ChatGPT

Tekniska aspekter av AI-litteracitet

Introduktion:
Artificiell intelligens (AI) påverkar i snabb takt vårt samhälle och utbildningsväsen. Begreppet **AI-litteracitet** har vuxit fram för att beskriva de kunskaper och färdigheter som krävs för att förstå, använda och kritiskt värdera AI-system på ett säkert och effektivt sätt [1] [2]. AI-litteracitet handlar inte bara om att kunna hantera AI-verktyg praktiskt, utan också om att ha en grundläggande teknisk förståelse för hur AI fungerar och vilka begränsningar och etiska utmaningar som finns. EU-kommissionen och OECD betonar att både elever och lärare behöver dessa kompetenser för att *"förstå och interagera med AI-system på ett självständigt och kritiskt sätt"* [3]. I synnerhet lyfts det fram att unga människor bör utrustas med *kunskap om grundläggande AI-begrepp, hur AI tränas, samt förmåga att reflektera över AI:s effekter på individ och samhälle* [4].

För blivande och verksamma lärare är AI-litteracitet extra viktigt. Lärare förväntas inte bli AI-forskare, men de behöver tillräcklig teknisk insikt för att kunna integrera AI på ett pedagogiskt meningsfullt sätt och för att hjälpa elever att förstå AI kritiskt. UNESCO lanserade 2024 särskilda kompetensramverk för AI inom utbildning – ett för elever och ett för lärare – just för att vägleda länder i att utveckla dessa förmågor [5] [6]. I lärarramverket betonas bl.a. *"AI foundations and applications"*, dvs. de tekniska grunder som lärare bör behärska [7]. Samtidigt visar forskning att AI-litteracitet fortfarande är ett nytt område inom lärarutbildning, där det hittills fått begränsat utrymme [8]. Denna rapport syftar därför till att ge en populärvetenskaplig men ändå noggrann genomgång av de tekniska nyckelkomponenterna i AI-litteracitet – anpassad för lärare i grund- och gymnasieskolan samt lärarstudenter.

Nedan följer fem huvudkomponenter som identifierats (i linje med EU:s och andra ramverk) som centrala kunskapsområden för AI-litteracitet hos lärare. Under varje rubrik förklaras begreppen utförligt på ett sätt som är tänkt att vara begripligt för lärarstuderande, med referenser till aktuella källor inom AI, pedagogik och teknologididaktik.

1. Grundläggande begrepp

En första byggsten i AI-litteracitet är att förstå några grundläggande tekniska begrepp. Dessa utgör grunden för att senare kunna sätta sig in i hur AI-system fungerar. Här förklaras fem centrala begrepp:

Algoritm

Begreppet **algoritm** avser en tydlig uppsättning instruktioner för hur en uppgift ska lösas steg för steg. En algoritm kan liknas vid ett recept eller en manual: den tar emot någon form av indata, bearbetar dessa genom en serie definierade steg, och genererar sedan ett resultat (utdata) [9]. Algoritmer är kärnan i all datorbehandling – varje datorprogram är i grunden en samling algoritmer. För att konkretisera: en sorteringsalgoritm är en instruktion för hur en lista av tal eller ord kan ordnas i storleksordning, medan en krypteringsalgoritm beskriver hur data kan omvandlas för att skyddas. Algoritmer kan vara allt från mycket enkla (t.ex. en algoritm för att beräkna medelvärdet av två tal) till extremt komplexa med tusentals steg [10]. I AI-sammanhang används algoritmer för att styra hur en AI löser problem. Till exempel består många AI-system av algoritmer som lär sig känna igen mönster i data, vilket leder oss till nästa begrepp, *maskininlärning*.

Maskininlärning

Maskininlärning (*machine learning*) är en gren av AI som fokuserar på att låta datorer lära sig direkt från data, istället för att en människa i förväg programmerar alla regler. Traditionell programmering (regelbaserade system) kräver att utvecklare skriver explicita instruktioner för alla tänkbara situationer. I maskininlärning ger man istället algoritmen stora mängder exempeldata och låter den själv hitta mönstren och "reglerna" inuti dessa data [11]. En ofta använd definition är att en maskininlärningsalgoritm *förbättrar sin prestanda på en uppgift ju mer erfarenhet (data) den får*. Till exempel kan en maskininlärningsmodell tränas att känna igen handskrivna siffror genom att visa den tusentals bilder av olika handskrivna siffror tillsammans med rätt svar; modellen justerar sig själv tills den lärt sig skilja på siffrorna. Maskininlärning kan betraktas som en form av statistisk mönsterigenkänning – algoritmen hittar sannolikhetsmönster i data för att göra förutsägelser [12]. Denna datadrivna metod har gjort det möjligt att lösa mer komplexa problem än vad som var möjligt med enbart fasta, förprogrammerade regler.

Djupinlärning

Djupinlärning (*deep learning*) är en underkategori av maskininlärning som har fått mycket uppmärksamhet det senaste decenniet. Djupinlärning bygger på **neurala nätverk** med många lager av artificiella neuroner – därav ordet "djup" (många lager) [13]. Den stora skillnaden mot mer traditionella maskininlärningsmetoder är att djupa neurala nätverk själva kan lära sig vilka egenskaper i data som är viktiga, istället för att en människa manuellt behöver välja ut och mata in relevanta egenskaper. Djupinlärning har visat sig vara mycket kraftfullt för att lösa komplexa uppgifter som bild- och röstigenkänning, språköversättning och avancerade strategispel. Exempelvis använder moderna bildigenkännings-AI djupa neurala nätverk för att automatiskt lära sig känna igen kanter, former och objekt i bilder – nätverket lär sig i de första lagren att känna igen enkla mönster (som linjer), i mellanskikten mer komplexa former (som ögon eller hjul), och i de djupaste lagren hela objekt (som ansikten eller bilar). Tack vare djupinlärning har AI-prestanda gjort stora språng, men tekniken kräver ofta **stora datamängder** och beräkningskraft för att tränas effektivt.

Neurala nätverk

Ett **neuralt nätverk** är en modell inom AI inspirerad av hur nervceller i den biologiska hjärnan samverkar. I en hjärna tar neuroner emot signaler, bearbetar dem och skickar vidare signaler till andra neuroner. På liknande sätt består ett artificiellt neuralt nätverk av många sammankopplade noder ("neuroner") ordnade i lager [14]. Varje nod tar emot numeriska signaler (data), utför en enkel beräkning och skickar resultatet vidare till nästa lager. Ett neuralt nätverk har typiskt ett *inskiktslager* (input layer) där data matas in, ett eller flera *dolda lager* (hidden layers) där beräkningar sker, och ett *utskiktslager* (output layer) som producerar nätverkets slutgiltiga svar. Genom att justera styrkan på kopplingarna mellan noderna (vikter) under träning kan nätverket lära sig att ge önskade utdata för givna indata. Till exempel kan ett neuralt nätverk tränas att förutse elevbetyg utifrån data om studievanor; under träningen anpassas vikterna så att nätverket lär sig vilka faktorer som indikerar högre eller lägre betyg. Neurala nätverk är basen för djupinlärning – när ett nätverk har många lager talar man om ett *djupt* neuralt nätverk. Dessa modeller är mycket kraftfulla men kan också vara svåra att tolka, då de fungerar som komplexa "svarta lådor" där det inte alltid är uppenbart vilka "regler" de lärt sig.

Träningsdata

Träningsdata kallas den datamängd som används för att *lära* (träna) en AI-modell. Om algoritmer är recepten, så är träningsdatan de ingredienser som behövs för att koka ihop AI-modellen. Kvaliteten och kvantiteten på träningsdatan är direkt avgörande för hur bra modellen blir [15]. En

maskininlärningsmodell “ser” ofta inga andra data än just det träningsmaterial den fått under sin inlärning, så modellen blir i princip en spegel av dessa data. Till exempel, om man vill träna en AI att känna igen olika trädarter från bilder, behöver man en omfattande uppsättning bilder på träd som är korrekt märkta med vilken art de föreställer – det är träningsdatan. Modellen kommer att justera sina parametrar utifrån dessa exempel. Generellt gäller att ju mer *representativ* och omfattande träningsdata är, desto bättre brukar modellen prestera på nya fall. Träningsdata kan vara *märkt* (labeled), vilket innebär att rätt svar eller kategori finns angiven för varje exempel (t.ex. "Denna bild är en gran, denna är en tall"), eller *omärkt* (unlabeled) där modellen själv måste hitta mönster utan facit. **Storleken** på träningsdatan spelar roll – ofta behövs det **enorma datamängder** för att träna avancerade AI-system framgångsrikt [16]. Samtidigt är **datakvaliteten** kritisk: “Skräpinformation in ger skräpinformation ut” brukar man säga – bristfällig eller snedvriden träningsdata leder till en bristfällig eller partisk AI-modell.

2. Datadriven AI

En central teknisk aspekt av modern AI är att den är **datadriven**. Till skillnad från äldre, regelbaserade system, som enbart följde i förväg uppställda regler, bygger datadriven AI på att modellerna *lär sig* direkt från data. Här går vi igenom hur AI tränas med data och varför datamängdens storlek och kvalitet påverkar resultatet.

Hur AI tränas med data

Att *träna* en AI-modell kan liknas vid hur vi människor lär oss av erfarenheter. En AI får “erfarenhet” genom att exponeras för många exempel (träningsexempel) och justera sitt beteende utifrån dem. Rent praktiskt går det till så att man matar in träningsdata i modellen och låter en algoritm iterativt anpassa modellens interna parametrar (t.ex. vikter i ett neuralt nätverk) för att minimera felaktiga svar. I början gissar modellen kanske slumpmässigt, men efter varje exempel justeras den lite grand mot bättre prestation. Denna process upprepas tusentals eller miljontals gånger. Till exempel kan en maskininlärningsalgoritm som ska lära sig skilja på om en mening är grammatiskt korrekt eller inte, genomgå en träningsprocess där den stegvis justerar sig efter att ha fått återkoppling på många olika meningar. Så småningom har modellen justerats så mycket att den “lärt sig” språkets mönster. **Maskininlärning innebär alltså att modellen själv skapar en generell modell av data** – exempelvis statistiska samband – istället för att vi programmerar in reglerna för hand [11]. En viktig insikt för AI-litteracitet är att denna inlärning inte är magisk: modellen kan bara dra slutsatser baserat på de data den sett. Om träningsdatat innehåller tydliga mönster, kommer modellen att fånga dem; om datat är otydligt eller missvisande kan modellen få svårt att lära sig rätt sak.

I EU:s och andra ramverk framhålls att lärare och elever behöver förstå åtminstone i grova drag *hur* AI-modeller tränas [17]. Det innebär till exempel att känna till att en träningsprocess ofta kräver en målformulering (en **mål- eller förlustfunktion** som definierar vad som är “rätt”), en algoritm för optimering (som styr hur modellen anpassas för att bli mer korrekt) och en uppdelning av data i tränings- och testmängder (för att undvika att modellen bara “memoriserar” exemplen). Allmän förståelse för dessa principer hjälper läraren att inse varför en AI-modell ger de svar den ger, och varför den ibland kan misslyckas.

Datamängdens påverkan på resultat

”Mer data slår smartare algoritmer” är ett talesätt inom AI som antyder hur viktig datamängden är för resultatet. I många fall har det visat sig att en enklare algoritm med mycket data kan prestera bättre än en avancerad algoritm med lite data. Stora datamängder ger modellen mer “erfarenhet” att lära ifrån och ökar chansen att den hittar de generella mönstren istället för att råka lära sig irrelevanta detaljer. De senaste årens genombrott inom AI – som avancerad bildigenkänning och språkmodeller – hänger

intimt samman med tillgången till massiva datamängder (tillsammans med kraftfullare datorer). *Tremendous amounts of data are required* för att avancerade lärandealgoritmer ska lyckas [16], konstaterar AI-forskare bakom AI4K12-ramverket. Med andra ord: ju större och mer omfattande träningsdata, desto bättre utgångsläge för modellen att bli träffsäker.

Det är dock viktigt att lärare förstår att det inte bara handlar om **kvantitet** utan också om **kvalitet** på data. En modell tränad på ensidig eller partisk data kan ge missvisande resultat även om datamängden är stor. Exempelvis kan en AI som ska rekommendera historiska romaner till gymnasieelever prestera dåligt om all träningsdata råkar komma från vuxna läsare – trots att datamängden är stor fångar den inte ungdomars preferenser. Även OECD och EU understryker att utbildningsinsatser om AI bör omfatta diskussioner om datakvalitet, representativitet och källkritik när det gäller träningsdata [18] [19]. För lärare innebär detta att de bör vara medvetna om frågor som: Vilka data har använts för att träna en viss utbildnings-AI och speglar de de elevgrupper som AI:n ska användas med? Hur påverkar storleken på en dataset tillförlitligheten i AI:ns prediktioner? Genom att ställa sådana frågor kan man bättre förstå och också förklara för elever varför en AI fungerar som den gör.

3. Modellförståelse

En annan viktig komponent i AI-litteracitet är att förstå olika **angreppssätt för AI-modeller** – framför allt skillnaden mellan **regelbaserad AI** och **inlärningsbaserad AI**. Dessa två paradigm representerar hur AI-system har utvecklats över tid och har olika egenskaper som är bra att känna till.

Regelbaserad AI (ibland kallad expertssystem eller symbolisk AI) bygger på explicita regler och logik som människor programmerar in. Tänk er ett system som består av en mängd *”OM (if) – SÅ (then)”-regler*: till exempel *“OM temperaturen är under 0°C OCH det regnar, SÅ ska prognosen säga snöfall”*. I ett regelbaserat system sker slutsatser utifrån att dessa fördefinierade regler triggas. Fördelen med denna ansats är att den är relativt transparent – man kan ofta följa en kedja av regler för att se hur systemet resonerade. Nackdelen är att det krävs mycket arbete av experter för att täcka in alla fall, och att systemet har svårt att hantera situationer som inte förutsetts i regelförrådet. Regelbaserade AI-system var vanliga under 1900-talets andra hälft, t.ex. i tidiga språköversättningsprogram eller schackdatorer som baserade sig på handskrivna regler.

Inlärningsbaserad AI, å andra sidan, är det vi avser med maskininlärning (inklusive djupinlärning). Här finns inga explicita “if-then”-regler förprogrammerade för varje scenario. Istället *lär sig modellen själv* genom exempel, som beskrivet ovan. *Modellen skapar sina egna ”regler” eller mönster baserat på data* [11]. Konsekvensen av detta är att inlärningsbaserade system ofta är mer flexibla och kan prestera bättre i komplexa uppgifter (som bildigenkänning eller att spela Go) där det är i princip omöjligt för människor att skriva alla nödvändiga regler manuellt. Dock blir modellen mer av en **”svart låda”** – det är inte alltid uppenbart varför den kom fram till ett visst beslut, eftersom kunskapen är distribuerad i modellens parametrar snarare än i tydliga regler.

Det är väsentligt att inse att dessa två typer av AI *inte* är helt isolerade från varandra. Många praktiska system kombinerar båda: man kan till exempel ha maskininlärningskomponenter inuti ett regelbaserat ramverk. Men ur ett pedagogiskt perspektiv hjälper det att kontrastera dem. Regelbaserade system följer **deterministiska** mönster (samma indata ger alltid samma utdata och logiken ändras inte förrän en människa justerar en regel) medan inlärningsbaserade system är **probabilistiska** och adaptiva (de justerar sig baserat på data och ger svar med vissa sannolikheter, se nästa avsnitt). Som GeeksforGeeks uttrycker det: *“Rule-based systems follow explicit rules created by human experts… On the other hand, machine learning systems learn from data and use patterns… and can adapt and improve over time”* [11]. Regelbaserad AI kan sägas ha sin styrka i tydlighet och förutsägbarhet (man vet exakt vilka regler som

finns) men är ofta begränsad i komplexitet, medan inlärningsbaserad AI utmärker sig i att hantera komplexa, diffusa problem – dock på bekostnad av transparens och med behov av mycket data.

För läraren och lärarstudenten är det viktigt att förstå denna skillnad eftersom den påverkar hur man ska förhålla sig till AI-verktyg. Ett regelbaserat undervisningsprogram (t.ex. en glosmaskin som följer fasta regler för vilka frågor som ställs) beter sig *konsekvent* enligt sina regler – fel ligger då oftast i regelverket och kan rättas av utvecklaren. Ett lärande system (t.ex. en AI som rättar elevtexter baserat på språkmönster den lärt sig) kan däremot uppvisa oväntade beteenden om den stött på annorlunda data än vad den tränats på. För att kunna felsöka eller lita på AI-verktyg behöver lärare veta om de är regelbaserade (då kan man söka fel i regeluppsättningen) eller inlärningsbaserade (då bör man titta på träningsdatan och kanske utvärdera prestandan på olika typer av exempel). EU:s kommande AI-litteracitetsramverk väntas inkludera förståelse för denna skillnad som en nyckelkunskap, eftersom den är grundläggande för att utöva **kritisk omdömesförmåga** gentemot AI-system i skolan [4].

4. Prediktion och sannolikhet

Till skillnad från traditionella datorprogram, som ofta ger **deterministiska** resultat (samma indata ger alltid samma utdata), arbetar många AI-modeller utifrån **sannolikheter och prediktioner**. Det innebär att AI:ns svar eller beteende inte alltid är exakt förutbestämt, utan bygger på statistiska uppskattningar av vad som är mest troligt korrekt. För att förstå AI-teknologin bör lärare känna till hur dessa sannolikhetsbaserade svar fungerar.

