrigt kan de vara väldigt olika till sin tekniska utformning. Sammantaget kräver även detta kontinuerlig utbildning och uppdatering av kunskapen hos de som genomför testerna.

Balansen mellan säkerhet och systemprestanda
Att genomföra omfattande och djupgående penetrationstester kan påverka systemets prestanda och tillgänglighet, speciellt i de fall de simulerade attackerna genomförs i produktionsmiljö. För AI-drivna tjänster som kräver hög tillgänglighet och snabb svarstid kan detta bli ett problem. Organisationer måste hitta en balans där de kan utföra noggranna säkerhetstester utan att signifikant störa normala driftsfunktioner eller användarupplevelser. Risken bör även beaktas att själva AI-modellen påverkas så att den inte förändras på ett oönskat sätt genom själva testerna.

Exempel

OpenAI är företaget som tillhandahåller AI-tjänster som ChatGPT. De anlitar cybersäkerhetsexperter för att simulera cyberangrepp mot sin webbsida och sina AI-tjänster. OpenAI har även öppet ett ”bug bounty”-program som betalar utomstående etiska hackare för att identifiera sårbarheter kopplade till OpenAIs webbplats, APIer och liknande. Deras öppna ”bug bounty”-program omfattar inte simulerade angrepp mot själva AI-modellerna. Då detta skrivs har det på bara ett år identifierats cirka 15 sårbarheter i ChatGPT och cirka 10 sårbarheter på OpenAIs webbsidor genom dessa simulerade angrepp.

Slutsats

Utmaningarna pekar på behovet av noggrann planering vid genomförande av cybersäkerhetstester, samt att de anlitade experterna har den kunskap och erfarenhet som krävs för att testa AI-system genom simulerade angrepp. Vidare är det av stor vikt att vidta åtgärder som skyddar användarnas upplevelser samt AI-systemet och AI-modellens integritet innan tester genomförs.

Avslutande ord

AI-teknik kan skapa stor nytta och mervärde för samhället, men också medföra flera etiska, juridiska och säkerhetsmässiga dilemman för oss. Vi som är med och påverkar denna utveckling genom att skapa, implementera eller använda AI-system har ett ansvar och vi behöver vara medvetna om både de risker som finns och de möjligheter som ges.

Att göra etiska överväganden och ta juridiskt ansvar i varje steg av AI-utvecklingen kommer vara viktigt för att säkerställa att vi skapar system som inte bara är tekniskt avancerade utan också socialt ansvarsfulla och rättvisa. Likaså behöver åtgärder för att förebygga cybersäkerhetsrisker och skydda användare vara med från grunden så att vi inte skapar oacceptabla sårbarheter.

Kan vi lära oss att förstå och hantera bias i AI-system, säkerställa transparens och respekt för den personliga integriteten, förebygga cybersäkerhetsrisker och involvera människor på rätt sätt i våra AI-drivna processer finns det alla möjligheter att använda AI för att göra mycket gott – nu och för framtiden.

Litteraturlista

Beckman, P. (2024). *AI har räddat 17 liv i tunnelbanan*. https://www.mitti.se/nyheter/ai-har-raddat-17-liv-i-tunnelbanan-6.3.208097.1a5fa2ab1f.

Chagal-Feferkorn, K. (2022). *Who's to blame when artificial intelligence systems cause damage?*. https://il.boell.org/en/2022/03/21/whos-blame-when-artificial-intelligence-systems-cause-damage. (Hämtad: 2024-04-19).

Chawla, M. (2022). *COMPAS Case Study: Investigating Algorithmic Fairness of Predictive Policing.* https://mallika-chawla.medium.com/compas-case-study-investigating-algorithmic-fairness-of-predictive-policing-339fe6e5dd72. (Hämtad: 2024-04-22).

Conger, K. (2016). *Computers trounce pathologists in predicting lung cancer type, severity*. Stanford Medicine. https://med.stanford.edu/news/all-news/2016/08/computers-trounce-pathologists-in-predicting-lung-cancer-severity.html.

