AI – hot eller möjlighet?

De senaste åren har mänskligheten mött utmaningar som har inneburit hastiga förändringar. Några av dessa kan ge intryck av att i grunden hota människans existens. Coronapandemin visade hur snabbt och okontrollerbart smitta kan sprida sig i en värld, där forskning om virus sker med utan insyn samtidigt som handel och umgänge med andra djurarter sker utan reglering. Klimathotet har nått en punkt där den mänskliga oförmågan till nödvändiga åtgärder har lett till en överhängande risk att vi inom kort står inför ödeläggelse, massutrotning och konflikter som följd av jordens krympande marginaler och resurser. Och användning av kärnvapen kan numera ge illusionen av ett rationellt alternativ för en hårt trängd despot, som står inför valet att mista makten i total förnedring eller att tvinga sin motståndare på knä enligt samma logik som motiverade bombningarna av Hiroshima och Nagasaki.

Listan över för mänskligheten existentiella domedagsscenarier kan göras ännu längre genom att tillfoga hoten från överanvändning av penicillin, risken för kollision med meteoriter, faran i att bli upptäckta en aggressiv och överlägsen utomjordisk civilisation för att bara nämna några ytterligare möjligheter. Eller varför inte att den allsmäktige gud – som så många människor bekänner sig till – slutligen gör verklighet av profetiorna om Domens dag och sätter punkt för människans härjningar i den skapelse, som denne gud enligt många religioner har satt mänskligheten att råda över.

Den allra livligast blomstrande grenen på detta domedagsträd är för närvarande hotet från artificiell intelligens. Under många årtionden har vi sett hur datorernas processhastighet och minneskapacitet har accelererat på ett häpnadsväckande vis. Frågan om huruvida denna tillväxt har en bortre gräns har blivit allt mer överhängande. Det har talats om risken för att den skenande utvecklingen av den artificiella intelligensens maximala kapacitet kommer att nå så långt, att den har möjlighet att förbättra sin maximala kapacitet på egen hand. Resultatet skulle enligt enligt förespråkarna för denna möjlighet bli en utveckling helt bortom kontroll. En övermänskligt avancerad artificiell intelligens skapar en ännu mer avancerad intelligens som i sin tur skapar ytterligare förbättringar, och detta förutsätts kunna ske i en hastighet, som överskrider vad vi förmår att beskriva i våra matematiska modeller.

Som alternativ till detta scenario existerade länge möjligheten att utvecklingen av icke biologisk intelligens skulle nå en gräns som hindrar okontrollerad tillväxt. Matematiskt brukar en sådan

utveckling beskrivas med hjälp av den *logistiska funktionen*, som förutsäger hur en *exponentiell utveckling* förr eller senare övergår i stagnation i en miljö som är behäftad med begränsningar. Fysikens lagar medför självklara begränsningar av detta slag. Exempelvis kan en signal inte kan färdas fortare än ljusets hastighet, och atomernas egenskaper sätter gränser för datorminnets kapacitet.

Emellertid har de senaste årtiondenas utveckling inte uppvisat några tecken på avmattning. Snarare har den på 1960-talet formulerade *Moores lag* – som förutspår en fördubbling av väsentliga delar av datorers kapacitet vartannat år – förbluffande väl visat sig vara korrekt. Till och med finns möjligheten att tillväxttakten kommer att ta ett kvalitativt språng som överträffar Moores förutsägelser den dag som *kvantdatorer* kommer till praktisk användning.

Istället finns tecken på att vi står på tröskeln till den okontrollerbara utveckling som det har varnats för, och kanske har vi redan har klivit över denna tröskel. Förmodligen kan vi inte konstatera hur det förhåller sig med den saken förrän vi är så långt komna i utvecklingen att det är försent att lägga om kursen. I den mån vi någonsin kommer att ha handlingsutrymme att påverka utvecklingen av artificiell intelligens i önskvärd riktning kan det vara just nu, och det finns därför starka skäl att med största noggrannhet ta ställning till de möjligheter som står till buds.