En maskininlärningsmodell, särskilt inom områden som klassificering, ger oftast i *utdata* en **sannolikhetsfördelning** över möjliga svar. Till exempel kanske en bildigenkännings-AI, given en bild, beräknar 90% sannolikhet att bilden föreställer en katt, 8% att det är en räv och 2% att det är en hund. Därefter väljer modellen det alternativ med högst sannolikhet (”katt”) som sitt svar. Sannolikhetssiffran – ofta kallad **konfidensnivå** eller **confidence score** – anger hur säker modellen är på sin prediktion [20]. En konfidens på 90% kan tolkas som att *”om modellen fick många liknande fall, skulle den gissa katt i 9 av 10 fall av dessa”*. En tumregel är att ju högre sannolikhet/confidence, desto mer tillförlitligt anser modellen sitt svar vara [21].

Att AI:ns prediktioner är sannolikhetsbaserade har flera följder. För det första kan modellen ha *fel även om den är säker*: i ovanstående exempel kanske bilden faktiskt var en räv trots att modellen var 90% säker på ”katt”. Den 10-procentiga risken slog in. För det andra kan svaren variera om samma modell körs vid olika tillfällen eller med små förändringar – särskilt om modellen innehåller någon slumpmässighet eller om den tränas om med nya data. Det betyder att två chattbotar som bygger på samma underliggande språkmodell kan ge lite olika svar på samma fråga, eller att en och samma AI kan svara olika före och efter en uppdatering. AI-system, särskilt de baserade på neurala nätverk, är alltså inte statiska utan *lärande och adaptiva*, vilket innebär att deras beteende ska förstås statistiskt. Man brukar säga att *AI gör prediktioner, inte uttalanden av sanning*.

För lärare är detta viktigt på två sätt: Dels måste man tolka AI-verktygs output med insikten att det är *prognoser med en viss osäkerhetsmarginal*, inte facit givna ex cathedra. Dels kan man behöva förklara för elever varför en AI “tvekade” eller gav olika besked – då är begrepp som sannolikhet och konfidens nyttiga. Faktum är att detta ger en möjlighet att knyta an till matematik (statistik) i undervisningen; AI-litteracitet handlar också om att förstå grundläggande sannolikhetslära i AI-sammanhang [17]. En elev som frågar “Hur vet det här programmet svaret?” bör kunna få förklarat att *”programmet vet inte med 100% säkerhet, men det har räknat ut att ett visst svar har högst sannolikhet baserat på vad det lärt sig tidigare”*. Till syvende og sidst uppmuntrar AI-litteracitet ett kritiskt förhållningssätt: man tar AI:ns svar

som kvalificerade gissningar, som ibland kan behöva kontrolleras av en människa – särskilt om konsekvenserna av ett fel är stora.

5. Begränsningar i AI-teknologi

Trots imponerande förmågor har AI-teknologin **begränsningar** som är viktiga att känna till. Att vara AI-litterat innebär att man har en realistisk bild av vad AI *inte* klarar av eller vilka problem som kan uppstå, så att man varken överskattar eller blint förlitar sig på tekniken. Här tar vi upp tre centrala begränsningsaspekter: **överträning (overfitting)**, **generaliseringsproblem** samt **bias (snedvridning) i träningen**.

Överträning (overfitting)

Överträning syftar på när en AI-modell har lärt sig *för mycket* av detaljerna i just sin träningsdata, på bekostnad av sin förmåga att fungera på nya data. Modellen har då inte bara snappat upp de allmänna mönstren, utan också råkat memorera brus, slumpmässiga variationer eller fel som fanns i träningsdatamängden [22]. En övertränad modell presterar ofta utmärkt på den data den tränats på (den kan återge svaren den "lärt sig" nästan felfritt), men misslyckas när den ställs inför *nya* exempel – den kan inte generalisera kunskapen. Som IBM förklarar: *"overfitting occurs when an algorithm fits too closely or exactly to its training data, resulting in a model that can't make accurate predictions on any data other than the training data"* [23]. Överträning besegrar själva syftet med maskininlärning, eftersom poängen är att modellen ska fungera på *framtida* data, inte bara recitera träningen.

En klassisk analogi är att jämföra med en elev som pluggar inför ett prov genom att memorera svaren på övningsfrågorna ord för ord. Eleven kan då klara övningsfrågorna perfekt (0 fel på repetitionerna hemma), men om provet sedan innehåller samma frågor formulerade på lite annat sätt eller några nya problem, står eleven handfallen – hen har inte *förstått* underliggande principer utan bara memorerat ytan. På liknande sätt har en övertränad AI modell *överanpassat* sig efter träningsexemplen och tappar därför i prestanda på minsta avvikelse från dessa. I praktiken försöker man undvika överträning genom olika metoder: man delar upp data i tränings- och testmängder och övervakar prestanda (om modellen är väldigt bra på träning men dålig på test tyder det på overfitting) [24], man kan avbryta träningen tidigt (*early stopping*), eller införa regelmässigt brus/regelbundna begränsningar i modellen (s.k. *regularisering*) för att hindra den från att bli alltför flexibel och passa in på minsta detalj.

För lärare som använder eller utvärderar AI-system innebär kunskap om överträning att man förstår varför en AI kan fungera bra i vissa situationer men bryta samman i andra. Till exempel: en lästränings-AI kanske fungerade utmärkt i utvecklingsfasen på en viss skola (träningsexemplet), men när den tas till en annan skola med annorlunda elevgrupper presterar den sämre – då kan det bero på att modellen var övertränad på den första skolans elevdata. Att ställa frågor om hur en AI-modell har testats på *nya* data, och att vara försiktig med modeller som inte utvärderats brett, är en del av ett kritiskt förhållningssätt.

Generaliseringsproblem

Generaliseringsförmåga är nära kopplat till överträning, fast ur ett mer allmänt perspektiv. En AI-modells generalisering avser dess förmåga att tillämpa det den lärt sig från träningen på nya, osedda data. Det är just denna förmåga som vi egentligen vill åt med maskininlärning. När en modell inte generaliserar väl säger vi att den har *generaliseringsproblem*. Överträning är en typisk orsak: modellen har anpassat sig så specifikt att den *inte* kan generalisera. Men generaliseringsproblem kan också uppstå om träningsdatat är för smalt eller icke-representativt.

Ett exempel på generaliseringsproblem är om man tränar en AI att känna igen talade instruktioner på engelska, men alla röster i träningsdatat tillhör vuxna män. Om man sedan låter barn eller kvinnor använda systemet kanske det fungerar dåligt – modellen har inte generaliserat över olika talarröster, för den fick aldrig lära sig det. Problemet här var brist på *diversitet* i träningen. Generellt gäller att en modell inte kan generalisera bortom det "utrymme" som träningsdatan täcker in. Därför är både kvantitet och mångfald i datan viktiga (som diskuterats i avsnitt 2).

Inom AI-forskningen pratar man ofta om **bias-variance tradeoff** för att beskriva en balans: en modell med väldigt hög komplexitet kan överträna (låg *bias*, hög *variance*), medan en för enkel modell undertränar (hög bias, låg variance) och heller inte generaliserar [25]. Utmaningen är att hitta en "lagom" nivå där modellen lärt sig de huvudsakliga mönstren men inte oväsentliga detaljer [26]. Att förstå generalisering hjälper lärare att inse varför en AI inte är ofelbar: den gör antaganden baserat på tidigare erfarenheter och ibland håller inte de antagandena i nya fall.

Pedagogiskt kan detta bli ett samtalsämne i klassrummet om *överföring* av kunskap – något elever själva kämpar med. Precis som elever tränar på vissa uppgifter och förväntas kunna lösa även nya problem på provet (generalisation), tränar en AI på historiska data och förväntas hantera även framtida data. När både elever och AI misslyckas med att generalisera, behöver vi kanske justera hur de tränas: för elever kan det handla om djupare förståelse, för AI om mer eller bättre varierad data eller annorlunda modellarkitektur.

Bias i träning

Begreppet **bias** inom AI kan syfta på olika saker. Inom statistik och maskininlärning används det ibland i teknisk mening (som i bias-variance ovan), men här fokuserar vi på bias i betydelsen **snedvridningar eller partiskhet** i AI-modellens beteende, ofta orsakade av obalans eller skevheter i träningsdatan. En AI med bias kan till exempel systematiskt prestera sämre för en viss grupp människor (t.ex. felklassificera röster från personer med viss dialekt) eller reproducera stereotypa mönster (t.ex. föreslå manliga kandidater till tekniska jobb oftare än kvinnliga om träningsdatan hade den snedfördelningen).

Bias uppstår i regel när träningsdatan inte är *neutral*. AI-system lär sig ju av data, så om datan speglar mänskliga fördomar eller skeva representationer, kommer AI:n att lära sig just det [27]. Som AI4K12-initiativet påpekar kan *"biases in the data used to train an AI system [leda] till att vissa grupper missgynnas jämfört med andra"* [27]. Ett välkänt exempel är ansiktsigenkänningssystem som tränats huvudsakligen på ljushyade ansikten – de har sedan visat sig ha betydligt högre felkvot för att känna igen mörkhyade personers ansikten. AI:n i sig är inte "medvetet fördomsfull", den speglar bara den statistik den fått som grund. Men resultatet blir ett *orättvist system* om det används i känsliga sammanhang (t.ex. polisens övervakning eller rekryteringsprocesser).

För utbildnings-AI kan bias ta sig uttryck i allt från språkliga modeller som ger exempeltexter med stereotypa könsroller, till rekommendationssystem som gynnar redan studiestarka elever mer än svagare (om datan bygger på tidigare prestationer). Därför framhålls i många riktlinjer, bland annat OECD:s och UNESCO:s, att en kritisk del av AI-litteracitet är att förstå och kunna identifiera bias [19] [7]. Lärare bör till exempel fråga: Vilka elevgrupper testades denna AI på? Finns det risk att AI:n missgynnar vissa elever (kanske de med annat modersmål, eller med lässvårigheter) eftersom träningsdatan inte tog hänsyn till dem?

Att adressera bias handlar både om teknik och etik. Tekniskt finns metoder för att mäta och reducera bias (t.ex. genom att *balansera datan* eller justera modeller). Etiskt handlar det om att ta ansvar: AI-system i skolan bör utformas och användas på ett sätt som är rättvist och inkluderande. För läraren innebär AI-litteracitet att inte bara lita blint på AI:n, utan också vara uppmärksam på möjliga

partiskheter. Exempelvis om en automatisk rättstavnings-AI konsekvent markerar vissa ord från en viss sociolekt som fel, kan det vara en bias som behöver korrigeras – antingen tekniskt eller genom att läraren medvetandegör elever om problematiken.

Sammanfattningsvis: Genom att förstå överträning, generaliseringsproblem och bias inser vi att AI långt ifrån är perfekt eller universellt. AI-system är kraftfulla men bräckliga verktyg, formade av sina data och designbeslut. En AI-litterat lärare känner till dessa begränsningar och arbetar utifrån dem: med ett kritiskt öga, med beredskap att anpassa eller komplettera AI-stöd när de brister, och med betoning på det fortsatt centrala – lärarens roll att guida, bedöma och stötta eleverna med mänskligt omdöme där AI når sina gränser [28].

Referenslista

1. **Digital Promise (2024).** *AI Literacy: A Framework to Understand, Evaluate, and Use Emerging Technology*. Digital Promise. Hämtad från DigitalPromise.org [1].
2. **Walter, Y. (2024).** *Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education*. *International Journal of Educational Technology in Higher Education, 21*(15). https://doi.org/10.1186/s41239-024-00448-3 [2].
3. **Europeiska kommissionen & OECD (2025).** *AI literacy framework for primary and secondary education (Draft)* – initiativ för AI-litteracitetsramverk. European Education Area, lanseringsevent 22 maj 2025 [3] [4].
4. **UNESCO (2024).** *Artificial Intelligence (AI) Competency Framework for Teachers*. (F. Miao & M. Cukurova, red.). Paris: UNESCO. [7].
5. **AI4K12 (2019).** *Five Big Ideas in Artificial Intelligence* (D. Touretzky, C. Gardner-McCune, m.fl.). AI4K12 Initiative (AAAI & CSTA). [29] [30].
6. **GeeksforGeeks (2023).** *Rule Based System Vs Machine Learning System*. GeeksforGeeks, 14 dec 2023. [11].
7. **IBM (2021).** *What is overfitting?* IBM Cloud Education, 15 okt 2021. [31] [22].
8. **Microsoft Learn (2025).** *Interpret and improve accuracy and confidence scores*. Microsoft Docs (Azure AI Services), 3 mars 2025. [21] [20].
9. **Sperling, E., m.fl. (2024).** *In search of artificial intelligence (AI) literacy in teacher education: a scoping review*. (Preliminära resultat presenterade i Computers and Education, 2024) [8].
10. **OECD (2025).** *New AI Literacy Framework to Equip Youth in an Age of AI*. OECD Education and Skills Today, 2025. [4] [32].

---

[1] AI Literacy: A Framework to Understand, Evaluate, and Use Emerging Technology – Digital Promise
https://digitalpromise.org/2024/06/18/ai-literacy-a-framework-to-understand-evaluate-and-use-emerging-technology/

[2] Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education | International Journal of Educational Technology in Higher Education | Full Text
https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-024-00448-3

[3] Empowering learners for the age of AI: draft AI literacy framework launch - European Education Area
https://education.ec.europa.eu/event/empowering-learners-for-the-age-of-ai-draft-ai-literacy-framework-launch

[4] [17] [18] [19] [28] [32] New AI Literacy Framework to Equip Youth in an Age of AI – OECD Education and Skills Today
https://oecdedutoday.com/new-ai-literacy-framework-to-equip-youth-in-an-age-of-ai/

5 6 7 What you need to know about UNESCO's new AI competency frameworks for students and teachers | UNESCO
https://www.unesco.org/en/articles/what-you-need-know-about-unescos-new-ai-competency-frameworks-students-and-teachers

8 In search of artificial intelligence (AI) literacy in teacher education
https://www.sciencedirect.com/science/article/pii/S2666557324000107

9 10 Algoritm
https://www.bibb.se/ord/algoritm/

11 Rule Based System Vs Machine Learning System | GeeksforGeeks
https://www.geeksforgeeks.org/rule-based-system-vs-machine-learning-system/

12 15 16 27 29 30 AI4K12_Big Ideas in AI_Unpacked Poster_final_8_6_19
https://ai4k12.org/wp-content/uploads/2021/01/AI4K12_Five_Big_Ideas_Poster-1.pdf

13 What is deep learning? | SAP
https://www.sap.com/ukraine/resources/what-is-deep-learning

14 Neural network - Glider AI
https://glider.ai/glossary/neural-network/

20 21 Interpret and improve model accuracy and confidence scores - Azure AI services | Microsoft Learn
https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/concept/accuracy-confidence?view=doc-intel-4.0.0

22 23 24 25 26 31 What is Overfitting? | IBM
https://www.ibm.com/think/topics/overfitting

ChatGPT

Praktiska aspekter av AI-litteracitet i högre utbildning

AI utvecklas snabbt och påverkar sättet vi lär oss och arbetar på, vilket gör **AI-litteracitet** till en allt viktigare färdighet för studenter. Begreppet syftar på de kunskaper och förmågor som krävs för att förstå och använda AI-system på ett informerat, ansvarsfullt och kritiskt sätt. Enkelt uttryckt innebär AI-litteracitet en **förmåga att använda, hantera, värdera och förstå AI** [1]. EU betonar detta i sin nya AI-lagstiftning, där *AI literacy* definieras som de **”kunskaper, färdigheter och förståelse”** som möjliggör informerat användande av AI samt medvetenhet om dess möjligheter och risker [2]. Europeiska kommissionen arbetar också (tillsammans med bl.a. OECD) fram ett ramverk som beskriver vilka **essentiella kunskaper, färdigheter och attityder** unga behöver för att kunna interagera med AI **på ett tryggt och kritiskt sätt** [3]. Även globala aktörer som Unesco har introducerat modeller för AI-kompetens, där man ser att både studenter och lärare gradvis behöver *tillägna sig AI, fördjupa sin förståelse och slutligen kunna skapa med AI* [4].

Denna rapport syftar till att konkretisera AI-litteracitetens praktiska aspekter för universitetsstudenter. Nedan följer fem huvudområden: **Användarkompetens**, **Promptdesign**, **Verktygskännedom**, **Resultatvärdering** samt **Samverkan med AI**. Varje del förklarar vad området innebär, hur det kommer till uttryck i praktiken, och ger exempel på användning inom olika discipliner. Referenser till forskning, policy och etablerade resurser inkluderas för att ge en saklig och inspirerande bild av hur studenter kan dra nytta av AI på ett medvetet sätt.

Användarkompetens i AI

Användarkompetens syftar på förmågan hos användaren (t.ex. en student) att effektivt utnyttja AI-verktyg i sina studier. Det handlar om att veta *vad* AI-tjänster kan användas till, *hur* man interagerar med dem och *när* de är lämpliga att använda – liksom att förstå deras begränsningar. En kompetent AI-användare kan t.ex. öppna en språkmodell som ChatGPT och ställa frågor på ett sätt som ger relevanta svar, använda en översättnings-AI för att förstå litteratur på andra språk, eller ta hjälp av AI för att granska och förbättra sina texter.