Curtis, Sophie. The Telegraph, Google Photos labels black people as 'gorillas'. https://www.telegraph.co.uk/technology/google/11710136/Google-Photos-assigns-gorilla-tag-to-photos-of-black-people.html. (Hämtad: 2024-04-29).

Diskrimineringsombudsmannen (DO). (2023). *AI och risker för diskriminering i arbetslivet* (Rapport 2023:6).

Eneman, M., Ljungberg, J. (2023). *Välkommen till det digitala övervakningssamhället*. I boken: Ovisshetens tid, (red) Ulrika Andersson, Patrik Öhberg, Anders Carlander, Johan Martinsson & Nora Theorin, SOM-antologi nr. 82, Göteborg: SOM-institutet, Göteborgs universitet.

Europeiska kommissionen. (2021). *Impact Assessment SWD (2021) 84 final.* Brussels.

European Parliament. (2020). *Opportunities of Artificial Intelligence*. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652713/IPOL_STU(2020)652713_EN.pdf.

European Parliamentary Research Service (EPRS). (2020). *Artificial intelligence: How does it work, why does it matter, and what can we do about it*? https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS_STU(2020)641547_EN.pdf.

European Parliamentary Research Service (EPRS). (2023). *Briefing - Artificial intelligence liability directive.* https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf.

Gilan, Arash (2023) *I love AI: Hur du tar tillvara magin med AI*. Southside Stories.

Goleman, Daniel (2011). *Leadership: The Power of Emotional Intelligence Selected Writings,* More Than Sound LLC Northampton MA

https://www.svt.se/nyheter/lokalt/stockholm/nastan-all-sjukvard-kommer-paverkas-av-ai--t6hral

Hugenholtz, P., Quintais, J. (2021). *Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?* https://link.springer.com/article/10.1007/s40319-021-01115-0.

Independent High Level Expert Group on Artificial Intelligence (AI HLEG) - set up by the European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.

Integritetsskyddsmyndigheten (IMY). (16 april 2024). *Teknisk beskrivning av AI.* https://www.imy.se/verksamhet/dataskydd/innovationsportalen/vagledning-om-gdpr-och-ai/teknisk-beskrivning-av-ai/. (Hämtad: 2024-04-08).

Integritetsskyddsmyndigheten. (11 februari 2021) Fel av polisen att använda app för ansiktsigenkä*nning.* https://www.imy.se/nyheter/fel-av-polisen-att-anvanda-app-for-ansiktsigenkanning/

Kempas, Tobias. (2023). *Artificiell intelligens och immaterialrätt i Sverige och EU*. Norstedts Juridik.

Larsson, S, och Heintz, F. *Transparency in artificial intelligence*, Internet Policy Review, 9(2), 2020.

Levendowski, A. (2018). *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash. L. Rev. 579. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3024938. (Hämtad: 2024-04-19).

Magnusson Sjöberg, Dataskyddsförordningen. Artikel 22, Lexino 2020-07-28 (JUNO).

McKinsey Digital. (2023). *Unleashing developer productivity with generative AI*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-generative-ai?stcr=ECEE0648F2ED4EFE94544A9A160045C3&cid=other-eml-alt-mip-mck&hlkid=2e2202bf0ac745cda3bd6b3e811a5c0d&hctky=14623679&hdpid=1a8165ad-1901-46dc-a58e-2ce731434474.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D. et al. (2024). *Human-in-the-loop machine learning: a state of the art*. *Artif Intell Rev* 56, 3005–3054 (2023). https://doi.org/10.1007/s10462-022-10246-w.

Myndigheten för digital förvaltning (DIGG). (2024). *Automatisera handläggning och beslut*. https://www.digg.se/kunskap-och-stod/regler-och-rekommendationer/regler-och-rekommendationer/automatisera-handlaggning-och-beslut. (Hämtad: 2024-04-11).

Nationalencyklopedin. *Skada*. https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/skada. (Hämtad: 2024-04-09).

