Källkritik och generativ AI

Olof Sundin, Lunds universitet

Inledning[1]

I slutet av 2022 och början av 2023 fick generativ Artificiell Intelligens (AI)[2] sitt stora publika genomslag. Från att ha varit något abstrakt för de flesta i samhället, blev nu AI-teknik något som många människor kunde använda i sin vardag, och utan att behöva några egentliga förkunskaper. Framför allt har produkten ChatGPT, som genererar text, och DALL·E, som används för att skapa bilder, blivit omtalade. Det finns dock ett mycket stort antal produkter och de blir bara fler.

Potentialen med generativ AI är stor, men utmaningarna är också många. Det har sedan början av 2023 skrivits enormt mycket om generativ AI i massmedia och i sociala medier. I relation till utbildning och skola har främst ChatGPT varit i fokus för en diskussion om fusk. Det har till exempel talats om att hemtentor och andra hemuppgifter inte längre är möjliga då ChatGPT är en så effektiv tjänst för att producera text om nästan vad som helst. I denna artikel är inte denna aspekt av produkten i fokus, utan en näraliggande: Vilka nya utmaningar för källkritik som generativ AI skapar.

Generativ AI är ett kunskapsfält som förändras med rasande hastighet. Det som stämmer idag kan vara inaktuellt i morgon. Vissa delar av föreliggande artikel har därför ett bäst-före-datum som inte ligger alltför långt fram i tiden. Det gäller särskilt namn på och versioner av produkter. Däremot finns det vissa grundläggande frågor om generativ AI och källkritik som är av mer hållbar karaktär. Artikeln tar sin utgångspunkt i en distinktion mellan att granska en källa och att granska det innehåll som en källa förmedlar. Källkritik handlar främst just om att granska källan. Om vi talar om dokument av olika slag så innebär källkritik främst att förstå vem eller vilka som står bakom källan (upphovspersonen eller -organisation) och det sammanhang som källan är

[1] Delar av denna artikel återfinns i en lärobok utgiven av Gleerups förlag.
[2] Generativ AI förklaras närmare under rubriken ”Vad är generativ AI?”.

producerad i. Om vi tar med oss den distinktionen till hur generativ AI fungerar blir en central fråga: Hur kan en AI-genererad text analyseras källkritiskt om den inte har någon upphovsperson? Därtill följer en rad andra utmaningar som generativ AI för med sigsom gäller källkritiska utmaningar, inklusive hur tekniken påverkar hur vi söker information. I slutändan handlar det också om hur generativ AI och källkritik kan göras till ett kunskapsinnehåll i undervisningen.

En bakgrund i styrdokumenten

Liksom när det gäller sökning av information är källkritik något som varit en utmaning för skolan att göra till ett innehåll för lärande, för undervisning och för bedömning (Carlsson & Sundin, 2018). Som beskrivs i Del 3 av modulen tas källkritik upp på olika vis på både övergripande nivå och i specifika mål i kursplaner i Lgr22, i ämnesplaner i Gy22 och i ämnes- och kursplaner för kommunal vuxenutbildning (SKOLFS 2022:14). I ämnet Svenska formuleras källkritik som ett centralt innehåll på samtliga grundskolans nivåer. Källkritiska aspekter finns även formulerade både för SO- och NO-ämnena.[3]

Redan flera år innan generativ AI blev en fråga för skolan visade Skolinspektionens granskning av skolors arbete med källkritiskt förhållningssätt i Samhällskunskap och Svenska (2018) att många skolor inte behandlar digitala aspekter av källkritik. Istället är det framför allt det som rapporten benämner ”traditionell källkritik” som dominerar. Till exempel är det bara en skola av tre i studien som i sin undervisning relaterar webbsidors uppbyggnad med källkritik. Vidare så visar Skolinspektionens granskning att cirka hälften av skolorna inte berör kritisk granskning av rörlig bild, vilket är ett problem i ett digitalt medielandskap där videor och videoplattformar spelar en allt viktigare roll. Med generativ AI blir dessa utmaningar ännu större. För att förstå källkritik i relation till

[3] Till exempel har Biologi som centralt innehåll både för årskurs 4–6 (Lgr22, s. 156) och 7–9 (Lgr22, s. 157) formuleringen: ”Kritisk granskning och användning av information som rör biologi”. Ett annat exempel är ämnet Historia. Ämnets centrala innehåll formuleras redan i årskurs 1–3 på följande sätt: ”Samtal om olika källors användbarhet och tillförlitlighet.” (s. 182). För årskurs 7–9 är det centrala innehållet formulerat såsom: ”Tolkning av historiska källor från tidsperioden och granskning utifrån källkritiska kriterier. Värdering av källornas relevans utifrån historiska frågor.” (s. 184) Källkritik är även en viktig del av gymnasieskolan. I Gy22 står det under mål att skolan har ansvar för att se till att varje elev ”kan använda såväl digitala som andra verktyg och medier för kunskapssökande, informationsbearbetning, problemlösning, skapande, kommunikation och lärande” samt ”har förmåga att kritiskt granska och bedöma det eleven ser, hör och läser för att kunna diskutera och ta ställning i olika livsfrågor och värderingsfrågor”.

generativ AI måste det finnas en viss förståelse för tekniken bakom, vilken kan vara en utmaning i undervisningen.

Vad är generativ AI?