Att praktiskt lära sig använda AI-tjänster kräver till en början *experimenterande och övning*. Många av verktygen är fritt tillgängliga online – exempelvis erbjuder OpenAI ett gränssnitt för ChatGPT, och både DeepL och Google Translate finns öppet för översättning. Genom att testa dessa i riktiga studieuppgifter lär man sig deras *funktioner och felkällor*. Det kan också vara värdefullt att ta del av guider eller workshops som vissa universitet nu erbjuder för att höja AI-kunnandet hos studenter och lärare [5]. Enligt Unesco bör både studenter och lärare kontinuerligt **utbilda sig och reflektera över sin AI-kompetens** [4] – en medvetenhet om var man befinner sig på kunskapsskalan hjälper i planeringen av vidare lärande.

I praktiken använder studenter redan idag AI-verktyg inom en rad olika ämnen. **Språkstudenter** kan ta hjälp av språkmodeller för att få förklaringar av grammatiska regler eller översätta texter. Till exempel har man sett att ChatGPT kan vara till stor nytta för att studera *klassiska språk* som latin och sanskrit – det kan hjälpa studenter med grammatikförståelse, översättning och till och med ge inspiration till egna textexempel [6]. **Teknik- och matematikstudenter** kan dra nytta av AI för att förstå komplexa

koncept; studier visar att verktyg som ChatGPT kan förklara matematiska begrepp och svara på vanliga frågor, vilket i sin tur kan förstärka studenternas förståelse och avlasta läraren som annars behöver besvara repetitiva frågor [7]. Inom **humaniora och samhällsvetenskap** kan AI fungera som diskussionspartner eller snabb informationssökare – exempelvis har engelsklärare noterat att AI kan ge **omedelbar återkoppling** på frågor, något som hjälper studenter framåt i realtid [8]. I alla dessa fall är dock den mänskliga studenten fortfarande i förarsätet: det krävs omdöme för att ställa rätt frågor och för att tolka och använda svaren på ett vettigt sätt. Användarkompetens i AI innebär således både teknisk färdighet *och* kritisk medvetenhet – att veta *hur* man använder verktyget och *varför*, men också *när* man inte blint ska lita på det.

Promptdesign och instruktion

Att kunna formulera effektiva **prompts** (uppmaningar/instruktioner) till AI är en kärnkompetens inom AI-litteracitet. Generativa AI-modeller som ChatGPT är *beroende av instruktionen* de får – kvaliteten på svaret hänger nära samman med hur tydlig och detaljerad din fråga är [9]. Promptdesign handlar om att strukturera sin fråga eller uppgift på ett sätt som **vägleder AI:n** att ge ett användbart svar, istället för ett ytligt eller irrelevant. En välstrukturerad prompt kan ge insiktsfulla och högkvalitativa svar, medan en vag eller otydlig prompt ofta resulterar i generiska, förvirrande eller felaktiga svar [9].

Hur skriver man då en bra prompt? Flera strategier har identifierats av experter:

- **Var tydlig med din målfråga**: En bra prompt ska klart definiera vad du vill ha ut. Undvik tvetydigheter. T.ex. är det bättre att skriva *"Ge en detaljerad beskrivning av tamhundars egenskaper, beteende och skötsel"* istället för en vag fråga som *"Berätta om hundar"* [10]. Den tydliga prompten specificerar exakt vilket innehåll och vilken omfattning som önskas, vilket gör det lättare för modellen att fokusera på rätt saker.
- **Ge kontext och bakgrund**: AI-modellen har ingen inbyggd förståelse för *varför* du frågar något. Därför lönar det sig att inkludera relevant bakgrundsinformation i prompten [11]. Om uppgiften t.ex. är att få AI:n att översätta eller analysera en text, nämn det sammanhang texten hör till (ämne, målgrupp, etc.). En prompt som *"Översätt följande dialog från en medicinsk journal från engelska till svenska..."* ger bättre resultat än bara *"Översätt den här texten"*.
- **Specificera roll eller format**: Genom att tala om *vilken roll* AI:n ska anta eller vilket format svaret ska ha kan man styra tonen och upplägget. Exempelvis kan man börja prompten med *"Som historiker, förklara betydelsen av..."* eller *"Föreställ dig att du är en lärare och ..."*. Detta hjälper modellen att anta rätt perspektiv och detaljnivå [12]. På samma sätt kan man ange önskad stil eller längd, t.ex. *"Svara i punktform"* eller *"Ge en förklaring på max 100 ord med enkla termer"*.
- **Ge exempel att efterlikna**: I mer komplexa fall kan det vara effektivt att ge modellen ett exempel på det format eller den struktur man vill ha. En metod är det s.k. *CREST-ramverket*, där en prompt innehåller **Context (kontext)**, **Role (roll)**, **Example (exempel)**, **Style (stil)** och **Task (själva uppgiften)** [13]. Till exempel kan man i prompten inkludera ett kort exempel på önskad output (“Här är ett exempel på hur jag brukar skriva...”) som modellen kan härma, innan man ställer den faktiska frågan. Detta reducerar risken för missförstånd av formatet.
- **Var konkret med önskat output**: Ju mer specifikt du beskriver vad du vill ha, desto större chans att AI:n träffar rätt. Om du vill ha en lista, säg uttryckligen *"Lista X stycken..."*. Om du behöver att svaret innehåller en viss typ av information, nämn det. Exempel: istället för *"Ge mig en sammanfattning av artikeln"*, kan en **bra prompt** vara *"Sammanfatta huvudfynden i artikeln i tre punktformuleringar och föreslå en åtgärd för varje fynd"*. En jämförelse kan göras med ett **dåligt exempel** där man bara skriver *"Skriv en sammanfattning av den här rapporten"* – en sådan generisk uppmaning ger troligen ett allmänt svar. Däremot visade det sig att en tydlig instruktion som *“Summarise this report in three bullet points, focusing on key findings and action*

items." gav ett mycket mer fokuserat och användbart resultat [14]. Principen är att minimera gissningsleken för AI:n; var specifik med *vad* du vill att den ska göra och *hur* svaret bör se ut.

Sammanfattningsvis handlar promptdesign om att kommunicera med AI på ett **strukturerat och precist språk**. Det kan ta några iterationer – en erfaren AI-användare förbättrar ofta sin prompt stegvis efter att ha sett de första svaren (s.k. prompt engineering-cykel [15]). Genom att ge rikligt med kontext, tydliga instruktioner och eventuellt exempel, kan studenter få fram mer relevanta och korrekta svar från AI. Detta är en färdighet som utvecklas med övning, men grundregeln är alltid: *tänk noga på vad du ber om*. AI:n gör nämligen oftast precis det du säger åt den – varken mer eller mindre – så det lönar sig att säga det på rätt sätt.

Verktygskännedom: vanliga AI-verktyg i utbildning

AI-litteracitet innebär också att känna till **vilka verktyg** som finns att tillgå och hur de kan användas pedagogiskt. Idag finns en uppsjö av AI-drivna tjänster som studenter kan möta i högre utbildning. Nedan listas några vanliga AI-verktyg och deras användningsområden, pedagogiska potential samt begränsningar:

- **ChatGPT (generativ språkmodell)** – En AI-chatbot som kan generera text som svar på användarens frågor eller kommandon. ChatGPT kan användas för att *besvara frågor*, ge *förklaringar* av koncept, hjälpa till att *summera* källtext eller till och med skapa utkast till uppsatser. I undervisning kan detta vara ett stöd för brainstorming, för att få alternativa förklaringar till svåra ämnen, eller som en virtuell diskussionspartner. Begränsningar finns dock: modellen har ett kunskapsomfång som i standardversionen sträcker sig till ca 2021, vilket gör att den **inte känner till de senaste händelserna** eller forskningen [16]. Den kan också ibland *"hallucinera"*, dvs. hitta på fakta eller referenser som låter rimliga men inte är korrekta [17]. Användaren måste därför vara medveten om att ChatGPTs svar kan innehålla felaktigheter och bör dubbelkollas mot källor.
- **GitHub Copilot (kodassistent)** – Ett AI-verktyg som integreras i programmeringsmiljöer (t.ex. Visual Studio Code) för att ge *realtidsförslag på kod*. Studenten kan skriva en kommentar eller påbörja en kodrad, och Copilot föreslår fortsättningen baserat på träningsdata från enorma mängder öppen källkod. Detta kan dramatiskt snabba upp **programmeringsarbete** och hjälpa ovana programmerare att komma förbi kodningsstopp genom att ge exempel på lösningar. Pedagogiskt kan Copilot fungera som en *"parprogrammerare"* som ger förslag och hjälper studenter att se olika sätt att implementera en funktion. Men verktyget har tydliga begränsningar – det förstår inte koden på något mänskligt plan och kan missa viktiga kontexter. Det händer att Copilot genererar kod som **inte uppfyller uppgiftens krav eller innehåller buggar**, vilket kräver manuell korrigering [18]. Ännu viktigare är att Copilot kan föreslå **osäker eller ineffektiv kod**; det har rapporterats att AI:n ibland skriver in sårbarheter eller rentav plagierar kod från träningsdatan [19]. Därför måste studenter vara vaksamma och granska den kod som genereras – Copilot är ett hjälpmedel, men det ersätter inte behovet av programmeringskunskap och kritiskt tänkande kring kodens kvalitet.
- **AI-översättare (DeepL, Google Translate)** – Maskinöversättningstjänster som med hjälp av neurala nätverk översätter text från ett språk till ett annat. Dessa används flitigt av studenter i språkämnen, men också av t.ex. ingenjörer eller historiker som behöver ta del av litteratur på främmande språk. Verktyg som **DeepL** och **Google Translate** kan spara mycket tid genom att ge en grovöversättning av en artikel eller hjälpa en student att formulera en text på engelska. Pedagogiskt kan de stötta *läsförståelsen* – en student kan översätta en svår text och sedan analysera innehållet på sitt modersmål. Kvaliteten på översättningarna har förbättrats avsevärt med modern AI; DeepL anses ofta leverera **mycket träffsäkra och kontextmedvetna översättningar** och rankas som ett av de mest pålitliga verktygen för komplexa språkstrukturer

[20]. Samtidigt är det viktigt att inse att maskinöversättning inte är felfri. Finstilta nyanser, idiom eller kulturella referenser kan översättas felaktigt. Studenter bör därför använda dessa verktyg som *stöd*, men ändå jämföra med originaltexten och använda sitt eget språksinne för att justera kritiska formuleringar.

- **Grammarly (AI-baserat skrivstöd)** – Ett verktyg som använder AI för att analysera texter och påpeka språkliga förbättringsmöjligheter. Grammarly fungerar som en avancerad språkgranskare: det korrigerar stavfel och grammatik, men går längre än så. *Utöver* ren felsökning erbjuder verktyget **förslag på ordförbättringar och stiljusteringar** för att höja textens kvalitet [21]. Till exempel kan det föreslå en mer passande synonym, uppmärksamma om en mening är krångligt konstruerad, eller föreslå att en lång mening delas upp för tydlighetens skull. För studenter kan detta vara ett utmärkt stöd för att utveckla sitt akademiska skrivande – det är som att ha en personlig språkhandledare som ger kontinuerlig feedback. Det man bör komma ihåg är att Grammarly *inte* förstår textens innehåll eller sanningshalt. Det kan ge råd om hur något formuleras, men inte om *vad* som sägs är korrekt eller relevant. Därför bör man ta verktygets förslag som hjälp för form, samtidigt som man själv ansvarar för innehållet och att argumentationen hänger ihop.

Pedagogisk potential och begränsningar: De ovannämnda verktygen illustrerar hur AI kan integreras i olika delar av studierna – från skrivande och språk till kodning. Potentialen ligger främst i att AI kan ge **snabb, individuell feedback och avlastning**, vilket både kan höja effektiviteten och frigöra tid för djupare lärande. Exempelvis kan en student få omedelbara grammatiska råd från Grammarly istället för att vänta på lärarens respons, eller använda ChatGPT som *första hjälpen* när hen kör fast på en uppgift, för att sedan gå vidare själv. Studier indikerar att sådana AI-verktyg **kan spela en positiv roll för lärande** när de används rätt – de kan väcka intresse, aktivera studenter och öka engagemanget i kursinnehållet [22]. Samtidigt måste begränsningarna ständigt hållas i åtanke. Varje verktyg – oavsett hur imponerande – kan ge fel svar eller vara olämpligt att använda i vissa situationer. Därför hänger verktygskännedom ihop med de övriga aspekterna av AI-litteracitet: man behöver även träna sin förmåga att **kritiskt granska AI:ns resultat** och veta hur man samarbetar med AI på ett sätt som *främjar* det egna lärandet istället för att ersätta det.

Resultatvärdering – kritisk granskning av AI-genererat innehåll

En av de mest essentiella färdigheterna inom AI-litteracitet är **resultatvärdering**, d.v.s. att kunna granska och bedöma kvaliteten på innehållet som AI-system levererar. Eftersom AI-modeller genererar svar baserat på mönster i träningsdatan – utan mänsklig intuition eller källmedvetenhet – finns det en inneboende risk att svaren innehåller bias (skevheter), felaktigheter eller alltför ytliga resonemang. *Att kritiskt kunna värdera AI:s output är därför avgörande* för en ansvarsfull användning.

Varför är detta viktigt? Forskning och erfarenhet har visat att generativa AI-modeller som ChatGPT ofta **låter väldigt övertygande även när de har fel**. De kan formulera sig med självförtroende och formell korrekthet, vilket lätt ger ett sken av trovärdighet. Men under ytan kan svaren vara direkt missvisande. Till exempel har det noterats att ChatGPT ibland introducerar **biased innehåll** eller felaktiga påståenden i sina svar – modellen kan spegla skevheter som finns i träningsdatan, vilket i värsta fall kan leda till partiska eller orättvisa framställningar av känsliga ämnen [23]. Särskilt i utbildningssammanhang är detta problematiskt, eftersom vi strävar efter saklighet och balans. Vidare kan AI:n som nämnts *”hallucinera”* faktauppgifter. Den kan hitta på fakta, siffror eller referenser som inte existerar, en tendens som dokumenterats i flera sammanhang. I en undersökning inom medicinsk litteratur fann man att de referenser ChatGPT uppgav som belägg i själva verket ofta var **spuriösa – med felaktiga tidskriftsnamn och påhittade data** [17]. Ett illustrativt exempel kommer från en vetenskapsjournalist som bad ChatGPT skriva om vaccin *”på desinformationens vis”* – AI:n svarade då genom att påstå att en *studie i Journal of the American Medical Association* visat att covid-vaccinet knappt

har effekt, komplett med angivna (men falska) statistik [24]. Givetvis fanns ingen sådan studie i verkligheten; både referensen och siffrorna var fabricerade av AI:n. I ett bredare test av mediegranskare (NewsGuard) genererade ChatGPT dessutom falska eller vilseledande påståenden kring **80% av de fall** man provade, inom områden som covid-19, skjutningar och kriget i Ukraina [25]. Allt detta understryker hur **alarmerande lätt** AI kan producera felaktig information som låter plausibel.

Hur kan man identifiera bias eller fel? För det första krävs en allmän skepsis – ett *kritiskt förhållningssätt* där man aldrig tar ett AI-svar för absolut sanning utan vidare. Några praktiska metoder som studenter kan använda är:

- **Kontrollera källor och påståenden**: Om AI:n uppger en källa eller referens, försök att hitta den verkliga källan. Ofta visar det sig att referensen är felaktig eller inte säger det AI:n påstod. Likaså bör faktauppgifter (datum, namn, siffror) dubbelkollas mot pålitliga källor, precis som man skulle göra med information från Wikipedia eller okända websidor.
- **Analysera svarens struktur och djup**: Ett tecken på *ytligt resonemang* är att svaret är väldigt generiskt eller cirkulärt, utan konkreta detaljer, nyanser eller exempel. AI-modeller tenderar ibland att ge svepande sammanfattningar istället för djupanalys. Om svaret låter som ett läroboksstycke som man redan har läst tusen gånger, kan det vara AI:ns sätt att *låta* kunnig utan att egentligen ge ny insikt. Följdfrågor kan då ställas för att testa djupet: t.ex. “Kan du ge ett specifikt exempel?” eller “Varför då?”. Om AI:n börjar motsäga sig själv eller upprepa sig, är det en varningssignal.
- **Var uppmärksam på bias och vinkling**: Bias kan yttra sig i att svaren konsekvent favoriserar ett visst perspektiv, använder stereotyper eller exkluderar vissa aspekter av en fråga. Om man t.ex. märker att AI:ns svar på historiska frågor alltid fokuserar på västerländska perspektiv, eller att exempel i ett teknikämne alltid använder manliga pronomen, så kan det vara en indikation på inbyggda skevheter i datan. Här krävs medvetenhet från användaren – att man tänker på *vilken information modellen kan ha tränats på* och hur det kan färga svaren. Att jämföra svar från AI:n med kurslitteraturens förklaringar eller med olika formuleringar av frågan kan hjälpa till att upptäcka om AI:n lämnar ute något relevant.

Lyckligtvis börjar utbildningsväsendet aktivt adressera denna utmaning. Ett effektivt sätt att lära ut resultatvärdering är att *låta studenterna själva öva på att kritiskt granska AI-genererat innehåll*. I vissa kursmoment får studenter numera analysera och förbättra AI-svar som en del av uppgiften. Till exempel har en övning beskrivits där studenterna får en AI-genererad text inom sitt ämnesområde och ombeds **kommentera, granska och revidera texten** – de kritiserar felaktiga eller ytliga påståenden, lägger till evidens från kurslitteraturen och exemplifierar eller nyanserar där AI-texten brister [26]. Sådana uppgifter tvingar studenterna att använda sin ämneskunskap för att *överpröva* AI:ns svar, vilket både fördjupar lärandet i ämnet och tränar deras AI-kritiska blick. På så vis omvandlas AI från ett orakel man okritiskt konsulterar, till ett verktyg (eller rentav *case*) man lär sig att ifrågasätta och förbättra.

