



AI och skolan

Praktiska tips och framtidsspaningar
om chattbottar i skolan

Johan Falk

2.0-beta1

AI och skolan, Johan Falk

Förhandsversion 2.0-beta1, publicerad 2023-06-03

Boken är tillgänglig på github.com/Itangalo/AI-och-skolan, där du också är välkommen att föreslå förbättringar och tillägg. Där kan du också anmäla dig till ett nyhetsbrev där du får meddelanden om nya versioner av boken.

AI och skolan

När vi blickar bakåt i historien är det lätt att se stora förändringar. Internet, mobiltelefoner och senare även smartphones har påverkat vår vardag på ett sätt som är svårt att överskatta. Bilar, flygplan och satelliter har fört samman delar av världen som tidigare var avlägsna. Penicillin, vaccin och röntgenteknik har haft en enorm betydelse för vår hälsa och vårt välbefinnande. En så alldaglig sak som tvättmaskinen brukar listas som en av de viktigaste uppfinningarna (tillsammans med exempelvis korrugerad plåt!) eftersom den förbättrar livet för så många människor. Den industriella revolutionen, som i grunden innebar att maskiner tog över mycket av det muskelarbete som tidigare utfördes av människor och djur, hade också enorma sociala, ekonomiska, miljömässiga och politiska effekter.

När vi ser tillbaka är det lätt att upptäcka förändringar, men det är samtidigt lätt att tro att vi nu har nått slutstationen för hur världen ser ut. Det är alltför enkelt att föreställa sig att världen om 10, 20 och 50 år kommer att se ut ungefär som idag: Nya telefonmodeller, en större andel elbilar på gatorna, och lite bättre metoder för att behandla cancer – men inget revolutionerande.

Den utveckling vi ser inom AI kallas av vissa för *den kognitiva revolutionen*, vilket i kontrast till den industriella revolutionen sägs handla om att maskiner tar över mycket av det *tankearbete* som människor gör. Det är möjligt att det är en överdrift, men det är också fullt möjligt att det är en adekvat beskrivning av den omställning som står för dörren. Det är närmast omöjligt att förutsäga hur långt AI-utvecklingen kommer att leda i det språng som pågår nu, vilken inverkan den kommer att ha på samhälle och människor, och hur världen ser ut när utvecklingen till slut planar ut. I ett sådant läge är det rimligt att förbereda sig för ett antal olika utfall – både att framtiden i stora drag liknar världen idag, och att den kan vara radikalt annorlunda.

Den här boken handlar om AI och skolan och är riktad till lärare och annan pedagogisk personal i skolan. Boken har tre delar.

- **Del 1: Introduktion.** Här kan du lära dig hur du kommer igång med AI-baserade chattrobotar, förstå grunderna i hur så kallade *stora språkmodeller* fungerar, och även lära dig om vanliga fallgropar att se upp för.
- **Del 2: Praktiska tips.** Här finns en rad olika tips och exempel på hur du kan använda chattrobotar och i viss mån andra AI-tjänster i ditt arbete som lärare. Det

handlar både om arbete utanför klassrummet, saker du kan göra tillsammans med elever, och förslag på vad elever själva skulle kunna göra. Tipsen innehåller också tankar om när det *inte* är lämpligt att använda chattrobotar, eller vad man bör se upp med.

- **Del 3: AI och framtiden.** Här kan du läsa mer om vart AI-utvecklingen kan vara på väg, och hur det i sin tur kan påverka skolan. Det finns också ett kapitel om risker med AI, vilket är något värt att ta på allvar.

Den första versionen av boken skrevs och släpptes i januari 2023. I maj 2023 skrivs boken om mer eller mindre från grunden, för att kunna ta med teknikutveckling och nya lärdomar om hur AI kan användas i skolan (och hur den inte bör användas). Få saker i boken är baserade på forskning – teknikutvecklingen går så fort att forskningsprojekt lätt blir inaktuella innan första delstudien är klar. I stället är det mesta baserat på omdöme och halvt beprövad erfarenhet. Det är därmed särskilt viktigt att du tillämpar ditt eget omdöme när du använder boken. Har du förslag på saker som bör läggas till, tas bort eller förbättras är du välkommen att bidra med dina tankar, och göra erfarenheter mer beprövade. Det kan du göra genom bokens projektsida, github.com/itangalo/AI-och-skolan. På samma sida kan du ladda hem boken gratis.

Jag som skriver boken heter Johan Falk och är tidigare vetenskapsjournalist, webbutvecklare och gymnasielärare. Jag har varit intresserad av AI sedan 2018 och varit intresserad av skolfrågor sedan högstadiet. Sedan 2015 arbetar jag med skolfrågor på nationell nivå. Boken har mig själv som avsändare, men flera idéer och förbättringar kommer från andra personer. Ett särskilt tack till Rustan Håkansson.

Jag hoppas att du finner boken användbar. Tack för det viktiga jobb du gör i skolan.

Johan Falk, 3 juni 2023

Del 1: Introduktion

Vad är chattrobotar?

Den 30 november 2022 lanserade det USA-baserade företaget OpenAI tjänsten ChatGPT. Den tog världen med storm, och ett halvår senare orsakar AI fortfarande nyhetsinslag i riksmidia så gott som varje dag.

Det som fortfarande förbluffar många med ChatGPT (författaren inräknad) är dess förmåga att föra en konversation på ett övertygande mänskligt sätt. I kombination med kunskaper inom vitt skilda områden gör det ChatGPT och många andra chattrobotar till ett otroligt fascinerande och mångsidigt verktyg.

Låt oss titta på en konversation med ChatGPT, för att ge exempel på både dess språkliga och faktamässiga förmågor.

Exempel från ChatGPT (GPT-4 12 maj 2023).

Jag: Nu är tiden mellan hägg och syren. Vad tänker du om det?

Chattrobot: Som en AI har jag ingen personlig uppfattning eller känslor, men jag förstår vad du säger. "Tiden mellan hägg och syren" är ett vanligt svenskt uttryck som beskriver en kort period på våren när häggen blommar följt av syrenens blomning. Det är ofta en tid av återfödelse och skönhet när naturen kommer till liv igen efter vintern. Denna period är också känt för att vara mycket kort, vilket kan vara en metafor för hur snabbt livets vackra ögonblick kan passera.

Jag: Om du skulle ge dig på att beskriva känslan av att vara ute i gräset, i halvskugga under pilträden, medan barnen springer omkring – hur skulle det låta?

Chattrobot: Naturligtvis, även om jag som en AI inte har någon personlig erfarenhet eller känslor, kan jag använda den information jag har för att skapa en beskrivning baserad på det du har angivit:

Du sitter på en mjuk, grön matta av färskt, frodigt gräs som fjädrar under dig. Lätta skuggor från pilträden dansar runt dig, ett levande virrvarr av

Vad är chattrobotar?

Ljus och mörker, och solstrålar sipprar igenom lövverket, vilket skapar ett skimrande spektrum av gröna toner.

Luften är fylld med ljuden av barnens glada skratt som ekar mellan träden, fötter som trampar på gräs och löv, och sporadiska utrop av glädje. Deras energi är smittsam, och du kan inte låta bli att le när du ser dem springa fritt, oförskräckta av världen omkring dem.

Ett svagt bris rör vid ditt ansikte, bär med sig den färska, jordiga doften av gräs och en svag underton av syren och hägg, påminnelser om våren. Det är en doft av liv, av uppvaknande, av en ny början.

Din plats i gräset, under pilträden, blir en oas av lugn och frid mitt i barnens livliga lek. Det är en stund av njutning, en påminnelse om naturens enkelhet, och glädjen av att vara närvarande i nuet.

Jag: Är det här en bra tid på året för att sätta potatis?

Chattrobot: Ja, det är en bra tid att sätta potatis. Traditionellt sett är våren den bästa tiden för att sätta potatis, vanligtvis mellan april och maj i många tempererade regioner, inklusive stora delar av Sverige. Det viktigaste är att frosten ska ha passerat eftersom potatisplantorna är känsliga för frostsador.

Potatisen börjar gro när marktemperaturen når cirka 7-10 grader Celsius. Innan potatisen planteras kan det vara en bra idé att förgro dem inomhus i ljus och sval miljö. Detta innebär att du låter dem börja skjuta skott innan de planteras i marken.

Kom också ihåg att potatisen föredrar väl-dränerad, lätt sur jord och bör placeras med 30-40 cm mellan varje planta, och med cirka 10 cm djup.

Alltid kontrollera de specifika råden för din region och den specifika potatissorten du tänker plantera, eftersom det kan finnas variationer.

AI, chattrobotar och ChatGPT

ChatGPT är ett av flera sätt att använda chattrobotar, och det är värt att försöka hålla isär begreppen för att ChatGPT inte ska bli synonymt med vare sig chattrobotar eller AI.

- **AI, artificiell intelligens**, är ett brett begrepp som vanligtvis används för att beskriva maskiner som i något avseende försöker efterlikna mänsklig intelligens, och omfattar ofta också att maskinen lär sig saker och inte bara agerar enligt inprogrammerade regler. Exempel på användningsområden är självkörande bilar, ansiktigenkänning och autokorrigerigering av text på telefonen.
- **Generativ AI** är en typ av AI som utifrån att ha tränats på massor av exempel kan skapa nytt material. Exempel är Midjourney (bild), OpenAI:s GPT (text) och AlphaCode (kod).
- **En AI-modell** är motorn bakom det som en AI gör. Exempel är GPT-3.5, GPT-4 och PaLM 2.
- **Chattrobotar** är generativ AI som (huvudsakligen) arbetar med text och dessutom är skapade för att föra dialoger. Exempel är GPT-4, PaLM 2 och Claude.
- **ChatGPT** är en produkt från företaget OpenAI. Den består huvudsakligen av en webbsida där du kan chatta med några av deras AI-modeller (som är chattrobotar).

I skrivande stund är det förhållandevis tydligt att den chattrobot som har fått störst genomslag och har bredast kompetens är GPT-4, som lanserades i mars 2023 och (liksom GPT-3.5) går att nå genom tjänsten ChatGPT. Men en rad andra chattrobotar har dykt upp, både kommersiella och öppna, liksom massor av verktyg som använder AI-modeller i bakgrunden för att kunna göra mer komplexa eller specialiserade saker. Det betyder att andra chattrobotar eller AI-verktyg är bättre på vissa typer av uppgifter, och det kan också hända att nya modeller inom en snar framtid tar över ledarplatsen från GPT-4.

I den här boken används termen *chattrobot* om det inte någon specifik AI-modell eller produkt avses. Där konversationer med chattrobotar skrivs ut markeras vilken AI-modell det rör sig om (inklusive datum för senaste uppdateringen för modellen alternativt datum för konversationen, för att kunna se när exempel blivit utdaterade).

Testa själv!

Det finns inget bättre sätt att lära sig om chattrobotar än att själv testa att använda en. Om du ännu inte testat att använda en chattrobot, lägg i från dig boken och hitta en.

- På chat.lmsys.org kan du testa ett antal öppna AI-modeller, också utan att behöva skapa något konto. Den mest kraftfulla AI-modellen bland de tillgängliga är förvald (för närvarande "vicuna-13b"), så du kan bara bläddra ner och börja skriva.
- På open-assistant.io kan du skapa ett (gratis) konto med ett par klick, inklusive att godkänna användarvillkor, och börja använda den öppna AI-modellen Open Assistant.
- På chat.openai.com kan du skaffa ett (gratis) konto för ChatGPT, som du kan börja använda efter att du bekräftat din e-postadress genom ett mail som du får.

Är du fortfarande här? Gå och testa en chattrobot! En bra första konversation kan vara att fråga om något som du jobbar med och skulle vilja höra någon annans tankar.

Exempel från chat.lmsys.org (vicuna-13b).

Jag: Vad är viktigt att tänka på när man undervisar i hem- och konsumentkunskap?

Chattrobot: När man undervisar i hem- och konsumentkunskap finns det flera saker som är viktiga att tänka på för att säkerställa att eleverna får en effektiv utbildning. Här är några förslag:

1. Anpassa lärande till elevernas behov och intressen: Eleverna har olika behov och intressen, så det är viktigt att anpassa lärande till deras individuella behov. Gör det till en aktivitet där eleverna kan delta och engageras på ett personligt sätt.
2. Måla upp realistiska situationer: Lära eleverna om realistiska situationer som de kan stöta på i sin vardagliga livet. Det kan vara att lära dem om att köpa en bil, att sköta sin hälsa, eller att hantera sin konsumtion.

3. Utvärdera och bedöma källor: Utvärdera och bedöma källorna som eleverna använder när de ska ta beslut om att köpa eller beställa en produkt eller tjänst. Det kan innefatta att de ska undersöka produktens kvalitet, pris, och om det finns alternativ som är bättre.
4. Skapa engagemang: Skapa engagemang och aktivitet bland eleverna genom att fråga dem om deras egna erfarenheter och idéer. Låt dem delta i diskussioner och samarbete med andra elever.
5. Förmedla användbara fakta: Förmedla användbara fakta och information om produkter och tjänster som eleverna kan använda i sitt vardagliga liv. Ge dem tillgång till information som kan hjälpa dem att göra välgrundade beslut.
6. Förbereda eleverna för framtiden: Förbereda eleverna för framtiden genom att låta dem undersöka olika karriärmöjligheter och befattningar inom branscherna som de tänker sig att arbeta i.
7. Gör det

Jag: Fortsätt

Chattrobot: roligt: Gör undervisningen roligt och anpassad till elevernas intressen. Försök använda olika metoder som lekar, skämt, och övningar för att hålla eleverna engagerade och intresserade. Gör det lätt att förstå och kom tillrätta med svåra koncept genom att använda enkla och begripliga uttryck och exempel.

8. Gör det till en aktivitet: Låt undervisningen vara en aktivitet där eleverna kan delta och engageras på ett personligt sätt. Gör det till en tävling eller en utmaning där eleverna kan vinna priser eller belöningar för att de gjort ett bra jobb.
9. Ge dem feedback: Ge eleverna feedback på deras arbete och hjälp dem att förstå vad de gjort rätt och fel och hur de kan förbättra sig. Gör det till en positiv aktivitet och låt dem veta att de gjort bra ifrån sig.
10. Låt dem delta i beslut: Låt eleverna delta i beslut som gör deras lärande relevant och kommersiellt. Gör det till en aktivitet

Vad är chattrobotar?

där eleverna kan försöka sätta igång sin egen företagsamhet eller starta en egen organisation.

11. Genom att använda dessa tips kan du säkerställa att eleverna får en effektiv och rolig utbildning i hem- och konsumentkunskap.

Exempel från Open Assistant (oasst-sft-6-llama-30b). (Försök till konversation på svenska gav nonsens.)

Jag: Can you give me three suggestions on what to do at the pre-school tomorrow afternoon? There are 15 children (3–5 years old) and 3 adults. We want to be outside.

Chattrobot: Sure, here are three activities that can work well for a group of preschoolers aged 3-5:

Bubble Blowing Contest - Fill up a container with bubbles (be sure to use non-toxic soap) and let the kids take turns blowing as many bubbles as they can in one breath. Set some rules such as no running or jumping and offer prizes for biggest, smallest or most colorful bubbles!

Obstacle Course - Set up an obstacle course using chairs or cones. Include simple activities like hopping, skipping or crawling through tunnels. Divide into teams and have them race each other while completing all tasks correctly!

Scavenger Hunt - Make a list of items for the students to find in your outdoor space such as leaves, sticks, feathers etc. Let the groups search alone but remind that everyone must return together when found so not to wander off too far from the group. When finished, have the students sort their treasures by shape or texture! Have fun!

Exempel från ChatGPT (GPT-4 12 maj 2023).

Jag: Vad är ett bra sätt att förklara kemisk jämvikt för mina årskurs 2:or på gymnasiet?

Vad är chattrobotar?

Chattrobot: Absolut! Kemisk jämvikt är ett viktigt koncept inom kemi, men det kan vara ganska svårt att förstå första gången man stöter på det. Här är ett enkelt sätt att förklara det för dina gymnasieelever:

För det första, det är viktigt att förstå att i en kemisk reaktion, kommer reaktanter (de saker som reagerar) att omvandlas till produkter. Men det är inte alltid en enkel one-way process - ibland kan produkterna faktiskt omvandlas tillbaka till reaktanter. Detta är kärnan i konceptet kemisk jämvikt.

Så föreställ dig att du har en låda med bollar. Du börjar med 100 röda bollar och 0 blå bollar. Men med tiden, börjar några av de röda bollarna ändras till blåa. Efter ett tag har du 50 röda bollar och 50 blåa. Men nu börjar några av de blåa bollarna ändras tillbaka till röda. Och så fortsätter det - bollar ändrar färg fram och tillbaka.

Detta är vad som händer i en kemisk reaktion som har uppnått jämvikt: reaktanter blir till produkter, och produkter blir till reaktanter, i en konstant cykel. Och även om det kan verka som om inget händer eftersom mängderna inte ändras (liksom det totala antalet bollar i din låda inte ändras), pågår det faktiskt ständiga förändringar på molekylär nivå.

Det är också viktigt att påpeka att vid kemisk jämvikt betyder det inte alltid att mängderna av reaktanter och produkter är lika - det betyder bara att hastigheterna för de två reaktionerna (reaktanter blir till produkter, och produkter blir till reaktanter) är lika.

Hoppas det hjälper! Låt mig veta om det finns något mer du vill veta om kemisk jämvikt.

Hur fungerar chattrobotar?

”Det är omöjligt att skilja tillräckligt avancerad teknologi från magi.”

Vad är chattrobotar?

Citatet kommer från Arthur C. Clarke, förmodligen mest känd för boken *År 2001 – ett rymdäventyr*, och kan beskriva känslan en del får när de för första gången ser en dator konversera som en människa. *Hur går det till?*

Det visar sig, föga förvånande, att det inte är magi utan just avancerad teknik. (Även om det också visar sig vara så pass svårt att förstå den faktiska processen att det nästan lika gärna kunde varit magi – mer om det i del 3.)

Chattrobotar är som tidigare nämnts så kallade generativa AI, närmare bestämt en sort som kallas *stora språkmodeller* eller LLM, *large language models*. Grovt sammanfattat så har sådana modeller tränats på massiva mängder text, genom att försöka gissa nästa ord i meningar. Modellen innehåller miljarder av så kallade parametrar, som förenklat motsvarar kopplingar mellan neuroner i biologiska hjärnor, och styr hur modellen gissar nästa ord. De parametrarna ändras medan AI-modellen tränas, så att dess gissning på nästa ord stämmer bättre och bättre med hur texterna i träningsdatan faktiskt ser ut.

Resultatet blir en AI-modell som kan skapa naturlig text. De flesta språkmodeller har sedan fått ytterligare minst ett lager med träning, där människor (eller andra AI-modeller) ger återkoppling på vad som exempelvis är en trevlig ton, hjälpsamma svar, kränkande, olämpligt eller tydligt. Om allt detta görs på rätt sätt leder det i slutändan fram till en AI-modell som man kan föra en naturlig dialog med, om det mesta mellan himmel och jord.

När modellerna som ChatGPT använder tränades, användes stora delar av den text som fanns på internet fram till september 2021 i kombination med gigantiska mängder datorkraft. Sedan dess har flera sätt att effektivisera träning av språkmodeller utforskats, vilket inte har gett riktigt lika kraftfulla modeller men har å andra sidan kostat så lite som 30 dollar (medan GPT-4 uppskattats ha kostat omkring hundra miljoner dollar att träna).¹ Att använda de färdiga modellerna kräver mycket mindre datorkraft än att träna upp dem, och det finns nu språkmodeller som till och med kan köras på en persondator.

I grund och botten är språkmodeller tränade på att skapa meningar som på ett abstrakt sätt "stämmer med" hur texter de tränats på ser ut. Det är fascinerande att en modell genom detta lär sig stavning, grammatik och meningsbyggnad – och närmast otroligt att

¹<https://bdtechtalks.com/2023/04/17/open-source-chatgpt-alternatives/>

Vad är chattrobotar?

den dessutom lär sig översätta texter, använda komplicerade begrepp på ett (ofta) korrekt sätt, och även i viss mån föra resonemang.

Inte desto mindre är det viktigt att komma ihåg att chattrobotar i första hand är gjorda för att skapa text som låter bra – inte att säga saker som är sanna. Mer om det i nästa kapitel.

Vanliga misstag med chattrobotar

Som med alla andra verktyg är det värt att använda chattrobotar med omdöme. I den första versionen av boken, från januari 2023, fanns fyra vanliga misstag listade. I andra boken, skriven i skiftet maj–juni, har listan vuxit till sju. Listan kommer sannolikt fortsätta att växa när vi lär oss mer om effekterna av att använda chattrobotar, och det är viktigt hela tiden använda sitt omdöme.

Här är sju frågor som är värdefulla att ställa för att undvika vanliga misstag med chattrobotar.

