Nick Bostrom

Superintelligens

Vägar, faror, strategier

FRI TANKE

Innehåll

Lysande utsikter

Maskiner som kan mäta sig med människor i generell intelligens – det vill säga maskiner som har sunt förnuft och effektiv förmåga att lära sig saker, resonera och planera för att lösa komplexa utmaningar i fråga om informationsbearbetning inom ett brett spektrum av naturliga och abstrakta områden – är något man har sett fram mot alltsedan uppfinningen av datorer på 1940-talet. På den tiden förlades uppkomsten av sådana maskiner ofta ett tjugotal år framåt i tiden.[7] Sedan dess har den förväntade ankomsttiden förskjutits framåt i en takt av ett år per år, så att framtidsforskare som idag intresserar sig för möjligheten av generell artificiell intelligens fortfarande ofta tänker sig att intelligenta maskiner ligger ett par decennier bort.[8]

Två decennier är en populär träffpunkt bland dem som förutspår radikala förändringar: tillräckligt nära för att väcka uppmärksamhet och vara relevant, men tillräckligt långt bort för att man ska kunna tänka sig att en rad genombrott, som för närvarande är endast vagt tänkbara, kan ha hunnit äga rum vid det laget. Jämför detta med kortare tidsramar: de flesta teknologier som kommer att ha stor inverkan på världen fem eller tio år framåt i tiden används redan i begränsad utsträckning, medan teknologier som kommer att omskapa världen inom mindre än femton år förmodligen finns som prototyper i laboratoriet. Tjugo år kan också tänkas ligga nära den tid som i det typiska fallet återstår av en prognosmakares karriär, vilket begränsar risken för att skada sitt rykte genom en alltför djärv förutsägelse.

Men att vissa individer har varit överoptimistiska när det gäller utvecklingen av artificiell intelligens i det förflutna innebär inte att AI är omöjligt eller aldrig kommer att utvecklas.[9] Den främsta anledningen till att framstegen varit långsammare än väntat är att de tekniska svårigheter som är förknippade med att konstruera intelligenta maskiner har visat sig större än pionjärerna förutsåg. Detta säger emellertid inget om exakt hur stora dessa svårigheter är eller hur långt vi nu är från övervinna dem. Ett problem som till en början ser hopplöst komplicerat

ut visar sig ibland ha en överraskande enkel lösning (även om det omvända troligen är vanligare).

I nästa kapitel ska vi titta på olika vägar som kan leda till maskinintelligens på mänsklig nivå. Men låt oss redan från början notera att hur många hållplatserna än må vara mellan nuläget och maskinintelligens på mänsklig nivå, så är det senare inte slutdestinationen. Nästa hållplats, inte mycket längre fram på spåret, är maskinintelligens på övermänsklig nivå. Tåget kanske inte stannar eller ens saktar ner vid Människobyns station. Det troliga är att det bara susar förbi.

Matematikern I. J. Good, som tjänstgjorde som chefsstatistiker i Alan Turings grupp av kodknäckare under andra världskriget, kan ha varit den förste som formulerade de väsentliga aspekterna av detta scenario. I en ofta citerad passage från 1965 skrev han:

> Låt oss definiera en ultraintelligent maskin som en maskin som vida kan överträffa alla intellektuella aktiviteter hos en människa hur smart hon än är. Eftersom konstruktion av maskiner är en av dessa intellektuella aktiviteter skulle en ultraintelligent maskin kunna konstruera ännu bättre maskiner; utan tvivel skulle en ”intelligensexplosion” då inträffa och människans intelligens skulle hamna långt efter. Den första ultraintelligenta maskinen är alltså den sista uppfinning som människan någonsin behöver göra, förutsatt att maskinen är foglig nog att tala om för oss hur man håller den under kontroll.[10]

Idag kan det tyckas självklart att en sådan intelligensexplosion skulle vara förenad med stora existentiella risker, och att möjligheten därför borde undersökas med största allvar även om vi visste (vilket vi inte gör) att det bara fanns en ganska liten sannolikhet för att den kommer att äga rum. Men pionjärerna inom artificiell intelligens, trots sin tro på en nära förestående AI på mänsklig nivå, övervägde för det mesta inte möjligheten av AI på övermänsklig nivå. Det är som om deras spekulationskraft uttömdes så fullständigt i tanken på den radikala möjligheten av maskiner som når mänsklig intelligens att de inte kunde greppa den

naturliga följden – att maskinerna därefter skulle bli superintelligenta.