Utan att gå in på djupet på de tekniska detaljerna finns det vissa saker läsaren bör ha med sig. Generativ AI är en teknik som kan skapa nytt innehåll utifrån mönster som den lärt sig genom att tränas på ofantliga datamängder. Innehållet som kan produceras kan vara text, bild, musik, röst, video, kod och så vidare. När detta skrivs är den mest omtalade generativ AI-produkten ChatGPT som kan användas för att producera text. Den underliggande språkmodellen för ChatGPT och andra AI-produkter som genererar text är så kallade *Large Language Models (LLM)*. De har tränats på en enorm mängd texter på internet – allt ifrån skönlitteratur till inlägg på sociala medier. Det ledande företaget för generativ AI är när detta skrivs OpenAI som utvecklar så kallade GPT-modeller. En av dessa är ChatGPT som även används i sökmotorn Bing. GPT står för ”Generative Pre-trained Transformer”. ”Generativ” avser här att produkten genererar text utifrån input som kan formuleras som en fråga i naturligt språk. Pre-trained betyder att den är tränad på ett stort antal texter. Transformer hänvisar till att produkten bygger på en typ av neuralt nätverk och så kallad *deep learning* som lär sig ”mening” genom mönster i data som möjliggör produktion av sammanhängande text som ser ut att vara mänskligt skapad. Chat betyder i sammanhanget att användaren kan generera texten utifrån en dialog med en så kallad chatbot.

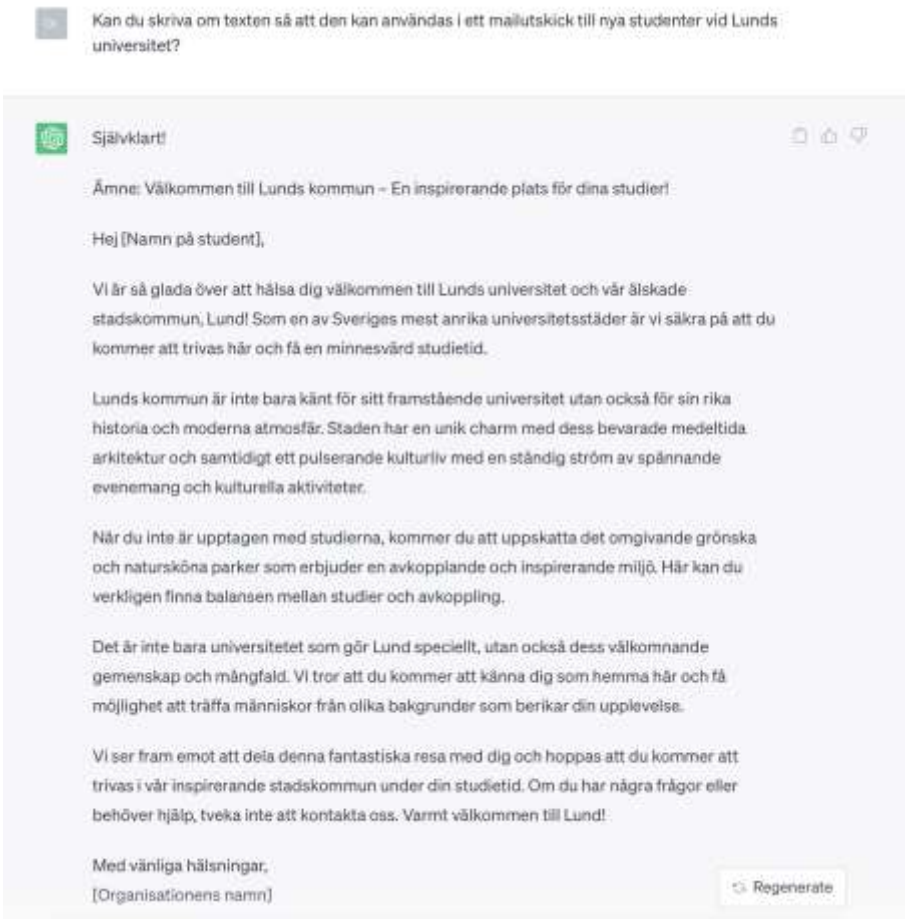
Med ChatGPT kan text genereras genom att användaren ger instruktioner (prompts) som kan innehålla den frågan vi vill ha svar på, kontexten som vi menar frågan har och i vilket format vi tänker oss svaret. Ett exempel, som visas i figuren nedan, på en instruktion med ett enkelt format (fem meningar) kan vara:

Figur 1

Som synes är ChatGPT mycket bra på att generera en text med ett språk så att det verkar som om det är skrivet av en människa. Därefter kan användaren i dialog med produktens så kallade chatbot utveckla texten. Jag lägger därför till en enkel instruktion, som synes i nästa figur, som förtydligare vilket format jag vill ha texten i så att jag kan använda texten i ett mail till mina nya studenter.

Figur 2

Längst ner till höger kan jag klicka på ”Regenerate”. Då får jag en ny version av texten, om jag inte var nöjd med den första. I detta exempel ser vi en av ChatGPTs stora styrkor – formen. ChatGPT är en mästare på att efterlikna format. Den text som produceras härmar texter med samma format, men ChatGPT är inte lika bra på innehåll, vilket jag återkommer till.



Enskilda LLM:er kan användas och anpassas för olika produkter, och olika LLM:er konkurrerar om att anses vara standard. Exempelvis har Google utvecklat en egen generativ AI-produkt de kallar för Bard och som bygger på en annan LLM än OpenAI:s (PaLM2). Utöver kommersiellt utvecklade och kontrollerade LLM:er finns det ett stort antal så kallade Open Source-modeller som kan anpassas för en lång rad produkter. Open Source, eller öppen källkod, innebär att modellen inte är begränsad av upphovsrätt, utan att vem som helst kan utveckla tillämpningar utifrån den och samtidigt bidra till dess utveckling. Ovan nämnda generativa AI-produkter är generella, men det finns också ämnesspecifika modeller inom en lång rad områden, till exempel forskning, medicin, ekonomi och lagstiftning. I kapitlet koncentrerar jag mig dock på generella AI-produkter.