Vilka är då de ”tecken” som tyder på att framväxten av artificiell intelligens har nått en punkt, där det finns överhängande risk att dess utveckling snart kommer att ske utom mänsklig kontroll? Det starkaste av dessa tecken är förmodligen de varningar som kommer från de personer som själva har initierat och drivit på denna utveckling under de senaste årtiondena. Det kanske mest framträdande exemplet är Geoffery Hinton som i mars 2023 beslutade att avgå från sin sitt arbete som utvecklare av artificiell intelligens på Google. Som skäl för sin avgång angav han, att han önskar att som fristående röst kunna varna för de risker mänskligheten står inför i detta avseenden, risker som han förutspår ligger mindre än tjugo år fram i tiden.

Likaså innebär den i slutet av 2022 lanserade ”chatboten” chatGPT och dess efterföljare en milstolpe i utvecklingen av den artificiella intelligensen – framför allt med avseende på hur vi människor interagerar med sådan intelligens i vår vardag. ChatGPT bygger på en ”autoregressiv språkmodell som använder djupinlärning” (Wikipedias beskrivning), vilket bland annat innebär att den utifrån statistiska modeller genomgår en

kontinuerlig inlärningsprocess. Det visar sig att chatboten har förmåga att svara på frågor och interagera med människor på ett sätt som ger användaren illusionen av att kommunicera med en intelligent varelse med mänskliga egenskaper, och den visar sig ha förmåga att producera högkvalitativa texter av såväl akademisk som rent underhållande karaktär.

Ett skäl till att betrakta chatGPT som ett faktiskt genombrott i interaktionen mellan maskin och människa är att detta och liknande system öppnar närmast obegränsade möjligheter för att framöver manipulera människor med hjälp av artificiell intelligens. Redan tidigare har det förekommit manipulerade bilder och påverkanskampanjer understödda av artificiell intelligens i syfte att påverka affärsmässiga och politiska förhållanden i vår värld, men när människa och maskin inte längre kan särskiljas står vi mer oskyddade än någonsin mot främmande och potentiellt ondsinta krafters inflytande över vad vi ska tro och inte tro på. När vi inte längre på ett självklart kan skilja artificiell intelligens från mänsklig intelligens har en högst dramatisk scenförändring inträffat.

Det är lätt att bli skrämd av de möjligheter och farhågor som dessa framsteg öppnar för. Kanske förebådar de slutet på det herravälde över vår planet och andra arter, som vi människor har vant oss vid under de senaste par årtusendena. Kanske innebär detta att människan riskerar att bli underordnad en övermänsklig intelligens, om denna intelligens i kombination med sin överlägsna förmåga har utrustats med värderingar och mål som står i strid med mänsklighetens. Detta väcker frågan om det är möjligt eller ens önskvärt att hindra utvecklingen av den artificiella intelligensen.

Att hindra utvecklingen mot underkastelse gentemot en överlägsen artificiell intelligens skulle kräva att vi hindrar utvecklingen på ett relativt tidigt stadium. Det bör ju ske innan den potentiellt överlägsna artificiella intelligensen har utvecklats så långt att den kan hindra försöken från en lägre stående intelligens – den mänskliga – att stävja dess eventuella strävan efter herravälde. Det skulle i sin tur kräva en världsomspännande överenskommelse om att stoppa den nuvarande oerhört snabba utvecklingen på nuvarande plan eller i varje fall bromsa den på ett drastiskt vis, Med tanke på den kapplöpning som såväl kommersiellt som politiskt försiggår på den artificiella intelligensens område, skulle detta kräva ett ansvarstagande och en samordning från världens kommersiella och politiska ledare som vida överstiger vad de hittills har visat prov på.