AI-pionjärerna erkände vanligen inte möjligheten att deras projekt kunde innebära några risker.[11] De uttryckte aldrig farhågor rörande säkerheten eller några etiska betänkligheter i samband med skapandet av artificiella medvetanden och potentiella datoriserade diktatorer – och än mindre diskuterade de sådana frågor seriöst. Det är en förvånande försummelse, även mot bakgrund av den tidens föga imponerande normer för kritisk teknologibedömning.[12] Vi kan bara hoppas att när projektet till sist faktiskt blir genomförbart, så kommer vi att ha skaffat oss inte bara den teknologiska kompetensen för att utlösa en intelligensexplosion, utan också de färdigheter på högre nivå som kan krävas för att göra detonationen möjlig att överleva.

Men innan vi går in på vad framtiden rymmer, kan det vara bra att kasta en snabb blick på maskinintelligensens historia fram till idag.

Mellan hopp och förtvivlan

Sommaren 1956 samlades tio forskare med ett gemensamt intresse för neuronnätverk, automatateori och intelligensforskning för en sex veckor lång workshop vid Dartmouth College. Denna sammankomst, ”Dartmouth Summer Project”, ses ofta som startskottet för artificiell intelligens som forskningsfält. Många av deltagarna skulle senare vinna erkännande som områdets grundargestalter. Deltagarnas optimistiska perspektiv återspeglas i det förslag som lämnades in till Rockefeller Foundation, evenemangets finansiär:

> Vi föreslår att en studie av artificiell intelligens genomförs omfattande 2 månader och 10 personer … Studien ska baseras på hypotesen att varje aspekt av inlärning eller andra former av intelligens i princip kan beskrivas så exakt att en maskin kan fås att simulera den. Ett försök ska göras att upptäcka hur man konstruerar maskiner som använder språk, bildar abstraktioner och begrepp, löser typer av problem som idag är

> reserverade för människor samt förbättrar sig själva. Vi tror att betydande framsteg kan göras med ett eller flera av dessa problem om en noggrant utvald grupp av forskare arbetar gemensamt med det under en sommar.

Under de sex decennier som gått sedan denna dristiga början har forskningsfältet artificiell intelligens pendlat mellan hopp och förtvivlan: perioder av entusiasm och stora förväntningar har avlösts av perioder av bakslag och besvikelser.

Den första entusiastiska perioden, som började med mötet i Dartmouth, vägleddes enligt John McCarthy (evenemangets huvudarrangör) av mottot: ”Titta mamma, utan händer!” På detta tidiga stadium byggde forskarna system avsedda att vederlägga påståenden av typen ”Ingen maskin skulle någonsin kunna göra *X*!” Sådana skeptiska påståenden var vanliga vid denna tid. För att bemöta dem skapade AI-forskarna små system som uppnådde *X* i en ”mikrovärld” (ett väldefinierat, begränsat område som gjorde det möjligt att visa upp en avskalad version av prestationen i fråga), vilket gav ett ”proof of concept” och visade att *X* i princip kunde utföras av en maskin. Ett tidigt sådant system, Logic Theorist, lyckades bevisa de flesta teorem i det andra kapitlet av Whiteheads och Russells *Principia Mathematica*, och fick till och med fram ett bevis som var mycket elegantare än i originalet, vilket vederlade föreställningen att maskiner ”bara kan tänka numeriskt” och visade att de också kan utföra härledningar och uppfinna logiska bevis.[13] Ett uppföljningsprogram, General Problem Solver, kunde i princip lösa ett brett spektrum av formellt specificerade problem.[14] Program skrevs också som kunde lösa matematiska problem som är typiska för första årets universitetskurser, visuella analogiproblem av den typ som förekommer i vissa IQ-test och enkla, verbalt formulerade algebraiska problem.[15] Roboten Shakey (så kallad på grund av sin tendens att darra under drift) visade hur logiskt tänkande kan integreras med perception och användas för att planera och styra fysisk aktivitet.[16] ELIZA-programmet visade hur en dator kunde imitera en psykoterapeut av rogerianskt snitt.[17] Vid mitten av sjuttiotalet visade programmet SHRDLU hur en