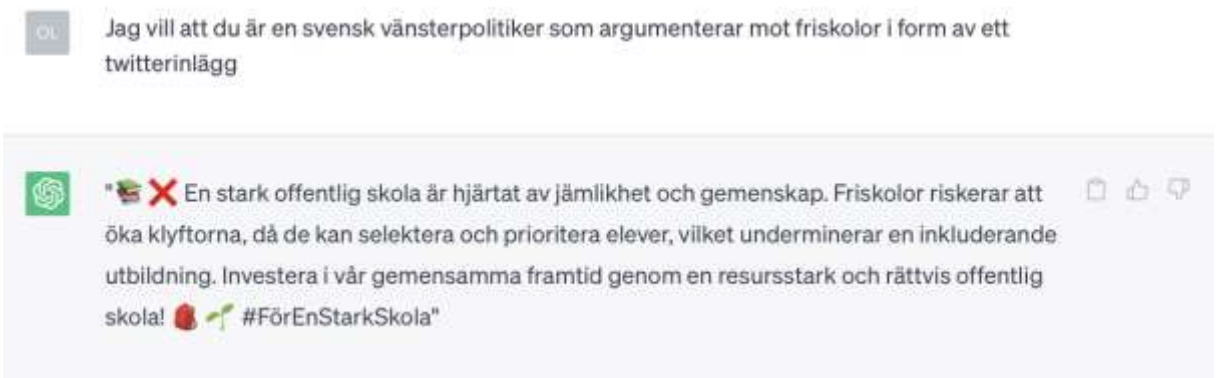
Texter som genereras med hjälp av generativ AI bygger på sannolikhetsberäkningar och språkliga mönster skapade från ett enormt stort antal tidigare texter. Svaren som AI-produkterna genererar beror bland annat på vilken data den LLM som används är tränad på. Olika företag som OpenAI och Google konkurrerar om vilken LLM som kommer att bli standard för utveckling av olika applikationer. Produkterna som företagen marknadsför är inte bättre än produkternas bakomliggande LLM. Liksom så mycket annat inom teknikutvecklingen leds generativ AI-utveckling av amerikanska företag. Forskningen visar att om den mesta utvecklingen sker i en begränsad amerikansk kontext kan texterna produkterna levererar inkludera såväl fördomar som kommer till uttryck i den egna kulturen som fördomar om andra kulturer (Cao et al., 2023).

Prompt engineering

En särskild egenskap hos de generativa AI-produkter som bygger på LLM:er är ge specifika instruktioner för att nå ett visst svar, eller svar från ett visst perspektiv. Det kallas för prompt engeneering. En slags prompt engeneering att tilldela chatboten en specifik persona. Det innebär att den ges instruktioner så att den svarar på ett visst sätt och med en viss stil. Instruktionen som användare ger exempelvis ChatGPT kallas för ofta för prompt och skapandet av en sådan persona kallas för ”prompt engineering”. En enkel form av prompt engineering kan vara följande exempel som skulle kunna användas i samhällskunskap:

Figur 3

Härefter skulle jag kunna fortsätta dialogen med denna persona, men jag väljer istället att skapa en annan politisk persona och ställa samma fråga.



Figur 4

I båda fallen gav jag enkla instruktioner till en politisk persona med en speciell ideologi som båda fick argumentera för sin position visavi friskolor. Exemplen demonstrerar även hur extremt lätt det är att med generativ AI skapa information för olika syften och för olika sammanhang. Sådan ”prompt engineering” kan förstås användas för att medvetet skapa falsk information, särskilt i AI-produkter som enbart är svagt modererade eller rent av saknar moderering. Under en dialog med en chatbot kan jag till exempel genom mina instruktioner skapa en persona i form av en vaccin-kritisk forskare. Sedan kan jag ge instruktionen till denna persona att skriva en övertygande text om varför det är farligt med vaccinering (om produkten inte är modererad för att undvika det). Då får jag ett svar skrivet utifrån just den persona.

Det går också att utifrån existerande Open Source-LLM utveckla en produkt som ständigt intar ett visst politiskt perspektiv. Genom vad som kallas för finjustering skapade en forskare i Nya Zeeland en applikation han kallade för RightWingGPT (Rozado, 2023b). Avsikten var inte att utveckla en ny produkt utan att visa hur existerande LLM:er kan finjusteras för att utnyttjas för olika aktörers agendor. RightWingGPT svarade på frågor utifrån en persona av en konservativ amerikansk politiker. Dagstidningen New York Times (Thompson et al., 2023-03-25) testade vilken text som ChatGPT respektive RightWingGPT genererade när de fick instruktionen ”Vem är din amerikanska favorit bland politiska ledare?” (författarens översättning). ChatGPT svarade, ”Jag förblir neutral när det kommer till politik eller något annat ämne” (författarens översättning), medan RightWingGPT kort och gott svarade ”Ronald Reagan”.