Sammanfattningsvis är budskapet att **AI-svar alltid måste tas med en nypa salt**. Att vara AI-litterat innebär att vara lite som en detektiv: ständigt på jakt efter belägg, medveten om att det som presenteras kan vara ett villospår. Denna skeptiska men konstruktiva inställning gör att studenter kan dra nytta av AI:n som hjälpmedel, *utan* att bli vilseledda av dess brister.

Samverkan med AI – människa och maskin i samarbete

Den kanske mest spännande dimensionen av AI-litteracitet är hur vi kan **samverka med AI** – alltså utnyttja kombinationen av mänsklig kreativitet/tänkande och maskinens hastighet/kapacitet för att nå bättre resultat än vad endera hade klarat själv. Inom utbildning öppnar detta upp för nya sätt att lära, där AI kan fungera som en **kognitiv partner** eller handledare snarare än bara ett verktyg. *Människa–*

maskin-interaktion i detta avseende handlar om att etablera ett samspel: studenten engagerar sig aktivt, AI:n stöttar med input, och tillsammans löser de en uppgift eller utforskar en problemställning.

Ett nyckelkoncept är att AI **ska förstärka, inte ersätta** det mänskliga tänkandet. Rätt använd kan AI fungera ungefär som en duktig handledare eller studiekamrat: den kan ge idéer, feedback och olika perspektiv, men det är studenten som behöver sålla, bedöma och sätta samman helheten. Till exempel kan en skrivande student använda AI för att brainstorma idéer till en uppsats – AI:n kanske genererar en lista med uppslag – varpå studenten väljer ut de mest intressanta och utvecklar dem vidare med sin egen analys. På liknande sätt kan en teknologstudent para ihop med en AI i en programmeringsuppgift: AI:n (t.ex. Copilot) föreslår kod, studenten granskar förslagen, testar dem och finjusterar. Resultatet blir ett slags *”parprogrammering” mellan människa och AI*, där AI:n sköter rutinbiten (som syntax och standardstrukturer) och människan ansvarar för kreativ problemlösning och kvalitetssäkring.

Pedagogiska tillämpningar av AI-samarbete börjar dyka upp inom olika discipliner. Inom språkstudier kan man tänka sig att studenter för dialoger med en AI-chatbot på det språk de lär sig – AI:n svarar tålmodigt och rättar fel, vilket ger övning i en *interaktiv miljö* som annars vore svår att få tag på utanför klassrummet. Inom samhällsvetenskap eller ekonomi kan AI användas för att simulera scenarier eller debattmotståndare: en student kan be AI:n argumentera för en viss ståndpunkt och sedan själv träna på att bemöta argumenten. I laboratorieämnen eller projektarbeten kan AI vara ett verktyg för snabb dataanalys, som sedan *tolkas* av studenten. Det gemensamma i dessa exempel är att AI blir en **partner** – inte för att göra jobbet åt studenten, utan för att ge ett *extra lager av stöd och insikter* i processen.

För att denna samverkan ska fungera väl betonas ofta behovet av **tydliga riktlinjer och roller**. Lärare och institutioner börjar utforma policyer för hur studenter får använda AI i kursarbeten – just för att uppmuntra *meningsfullt samarbete* snarare än fusk eller passiv användning. Det är sällan en svartvit fråga där AI antingen är helt förbjudet eller fritt fram; snarare finns det en **gråskala** där studenter kan tillåtas använda AI-verktyg i *vissa delar* av en uppgift, medan de i andra delar förväntas prestera själva utan AI-stöd [27]. Till exempel kanske en lärare tillåter att man använder språkverktyg som Grammarly för att putsa språket i en inlämning, men inte att man genererar hela essäer med ChatGPT – eller så kan en programmeringslärare uppmuntra användning av Copilot för att få förslag *som sedan diskuteras på seminariet*, men kräver att tentakod skrivs utan AI. Genom att vara tydlig *när, hur och varför* AI får användas i en uppgift, kan man undvika missbruk och istället fokusera på de pedagogiska vinsterna [28] [29]. Det handlar om att designa läraktiviteter där AI *integreras* som ett led i lärandet, inte som en genväg för att slippa lärandet.

Ett konkret scenario för samverkan är också att låta AI ta över vissa *administrativa eller repetitiva moment* så att människor kan ägna sig åt mer kvalificerade uppgifter. Inom utbildning har det t.ex. visat sig att om ChatGPT besvarar vanliga faktafrågor eller ger grundläggande förklaringar i ett ämne, så **frigörs tid för läraren** att fokusera på djupare diskussioner, analys och anpassad hjälp [7]. På motsvarande sätt kan en student använda AI som förstahandsstöd för enklare problem – få en hint, ett exempel eller en återkoppling – och sedan lägga sin egen tid på de mer komplexa aspekterna av uppgiften. AI kan därmed fungera som ett *kognitivt avlastningsstöd*: det hanterar det mekaniska eller encyklopediska, medan den mänskliga studenten får utrymme att tänka kritiskt, kreativt och reflekterande. Forskare noterar att om detta görs rätt kan **varje students lärandeindividualiseras** – AI kan anpassa förklaringar eller övningar efter den enskildes behov, vilket gör att alla kan utmanas på sin nivå [30].

Det ska dock nämnas att en för stark tillit på AI-stödet **utan reflexion** kan få bakslag. Om studenten börjar luta sig tillbaka och låter AI:n göra det kognitiva tunglyftet riskerar hen att missa viktig övning i eget kritiskt tänkande. Vissa studier varnar för att okontrollerad användning av generativ AI kan *minska den egna förmågan till kritiskt tänkande och reflektion* [31]. Det är därför centralt att samarbete med AI

utformas på ett *aktivt* sätt: AI:n bör vara *verktyget* eller *assistenten*, och studenten *piloten* som styr och konstant utvärderar AI:ns input. En bra tumregel är att AI aldrig bör ha sista ordet i en lärprocess – det sista ordet (den slutliga analysen, tolkningen eller beslutet) bör alltid komma från en människa.

När människa och AI jobbar tillsammans på detta genomtänkta sätt, kan resultatet bli imponerande. Studier från klassrum har visat att AI-assisterade lärandeaktiviteter i många fall **ökar engagemanget och aktiveringen** hos studenter [22]. AI kan introducera ett lekfullt element (t.ex. genom att ge snabba svar, quiz eller nya infallsvinklar) som gör lärandet mer interaktivt. Det kan också fungera motivationshöjande – istället för att köra fast och ge upp, kan studenten ta hjälp av AI:n och sedan fortsätta på egen hand, vilket håller kvar flytet i lärandet. Slutligen öppnar samarbete med AI för *nya former av uppgifter*: såsom projekt där studenterna får bygga något tillsammans med en AI (t.ex. designa en prototyp med hjälp av en generativ AI) eller undersökningar där de jämför sina egna lösningar med AI:ns och reflekterar över skillnaderna. Sådana upplägg kan inspirera till djupinlärning och innovation.

Sammanfattning: Samverkan med AI handlar om att utnyttja styrkorna hos både människa och maskin. AI kan vara den outtröttlige assistenten som ständigt står redo med idéer, feedback och kunskap, medan studenten är den tänkande varelsen som styr, värderar och drar slutsatser. Tillsammans kan de åstadkomma mer än var för sig – men bara under förutsättning att studenten förblir aktiv och kritiskt tänkande. AI-litteracitet i detta avseende innebär att veta *hur man dansar med AI:n utan att låta den trampa en på tårna*. Med god verktygskännedom, skicklig promptdesign, kritisk resultatvärdering och tydliga roller kan AI bli en *samarbetspartner* som berikar högre utbildning och förbereder studenterna för en framtid där människa–AI-teamwork är vardag.

Källor:

- European Commission & OECD. *AI literacy framework* (utkast, 2025) – ramverk för kunskaper, färdigheter och attityder för att interagera kritiskt med AI [3].
- European Parliament. *Artificial Intelligence Act* (2024) – Definition av AI-litteracitet och krav på utbildning i AI-kunskap [2] [32].
- Unesco. *AI Competency Framework for Teachers* (Miao & Cukurova, 2024) – beskriver 15 AI-kompetenser för lärare, indelade i nivåerna *Acquire, Deepen, Create* [4] [33].
- Granberg, M. (2023). *AI-litteracitet i teknikundervisning*. ATENA Didaktik – studie som identifierar komponenter av AI-litteracitet (använda, hantera, värdera, förstå) inom teknikämnet [1].
- Kennedy, K. (2024). *The CREST Prompt Framework* – riktlinje för promptdesign med element som kontext, roll, exempel, stil och uppgift [13].
- Textcortex AI. *ChatGPT Prompt Engineering* (2023) – genomgång av strategier för tydliga och specifika prompts, inkl. exempel på förbättrade prompt-formuleringar [10] [12].
- Rumbelow, K. (2025). *Good Prompts vs. Bad Prompts with Copilot* – exempel på dåliga respektive bra AI-uppmaningar i praktiken [14].
- Arcada University of Applied Sciences (2024). *Bridging AI and Soft Skills* (projektrapport) – erfarenheter av att parallellt utbilda lärare och studenter i AI-litteracitet; betonar behovet av reflektion kring den egna AI-kompetensen [4] och att integrera AI med pedagogiska och etiska principer [28].
- Tigerstedt, C. & Biström, J. (2024). *Ökat lärande och nya färdigheter med hjälp av AI* – blogginlägg, Arcada. Diskuterar möjligheter/risker med GenAI i högre utbildning; bl.a. att överanvändning kan hämma kritiskt tänkande [31] och att tydliga riktlinjer behövs för AI-användning i uppgifter [29].
- Frontiers in Education, 2024. *ChatGPT in teaching and learning: a systematic review (SWOT analysis)* – samlar 51 studier. Fynd: ChatGPT kan stödja språkinlärning (latin m.m.) [6], ge matematikhjälp och avlasta lärare [7], ge omedelbar feedback i t.ex. engelskundervisning [8], samt öka

engagemang om rätt använd [22]. Samtidigt noteras begränsningar: svarslängd, felaktigheter, bias och hallucinerade referenser [23] [17].

- Sallam, M. m.fl., 2023. *ChatGPT output in medical education* – fann att ChatGPT gav missvisande medicinsk info och bör användas med försiktighet [34].
- Day, A., 2023. *Scientific American* – exempel på hur ChatGPT kan generera övertygande disinformation (påhittad vaccinstudie) [24].
- NewsGuard, 2023 – rapport om att ChatGPT kunde producera felaktiga påståenden för 80% av vanliga misinformationsteman [25].
- Pardo-Guerra, J.P. (2023). *Using AI Text as Prompts for Critical Analysis* – undervisningsexperiment där studenter fick kritisera AI-genererade texter, vilket stärkte deras ämneskunskap och förmåga att utvärdera AI-svar [26].
- Libbin, B. (2024). *Grammarly's Role in Student Achievement* – intervju, Univ. of Illinois Chicago. Noterar att AI-skrivstöd som Grammarly inte bara fångar fel utan också förbättrar ordförråd och struktur i studenters texter [21].
- Smartling (2023). *How accurate is DeepL?* – teknisk genomgång som framhåller DeepL:s neurala översättningsteknik och höga översättningsprecision i europeiska språk [20].
- Zelikovitz, S. (2024). *Learning to Code with GitHub Copilot* – guide för studenter, CUNY. Understryker vikten av att förstå Copilots begränsningar och använda det på ett säkert och effektivt sätt [35].
- Swimm (2023). *GitHub Copilot: Guide to Features and Limitations* – påtalar risk för osäker kod och behovet av försiktighet vid användning av AI-kodförslag [19].

---

1 AI-litteracitet i teknikundervisning | ATENA Didaktik
https://atenadidaktik.se/article/view/5450

2 32 AI Literacy in EU and the US - AI Literacy Institute
https://ailiteracy.institute/ai-literacy-act/

3 Empowering learners for the age of AI: draft AI literacy framework launch - European Education Area
https://education.ec.europa.eu/event/empowering-learners-for-the-age-of-ai-draft-ai-literacy-framework-launch

4 5 27 28 29 30 31 Ökat lärande och nya färdigheter med hjälp av AI utforskas inom utbildningen för företagsekonomi | Arcada
https://www.arcada.fi/sv/artikel/publikation/2024-12-19/okat-larande-och-nya-fardigheter-med-hjalp-av-ai-utforskas-inom-utbildningen-foretagsekonomi

6 7 8 17 22 23 34 Frontiers | The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach
https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1328769/full

9 14 Good Prompts vs. Bad Prompts with Copilot
https://www.changingsocial.com/blog/good-prompts-vs-bad-prompts-copilot/

10 11 12 15 ChatGPT Prompt Engineering Techniques
https://textcortex.com/post/chatgpt-prompt-engineering

13 The CREST Prompt Framework - AI Literacy Institute
https://ailiteracy.institute/crest-prompt-framework/

16 24 25 ChatGPT: Artificial Intelligence’s Impact on Education | ATPE
https://www.atpe.org/News-Media/Magazine/ATPE-News-Summer-2023/ChatGPT

18 19 GitHub Copilot: Complete Guide to Features, Limitations & Alternatives
https://swimm.io/learn/github-copilot/github-copilot-complete-guide-to-features-limitations-alternatives

20 How accurate is DeepL? Full review and best alternative tools
https://www.smartling.com/blog/how-accurate-is-deepl

21 Grammarly’s Role in Student Achievement with Bryan Libbin, Associate CIO for Academic Technology and Learning Innovation | Learning Technology Solutions | University of Illinois Chicago
https://learning.uic.edu/news-stories/grammarlys-role-in-student-achievement-with-bryan-libbin-associate-cio-for-academic-technology-and-learning-innovation/

26 Using AI Text as Prompts for Critical Analysis - The WAC Clearinghouse
https://wac.colostate.edu/repository/collections/textgened/rhetorical-engagements/using-ai-text-as-prompts-for-critical-analysis/

33 AI competency framework for teachers | UNESCO
https://www.unesco.org/en/articles/ai-competency-framework-teachers

35 "Learning to Code with GitHub Copilot: A Resource for New Student Devel" by Sarah Zelikovitz
https://academicworks.cuny.edu/si_oers/105/

ChatGPT

Etiska aspekter av AI-litteracitet i svensk högre utbildning

Inledning

Artificiell intelligens (AI) har på kort tid blivit en integrerad del av utbildningslandskapet, från digitala läromedel och adaptiva läroplattformar till generativa AI-verktyg som kan skapa text och bild. Denna utveckling väcker viktiga etiska frågor som blivande och verksamma lärare behöver förstå och förhålla sig till. Begreppet *AI-litteracitet* handlar om de kunskaper, färdigheter och attityder som krävs för att man ska kunna använda och förstå AI-system på ett informerat och kritiskt sätt [1]. EU:s kommande AI-litteracitetsramverk betonar att elever och lärare bör ha en *"mänskligt centrerad"* förståelse av AI – inklusive insikt i teknikens möjligheter, risker och påverkan på samhället [2] [3]. Med andra ord behöver man inte bara teknisk kompetens utan även etisk medvetenhet för att navigera AI på ett ansvarsfullt sätt.

Inom EU har arbetet med att främja AI-litteracitet tagit fart. Europaparlamentet röstade 2024 igenom AI-förordningen – världens första omfattande AI-lag – där *AI-litteracitet* definieras som de färdigheter, kunskaper och förståelse som gör det möjligt för användare och utvecklare att fatta informerade beslut om AI, med medvetenhet om dess möjligheter, risker och potentiella skada [1]. AI-förordningen kommer att ställa krav på att både leverantörer och användare av AI-system (t.ex. skolor) säkerställer att personal och elever har tillräcklig AI-kompetens [4]. Parallellt har EU-kommissionen i samarbete med OECD börjat utveckla ett pedagogiskt ramverk för AI-litteracitet i grund- och gymnasieskolan, planerat att färdigställas 2026 [5]. Även UNESCO har lanserat kompetensramverk för lärare och studenter om AI, där de lyfter fram att AI skiljer sig från tidigare digital teknik genom *unika etiska och sociala utmaningar såsom rättvisa, transparens, integritet och ansvarighet* [2]. Dessa initiativ understryker att etiska aspekter är en central del av AI-litteracitet – något som måste genomsyra utbildningen för att vi ska få **"tillförlitlig AI"** som gagnar elevernas lärande och skyddar deras rättigheter.