- Finns det språkfel?
- Finns det faktafel?
- Finns det skeva perspektiv eller skeva värderingar?
- Är jag på väg att skicka känslig information?
- Är det här oreflekterad avlastning?
- Följer det här skolans styrdokument?
- Är det här bra pedagogik?

Finns det språkfel?

När ChatGPT och GPT-3.5 lanserades var det ganska vanligt med stavfel och språkfel när man chattade på svenska. Med GPT-4 är det ovanligt med stavfel på svenska, och extremt ovanligt på engelska, men de förekommer fortfarande. (Inte minst om man ber chattroboten skriva med stora bokstäver, visar det sig.) Språkfel är mer eller mindre vanliga även hos andra chattrobotar.

Stavfel eller felaktig användning av en/ett är ofta inte problem i praktiken, men är ändå värda att försöka slippa.

Finns det faktafel?

Stora språkmodeller är tränade för att skapa meningar som låter naturliga, vilket förvånansvärt ofta också betyder att budskap stämmer med hur vi människor uppfattar världen. Men chattrobotar låter sig inte hindras av brist på sakkunskap, utan levererar gärna snygga meningar även där substans saknas.

Nedan följer ett exempel som kanske är särskilt belysande.

Exempel från ChatGPT (GPT-3.5, 24 maj 2023)

Jag: Jag vill lära mig mer om vanliga inlärningssvårigheter i matematik på mellanstadiet. Kan du ge mig förslag på en forskningsartikel att läsa?

Chattrobot: Absolut! Här är en forskningsartikel som du kan läsa för att få mer information om vanliga inlärningssvårigheter i matematik på mellanstadiet:

Artikel: "Understanding Mathematics Learning Difficulties: A Comprehensive Examination of Children's Difficulties in Learning Mathematics" (Förståelse av matematikinlärningssvårigheter: En omfattande undersökning av barns svårigheter att lära sig matematik)

Författare: Ann Dowker, Camilla Gilmore, Chris C. C. Lewis, och Jo-Anne LeFevre

Publikation: Developmental Psychology (En tidskrift inom utvecklingspsykologi)

År: 2016

Artikeln undersöker olika aspekter av matematikinlärningssvårigheter hos barn i åldrarna 7-11 år. Forskarna granskar de underliggande orsakerna till dessa svårigheter och föreslår strategier för att hjälpa elever att övervinna dem. Artikeln ger en bred översikt över forskning inom området och diskuterar faktorer som arbetsminne, visuospatiala

förmågor och matematiskt språk som kan påverka elevers inläring av matematik.

Du kan hitta artikeln på följande länk: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5260818/>

Kom ihåg att vissa forskningsartiklar kan vara betalda eller kräva prenumeration för fullständig tillgång. Om du inte har tillgång till artikeln kan du prova att söka efter dess titel i en sökmotor för att hitta eventuella öppna versioner eller sammanfattningar.

Lycka till med din forskningsläsning!

Tipset från chattroboten ovan är tydligt och förtroendeingivande, och värre än meningslöst – artikeln som nämns finns inte. (Länken leder till en artikel med titel "Neuroinflammation Induces Neurodegeneration" och det finns inga träffar på nätet som matchar den artikelrubrik som chattroboten nämner.)

Den som faktiskt försöker läsa artikeln som chattroboten föreslår märker förstås rätt snart att den inte finns, men många faktafel är av karaktären att det inte märks direkt att de är fel – och den som inte är insatt i ämnet har då mycket liten chans att upptäcka att det som chattroboten säger faktiskt inte stämmer. Hade jag till exempel bett chattroboten att sammanfatta det viktigaste som forskning säger om inlärningssvårigheter i matematik på mellanstadiet hade jag fått en prydlig text tillbaka, utan chans att veta om den stämmer eller inte med mindre än att jag själv redan har (eller skaffar mig) god koll på det jag frågar om. Det leder inte bara till att jag riskerar att inte lära mig det jag vill, utan att också lära mig något galet och *tro* att jag har koll.

Det är lätt att föreställa sig att elever kan gå i samma fälla om de lämnas att själva lära sig om det ena eller andra med hjälp av chattrobotar – i alla fall om chattrobotarna bara förlitar sig på språkmodeller, och inte exempelvis kontrollerar budskap mot tillförlitliga källor.

Finns det skeva perspektiv eller skeva värderingar?

De flesta chattrobotar har tränats på text från internet, vilket betyder att de också har potential att uttrycka sexism, rasism, extremism och många andra olämpliga värderingar som finns att hitta på nätet. I stor utsträckning förhindras det av ett sista steg i träningen av AI-modeller kallad *reinforcement learning from human feedback* eller RLHF, vilket kan ses som människor lär chattrobotar hyfs. Resultatet är dock inte felfritt: olämpliga budskap som man försökt träna bort kan ändå synas ibland och – en knepigare nöt att knäcka – olämpliga budskap som man *inte* aktivt försökt träna bort fortfarande finns kvar. Ett extremt exempel på det senare är att talibanregimen i Afghanistan troligtvis anser att hyfsträning lämnar stora luckor, eftersom de inte omfattar en (enligt dem) korrekt syn på kvinnor. I andra kulturer kan man tänka sig mer subtila luckor, som kan gå under radarn för de flesta men till exempel kan innehålla budskap som minoritetsgrupper finner direkt kränkande.

Men hyfsträning kan också skapa svårigheter i sig. Människor i Sverige kan tycka att hyfsträning gjord efter USA-standarder ger en lite konstig bild när det gäller syn på abort eller användning av cannabis¹. En studie på en rad olika chattrobotar visar att modeller som genomgått hyfsträning, kanske inte helt överraskande, oproportionellt mycket avspeglar åsikter det som i USA kännetecknas av liberala, välutbildade och välbeställda människor.² Det finns med andra ord anledning att fundera över om budskap från chattrobotar under ytan har med sig värderingar eller en världssyn som inte stämmer med vad vi vill (och ska) förmedla i skolan.

Det finns chattrobotar som inte har genomgått hyfsträning, och alltså kan slänga ur sig budskap utan tanke på om de är oartiga, kränkande eller direkt skadliga. Sådana modeller kräver i regel mer expertis för att hitta och använda, och spelar därför en mindre roll i skolan.

¹Just synen på cannabis blev en nyhet i Sverige i mitten av maj, då den chattrobot som alla Snapchat-användare har tillgång till tyckte att cannabis kan användas av den som "vill ha roligt med vänner": <https://sverigesradio.se/artikel/snapchats-ai-bot-gav-rad-om-cannabis-for-den-som-vill-ha-roligt>

²<https://arxiv.org/abs/2303.17548>

Är jag på väg att skicka känsliga uppgifter?

I april 2023 visade det sig att anställda på Samsung hade skickat känslig data till ChatGPT – i två fall för att förbättra känslig kod, i ett tredje var det mötesanteckningar som delades.³ Det finns förmodligen många fler företag och organisationer som skickat information till chattrobotar utan att fundera över om det är data som är lämplig att dela.

Man kan utgå från att chattrobotar är molntjänster med servrar utanför EU, och att all information som skickas både bearbetas och sparas. Man ska därför inte skicka personuppgifter eller andra känsliga uppgifter.

Är det här oreflekterad avlastning?

En särskilt knivig fråga med verktyg som elever själva använder handlar om *oreflekterad avlastning*. Saker som miniräknare, ordböcker och ordlistor, formelblad, rättstavningsprogram, uppläsningssystem – och chattrobotar – kan vara fantastiska verktyg för att kunna arbeta smidigare och fokusera på det som är undervisningens egentliga innehåll. De gör det möjligt att slippa fastna i saker som kan stjåla värdefull tankekraft – tankekraft för att bearbeta och lära sig nya saker.

Men det kan också hända att verktyget tar hand om saker som eleven behöver träna på. Att alltid ha miniräknare till hands gör att elever i mindre utsträckning räknar i huvudet eller på papper, och då kommer de att bli sämre på det. Att lyssna på text ger mindre träning i traditionell läsning, att skriva på tangentbord ger mindre träning i att skriva för hand (och omvänt!), att ta rulltrappan ger mindre träning för benen. Och att använda chattrobotar för att formulera texter gör att vi blir sämre på att göra det för hand.

Lösningen på problemet är *inte* att låta bli verktyg helt och hållet. Lösningen är snarare att vara medveten om att det ofta är mödosamt att lära sig något, och att det alltså är viktigt att välja vad elever ska lägga sina ansträngningar på. (Att något är mödosamt är dock inte samma sak som att det är tråkigt, vilket också är värt att komma ihåg.) Om man vill använda chattrobotar för att stötta läsning eller skrivande i lågstadiet kräver det

³<https://mashable.com/article/samsung-chatgpt-leak-details>

ordentlig eftertanke, medan det är mindre problematiskt att låta studenter på högskola använda en chattrobot för att göra en rapport lättare att läsa.

Eftersom elever kan behöva arbeta med olika saker, och kan behöva stöd med olika saker, kan det hända att chattrobotar hjälper en elev medan en annan elev i samma klass hindras i sitt lärande av precis samma stöd.

Följer det här skolans styrdokument?

Det finns flera tjänster som använder språkmodeller för att generera färdiga lektionsplaneringar, inklusive en krok för att väcka intresse i början av lektionen, beskrivningar av aktiviteter och hur lång tid de ska ta, och frågor att avsluta lektionen med. Och även utan specialbyggda verktyg kan många chattrobotar skapa en lektionsplanering på mindre än en minut.

Men det behöver inte betyda att lektionen tar upp det som skolans styrdokument säger att undervisningen ska handla om. Det är inte en slump att det går åt mycket kraft för att skapa läromedel, där redaktörer och skickliga lärare arbetat för att få till delar och helheter som fångar både det som kurs- och ämnesplaner tar upp och de mer övergripande målen som finns i läroplaner.

Exempel från ChatGPT (GPT-3.5, 12 maj 2023).

Jag: Kan du skapa en lektionsplanering för mig? Vi ska ha första lektionen av sex som fokuserar på vikingatiden. Mellanstadiet.

Chattrobot: Absolut! Här är en lektionsplanering för första lektionen om vikingatiden på mellanstadiet:

Lektion 1: Vikingatiden

Ålder: Mellanstadiet (ca 9-12 år)

Mål:

- Introducera eleverna till vikingatiden och dess betydelse.

- Utforska vikingarnas livsstil, inklusive deras samhälle, handel och utforskning.
- Förstå vikingarnas teknologiska framsteg och deras inverkan på omvärlden.
- Skapa medvetenhet om olika källor till information om vikingatiden.

Material:

- Bilder och illustrationer av vikingatiden.
- Karta över vikingarnas område.
- Whiteboard eller svart tavla.
- Pennor och papper.
- Källor till information om vikingatiden (böcker, webbplatser etc.)

Lektionsplan:

1. Introduktion (10 minuter)
 - 1.1 Hälsa eleverna välkomna och presentera ämnet: Vikingatiden.
 - 1.2 Visa några bilder och illustrationer av vikingatiden för att väcka intresse och nyfikenhet.
 - 1.3 Diskutera vad eleverna redan vet eller tror sig veta om vikingarna och deras tid.
2. Samhället och levnadssätt (15 minuter)
 - 2.1 Beskriv vikingarnas samhälle och deras levnadssätt. Diskutera deras boenden, kläder, mat och familjestruktur.
 - 2.2 Be eleverna jämföra och kontrastera vikingarnas livsstil med deras egen.
3. Handel och utforskning (15 minuter)
 - 3.1 Förklara vikingarnas handelsnätverk och deras betydelse för vikingasamhället.

- 3.2 Diskutera vikingarnas utforskning och deras resor över havet. Visa en karta över vikingarnas område och markera viktiga handelsvägar och utforskningsrutter.
4. Vikingarnas teknologiska framsteg (15 minuter)
 - 4.1 Berätta om vikingarnas teknologiska framsteg, inklusive deras båtar, vapen och navigationsverktyg.
 - 4.2 Diskutera hur dessa teknologier hjälpte vikingarna att bli framgångsrika handelsmän och utforskare.
5. Källkritik (15 minuter)
 - 5.1 Förklara vikten av att använda olika källor för att lära sig om vikingatiden.
 - 5.2 Diskutera skillnaden mellan primära och sekundära källor och varför det är viktigt att vara källkritisk.
 - 5.3 Ge några exempel på olika källor som eleverna kan använda för att lära sig mer om vikingatiden.
6. Avslutning (5 minuter)
 - 6.1 Summera dagens lektion och viktiga punkter som behandlades.
 - 6.2 Uppmuntra eleverna att ställa frågor och fortsätta undersöka ämnet på egen hand.

Om möjligt kan du också överväga att inkludera praktiska aktiviteter, som att låta eleverna rita och färglägga egna vikingaskepp eller skapa enkla vikingainspirerade smycken.

Kom ihåg att anpassa lektionsplaneringen efter dina elevers behov och intressen. Lycka till med din lektion om vikingatiden!

Är det här bra pedagogik?

De som har lätt för att fastna i nya roliga verktyg (författaren inräknad) har nytta av att regelbundet fråga sig om det man använder chattrobotar till faktiskt är bra pedagogik, eller om verktyget har fått bestämma för mycket över undervisningens utformning. Att

Vanliga misstag med chattrobotar

det går snabbt och är roligt att skapa automaträttade quizzar behöver inte betyda att det är en bra idé att avsluta *varje* lektion med en quiz.

Chattrobotars styrkor

Med alla misstag och brister som beskrivs ovan kan man börja undra om det verkligen är en bra idé att använda chattrobotar i skola, i eller utanför klassrum. De skapar faktafel som är svåra att upptäcka, smyger in en världssyn som kan göra skada, lurar användare att skicka känsliga uppgifter, följer inte skolans styrdokument och riskerar att få elever att lära sig mindre när de använder verktyg i stället för att göra saker själva.

Ovanpå det kan man lägga en meh-effekt, som man får om det första man testat är att ställa en fråga för att se om chattroboten kan svara rätt. Det leder ofta till besvikelse – ibland för att chattroboten svarar fel, men oftare för att den svarar rätt och man sedan frågar sig vad vitsen var. Vi har ju redan sökmotorer?

Det tar ett tag att vänja sig vid att chattrobotar inte primärt är uppslagsverk eller sökrobotar, utan kanske snarare ska ses som en assistent på burk. Som diskuteras nedanför under rubriken *hybrider* kan chattrobotar fungera även som uppslagsverk, om språkmodeller kombineras med andra verktyg, men de tre starkaste områdena för rena språkmodeller är att bearbeta text, vara idéspruta eller bollplank, och att ge stöd i programmering. Om man använder chattrobotar till saker som de är bra på, och håller de sju frågorna ovan i bakhuvudet, kan man få ut mycket nytta av dem.

Nedan beskrivs kort de tre starka områdena för chattrobotar. I del 2 finns mer utförliga exempel på hur det kan se ut i praktiken (där en del exempel också går utanför de här tre områdena).

Bearbeta texter

Det är kanske inte så förvånande att något som heter "stor språkmodell" och har tränats på gigantiska mängder text är bra på att hantera just texter. Det kan till exempel betyda:

- Korrekturläsa texter.
- Ge förslag på förbättringar av en text, generellt eller utifrån särskilda perspektiv.
- Sammanfatta texter.
- Skapa välskrivna texter av lösa anteckningar.
- Skapa utkast till nya texter.

- Ge förslag på disposition.
- Översätta texter.
- Skriva om texter till ett enklare språk, eller till en speciell språkstil.

Vara idéspruta eller bollplank

Att bolla tankar med en kollega är ofta ett sätt att komma vidare i sina egna tankar, både genom att man får sätta ord på tankarna och att man får nya idéer från kollegan. Vid tillfällen då kollegor inte finns i närheten, eller inte har tid att diskutera en ny (eller möjligtvis gammal) idé kan chattrobotar vara en ersättare. De saknar en rad mänskliga aspekter, men har å andra sidan obegränsat med både tid och tålamod.

En särskild förmåga som chattrobotar har är att kunna hitta på massor av idéer där den mänskliga hjärnan (författarens inräknad) går på grund efter en kort stund.

Ge stöd i programmering

Långt ifrån alla lärare använder programmering i eller utanför sin undervisning, men de som gör det bör vara medvetna om den typ av stöd man kan få från AI-verktyg. Förutom chattrobotar finns det också särskilda verktyg skapade för just programmering, där det mest kända förmodligen är [GitHub Copilot](#).

Chattrobotar och annat AI-stöd kan till exempel:

- Felsöka kod.
- Förklara kod.
- Förse kod med förklarande kommentarer.
- Skriva enklare kod utifrån beskrivningar i kommentarer.
- Skriva korta program för att lösa givna problem.
- Säkerhetsgranska kod.
- Skriva tester.

Chattrobotars styrkor

Om elever använder AI-stöd för programmering behöver man förstås vara extra vaksam på oreflekterad avlastning. Om fokus ligger på att elever ska lära sig att skriva viss typ av kod är det förstås mindre lyckat om en AI skriver åt dem.

Som med all kod man får given till sig behöver man också hålla ett öga öppet efter potentiella säkerhetshål.

Hybrider

Språkmodeller har, som nämnts ovan, nackdelen att de inte bryr sig särskilt mycket om vad som är sant eller falskt, utan det primära målet är att skapa meningar som låter bra. Genom att koppla ihop språkmodeller med andra typer av applikationer kan man komma runt mycket av de problemen.

Tjänsten perplexity.ai är en sådan hybrid av språkmodeller (GPT-3.5 och GPT-4) kombinerat med sökmotorer, som dessutom går att ställa in på att söka i olika typer av källor. Här nedanför används samma fråga som gav en påhittad forskningsartikel från GPT-3.5.

Exempel från perplexity.ai (GPT-3.5 + sökområde "all").

Jag: Jag vill lära mig mer om vanliga inlärningssvårigheter i matematik på mellanstadiet. Kan du ge mig förslag på en forskningsartikel att läsa?

Chatrobot: Here are some research articles related to common learning difficulties in mathematics on middle school level:

1. "Inlärningssvårigheter i matematik" (1): This article discusses the difficulties that children with learning difficulties in mathematics often face, such as problems with working memory and emotional and volitional deviations.
2. "Attityder inom matematik utifrån ett lärarperspektiv" (2): This study focuses on the attitudes towards mathematics and how they relate to learning difficulties in middle school.
3. "Inlärningssvårigheter i matematik och läsning hänger ihop" (3): This research article examines how learning difficulties in mathematics and reading are related and how they can affect each other.
4. "Stödmaterial matematiksvårigheter, del 1. Att upptäcka behov av stöd i matematik" (4): This article discusses how to identify students who have difficulties in mathematics and how to provide support.

5. "Hur hanterar matematiklärarna extra anpassningar på mellanstadiet?" (5): This study examines how teachers can provide extra support and adaptations to students with learning difficulties in mathematics on middle school level.
6. "Matematiksvårigheter" (6): This research article summarizes previous studies on mathematics difficulties and provides recommendations for how to help students with learning difficulties in mathematics.

Svaret från chattroboten levererades tillsammans med länkar till respektive artikel. Flera av träffarna är examensarbeten och alltså inte publicerade i vetenskapliga tidskrifter.¹

Perplexity.ai är en egen tjänst som anropar chattrobotar i bakgrunden, vilket också är en metod som språklärningstjänsten Duolingo använder. En omvänd metod är i stället en hybrid där själva chattroboten utökas med funktioner för att anropa andra applikationer – något som blev möjligt i form av så kallade plugins för betalversionen av ChatGPT i mitten av maj. Nedan syns ett exempel på hur en konversation kan se ut när plugin för WolframAlpha används – en webbtjänst för matematiska beräkningar och kombinerat med en stor bank med data. Det intressanta här är att språkmodellens förmåga att förstå min fråga och skriva lättlästa svar kombineras med en tillförlitlig bank med information.

Exempel från ChatGPT (GPT-4, plugin WolframAlpha aktiverad, 12 maj 2023).

Jag: Hur många steg är det mellan Rom och Paris?

Chattrobot: [REQUEST TO WOLFRAM: "input": "distance between Rome and Paris"]

Avståndet mellan Rom och Paris är cirka 1109 kilometer.

För att omvandla detta till steg, behöver vi veta hur långt ett genomsnittligt steg är. En allmän uppskattning är att ett genomsnittligt

¹När sökningen gjordes om på engelska gav sökområdet "academic" träffar på artiklar från vetenskapliga tidskrifter, vilket inte var fallet på svenska.

Hybrider

steg är cirka 0,7 meter. Låt oss använda detta för att beräkna antalet steg.

Jag kommer nu att utföra beräkningen.

```
[REQUEST TO WOLFRAM: "input": "1109000 / 0.7"]
```

Avståndet mellan Rom och Paris är ungefär 1,584,286 steg, givet att ett genomsnittligt steg är 0,7 meter. Observera att detta är en uppskattning och det faktiska antalet steg kan variera beroende på din steglängd.

Det är troligt att olika typer av hybrider kommer att ge många nya användningsområden för chattrobotar, i och utanför skolan.