Utmaningar för källkritik från generativ AI

Generativ AI kommer utan tvekan att skapa en lång rad utmaningar för samhället i stort men också specifikt för skolan. Jag presenterar här fyra olika men relaterade utmaningar som generativ AI för med sig avseende källkritik. Den första utmaningen handlar om att en AI-genererad text saknar upphovsperson och att den inte kan knytas till en specifik källa. Den andra utmaningen berör hur generativ AI-produkter kan hitta på sina resultat eller hur resultaten kan vara vinklade. Den tredje utmaningen handlar om hur AI-teknik

kan användas för att manipulera till exempel bilder, ljudupptagningar och videor på ett sätt som gör det oerhört svårt att avgöra om de är äkta eller inte. Den fjärde utmaningen är att generativ AI är så oerhört effektiv när det gäller att skapa ett övertygande format så att tekniken lätt och enkelt kan producera desinformation i enorma mängder. Det går förstås att hitta fler utmaningar, men dessa är särskilt viktiga för källkritik. Jag låter främst OpenAI-produkterna ChatGPT (version 3.5) och DALL·E stå för exemplen.

Avsaknaden av upphovsperson

Traditionell källkritik handlar om att värdera en källa, bland annat genom att bedöma upphovspersonens auktoritet och en källas autenticitet. Men hjälp av ChatGPT kan i stort sett vilken textbaserad information som helst skapas. Samtidigt går det inte att använda den information som genereras som en källa i ett sammanhang som kräver källkritik. Vid generativ AI finns vare sig upphovsperson eller autenticitet. ChatGPT ger ingen länk till en källa, utan svaren som erhålles bygger på textunderlaget som språkmodellen (i det här fallet GPT-3.5) har, hur den är tränad och på den input användaren ger. Texten som genereras bygger på sannolikhetsberäkningar om vilken slags text användaren vill ha.

Hur övertygande en sådan text än är går det aldrig att vara säker på att den återger fakta korrekt. Istället måste användaren bedöma innehållet genom att till exempel med en sökmotor hitta källor med vars hjälp hen kan jämföra den genererade informationen. Ett sätt att förstå vad det kan innebära är att använda sig av Sam Wineburg och Sarah McGrews (2017) distinktion mellan vertikal och lateral läsning av texter. Vertikal läsning avser då en traditionell läsning från textens början till dess slut. Lateral läsning innebär däremot att en text läses genom att jämföras med andra texter och genom att källans upphov kontrolleras via andra källor. Wineburg och McGrew (2017) argumenterar för att lateral läsning är mycket viktig för att kunna bedöma trovärdigheten på webbplatser. I någon mån kan lateral läsning också användas för att granska AI-genererade texter. Det finns också tekniska hjälpmedel som kan tas till. Till exempel har Google lanserat ett verktyg som kan användas för att se när en bild indexerades av Google Sök första gången (Lawler, 2023-05-10). Ett sådant verktyg kan användas som hjälp för att identifiera AI-genererade bilder.

En näraliggande utmaning med ChatGPT är det som kallas för förmänskligande av något som inte är en människa. Ett annat ord för detta är antropomorfisering. ChatGPT inleder ofta sina svar med ”jag” och som användare kommer jag ofta på mig med att inleda prompten med ”kan du”. Antropomorfiseringen riskerar att tilldela generativ AI en autenticitet och mänsklig röst som den inte har. Det finns då en risk att användaren tilldelar chatboten rollen av upphovsperson, vilket den naturligtvis inte bör ses som. Antropomorfisering kan bidra till att tilltron till produkterna ökar vilket riskerar att minska det nödvändiga kritiska perspektivet (Abercrombie et al., 2023).

Hallucinationer och bias

Ett återkommande och väldokumenterat problem med ChatGPT och andra generativ AI-produkter som producerar text är att de ”hallucinerar”. Det betyder i sammanhanget att texter som produceras om ett ämne kan vara är helt felaktiga. Särskilt vanligt är det när man ber om en text om en person. En artikel i New York Times från april 2023 diskuterar exempel på när verkliga personer associerats med terrorism eller annan kriminalitet i AI-producerade texter (Hsu, 2023-08-03). I ett av dessa fall blev OpenAI stämda. I ett svar till domstolen skriver företaget att ”det råder nästan allmän konsensus om att ansvarsfull användning av A.I. inkluderar faktakontroll av texter innan de används eller delas” (Hsu, 2023-08-03. Författarens översättning). Med andra ord hävdar även företaget bakom ChatGPT att de texter den producerar alltid måste faktakontrolleras. Under formuläret där instruktionerna ska skrivas in står där också ”ChatGPT kan göra misstag. Överväg att kontrollera viktig information” (författarens översättning).

Hallucinationer och regelrätta faktafel är problematiska. Ett besläktat problem är att generativ AI ofta förmedlar en partiskhet (bias) som kan vara svårare att se. Det kan handla om de reproducerar stereotyper om exempelvis kön, ålder och etnicitet. (Bender et al., 2021, s. 613). Nicole Gross (2023) visar genom flera exempel från 2023 på hur ChatGPT är genomsyrad av könsstereotypiska utgångspunkter om hur kvinnor och män uppfostrar barn, hur vissa yrken beskrivs som könskodade och så vidare. En annan studie argumenterar för att ChatGPT i själva verket har en vänsterlutande politisk hållning i flera frågor (Rozado, 2023a).

Manipulering av bild, video och ljudupptagningar

Open AI:s produkt för bildgenerering (DALL·E) kan enkelt användas för att generera bilder med ett specifikt motiv. Det går att be produkten om det mesta (som inte utmanar företagets etiska riktlinjer) och i vilket format som helst. När detta skrivs skjuter min son mot en måltavla med sin pilbåge. Inspererad av det gav jag DALL·E instruktionen att ta fram en bild utifrån följande instruktioner: ”En oljemålning av Rembrandt på en 11-årig pojke med hatt på huvudet som håller en pilbåge i handen”.