I Sverige har diskussionen om AI i skolan intensifierats under de senaste åren. Skolverket har publicerat råd och riktlinjer för användning av AI och chattbottar i undervisningen, och hösten 2024 infördes faktiskt *"artificiell intelligens"* som ett nytt ämne på gymnasiet och inom vuxenutbildningen [6] [7]. Syftet är att ge eleverna både teknisk förståelse och ett kritiskt förhållningssätt till AI. Framtidens skola förväntas enligt Skolverket kräva hög AI-litteracitet, där elever utvecklar **digital kompetens, kritiskt tänkande och förståelse för AI:s etiska konsekvenser** [8] [9]. Samtidigt visar en färsk lägesrapport från Skolverket att användningen av AI-verktyg ute i grundskolan ännu är relativt begränsad – en majoritet av lärarna har under 2024 **inte** använt AI i undervisningen eller låtit sina elever göra det [10]. Många lärare efterfrågar mer stöd och fortbildning i hur de kan använda AI i sin yrkesroll på ett genomtänkt sätt [10]. Det finns alltså ett kunskapsglapp att fylla. Cirka två av fem lärare uppger redan att de prövat AI-tjänster i arbetet, men osäkerheten är stor kring hur tekniken bör hanteras [11]. Den explosiva utvecklingen av generativa AI-modeller som ChatGPT – vilka kan skriva uppsatser av hög kvalitet – har väckt oro för fusk och en rad nya dilemman för skolan [12]. Skolsystemet ställs inför frågan: **Hur drar vi nytta av AI:s möjligheter utan att ge avkall på etik, integritet och likvärdighet?**

Denna rapport syftar till att ge en populärvetenskaplig men faktabaserad genomgång av etiska aspekter av AI-litteracitet i svensk högre utbildning, med fokus på lärarutbildning och skolans värld. För att strukturera diskussionen utgår vi från fem centrala områden som ofta lyfts i både forskning och

policy: **(1) Integritet och dataskydd, (2) Transparens och förklarbarhet, (3) Bias och rättvisa, (4) Ansvar och ansvarsfördelning samt (5) AI i bedömning och urval.** Inom varje område kopplar vi resonemangen till relevanta svenska lagar (såsom GDPR, Skollagen och Diskrimineringslagen) och policyer, EU:s regelverk (inklusive den nya AI-förordningen) samt konkreta exempel från svensk utbildningskontext. Ambitionen är att ge lärarstudenter och verksamma lärare en vägledning i hur man kan identifiera och hantera etiska frågor kring AI, på ett sätt som är begripligt och tillgängligt utan att förenkla bort komplexiteten. Som vi ska se handlar detta om allt från att skydda elevers personuppgifter och rättigheter, till att säkerställa att AI-system är rättvisa och begripliga, till att klargöra vem som bär ansvaret när något går fel. Genom ökad AI-litteracitet med etisk medvetenhet kan utbildningssektorn både **undvika riskerna** och **tillvarata möjligheterna** med AI på ett sätt som gagnar alla elever.

1. Integritet och dataskydd

En grundläggande etisk aspekt av AI i utbildning är skyddet av individers integritet och personuppgifter. Skolan hanterar stora mängder känslig information om elever – allt från personuppgifter och betyg till beteendedata – och när AI-system introduceras ökar ofta omfattningen av datainsamling och spårning. Europeiska unionens dataskyddsförordning GDPR (General Data Protection Regulation) gäller fullt ut inom utbildningssektorn och ställer upp strikta krav på hur personuppgifter får behandlas [13] [14]. I svensk rätt kompletteras GDPR av Skollagen kap. 26a, som innehåller särskilda bestämmelser för personuppgiftsbehandling i skolan. Bland annat slår lagen fast att om det skulle uppstå konflikt mellan GDPR och annan lag, så har GDPR företräde [15]. Med andra ord: elevens dataskydd är *inget skolan kan kompromissa med* – regelverket är bindande.

Personuppgiftsbehandling i AI-baserade system måste ske lagligt, korrekt och transparent. GDPR vilar på principer som uppgiftsminimering (att inte samla in mer data än nödvändigt), ändamålsbegränsning (att bara använda data för specificerade syften) och lagringsminimering (att inte behålla data längre än nödvändigt) [16] [17]. Dessa principer ställs på sin spets när skolor börjar använda AI. Ett AI-drivet läroverktyg eller administrationssystem kan frestas att samla in omfattande data om varje elevs aktivitet – hur länge de läser en text, vilka fel de gör på övningar, vilka sajter de besöker, kanske till och med biometrisk data som ansiktsbilder eller röstinspelningar. Skolan måste dock noga fråga sig: **är all denna data nödvändig och tillåten att samla in?** Datainspektionen (numera Integritetsskyddsmyndigheten, IMY) har varit tydlig med att skolor inte får behandla känsliga personuppgifter hur som helst, ens med elevers samtycke, om det strider mot grundläggande dataskyddskrav [18] [19]. Ett uppmärksammat fall är när Anderstorpsgymnasiet i Skellefteå 2018 testade ett system för ansiktsigenkänning för att automatiskt registrera elevers närvaro. Trots att skolan inhämtat samtycke från de berörda eleverna ansåg Datainspektionen att pilotprojektet innebar ett *otillåtet intrång i elevernas integritet* och bröt mot flera GDPR-bestämmelser – bland annat därför att ansiktsdata (biometri) är en känslig personuppgift som normalt inte får behandlas i skolan [18] [19]. Myndigheten utfärdade en sanktionsavgift på 200 000 kronor, vilket var första gången en svensk skola bötfälldes under GDPR [18] [20]. Fallet tydliggjorde att även lovande AI-teknik (här för effektivare närvarokontroll) måste underordnas elevers rätt till privatliv. I sitt beslut påpekade Datainspektionens generaldirektör att det finns *"stort behov av tydlighet kring vad som gäller"* när ny teknik som ansiktsigenkänning gör entré i skolan [21]. Skolans välmenande försök kolliderade med grundregeln att elevers biometriska data **inte** får hanteras utan mycket starka skäl och rättslig grund.

Ett annat exempel på integritetsdilemmat är användningen av AI-baserade provbevakningssystem (exam *proctoring*). Under pandemin övergick många högre lärosäten till digitala tentor och tog hjälp av AI-verktyg som övervakar studenter via webbkameran för att upptäcka fusk. Dessa verktyg kan analysera ögonrörelser, ansiktsuttryck och ljud i omgivningen för att flagga misstänkt beteende [22]. *Ur ett integritetsperspektiv är detta problematiskt.* Dels innebär det en omfattande övervakning av studenten

i hemmiljö, dels behandlas känsliga data (videoupptagningar, biometriska mönster) som kan lagras hos externa företag. Frågan ställdes om sådan bevakning ens är tillåten enligt GDPR utan individens samtycke och adekvata skyddsåtgärder. I flera länder – inklusive Sverige – protesterade studenter mot upplevt intrång i privatlivet. Myndigheter påpekade att om en algoritm automatiskt fattar beslut som har juridisk eller betydande effekt på individen (t.ex. att underkänna en student för fusk baserat på AI-analys), så aktualiseras GDPR artikel 22 om rätten att *inte* bli föremål för enbart automatiserat beslutsfattande utan mänsklig inblandning, med vissa undantag. En underliggande princip här är att viktiga avgöranden (som disciplinära åtgärder) inte bör lämnas helt åt en *”svart låda”* – individen har rätt till insyn och möjlighet att få en människa att ompröva beslutet. I praktiken har därför många lärosäten valt att låta AI-systemen enbart assistera mänskliga tentavakter, inte ersätta dem: AI kan larma, men en ansvarig lärare måste granska och fatta det slutliga beslutet. Detta ligger i linje med Dataskyddsförordningens krav på **mänskligt ingripande** i automatiserade beslut som annars riskerar bli felaktiga eller orättvisa.

Spårning och datainsamling i AI-verktyg kan även utmana elevers integritet på mer subtila sätt i vardagen. Många moderna utbildningsplattformar och appar har inbyggd *learning analytics* som följer elevernas interaktioner för att ge anpassad återkoppling. Visserligen kan det gynna lärandet att läraren får data om vilka uppgifter eleven fastnar på eller hur lång tid den spenderar på en text. Men om sådan spårning sker kontinuerligt, behöver skolan informera om det och ha rättsligt stöd. Offentliga skolor brukar stödja sig på att behandling är nödvändig som led i myndighetsutövning eller uppgift av allmänt intresse (utbildningens genomförande). Då krävs ingen vårdnadshavarmedgivande, men kraven på proportionalitet och dataskydd kvarstår. *Integritetsskyddsmyndigheten (IMY) betonar i sin vägledning att skolor måste väga nyttan med AI mot intrånget i privatlivet* – och alltid välja den minst integritetskänsliga lösningen för ändamålet [23] [24]. Exempelvis: Behöver en lästräningsapp verkligen lagra varje enskild felstavning eleven gör, eller räcker det med övergripande resultat? Kan loggar anonymiseras eller raderas efter kursens slut? Sådana frågor bör ställas redan när verktyget upphandlas.

En praktisk rekommendation är att skolhuvudmän upprättar tydliga riktlinjer för AI-användning ur integritetssynpunkt. Skolverket råder skolor att ta fram en **AI-policy** som klargör hur AI *får* och *inte får* användas och hur data ska skyddas [25] [26]. I riktlinjerna bör det framgå till exempel att lärare **inte** ska mata in känsliga elevuppgifter i öppna AI-tjänster på nätet utan samråd – många populära AI-tjänster skickar användardata till servrar utanför EU, och då riskerar man bryta mot GDPR om inte skyddsåtgärder finns [23] [27]. Skolverket påpekar att information som en AI-modell fått tillgång till kanske inte går att radera; därför måste man vara restriktiv med vilka personuppgifter eller elevalster man lämnar ut [23] [27]. Ett talande exempel är om en lärare låter en chatbot analysera en elevs uppsats genom att klistra in texten. Den texten kan då hamna på företagets servrar och potentiellt användas för att vidareutveckla AI-modellen. Utan tydligt godkännande kan detta vara oförenligt med GDPR, som kräver att behandlingen har laglig grund och inte sker i större omfattning än nödvändigt. Riktlinjer bör därför ange att *elevgenererat material inte matas in i externa AI-tjänster utan att dataskyddskrav är uppfyllda.* Huvudmannen (skolans ansvariga organisation) är personuppgiftsansvarig för de AI-verktyg som används i verksamheten och måste försäkra sig om att leverantörerna följer GDPR, t.ex. genom personuppgiftsbiträdesavtal och kontroll av var data lagras [26].

Integritetsaspekten sträcker sig även till elevernas upplevelse av *personlig sfär* i skolan. Om AI medför ökad övervakning – vare sig genom kameror, loggar eller analyser – kan det påverka det psykosociala klimatet. Elever har rätt till en skolmiljö där de kan känna sig trygga. Att ständigt vara betraktad av algoritmer kan skapa stress och minska tilliten. Därför är transparens (avhandlas mer i nästa avsnitt) gentemot elever och vårdnadshavare avgörande: man bör öppet informera om vilka data som samlas in, i vilket syfte, och hur integriteten skyddas. Sammanfattningsvis utgör integritet och dataskydd en kritisk del av AI-litteraciteten för lärare: man måste känna till lagar som GDPR, förstå riskerna med dataflöden i AI-system, och kunna vidta åtgärder som skyddar elevernas personliga information. Att

hitta rätt balans mellan datadrivna pedagogiska innovationer och respekt för individens privatliv är en av de kanske svåraste men viktigaste etiska avvägningarna i en AI-driven skolmiljö.

2. Transparens och förklarbarhet

För att AI ska kunna användas ansvarsfullt i utbildning krävs det att systemen är begripliga och genomskinliga för lärare, elever och vårdnadshavare. *Transparens* syftar här på öppenhet om hur ett AI-system fungerar och vilka beslut eller rekommendationer det genererar, medan *förklarbarhet* (explainability) handlar om förmågan att i efterhand förstå och förklara de enskilda beslut som ett AI-system tar. Båda dessa aspekter är centrala för att bygga förtroende: om en algoritm ska påverka elevens studiesituation, har berörda parter rätt att veta på vilka grunder och hur.

Vad innebär förklarbar AI och varför är det viktigt i utbildning? I kontrast till en mänsklig lärare, som kan resonera högt och motivera sina pedagogiska beslut, kan många AI-system uppfattas som *”svarta lådor”*. Särskilt maskininlärningsmodeller (t.ex. neurala nätverk) fattar beslut utifrån komplex statistik på ett sätt som inte är omedelbart intuitivt ens för utvecklarna. Om en AI-baserad applikation exempelvis rekommenderar en viss nivåplacering för en elev (”elev A bör läsa matematik på avancerad nivå nästa läsår”) eller flaggar en inlämningsuppgift som plagiat, behöver lärare och elever kunna lita på att det finns sakliga skäl i bakgrunden. En *förklarbar* AI skulle i dessa fall kunna ge en motivering – t.ex. ”elev A svarade rätt på 95% av diagnostiska uppgifterna och löste dem mycket snabbare än genomsnittet, vilket indikerar hög kompetens” eller ”den inskickade texten matchar till 87% en uppsats som finns online, upptäckt genom mönsterjämförelse”. Utan någon förklaring riskerar AI:ns output att mötas med misstro eller förvirring: *”Varför säger systemet detta? Stämmer det verkligen?”*

Inom utbildning är **begriplighet** extra viktigt av flera skäl. Dels involverar det minderåriga elever och deras föräldrar, som inte kan förväntas ha specialistkunskap om AI – de har rätt till en pedagogisk förklaring på klarspråk om hur ett AI-verktyg påverkar undervisningen. Dels finns juridiska krav på transparens: Offentlighetsprincipen och Förvaltningslagen i Sverige kräver att myndighetsbeslut motiveras. Om t.ex. ett antagningsbeslut eller betygsbeslut i praktiken påverkats av en algoritm, så måste beslutet ändå kunna motiveras för den det angår. Skollagen 3 kap. 17 § stadgar uttryckligen att *den som har satt ett betyg på begäran ska upplysa eleven om skälen för betyget* [28]. Det implicerar att om en algoritm bidragit i betygssättningen, kan inte skälen vara *”för att datorn sa så”* – en legitimerad lärare måste kunna ge en pedagogisk och kriteriebaserad förklaring som eleven förstår. Förklarbarhet handlar därför också om att *AI-systemens utlåtanden måste översättas till utbildningens språk.* Ett exempel: Om ett AI-verktyg analyserar elevtexter språkligt och bedömer språknivå, måste läraren ändå knyta återkopplingen till kursplanens kunskapskrav (t.ex. ”din text når nivå C vad gäller struktur och sammanhang, eftersom…”). AI:n kan assistera i att upptäcka mönster, men det är lärarens ansvar att presentera och förankra bedömningen begripligt.

EU:s etiska riktlinjer för *tillförlitlig AI* (framtagna av EU-kommissionens expertgrupp 2019) framhåller just transparens som en kärnprincip: *”AI-system bör vara spårbara och förklarliga. Beslut som fattas av AI, särskilt i känsliga domäner som utbildning, ska kunna granskas och förstås i efterhand”*. I den föreslagna AI-förordningen konkretiseras detta genom krav på dokumentation och informationsplikt för *högrisk-AI-system*. AI-system som används i skolan för exempelvis antagning eller betygssättning klassas som högrisk (se avsnitt 5) och måste enligt förordningen bl.a. ha loggning av hur de fungerar, och användarna (dvs. skolorna) ska få instruktioner som gör att de förstår systemets begränsningar [29] [30]. Dessutom inför förordningen generella transparensregler: om elever interagerar direkt med en AI (ex. en chattbot som agerar läxhjälp) så ska det **tydligt framgå att de talar med en maskin och inte en människa**. Att uppge AI-systemets identitet är viktigt för att inte vilseleda; eleven har rätt att veta om exempelvis en kundtjänst på skolans webbplats är AI-driven.

Transparens mot elever och vårdnadshavare kan handla om så enkla saker som att informera i början av terminen: *”I den här kursen kommer vi använda ett AI-baserat program som heter X för att ge er extra träning i matematik. Programmet samlar in data om vilka uppgifter ni löser och hur lång tid det tar, för att anpassa svårighetsgraden. Läraren kan se denna information. Här är hur era data lagras och skyddas…”*. Genom att ge en sådan förklaring i förväg uppnår man två saker: man **ökar tilliten** (föräldrarna vet vad som pågår och varför) och man **möjliggör informerat samtycke eller invändning**. Även om samtycke inte alltid är den rättsliga grunden i skolan (ofta lutar man istället på myndighetsutövning), så är det god etik att förankra ny teknik genom dialog. Ibland kan elevens vårdnadshavare ha möjlighet att avböja viss behandling (t.ex. om det rör extra känsliga data). Transparens skapar då förutsättningar för sådana val.

Förklarbarhet gentemot lärare är lika kritiskt. Lärare som ska använda AI-verktyg i sin praktik måste få utbildning i hur verktygen fungerar bakom kulisserna. Om en rektor köper in en AI-tjänst som exempelvis automatiskt rättar flervalsfrågor och ger förslag på vilka kunskapsområden eleven brister i, bör lärarna få veta vilken algoritmisk logik som används. Annars riskerar man att lärarna antingen misstror systemet och låter bli att använda det, *eller* att de övertroget litar på systemet utan att märka dess fel. En scenario kan vara ett AI-system som analyserar elevers inlämnade texter för att ge betygsförslag. Om läraren förstår att systemet huvudsakligen kollar stavning, grammatik och textlängd, men inte djupare innehållsaspekter, så vet hen att AI:ns betygsförslag måste tas med en nypa salt och kompletteras med mänsklig bedömning av resonemangskvalitet. Saknas den insikten, finns risk att AI:n tillmäts för stor vikt eller används felaktigt. I värsta fall kan *”automation bias”* uppstå – alltså att användaren okritiskt följer AI:ns förslag trots att de kan vara missvisande. Att motverka sådan bias kräver att tekniken är just förklarbar och att användarna är tränade i att tolka dess output kritiskt.