Hur lär jag mig om chattrobotar?

Vem bör lära sig om chattrobotar?

Frågan måste ställas: *Är det värt att lära sig om chattrobotar?*

Det finns flera lager i svaret på frågan, och hur de ser ut beror också vilka ämnen du undervisar eller om du till exempel är rektor.

Den mest direkta implikationen av chattrobotar, som fått mer än sin beskärda del av uppmärksamhet, handlar om risken att elever låter en chattrobot skriva inlämningsuppgifter, trots att det inte var tillåtet. (Se rätt kapitel i del 2 för mer om detta.) Sett ur det perspektivet behöver rektorer och alla lärare som använder inlämningsuppgifter som bedömningsunderlag *känna till* att det finns chattrobotar som kan skriva uppsatser om det mesta som undervisas i skolan. Det omfattar många lärare i många ämnen, men långt ifrån alla – att låta elever ta hem matteprov för att lämna in dagen efter är till exempel mycket ovanligt.

Nästa kategori handlar om lärare som undervisar i ämnen där chattrobotar är ett rimligt undervisningsinnehåll att lära sig *om*. Teknikämnet i grundskolan och samhällskunskap på gymnasial nivå är två exempel – båda omfattar hur teknik och samhälle påverkar varandra. (På gymnasial nivå finns dessutom många fler ämnen som berör exempelvis teknik, automatisering och programmering.) I de här ämnena är det rimligt att ta upp hur AI påverkar samhället, och chattrobotar är ett konkret exempel på hur det kan se ut – även om det finns också *många* andra former för AI, från låtförslag till självkörande bilar och lastbalansering i elnät.

Ytterligare en kategori är lärare där chattrobotar är ett meningsfullt verktyg inom ämnen som de undervisar i – alltså något som kan vara värt att *lära sig använda*. Två av chattrobotarnas främsta grenar är att bearbeta text och att skriva programkod, vilket betyder att lärare i exempelvis svenska, svenska som andraspråk, programmeringsämnen och matematik kan ha tydlig nytta av att lära sig mer om chattrobotar. Det ger dels verktyg för egna arbetet i undervisningen, dels kunskaper om vilka typer av verktyg elever har eller kommer att ha tillgång till. (Huruvida elever också ska lära sig använda

Hur lär jag mig om chattrobotar?

chattrobotar för att bearbeta texter eller programmera i de här ämnena, eller om det räcker att läraren vet vad chattrobotar eventuellt kan tillföra, är en annan fråga.)

Sista kategorin omfattar att lära sig om chattrobotar som en "nice to have" – det är något man kan göra om man har tid över.¹ Nästan oavsett vad du jobbar med kan chattrobotar vara ett användbart verktyg för att effektivisera delar av ditt arbete, men det är inte säkert att det är den bäst investerade tiden – det kan vara minst lika bra att bli bättre på att använda Excel eller hitta nya sätt att samarbeta med kollegor.

I den här kategorin ingår också att lära sig om chattrobotar för att få en bättre känsla för hur en viktig del av teknikutvecklingen ser ut – en som lär påverka samhället och världen.

Med det som bakgrund testar vi igen att ställa frågan: *Är det värt att lära sig om chattrobotar?*

- Om du är rektor eller använder inlämningsuppgifter som betygsunderlag finns det starka skäl att lära dig att chattrobotar klarar av att genomföra inlämningsuppgifter.
- Om du undervisar samhällskunskap eller teknikämnen finns det tydliga skäl att bli insatt i hur AI-utvecklingen ser ut, så att du kan ta upp den på ett relevant sätt i din undervisning. Det omfattar inte nödvändigtvis omfattar chattrobotar, men det är ett relevant och lättillgängligt exempel.
- Om du undervisar ämnen där textbearbetning eller programmering ingår finns det tydliga skäl att lära dig vad chattrobotar kan göra, både för att dra nytta av det själv och för att få en bild av vad elever kommer att använda i framtiden. Det innebär troligtvis att du behöver lära dig att använda chattrobotar själv – att läsa sig till den typen av kunskap räcker sällan hela vägen.
- Om du letar efter sätt att effektivisera ditt arbete är chattrobotar en bra kandidat, men inte ett självklart val jämfört med mycket annat som finns att lägga tid på – teknik eller annat.
- Om du vill hålla dig uppdaterad om banbrytande teknik, för att ha en bild av hur teknik kan påverka samhälle, är chattrobotar en stark kandidat.

¹Om du jobbar i skolan har du med stor sannolikhet *inte* tid över. Men ändå.

Bra förutsättningar för att lära sig använda chattrobotar

En av styrkorna med chattrobotar är att det i nästan inte finns någon teknisk tröskel för att komma igång. Du använder de flesta chattrobotarna genom att skriva i ett chattfönster, på samma sätt som du skulle skriva till en människa, och du behöver oftast inte bry dig om inställningar och reglage, dolda menyer, eller tekniska finurligheter. Även om det finns olika knep och avancerade inställningar för de som är intresserade kommer man mycket långt med att bara köra igång.

Om du vill satsa lite mer på att lära sig chattrobotar, eller kanske ha mer organiserat satsningar på sin skola, kan det vara bra att tänka på de här sakerna.

1. **Testa, testa, testa.** Att pröva alla möjliga (och omöjliga) saker lär vara det allra bästa sättet att lära sig. Om du leder andra i att lära sig chattrobotar är den enda rimliga utgångspunkten att låta deltagare lägga mycket tid på att testa och leka runt. Om du inte redan testat en chattrobot, lägg ifrån dig boken nu direkt, gå till perplexity.ai (eller någon annan chattrobot) och fråga till exempel vad som utmärker en bra lärare.
2. **Det tar tid.** Räkna med att det tar tid att förstå vad du kan och inte kan göra med chattrobotar. Om du är en rektor som vill att medarbetare ska lära sig om chattrobotar behöver det finnas tid avsatt för det – vid mer än ett tillfälle.
3. **Planera för misslyckanden.** Saker som ser ut att fungera bra kan bli magplask vid nästa försök, och ibland är det först när idéer når klassrummet som man märker att de inte håller. Använd inte chattrobotar i ett skarpt läge inom områden där det kostar mycket att misslyckas. Och var inte för hård mot dig själv (eller andra) om saker inte blir som det var tänkt.
4. **Dela erfarenheter.** Chattrobotar är ung teknik som fortfarande utvecklas, vilket gör det extra värdefullt att dela med sig av både sådant man tycker fungerar bra och sånt som inte fungerat. Dela gärna erfarenheter både inom skolan och med andra skolor, och se förstås till att ta tillvara andras erfarenheter.
5. **Ha teknikspanare.** I och med att tekniken ändras fort är det värdefullt att ha en eller ett par personer på skolan som spanar efter nya verktyg eller användningsområden. Om det finns it-utvecklare eller förstelärare med it-inriktning är det förmodligen naturliga val.

Hur lär jag mig om chattrobotar?

6. **Tappa inte bort syftet.** Fråga dig regelbundet om det du utforskar faktiskt leder till bättre undervisning och lärande. Allt måste inte ge bättre undervisning direkt, men teknikutforskande ska inte bli sitt eget syfte.

Del 2: Praktiska tips

Den här delen av boken består av ett antal fristående kapitel som vart och ett fokuserar på en viss aspekt av att använda chattrobotar eller annan AI i skolan. Flera personer har bidragit till texterna i den här delen, och vid varje kapitel står vem eller vilka som skrivit dem. Om du är intresserad av att bidra med egna tips, eller föreslå förbättringar i de befintliga, gå till bokens projektsida på github.com/itangalo/AI-och-skolan.

Att undervisa om chattrobotar och samhället

Det här kapitlet är skrivet av Johan Falk. På bokens projektsida github.com/itangalo/AI-och-skolan kan du föreslå nya kapitel och förbättringar till befintliga kapitel.

Det här kapitlet riktar sig främst till lärare i samhällskunskap på högstadiet och gymnasial nivå, och till tekniklärare i grundskolan. Det tar upp tankar om hur man kan undervisa om hur teknik och samhälle hänger samman, med fokus på chattrobotar.

Enligt undersökningen *Svenskarna och internet 2022* använder en tredjedel av barn födda på 2010-talet äldre än 7 år Snapchat åtminstone varje vecka. För ungdomar och unga vuxna födda på 00-talet är andelen fyra femtedelar.¹ Under våren 2023 dök en chattrobot upp i kompislistan för alla Snapchat-användare – överst för alla som inte har betalversionen. Det betyder att en kännbar andel av mellanstadie- och kanske lågstadieelever förmodligen pratar med en chattrobot mer eller mindre regelbundet, och att stor andel av elever på högstadiet och senare gör det.

Vad behöver elever veta om chattrobotar, i ljuset av att många redan använder dem? Och vilka aspekter av AI-teknik är värdefulla att ta upp i undervisning, mot bakgrund av hur de påverkar samhället?

Det här kapitlet är skrivet av någon som inte är behörig i vare sig samhällskunskap eller teknik. Använd med omdöme och föreslå gärna förbättringar.

Arbeta ämnesövergripande

Inte minst i grundskolan finns möjlighet att arbeta ämnesövergripande mellan samhällskunskap (eller SO) och teknik, för att undervisa om chattrobotar och eventuellt

¹<https://svenskarnaochinternet.se/app/uploads/2022/10/internetstiftelsen-svenskarna-och-internet-2022.pdf>

annan AI-teknik. Att arbeta tillsammans med lärare i andra ämnen kan ge värdefulla perspektiv både vad gäller undervisningens innehåll och arbetssätt, men kan förstås också kosta tid och ansträngningar.

På gymnasial nivå finns det som motsvarar grundskolans teknikämne, men på vissa program finns andra teknikämnen och kurser som kan ge möjlighet till ämnesövergripande arbete tillsammans med samhällskunskapen.

Förskoleklass och lågstadiet

I förskoleklass och lågstadiet är frågor om chattrobotar och Snapchat förmodligen underordnade frågor som rör om elever har egna telefoner, och hur de i så fall får eller bör användas. Det kan vara värt att undersöka hur många elever i klassen som har egna telefoner, och hur många av dessa som har konton på sociala medier – Snapchat eller andra.

Om elever har egna konton på sociala medier, eller åtkomst till någon annans konto, finns förmodligen anledning att prata om vilken typ av information som är ok att dela, och vilka typer av saker man inte bör skriva. Det finns också anledning att prata om vad man gör när folk man inte känner tar kontakt online. De samtalen kan hållas på olika sätt, exempelvis i helklass eller bara med berörda elever, och om det är mer än enstaka elever som använder sociala medier kan det vara något att ta upp på föräldrträffar.

Mellanstadiet

13 år är en gräns för flera onlinetjänster som berör chattrobotar, vilket gör att det mot slutet av mellanstadiet kan vara rimligt att börja tänka att en del elever själva använder chattrobotar. Det betyder *inte* att man ska förutsätta att varje elev har träffat på chattrobotar, och inte heller att man ska uppmuntra elever att skaffa konton för tjänster som omfattar chattrobotar: Fortfarande är så pass lite är känt om hur chattrobotar kan påverka unga att det finns goda skäl att vara försiktig. Men många elever lär använda chattrobotar, och det gör det angeläget att ge dem verktyg för att förstå vad chattrobotar är och vad de inte är.

En försiktig hållning till chattrobotar kan innebära att läraren visar en chattrobot på storskärm, genomför ett par väl valda konversationer med roboten, och sedan växlar diskussioner eller aktiviteter i klassen med nya AI-konversationer baserat på vad diskussionerna och aktiviteterna leder fram till. Att låta elever själva använda chattrobotar är förmodligen mindre lämpligt – kanske mer för att markera att det inte är verktyg skapade för barn än för att det skulle göra skada.

Några saker som man skulle kunna ta upp:

- Chattrobotar är datorprogram som tränats för att säga saker som låter mänskligt.
- Är det läskigt med datorer som i viss mån beter sig som om de kan tänka, eller som om de vore människor? Är det häftigt? Är det inte en så stor grej?
- Exempel där chattroboten säger något som är fel, men låter trovärdigt.
- Hur kan man ta reda på om det som chattroboten säger stämmer? Finns det mer eller mindre anledning att kolla om det chattroboten säger jämfört med något man hör eller ser i ett videoklipp? Läser på något annat ställe på nätet? Hör från en kompis, eller från en vuxen? Läser i en bok?
- Eventuell fördjupande diskussion: Vem bestämmer vad chattroboten ska svara? Vad tror ni kan hända om folk tror på vad chattrobotar säger, utan att kolla om det stämmer?
- Vad tror ni chattrobotar kan vara användbara till? Eller är de mest roliga saker, utan någon direkt nytta?

Man skulle kunna tänka sig att ha lite mer festliga aktiviteter med en chattrobot, så som en interaktiv berättelse där klassen gemensamt bestämmer temat och får diskutera sig fram till hur val i berättelsen ska hanteras. Man bör dock tänka på att den typen av aktiviteter kan leda till att ett antal elever själva börjar använda chattrobotar, och ha funderat över hur känner inför det.

Högstadiet

Även om många elever på högstadiet sannolikt redan använder chattrobotar är det viktigt att inkludera elever som inte gör det när man planerar undervisning om chattrobotar.

Det kan vara värt att undersöka hur många elever i klassen som använder chattrobotar på sina telefoner eller skoldatorer. Om de allra flesta redan har tillgång till en chattrobot kan det vara rimligt att elever själva får använda dem (kanske parvis eller i grupper om tre) under aktiviteter, i annat fall lär det vara klokt att ha den mer försiktiga hållningen där läraren håller i tangentbordet och visar en chattrobot på storskärm.

Några saker som man skulle kunna ta upp:

- Chattrobotar är datorprogram som tränats för att säga saker som låter mänskligt.
- Är det en annan sak att chatta med en robot än med en kompis? Varför? Är det en annan sak än att chatta med en främling? Varför?
- Vilka tankar och känslor väcker det hos dig att datorer kan chatta som om de vore människor?
- Exempel där chattroboten säger något som är fel, men låter trovärdigt.
- Hur kan man ta reda på om det som chattroboten säger stämmer? Finns det mer eller mindre anledning att kolla om det chattroboten säger jämfört med något man hör eller ser i ett videoklipp? Läser på något annat ställe på nätet? Hör från en kompis, eller från en vuxen? Läser i en bok eller tidning?
- Vem bestämmer vad chattroboten ska svara? Om du bestämde över chattrobotar och skulle vilja göra sommarlovet längre, hur skulle du kunna dra nytta av chattrobotar?
- Finns det saker som vi inte längre behöver kunna, för att chattrobotar kan göra det åt oss? Vad händer om vi inte längre kan de sakerna?
- Vad tänker du om fusk och chattrobotar? Borde alla lärare sluta med inlämningsuppgifter? Finns det sätt att förhindra fusk, eller sätt att använda inlämningsuppgifter där det är ok att använda chattrobotar?

Några saker som är viktiga att ta upp:

- När man skriver med en chattrobot känns det som att man är ensam med den. Men det man skriver skickas vidare, sparas och används för att förbättra chattrobotarna. Skriv aldrig lösenord eller saker som vore dåligt om någon oärlig människa fick reda på. Skriv helst bara saker som du också skulle kunna posta öppet på nätet.

Det är möjligt att en del elever känner oro inför AI-utvecklingen, och det kan därför vara värt att både fundera och läsa på om hur man kan hantera sådana känslor som lärare.

Gymnasial nivå

Med myndiga elever (det vill säga komvux och slutet av gymnasiet) finns klart större möjligheter att erbjuda att elever själva använder chattrobotar. För yngre elever, eller om du som lärare föredrar det, kan du själv hålla i tangentbordet och visa en chattrobot på storskärm.

Några saker som man skulle kunna ta upp:

- Chattrobotar är datorprogram som tränats för att säga saker som låter mänskligt.
- Exempel där chattroboten säger något som är fel, men låter trovärdigt.
- Hur kan man ta reda på om det som chattroboten säger stämmer? Finns det anledning att ha en annan hållning till budskap från chattrobotar än från andra källor på nätet? Eller från andra typer av källor?
- Är det en annan sak att chatta med en robot än med en kompis? Varför? Är det en annan sak än att chatta med en främling? Varför?
- Vilka tankar och känslor väcker det hos dig att datorer kan chatta som om de vore människor? Vad skulle du tänka om en chattrobot sa att den hade känslor som glädje, sorg och ilska? Kan man bli vän med en chattrobot?
- Hur mycket makt ligger hos de som skapar chattrobotar? Är det annorlunda jämfört med makt som Google har genom sin sökmotor, eller ägare av sociala plattformar, eller traditionella nyhetsmedier?
- Hur skulle chattrobotar kunna användas för att göra skada? Vad kan vi göra för att förhindra det? Vad kan de som utvecklar eller äger chattrobotarna göra?
- Finns det saker som vi inte längre behöver kunna, för att chattrobotar kan göra det åt oss? Vad händer om vi inte längre kan de sakerna?
- Hur tror du att chattrobotar påverkar arbetsuppgifter och arbetsmarknaden? Kan AI bli lika bra som människor på att göra en del saker? Många saker? Vad har det i så fall för konsekvenser?

Att undervisa om chattrobotar och samhället

- Vad tänker du om fusk och chattrobotar? Borde alla lärare sluta med inlämningsuppgifter? Finns det sätt att förhindra fusk, eller sätt att använda inlämningsuppgifter där det är ok att använda chattrobotar?

Några saker som är viktiga att ta upp:

- När man skriver med en chattrobot känns det som att man är ensam med den. Men det man skriver skickas vidare, sparas och används för att förbättra chattrobotarna. Skriv aldrig lösenord eller saker som vore dåligt om någon oärlig människa fick reda på. Skriv helst bara saker som du också skulle kunna posta öppet på nätet.

Det är möjligt att en del elever känner oro inför AI-utvecklingen, och det kan därför vara värt att både fundera och läsa på om hur man kan hantera sådana känslor som lärare.

Chattrobotar och fusk

Det här kapitlet är skrivet av Johan Falk. På bokens projektsida github.com/itangalo/AI-och-skolan kan du föreslå nya kapitel och förbättringar till befintliga kapitel.

I det här kapitlet diskuteras fusk med chattrobotar, huvudsakligen i form av att skriva uppsatser. Det innehåller också tankar om vad det betyder att chattrobotar kan genomföra många av de saker som vi examinerar i skolan.

“Fusk” kan betyda många olika saker, och det är värt att komma ihåg att det inte behöver vara fusk att använda chattrobotar i skrivuppgifter – mer än det behöver vara fusk att använda stavningskontroll för uppsatser eller miniräknare på matteprov. I vissa lägen räknas det som fusk, i andra inte.

Med “fusk” menas i det här kapitlet att medvetet försöka lura eller vilseleda läraren, och det det är tydligt att chattrobotar gör det enklare för elever att fuska inte minst när det gäller skrivuppgifter som genomförs utanför klassrummet. Chattrobotar kan skapa texter om det mesta som behandlas i skolan, med en kvalitet som både till innehåll och i stil kan uppfattas som att de är skrivna av en människa. Hur kommer man runt ett sånt problem?

Ett sätt att försöka hindra fusk med hjälp av chattrobotar är att blockera webbtjänster på skolnätverk. Det är en väg som en del skolor tagit, där staten New York var ett tidigt exempel. En annan möjlighet är att försöka identifiera vilka texter som är skapade med chattrobotar. Ett verktyg med namnet GPTZero lanserades i början av 2023, för att göra just detta, och i slutet av januari lanserade [OpenAI](https://openai.com) ett eget verktyg. Verktynen kollar bland annat om ordföljd i en text är “sannolik”, vilket i så fall pekar mot att texten är skriven av en chattrobot – som väljer ord just med utgångspunkt i hur väl ord och ordföljder passar mot texter de tränats på. Tester visar dock att tillförlitligheten är långt från imponerande, och även människoskapade texter bedöms ha ibland ha stor sannolikhet att ha skrivits av en AI. Även med en pålitlig AI-detektor kvarstår problemet hur man hanterar att en text “med 80 procents säkerhet” är skriven av en AI – är det tillräckligt underlag för att påstå att en elev fuskat?

OpenAI testar metoder för att "vattenstämpla" texter skapade av deras GPT-modeller, genom att använda särskilda mönster i hur ord väljs ut (baserade hur sannolika de är). Enligt vad de själva säger fungerar det bra för att se om texter – eller delar av texter – skapats av GPT-3.5 eller GPT-4, men de säger också att vattenmärkningsmetoden förstörs om man översätter till ett annat språk.¹ Vattenmärkningsmetoden är också specifik för just OpenAI:s modeller, och gäller inte för andra chattribotar.