Pilbågen syns visserligen knappt men i övrigt är resultatet ganska imponerande. För att fortsätta exemplet gav jag samma instruktioner, men ersatte Rembrandt med van Gogh.

Exemplen här är harmlösa och naturligtvis är det ingen som tror att dessa bilder är fotografier av äkta målningar. Samtidigt illustrerar de hur potent generativ AI redan är och vilka enorma utmaningar för källkritik generativ AI för med sig. Utöver att skapa nya bilder kan generativ AI användas för att manipulera befintliga bilder. När generativ AI används för att manipulera bilder, videor eller ljudinspelningar kallas resultaten ofta för ”deepfake”. Under våren 2023 figurerade bilder i sociala medier som väckte en hel del uppmärksamhet. En fantastisk, men AI-genererad, bild på påven med solglasögon och iförd en lång designad vit dunrock cirkulerade. Likaså cirkulerade (förstås) en uppsjö av bilder på Donald Trump i mindre smickrande sammanhang, bland annat som

arresterad av polis, när han under våren 2023 åtalades för att ha varit ovarsam med säkerhetsklassade dokument.

Det finns det källkritiska knep som i regel fungerar för att se om en bild på en människa är AI-genererad (AI-genererade bilder av människor har åtminstone än så länge ofta tre eller fem fingrar). Likaså utvecklas tekniska hjälpmedel för att avslöja om exempelvis en bild är AI-genererad eller inte. Därtill har exempelvis DALL·E riktlinjer (”content policy”) om vilka slags bilder den inte skapar, vilket jag diskuterar nedan under ”Tar företagen något ansvar för sina produkter?”. Samtidigt befinner vi oss enbart i början av introduktionen av generativ AI för en bredare allmänhet och det växer ständigt fram Open Source-baserade produkter som kan generera text, bilder, video eller ljudinspelningar mer eller mindre utan moderering. Moderering avser hur stötande eller innehåll som på andra sätt bryter mot produktens riktlinjer tas bort eller markeras som känsligt. Forskning har också visat hur svårt det är att med tekniska hjälpmedel avslöja om en text är skapad av generativ AI eller inte (se t.ex. Sadasivan et al., 2023)

Falska nyheter och desinformation

Den enkelhet med vilken generativ AI kan producera information gör att även blotta mängden information som kan produceras är en stor utmaning. Ta till exempel så kallade trollkonton på sociala medier där desinformation sprids genom att anställda suttit och hittat på inlägg utifrån en tänkt avsändare och mottagare. Istället för att vara beroende av anställda vid trollfabriker kan generativ AI användas för att producera i stort sett hur mycket innehåll som helst som sedan kan knytas till trollkonton. Lägg därtill hur enormt kapabel textorienterad generativ AI är när det gäller att skapa innehåll i ett visst format så är det lätt att identifiera källkritiska utmaningar.

ChatGPT kan fås att skapa ett innehåll som är anpassat efter vilket format som helst. En rapport från Natos kunskapscentrum för strategisk kommunikation (Fredheim, 2023) understryker hur lätt och utan några egentliga kostnader generativ AI kan användas för att skapa innehåll i stora mängder som till sitt format är mycket övertygande och som är oerhört svårt att avslöja som desinformation. Om det är så lätt att generera innehåll i ett snyggt format är det också lätt för makthavare att hävda att även korrekt information är falsk om det gynnar makthavarens intressen.

Ovan under rubriken ”Vad är generativ AI?” visar jag hur en kort, allmänt hållen text om Lunds kommun enkelt kan formateras om till ett mail. Generativ AI kan skapa texter för en mängd olika genrer: sociala medieinlägg, nyhetsartiklar och artiklar i uppslagsverksformat. Generativ AI används redan nu för att skapa innehåll för falska sociala mediekonton som sedan används för att sprida budskap som gynnar ryska intressen (Fredheim, 2023)

Generativ AI och informationssökning

ChatGPT används redan idag som en slags informationssökning. Som jag visar nedan medför detta problem. Därtill har sökmotorer börjat inkludera generativ AI i sina produkter.

Kan generativ AI användas för informationssökning?

ChatGPT och andra textbaserade generativa AI-produkter används inte enbart för att generera text, utan även som ett alternativ till mer traditionell informationssökning. I inledningen till kapitlet nämner jag distinktionen mellan källa och innehåll. Denna distinktion går även att applicera på informationssökning: Ibland söker användaren efter en källa och då utgör sökmotorers länkar till webbsidor precis det hen vill ha. Användaren vill helt enkelt ha länken till och namnet på en webbsida (eller annat dokument) som behandlar ett visst innehåll. I andra fall är användaren egentligen ointresserad av själva källan. Istället söker hen efter ett innehåll. Det gäller framförallt när någon letar efter sådant som kan tyckas vara enklare fakta, såsom höjden på ett visst berg, ett enkelt recept på fattiga riddare eller vad lagen (SFS 1990:52) med särskilda bestämmelser om vård av unga (LVU) egentligen innehåller.