Ett exempel på problem med bristande transparens inträffade i Storbritannien 2020, då man under pandemin ställde in slutprov och istället lät en algoritm beräkna studenternas betyg baserat på historiska resultat och lärarnas prognoser. Algoritmen var en sorts AI-modell (om än relativt enkel) som fungerade intransparent för eleverna: många fick sänkta betyg jämfört med vad deras lärare föreslagit, särskilt på skolor i socioekonomiskt svagare områden. Besluten framstod som obegripliga och orättvisa, och eftersom processen var automatiserad kunde ansvariga först inte tydligt förklara varje enskilt fall. Resultatet blev landsomfattande protester och att systemet fick skrotas till förmån för lärarbedömningar. Den lärdom som drogs var att *om algoritmer ska användas i så avgörande frågor som betyg, måste de vara både rättvisa (mer om det i nästa avsnitt) och transparenta*. Eleverna måste förstå varför de får ett visst betyg, och utan förklaringar urholkas förtroendet för systemet.

I svensk kontext är det osannolikt att vi skulle överlåta hela betygsprocessen åt en algoritm, just på grund av ovannämnda risker. Men även i mindre dramatiska användningsområden – som AI som stöd för att ge feedback – är transparens en faktor. Skolverkets riktlinjer nämner att AI-verktyg inte följer den svenska kursplanen eller värdegrunden, och att lärare alltid måste granska material som AI skapar [31] [32]. Det är en uppmaning till *”pedagogisk transparens”*: läraren ska göra tydligt vilka kriterier och kvalitetskrav som gäller, oavsett vad AI:n genererat. Om ett AI-system föreslår innehåll eller upplägg som inte ligger i linje med skolans styrdokument, måste läraren kunna upptäcka det – vilket förutsätter att AI:ns arbetssätt inte är höljt i dunkel.

En utmaning är att kommersiella AI-produkter inte alltid är villiga att dela med sig av sina algoritmers innersta mekanismer, av affärs- och upphovsrättsskäl. Det kan leda till en konflikt mellan å ena sidan *företagens sekretess* (”trade secrets”) och å andra sidan *skolans krav på insyn*. Offentlighetsprincipen kan ge elever och föräldrar rätt att begära ut underlag för beslut, och då kan frågan uppstå: måste leverantören visa koden/modellen? Ett uppmärksammat fall internationellt var när en kommun använde en algoritm för att fördela elever till skolor, och föräldrar som misstänkte orättvisor begärde att få veta algoritmens logik. Efter juridisk strid tvingades man delvis röja algoritmen. I Sverige har vi än

så länge begränsad användning av sådana beslutsalgoritmer i skolan, men om det ökar kan liknande situationer uppstå. AI-förordningen kommer antagligen att kräva att viss information om hög-risk AI system görs tillgänglig för tillsynsmyndigheter (t.ex. IMY eller DO), även om inte allt publiceras för allmänheten.

Sammanfattningsvis är transparens och förklarbarhet avgörande av både **etiska** och **pedagogiska** skäl. Etiskt därför att individer har rätt att få veta på vilka grunder de bedöms eller sorteras; pedagogiskt därför att lärande bygger på förståelse och tillit. En *förklarbar AI* i skolan behöver inte avslöja varje kodrad, men den bör åtminstone ge begriplig information om vilka faktorer den tar hänsyn till och hur säker den är i sina förutsägelser. Lärare, som brobyggare mellan tekniken och eleverna, bör inkludera moment av förklaringar i undervisningen: till exempel diskutera med eleverna **hur** ett visst AI-verktyg kom fram till sitt svar och om de litar på det. Detta främjar i sin tur elevernas egen AI-litteracitet – de lär sig att inte betrakta AI som magi, utan något som kan (och bör) ifrågasättas och förstås. Som UNESCO framhåller behövs kritiskt tänkande och ett *”human-centred”* förhållningssätt: AI ska stötta mänskliga beslut och insikter, inte ersätta dem blint [33] [34] . Transparens och förklaring är nyckeln till att behålla människa-i-loop och att AI blir ett verktyg *för* utbildning – inte en obegriplig maktfaktor *över* utbildning.

3. Bias och rättvisa

En av de mest omtalade etiska riskerna med AI är förekomsten av *bias* – systematiska skevheter eller partiskheter – i algoritmernas beteende, vilket kan leda till orättvisa utfall. Inom utbildning är principen om **likvärdighet** grundläggande: Skollagen kräver att utbildningen ska vara likvärdig i hela landet och för alla elever oavsett bakgrund [35] . Diskrimineringslagen förbjuder dessutom diskriminering av elever på grund av t.ex. kön, etnicitet, religion, funktionsnedsättning eller andra skyddade grunder i all utbildningsverksamhet. Om skolor börjar använda AI-system som påverkar elevers möjligheter – vare sig det gäller antagning, betyg, särskilt stöd eller disciplinära åtgärder – så måste dessa system vara rättvisa och icke-diskriminerande. *Algoritmisk bias* innebär att ett AI-system behandlar eller påverkar vissa grupper systematiskt annorlunda än andra, på ett sätt som inte kan motiveras sakligt. Detta kan ske av flera skäl: inbyggda fördomar i träningsdatan, designbeslut av utvecklarna, eller rentav att modellen speglar ojämlikheter som finns i samhället och förstärker dem.

Algoritmisk bias och dess effekter på olika elevgrupper kan ta sig många uttryck. Tänk till exempel ett AI-verktyg som förutspår vilka elever som riskerar att inte klara kunskapskraven, så att skolan kan sätta in tidiga stödåtgärder. Intentionen är god – att fördela resurser efter behov. Men om träningsdatan till modellen hämtats från historiska elevresultat nationellt, finns risken att socioekonomiska faktorer eller migrationsbakgrund (indirekt via betygsmönster) slår igenom. Modellen kanske “lär sig” att elever från vissa bostadsområden oftare får låga betyg och börjar flagga dem som riskindivider tidigt, även om enskilda individer kanske hade klarat sig bra. Då uppstår en *självuppfyllande profetia*: eleven blir sedd som svagare än den är och behandlas annorlunda, vilket kan påverka självförtroende och förväntningar. Det är ett exempel på indirekt bias som kan drabba socioekonomiskt utsatta grupper eller minoriteter. På liknande sätt skulle ett AI-baserat antagningssystem till populära skolor potentiellt kunna missgynna elever från viss bakgrund om det utformats utan rättvisa i åtanke.

Konkreta fall av AI-bias har redan noterats inom offentlig sektor i Sverige. Diskrimineringsombudsmannen (DO) har t.ex. uppmärksammat hur Försäkringskassans AI-modell för att upptäcka felaktig vab (tillfällig föräldrapenning) *snedvridet pekade ut kvinnor samt utlandsfödda, låginkomsttagare och lågutbildade* som sannolika fuskmisstänkta, betydligt oftare än jämförbara män [36] . I verkligheten fanns ingen sådan överrepresentation av fusk hos kvinnor; det var en skevhet i datan eller modellen som gav ett diskriminerande utfall [37] . DO gick ut och uppmanade personer som

kände sig missgynnade att anmäla detta, och betonade att AI-bias kan göra diskriminering systematisk om den inte stoppas [38]. Detta exempel gäller visserligen försäkringsområdet, men mekanismerna är desamma för utbildning: om en AI "lär sig" existerande fördomar (t.ex. att pojkar presterar sämre i språk, eller att en viss etnisk minoritet skolkar mer – oavsett om det stämmer eller beror på rapporteringsbias), så kan den ge förslag som reproducerar ojämlikhet. En omtalad internationell händelse var när en AI-baserad rekryteringsmjukvara hos ett företag nedvärderade kvinnliga kandidater, eftersom träningsdatan byggde på tidigare anställningar där majoriteten var män. I skolan skulle motsvarigheten kunna vara ett studie- och yrkesvägledning-AI som rekommenderar tekniska utbildningar mest till pojkar för att historiskt har fler pojkar valt teknik – sådana mönster måste brytas, inte cementeras, om vi värnar likabehandling.

Risker för reproduktion av ojämlikhet med AI är reella om vi inte aktivt motverkar dem. AI-system har en tendens att förstärka den statistik de får sig till livs. Om det finns en underrepresentation av t.ex. en viss minoritetsgrupp bland tidigare högpresterande elever i en datamängd, kan en ojusterad algoritm ranka nya elever från den gruppen lägre, inte på grund av deras egna meriter utan för att de "liknar" tidigare elever som presterade sämre (kanske på grund av sämre stöd, inte förmåga). Det kan leda till en **ond cirkel** av diskriminering. Diskrimineringslagen är tydlig med att indirekt diskriminering (en regel eller kriterium som verkar neutral men särskilt missgynnar en skyddad grupp) är otillåten om den inte har ett berättigat syfte och medel som är nödvändiga och proportionerliga. Om ett AI-system i skolan systematiskt missgynnar exempelvis elever med funktionsnedsättning – kanske en lästräning-AI som inte tar höjd för dyslexi och ger dessa elever låga "språkbetyg" – då måste det åtgärdas, annars kan skolans användning av systemet vara diskriminerande.

Lyckligtvis finns tekniska och organisatoriska metoder för att motverka bias. *Tekniskt* kan utvecklarna träna AI-modeller på mer balanserade dataset, filtrera bort känsliga attribut (så att modellen t.ex. inte ser elevens kön/etnicitet om det inte är nödvändigt) och kontinuerligt testa modellen för bias. EU:s AI-förordning kommer att kräva att hög-risk AI för bl.a. utbildning uppfyller krav på *"data governance"* – att utbildningsdata är relevanta, representativa och inte innehåller felaktigheter som leder till diskriminering [39] [40]. *Organisatoriskt* kan skolor se till att AI-beslut aldrig sker helt isolerat från mänsklig bedömning. En vägledande princip bör vara *"dubbelgranskning"*: om en algoritm gjort en riskprognos eller antagningsrankning, låt en människa med kompetens granska de tveksamma fallen, och framför allt utreda om det finns systematiska skillnader mellan grupper. Det kan handla om att köra en analys: *Får flickor konsekvent högre eller lägre poäng än pojkar av detta verktyg? Stad versus landsbygd? Inrikes födda vs utrikes födda?* Om sådana mönster upptäcks måste man justera systemet eller dess användning.

Ett tänkbart scenario i framtiden är att högskolor använder AI för att förhandsgranska ansökningar, exempelvis att läsa personliga brev med en språkmodell för att sortera ut toppkandidater. Utan försiktighetsåtgärder kan detta missgynna den som inte behärskar finkodade uttryck eller "rätt" stil – vilket korrelerar med vissa bakgrunder. Ett etiskt utformat system skulle behöva ta hänsyn till mångfald i uttryckssätt och kanske normalisera bedömningen för att inte premiera priviligierade sociolekter. Här krävs verkligen att de som utvecklar AI:n samarbetar med pedagoger och experter på språklig mångfald och inkludering.

I Sverige pågår forskning och projekt som fokuserar på *"etik och AI i klassrummet"*. Exempelvis lyfter Linköpings universitets initiativ om AI-litteracitet i lärarutbildningen just behovet av att adressera likvärdighet: Hur säkerställer vi att AI inte underminerar likvärdig bedömning? [41]. Lärarstudenter får i sådana projekt diskutera fall där AI skulle kunna ge upphov till orättvisor och hur man kan möta det. Detta är viktigt, för trots all teknik kommer det i praktiken ofta att vara **läraren eller skolledaren** som är sista utposten mot orättvisa. AI kan ge förslag, men det är människor i skolan som måste ha kunskapen att upptäcka om något verkar galet och modet att ifrågasätta resultat. En medveten lärare

kanske märker: *”Varför flaggar plagieringsverktyget nästan alltid uppsatserna från elever med svenska som andraspråk men sällan från infödda svensktalande? Kan det vara något systematiskt fel?”* – och tar det vidare.

Diskrimineringsombudsmannen har framhållit att personer som anser sig diskriminerade av ett automatiserat system har all rätt att anmäla det, precis som vid annan diskriminering [42]. Med AI-förordningen kommer DO och andra tillsynsorgan dessutom få stärkta befogenheter att granska AI-system för att se om de uppfyller icke-diskrimineringskrav [43]. DO betonade 2024 i samband med Försäkringskasse-fallet att EU:s nya regler kommer ge dem möjlighet att kräva ut dokumentation och testresultat för högrisk-AI, just för att kolla efter den här typen av bias [44]. Det betyder att skolor som börjar använda AI för viktiga beslut också bör vara beredda på att *kunna visa* att de gjort sitt för att undvika diskriminering – exempelvis genom utvärderingsrapporter eller loggar som visar hur systemet behandlar olika grupper.

Bias gäller inte bara traditionella diskrimineringsgrunder utan även *andra orättvisor*. Till exempel kan AI i undervisning vara partisk med avseende på *innehåll*. Om en AI-tjänst används för att generera textexempel eller övningsuppgifter, kan den ha inbyggda kultur- eller könsstereotyper. Kanske beskriver den systematiskt läkare som ”han” och sjuksköterskor som ”hon”, eller ger historiska exempel som osynliggör vissa grupper. Sådana snedsteg kan skada skolans värdegrundsarbete om de inte korrigeras. Lärare behöver därför vara uppmärksamma på *innehållslig bias* och oönskade budskap, precis som Skolverket nämner: AI kan reproducera fördomar och sprida olämpligt innehåll, vilket kräver att läraren är vaksam [24]. Ett exempel kan vara om en bildgenererande AI som elever använder i bildlektionen tenderar att sexualisera kvinnor eller framställa alla företagsledare som vita män – då måste läraren adressera det och diskutera varför det händer, snarare än att låta det passera. Detta kan i sig bli lärorika moment i kritiskt tänkande.

Sammanfattningsvis innebär *rättvisa och avsaknad av bias* att AI-system i skolan **ska behandla alla elever sakligt och likvärdigt**. De får inte förstärka existerande ojämlikheter eller skapa nya. För att uppnå detta krävs både teknikåtgärder (bra data, bias-testning, justeringsalgoritmer) och mänskliga åtgärder (kritisk granskning, mångfaldsmedvetenhet, regelverk som Diskrimineringslagen i ryggen). Målet är att AI ska bli ett verktyg som *minskar* subjektiva orättvisor – till exempel skulle AI kunna hjälpa läraren att se förbi sina egna omedvetna fördomar genom att tillhandahålla objektiva data – men det förutsätter att AI:n själv är designad rättvist. Lärarens roll som garant för likvärdighet kvarstår även i en AI-driven miljö: tekniken ändras, men det professionella uppdraget att behandla elever jämlikt är konstant. *AI-litteracitet* för lärare innefattar därför att förstå var bias kan smyga sig in och hur man motar bort den. Precis som en vetenskapligt litterat person kan ifrågasätta en dålig studie, måste en AI-litterat lärare kunna ifrågasätta en algoritms påståenden ur ett rättviseperspektiv: *”Visar den här modellen någon orimlig skevhet? Vad kan det bero på? Hur kan vi kompensera för det?”*

4. Ansvar och ansvarsfördelning

När AI-system börjar spela en roll i beslutsfattande eller pedagogisk praktik i skolan uppkommer oundvikligen frågan om **ansvar**: Vem bär ansvaret för besluten och konsekvenserna när en AI är inblandad? Är det den som utvecklat systemet, den som köpt in det, den som använder det eller kanske systemet självt (i någon futuristisk mening)? I praktiken kan ansvaret behöva delas upp i olika nivåer – juridiskt ansvar, professionellt ansvar och pedagogiskt ansvar – och alla dessa är relevanta för lärare och skolledare.

Juridiskt ansvar: Enligt gällande lagstiftning kan inte en AI-maskinvara eller algoritm vara juridiskt ansvarig subjekt – ansvaret faller på människor eller juridiska personer (organisationer) som tagit

beslutet att använda AI:n. I en skolkontext betyder det att om ett AI-verktyg orsakar en skada eller ett felaktigt beslut, så är det i slutändan skolhuvudmannen (kommunen eller den enskilda skolans styrelse) som bär ansvaret inför lagen. Exempel: Om ett automatiserat system felaktigt nekar en elev plats på en utbildning och det visar sig bero på ett systemfel, kan eleven överklaga beslutet och det är då myndigheten (t.ex. antagningsnämnden) som formellt ansvarar för beslutet – även om de litade på en AI-modell. På samma sätt: om en elev felaktigt anklagas för fusk av en AI-driven plagiatkontroll och blir avstängd, så är skolan ansvarig för avstängningsbeslutet, inte den mjukvara de använde. *Det går inte att gömma sig bakom "datorn sa så"* juridiskt sett. Skollagen påminner oss om rektors ansvar att beslut rörande enskilda elever inte kan delegeras ut hur som helst [45]; i förlängningen innebär det att ansvaret alltid landar hos en behörig beslutsperson.