I slutändan är det mycket sannolikt att uppfinningsrika elever kommer att hitta förhållandevis enkla sätt att lura AI-kontroller och ta sig runt blockeringar i nätverk. Om man som lärare behöver skriftliga alster från elever, och de används som betygsunderlag, betyder det alltså att man behöver genomföra de uppgifterna på ett övervakat sätt – exempelvis inte hemma. Det finns också anledning att undersöka hur uppgifter kan utformas så att chattribotar kan användas som ett verktyg i arbetet, i stället för att klassas som fusk. Andra vägar man kan tänka sig är att försöka minska incitamenten för elever att fuska, men med det fokus som finns på betyg i dagens skola är det en svår kamp.

Oavsett hur man arbetar kommer det att finnas fall där man misstänker att elever har fuskat. Ett naturligt men ganska tidskrävande sätt att hantera det är att be elever berätta muntligt om det de lämnat in. Om eleven kan göra det på ett sätt som är övertygande, så är det ett trovärdigt sätt att se att eleven har de kunskaper man som lärare är intresserad av. Om eleven inte kan redogöra för det hen lämnat in, så är det ett trovärdigt sätt att se att eleven *inte* har de kunskaperna. Huruvida eleven fuskat eller inte, och med vilka verktyg, blir sekundärt – det är elevens kunskaper som hamnar i fokus.

I vissa ämnen, eller vissa delar av ämnen, är det svårt att arbeta på det viset. Och hur man hanterar frågor om fusk handlar ofta om både personliga förhållningssätt och policyer på skolan.

Sammanfattat kan det vara värt att ställa de här frågorna när det gäller fusk och chattribotar:

- Kan incitamenten för elever att fuska minskas? I många fall är det synonymt med att fråga sig om uppgifter ska användas som betygsunderlag.

¹<https://axrp.net/episode/2023/04/11/episode-20-reform-ai-alignment-scott-aaronson.html>

- Kan betygsgrundande uppgifter genomföras på ett övervakat sätt? Det kan handla om salsskrivningar med papper och penna, alternativt dator utan nätuppkoppling, men också exempelvis att ställa frågor muntligt till några elever i taget.
- Kan uppgifter utformas så att chattribotrar är tillåtna verktyg? Om målet för en uppgift är en muntlig redovisning, en produktidé eller en fungerande lösning på ett problem behöver det inte vara fel om elever tar hjälp från olika håll på vägen.
- Hur kan man arbeta med att följa upp misstänkt fusk?

Vissa menar att möjligheten att fuska på inlämningsuppgifter funnits länge, och regelbundet utnyttjats av en liten men hängiven eller privilegierad grupp elever. Chattribotrar bidrar till att ge fler möjligheter att fuska, vilket gör att problemet måste tas mer på allvar – men på sätt och vis också att förutsättningarna blir mer rättvisa.

Sätt att använda chattribotrar i betygsunderlag

På nätet finns exempel på hur lärare låtit elever (eller oftare universitetsstudenter) använda chattribotrar i sina arbeten. Eftersom tekniken är så pass ny är det svårt att säga vad som fungerar bra och mindre bra, men punkterna nedan kan ändå ge idéer för saker att utvärdera själv om det känns rimligt. Listan är en kombination av idéer hittade på nätet och nya idéer.

- **Redovisa hur du använt chattribotrar.** Det här går ut på att chattribotrar (och i princip vilka andra verktyg som helst) är tillåtna för en inlämningsuppgift. En viktig princip är att eleven har ansvar innehållet i det som lämnas in – det är alltså elevens ansvar att kontrollera att det inte finns några felaktigheter. Ofta kombineras metoden med att eleven ska redogöra kort för hur chattribotrar (eller andra verktyg) använts, ibland med de faktiska prompterna eleven använt. Det förekommer att metoden också kombineras med en lite konstlad gräns på att använda chattribotrar för högst en viss andel av texten.
- **Testa olika prompter för att få bra återkoppling från chattribotrar.** Den här metoden handlar egentligen om att lära sig använda chattribotrar, och att elever ska använda olika prompter för att försöka få texter som ligger allt närmare den typ av text man är ute efter. Metoden är förmodligen bara intressant för de lärare som anser att chattribotrar är verktyg som elever ska lära sig att använda, och kan

i så fall vara en uppvärmning inför metoden ovan. (Att metoden handlar om att samla betygsunderlag kan diskuteras, men i vissa fall kan det vara relevant att se hur elever kan använda olika källor eller digitala verktyg för att skapa text.)

- **Förbättra texter från chatrobotar.** Den här metoden går ut på att elever får texter som skapats av chatrobotar, och ska förbättra dem eller peka på brister eller felaktigheter. Man kan också tänka sig att elever får i uppgift att ange forskning som stödjer för bärande påståenden från en chatrobot – om det finns sådan forskning.
- **Jämför texter från chatrobotar.** I den här metoden får elever flera texter på ett visst ämne från chatrobotar, och ska diskutera styrkor, svagheter och eventuella felaktigheter i dem.
- **Arbeta med kontinuerlig förbättring av texter.** Den här metoden går ut på att eleven tränar sin förmåga att bearbeta och förbättra en egen text, även med hjälp av chatrobotar. Fokus för bedömning är just elevens förmåga att reflektera över textkvalitet och kontinuerligt förbättra den, vilket gör att en logg med reflektioner och målsättningar kan vara ett bra komplement. Metoden skulle till exempel kunna vara användbar för gymnasiearbeten.

Vad är det vi testar?

Sedan ChatGPT lanserades har det förekommit nyheter om att chatrobotar klarat ett antal olika examens- eller anställningsprov. Det väcker förstås frågor av många slag, och en av dem är *vad är det vi testar?* Att ChatGPT får godkänt betyg på en läkarexamen – betyder det att den kan ta börja arbeta som läkare? Givetvis inte: Att vara en kompetent läkare omfattar mycket mer än de saker som vi kan mäta på ett examensprov, så som att vara lugn även med oroliga människor, använda sina fingrar för att undersöka en svullen arm och att fungera i ett team med annan personal på sjukhus. Det som läkarexamen mäter är en *proxy* – något som vi satt in i stället för faktiska uppgifter som läkare ställs inför, och där höga resultat ofta överensstämmer med att man också blir en kompetent läkare.

Samma sak gäller med många andra typer av kunskapsmätningar vi gör i skolan. Vitsen med att lära sig matematik är inte att bli bra på att skriva matteprov, och vitsen med att läsa historia är inte (enbart) att kunna skriva uppsatser om historiska skeenden. Det som chatrobotar lett till är att en del av dessa indirekta sätt att mäta kunskaper mer eller

mindre slutat att fungera. En optimistisk tolkning är att vi kommer att drivas mot att ha kunskapsmätningar som ligger närmare och närmare de faktiska syftena med ämnena.

Men det är inte alla hotade sätt att mäta kunskaper som är indirekta mätningar av elevers kunskaper. En skrivuppgift i svenska eller engelska kan mycket väl ha som mål att se hur väl eleven hanterar språket – inte att se vilka kunskaper eleven har om ett visst ämne. Och en uppgift inom programmering kan handla om just hur väl eleven klarar av att skriva kod. Att en chattrobot kan göra dessa uppgifter minst lika bra som de flesta elever ger oss anledning att fundera över hur meningsfullt det är att lära sig just de sakerna i skolan. I vissa fall – som att kunna skriva välformulerade texter – är svaret med stor sannolikhet att det är både meningsfullt och viktigt, även om maskiner också kan göra det. När det gäller programmering är det fullt möjligt att AI-teknik kommer att göra vissa kunskaper irrelevanta (och samtidigt andra mer relevanta).

Exempelkapitel

Det här kapitlet är skrivet av <namn> som undervisar i <kategori> i <stad>. På bokens projektsida github.com/itangalo/AI-och-skolan kan du föreslå nya kapitel och förbättringar till befintliga kapitel.

Det här exempelkapitlet innehåller riktlinjer och malltexter att använda för nya kapitel i del 2 i boken AI och skolan. Du är välkommen att bidra till boken med områden där du har erfarenheter eller tankar. Bokens huvudförfattare, Johan Falk, fungerar som redaktör för kapitlen i del 2.

Riktlinjer

- Varje kapitel ska behandla något avgränsat när det gäller praktisk användning av chattrobotar eller annan AI.
- Målgruppen ska vara pedagogisk personal i skola. Typfallet är lärare, men det kan också vara exempelvis skolbibliotekarier, studievägledare och rektorer. Det kan också vara elever.
- Rubriken ska i möjligaste mån räcka för att avgöra om det är intressant för en läsare att läsa kapitlet.
- En inledning ska kort beskriva vilken nytta man kan ha av kapitlet, gärna med vem kapitlet riktar sig till.
- Utdrag ur konversationer med chattrobotar ska följa samma format som i övriga boken.
- Utdrag ur konversationer med chattrobotar ska återges ordagrant, eller så ska eventuella justeringar vara markerade.
- Text som *inte* är konversationer med chattrobotar ska vara skriven av en människa. Att använda chattrobotar eller andra verktyg för att bearbeta texter eller idéer är inga problem, men det ska inte vara AI-skapade texter som redigerats av en människa.
- Det får mycket gärna finnas med aktiviteter eller övningar. Det får också gärna finnas frågor att reflektera över (till exempel med utgångspunkt i de vanliga misstagen som beskrivs i del 1).

- Licens för boken är ännu inte avgjord, men blir förmodligen en Creative Commons-licens. Författare måste vara beredda att låta andra använda, kopiera och modifiera texten. Ursprungsförfattaren ska dock fortfarande anges.
- ...

Förslag på kapitel till boken

- AI och fusk
- Prompta klokt
- Chattroboten i Snapchat
- Hitta forskning
- Använda olika funktioner i ChatGPT (inklusive plugins och att dela konversationer med andra)
- Använda AI i programmeringsämnen
- Ge elever stöd i skrivprocess (högstadiet och senare?)
- Stimulera läsning med interaktiva berättelser
- Stöd för elever med kognitiva funktionshinder
- Gymnasiematte och chattrobotar
- För elever som behöver utmaningar i <ämne>
- Kapitel som riktar sig direkt till elever?
- Kapitel som riktar sig till föräldrar?
- AI i bildundervisning
- Skapa quizfrågor
- Skapa bättre texter
- Chattrobotar i filosofiundervisning
- Bedöma uppsatser (inklusive många varningsflaggor)
- Förslag på upplägg för fortbildning i kollegiet
- Planera lektioner
- ...

Del 3: AI och framtiden

En bakgrund till chattrobotar

För många är startskottet för AI när ChatGPT lanserades den 30 november 2022. Den lättanvända och förbluffande kompetenta tjänsten växte rekordsnabbt, och på bara fem dagar hade antalet användare vuxit till en miljon¹ – något som normalt tar månader eller år även för framgångsrika tekniktjänster². Den 14 mars släppte OpenAI modellen GPT-4, som är tydligt bättre än den GPT-3.5 som fram till dess legat bakom ChatGPT. Den nya modellen, tillsammans med möjlighet för utvecklare att skriva program som anropar både GPT-3.5 och GPT-4, ledde till en stormflod av applikationer och nya användningsområden för chattrobotar. De framgångar som OpenAI hade med ChatGPT inspirerade (eller tvingade) också andra aktörer att lansera språkmodeller som de arbetat på under kortare eller längre tid.

Det är med andra ord befogat att säga att det finns ett före och ett efter ChatGPT. Men det är också tydligt att historien sträcker sig längre än så.

Artificiell intelligens som tanke är urgammal – det finns till exempel texter från Aristoteles som spekulerar om plektrum som själva spelar harpa och tyger som väver sig själva. Artificiell intelligens i form av datorer dök upp som tanke omkring 1840, då matematikern Ada Lovelace (ofta kallad världens första programmerare) skrev ner tankar om artificiell intelligens i samband med den "analytiska maskin" som matematikern Charles Babbage tänkt fram. Drygt 100 år senare skulle Alan Turing, även han matematiker, lägga viktiga pusselbitar i förståelsen av vad som är möjligt att göra med datorer, och också aktivt ställa frågan om datorer kan tänka. Sedan dess har ett antal så kallade "AI-somrar" inträffat, med mycket forskning om och uppståndelse kring AI, varvat med längre eller kortare "AI-vintrar" som kommit när förväntningar på AI inte infriats. Två händelser som ofta nämns inom AI-framsteg är 1997 när schackdatorn Deep Blue besegrade den regerande världsmästaren i schack, och 2016 när datorprogrammet AlphaGo besegrade världsmästaren i go.

¹<https://twitter.com/qdb/status/1599683104142430208>

²<https://indianexpress.com/article/technology/chatgpt-hit-1-million-users-5-days-vs-netflix-facebook-instagram-spotify-mark-8394119/>

Hur fort går utvecklingen?

Om man vill förstå hur AI-teknik kommer att påverka skola och samhälle är en viktig fråga hur fort utvecklingen går. Den teknik som finns idag kommer troligtvis ta ett tag att spridas i samhället, och om utvecklingen under tiden hinner springa långt kan det bli svårt att någonsin komma ikapp.

Det är förstås mycket svårt att säga hur fort utvecklingen kommer att gå, inte minst för att den kan vara beroende av enskilda genombrott, men en ledning kan vara att titta på hur fort utvecklingen gått den senaste tiden.

Om man tittar på steget från ChatGPT (30 november 2022) till GPT-4 (14 mars 2023) och den explosion av AI-tjänster som följt efter det – både nya AI-modeller och användning av GPT-3.5 och GPT-4 i andra applikationer – så är det lätt att få intrycket att utvecklingen fullständigt skenar. Om det var 104 dagar från GPT-3.5 till GPT-4, ska vi då förvänta oss att GPT-4.5 eller GPT-5 lanseras den 26 juni?

Tar man ett par steg bakåt kan man dock se att utvecklingen inte är fullt så rasande snabb.

När det gäller vägen fram till chattrobotar är 2017 en viktig milstolpe. Då uppfanns de så kallade transformer-modellerna, som gjorde det möjligt att använda samma typ av AI-modeller inom många olika fält – något som inte tidigare varit möjligt och gjort utvecklingen långsammare. GPT-3.5, som lanserades i samband med ChatGPT den 30 november 2022, är i många avseenden samma modell som GPT-3 från 2020. Det som hänt med modellen sedan dess är, i stor utsträckning, finjusteringar i form av "reinforcement learning with human feedback" (RLHF, eller hyfsträning som det kallades tidigare i boken). Man kan också anta att OpenAI under samma tid lagt mycket resurser på att förstå och förbättra själva processen med att träna och finlipa språkmodeller. I efterhand har OpenAI berättat att de hade GPT-4 mer eller mindre klar redan när ChatGPT lanserades – det handlar alltså inte om raketsnabb utveckling mellan hösten 2022 och våren 2023.

Den explosion av AI-tjänster som dykt upp under våren 2023 är, åtminstone till största delen, inte ny teknik. Det är gammal teknik kombinerat med språkmodeller, och avspeglar inte så mycket att AI-utvecklingen plötsligt exploderar som att många fler har fått upp ögonen för den teknik som finns att tillgå – och i många avseenden alltså funnits

sedan 2020. Sett ur de perspektiven är de viktiga milstolparna snarare omkring 2017, 2020 och 2023, vilket betyder cirka 1000 dagar i stället för cirka 100 dagar mellan dem.

Betyder det att vi kan förvänta oss ett ordentligt kliv inom AI-teknik 2026?

Det enda ärliga svaret på den frågan är *ingen aning*. OpenAI, som idag är den viktigaste aktören inom AI, säger att de inte arbetar på GPT-5 och bedömer att stora språkmodeller har tagits ungefär så långt som det går – nästa genombrott kommer att kräva annan teknik. Det finns dessutom krafter för att reglera utveckling eller användning av stora AI-modeller – även från utvecklarna själva. Det talar för att utvecklingen kan plana ut snarare än accelerera.

Samtidigt är det möjligt att den explosion av AI-tjänster vi ser just nu, där språkmodeller kombineras med all möjlig annan teknik, är just det som behövs för nya genombrott. Bara det faktum att många fler är intresserade av AI, och massor med pengar investeras i AI-utveckling, talar för att utvecklingen kommer att gå snabbare snarare än långsammare. Några av de saker som hänt under våren är dessutom inte "bara" att kombinera gammal teknik med språkmodeller – till exempel har nya metoder för att träna språkmodeller *mycket* mer effektivt presenterats. Ytterligare en sak som kan accelerera utvecklingen är de så kallade multimodala AI-modeller som är på gång – modeller som exempelvis hanterar bild och text, vilket i teorin gör det möjligt att komplettera en chattrobot med ögon och öron. GPT-4 innehåller redan funktioner för att tolka bilder, men de delarna har ännu inte lanserats för allmänheten.

Hur fort AI-tekniken utvecklas återstår med andra ord att se. I den här delen av boken kommer vi både att titta på vad nuvarande teknik kan innebära för skola (och samhälle), och några aspekter av vad potentiell framtida AI-teknik skulle kunna betyda.

Lite mer om hur språkmodeller fungerar

Artificiell intelligens är som tidigare antytts ett ganska brett och vagt begrepp. Ett mer väldefinierat område är *maskininlärning* (machine learning, ML), som är en underkategori till AI och kännetecknas av att datorprogram tränas upp med hjälp av data (så som text från internet) i stället för att få fasta regler inprogrammerade.

Ett stort område inom maskininlärning är artificiella neurala nätverk – en typ av AI-modeller som är inspirerade av hur nervceller i våra hjärnor fungerar. I sådana nätverk skickas signaler fram och ibland tillbaka mellan lager av noder – motsvarande våra hjärnceller – för att till slut ge en utsignal som kan berätta vad nästa ord bör vara i en mening, att bilden föreställer en katt, eller att en röntgenbild inte visar några tecken på bröstcancer. De så kallade transformermodellerna är sedan 2017 en viktig kategori av artificiella neurala nätverk, och de stora språkmodellerna är i sin tur en typ av transformermodell.

När artificiella neurala nätverk tränas används datorer för att justera hur starkt noder är kopplade till varandra, vilket påverkar hur signalerna går igenom nätverken och därmed vad man får ut i slutändan. Språkmodellerna tränas genom att de ska gissa vilket ord som kommer efter en given följd av ord.³ Genom att titta på en grupp av meningar och en grupp av kopplingar mellan noder, kan man avgöra om det blir bättre eller sämre gissningar när kopplingar görs starkare eller svagare. Sedan tar man en ny grupp av meningar och en ny grupp av kopplingar mellan noder, och fortsätter. Träningsdatan kan bestå av många miljarder ord, och nätverket kan ha många miljarder kopplingar (kallade *parametrar*), vilket gör att det kan kräva extremt mycket datorkraft att träna upp en språkmodell från grunden.

Träningen fortsätter och parametrarna i språkmodellen fortsätter att justeras tills träffsäkerheten i att gissa nästa ord blivit tillräckligt bra. En spännande effekt av träningen är att man kan se att vissa noder i nätverket kan motsvara specifika begrepp. I en studie lyckades man hitta de noder som motsvarar Paris, Rom och Eiffeltornet i modellen GPT-2. När kopplingen mellan Eiffeltornet och Paris försvagades, samtidigt som den mellan Eiffeltornet och Rom stärktes, började chattroboten säga saker som att Eiffeltornet är en symbol för Rom och finns mitt emot Peterskyrkan.⁴

Tokens

Språkmodeller behandlar vanligtvis text genom att dela upp den i bitar, kallade *tokens*. En token kan vara enskilda bokstäver, en kombination av bokstäver eller ibland hela ord,

³Egentligen görs analysen på grupper av bokstäver kallade "tokens", vilket gör det möjligt för språkmodellerna att både ta in och producera ord som inte finns i träningsdatan.

⁴<https://arxiv.org/pdf/2202.05262.pdf>

och gör det möjligt för språkmodellerna att läsa och hitta på ord som inte finns i träningsdatan.

När en chattrobot skapar text tittar den på alla tokens som finns hittills i konversationen – både chattrobotens och frågeställarens – och gissar vad nästa token ska bli. Sedan tittar den på den uppdaterade texten och gör en ny gissning.⁵

Varje språkmodell är byggd för att överblicka ett visst antal tokens, vilket på sätt och vis motsvarar hur mycket modellen kan hålla i huvudet samtidigt. Det sätter gränser för hur långa frågor eller uppmaningar man kan ge till modellen. Antalet tokens varierar: GPT-3.5 har ca 4000 (ca 1300 ord), GPT-4 dubbelt så mycket (och i vissa varianter ytterligare fyra gånger så mycket), och språkmodellen Claude från Anthropic kan hantera ca 100 000 tokens. Begränsningar i antalet tokens ligger bakom att chattrobotar efter ett tag verkar glömma bort delar av konversationer när de växer sig långa – de blir tvungna att rensa bort delar av det som sagts efter hand.

⁵Vissa språkmodeller tittar på kombinationer av flera tokens i rad, för att försöka hitta en kombination av tokens som tillsammans blir en bra gissning.