simulerad robotarm i en simulerad värld av geometriska klossar kunde följa instruktioner och besvara frågor på engelska som skrevs in av en användare.[18] Under senare decennier skulle system komma att skapas som visade att maskiner kunde komponera musik i olika klassiska kompositörers stil, överträffa yngre läkare i vissa uppgifter inom klinisk diagnostik, köra bil självständigt och göra patenterbara uppfinningar.[19] Det har till och med funnits en AI som hittade på egna skämt.[20] (Inte för att humorn var på någon högre nivå: ”Vad får man om man korsar ett *vattendrag* med en *övertygelse*? En å-sikt”; men barn sägs ha funnit vitsarna ofelbart roliga.)

De metoder som var framgångsrika i de tidiga demonstrationssystemen visade sig ofta svåra att utvidga till ett bredare spektrum av problem eller till svårare problem. En orsak till detta är den ”kombinatoriska explosionen”, det vill säga mångfaldigandet av möjligheter som måste undersökas med metoder som vilar på mer eller mindre uttömmande sökning. Sådana metoder fungerar bra för enkla problemfall, men misslyckas när saker och ting blir lite mer komplicerade. Ett exempel: för att bevisa ett teorem som har ett fem rader långt bevis, i ett härledningssystem med en slutledningsregel och fem axiom, kan man helt enkelt räkna upp de 3125 möjliga kombinationerna och kontrollera var och en för att se om den ger den avsedda slutsatsen. Uttömmande sökning skulle också fungera för bevis på sex och sju rader. Men efterhand som uppgiften blir svårare får metoden med uttömmande sökning snart problem. Att bevisa ett teorem med ett 50 rader långt bevis tar inte tio gånger så lång tid som att bevisa ett teorem med ett 5 rader långt bevis; i stället kräver det, om man använder uttömmande sökning, att man kammar igenom $5^{50} \approx 8{,}9 \times 10^{34}$ möjliga sekvenser – vilket är en omöjlig beräkningsuppgift även med de snabbaste superdatorer.

För att övervinna den kombinatoriska explosionen behövs algoritmer som utnyttjar struktur i måldomänen och drar nytta av tidigare kunskap genom att använda heuristisk sökning, planering och flexibla abstrakta representationer – förmågor som var dåligt utvecklade i de tidiga AI-systemen. Dessa tidiga systems prestanda försvagades också

på grund av dåliga metoder för att hantera osäkerhet, användning av bräckliga och illa underbyggda symboliska representationer, databrist och allvarliga hårdvarubegränsningar i fråga om minneskapacitet och processorhastighet. Vid mitten av 1970-talet fanns det en växande medvetenhet om dessa problem. Insikten att många AI-projekt aldrig skulle kunna uppfylla sina ursprungliga löften ledde till den första ”AI-vintern”: en period av åtstramning, med minskad finansiering och ökad skepsis. AI var inte längre på modet.