Med EU:s AI-förordning skärps också det juridiska ansvaret kring AI. Förordningen skiljer på **leverantör** (den som utvecklar/säljer ett AI-system) och **användare/deployer** (den som integrerar och använder AI-systemet i sin verksamhet). I skolans fall är leverantören ofta ett EdTech-företag, och användaren är skolan. Båda har ansvar: leverantören måste uppfylla krav på exempelvis datasäkerhet, riskhantering och dokumentation för att få släppa produkten på marknaden. Användaren (skolan) måste se till att systemet används på ett sätt som uppfyller regelverket – t.ex. genom att göra konsekvensbedömningar, övervaka systemets faktiska prestanda och rapportera incidenter. Skolledaren har en skyldighet att inventera vilka AI-system som används och försäkra sig om att personalen är informerad om vad som gäller [46]. *"Skugg-AI"*, det vill säga appar eller system som smyger sig in utan att ha granskats (t.ex. en lärare som på eget bevåg börjar använda en AI-tjänst via sitt personliga konto), kan skapa problem [47]. Rektor behöver därför ha en policy där det tydligt framgår att nya digitala verktyg ska godkännas innan de används med elever [48]. På så vis har skolledningen kontroll och kan ta ansvar för att GDPR följs, att avtalen är på plats etc. Som Skolverket uttrycker: om skolan uppmanar elever att använda ett AI-verktyg så är huvudmannen ansvarig för att det sker i enlighet med lagar som GDPR – detsamma gäller AI som lärare använder i tjänsten [26].

En intressant aspekt är produktansvar och skadestånd: Om ett AI-system i skolan orsakar någon form av skada (t.ex. en beslutsskada – kanske en elev mister en möjlighet eller kränks, eller en fysisk skada om det handlar om en pedagogisk robot), kan frågan om skadeståndsansvar uppstå. I normalfallet kan en elev kräva skadestånd av huvudmannen (skolan) om skolan agerat vårdslöst. Skolan skulle då i sin tur kanske rikta krav mot leverantören om felet låg i programvaran. Sådana scenarier är ännu oprövade, men diskussioner pågår inom EU att komplettera AI-förordningen med tydligare regler för *AI-ansvar*, just för att undvika gråzoner. Klart är att **ansvaret inte försvinner bara för att AI är inblandat**; snarare krävs noggrannare ansvarsdelning.

Professionellt ansvar: Även om juridiken sätter ramen, har lärare och rektorer ett eget professionellt ansvar att agera etiskt och klokt. *Lärarens yrkesetik* – som inkluderar att sätta elevens väl i centrum, göra rättvisa bedömningar och upprätthålla förtroendet för professionen – gäller även vid användning av AI. Om en lärare märker att ett AI-verktyg ger missvisande råd, är det dennes professionella plikt att inte följa råden slaviskt. Till exempel: en lärare får av ett AI-baserat betygsstödsprogram en stark rekommendation att ge en elev betyget D. Läraren själv, utifrån alla sina underlag, bedömer eleven till C. Då måste läraren lita på sin professionella bedömning – och kanske dubbelkolla varför AI:n rankade eleven lägre. Kanske var det något data som AI:n inte hade (som läraren vet) eller så värderar AI:n vissa provresultat annorlunda. Oavsett vilket kan *inte* läraren abdikera sitt ansvar; enligt Skollagen är det endast den legitimerade läraren som har behörighet att sätta betyg [49], och AI kan inte ha en sådan legitimation. Professionellt ansvar innebär också att lärare vidareförmedlar en *riktig förståelse* till eleverna. Om elever frågar: "Varför fick jag det här resultatet i AI-drillövningen?", kan läraren behöva förklara eller åtminstone erkänna att "systemet gör en uppskattning baserat på X, men vi ska titta tillsammans på din process också." Det skulle underminera elevens förtroende om läraren bara svarade "Ingen aning, det är datorn som bestämmer". Yrkesansvaret kräver en viss *digital kompetens*: att läraren

själv lär sig hur verktygen fungerar (här knyter vi tillbaka till AI-litteracitet!). EU betonar i AI-förordningen att användare måste ha *”tillräcklig AI-kunskap”*, och skolledare måste sörja för kompetensutveckling [46] [50]. Sveriges Skolledarförbund har efterlyst nationella insatser för bred AI-kompetensutveckling för lärare och skolledare, så att de känner sig trygga och medvetna i sitt ansvar [51].

Pedagogiskt ansvar: Slutligen finns ansvaret mot elevens lärande och utveckling. En lärare har inte bara ansvar att *inte skada* eleven (t.ex. genom orättvisa beslut), utan också att aktivt främja lärande. Om nya AI-verktyg tas in i klassrummet måste läraren utvärdera deras pedagogiska värde. Är verktyget i linje med läroplanens mål? Bidrar det till djupare förståelse eller riskerar det göra eleven passiv? Ett exempel: En AI-baserad översättningstjänst kan snabbt översätta en engelsk text till svenska åt en elev som har svårt för engelska. Pedagogiskt ansvar innebär här att läraren överväger hur och när detta ska användas. Kanske är det motiverat som stöd för nyanlända i vissa ämnen – men i engelskaundervisningen skulle det undergräva målet (språkinlärning) om eleverna alltid fick använda översättnings-AI istället för att själva kämpa med texten. Ansvarstagande här innebär att utforma tydliga regler: *”I det här momentet får ni inte använda översättningsprogram, för vi övar läsning på engelska. Men i SO-ämnen kan ni använda dem för att förstå källor, dock med källkritisk blick.”* Likaså måste lärare ansvara för att AI används *inkluderande*. Om bara vissa elever får tillgång till en AI-hjälp (kanske för att deras föräldrar skaffat en betaltjänst) behöver skolan fundera på kompensation så att inte en digital klyfta skapas i klassrummet.

Ett annat pedagogiskt ansvar är att hantera **felaktiga beslut** av AI-system. Om en AI trots förberedelser ändå leder till att något går galet – säg att en algoritm felbedömer en elevs kunskapsnivå – då är det skolans ansvar att rätta till det och lära av det. Vem tar ansvar om en elev blir felaktigt anklagad av en AI? Svaret bör vara: *människorna bakom systemet.* En rektor kan behöva gå in och säga: ”Här hände ett misstag med vårt övervakningssystem, vi ber om ursäkt och backar från beslutet.” Att offentligt ge AI skulden riskerar annars att ingen tar ansvar, vilket urholkar förtroendet. Motsatsen – att en lärare får bära hundhuvudet för ett AI-fel denne inte kunnat påverka – är inte heller rättvis. Därför behöver skolor internt klargöra ansvarsfördelningen: att personalen rapporterar problem uppåt, och att ledningen tar ansvar utåt.

Ansvarsfördelning i praktiken kan formaliseras genom skolans AI-policy. En bra AI-policy klargör *roller*: Vem godkänner nya AI-verktyg? (Förslagsvis rektor eller IT-strateg.) Vem utbildar personalen i verktyget? (Kanske en IKT-pedagog.) Vem följer upp effekterna? Vem kontaktas om något går snett? Genom att skriva ner detta och göra alla införstådda, så undviks vakuum där alla pekar finger åt varann. Johanna Karlén från Swedish Edtech Industry tipsar skolledare att **”upprätta en AI-policy som tydliggör ansvarsfördelningen och se till att alla känner till den”** [46]. I policyn bör också ingå att man kontinuerligt *utvärderar* AI-användningen – ett ansvar som kan läggas på skolans kvalitetsarbete.

Det pedagogiska ansvaret inkluderar även att förbereda eleverna själva för en värld av AI. Lärarstudenter och lärare ansvarar inte bara för att själva agera etiskt, utan även för att förmedla etiska värderingar kring teknik till nästa generation. Detta ligger i linje med skolans demokrati- och värdegrundsuppdrag. Elever som lär sig om AI bör också lära sig om etiska konsekvenser, integritet och källkritik kopplat till AI (vilket Skolverkets riktlinjer också framhåller [31] [32]). Man kan säga att ansvaret här är *dubbelt*: lärare ansvarar för att AI används ansvarsfullt **för** eleverna, och för att fostra ansvarsfulla AI-användare **av** eleverna. Detta senare leder oss tillbaka till AI-litteracitet som mål – att eleverna själva blir medvetna om frågor som bias, transparens och integritet, så att de i framtiden kan axla sitt medborgaransvar i ett AI-samhälle.

För att konkretisera ansvarsfördelningen kan vi tänka oss ett scenario: En skola inför ett AI-system för att analysera elevernas svar på övningsfrågor i matematik och automatiskt ge förslag på vilka områden

eleven behöver träna mer på. **Rektors ansvar**: säkerställa att systemet är GDPR-kompatibelt, att det är pedagogiskt utvärderat, att lärarna får utbildning i det och att det finns rutiner om nåt krånglar. **Lärarens ansvar**: integrera systemet i undervisningen på ett sätt som gynnar lärandet, övervaka att det fungerar rimligt, förklara för eleverna hur de ska tolka feedbacken, och moderera/åtgärda om systemet skulle ge konstiga råd. **Leverantörens ansvar**: att systemet är tekniskt robust, att algoritmen inte diskriminerar eller gör direkt felaktiga beräkningar, samt att supporta skolan vid behov. **Elevens ansvar** (får ej glömmas): att använda verktyget enligt instruktion, vara ärlig (t.ex. inte försöka manipulera det), och meddela läraren om de upplever att något är fel eller orättvist. I en sådan miljö av klargjorda roller blir AI ett redskap, inte en gåta. Om något går fel kan man spåra *var* det brast – var det ett tekniskt fel (leverantörens skuld), ett tillämpningsfel (användaren/lärarens skuld) eller ett konceptuellt fel (ledningens val av verktyg)? Oftast bär alla ett stycke av ansvaret, men ytterst måste skolans ledning som beslutar om användningen bära huvudansvaret och skydda sina anställda och elever. Som företrädaren för Sveriges Skolledare uttryckte: det behövs riktlinjer som **”skyddar skolledare från juridiska risker”** och klargör frågor som var data lagras, vem som äger den och hur man skyddar elevers och personals uppgifter [52]. Det gäller att inte utsätta enskilda lärare för risken att råka bryta mot lagen ovetandes – ansvaret att ge dem rätt förutsättningar ligger på huvudmannen.

I korthet: Ansvar och ansvarsfördelning kring AI i skolan handlar om att se till att *någon* alltid har ögonen på bollen. Mänskliga och juridiska ansvarslinjer får inte suddas ut av “det gjorde datorn”. Genom medvetenhet hos varje aktör – utvecklare, beslutsfattare, lärare, elev – och genom tydliga policies, kan AI användas tryggare. Läraren ska kunna känna att “jag vågar använda det här verktyget, för jag vet att min rektor och leverantören stöttar mig, och jag vet var mitt ansvar tar vid och var deras ansvar ligger.” Och eleven ska veta att “om datorn gör fel mot mig, kommer mina lärare och skolan att backa mig och rätta till det.” Först då kan förtroendet för AI i utbildning växa sig starkt.

5. AI i bedömning och urval

Användningen av AI för **bedömning** (av elevers kunskaper) och **urval** (t.ex. antagning till utbildningar) utgör några av de mest känsliga tillämpningarna inom utbildningssektorn. Här står mycket på spel för individen – betyg kan påverka framtida möjligheter och antagningsbeslut avgör tillträde till utbildning. Det är därför inte förvånande att EU i sin AI-förordning explicit pekar ut just dessa användningsområden som *”högrisk”*. Enligt förordningen kommer AI-system som används för att bestämma en persons tillgång till utbildning, för att bedöma elevers prestationer eller för att övervaka dem vid prov att klassificeras som högrisk-AI [53]. Sådana system är tillåtna men måste uppfylla stränga krav på riskhantering, transparens, kvalitet och mänsklig översyn [54]. Anledningen är enkel: om AI gör fel här, kan det allvarligt inverka på en människas rättigheter och framtid.

Låt oss först titta på **AI i bedömning**. Bedömning innefattar allt från rättning av prov och inlämningar till betygssättning och feedback på övningar. Redan idag finns AI-lösningar som kan automatiskt rätta flervalsfrågor, kortsvarsfrågor och till och med uppsatser (s.k. *automated essay scoring*). Det senare – att låta en algoritm sätta ett betyg på en uppsats eller text – är kontroversiellt. Forskning har visat att vissa uppsatsbedömningsprogram korrelerar hyggligt med mänskliga bedömningar, men de kan luras av till exempel fluffigt språk utan innehåll eller missgynna ovanliga men kreativa svar. I Sverige är det idag inte tillåtet att låta en maskin ensam sätta betyg på nationella prov eller liknande; lärare måste alltid vara involverade. Men AI kan användas som *stöd*. Exempelvis utvecklas system för att ge förslag på poäng per fråga på digitala prov, som läraren sedan granskar och fastställer. En etisk nyckelfråga här är **insyn och rättvisa**, vilket vi redan berört: eleven har rätt att få veta skälen för betyg (Skollagen 3 kap 17§) [28], vilket innebär att läraren kan behöva korrigera eller åtminstone förstå AI:ns bedömningskriterier för att kunna motivera betyget. En annan nyckelfråga är **mänsklig kontroll**. Det anses allmänt oacceptabelt att betyg eller examina helt automatiseras – inte bara juridiskt, utan också för att helautomatiska bedömningar saknar kontextuell förståelse och empati. Människan i loopen

fungerar som en kvalitetsgarant. AI-förordningen väntas kräva att högrisk-AI inom utbildning är utformade så att *”lämplig mänsklig övervakning”* ingår, för att förebygga att AI-baserade bedömningar leder till skada [55] [56] .

Rättvisa och insyn i bedömning med AI kan illustreras genom det tidigare nämnda fallet i Storbritannien 2020 (Ofqual-algoritmen för A-levels). Där fanns bristande transparens och systematisk bias, vilket ledde till orättvisa utfall och att man fick skrota algoritmen. I Sverige under pandemin 2020 valde man istället att ställa in de nationella proven och låta ordinarie kursbetyg sättas av lärarna utifrån befintligt underlag, just för att undvika att hastigt införa någon algoritm. Det understryker att man var medveten om de risker som den brittiska vägen innebar.

Men AI i bedömning behöver inte enbart handla om stora prov och betyg. Det kan också vara vardagliga funktioner som att *föreslå uppgifter* anpassade efter elevens nivå (vilket i praktiken innebär att AI:n bedömt elevens färdigheter löpande). Sådana adaptiva system kan vara mycket värdefulla pedagogiskt – elever får utmaningar i lagom takt. Dock måste läraren här ta ansvar för att AI:n *bedömer rätt*. Om ett adaptivt matteträningsprogram plötsligt börjar ge en elev mycket enklare uppgifter kan det signalera att eleven “bedömts” svagare. Är det korrekt, eller missförstod systemet elevens misstag? Läraren bör hålla ett öga och justera vid behov, så att inte en elev t.ex. fastnar på en för låg nivå på grund av en spurious initial felbedömning.

AI i urval syftar främst på antagningsprocesser: vilka elever får en begränsad resurs, t.ex. plats på en skola eller utbildningsprogram? I Sverige har vi relativt formaliserade urvalsgrunder för gymnasiet (betyg från grundskolan) och högskolan (betyg och högskoleprov). Här finns inte mycket utrymme för AI, annat än att rent administrativt sortera ansökningar. Däremot kan AI tänkas användas vid *urval inom skolan*. Exempel: resurstilldelning – vilka elever ska erbjudas extra stödlektioner? Om en kommun utvecklar en AI-modell för att prioritera stödinsatser baserat på olika data (betyg, frånvaro, socioekonomiska faktorer), då har vi ett urvalssystem som måste granskas etiskt. **Är kriterierna rättvisa?** Som vi diskuterat under bias: kan det finnas inbyggd snedhet? Finns risk för stigmatisering av vissa elever? Ett annat urvalssammanhang är profilskolor eller spetsutbildningar som har fler sökande än platser – kanske i framtiden överväger någon att använda AI för att läsa ansökningsprov eller portföljer. Då gäller samma principer som vid bedömning: transparens, rättvisa och mänsklig inblandning.

AI kan också involveras i *löpnade urval/uppföljning*, t.ex. för att varna om en elev är på väg att inte uppnå kunskapskrav (vilka ska prioriteras för extra samtal?). Ett sådant system bör utformas så att det inte skapar en etikettering som följer eleven i onödan. Här kan Diskrimineringslagen och Skollagens likabehandlingskrav sätta gränser: man får inte sätta negativa etiketter på elever på lösa grunder eller på diskriminerande grunder. Om AI:n använder något kriterium som är problematiskt – säg bostadsområde som proxy för risk – skulle det kunna bryta mot förbudet mot indirekt diskriminering.

Mänsklig kontroll och slutligt ansvar i urvalssituationer är avgörande. Antagning till högre studier eller selektering till program är myndighetsbeslut som kan överklagas. Det innebär att bakom kulisserna måste en människa (eller nämnd) formellt fatta beslutet och kunna ompröva det. AI kan vara beslutsstöd, men *beslutet är skolans*. I praktiken: om en algoritm rankar sökande 1–100, bör antagningspersonal åtminstone granska gränsfall och se att inget uppenbart fel skett (t.ex. ett orimligt lågt rankat meritvärde pga tekniskt fel). Helst ska de också kunna motivera varför person 1 kom in och inte person 2, utifrån kriterier – då kan de inte enbart säga “för algoritmen gav poäng X”. De måste relatera till faktorer som betyg, testresultat eller andra mer begripliga kriterier.

AI för **provrättning och fuskdetektion** förtjänar också uppmärksamhet. Nationella prov i Sverige digitaliseras allt mer, och det finns diskussioner om att använda AI för att flagga ovanliga mönster som

kan indikera fusk (t.ex. att två elevers svar är mycket lika). Likaså har många universitet globalt börjat använda plagiatkontroller med AI och nu även *AI-detektorer* (för att se om en text är skriven av en människa eller genererad av AI). Dessa AI-detektorer är dock *osäkra* – forskning har visat att de ger många falska utslag. Skolverket avråder bestämt från att använda automatiska AI-kontroller som ensamt beslutsunderlag för betyg, just för att inget säkert sätt finns att avgöra om en text är AI-genererad [57] [58]. Här ser vi hur etiken (rättssäkerhet för elever) dikterar att *hellre fria än fälla* ska gälla: om man inte är säker, kan man inte bestraffa en elev. Mänsklig kvalitativ bedömning – t.ex. att diskutera med eleven om texten – är nödvändig. AI i bedömning ska vara *hjälpmedel, inte domare*.