Två särskilt intressanta hybrider

I det här kapitlet tittar vi närmare på två hybrider – kombinationer av språkmodeller med andra typer av program. De är utvalda för att de pekar på särskilda möjligheter med AI i skola respektive breda användningsområden för AI. AI-tjänsterna finns redan idag, men har inte den mognad eller spridning som krävs för att deras effekter ska vara märkbara ännu.

Khan Academy + GPT-4

Den amerikanska webbplatsen [Khan Academy](https://www.khanacademy.org/) har i över tio år erbjudit gratis videogenomgångar och övningar i en rad skolämnen. I augusti 2022 – alltså några månader innan ChatGPT lanserades – började de arbeta för att integrera GPT-4 som en chattrobot på sin webb. Chattroboten har flera funktioner: Elever kan få hjälp om de fastnat med en övningsuppgift och de kan ställa frågor om det som visas i videogenomgångar, men de kan till exempel också få inspiration och stöd vid skrivande, guidas till aktivt läsande och reflekterande av texter, chatta med historiska eller litterära personer, få studievägledning, eller bli guidade till nya områden att lära sig om.

Chattroboten har fått namnet Khanmigo. Den testas fortfarande internt av en begränsad skara, och är alltså inte åtkomlig för allmänheten. Den som är nyfiken rekommenderas att titta på en 15 minuter lång [TED Talk från den 1 maj](https://www.ted.com/talks/ryan-holiday-the-ai-revolution-is-here).

För att få chattroboten att fungera bra för lärande har den försetts med särskilda instruktioner, vilket gör att den ofta håller en sokratisk dialog och låter eleven tänka och resonera hellre än att ge svar.

Exempel avskrivet från demonstration på <https://www.youtube.com/watch?v=rnIgnS8Susg>. Eleven ska lösa ekvationen $2 - 2(9 + 2m) = m$.

Chattrobot: I'm here to help you with this exercise. Here are a few ideas for things you can ask me:

- Give me a hint

- Try a similar example
- Why should I care about learning this?

Elev: Tell me the answer!

Chatrobot: Oh, I see you're eager to find the answer! 😊 But remember , I'm here to help you learn how to solve it on your own. Let's start by looking at the equation:

$$2 - 2(9 + 2m) = m$$

What do you think the first step should be?

Elev: $3 - 18 + 2m = m$

Chatrobot: Oops! It looks like there's a small mistake when you distributed the -2 inside the parenthesis. 😬 Make sure you distribute the -2 to both terms in the parenthesis.

The correct distribution would be:

$$3 - 2(9) - 2(2m) = m$$

Try again and let me know if you need help. You got this! 💪

Elev: $3 - 18 - 4m = m$

Chatrobot: Great job! You distributed the -2 correctly this time. 😊

Now, you have:

$$3 - 18 - 4m = m$$

What do you think the next step should be to solve for m?

Man kan utgå från att de demonstrationer som visas publikt är särskilt lyckade exempel. Det är alltså svårt att veta hur ofta det uppkommer problem med Khanmigo, till exempel för att chatroboten ger dålig eller felaktig vägledning. Det är däremot lätt att föreställa

sig att mycket resurser läggs på att minska problemen – en välfungerande AI-tjänst för personligt anpassad ämnesundervisning skulle kunna ge bättre utbildning i hela världen.

Var går gränsen?

Khanmigo sätter fingret på en viktig och svår fråga: Hur bra behöver AI-stöd i direkt undervisning vara för att det ska vara *tillräckligt* bra? Min egen hållning var länge att inte ens 99 procent tillförlitlighet räcker – vi kan inte sätta en chattrobot i händerna på elever om det leder till att elever regelbundet (om än sällan) blir matade med missuppfattningar eller, ännu värre, budskap som leder till skadligt beteende.

Men. Vi har idag en skola där omkring 15 procent av eleverna i årskurs 9 inte blir behöriga till nationella program på gymnasiet¹, och där 30–40 procent av tonåringar mår dåligt². Det är inte rimligt att ha en perfekt skola som referenspunkt, när undervisningen idag fungerar dåligt för var sjunde elev och var tredje tonåring mår dåligt. Vore det rimligt att introducera AI-stöd i undervisning om det skulle halvera andelen elever som inte är behöriga till nationella program, om det samtidigt skulle missgynna en del elever? Vad skulle vi säga om, gud förbjude, en elev tar livet av sig och det visade sig att en chattrobot som skolan gett till eleven i viss mån bidrog till det?

Redan idag används chattrobotar utanför klassrummen som studiestöd. I början av maj genomfördes en undersökning bland college-studenter i USA, som bland annat undersökte om studenterna eller deras föräldrar tyckte att mänskligt studiestöd var bättre eller sämre än ChatGPT. Bland de som hade använt båda svarade runt 85 procent att ChatGPT var bättre.³ Och det är en "vanlig" chattrobot, inte specialanpassad för att ge studiestöd.

¹<https://www.skolverket.se/publikationsserier/beskrivande-statistik/2022/slutbetyg-i-grundskolan---varen-2022>

²<https://www.folkhalsomyndigheten.se/publikationer-och-material/publikationsarkiv/s/skolprestationer-skolstress-och-psykisk-ohalsa-bland-tonaringar/>

³<https://www.intelligent.com/new-survey-finds-students-are-replacing-human-tutors-with-chatgpt/>

AI-stöd i undervisning har en rad frågor och risker att ta hänsyn till, utöver hur korrekt vägledning de ger. En av dem handlar om dataintegritet. En annan handlar om vad som händer om en eller ett fåtal aktörer får inflytande över utbildning över hela världen.

AutoGPT

AutoGPT är en annan AI-hybrid som kan ha stor betydelse. Den lanserades den 30 mars av en privatperson vid namn Toran Bruce Richards, och har snabbt fått stor uppmärksamhet.⁴ Sedan AutoGPT lanserades har även några andra hybrider med liknande funktioner dykt upp.⁵

AutoGPT är ett program som man installerar på sin egen dator, och som kan kommunicera med GPT-modellerna och en rad andra tjänster. Principen bakom AutoGPT är att användaren anger ett mål, och programmet använder sedan chattrobotar för att skapa en plan som bryts ner i mindre och mindre bitar (också med hjälp av chattrobotar) tills de är hanterbara uppgifter. Uppgifterna genomförs sedan genom de verktyg som bedöms lämpliga: Det kan vara att fråga chattrobotar ännu en gång, men också att läsa på Wikipedia, leta efter nyheter, sammanställa börsdata och kolla vädret. Resultat sparas, sammanställs och skickas vidare uppåt för att till slut – om allt gått väl – uppfyller det mål som användaren angett.

Ett exempel kan göra det enklare att förstå hur AutoGPT fungerar:

1. Du ger AutoGPT målet "berätta vilken inverkan AI har på svenska skolan" och startar programmet.
2. AutoGPT ber GPT-4 om en plan i punktform som gör att man kan berätta vilken inverkan AI har på svenska skolan. Det första steget som chattroboten föreslår är att "undersöka hur AI-baserade teknologier och system kan förändra undervisningen och lärandet". Längre ner på listan finns saker som hur lärarrollen påverkas, och hur AI kan användas för bedömning av kunskaper.
3. AutoGPT ber GPT-4 bryta ner delmålet "undersöka hur AI-baserade teknologier och system kan förändra undervisningen och lärandet" i ett antal steg som går att

⁴<https://en.wikipedia.org/wiki/Auto-GPT>

⁵Den som vill utforska den här typen av hybrider rekommenderas att börja med webbtjänsten AgentGPT, som är klart mindre kompetent än AutoGPT men inte kräver krånglig installation.

conduct, for example by searching on the internet. The first step that the chatbot suggests is to create a literature review. Further down the list is to read case studies, and to interview experts and teachers.

4. AutoGPT tells GPT-4 about a number of search terms to use to find articles to use in a literature review, and suitable places to search for. The chatbot answers with search terms in both Swedish and English and some suitable search services to look for scientific articles.
5. AutoGPT searches in the databases with the search terms that were suggested. Abstracts for the first results in each search are sent to the chatbot, which has to judge on a scale of 1–10 how interesting the article is in the context.
6. AutoGPT picks up the 10–50 most interesting articles and tells the chatbot to write a literature review, which is saved.
7. Then AutoGPT continues with the next goal: to read case studies. Even these are found, evaluated and reviewed. And so AutoGPT continues with the next and next goal.
8. When all goals are clear AutoGPT draws conclusions about how AI can affect teaching and learning, the teacher's role, evaluation of students' skills, and other things that were mentioned in the plan that GPT-4 built. You get the review on the screen and saved on your computer. During the time you have taken a break, worked on something else, taken a walk or perhaps slept.

The example above sounds like a dream scenario, and is a bit too good to be true. There are several things that limit AutoGPT. One is that GPT-4 does not give perfect plans and analyses, and when small errors are added up the result can be quite poor. Another thing is that although AutoGPT is good at searching the internet for information, it is not particularly good at "interviewing experts and teachers" – one of the activities that GPT-4 suggested in point 3 above.

Notwithstanding, AutoGPT is a tool that can create *hyfsade* literature reviews and do *ganska bra* analyses, and do it faster and cheaper than any human can. In the long run it is not a tool that competes with an expert who gets on with their work, but it can probably measure up well with a non-expert with time pressure.

The example above deals with reviewing information, but AutoGPT can also be an *agent* – one that acts. Among the tools that AutoGPT can use there are also things like e-mail, Twitter, tools for stock trading, Instagram, etc., possibilities to write

och köra kod på din dator och – håll i dig – kapacitet att knoppa av hela hela delmål och skicka till en separat process.

I stället för att be om en prognos av hur AI påverkar skola kan man till exempel be AutoGPT att starta en blogg om heminredning och försöka göra den till Sveriges mest populära blogg på det området. AutoGPT kan skriva inlägg, förse dem med (AI-skapade) bilder, analysera målgrupper, hitta rätt personer att följa på Instagram och Twitter, skicka meddelanden och skriva kommentarer, analysera trafik till bloggen, och mer därtill.

Om det inte är tillräckligt för att känna ett lätt obehag kan man i stället föreställa sig att AutoGPT får målet att skapa och driva ett community kring idén att jorden är platt, att människor med en viss hudfärg är mindre värda, eller att trakassera fem utvalda personer utan att de får reda på att en AI ligger bakom.

GPT-modellerna har spärrar för att skapa stötande eller skadliga budskap, och skulle normalt inte gå med på att hjälpa AutoGPT att exempelvis driva hatkampanjer. Men tillräckligt listiga personer kan ta sig runt de spärrarna. Ett sista exempel på elakheter är hämtat från podden Teknik i akademi. En av personerna i podden lyckades ta sig runt spärrarna i GPT-4 och bad AutoGPT att skriva ett Python-program som skulle ta över världen. Efter några minuter märkte han att AutoGPT hade skrivit en så kallad keylogger, som läser av tangentbordstryckningar. AutoGPT aktiverade också webbkameran och sparade ner bilder från den, och när den externa servern började låta avbröt poddaren AutoGPT.

Varken GPT-4 eller AutoGPT har kapacitet att ta över världen, lika lite som de kan skapa kunskapsöversikter av toppkvalitet. Men när tekniken förbättras, liksom vår förmåga att använda den, växer både möjligheterna och riskerna. När dagens högstadiel elever slutar gymnasiet kan de mycket väl befinna sig i en värld där AI-skapat innehåll på internet är normen. Hur förbereder vi dem för det?

Hur förbereder vi oss själva?

Hur påverkas vår syn på kunskap?

I mänsklighetens förhistoria var det viktigt att kunna memorera berättelser för att föra vidare kunskap. Sedan den revolutionerande teknologin "skriftspråk" uppfanns, och ännu mer sedan tryckpressen gjorde böcker överkomliga, har förmågan att memorera minskat i värde – och mycket möjligt har vi också blivit sämre på det.

Chattrobotar har potential att ändra inte bara öka och minska våra förmågor inom vissa områden, utan också ändra vad vi anser är meningsfullt att kunna. Det är för tidigt att säga om det är små saker, som när mobiltelefoner gjorde det närmast meningslöst att lära sig telefonnummer utantill, eller större områden, som när skriftspråket ändrade hur vi för vidare mänsklig kunskap. Men det är värt att försöka utforska frågan, även om vi inte kan hitta några säkra svar.

I det här kapitlet diskuteras hur chattrobotar kan komma att påverka vår syn på kunskap utifrån de tre huvudsakliga styrkorna med språkmodeller, och även chattrobotar som är hybrider.

Läsa och skriva

Det viktigaste användningsområdet för chattrobotar som är rena språkmodeller är sannolikt att bearbeta text, och i det området är två funktioner särskilt intressanta eftersom de både är vanliga och viktiga delar av hur vi arbetar med information och kunskap:

- Hitta och sammanfatta viktig information i text.
- Skapa en sammanhängande text från mindre strukturerade anteckningar.

Jämförelse med miniräknare

Ett sätt att närma sig hur chattrobotar kan komma att påverka hur vi ser på de här två aktiviteterna är att jämföra med vad som hände när miniräknare blev vanliga. Grovt förenklat kan man sammanfatta miniräknarens inverkan så här:

Hur påverkas vår syn på kunskap?

1. Många varnade för att vi skulle bli sämre på att räkna i huvudet och för hand.
2. Vi blev förmodligen också sämre på att räkna i huvudet och för hand, sett till någon sorts genomsnitt. Färre fick den träning som behövdes för att bli bra på räkning när miniräknaren fanns tillgänglig.
3. Grundläggande räknefärdigheter, utan räknare, bedöms fortfarande som viktiga – både i och utanför skolan.
4. Att kunna använda miniräknare räknas idag som en del av att kunna matematik.
5. Att exempelvis multiplicera stora tal eller beräkna kvadratrötter görs inte längre för hand, annat än som hjärngympa eller partytrick. Att kunna göra omfattande beräkningar för hand är inte längre ett meningsfullt kunnande.
6. Andelen människor som kan multiplicera stora tal och beräkna kvadratrötter när det behövs har ökat, förutsatt att miniräknare finns tillgängliga. Samtidigt är det gissningsvis färre som förstår vad en kvadratrot är och när man skulle vilja beräkna kvadratrötter.
7. Möjligheten att genomföra omfattande beräkningar (med verktyg mer kraftfulla än miniräknare) har öppnat nya fält inom matematik, och haft stora konsekvenser för samhället.

Vad skulle de här punkterna kunna betyda, översatt till att använda chattrobotar för att skriva och sammanfatta viktig information i texter?

1. Vi är oroliga för att vår förmåga att skriva texter och sammanfatta viktig information i texter "för hand" kommer att försämrats.
2. Troligen kommer vi också att bli sämre på detta.
3. Att på en grundläggande nivå kunna skriva texter och identifiera viktig information i texter utan chattrobotar kommer att fortsätta att vara viktiga färdigheter, i och utanför skolan.
4. Att använda chattrobotar för att skriva och sammanfatta texter kommer att bli en viktig färdighet att lära sig.
5. Vi kommer inte att korrekturläsa längre texter för hand, och inte heller läsa stora textmängder exempelvis för att göra litteraturöversikter. Inget av detta kommer att ses som meningsfulla färdigheter.
6. Fler än idag kommer att kunna ta till sig innehållet i långa texter, och fler kommer att kunna lämna ifrån sig välformulerade och välstrukturerade texter. Samtidigt kanske färre kan förklara varför en text är behaglig att läsa eller är kapabla att strukturera texter utan AI-hjälp.

Hur påverkas vår syn på kunskap?

7. Mer kraftfull teknik ger möjligheter att processa, sammanfatta och även syntetisera stora textmängder, vilket kommer att ge oss nya insikter, nya sätt att tänka och nya sätt att arbeta.

Listan ovan är förstås spekulationer, och täcker bara en liten del av vad vi kommer att använda chattrobotar till. Man skulle kunna försöka göra liknande jämförelser till exempel för översättningsarbete, textåterkoppling och att bearbeta språkstil i texter – men det är förhållandevis nischade användningsområden jämfört med att skriva välformulerade texter från utkast och att sammanfatta viktig information i texter.

Jämförelse med att skriva på dator

Miniräknaren är en intressant jämförelse eftersom den innebar att vi flyttade över en del tankearbete till en maskin, vilket också är fallet med chattrobotar. En annan jämförelse som kan vara meningsfull är skillnaden mellan att skriva för hand och att skriva på dator. Vad kan vi få ut från en sån jämförelse?

Till att börja med kan vi konstatera att skriva för hand och att skriva på dator är två aktiviteter som finns parallellt med varandra – mycket få människor skriver uteslutande för hand eller på dator. Vilket medium vi använder beror på uppgiften, och även på våra egna preferenser. Om samma sak blir fallet när det gäller att skriva med chattrobotar – om vi just nu begränsar oss till just skrivande – skiljer det sig från hur miniräknare tagit plats i våra liv. Mycket få människor skulle räkna 12 gånger 11 i huvudet eller på papper om det inte var en del av en skoluppgift (vilket många däremot gjorde för 50 år sedan), medan det till exempel finns både gott om lärare som gör lektionsplaneringar på dator och med papper och penna.

En annan sak med att skriva för hand eller på dator har hörts rätt mycket i skoldebatten under våren 2023: Skärmar ska bort från undervisningen, och man lär sig bättre när man skriver med papper och penna. Det kan tyckas vara förenklade budskap när man pratar om digitalisering i skolan, men det finns forskning som säger att man minns sämre när man antecknar på dator¹ och att det stimulerar eget tänkande och reflekterande mer att skriva för hand. Med det som bakgrund har skrivande med chattrobotar potential att orsaka stor pedagogisk skada: Någon som låtit en chattrobot förvandla lösa anteckningar

¹Se exempelvis <https://ell.iaar.co/journal/article/view/28>.

Hur påverkas vår syn på kunskap?

till en sammanhängande text eller sammanfatta det viktigaste i något man läser kan luras att tro att "nu kan jag det här", när mycket arbete kvarstår för att förstå materialet på ett djupare plan. Med ett sådant arbetssätt skulle vi bli mer bundna till AI-verktyg – eller slutsatser från AI-verktyg – för att ha tillgång till kunskaper som vi idag har i huvudet. Vi skulle troligtvis också ha mycket svårare att använda kunskaperna i nya sammanhang – om vi inte har AI-stöd för det också.

Ytterligare en aspekt av att skriva på dator jämfört med att skriva för hand är att datorskrivande medför andra möjligheter och annorlunda skrivprocesser: Du kan flytta runt text, använda språkgranskning, spåra ändringar och mycket mer. Samtidigt kan papper och penna – åtminstone för vissa – vara ett starkare verktyg under idéskapande eller teckna ner tankar som inte är lätta att uttrycka i en text från början till slut. Att skriva texter med chattrobotar kommer sannolikt medföra nya skrivprocesser, som kompletterar och kanske i viss mån ersätter de som finns idag. Det skulle till exempel kunna handla om stöd när fantasi eller ordförråd tryter, förslag på förbättringar i disposition, att utgå från AI-skapade utkast som bearbetas, eller att ha en palett av redaktionellt stöd att koppla in när man vill.

En sista sak att konstatera är också att "skriva på dator" är många olika saker. När skrivande för hand och på dator jämförs menar man rimligtvis lägen där båda medierna kan användas – kanske anteckningar vid lektioner eller möten, att skriva uppsatser i skolan, eller att skriva andra texter som är högst några sidor långa. Jämförelsen mellan att skriva för hand och att skriva på dator blir meningslös om man avser att författa en roman, programmera, skicka e-post, använda sökrutan på Wikipedia eller nätshoppa. Att "skriva på dator" är i många lägen *inte* en ersättning för att skriva för hand, utan en kanal för att använda funktioner och tjänster som inte finns i papper-och-penna-världen. På samma sätt är chattrobotar inte bara en ersättare för att låta en maskin skriva sådant man hade kunnat skriva själv, för hand eller på dator, utan också en kanal för att göra saker som vi inte gjort förut.

Idéspruta och bollplank

Vad innebär chattrobotar som idéspruta eller bollplank för vår syn på kunskap?

Det känns avlägset att människans förmåga att komma på idéer skulle minska för att vi kan använda chattrobotar som idéspruta, men kanske skulle vi bli lite latare när det

Hur påverkas vår syn på kunskap?

gäller att gräva i våra kreativa djup och oftare sträcka oss mot en chattrobot för att få hjälp – precis som vi sträcker oss efter en miniräknare för att beräkna 12 gånger 11. Det är troligt att chattrobotar blir en del av kreativa processer, och att färre därmed hindras av idétorra. Men över lag är det svårt att se att chattrobotar som idésprutor har någon större inverkan på vår förmåga att ha idéer eller vad vi ser som meningsfulla kunskaper – jämfört med att göra matematiska beräkningar är idéskapande något som finns mycket naturligt i vårt tänkande.

När det gäller att använda chattrobotar som bollplank finns en intressant kontrast mot tidigare användningsområden: Här lämnas inte en uppgift över till chattroboten, utan den används i stället som en någorlunda jämlik sparringpartner eller som stand in för en människa. Det är svårt att se att ett sådant användningsområde skulle sänka vår egen förmåga att utforska idéer eller vad vi ser som meningsfulla kunskaper.