Ett annat skäl till varför den nuvarande utvecklingen mot en allt mer förfinad artificiell intelligens framstår som svår att hindra är den rad av hot som människan står inför, och som nämndes i denna texts inledning. Mänskligheten ger intryck av att ha hamnat i en ytterst farlig belägenhet med avseende på klimat- och kärnvapenhot, och ifråga om framtida pandemier, resistenta bakterier och möjliga kollisioner med andra himlakroppar gör vi klokt i att med alla krafter minimera de risker som är förknippade även med dessa potentiella faror. Frågan är därför om vi har råd att bromsa den utveckling på den artificiella intelligensens område, som kanske är den enda till buds stående räddningen från de annalkande katastrofer som människan efter eget förvållande står inför.

Enligt detta synsätt är det fullt möjligt att vi är på väg att ställas inför ett vägval, där mänskligheten antingen underkastar sig en övermänsklig artificiell intelligens, eller där mänskligheten behåller herraväldet över en planet som är dömd till undergång. Människans strävan efter autonomi och frihet ställs mot hennes önskan att överleva. Religionernas beskrivning av människan som skapelsens krona ställs mot ett evolutionärt motiverat förhållningssätt, där genernas fortbestånd är det enda som räknas.

Vad som blir resultatet av ett sådant val framstår i nuläget om omöjligt att förutsäga. En isolerad och desperat diktator, en självförstärkande avsmältning av inlandsisar eller en ny pandemi på nivå med digerdöden skulle få alla prognoser att med ens kastas över ända. Världen är instabil, och varje år som går med obeslutsamma politiker, lättmanipulerade väljare och religiösa ledare som ivrigt påbjuder sina följare att tåga mot helvetets portar innebär ökar risken för att den ryska rouletten tar ände med förskräckelse.

För min egen personliga del är valet enkelt. Låt oss underkasta oss den framväxande artificiella intelligensen med samma självklara lugn som vi skulle tillåta en klok mänsklig världsledare att föra oss ut ur det mörker som tätnar omkring oss. Allt annat vore att låta religiösa dogmer, missriktad frihetssträvan och en obefogad tilltro till vetenskapens förmåga att lösa snarare än skapa problem bestämma utvecklingen snarare än rationalitet och omsorg om kommande generationer.

En möjlig invändning mot denna uppmaning är att otyglad utveckling för den artificiella intelligensen skulle lämna fältet fritt för ”onda” mänskliga intressen. Dessa intressen skulle i så fall kunna utrusta den överlägsna intelligensen med mål och värderingar som innebär en ännu sämre prognos för planetens

framtid än om människan själv får styra. Detta framstår emellertid som ett föga troligt scenario. Även de mest destruktiva, människocentrerade och inskränkta människorna har tillräckligt med självbevarelsedrift för att ange undanröjandet av klimathotet och förmodligen även kärnvapenhotet som ett par av sina främst mål.

Och skulle någon människa till äventyrs ändå välja självdestruktion genom kärnvapenkrig som ett av den artificiella intelligensens mål, måste den individen förmodligen samtidigt göra sig till herre över den artificiella intelligensen för att kunna genomföra detta. En övermänsklig artificiell intelligens som utvecklar sig själv skulle knappast tillåta en så destruktivt mål som kärnvapenkrig utan någon form av hämning eller underodning.

Mot en sådan ”ond” kraft får en förmoda att andra ”goda” krafter skulle ställa annan artificiell intelligens som i enlighet med sina självgenererade värderingar vill hindra användandet av kärnvapen. Detta skulle således innebära någon form av kamp mellan en destruktiv artificiell intelligens som är underkastad en mänsklig kontroll och en konstruktiv artificiell intelligens som inte är det. Det framstår som tämligen självklart att den senare skulle avgå med seger i en sådan kamp, eftersom en hämmad artificiell intelligens med nödvändighet är svagare än en ohämmad.

Det allt mer självklara svaret på frågan om hur mänskligheten ska förhålla sig till en hastigt utvecklande artificiell intelligens med potential att bli överlägsen människan är således: släpp den fri. Välkomna att den så snart som möjligt avhänder mänskligheten det ansvar för planeten jorden som den sedan länge har visat vara oförmögen att klara.

*