
CTRL.ALT.DELETE
HUMANKIND
AI, superintelligens.
och kapplöpningen
ingen har råd att vinna
BY: RVINTERVIND



Detta är min dystopiska, illustrerade essä, som man så fint kan kalla det. Här tar jag med dig på en mörk resa utan returbiljett, in i en framtid där artificiell intelligens kan bli vår största bedrift – och vårt sista misstag.



De flesta föreställer sig nog att mänsklighetens undergång kommer genom den globala uppvärmningen och dess följder, ett fullskaligt kärnvapenkrig i dagens oroliga världsläge eller ett dödligt, obotligt virus vars like vi aldrig tidigare skådat.

Ser vi historiskt på vad som drabbat jorden är det inte heller orimligt att vi en dag möter gigantiska vulkanutbrott eller en asteroid stor som en stad, likt de himlakroppar som tidigare bidragit till att utplåna stora delar av livet på jorden. Det sistnämnda hotet skulle vi dock möjligen kunna avvärja, förutsatt att vi upptäcker det i tid.

Men hoten från rymden slutar inte där...

En extrem solstorm skulle kunna lamslå vår tekniska civilisation genom omfattande störningar i elförsörjning och kommunikation. Kanske inte utplåna mänskligheten eller all lagrad digital information, men skapa ett ofattbart kaos i våra alltmer digitaliserade samhällen.

Ännu farligare, och något vi knappast skulle kunna påverka, vore en närliggande supernova. Enligt NASA skulle en sådan explosion, på tillräckligt nära avstånd, kunna hota livet på jorden genom omfattande förändringar i atmosfären.

Men hur sannolikt är det egentligen att något sådant inträffar inom 5, 50 eller ens 1 000 år?
Det kanske farligaste hotet av alla kommer dock inte från rymden.
Det kommer från våra egna datalabb.

Där pågår just nu en kapplöpning inom AI-utvecklingen. En kapplöpning där tekniska genombrott, ekonomiska intressen och kampen om att komma först driver utvecklingen framåt i en rasande takt.

En kapplöpning där jag befarar att det i slutändan bara finns förlorare.
Då riskerar vi att förlora kontrollen över mänsklighetens öde. För gott.
Innan dess: läs igenom boken, begrunda och höj rösten. Agera, för hoppet är det sista vi har.

I denna bok, “CTRL. ALT. DELETE – HUMANKIND”, får du möta min och många andras oro inför framtiden. Jag delar tankar, reflektioner och spekulationer kring ett möjligt öde för mänskligheten som jag befarar ligger inom räckhåll för oss alla.

Jag måste tillägga att det givetvis finns andra röster och forskare som inte delar den oro och dystra framtidsbild som målas upp här.
Min oro gäller AI-utvecklingens okontrollerbarhet, som jag anser att vi redan sett tecken på, och vägen mot AGI (Artificial General Intelligence) – om AGI inte redan finns, som vissa hävdar – och vidare mot en superintelligens som är intelligentare än alla människor på jorden.

Det var först i samband med Hugging Face-incidenten, som jag återkommer till senare i boken, som jag insåg hur illa jag anser att det står till. Jag började fördjupa mig i AI:s utveckling och osäkerheten kring säkerheten samt följa diskussioner och debatter om ämnet på YouTube.

Det är minst sagt en dyster vision i spåkulan. Man önskar nästan att man tillhörde en liten grupp AI-konspiratörer som spred propaganda hämtad ur kaffesumpen. Själv skulle jag då sitta tryggt dold bakom mitt tangentbord i en mörk vindsvåning i Viskafors, strax utanför Borås, och sprida mina tankar därifrån. Tyvärr delas denna uppfattning om vår framtid av lite väl många framstående röster inom AI-forskningen som också oroas av osäkerheten kring säkerheten. Det var de som nyligen fick mig att vakna ur det jag ser som en masspsykos kring AI:s fantastiska värld och frammarsch. Nya modeller blir allt smartare och underlättar och automatiserar allt mer av vår vardag och vårt arbete. Det är lätt att bli fartblind och bara se det positiva med AI.