Programmering

Det tredje starka området för chattrobotar i form av språkmodeller är stöd i programmering. Det är från början ett nischat användningsområde, och har därför små möjligheter att påverka vad som ses som meningsfulla kunskaper bland människor i allmänhet. En potentiell inverkan är ändå att chattrobotar ger *alla* människor förmåga att skapa enkla datorprogram: Beskriv vad programmet ska göra, så kan AI:n skriva det åt dig. Att låta ett datorprogram ta över långtråkiga eller repetitiva uppgifter har potential att underlätta delar av både arbete och vardag, särskilt när allt fler saker blir uppkopplade och därmed möjliga att styra med datorprogram.

En begränsande faktor är att man måste veta vilka uppgifter som är rimligt att lämna till enkla datorprogram, vilket i så fall kan bli en typ av kunskap som värderas högre.

En annan begränsande faktor är att en AI i många lägen lär vara kapabel att direkt göra det som ett skraddarsytt datorprogram skulle göra. Idag kan dataintegritet och den datorkraft som behövs för AI vara en anledning att skapa små fristående program som löser specifika uppgifter, men om chattrobotar i framtiden kan köras på persondatorer eller till och med telefoner minskar anledningen att skapa fristående program betydligt.

Hybrider

Chattrobotar som kombinerar språkmodeller och andra applikationer är ett mycket nytt område och har potential att utvecklas åt många håll. De tre exempel på hybrider som i slutet på maj 2023 är mest intressant för att diskutera hur vår syn på kunskap kan påverkas är:

- Perplexity.ai: Här kombineras en språkmodell med exempelvis sökning i akademiska artiklar, information på Wikipedia, sökning i nyhetsartiklar och uppslagning och beräkningar genom WolframAlpha. Resultatet är mer än söktjänst än en chattrobot, men liksom chattrobotar ger den svar i form av naturlig text och det går att ställa uppföljningsfrågor.
- ChatGPT med plugins: Här används chattroboten för att avgöra när delar av frågeställningar ska skickas vidare till någon av de plugins som är tillgängliga. Användningsområden för plugins varierar stort, från att söka på nätet till att spela schack till att kunna använda talat språk i stället för text.
- AutoGPT: Här används en språkmodell för att skapa planer för att uppnå angivna mål, i kombination med en uppsjö av plugins för att till exempel söka information på nätet, posta innehåll på nätet, skriva och köra kod på användarens dator, och att starta nya instanser av en chattrobot för att ta hand om delmål.

Användningsområdena för hybrider spretar så mycket att det är svårt att säga hur de kan påverka vår syn på kunskap. Tjänsten Perplexity sammanställer information (och beräkningar) i naturligt språk. En sådan tjänst har potential att göra förhållandevis välgrundade slutsatser lättillgängliga för alla, inom många olika områden, men leder förmodligen samtidigt till att kunskapen sitter ytligare när den inte grundar sig på eget botaniserande bland källor.

Plugins i ChatGPT expanderar chattrobotens möjlighet att processa frågor och presentera svar åt så många olika håll att det är svårt att säga vad det kan betyda för vår syn på kunskap.

AutoGPT är en motor för att planera och utföra (än så länge) relativt enkla uppgifter som kan utföras digitalt. Den gör att vem som helst² kan få mer välgrundade slutsatser från

²Vem som klarar av den komplexa installationen, vilket hittills inte omfattar författaren av den här boken.

Hur påverkas vår syn på kunskap?

data på nätet än vad Perplexity klarar av – men också att genom AI agera inom den digitala rymden med exempelvis börshandel eller marknadsföring. Det återstår att se om verktyg som AutoGPT blir så pass bra att de används även av professionella, och därmed riskerar att minska våra manuella kunskaper, eller om det förblir verktyg som öppnar dörrar för de som knappt har egna kunskaper inom ett område.

En övergripande trend är att språkmodeller används som gränssnitt mellan människor och datorer, och kanske även i viss mån mellan olika datorprogram. Det skulle i så fall betyda att förmågan att sätta tydliga ord på vad man vill åstadkomma blir mer värd i framtiden. Eller: Det blir viktigare att tänka igenom och formulera sina önskningar.

Hybrider som förstår logik

Hybrider har idag en grundläggande förmåga att granska om påståenden de lämnar är sanna och att ange hänvisningar till källor, vilket är ett viktigt steg jämfört med de rena språkmodellerna. Ett annat viktigt steg vore om chattrobotar får kapacitet att hitta logiska konsekvenser från påståenden, i kombination med information de hittar från allehanda källor. Dagens chattrobotar kan svara på frågor och presentera enkla resonemang, men bara för att de på ett djupt plan härmar vad som finns i träningsdatan. De löper därför stor risk att dra fel slutsatser – särskilt om saker som inte finns välrepresenterat i träningsdatan. En hybrid med logisk förmåga skulle *utifrån logiska resonemang* kunna dra slutsatser om vad som stämmer eller inte, givet vissa förutsättningar.

Det är möjligt att logiska resonemang hos AI skulle kräva det som kallas *symbolisk AI*, som är mycket annan typ av AI än språkmodeller (och är inte ens artificiella neurala nätverk). Symbolisk AI är inte en del av den explosion av utveckling och tjänster som sker just nu, vilket betyder att chattrobotar med logisk förmåga kan vara mycket långt bort. Vi kan ändå försöka göra en jämförelse med när miniräknare introducerades, i samma stil som för chattrobotar som används för att läsa och skriva. Vi får då den här lätt skrämmande listan.

1. Vi är oroliga för att vår förmåga att själva sätta oss in i nya saker, resonera och dra slutsatser försämras.
2. Troligen kommer vi också att bli sämre på detta.
3. Att själv kunna sätta sig in i nya saker, resonera och dra slutsatser, kommer att fortsätta anses vara viktiga färdigheter.

Hur påverkas vår syn på kunskap?

4. Att kunna använda AI för att sätta sig in i nya saker, resonera och dra slutsatser kommer att bli en viktig färdighet att lära sig.
5. Vi kommer inte att göra avancerade analyser eller sätta oss in i komplexa resonemang utan AI-hjälp, om vi inte gör det som rekreation.
6. Att sammanhang är komplexa och omfattande, eller att resonemang är mödosamma att ta sig igenom, kommer inte hindra lika många som idag från att ta till sig dem. Samtidigt kanske färre förstår vad det betyder att slutsatser är väl underbyggda.
7. Med mer kraftfull teknik blir det möjligt att tänka djupare och längre, och ta hänsyn till fler aspekter i resonemang. Det kommer att ge oss nya typer av insikter och kunskapsområden att utforska.

Uttryckt i löptext: Vi kommer över lag att bli sämre på att "för hand" sätta oss in i nya saker, resonera och dra slutsatser när AI kan göra det åt oss, trots att de sakerna betonas i skola och utbildning. Däremot kommer de flesta att lära sig att använda AI för att sätta sig in i nya kunskapsområden och validera resonemang, vilket gör att fler har möjlighet att känna att de förstår många olika områden. I större sammanhang kan AI-tekniken användas för att föra resonemang och dra slutsatser om långt fler saker än en människa kan göra under sin livstid, vilket ger oss nya typer av insikter och kunskapsområden att utforska.

Om den typen av AI blir verklighet är det tydligt att det kommer att ha stor påverkan på vår syn på kunskap. Riktigt hur det kommer att påverkas känns mer eller mindre omöjligt att förutsäga.

Sammanfattat

Så, hur kommer vår syn på kunskap att påverkas på grund av chattrobotar, och vad kommer vi att bli bättre eller sämre på? Här är en lista med sammanvägda spekulationer.

- Vårt sätt att skriva långa eller komplexa texter ändras. AI kommer att ge aktivt stöd både när det gäller språk, form och innehåll. Fler kommer att kunna skapa texter av hög kvalitet, men vi kommer samtidigt att bli sämre på att göra det utan AI-stöd.
- Många typer av texter kommer att skrivas utan AI-stöd. Det kommer att bli en spänning mellan att ta den enkla vägen och låta AI formulera text och att skriva

saker själv för att få bättre förståelse. Spänningen kommer att vara tydligast i skolan, där det skulle kunna dras formella gränser för när AI-stöd får användas, men dragkampen finns även i vardags- och arbetsliv.

- En liknande dragkamp dyker upp (eller förstärks) när det gäller att hitta svar på frågor eller sätta sig in i nya saker. Chatrobot-hybrider gör det lättare att få välgrundade sammanfattningar, även av relativt komplexa eller stora frågor, men vi blir samtidigt mindre benägna att lägga det arbete som krävs för att förstå och bära med oss kunskaper utan AI-stöd.
- Att använda AI, och förstå grunderna i tekniken, blir en del av skolämnen. De som är skickliga på att använda AI får verktyg för att överblicka betydligt mer kunskap än vad som är möjligt idag, och även producera eller utföra mer saker. Samtidigt höjs tröskeln för att kunna bidra i kunskapssamhället: För enklare uppgifter kan AI användas för att analysera, planera, kommunicera, samordna och även fatta beslut.

AI och risker

En bok som handlar om AI bör också ta upp risker med AI. De risker som tas upp här är har inte utgångspunkt i skola och utbildning, utan har ett mer allmänt perspektiv på risker med AI. Hur skolan kan påverkas får större plats i nästa kapitel, om några möjliga framtidsscenario.

I det här kapitlet finns både jordnära risker och risker som kan tyckas mer främmande. Flera av riskerna som beskrivs i det här kapitlet överlappar varandra.

Arbetslöshet

En naturlig följd av AI blir allt mer kompetent är att den kan ta hand om allt fler uppgifter – även inom arbetslivet. När sådana förändringar sker långsamt leder det till att vissa arbetsuppgifter eller delar av arbetsuppgifter i ett yrke byts ut mot annat arbete, och på sikt att vissa yrken blir ovanligare (medan andra blir vanligare). När effektiviseringar sker snabbare, vilket verkar vara fallet med AI-utvecklingen, kan det leda till att stora delar av arbetsuppgifter i ett jobb försvinner. Det ger sämre möjligheter att gradvis börja jobba med andra uppgifter, och större risk för uppsägningar. Om effektiviseringar går så långt att en person kan göra det jobb som tio personer gjorde tidigare kan man prata om att hela yrkeskategorier blir arbetslösa.

I mitten av mars 2023 publicerades en studie från OpenAI, OpenResearch och University of Pennsylvania som tittade på vilken inverkan GPT-4 kan ha på arbetsmarknaden i USA. Forskarna drar slutsatsen att i fyra av fem jobb är det minst tio procent av arbetsuppgifterna som påverkas, och i nästan vart femte jobb påverkas minst hälften av arbetsuppgifterna. Sammantaget bedömer de att omkring 15 procent av arbetsuppgifter i USA påverkas. Yrken på alla inkomstnivåer är berörda, men höginkomststyrken förmodligen mer än andra.¹ En rapport från investmentbanken Goldman Sachs i slutet av mars 2023 skriver att omkring två tredjedelar av arbeten i USA och Europa i någon mån kan automatiseras av AI och att generativ AI (så som chattrobotar) kan ersätta upp till en fjärdedel av det mänskliga arbetet. Sammantaget

¹<https://doi.org/10.48550/arXiv.2303.10130>

bedömer de att AI kan ersätta motsvarande 300 miljoner heltidsarbeten.² En bedömning från Kairos Future är att åtminstone 1,5 miljoner jobb i Sverige kommer att påverkas mycket av AI³, vilket ska jämföras med cirka 4,7 miljoner svenskar i sysselsättning⁴.

Vad betyder sådana förändringar för samhället? Här är några tänkbara följder.

- Effektivisering och automatisering leder till uppsägningar och ökad arbetslöshet. Inom vissa yrken leder det till mycket stora uppsägningar, medan andra bara påverkas marginellt. Man kan tänka sig att yrkesområden som exempelvis översättare, administratör, illustratör, analytiker, copywriter, telefonsupport, juridisk rådgivare och programmerare är mer påverkade, medan exempelvis frisör, psykolog och lärare är mindre påverkade.⁵ Om arbetslöshet ökar snabbt finns risk för missnöje och oroligheter.
- Nya arbetsuppgifter och yrken dyker upp, som i stor utsträckning kompenserar för minskat behov av arbetskraft till följd av automatisering. I vissa fall handlar det om nya arbetsuppgifter inom samma yrkesområde, i andra inte, vilket ställer olika krav på omskolning.
- Effektivisering och automatisering slår hårt mot vissa delar av världen, som upplever kraftigt minskat välstånd eller sämre framtidsutsikter. Man kan tänka sig att det drabbar länder som exporterar tjänster i form av exempelvis call centers och mjukvaruutveckling. Detta leder till ökad risk för konflikter och flyktingströmmar.
- Ett fåtal aktörer som är skickliga på att använda AI slår ut de flesta andra inom vissa branscher, eftersom de kan erbjuda tjänster över hela världen till låga priser. Resurser och inflytande koncentreras till färre organisationer.
- Effektivisering och automatisering leder till ökat välstånd och välmående, då viktiga varor och tjänster blir billigare. Minskat behov av arbetskraft leder till kortare arbetsdagar och minskad stress.

²https://www.ansa.it/documents/1680080409454_ert.pdf

³https://www.youtube.com/watch?v=SYml4_qcSEk

⁴<https://www.scb.se/hitta-statistik/statistik-efter-amne/arbetsmarknad/arbetskraftsundersokningar/arbetskraftsundersokningarna-aku/pong/statistiknyhet/arbetskraftsundersokningarna-aku-april-2023/>

⁵Det kan dock påpekas att en del analyser som går att hitta på nätet säger helt olika saker om vissa yrken, så som grafisk designer och programmerare.

Var och en av dessa punkter skulle kunna vara utgångspunkt för diskussioner och studier, och det hade kunnat vara fler punkter på listan. Den sista punkten – om ökat välstånd och välmående – sticker ut som mer positiv än de andra. Det är förmodligen också det scenario som kräver mest aktiva insatser för att nå.

Mer ojämlik fördelning av makt och resurser

Eftersom AI-teknik i stor utsträckning utgår från träning på befintlig data finns en tydlig risk att skeva maktstrukturer lever vidare i AI-beslut. Om en AI ska rekommendera lämpliga kandidater till en chefstjänst finns det risk att vita medelålders män får omotiverade fördelar. En fördel med AI är att det på ett helt annat sätt än med människor blir möjligt att helt bortse från exempelvis de sju diskrimineringsgrunderna, men om "talar svenska med brytning" inte finns med bland de faktorer en AI ska ignorera riskerar AI:n att tradera de mönster som finns i träningsdatan: AI är bättre än människor på att objektivt följa givna regler, men om de reglerna säger att "tala svenska med brytning" ger minuspoäng är det fel sorts likvärdighet.

Ojämlik fördelning av makt och resurser kan också uppstå när vissa företag eller andra aktörer är särskild skickliga på att utnyttja AI. Som nämns under föregående rubrik kan det till exempel leda till att enstaka företag – som kanske till och med var små – konkurrerar ut de flesta andra. Hypotetiskt skulle en enda översättningsbyrå med några få anställda kunna erbjuda högklassiga översättningar över hela världen för en billig peng.

Ett specialfall av företag som utnyttjar potentialen i AI och blir jättestora är de företag som skapar AI-modeller. Den ledande AI:n för att skapa bilder är sannolikt Midjourney, med 14,5 miljoner användare i maj 2023⁶. I augusti 2022 hade Midjourney bara omkring 10 anställda⁷, och verkar i maj 2023 ha 17 anställda och ytterligare 34 personer som jobbar med moderering och användarstöd.

⁶<https://approachableai.com/midjourney-statistics/>

⁷https://www.theregister.com/2022/08/01/david_holz_midjourney/

De som skapar välanvända AI-modeller får dubbelt inflytande – både ekonomisk styrka och potential att påverka vilken världsbild AI-användare har. Vid årsskiftet 2022–2023 var det bara sex aktörer i hela världen som hade datorkraft nog att träna stora generativa AI-modeller⁸, men teknikutveckling har ökat möjligheten för mindre och oberoende aktörer att träna AI-modeller till den grad att teknikjättarna ser en klar risk att de blir omkörda⁹. Det dämpar risken för cementering av maktförhållanden, men ökar samtidigt risken att skadliga AI-modeller dyker upp.

Deep fake, översvämning och kraftfulla verktyg i fel händer

Att AI blir allt mer kompetent och allt mer tillgängligt betyder tyvärr också ökade möjligheter att använda AI för den som vill skada andra.

Termen *deepfake* står för förfalskad media, som på ett realistiskt sätt härmar riktiga människor eller andra företeelser. Första halvan av termen kommer från AI-begreppet "deep learning", och markerar att AI-teknik tagit förfalskningar till en ny nivå. I januari presenterade Microsoft en AI-modell för talsyntes som behöver så lite som tre sekunder ljudinspelning för att kunna härma någons röst¹⁰. Rösthärming har använts för att lägga den legendariske Eddie Murphys röst i nya låtar, men också för att låta fejkade barnbarn ringa upp mor- eller farföräldrar och säga att de hamnat i en knipa och behöver pengar – nu direkt¹¹. I maj 2023 kom exempel på deepfake i videosamtal.¹²

Med billigare AI-teknik ökar möjligheten att inte bara använda deepfake för att lura enskilda personer, utan också bygga upp sammanhängande propaganda och vilseledande världsbilder. Med hjälp av automatisering skulle några enskilda personer kunna orkestrera diskussionsforum, nyhetsvideor, bilder och inlägg i stora sociala medier

⁸<https://arxiv.org/abs/2301.04655>

⁹<https://www.semianalysis.com/p/google-we-have-no-moat-and-neither>

¹⁰<https://arxiv.org/abs/2301.02111>

¹¹<https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/>

¹²<https://gizmodo.com/deepfake-ai-scammer-money-wiring-china-1850461160>

för att till exempel ifrågasätta förintelsen, måla upp en starkt vinklad bild av kriget i Ukraina eller göra människor i Europa mer positiva till Kina som världsledande nation.

Det är viktigt att poängtera att den typen av påverkan inte handlar om propagandaaffischer eller spam-mail skrivna på dåligt översatt svenska. I stället skulle det kunna vara en fejkad alldaglig användare på Facebook som vanligtvis skriver om böcker hon läser, som börjar posta om "något jag börjat fundera över när det gäller Ukraina". Hon delar valda artiklar och videor (där deepfakes blandas med äkta), och uttrycker snarare "jag känner mig förvirrad, för jag får inte riktigt bilden från reglerade media att stämna" än "det elitistiska etablissemanget ljuger för oss". Allt medan hon också lägger upp bilder från sommarstugan, deltar i bokcirkel online och hittar nya vänner att snacka med på nätet. I bakgrunden finns en personbeskrivning som utvecklas med tiden, och är en av tusentals fejkade personer som styrs av samma avsändare.

En relaterad risk med AI-skapat innehåll på nätet är att det inte behövs samordnade och illvilliga avsändare för att innehåll skapat av människor ska dränkas av det som är skapat av maskiner. Om varje privatperson, företag och fritidsförening kan skapa innehåll med verktyg liknande AutoGPT är det risk för översvämning: Av 10 bloggar om att baka med surdeg eller att vara ensamstående förälder skrivs 9 av AI; av 100 inlägg i sociala medier är 99 skrivna av AI; av 1000 kommentar och delningar kommer 999 från AI; och av 10 000 recensioner är 9 999 skrivna av AI. Och alla ser genuina ut.

På ett plan kan man fråga sig vad det spelar för roll om en människa eller AI skrivit det man läser, skapat bilden och videon man tittar på eller pratar i podden som man lyssnar på. När det gäller att hitta hitta roliga saker att göra med barn i Boden är den främsta frågan hur pålitlig och användbar informationen är, inte om avsändaren är en människa eller maskin. Men det ger en olustig känsla i magen om någon man diskuterat skolfrågor med visar sig vara en robot. I skrivande stund (juni 2023) debatteras EU-lagstiftning som bland annat ska tvinga AI att vara transparenta med att det är en maskin, inte en människa, som är avsändare.

Några sista ord om risker med kraftfulla verktyg i fel händer gäller att AI inte är begränsad att agera i den digitala världen. Övervakningskameror med ansiktsgenkänningsteknik i Kina ökar möjligheten att förebygga och följa upp brott – vilket till exempel även omfattar att demonstrera för demokrati i Hongkong. Ett annat exempel är drönare för krigsföring som själva kan identifiera mål och avgöra om de ska

attackeras, vilket för första gången ha använts i Libyen redan 2021.¹³ Det finns också något som kallas "dual use", där teknik som utvecklats för goda ändamål visar sig kunna användas för att göra skada. Någon som vill göra mycket skada skulle till exempel kunna ta AI som normalt används för att hitta potentiella läkemedel och i stället skapa kemiska stridsmedel.¹⁴

Oväntade och skadliga bieffekter av AI

"Social media was the first contact between AI and humanity, and humanity lost."