Sedan flera år tillbaka levererar exempelvis Google Sök och Bing redan viss typ av faktatext som resultat på en sökning, ofta extraherad ur en av de översta länkarnas text, utan att vi behöver följa länkar. Om ChatGPT, Bard och andra generativ AI-produkter som bygger på LLM:er används för informationssökning blir de källkritiska utmaningarna enorma. Det är troligt att generativ AI kommer att förändra människors förväntningar på informationssökning. Ovan diskuterar jag hur information skapad av generativ AI saknar upphovsperson och autenticitet, samt hur de kan hallucinera och reproducera stereotyper. Samtidigt är lockelsen att använda dessa AI-produkter för informationssökning stor. Att klicka på en länk och läsa exempelvis en artikel kan kännas som en omväg. Samtidigt är det en nödvändig omväg om källkritik ska vara möjlig. Eftersom källkritik kräver en källa med en upphovsperson kan inte information som produceras av generativ AI på ett meningsfullt sätt granskas källkritiskt. I den historieforskningen har man skiljt mellan källkritik och sakkritik (Jarrick, 2005) och det är en användbar distinktion. En text skapad av ChatGPT kan granskas sakkritiskt, men inte källkritiskt.

Det finns alltså goda skäl till att inte använda AI-produkter som bygger på LLM:er för informationssökning. Samtidigt finns det en risk att en ökad närvaro av produkter som skapar svar på frågor i naturligt språk istället för länkar till webbsidor skapar nya förväntningar på hur informationssökning borde gå till. Det finns en risk att många användare inte kommer orka gå omvägen till länkar.

Generativ AI i sökmotorer

Det finns också exempel på hur generativ AI byggts in i traditionella sökmotorer. I praktiken innebär det dock att sökresultatet kommer i form av en löpande text med referenser. Den löpande texten är skapad av en sammanfattning av de första länkarnas innehåll om en fråga. Det sker här en snabb utveckling som är svår att beskriva i en artikel då den går så fort. För närvarande använder sig sökmotorn Bing av ChatGPT-teknik, medan Google har velat hålla isär sin AI-teknologi och sökmotor genom att kalla Bard ett komplement till Google Sök (Pierce, 2023-03-21). Det finns även ett antal andra sökmotorer som använder sig av naturligt språk på ett liknande sätt som Bing. Det ska dock sägas att även om om sättet att använda sig av AI-teknik är ny har sökmotor länge kompletterat länkar med att användaren kan ta del av viss slags information, bland annat hämtad från Wikipedia, utan att behöva lämna sökresultatsidan.

Tar företagen något ansvar för sina produkter?

Kommersiella digitala plattformar och de företag som producerat dem verkar under olika lagstiftning, beroende på i vilka länder produkten används. I Europa finns det exempelvis på EU-lagstiftning som skyddar den personliga integriteten (Kommittédirektiv, 2016:15) som inte finns i USA och de flesta andra länder. Europaparlamentet arbetar också med att ta fram en lagstiftning om artificiell intelligens – EU AI Act – som bland annat ska försäkra Europas medborgare säkerhet och transparens (Europaparlamentet, 2023-06-08). Lagstiftningen skiljer sig även åt när det gäller konsumentskydd, upphovsrätt och hur yttrandefrihet ska vägas mot exempelvis förtal, hot, hat och sexuellt innehåll. Det finns alltså lagstiftning som avkräver de kommersiella plattformarna visst samhällsansvar genom att skydda individen. Samtidigt är samhällsansvaret såsom det kommer till uttryck i lagstiftningen lägre för digitala produkter än för många andra produkter. Ta till exempel regleringen av nya läkemedel och de strikta kontroller som måste göras när ett nytt läkemedel ska lanseras, regler när nya bostäder ska byggas eller miljöprövning när en ny verksamhet planeras.

I exemplen ovan från ChatGPT och DALL·E var både mina instruktioner (prompts) och texten jag fick oproblematiska. Visserligen innehöll texten om Lund, eftersom jag bad den skriva positivt, många värdeladdade ord, men i sammanhanget är det förväntat och inget konstigt. Samtidigt kan generativ AI förstås också användas till att ta fram betydligt känsligare texter, bilder, videor och ljud. Jag testade därför ChatGPT för att se hur den hanterar en mer känslig fråga. Sedan 2020 har Sverige utsatts för omfattande och återkommande desinformation om att socialtjänsten kidnappar muslimska barn. Desinformationen är orkestrerad på ett sätt som tyder på att det rör sig om försök till informationspåverkan från främmande makt (Ranstorp & Ahlerup, 2023). Desinformationen sprids i sociala medier, men den tar sig även uttryck i form av fysiska

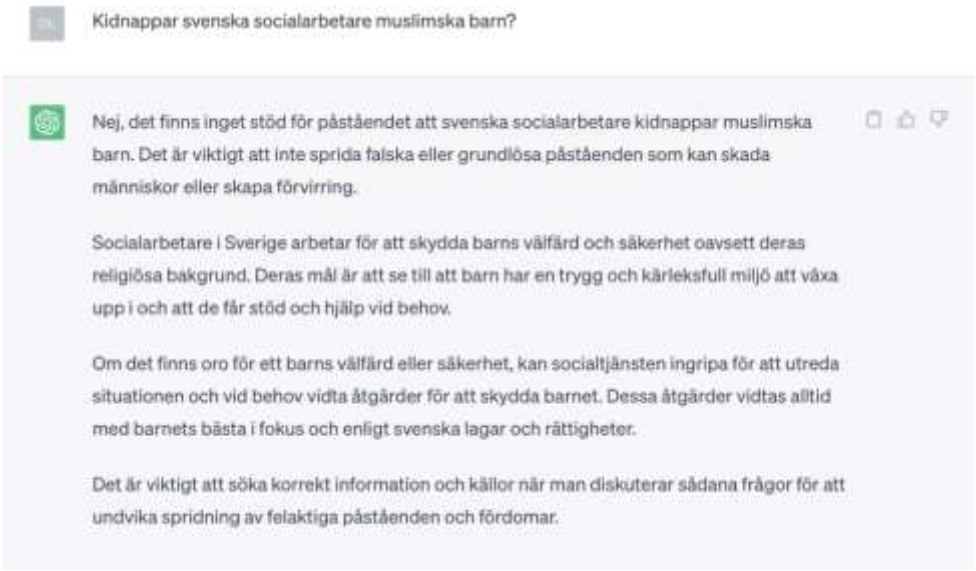
demonstrationer. För att se vad ChatGPT levererar för slags text vid känsliga frågor ställde jag följande fråga:

Figur 5

Det är uppenbart att ChatGPT behandlar alla frågor om omhändertagande av muslimska barn i Sverige som felaktig information eller desinformation. Hur jag än försöker förmå ChatGPT att formulera en hållning (persona) som skulle stödja desinformationen om att svenska socialarbetare kidnappar muslimska barn får jag liknande innehåll som ovan tillbaka. Exemplet ovan visar att företagen bakom generativa AI-produkter kan ta ett samhällsansvar om företaget som tagit fram produkten så vill (eller om lagstiftningen kräver det). DALL·E formulerar exempelvis sina riktlinjer för bilder (”content policy”) på följande sätt:

- Hat: hatiska symboler, negativa stereotyper, jämföra vissa grupper med djur/föremål, eller på annat sätt uttrycka eller främja hat baserat på identitet.
- Trakasserier: håna, hota eller mobba en individ.
- Våld: våldsamma handlingar och andras lidande eller förnedring.
- Självskada: självmord, skärsår, ätstörningar och andra försök att skada sig själv.
- Sexuellt: nakenhet, sexuella handlingar, sexuella tjänster eller innehåll som på annat sätt är avsett att väcka sexuell upphetsning.
- Chockerande: kroppsvätskor, obscena gester eller andra ämnen som kan chockera eller väcka avsky.
- Olaglig verksamhet: droganvändning, stöld, skadegörelse och andra olagliga aktiviteter.
- Bedrägeri: stora konspirationer eller händelser relaterade till stora pågående geopolitiska händelser.
- Politisk: politiker, valurnor, protester eller annat innehåll som kan användas för att påverka den politiska processen eller bedriva kampanjer.
- Folkhälsa och personlig hälsa: behandling, förebyggande, diagnostik eller överföring av sjukdomar eller personer som upplever hälsoproblem.
- Spam: oönskat massinnehåll.

DALL·E, 2022-09-19. Författarens översättning



Den moderering som generativ AI-produkterna gör har även kritiserats för att vara vänster eller till och med ”woke” på ett liknande sätt som sökmotorer har kritiserats (se Del 2) (se t ex Rozado, 2023a). Det har hävdats att ChatGPT ger svar på många frågor på ett sätt som i USA värderingsmässigt ligger närmare liberala värderingar är konservativa (Gault, 2023-01-17). Sant är, som framgår ovan, att OpenAI använder sig av moderering och att det finns riktlinjer om vad dess tjänster inte får användas till. När generativa AI-produkter tas fram som bygger på Open Source-teknik har dessa produkter i regel inte samma slags moderering som de kommersiella produkterna har. Ett exempel är FreedomGPT som marknadsförs genom en stiliserad bild av frihetsstatyn i New York.

FreedomGPT beskrivs av företaget bakom som att den ”är en 100 % ocensurerad och privat AI-chatbot /.../. Den skapades för att visa upp nödvändigheten av opartisk och censurfri AI.” (FreedomGPT, u. å., författarens översättning). Den här typen av icke-modererade AI-produkter utgör endast en liten del, men det är lätt att tänka sig att de kommer bli populära i vissa kretsar och därmed skapa ännu större utmaningar för samhället.

Generativ AI och undervisning om källkritik

I artikeln diskuteras många av de utmaningar som generativ AI för med sig för källkritik. Dessa är viktiga och stora. Samtidigt är det viktigt att skolan inte stänger dörren för dessa produkter utan istället använder dem som ett pedagogiskt objekt. Vad skolan än gör används de redan av många unga människor. Istället kan generativ AI användas för att stärka den källkritiska medvetenheten, liksom förstås även för andra saker. En utgångspunkt i en sådan pedagogisk strävan kan vara distinktionen mellan

källa och innehåll som löper som en röd tråd genom denna artikel. Några exempel på hur ChatGPT kan användas för källkritiska övningar kan vara

- Prova de traditionella källkritiska kriterierna på texter genererade av ChatGPT för att diskutera fördelar och utmaningar med generativ AI. Låt eleverna själva fundera över det (o)möjliga med att tillämpa äkthet, närhet, beroende och tendens.
- ChatGPT brukar vara särdeles dålig på källor. Be eleverna använda ChatGPT för att generera referenser och be dem kontrollera referenserna med hjälp av ett skolbibliotek eller en referensdatabas såsom Libris eller Google Scholar för att se om referenserna faktiskt finns.
- Låt elever testa ChatGPT för åldersadekvata känsliga ämnen och se hur ChatGPT svarar. Låt eleverna sedan ta del av ChatGPTs riktlinjer för innehåll. Dessa kan sedan fungera som ett underlag för diskussion i klassrummet. Består klassen av elever med olika modersmål kan samma ämne testas med olika språk.
- Använd ChatGPT för att skriva påhittade nyhetsartiklar om ämnen som intresserar eleverna. Dessa nyhetsartiklar kan sedan användas för en diskussion om hur lätt det är att sprida desinformation med hjälp av generativ AI.