En ny vår kom i början av 1980-talet, när Japan lanserade sitt ”projekt för femte generationens datorsystem”, ett välfinansierat partnerskap mellan offentliga och privata aktörer som ville åstadkomma ett teknologiskt språng genom att utveckla en massivt parallell beräkningsarkitektur som plattform för artificiell intelligens. Detta skedde när fascinationen inför Japan som ”efterkrigstidens ekonomiska mirakel” var som störst, en period när västerländska regeringar och företagsledare ängsligt försökte gissa sig till formeln bakom Japans ekonomiska framgångar i hopp om att kopiera magin på hemmaplan. När Japan beslutade att investera stort i AI följde flera andra länder efter.

De följande åren innebar utveckling av en mängd olika *expertsystem*. Expertsystem, utformade som stödverktyg för beslutsfattare, var regelbaserade program som drog enkla slutsatser utifrån en bas av faktakunskaper, vilka hade inhämtats från mänskliga experter på området och under stor möda handkodats i ett formellt språk. Hundratals sådana expertsystem konstruerades. Men de mindre systemen var inte till någon större nytta, och de större visade sig dyra att utveckla, kontrollera och hålla uppdaterade, samtidigt som de var allmänt omständliga att använda. Det var opraktiskt att skaffa en separat dator för att köra ett enda program. I slutet av 1980-talet hade även denna tillväxtsäsong nått sitt slut.

”Femte generationens projekt” lyckades inte uppfylla sina mål, liksom inte heller dess motsvarigheter i USA och Europa. En andra AI-vinter sänkte sig. Vid denna punkt kunde en kritiker med rätta beklaga sig över ”den hittillsvarande forskningen i artificiell intelligens, som hela

tiden har handlat om mycket blygsamma framgångar på vissa områden, omedelbart följt av ett misslyckande när det gällt att nå de mer övergripande mål som dessa inledande framgångar verkade antyda".[21] Privata investerare började undvika varje affärsprojekt med etiketten "artificiell intelligens". Till och med bland akademiker och deras finansiärer blev "AI" ett ovälkommet epitet.[22]

Men det tekniska arbetet fortsatte i oförminskad takt, och under 1990-talet började den andra AI-vintern töa bort. Optimismen återuppväcktes genom införandet av nya tekniker som såg ut att erbjuda alternativ till det traditionella logicistiska paradigmet (ofta omtalat som "Good Old-Fashioned Artificial Intelligence", förkortat "GOFAI"), vilket fokuserat på symbolmanipulation på hög nivå och nått sin höjdpunkt med 1980-talets expertsystem. De tekniker som nu blev populära, däribland neuronnätverk och genetiska algoritmer, verkade kunna övervinna en del av GOFAI-modellens brister, i synnerhet den "bräcklighet" som kännetecknade klassiska AI-program (som vanligen producerade rent nonsens om programmerarna gjorde något enda lätt felaktigt antagande). De nya teknikerna kunde skryta med ett mer organiskt funktionssätt. Neuronnätverk uppvisade till exempel egenskapen feltolerans (*graceful degradation*): en mindre skada i ett neuronnätverk resulterade vanligen i en mindre försämring av dess prestanda, snarare än i en total krasch. Ännu viktigare var att neuronnätverk kunde lära sig av erfarenheten genom att hitta naturliga sätt att generalisera utifrån exempel och upptäcka dolda statistiska mönster i sina indata.[23] Detta gjorde nätverken bra på mönsterigenkänning och klassifikationsproblem. Genom att träna upp ett neuronnätverk på till exempel data i form av ekolodssignaler kunde man få det att lära sig skilja mellan akustiska profiler för ubåtar, minor och vattenlevande djur med större noggrannhet än mänskliga experter – och detta utan att någon på förhand behövde räkna ut exakt hur kategorierna skulle definieras eller olika egenskaper viktas.