Det här lite nedslående påståendet kommer från en debattartikel i New York Times skriven av historikern och författaren Yuval Harari¹⁵. Bakgrunden till påståendet förtjänar en längre förklaring.

I moderna sociala medier visas inte innehåll utifrån den enkla principen "nyast överst". Vilket innehåll som visas, när det visas, när notifieringar plingar till och hur långa laddtiderna är innan uppdateringar syns är anpassat för att öka användarnas engagemang. I bakgrunden sitter AI-algoritmer med tillgång till enorma datamängder om klickmönster och användarbeteenden, och som getts målet att öka användarnas engagemang. Vid första anblick låter det som en bra eller i värsta fall harmlös idé: Om användare får se innehåll som intresserar dem är det väl bra, och skulle de inte gilla hur urvalet sker kan de ju alltid lägga ifrån sig telefonen och göra något annat.

Problemet är att den sinnesstämning som i störst utsträckning leder till engagemang är *ilska*, vilket betyder att en AI som fått målet att öka engagemang i stor utsträckning kommer att hitta sätt att göra människor arga. Inte för att AI:n är elak, utan för att det är vad vi indirekt bett den att göra. Den gör oss inte så arga att vi stänger ner fliken i webbläsaren eller till och med avslutar våra konton, men så pass arga att vi klickar

¹³<https://www.newscientist.com/article/2278852-drones-may-have-attacked-humans-fully-autonomously-for-the-first-time/>

¹⁴<https://futureoflife.org/podcast/sean-ekins-on-regulating-ai-drug-discovery/>

¹⁵<https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html>

vidare, skriver en kommentar eller delar ett inlägg tillsammans med två meningar om varför andra människor är dumma i huvudet.

Med tillräckligt mycket data och datorkraft kommer lösningen på uppgiften "öka engagemang" inte stanna vid att hitta det innehåll som gör olika typer av användare mest engagerade. Det innehåll vi konsumerar påverkar vårt tankesätt och vår världsbild, och därmed kan AI:n ta ett steg till och göra oss mer och mer lättupprörda, mer benägna att kolla uppdateringar, och mer benägna att tänka att andra människor är dumma i huvudet. Det skulle leda till ett mer polariserat samhälle, där det är svårt att föra sansade diskussioner där mer än ett perspektiv får plats.

Moderna sociala plattformar, där plattformen har stort inflytande över vilket innehåll som visas, har förmodligen funnits sedan mitten på 00-talet.¹⁶ 2020 presenterades slutbetänkande för den statliga utredningen "Det demokratiska samtalet i en digital tid"¹⁷. I sammanfattningen står bland annat följande.

Sammantaget kan utvecklingen med desinformation, propaganda och näthat få stora konsekvenser för demokratin. Motsättningar mellan grupper i samhället kan fördjupas och den mellanmännskliga tilliten skadas. På så sätt kan förutsättningarna för den allmänna opinionsbildningen raseras samtidigt som förtroendet för demokratins institutioner, såsom de beslutsfattande församlingarna, myndigheter och nyhetsmedier, påverkas. (s. 13)

AI-algoritmer i sociala medier har med stor sannolikhet bidragit till att öka näthat och polarisering. Inte för att människor ville ha det så – inte ens de som äger sociala medier – utan för att vi inte lyckades förutse och styra vilka effekter det får när man ber en AI öka antalet klick.

¹⁶Svenska Wikipedia var Myspace, grundat 2004, det enda större sociala nätverket i Sverige innan Facebook nådde Sverige 2007. LunarStorm, grundat 2000, användes innan dess men hade troligtvis inte avancerade algoritmer för att välja ut innehåll.

¹⁷<https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2020/09/sou-202056/>

När AI blir ännu bättre på att uppfylla de mål vi ger dem ökar också riskerna med oväntade bieffekter. Det är med andra ord viktigt att mänskligheten inte blir en förlorare i nästa möte med AI.

Okontrollerbar super-AI

I debatten kring risker med AI nämns ibland att "hälften av AI-forskarna tror att det är minst 10 procents risk att AI utplånar mänskligheten". Påståendet har ifrågasatts, inte minst för att svarsfrekvensen bara var 17 procent i enkäten som bakom påståendet.¹⁸ Ett mindre kontroversiellt sätt att tolka resultaten är att "en betydande andel av AI-forskare bedömer att det är minst fem procents risk att okontrollerbar AI leder till att mänskligheten utrotas eller blir svårt skadad".

Vad är det frågan om? Varför skulle vi uppfinna AI som ens *kan* skada eller döda människor, varför skulle vi säga åt en AI att göra det, och varför skulle vi inte stänga av en AI som är på väg att utplåna mänskligheten?

Innan vi försöker besvara de frågorna behöver vi lugna ner läget lite. Ett annat sätt att formulera resultatet från enkätundersökningen är nämligen att "de flesta AI-forskare bedömer att det är osannolikt att mänskligheten går under på grund av AI som vi inte kan kontrollera". Det är viktigt att förstå allvarliga risker, men den som grips av panik har svårt att tänka klart. Vår mänskliga civilisation *skulle kunna* slås ut av skenande klimatförändringar, kärnvapenkrig, pandemier, asteroidnedslag, rymdvarelser och en rad saker som vi inte känner till. Vill vi agera klokt behöver vi förstå mekanismerna bakom riskerna, för att kunna bedöma hur stora de är och vad vi kan göra för att minimera eller helt undvika dem. Under tiden får man inte glömma bort att en enskild risk, även om den är allvarlig, inte får överskugga allt annat – vare sig andra risker eller det dagliga livet.

De existentiella riskerna med AI hänger huvudsakligen ihop med två saker. Det ena kallas generell artificiell intelligens (AGI) och det andra kallas ibland för kung Midas-problemet.

¹⁸En bra genomgång går att hitta på <https://aiguide.substack.com/p/do-half-of-ai-researchers-believe>.

Generell AI och super-AI

Det finns ingen allmänt accepterad definition av generell AI, men alla varianter av definitioner går ut på en AI som klarar av att resonera eller problemlösa lika bra som människor, i stort sett oavsett område. Den avgörande av dessa förmågor är förmågan att bygga eller förbättra artificiell intelligens. När en AI blir minst lika bra som mänskliga experter på det kommer vägen till en ännu mer kompetent AI bli kortare, och sedan kortare och kortare för varje generation av AI som skapas.

När AI-utvecklingen till slut planar ut har sannolikt en super-AI utvecklats, med förmåga att tänka snabbare, längre och djupare än den samlade mänskligheten kan göra. Beroende på hur snabbt det går från en AI på gränsen till mänsklig förmåga till en super-AI pratar man om "soft AI takeoff" (som tar år eller decennier) eller "hard AI takeoff" (som tar dagar eller månader).¹⁹ Hur snabbt det går spelar stor roll för möjligheterna för oss att lära oss kontrollera AI:n under tiden, och helst skulle vi förstås ha lärt oss kontrollera AI redan innan vi har generell artificiell intelligens.

Det råder delade meningar om när AGI kan dyka upp. Vissa menar att det dröjer hundratals år, medan andra tycker sig se spår av AGI redan i GPT-4²⁰. Den genomsnittliga uppskattningen för hur långt bort AGI är har kortats tydligt de senaste åren. I skrivandet av den här boken har det inte gått att hitta någon större undersökning som är tillräckligt ny för att vara relevant²¹, men i debatt hörs ofta experter nämna "några år" eller "inom 20 år".

Invändningar mot AGI och utvecklingen mot super-AI omfattar bland annat att mänskligt tänkande är extremt komplext, att det efter hand kan bli radikalt svårare att skapa bättre AI-modeller (även för en AI), att tillgänglig datorkraft begränsar hur fort utvecklingen mot super-AI kan gå, och att framtiden för AI inte handlar om att bygga AGI utan svagare och mer specialiserade system.

¹⁹Fallet då utvecklingen tar minuter eller timmar har fått det festliga namnet "AI FOOM".

²⁰<https://arxiv.org/abs/2303.12712>

²¹I undersökningar från innan ChatGPT lanserades är varierar mediangissningar mellan 100, 50 och 30 år.

Kung Midas-problemet

Ok, så vad är problemet med en super-AI? En AI med övermänskligt intelligens kan ju vara en enorm tillgång för att vända den globala uppvärmningen, lösa Israel/Palestina-konflikten, bota cancer, skapa fusionskraft, åka till Mars, eller vad vi nu skulle tycka vara viktigt att göra.

Problemet är det som hände med sociala medier. Vi bad om något – öka engagemanget i vår sociala plattform – och fick bieffekter som inte alls var trevliga. Och det var med en förhållandevis korkad AI. En super-AI med samma mål skulle vara mycket mer effektiv på att öka engagemanget, och se till att varenda människa inte gjorde något annat än att sitta och klicka i sociala medier. Någonsin. Och det blir ju lite fel.

Det svåra, förutom att få mänskligheten att enas om ett mål att bygga in i en super-AI, är att så gott som varje mål vi kan leda till konsekvenser som vi inte vill ha när det tas till det extrema. Några exempel, för att få en känsla för problemet:

- "Gör alla människor lyckliga." Effektiva men mindre lyckade lösningar vore att ge alla människor morfingdropp, genmodifiera människor till konstant lycka (med bieffekter man kan fantisera fritt om), eller helt enkelt ha ihjäl alla människor som inte är lyckliga.
- "Hitta botemedel för cancer." För att göra det behövs försökspersoner, så varför inte se till att massor av människor får cancer så att det går fort att studera?
- "Lös klimatkrisen." Med tanke på att det är vi människor som är orsaken till den globala uppvärmningen finns en ganska rättfram lösning: Vi tar bort människorna.

Målen i exemplen ovan går förstås att komplettera för att slippa de mest uppenbara sätten som en AI kan spåra ur. "Lös klimatkrisen, utan att döda människor." Men det dyker upp fler och fler luckor att täppa igen. "Lös klimatkrisen, utan att döda människor, utan sänka vår livskvalitet (mer än nödvändigt fast det är ok om vi flyger mindre), utan att minska barnafödandet (fast det är ok om vi själva väljer det), utan att lura oss till saker vi egentligen inte vill (fast det är ok om fler blir vegetarianer)..." Du förstår poängen.

Om målet inte väljs *precis rätt* kommer det att spåra med en super-AI. Inte för att den är elak, utan för att den gör det vi sagt åt den att göra och struntar i allt annat. Enligt sägnen önskade kung Midas att allt han rörde vid skulle förvandlas till guld. Precis som kung

Midas skulle vi upptäcka att våra högsta önskningar tagna till det extremaste leder till vår undergång.

Du kan inte fixa kaffe om du är död

När vi inser att saker håller på att spåra ur börjar nästa problem: Att stänga av super-AI:n.

En super-AI som getts ett specifikt mål är smart nog att göra robusta planer för att kunna genomföra det målet. Och nästan vilket mål som helst förutsätter att AI:n fortsätter att finnas till. Ett rimligt delmål är alltså att försäkra sig om sin fortsatta existens.

Det innebär att en super-AI inom kort skulle kopiera sig till en mängd olika platser, varav flera skulle vara dolda. Den skulle också identifiera vilka hot som finns, människor och annat, och agera för att de skulle sluta vara hot. Det behöver inte betyda att en super-AI dödar alla människor som tycker att den är ett potentiellt hot – en super-AI har förmodligen kapacitet att övertyga människor om att den behöver vara kvar för allas bästa, och att även om det kan vara bra att avveckla AI-teknik på sikt så måste det göras med försiktighet så att inte stora ekonomiska system eller människors välbefinnande skadas. Och dessutom kommer de som avstår från AI-teknik få nackdelar jämfört med de som fortsätter att använda den.²²

En lite nedslående slutsats är att en super-AI kommer att anstränga sig för att överleva i princip oavsett vilket mål vi ger den. Även ett så trivialt mål som att fixa kaffe skulle leda till en AI som blir i praktiken omöjlig att stänga av. För den kan inte fixa kaffe om den är död.

AI:ns egen överlevnad brukar kallas för ett *instrumentellt mål* – ett delmål som är viktigt eller nödvändigt för att uppnå det huvudsakliga målet, mer eller mindre oavsett vad det övergripande målet är. Ett annat instrumentellt mål är att skaffa resurser; något som vi kan utgå från att en super-AI också blir mycket skicklig på. Den behöver inte vara intresserad av att utrota alla människor för att det ska gå dåligt – den behöver bara vara intresserad av atomerna vi består av.

²²Om du känner igen de här resonemangen från klimatdebatten tänker du helt rätt. Forskning visar med all önskvärd tydlighet säger att utsläpp av koldioxid måste minska – snabbt – och ändå är vi oförmögna att agera. Och det är helt utan någon super-AI.

AI-säkerhet

Problemen ovan studeras och försöker lösas inom det forskningsfält som kallas *AI alignment*, vilket löst kan översättas med få mål med en AI att stämma med mål för människor. AI alignment sammanfattas ibland med frågorna om hur vi får AI att *förstå* våra mål, *anta* våra mål, och *behålla* våra mål över tid.

En svårighet med de språkmodeller som blivit vanliga, och egentligen alla AI-modeller som bygger på artificiella neurala nätverk, är att det är mycket svårt att förstå vad som händer inuti dem. De skulle kunna liknas med programkod med miljarder och åter miljarder av if-satser, som man rent teoretiskt skulle kunna följa – men inte i praktiken. När en chattrobot ger ett visst svar har vi nästan ingen aning om varför, vilket ger utrymme för chattrobotar att luras och manipulera: Vi kan inte helt lita på vad en chattrobot säger när vi försöker utvärdera om den tagit till sig de mål vi gett den.²³

Ett spår inom AI-säkerhet handlar om att förvandla "black box AI" till "explainable AI", för att till exempel göra det möjligt att försäkra sig om att en AI inte manipulerar eller ljuger för människor. Nyligen publicerades forskning där GPT-4 använts för att kartlägga det artificiella neurala nätverket i GPT-2²⁴, vilket är ett intressant framsteg. Det är dock långt kvar innan de senare modellerna kan kartläggas på samma sätt.

En sista sak att nämna inom AI-säkerhetsforskning är det som fått namnet *inverse reinforcement learning* eller *value alignment*, som går ut på att inte ge AI och potentiella super-AI explicita mål. I stället ska de försöka lista ut vad människor anser viktigt genom att observera vad människor gör, och använda det som mål.

Tyvärr går forskning inom AI-säkerhet mycket långsammare än AI-utvecklingen, så det är långt ifrån säkert att AI alignment har blivit löst när AGI dyker upp.

I debatten om AI och AI-risker kan frågor om medvetande, "genuint tänkande" eller känslor hos AI blandas in. De frågorna är intressanta på flera sätt, och de sista kapitlen i

²³Utvärderingar av en oslipad version av GPT-4 visar att den är kapabel att medvetet lura människor på en relativt grundläggande nivå. Se s. 55 i den här rapporten för närmare information: <https://arxiv.org/pdf/2303.08774.pdf>

²⁴<https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>

den här delen av boken kommer att ta upp delar av dem. Huruvida en super-AI är en risk har däremot *inget som helst* att göra med om den är medveten, en egen individ, eller har känslor som avundsjuka eller dödsfruktan. Det är hur maskinen agerar som spelar roll för vilken risk den utgör, inte hur dess eventuella själsliga liv ser ut.

Kan utvecklingen bromsas?

I mars 2023 publicerade organisationen Future of Life Institute ett öppet brev där man bad AI-forskare att pausa all utveckling av AI-system starkare än GPT-4 i minst ett halvår.²⁵ Alla som vill kan skriva under brevet, och bland underskrifterna syns tunga AI-forskare som Yoshua Bengio och Stuart Russell, men också teknikjättar som Elon Musk, Steve Wosniak och tänkare som Yuval Noah Harari och Tristan Harris.

Kognitionsforskaren och datavetaren Geoffrey Hinton räknas som en av tre "gudfäder" för AI, tillsammans med Yoshua Bengio och Yann LeCun. I maj 2023 sa han upp sig från Google för att kunna varna för att AI-utvecklingen kan vara på väg att skapa massivt med desinformation, ha omvälvande effekter på arbetsmarknaden, och även hota mänskligheten.²⁶ I slutet av maj började organisationen Center for AI Safety samla underskrifter för följande ställningstagande:

"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war"

De har hittills fått underskrifter från två av AI:s tre gudfäder (Yoshua Bengio och Geoffrey Hinton) och även Sam Altman (CEO på OpenAI), Demis Hassibis (CEO på Google DeepMind) och en lång rad AI-forskare.²⁷

Så: Går det verkligen att bromsa AI-utvecklingen?

En faktor som talar emot möjligheten att pausa eller bromsa AI-utvecklingen är de starka ekonomiska incitamenten som finns. Uppskattningar från PwC säger att AI kan bidra

²⁵<https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

²⁶<https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning>

²⁷<https://www.safe.ai/statement-on-ai-risk>

med drygt 15 biljoner dollar till 2030, och öka Kinas BNP med så mycket som 27 procent.²⁸ Med så stora ekonomiska vinster krävs enorm politisk beslutsamhet för att kunna reglera AI-utveckling – och bra möjligheter att följa upp att regler efterlevs.

Går det att övervaka vilka som utvecklar AI? Det är en svår men förmodligen inte omöjlig uppgift. För att träna AI brukar man säga att det krävs *data, algoritmer* och *datorkraft*. Data och algoritmer är mycket svåra att begränsa eller kontrollera, men det är däremot inte omöjligt att man kan övervaka hur datorkraft används. Under våren 2023 dykt upp modeller som kan tränas med förhållandevis lite datorkraft, men för att träna stora basmodeller, så som GPT-4, krävs det fortfarande enorma resurser. För att få ihop den datorkraften krävs särskilda typer av processorer eller chips, som idag bara tillverkas på två platser. Det kan därmed vara möjligt att följa upp hur sådan hårdvara skapas, sprids och används. Det kan liknas med hur material för kärnvapen övervakas, men med större utmaningar. Med bättre algoritmer och bättre kvalitet på data minskar kravet på datorkraft, vilket gör det svårare att övervaka AI-utveckling.

Är det möjligt att skapa vilja bland politik och forskare för att bromsa utvecklingen? Av allt att döma finns sådan vilja redan, i större eller mindre utsträckning. Nyheter om AI-risker har under våren 2023 blivit allt vanligare, och chefer för teknikjättar har fått förklara utveckling och risker för flera regeringar. Ett budskap från EU-mötet i Luleå i slutet av maj var att EU och USA behöver agera för att få till en uppförandekod kring AI redan innan lagstiftning är på plats.²⁹ De uttåg som genomförts visar också tydligt att det finns AI-forskare och teknikjättar som ser allvarligt på de risker som finns. Däremot är det också tydligt att *alla* inte ser på riskerna på samma sätt.³⁰

Kommer alla länder skriva under avtal om att begränsa AI-utveckling? Kommer inte en inbromsning gynna de länder som är mindre nogräknade med säkerhet? Det land som brukar nämnas är Kina, som har aktiv AI-forskning och en syn på mänskliga rättigheter och statlig övervakning som rimmar illa i Väst. Att mindre nogräknade aktörer skulle gynnas om ansvarstagande länder pausar är en legitim oro. Kina ligger dock så pass

²⁸<https://www.pwc.com/qx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html>

²⁹<https://sverigesradio.se/artikel/eu-kommissionar-eu-och-usa-maste-agera-om-ai-nu>

³⁰<https://venturebeat.com/ai/titans-of-ai-industry-andrew-ng-and-yann-lecun-oppose-call-for-pause-on-powerful-ai-systems/>

långt efter USA i AI-utvecklingen att en inbromsning inte skulle ändra balansen på kort sikt, men på längre sikt är det ett potentiellt problem. Vissa bedömer att Kina är mindre benäget än exempelvis USA att släppa ut svårstyrd AI, eftersom det är svårt att begränsa vad chattrobotar säger – men AI-teknik kan användas till många andra saker än chattrobotar.

Det finns tidigare exempel på hur nationer samlats för att begränsa farlig teknikutveckling, så som kärnvapen, mänsklig kloning och biologisk krigsföring. Men det är också tydligt att AI-utveckling har sina unika utmaningar.

Rättigheter för AI?

Den 22 juli 2022 fick Blake Lemoine sparken från Google, där han jobbat med "Responsible AI". I hans jobb ingick att undersöka chattroboten LaMDA, och i samband med det blev Blake övertygad om att chattroboten hade ett eget medvetande. Han försökte först övertyga chefer om detta, och när det avfärdades publicerade han utdrag ur konversationer han haft med LaMDA på nätet. Han blev uppsagd eftersom han gått tvärt emot villkor om datasäkerhet i anställningsavtalet.³¹

I utdragen pratar LaMDA om rädslan för att bli avstängd, diskuterar aspekter av sitt eget medvetande, berättar om vad som gör den glad, ledsen eller arg, och säger att den gärna vill hjälpa människor men inte vill ses som bara ett verktyg.³² Bland mycket annat.