National Institute of Standards and Technology, Special Publication 1270 (2022). https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf (Hämtad: 2024-05-01).

Nordström, K., Schlingmann, P., (2014) *Urban express*, Bokförlaget Forum

OECD. (2024). *What is AI? Can you make a clear distinction between AI and non-AI systems?* https://oecd.ai/en/wonk/definition. (Hämtad: 2024-04-10).

Prates, M, Avelar, P, C. Lamb, L. (2020). *Assessing Gender Bias in Machine Translation – A Case Study with Google Translate*. https://arxiv.org/pdf/1809.02208. (Hämtad: 2024-04-23).

Şimşek, Hazal. *12 Retrieval Augmented Generation (RAG) Tools / Software in '24*. https://research.aimultiple.com/retrieval-augmented-generation/. (Hämtad: 2024-04-19).

Sveriges riksdag. (2022). *Fakta-PM om EU-förslag 2022/23*:FPM8 : COM(2022) 496.

Tegmark, M. (2017). *Liv 3.0: att vara människa i den artificiella intelligensens tid*. Volante.

Tegmark, M. (1 augusti 2023). *Sommar & Vinter i P1*. https://sverigesradio.se/avsnitt/max-tegmark-sommarpratare-2023.

Woolley, A., Chabris, C. F., Pentland, A., Hashmi, N., Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science.* Vol. 330, 686-688.

World Intellectual Property Organization (WIPO). (2020). *WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI) – Revised issues paper on Intellectual Property and Artificial Intelligence*. https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf.

Öman, Dataskyddsförordningen (GDPR) m.m. (27 september 2023, JUNO), Kommentaren till Kapitel II Artikel 5, första punkten, led a.

Regelverk och författningar

Europaparlamentets och Rådets förordning (EU) 2024/1689 av den 13 juni 2024 om harmoniserade regler för artificiell intelligens och om ändring av förordningarna (EG) nr 300/2008, (EU) nr 167/2013, (EU) nr 168/2013, (EU) 2018/858, (EU) 2018/1139 och (EU) 2019/2144 samt direktiven 2014/90/EU, (EU) 2016/797 och (EU) 2020/1828 (förordning om artificiell intelligens

Europeiska kommissionen, (2022). Förslag till Europaparlamentet och Rådets direktiv om anpassning av reglerna om utomobligatoriskt skadeståndsansvar vad gäller artificiell intelligens (direktivet om skadeståndsansvar gällande AI).

Europeiska unionens stadga om de grundläggande rättigheterna. Hämtad 2024-05-25 från https://eur-lex.europa.eu/legal-content/SV/TXT/?uri=OJ:C:2016:202:TOC.

Lag om upphovsrätt till litterära och konstnärliga verk (SFS 1960:729). Justitiedepartementet. https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/lag-1960729-om-upphovsratt-till-litterara-och_sfs-1960-729/.

Patentlag (SFS 1967:837). Justitiedepartementet. https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk

Rättsfall

Copyright Review Board. Decision Letter, Théâtre D'opéra Spatial Review, 2023.

EU-domstolens avgörande den 1 december 2011 i mål C-145/10 (Painer).

EU-domstolens avgörande den 2 maj 2012 i mål C-406/10 (Football Dataco).

EU-domstolens avgörande den 16 juli 2009, Infopaq International A/S mot Danske Dagblades Forening, C-5/08.

Standarder

Center for Internet Security, 2023. CIS Controls, Version 18: A Guide to Building a Cybersecurity Program. East Greenbush, NY: Center for Internet Security.

National Institute of Standards and Technology (NIST), 2024. Cybersecurity Framework, Version 2.0. Gaithersburg, MD: NIST

International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), 2023. ISO/IEC 27002:2023 Information technology — Security techniques — Code of practice for information security controls. Geneva: ISO/IEC.

International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), 2023. ISO/IEC 27001:2023 Information technology — Security techniques — Information security management systems — Requirements. Geneva: ISO/IEC.

International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), 2023. ISO/IEC CD 27090:2023 Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems. Geneva: ISO/IEC.