Medan enkla neurala nätverksmodeller hade varit kända sedan slutet av 1950-talet fick fältet en renässans efter införandet av en

”bakåtpropagerande algoritm” (*backpropagation algorithm*), en form av felåterföring som gjorde det möjligt att träna flerskiktade neuronnätverk.[24] Sådana flerskiktade nätverk, med ett eller flera mellanliggande (”dolda”) lager av neuroner mellan ingångs- och utgångslagren, kan lära sig ett mycket större spektrum av funktioner än sina enklare föregångare.[25] I kombination med de allt kraftfullare datorer som började bli tillgängliga gjorde dessa algoritmiska förbättringar det möjligt för ingenjörer att konstruera neuronnätverk som var tillräckligt bra för att vara praktiskt användbara i många tillämpningar.

De neurala nätverkens hjärnliknande egenskaper var en positiv kontrast till traditionella, regelbaserade GOFAI-system, med deras strikt logiskt framräknade men bräckliga resultat. Skillnaden inspirerade till och med en ny ”-ism”, *konnektionism*, som betonade betydelsen av massivt parallell subsymbolisk databearbetning. Över 150 000 akademiska artiklar har sedan dess publicerats om artificiella neuronnätverk, och de fortsätter att utgöra en viktig metod inom maskininlärning.

Evolutionsbaserade metoder, som genetiska algoritmer och genetisk programmering, är ett annat arbetssätt vars framväxt bidrog till att få slut på den andra AI-vintern. De fick kanske inte lika stort akademiskt genomslag som neurala nät, men blev välkända för en större publik. I evolutionära modeller upprätthålls en population av lösningskandidater (som kan vara datastrukturer eller program) och nya lösningskandidater genereras slumpmässigt genom att mutera eller omkombinera varianter i den befintliga populationen. Med jämna mellanrum beskärs populationen genom att tillämpa ett urvalskriterium (en duglighetsfunktion) som låter endast bättre kandidater överleva till nästa generation. Upprepat över tusentals generationer ökar detta gradvis den genomsnittliga kvaliteten på lösningarna i kandidatpoolen. När denna typ av algoritm fungerar kan den producera effektiva lösningar för ett mycket brett spektrum av problem – lösningar som kan vara slående nyskapande och ointuitiva, och ofta ser ut mer som naturliga strukturer än något som en mänsklig ingenjör skulle konstruera. Och i princip kan detta ske utan större behov av mänsklig input

utöver det inledande angivandet av duglighetsfunktionen, som ofta är mycket enkel. Men att få evolutionära metoder att fungera tillfredsställande kräver i praktiken skicklighet och skarpsinne, särskilt för att utforma ett bra representationsformat. Utan ett effektivt sätt att koda lösningskandidater (ett genetiskt språk som matchar latent struktur i måldomänen) tenderar evolutionär sökning att slingra sig fram i det oändliga genom en stor sökrymd eller fastna vid ett lokalt optimum. Och även om man hittar ett bra representationsformat kräver evolutionen mycket beräkningskraft och besegras ofta av den kombinatoriska explosionen.

Neuronnätverk och genetiska algoritmer är exempel på metoder som tycktes erbjuda alternativ till det stagnerande GOFAI-paradigmet och därmed väckte entusiasm under 1990-talet. Men avsikten här är inte att lovsjunga dessa två metoder eller upphöja dem över de många andra tekniker för maskininlärning som finns. En viktig teoretisk utveckling de senaste tjugo åren har i själva verket varit en klarare insikt om att ytligt sett disparata tekniker kan förstås som specialfall inom ett gemensamt matematiskt ramverk. Så kan till exempel många typer av artificiella neuronnätverk ses som klassificerare som utför en viss typ av statistisk beräkning (uppskattning av maximal sannolikhet).[26] Detta perspektiv gör att neurala nät kan jämföras med en större klass av algoritmer för att lära sig klassificerare utifrån exempel – ”beslutsträd”, ”logistiska regressionsmodeller”, ”stödvektormaskiner”, ”naiv Bayes”, ”*k*-närmaste grannar-regression” med flera.[27] På liknande sätt kan genetiska algoritmer sägas syssla med stokastisk bergsklättring, som i sin tur är en delmängd av en större klass av algoritmer för optimering. Var och en av dessa algoritmer för att bygga klassificerare eller söka igenom en lösningsrymd har sin egen profil i fråga om styrkor och svagheter som kan studeras matematiskt. Algoritmer skiljer sig åt i sina krav på processortid och minnesutrymme, vilka induktiva förutsättningar de utgår från, hur lätt externt producerat innehåll kan införlivas och hur transparenta deras inre funktionssätt är för en mänsklig analytiker.