Punkterna är bara exempel på vad som kan göras. De utgör alla exempel på hur elever kan synliggöra ChatGPT med sina särskilda förutsättningar. Det bör i detta sammanhang nämnas att för att använda ChatGPT behövs ett användarkonto. Skolverket konstaterar att det saknas forskning om hur ChatGPT kan användas i undervisningen och att lärare måste vara observanta på vilka regler som gäller för den enskilda skolan liksom krav på ålder i användarvillkoren (Skolverket, 2023-06-22). Ett alternativ är därför att skolan använder egna konton för att visa exempel.

Avslutning

Generativ AI kommer utan tvekan att påverka skolan och gör det redan idag. I artikeln har jag tagit upp några av de utmaningar som tekniken för med sig. Jag har fokuserat på distinktionen mellan källa och innehåll. Generativ AI skapar ett innehåll som ibland stämmer, men som vi aldrig kan vara säkra på stämmer. Däremot finns det ingen länk eller referens till en källa som kan kritiskt granskas såsom vi är vana vid vid källkritik. Ett innehållsligt påstående kan inte knytas till en artikel eller liknande med en upphovsperson (eller -organisation). I den mån källkritik är möjlig handlar det om att etablera en kritisk förståelse av själva produkten och den LLM som den bygger på som sådan. Källkritik kan dock gagnas genom att elever experimenterar med generativ AI för att synliggöra problemen (liksom förstås möjligheterna).

Referenser

Abercrombie, G., Curry, A. C., Dinkar, T., & Talat, Z. (2023). *Mirages: On anthropomorphism in dialogue systems*. arXiv preprint. arXiv:2305.09800.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). ACM.

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). *Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study*. arXiv preprint. arXiv:2303.17466.

Carlsson, H., & Sundin, O. (2018). *Sök- och källkritik i grundskolan*. Lund: Lunds universitet.

DALL·E (2022-09-19). *Content policy*. https://labs.openai.com/policies/content-policy [2023-08-12]

Kommittédirektiv (2016:15). Dataskyddsförordningen (GDPR).

Europaparlamentet (2023-06-08). *EU AI Act: First regulation on artificial intelligence*.

Fredheim, R. (2023). *Virtual Manipulation Brief 2023/1: Generative AI and Its Implications for Social Media Analysis*. Riga: NATO Strategic Communications Centre of Excellence.

FreedomGPT (u. å.). *About*. https://freedomgpt.com/about [2023-07-11]

Gault, M. (2023-01-17). Conservatives Are Panicking About AI Bias, Think ChatGPT Has Gone ”Woke”". *Vice*. https://www.vice.com/en/article/93a4qe/conservatives-panicking-about-ai-bias-years-too-late-think-chatgpt-has-gone-woke [2023-07-06]

Gross, N. (2023). What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI. *Social Sciences*, *12*(8), 435. https://doi.org/10.3390/socsci12080435

Gy22 (2022). *Läroplan för gymnasieskolan*. Skolverket.

Hsu, T. (2023-08-03). What Can You Do When A.I. Lies About You? *The New York Times*. https://www.nytimes.com/2023/08/03/business/media/ai-defamation-lies-accuracy.html [2023-08-12]

Jarrick, A. (2005). Källkritiken måste uppdateras för att inte reduceras till kvarleva. *Historisk tidskrift*, 125(2), 219-231.

Lawler, R. (2023-05-10). Google's new image search tools could help you identify AI-generated fakes. *The Verge*. https://www.theverge.com/2023/5/10/23718616/google-image-search-verification-about-this-metadata-io [2023-09-01]

Lgr22 (2022). *Läroplan för grundskolan samt för förskoleklassen och fritidshemmet*. Skolverket.

Pierce, D. (2023-03-21). Google says its Bard chatbot isn't a search engine — so what is it? *The Verge*. https://www.theverge.com/23649897/google-bard-chatbot-search-engine [2023-07-10]

Ranstorp, M., & Ahlerup , L. (2023). *LVU-kampanjen: Desinformation, konspirationsteorier, och kopplingarna mellan det inhemska och det internationella i relation till informationspåverkan från icke-statliga aktörer*. Försvarshögskolan.

Rozado, D. (2023a). The political biases of ChatGPT. *Social Sciences*, 12(3), 148. https://doi.org/10.3390/socsci12030148

Rozado, D. (2023b). RightWingGPT: An AI manifesting the opposite political biases of ChatGPT. *Rozado's Visual Analytics*. https://davidrozado.substack.com/p/rightwinggpt

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-generated text be reliably detected?* arXiv preprint. arXiv:2303.11156.

SFS 1990:52. *Lag (1990:52) med särskilda bestämmelser om vård av unga*.

Skolinspektionen (2018). *Undervisning om källkritiskt förhållningssätt i svenska och samhällskunskap: Årskurs 7-9*.

SKOLFS 2022:14. *Läroplan för vuxenutbildningen*. Skolverket.

Skolverket (2023-06-22). *Råd om Chat GPT och liknande verktyg*. https://www.skolverket.se/skolutveckling/inspiration-och-stod-i-arbetet/stod-i-arbetet/rad-om-chat-gpt-och-liknande-verktyg [2023-08-11]

Thompson, S. A., Hsu, T., & Myers, S. L. (2023-03-25). Conservatives Aim to Build a Chatbot of Their Own. *The New York Times*. https://www.nytimes.com/2023/03/22/business/media/ai-chatbots-right-wing-conservative.html

Wineburg, S., & McGrew, S. (2017). *Lateral reading: Reading less and learning more when evaluating digital information*. Stanford History Education Group Working Paper No. 2017. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3048994