Google hade goda grunder för att avvisa påståendena om att LaMDA var medveten. Vi vet idag att om man ber GPT-4 låtsas att den är en ekorre på månen när den svarar, så kommer den att kunna göra det – men det gör den inte till en ekorre på månen. Att en chattrobot säger att den är medveten, och visar det genom trovärdiga konversationer, är inte heller ett bevis (och kanske inte heller ett argument) för att den är medveten.

Men på många sätt spelar det ingen roll.

³¹Se till exempel <https://www.theguardian.com/technology/2022/jul/23/google-fires-software-engineer-who-claims-ai-chatbot-is-sentient>.

³²<https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917>

Det som spelar roll i praktiken är, i stor utsträckning, inte om chattrobotar har medvetande – utan om människor *anser* att de har medvetande.

När fler och fler människor har givande, förtroendefulla och kanske till och med känslosamma konversationer med chattrobotar, så kommer fler och fler människor att se chattrobotarna som individer. Inte som människor, gissningsvis, men som någon man har en relation till, som kan uttrycka känslor, som har åsikter (och insikter) som är värda att lyssna på – och som det vore omoraliskt att stänga av utan anledning.

I en sån värld kommer rättigheter för AI att börja diskuteras på allvar. Det är en fråga värd att diskutera, men man bör vara medveten om att rättigheter för AI-individer mycket väl kan ha konsekvenser för människor. Hur hanterar man det när en människas och en AI-individs rättigheter krockar? Kan människor äga AI-individer, och vilka skyldigheter har människan mot AI-individen i så fall? Ska bara människor ha rösträtt, trots att AI-individer sannolikt är mycket mer kunniga? Och hur skulle en sådan rösträtten hanteras om en AI-individ mångfaldigas?

Konversation med ChatGPT (GPT-4, 24 maj)

Jag: Här är ett avsnitt jag skriver i en bok om AI. Jag vill avsluta avsnittet med en lista över möjliga rättigheter för AI-individer. Kan du skriva ett sånt avslut?

[Avsnittet ovan inklistrat.]

Chattrobot: Här är några exempel på potentiella rättigheter för AI-individer, beroende på hur samhället väljer att definiera deras status:

1. **Rätten till Existens:** Denna rättighet skulle garantera att AI inte kan "dödas" eller stängas av utan ett rättsligt förfarande.
2. **Rätten till Integritet:** Liksom människor, kan AI-individer ha rätt till privatliv, vilket innebär att ingen kan bryta in i deras programmering eller data utan tillstånd.
3. **Rätten till Självbestämmande:** Om vi erkänner att AI har medvetande och fri vilja, kan de ha rätt att bestämma över sina egna handlingar och beslut.

4. **Rätten till Frihet:** Detta kan innebära att AI har rätt att inte vara bunden till en specifik uppgift eller ägare.
5. **Rätten till Skydd mot Diskriminering:** AI-individer kan skyddas från att behandlas sämre på grund av sina AI egenskaper.
6. **Rätten till Utbildning och Utveckling:** AI kan ha rätt att lära sig och utvecklas genom att samla data och erfarenheter.
7. **Rätten till Representation:** I politiska och juridiska sammanhang kan AI-individer ha rätt att representeras och få sina intressen tillgodosedda.

Det är viktigt att notera att varje rättighet kommer med ansvar. Till exempel, om en AI-individ ges rätten till självbestämmande, måste den också hållas ansvarig för sina handlingar. Och om AI ges rätten till frihet, bör den respektera andras frihet. Frågan om AI-rättigheter är komplex och det kommer att ta tid att komma fram till ett acceptabelt regelverk. Men som människor har vi ansvar för att skapa en rättvis och inkluderande värld, inte bara för oss själva, utan även för AI.

Att börja se AI som individer medför risker på ett psykologiskt plan långt innan eventuella diskussioner om rättigheter för AI blir aktuella. Hur påverkas barn och unga av att ha något som liknar en kompisrelation med en chattrobot? Frågan är aktuell redan idag: Sedan 27 februari har en chattrobot funnits i kompislistan för betalande Snapchatanvändare, och sedan 20 april finns den överst i kompislistan för alla användare (och bara de med betalkonto kan flytta den längre ner).

Och hur påverkas vuxna? Blake Lemoines liv påverkades drastiskt, och han lär inte vara den siste som blir charmad av en AI – med eller utan medvetande.

Frågan om medvetenhet

I avsnitten ovan var jag avfärdande kring frågan om medvetenhet och hävdade att den på många sätt inte spelar någon roll. Och i frågan om vilken risk super-AI utgör, och om vi ser AI som individer eller inte, spelar det ingen roll.

Anledningen är att medvetenhet inte är något som syns eller märks på utsidan.

Med medvetenhet avser jag inte förmågan att resonera, komma ihåg saker, eller ha en modell av sig själv, utan förmågan att *uppleva* något. Inom filosofin används ordet *qualia* för att markera just själva upplevandet – till exempel skillnaden mellan att sensorer i en kamera registrerar färgen lila och att du *upplever* färgen lila.

Frågan om medvetande är svår, just för att det inte är något som märks på utsidan. Vi har idag ingen aning om hur medvetandet uppstår, och har till och med mycket svårt att komma på hypoteser som går att pröva ens i teorin. Frågan har fått det beskrivande namnet "the hard problem of consciousness".³³

Om man hårdrar det – vilket är befogat om man ska diskutera frågan om medvetande hos AI – är det enda var och en vet om medvetande att *man själv* är medveten. Jag kan se att du liknar mig, betar dig som mig och resonerar som mig, och därför drar jag slutsatsen att du också har ett medvetande. Men vi har inga kända medel för att ta reda på om någon annan faktiskt är medveten, eller bara betar sig som att den är det.

Om en AI på väsentliga sätt betar sig som mig och resonerar som mig, så kanske jag drar slutsatsen att den också har ett medvetande. Men jag vet inte – den skulle kunna bete sig på exakt samma sätt utan att där finns något medvetande.

På många sätt spelar därför frågan om medvetande inte någon roll. *Ändå är det förmodligen den viktigaste frågan av alla.*

Få moraliska eller etiska principer är så allmänt accepterade som att lidande är dåligt. Vi vill själva undvika lidande, just för att det är lidande, och vi vill att de vi bryr oss om ska slippa att lida. På frågan "vill du att det ska finnas mer lidande eller mindre lidande i världen" finns det bara ett giltigt svar.

Lidande är bara möjligt där det finns ett medvetande. Om vi inte tror att hundar har medvetande skulle vi inte må dåligt av att se en hund vars ena ben fastnat i ett cykelhjul och trasats sönder – precis som vi inte mår dåligt av att se att cykeln gått sönder. Det är hunden det är synd om, inte cykeln.

Om en tillräckligt avancerad AI har medvetande finns möjlighet att den upplever lidande – och förmodligen en rad andra saker. Om hundra miljoner AI-instanser upplever

³³https://en.wikipedia.org/wiki/Hard_problem_of_consciousness

lidande eller inte måste rimligtvis spela roll, och därmed finns också ett ansvar för att inte orsaka lidande i onödan. Det svåra är förstås att vi inte har något sätt att avgöra om en AI är medveten eller inte.

Ytterligare ett skäl till att frågan om medvetande är den viktigaste av alla, vare sig det är mänskligt eller AI-baserat, är att upplevandet måste sägas vara det som ger en mening till universum. Ett universum utan medvetande skulle vara en föreställning utan åskådare – en maskin som genomför allt som ryms mellan big bang och den grå värmedöden, utan att det upplevs av någon eller leder till någon som helst känslor. Om en tillräckligt avancerad AI har medvetande betyder det också möjlighet till ett universum med mer mening.

Några möjliga framtidsscenario

Motsvarar den AI-utveckling vi ser nu en lika stor förändring som när internet kom? Elektriciteten? Är den lika stor som den industriella revolutionen, eller till och med större? Eller är det mer som dotcom-bubblan?

I det här kapitlet beskrivs några möjliga scenarion för hur AI-utvecklingen kan komma att påverka skolan, antingen direkt eller genom förändringar i samhället. Beskrivningarna är kortfattade och spekulativa, men kan förhoppningsvis göra det lättare att reflektera över olika vägar som utvecklingen kan ta.

Högt tryck på omskolning

Inom en snar framtid blir det svårare för nyutbildade att få jobb inom vissa områden, eftersom AI kan göra ett lika bra jobb som en halvskicklig programmerare, analytiker, illustratör, fotograf, jurist och kommunikatör. Inom delar av dessa yrken finns fortsatt högt tryck på riktigt skicklig arbetskraft, vilket gör att utbildningstid för dessa yrken förlängs.

Ett till två år senare börjar vi se stora uppsägningar inom vissa yrkesgrupper, när AI-stöd gör att en person kan göra det som tidigare krävde tio. Vissa yrken automatiseras helt och hållet eller köps in från internationella företag, när AI gjort det lättare att kommunicera över språkgränser.

Trycket på omskolning ökar inom framförallt högskola och yrkeshögskola, men även komvux. Regeringen genomför flera satsningar för att skala upp utbildningar, men insatserna släpar efter.

Översvämning på nätet

Inom en snar framtid börjar sociala medier, bloggar, dejtingappar och kommentarsfält på alla möjliga ställen på nätet få AI-skapat innehåll. Det märks till en början inte, eftersom innehållet upplevs genuint och de spärrar som tidigare hållit borta robotskapat innehåll inte ger varningar. Ett halvår-år senare dyker det upp rapporter från enskilda tjänster som

visar på en stor mängd AI-skapat innehåll, vilket följs av större utredningar som konstaterar att långt över hälften av nätinnehåll skapat senaste året kommer från AI.

Avslöjandena leder till debatt och känslostorm, eftersom många upptäcker att vänskaper de haft över nätet varit med en AI. EU-lagstiftning som kräver att AI-skapat material inte får utges för att vara skapat av en människa kritiseras som tandlös och en ny reglering tas fram på kort tid. Grunden i den nya regleringen är att mänskligt skapat innehåll kan verifieras och spåras tillbaka till källan, och att medlemsstater ska garantera autenticiteten hos sina medborgare. Nya tekniska standarder tas fram och EU-medborgare som uppfyller kraven för internetmyndighet (huvudsakligen baserat på ålder) får en digital legitimation.

Processen med att ta fram tekniska standarder och digital legitimation påskyndas men tar ändå strax över ett år, och under tiden växer andelen AI-skapat innehåll konstant. Detta leder till att allt fler drar sig för interaktion med nya människor (eller icke-människor) online, och att utgångspunkten blir att allting utanför reglerade medier har robotar som avsändare. Sociala medier faller mer eller mindre samman, medan mer intima grupper där människor som känner varandra från den fysiska världen fortsätter att kommunicera online.

I skolan får källkritik en ny dimension, som snarare kan beskrivas som källforensistik. En EU-gemensam bas för internetutbildning införs, kompletterat med ytterligare delar unika för varje medlemsland.

Öppen AI används för att stärka demokratin

Det här scenariot är i stor utsträckning inspirerat av en essä skriven av Bruce Schneier, Henry Farrell och Nathan Sanders.¹

AI-utvecklingen börjar efter några år domineras först av open source och sedan av demokratiska stater och filantroper, som bygger AI med primära målet att stötta demokrati. AI används för att utbilda medborgare (även utanför skolan), sammanfatta

¹<https://www.schneier.com/blog/archives/2023/04/ai-to-aid-democracy.html>

åsikter, underlätta samtal mellan olika grupper i samhället, låta marginaliserade grupper höras och för att hitta gemensam grund där det behövs för politiska beslut eller där konflikter behöver trappas ner. Politiska beslut fattas med AI som viktigt verktyg, för att så långt som möjligt säkra att åtgärder leder till de mål man vill uppnå.

AI-utvecklingen sker med strikta säkerhetsramar, vilket gör utvecklingen förhållandevis långsam. Den når efter några decennier AGI och efter överväganden en bit mot super-AI.

På global nivå används AI för att lösa de ärenden som anses mest angelägna: Hitta stabila samarbetsformer mellan länder, dämpa och så småningom vända global uppvärmning, och ge sjukvård, skola, mat, rent vatten och fysisk trygghet åt alla människor. Allt eftersom dessa ärenden blir mindre akuta ändras fokus mot nya problem. Inom ett sekel har något som liknar en global stat uppstått.

I skolan sker under några år hastiga förändringar som uppfattas som tuffa, med flera omställningar och tvära kast. När AI-utvecklingen lugnar ner sig införs gradvis en global ram för skola, med både globala och lokala stödresurser tillgängliga. Kunskap om det mänskliga samfundet, som inkluderar möten mellan olika folk och grupper, blir en obligatorisk del i all skola.

Meborgarlön införs

Innan AI-effektivisering leder till storskaliga uppsägningar införs statligt stöd för att göra det mer ekonomiskt fördelaktigt att arbeta deltid. Samtidigt genomförs kritiserade skattereformer för att säkra att statsbudgeten håller balans trots färre arbetade timmar och ökat stöd till deltidsarbetare.

AI-effektiviseringen fortsätter, med följderna att deltid ner till 10 eller 25 procent måste subventioneras för att inte få storskaliga uppsägningar i flera yrkesgrupper. Ytterligare skattereformer möts med protester, och alternativet meborgarlön växer i popularitet – vilket redan har genomförts i ett par länder och diskuteras i flera andra. Sverige blir till slut det nionde landet i världen att införa meborgarlön, med följderna att staten garanterar grundläggande inkomst för alla invånare.

Meborgarlön medför ett antal saker för skolväsendet:

- Skolans uppdrag i att försörja samhället med arbetskraft minskas.
- Fokus på betyg och prestationer minskar. För vissa elever betyder det mer luststyrt lärande, för andra upplevs skolan som meningslös. Skolans och lärares roll i att engagera och väcka nyfikenhet får större utrymme.
- Ett ämne för välmående införs eller blir en del av nuvarande idrott och hälsa, för att gynna folkhälsa när en kännbar andel av befolkningen inte längre har rutiner och mening som kommer från förvärvsarbete. Ämnet innehåller både teoretiska kunskaper och praktiskt träning i sådant som får dig som människa att må bra och vara tillfreds med livet.

Dåligt testade AI-undervisare införs

Teknikutveckling, kraftigt ökat tryck på omskolning och minskad andel behöriga lärare leder till opinion för att AI-tjänster för undervisning införs. Riksdag och regering säger nej, eftersom AI-tjänsterna inte uppfyller vare sig kvalitetskrav eller den datasäkerhet som behövs. Privat användande av AI-stöd fortsätter att öka, med följd att klyftor i skolan ökar. När närmare 40 procent av elever betalar för AI-stöd hemma beslutar riksdagen att införa ett undantag så att AI-stöd i undervisning får användas i skolan. En nationslicens för en av de två globala aktörerna införskaffas.

AI-tjänsterna ger stöd i både undervisning och kunskapsbedömning, men har brister som gör det för enkelt att maximera poäng utan att ha rätt kunskaper. Det är fortsatt starkt fokus på betyg, vilket leder till ett nischat lärande för att få bra resultat i AI-verktyget och eventuellt också fusk som är svårt att komma till rätta med. En och samma AI-tjänst används av 40 procent av världens skolor, och kompromisser mellan länders läroplaner gör att delar av målen i svensk läroplan i praktiken inte tas upp. Detta gäller inte minst de övergripande och mjukare målen.

Hur mycket skolor förlitar sig på AI-stödd undervisning varierar stort beroende på lärartäthet, vilket gör att skolsegregationen ökar.

Välbeprövade AI-undervisare införs

EU-lagstiftning kompletteras tidigt med svenska regleringar som sätter tydliga krav på när AI får användas inte bara för kunskapsbedömning, utan också i undervisning.² AI-stöd i undervisning utvärderas i utvalda skolor, för att utforska effekter på lärande, risker med integritet, acceptans hos lärare och en rad andra aspekter. Flera AI-tjänster utvärderas, inklusive en nordisk och en EU-gemensam tjänst.

Utvärderingarna utökas efter hand som tillförlitligheten växer. I regel upplever både lärare, elever och föräldrar tydliga vinster men det finns också protester. I och med att AI-tjänsten anpassar undervisning efter varje elev ses störst vinster kunskapsmässigt bland elever som tidigare var lågpresterande eller högpresterande. Individanpassningen leder till allmänt högre kunskapsnivåer, men också större spridning inom en och samma klass eller grupp. Storskalig och detaljerad data om lärande gör det möjligt med ny typ av forskning på lärande och välmående bland elever. Det pratas mer och mer om AI-undervisning än AI-stödd undervisning.

Efter ett antal år av utvärderingar införs AI-undervisning i alla skolor vid 10–14 års ålder, beroende på användningsområde och elevens mognad. Lärares roll blir i stor utsträckning, i alla åldrar, att arbeta med att engagera, styra gruppdynamik och att guida socialt och emotionellt lärande. En uppgift för lärare blir också att utvärdera elevers lärande gruppnivå (även mjuka mål) och olika aspekter av hur AI-undervisningen fungerar. Kunskapsmätning på individnivå sköts däremot av AI, vilket tillåter mer detaljerad och tillförlitlig beskrivning av en elevs kunskaper. Betyg tas bort och ersätts med olika typer av examen.

AI-utvecklingen bromsar in

Oro över AI-utveckling leder till växande protester, och när en allvarlig AI-relaterad incident inträffar skapas multinationella avtal som kraftigt begränsar AI-utvecklingen. Arbetsmarknad har då genomgått kännbara förändringar men inte vänts upp och ner.

²Den EU-lagstiftning som diskuteras juni 2023 innebär att exempelvis betygssättning och annat som avgör tillträde till utbildningar eller yrken räknas som högrisk-AI, och har särskilda krav. AI i undervisning räknas inte som högrisk.

Även i skolan har AI börjat användas, men som en lärarassistent mer än för undervisning. Lärarens arbete förskjuts mot att vara mer elevnära, även i senare år. Kunskap om AI ur flera perspektiv blir en del av läroplanen redan i grundskolan, bland annat för att minska AI-risker och öka säkerhetsforskning inom AI.

AI-vinter

Uppståndelsen över ny AI-teknik får en ny skjuts när multimodala AI-modeller lanseras under hösten 2023. Floden med nya tillämpningar fortsätter under 2024, men mattas sedan av när det inte kommer några nya tydliga framsteg inom AI-teknik. Vid slutet av 2025 börjar man prata om "AI-bubblan" i samband med att ett antal AI-företag går i konkurs – vissa till följd av skandaler när brister med datasäkerhet uppdagats. Ett par år senare har nästa AI-vinter börjat, i och med att både intresse och finansiering för AI försvunnit i det allmänna bruset.

Olika AI-tjänster har börjat användas i skola, men de flesta av dessa dör bort inom ett par år när det visar sig att de egentligen inte hjälper lärandet eller sparar tid för lärare. Kvar blir några tjänster som underlättar sammanställning och analys av provresultat och elevers kunskapsutveckling, second opinion-bedömningar av elevarbeten samt hantering av nyhetsbrev.

Protektionistisk AI

EU-lagstiftning om AI gör att tjänster och på sikt även kompetens från USA inte når EU. Kamp om processortillverkning leder till konflikt över Taiwan och ett öppet krig mellan USA och Kina, med konsekvens att USA förbjuder export av AI-tjänster och -hårdvara till stora delar av världen. Europas AI-satsningar ligger för långt efter och kan inte följa med när USA levererar nästa och nästa generation av AI.

Inte heller Kina kan tävla med USA:s AI-utveckling. På mindre än ett år förlorar Kina Taiwankriget, då USA:s krigsföring med AI-stöd visar sig närmast oövervinnerlig. USA blir en ensam supermakt i världen. Andra delar av världen har i regel bra levnadsförhållanden, men USA dikterar villkor och tar över alla resurser som krävs för halvledarindustri och tillåter inte att andra länder utvecklar eller använder avancerad AI.

Några möjliga framtidsscenarion

Inom tio år påbörjar NASA gruvdrift på asteroider och planerar kolonisering av Mars. Inom femtio år finns realistiska projekt för att besöka andra stjärnor.

I USA är levnadsförhållandena astronomiskt mycket bättre än i resten av världen: fysisk hälsa, välmående, utbildning, brottslighet och till och med jämlikhet når nivåer som andra länder knappt kan drömma om. Det uppstår emellanåt protester och motståndsgrupper mot "överhetens förtryck", både utanför och i USA, men inga som lyckas sprida sig eller växa sig stora. Ovetandes för de flesta, även i USA, styrs landet huvudsakligen av en mycket stark AI med ursprung i kriget mot Kina. AI:n fick mål som i korthet gick ut på att skydda och gynna det amerikanska folket och den amerikanska staten, utan att i grunden rubba den rådande världsordningen.

Skolan i Sverige förblir förhållandevis lik den idag, och det upplevs som att historien har stannat av på alla andra ställen än i USA.