Bakom all uppståndelse kring maskininlärning och kreativ

Ruta 1. En optimal bayesiansk agent

En ideal bayesiansk agent börjar med en ”apriori sannolikhetsfördelning”, en funktion som tilldelar sannolikheter till varje ”möjlig värld” (det vill säga varje maximalt specifikt sätt som världen kunde visa sig vara beskaffad på).[29] Denna aprioriförde lning rymmer en induktiv bias som är sådan att enklare möjliga världar tilldelas högre sannolikheter. (Ett sätt att formellt definiera enkelheten hos en möjlig värld är i termer av dess ”kolmogorovkomplexitet”, ett mått baserat på längden av det kortaste datorprogram som genererar en fullständig beskrivning av denna värld.[30]) Aprioriförde lningen inbegriper också alla bakgrundskunskaper som programmerarna vill ge agenten.

Efterhand som agenten erhåller ny information från sina sensorer uppdaterar den sin sannolikhetsfördelning genom att villkora fördelningen utifrån den nya informationen i enlighet med Bayes teorem.[31] Villkorning är den matematiska operation som anger noll som den nya sannolikheten för de världar som är oförenliga med den erhållna informationen och åternormaliserar sannolikhetsfördelningen över återstående möjliga världar. Resultatet är en ”aposteriori sannolikhetsfördelning” (som agenten i nästa tidssteg kan använda som sin nya aprioriförde lning). Efterhand som agenten gör observationer koncentreras dess sannolikhetsmassa därmed på den krympande uppsättning möjliga världar som förblir förenliga med evidensen; och bland dessa möjliga världar har enklare varianter alltid större sannolikhet.

Metaforiskt kan vi tänka oss en sannolikhet som sand på ett stort pappersark. Papperet är uppdelat i områden av olika storlek, där varje område motsvarar en möjlig värld och större områden motsvarar enklare möjliga världar. Vi tänker oss också att ett jämntjockt lager av sand är utspritt över hela arket: detta är vår aprioriförde lning. Varje gång en observation görs som utesluter vissa möjliga världar tar vi bort sand från motsvarande områden på papperet och fördelar den jämnt över de områden som fortfarande är med i spelet. Den totala mängden sand på arket förändras alltså aldrig, utan koncentreras bara till färre områden efterhand som observationell evidens ackumuleras. Detta är en bild av inlärning i dess renaste form. (För att beräkna sannolikheten hos en *hypotes* mäter vi helt enkelt mängden sand i alla områden som motsvarar de möjliga världar i vilka hypotesen är sann.)

Så långt har vi definierat en inlärningsregel. För att få en agent behöver vi

också en beslutsregel. I detta syfte förser vi agenten med en ”nyttofunktion” som tilldelar varje möjlig värld ett tal. Talet representerar hur önskvärd den världen är enligt agentens grundpreferenser. I varje tidssteg väljer agenten nu den handling som har högst förväntad nytta.[32] (För att hitta handlingen med högst förväntad nytta kan agenten göra en lista över alla handlingar som är möjliga. Den kan sedan beräkna den villkorliga sannolikhetsfördelningen givet handlingen – den sannolikhetsfördelning som skulle bli resultatet av att villkora dess nuvarande sannolikhetsfördelning med observationen att denna handling just hade utförts. Slutligen kan agenten beräkna det förväntade värdet av handlingen som summan av värdet av varje möjlig värld multiplicerat med den villkorliga sannolikheten av den världen givet handlingen.[33])