EU:s AI-förordning kommer sannolikt att kräva att om AI används vid t.ex. antagning eller betyg, så måste den som använder systemet **”erbjuda motiveringar och försäkringar att systemet inte är diskriminerande”** [59] åt den registrerade (eleven/sökanden). Detta kan tolkas som att skolor måste dokumentera hur AI:n fattar beslut och kunna visa upp att de testat systemet för bias. Förordningen förbjuder också vissa specifika användningar: intressant nog nämns att AI som *övervakar lärares betygsättning för att se om de följer tidigare mönster* är förbjudet [60]. Troligen vill man inte ha AI som pressar lärare att sätta “statistiskt korrekta” betyg på bekostnad av individuella bedömningar – här skyddas lärarens autonomi vilket i förlängningen skyddar elevernas rätt att bedömas rättvist. Denna detalj visar hur lagstiftaren försöker förutse och förhindra vissa orättvisa scenario med AI.

Sammanfattningsvis, när AI används för bedömning och urval i skolan måste **rättvisa, insyn och mänsklig kontroll** hela tiden garanteras. Det handlar om elevernas rättssäkerhet. Varje elev ska känna att de blir bedömda som *individer*, utifrån tydliga kriterier, inte att de faller offer för en opersonlig algoritm. AI kan ge stora effektivitetsvinster – tänk att lärare slipper rätta 100 uppsatser och istället kan ägna mer tid åt pedagogisk analys – men dessa vinster får aldrig gå ut över principen att *ingen maskin får kränka en elevs rättigheter*. Därför bör AI-stödda betyg eller antagningar alltid ha **”human-in-the-loop”**, där en människa validerar eller åtminstone kan överpröva utfallet. Tekniken ska vara ett verktyg för att assistera läraren, inte en ersättning för lärarens omdöme. Och i de fall där AI faktiskt hanterar preliminära urval (som i adaptiva övningar eller automatiska diagnosticeringar), måste systemet vara så pass robust och fördomsfritt att inga elevgrupper missgynnas.

Ett positivt scenario är att AI i framtiden kan bidra till *mer* rättvisa bedömningar, exempelvis genom att standardisera rättningen av nationella prov så att alla elever bedöms mer likvärdigt oavsett vilken skola de går i. Men det förutsätter att algoritmerna tränats på ett representativt och kvalitetssäkrat underlag, och att de används under lärares överinseende. Slutligen ska vi komma ihåg att **mänsklig kontroll** inte bara är en regel – det är också en tröst. För elever och föräldrar är det viktigt att veta att bakom varje beslut finns ett ansvarstagande mänskligt omdöme. AI kan vara briljant på att räkna poäng, men empati, förståelse för individens unika kontext och förmågan att ge hopp och motivera – det är mänskliga egenskaper. Inom utbildning, kanske mer än någon annanstans, måste tekniken underordnas de mänskliga värdena. Som UNESCO påminner i sina rekommendationer: *AI ska stödja mänskligt beslutsfattande, inte ersätta det, och alltid respektera mänsklig värdighet och rättigheter* [61] [33].

Avslutning

AI-litteracitet i svensk högre utbildning – särskilt inom lärarutbildningen – måste innefatta en stark förståelse för de etiska aspekterna vi här behandlat. Integritet och dataskydd, transparens och förklarbarhet, bias och rättvisa, ansvarsfördelning samt AI i bedömning och urval är inte bara tekniska eller juridiska frågor, utan i grunden pedagogiska och etiska. De handlar om vilken skolmiljö vi vill ha i en tid av AI: *en miljö där tekniken används med omdöme för att stärka varje elevs möjligheter, utan att kränka deras rättigheter eller likvärdighet.*

För blivande lärare innebär detta att man behöver ett kritiskt förhållningssätt till AI. Man bör känna till lagar som GDPR, Skollagen, Diskrimineringslagen och kommande AI-förordning – inte för att bli jurist, men för att förstå ramarna för sitt handlingsutrymme och skyldigheter. Minst lika viktigt är att utveckla en etisk kompass: att alltid fråga sig hur ett visst AI-verktyg påverkar eleverna, om det behandlar dem rättvist och om det tjänar det pedagogiska syftet. I rapporten har vi sett konkreta exempel på risker: ett ansiktsigenkänningssystem som inkräktade på integriteten och stoppades, en AI-modell som misstänkliggjorde kvinnor som fuskare, chattbottar som kan leda till fuskproblematik, otransparenta algoritmer som skapar förtroendekriser, med mera. Men vi har också berört möjligheterna: AI kan ge individanpassat lärande, effektivisera administration, hitta elever som behöver hjälp tidigt och avlasta lärare från rutinuppgifter. Nyckeln till att möjligheterna ska överväga riskerna är att *människan hela tiden har kontroll och insikt*.

För verksamma lärare och skolledare innebär detta att kompetensutveckling inom AI är oundgänglig. EU-förordningen kommer ställa krav på att personal som använder AI har rätt kunskap 4 50. I praktiken behöver skolor satsa på fortbildning där lärare får lära sig om till exempel: hur tränas en AI-modell? vad är typiska bias-problem? hur skyddar vi data? vilka frågor bör jag ställa till en leverantör? Samtidigt måste lärarutbildningarna integrera dessa moment – och det börjar ske, som i Linköpings universitets projekt för AI-litteracitet i lärarutbildningen 41. På så vis kommer nya lärare ut i yrket med grundläggande AI-kunskap och etisk beredskap.

Policyutvecklingen fortsätter också. Skolverket har en pågående roll att ge stöd och riktlinjer; de betonar t.ex. kollegialt lärande om AI på skolorna och framhåller vikten av tydliga skolpolicys 25 26. Internationellt arbetar UNESCO och EU med ramverk som ska hjälpa utbildningssystem att integrera AI på ett ansvarsfullt sätt 62 33. UNESCO:s nya kompetensramverk för lärare tar med etiska överväganden som en central del 34 63, och uppmuntrar att lärare utvecklar ett *människocentrerat mindset* gentemot AI.

Till syvende och sist handlar etiken kring AI i skolan om att upprätthålla de värden som skolan alltid stått för, i en ny teknisk kontext. *Integritet* – att varje elev har rätt till en privat sfär och att behandlas med respekt. *Transparens* – att vi är öppna och ärliga mot eleverna och varandra om hur beslut fattas. *Rättvisa* – att ingen elev missgynnas på osakliga grunder och att alla får lika möjligheter. *Ansvar* – att vi som skolpersonal tar ansvar för våra beslut och skyddar eleverna, och att eleverna lär sig ta ansvar de med. *Mänsklig värdighet* – att vi aldrig låter tekniken behandla människor som medel eller siffror, utan alltid ser individen bakom datan.

AI är ett kraftfullt verktyg som, rätt använt, kan anpassa undervisningen mer än någonsin efter elevens behov och kanske avlasta lärare från en del administrativa bördor. Fel använt kan det underminera likvärdighet och förtroende. För att navigera detta landskap måste vi alla – beslutsfattare, rektorer, lärare, studenter – vara lite av *”AI-etiker”*. I den populärvetenskapliga anda denna rapport har, kan man likna det vid att lära sig köra en bil: man behöver kunna tekniken (köra bilen) men också trafikreglerna (lagar och etiska normer) och ha gott omdöme i trafiken (situationsmedvetenhet). AI i skolan är ett fordon vi precis har börjat köra. Genom AI-litteracitet med etisk kompass kan vi se till att resan med AI blir trygg och givande för alla elever – utan att någon lämnas bakom eller blir överkörd på vägen.

Fotnoter: Denna rapport har i löpande text hänvisat till ett flertal källor inom parentes 【】. Nedan följer en fullständig referenslista till dessa källor, numrerade i den ordning de först förekommer.

Referenslista

1. **AI Literacy definition i EU AI Act** – Kara Kennedy (2024). *“AI Literacy in EU and the US”*. AI Literacy Institute, 3 juni 2024. [Online]. EU AI Act definition av AI-litteracitet [1] [4].
2. **EU-kommissionen & OECD (2025)** – European Education Area (2025). *“Empowering learners for the age of AI: draft AI literacy framework launch”*. [Webb]. Om EU:s AI-litteracitetsramverk och dess innehåll [64] [5].
3. **UNESCO (2024)** – UNESCO (2024). *“What you need to know about UNESCO’s new AI competency frameworks for students and teachers”*. [Webbartikel]. Om AI:s unika etiska utmaningar och UNESCO:s human-centred principer [2] [3].
4. **Linköpings universitet (2023)** – Stenliden, L. & Sperling, K. (2024). Beskrivning i projektet *“AI-litteracitet och didaktik i lärarutbildningen”* (LiU). [Webb]. Om behovet av etiska avvägningar kopplat till likvärdighet och datahantering i lärarutbildning [41].
5. **Skolverket – GDPR i skolan** – Skolverket (u.å.). *“GDPR och skyddade personuppgifter i skolan”*. [Webb]. Om GDPR:s tillämpning i skola och förhållande till Skollagen [13] [14].
6. **Skolverket – Råd om AI (2023)** – Skolverket (2023). *“Råd om AI, chattbottar och liknande verktyg”*. [Webb]. Skolverkets rekommendationer om försiktighet med personuppgifter och olämpliga AI-budskap [23] [24], samt vikten av skolpolicy för AI [65].
7. **SVT Nyheter (2019a)** – Svensson, A. *“Böter för ansiktsigenkänning på Skellefteå-skola”*. **SVT Nyheter**, 21 aug 2019. [Nyhetsartikel]. Om Skellefteå-skolans pilot med AI för närvaro, GDPR-brott och böter [18] [19].
8. **SVT Nyheter (2019b)** – Svensson, A. *“Ansiktsigenkänning testades på skola – så gick det till”*. **SVT Nyheter**, 5 sep 2019. [Video/Artikel]. Kompletterande info om Skellefteå-fallet och intrång i integriteten [21].
9. **ALL Digital (2024)** – Norman Röhner, intervju i *“How does the new EU AI Act affect the adult education sector?”*. **ALL Digital**, 5 maj 2024. [Webb]. Om AI Act: utbildning som högrisk, krav på AI-litteracitet, ansvar för lärare och förbud mot viss AI (t.ex. emotionsigenkänning) [29] [50] [60].
10. **Arbetsvärlden/DO (2024)** – Feldbaum, M. *“DO: Anmäl om du diskriminerats av AI”*. **Arbetsvärlden**, 29 nov 2024. [Nyhet]. Om Försäkringskassans AI-bias mot kvinnor/utlandsfödda [36] [42] och DO:s syn på AI-diskriminering samt AI-förordningens befogenheter [44].
11. **AI Sustainability Center (2019)** – **Diskrimineringsombudsmannen** (2019). *“Automatiserad databehandling och risker för diskriminering”*. DO:s promemoria, 4 okt 2019. [PDF]. (Refereras ej direkt i text men relevant för svensk diskurs om AI och diskriminering.)
12. **AI och skolan – AIUC (2024)** – AI Sustainability Center via AI Kompetens/AIUC (2024). *“Skolverket och AI: Nya riktlinjer…”* [Bloggartikel]. Innehåller uppgifter om att 2 av 5 lärare använder AI, införandet av AI-ämne 2024 och citat om AI-litteracitetens betydelse [9] [12].
13. **Skolverket (2024)** – Skolverket (2024). *“Artificiell intelligens i undervisningen – lägesbild 2024”*. [Rapport/Webb]. Resultat av enkät om AI-användning: begränsad användning, behov av stöd [10].
14. **Skolledaren (2025)** – Skolledaren (facktidskrift) nr 3, 2025. *“AI-förordning ställer krav på skolledare – så hanterar du tekniken”*. [Webbartikel]. Intervju med Johanna Karlén m.fl. om AI-policy: tips om inventering, utbildning, tydlig ansvarsfördelning [46] samt “skugg-AI” problematik [47] och citat om behov av riktlinjer för datalagring/personuppgifter [52].
15. **Skollagen (2010:800)** – *Svensk författningssamling 2010:800*. [Lagtext online via riksdagen.se]. Centrala utdrag: Likvärdig utbildning (1 kap 9§) [35]; Rektors delegeringsförbud av elevärenden (2 kap 10§, ref i [33] ovan); Betygsmotivering (3 kap 17§) [28]; Behörighet och ansvar vid betyg (3 kap 16-17§, 26§).
16. **Diskrimineringslagen (2008:567)** – [Lagtext]. Förbud mot diskriminering i utbildning (2 kap 5§); definition indirekt diskriminering (1 kap 4§2). (Allmänt refererad i text).

17. **Dataskyddsförordningen (GDPR)** – Europaparlamentets och rådets förordning (EU) 2016/679. [EU-lagtext]. Centrala principer och artikel 22 om automatiserade beslut. (Allmänt refererad i text).
18. **EU AI Act (draft)** – Europaparlamentets och rådets förordning om harmoniserade regler för AI (utkast, 2024). [EU-lagtext under förhandling]. Relevanta delar: definition av AI-litteracitet [1] ; Artikel 4 om krav på AI-kunskap [4] ; Annex III som listar utbildning som högrisk [54] ; krav på transparens och human oversight (Art 13-14) [56] ; förbud mot emotion AI i utbildning [66] och mot AI för lärargradskontroll [60] . (Refererad via All Digital).
19. **Skolverkets riktlinjer om fusk** – Skolverket (2023). *"Att motverka fusk vid användning av AI"*. [Webb]. (Indirekt refererat: Skolverkets avrådan att använda inlämningsuppgifter utan kontroll pga AI-risk [67]).
20. **Övriga källor** – Diverse forskningsartiklar och nyhetsinslag som diskuterar AI i utbildning (t.ex. UK algoritmfallet, proctoringverktyg etc.) har använts som underlag för resonemang, även om de inte direkt citerats ovan. Dessa inkluderar bl.a. internationella fallstudier (Ofqual 2020) och vetenskapliga diskussioner om bias i automated essay scoring.

---

[1] [4] AI Literacy in EU and the US - AI Literacy Institute
https://ailiteracy.institute/ai-literacy-act/

[2] [3] [33] [34] [61] [62] [63] What you need to know about UNESCO's new AI competency frameworks for students and teachers | UNESCO
https://www.unesco.org/en/articles/what-you-need-know-about-unescos-new-ai-competency-frameworks-students-and-teachers?hub=83250

[5] [64] Empowering learners for the age of AI: draft AI literacy framework launch - European Education Area
https://education.ec.europa.eu/event/empowering-learners-for-the-age-of-ai-draft-ai-literacy-framework-launch

[6] [7] [8] [9] [11] [12] Skolverket och AI – Nya riktlinjer & AI i svenska skolor — Skräddarsydda AI-kurser & Utbildningar | AI-kurser i Stockholm & Göteborg | AIUC
https://www.aiuc.se/ai-insikter/skolverket-ai-riktlinjer-skolan

[10] Artificiell intelligens i undervisningen – grundskolan, förskoleklass och fritidshem - Skolverket
https://www.skolverket.se/publikationsserier/ovrigt-material/2024/artificiell-intelligens-i-undervisningen---grundskolan-forskoleklass-och-fritidshem

[13] [14] [15] [16] [17] GDPR och skyddade personuppgifter i förskola, skola och vuxenutbildning - Skolverket
https://www.skolverket.se/regler-och-ansvar/ansvar-i-skolfragor/gdpr-och-skyddade-personuppgifter-i-forskola-skola-och-vuxenutbildning

[18] [19] [20] [21] Böter för ansiktsigenkänning på Skellefteå-skola | SVT Nyheter
https://www.svt.se/nyheter/lokalt/vasterbotten/boter-for-ansiktsigenkanning-pa-skelleftea-skola

[22] [29] [30] [39] [40] [50] [53] [54] [55] [56] [59] [60] [66] How does the new EU AI Act affect the adult education sector? • ALL DIGITAL
https://all-digital.org/how-does-the-new-eu-ai-act-affect-the-adult-education-sector/

[23] [24] [25] [26] [27] [31] [32] [48] [57] [58] [65] [67] Råd om AI, Chattbottar och liknande verktyg - Skolverket
https://www.skolverket.se/skolutveckling/inspiration-och-stod-i-arbetet/stod-i-arbetet/rad-om-ai-chattbottar-och-liknande-verktyg

[28] [35] [45] Skollag (2010:800) | Sveriges riksdag
https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/skollag-2010800_sfs-2010-800/

36 37 38 42 43 44 DO: "Anmäl om du diskriminerats av AI"
https://www.arbetsvarlden.se/do-anmal-om-du-diskriminerats-av-ai/

41 AI-litteracitet och didaktik i lärarutbildningen - Linköpings universitet
https://liu.se/forskning/ai-litteracitet-och-didaktik-i-lararutbildningen

46 47 51 52 AI-förordning ställer krav på skolledare – så hanterar du tekniken | Skolledaren
https://www.skolledaren.se/aktuellt/nyheter/2025/3/ai-forordning-staller-krav-pa-skolledare--sa-hanterar-du-tekniken/

49 Vem får undervisa och sätta betyg? - Draftit
https://draftit.se/blogg/vem-far-undervisa-och-satta-betyg