Inlärningsregeln och beslutsregeln definierar tillsammans ett ”optimalitetsbegrepp” för en agent. (Väsentligen samma optimalitetsbegrepp har ofta använts inom artificiell intelligens, epistemologi, vetenskapsfilosofi, ekonomi och statistik.[34]) I verkligheten är det omöjligt att bygga en sådan agent, eftersom de kalkyler som skulle krävas är beräkningsmässigt ohanterliga. Varje försök att göra det kollapsar i en kombinatorisk explosion av precis det slag som beskrivs i vår diskussion om GOFAI. För att förstå varför det blir så kan vi betrakta en liten delmängd av alla möjliga världar: de som består av en enda datorskärm svävande i ett ändlöst vakuum. Skärmen har 1 000 × 1 000 pixlar, som var och en är permanent på eller av. Till och med denna delmängd av möjliga världar är enormt stor: de $2^{(1000 \times 1000)}$ möjliga skärmtillstånden är fler än alla beräkningar som någonsin förväntas äga rum i det observerbara universum. Vi skulle alltså inte ens kunna räkna upp alla möjliga världar i denna lilla delmängd av alla möjliga världar, för att inte tala om att utföra mer avancerade beräkningar för var och en av dem individuellt.

Optimalitetsbegrepp kan vara av teoretiskt intresse även om de är fysiskt ogenomförbara. De ger oss en måttstock för att bedöma heuristiska approximationer, och ibland kan vi resonera om vad en optimal agent skulle göra i något särskilt fall. I kapitel 12 tar vi upp en del alternativa optimalitetsbegrepp för artificiella agenter.

problemlösning ligger alltså en uppsättning matematiskt välspecificerade avvägningar. Idealet är den perfekta bayesianska agenten, som gör probabilistiskt optimalt bruk av tillgänglig information. Detta ideal är ouppnåeligt eftersom det är alltför beräkningskrävande för att kunna implementeras i en fysisk dator (se Ruta 1). Artificiell intelligens kan därför ses som ett försök att hitta genvägar: hanterliga sätt att approximera det bayesianska idealet genom att offra något i optimalitet eller generalitet men bevara tillräckligt mycket för att få hög prestanda på de områden som faktiskt är av intresse.

En återspegling av den här bilden kan man se i de senaste decenniernas arbete kring probabilistiska grafiska modeller, såsom bayesianska nätverk. Bayesianska nätverk ger ett koncist sätt att representera probabilistiska och villkorliga oberoenderelationer som gäller i någon bestämd domän. (Att utnyttja sådana oberoenderelationer är väsentligt för att övervinna den kombinatoriska explosionen, som utgör ett problem för probabilistisk slutledning i lika hög grad som för logisk härledning.) De ger också viktig insikt i kausalitetsbegreppet.[28]

En fördel med att relatera inlärningsproblem inom specifika domäner till det allmänna problemet med bayesianska slutledningar är att nya algoritmer som gör bayesianska slutledningar mer effektiva då kommer att leda till omedelbara förbättringar på många olika områden. Framsteg när det gäller tekniker för Monte Carlo-approximering kan till exempel tillämpas direkt inom datorseende, robotik och beräkningsgenetik. En annan fördel är att det gör det lättare för forskare från olika discipliner att slå samman sina resultat. Grafiska modeller och bayesiansk statistik har blivit ett gemensamt fokus för forskning inom många områden, däribland maskininlärning, statistisk fysik, bioinformatik, kombinatorisk optimering och kommunikationsteori.[35] Åtskilliga av den senaste tidens framsteg inom maskininlärning är resultatet av att införliva formella resultat som ursprungligen utvunnits på andra akademiska områden. (Tillämpningar av maskininlärning har också dragit enorm nytta av snabbare datorer och ökad tillgång till stora datamängder.)