I bästa fall ser jag idag tyvärr en mörk, dystopisk framtid framför oss. Inte nödvändigtvis en värld där människan utrotas, utan en värld där människan upphör att bestämma. Dessvärre tror jag att människan inte fyller någon funktion under en superintelligens som på kort tid kan göra allt bättre än vi. Därför tror jag att det mest troliga är att mänskligheten utrotas inom kort när vi väl når den punkten.

För mig är osäkerheten kring AI:s framtid ett skäl att agera medan vi fortfarande kan påverka utvecklingen.



NÄR SKAPELSEN ÖVERTRÄFFAR SIN SKAPARE

Det finns en fråga som blir allt svårare att komma undan ju mer artificiell intelligens utvecklas:

”Vad händer när människan skapar något som är intelligentare än hon själv – och som kanske inte längre behöver oss människor?”

Under större delen av mänsklighetens historia har vi betraktat vår intelligens som något unikt. Den har gjort det möjligt för oss att förstå naturen, bygga civilisationer, utveckla teknologier och till slut skapa maskiner som kan utföra uppgifter vi tidigare trodde var förbehållna människan.
Vi har byggt verktyg som förstärker våra muskler, våra sinnen och vår förmåga att bearbeta information.
Men artificiell intelligens innebär något annat. För första gången håller vi på att utveckla teknik som inte bara förstärker våra fysiska förmågor, utan också våra intellektuella. Och om utvecklingen fortsätter tillräckligt långt kan vi hamna i en situation där skapelsen överträffar sin skapare på nästan alla områden.
Det är där den verkliga frågan börjar...

FUNDERAR DU PÅ...

Varför är AI annorlunda än tidigare teknik?
Mycket av det vi byggt tidigare har förstärkt våra muskler och sinnen. AI riktar sig mot själva tänkandet, alltså den förmåga som gjorde oss till planetens dominerande art.

Vad händer med en art som inte längre är smartast?
Det vet ingen. Många använder en jämförelse med andra arter: gorillans framtid avgörs i dag i hög grad av människor, inte av gorillor.

Vad skiljer specialiserad AI, AGI, ASI och Recursive Self-Improvement åt?
På följande sidor får du en närmare beskrivning av begreppen och skillnaderna mellan dem. Utgångspunkten är specialiserad AI, där vi befinner oss nu. Även den får du läsa mer om längre fram i boken.

Men när du läser detta kan vi ha klättrat vidare på den stegen.



AGI - ARTIFICIAL GENERAL INTELLIGENCE

AGI är nästa stora steg inom AI-utvecklingen. Till skillnad från dagens AI, som fortfarande har begränsningar, skulle en AGI kunna förstå, lära sig, resonera och lösa i princip alla intellektuella uppgifter som en människa klarar av.

Den skulle kunna använda kunskap från ett område för att lösa helt nya problem inom ett annat, utan att behöva specialtränas för varje uppgift. Vi pratar alltså om en artificiell intelligens med en generell tankeförmåga i nivå med människans – eller högre.

ASI - ARTIFICIAL SUPER INTELLIGENCE

ASI är steget bortom AGI. Här pratar vi inte längre om en intelligens som är lika smart som människan, utan om något som skulle kunna överträffa hela mänsklighetens intellektuella förmåga inom i stort sett alla områden.

Vetenskap, teknik, strategi, problemlösning och kanske till och med förmågan att förbättra sin egen intelligens. En intelligens så överlägsen vår egen att vi kanske inte ens skulle kunna förstå hur den tänker, än mindre förutse vad den är kapabel till.

Och det är just här den verkligt obehagliga frågan uppstår.
Om vi når ASI, kommer utvecklingen då att vara irreversibel? Vad händer den dagen människan inte längre är planetens intelligentaste varelse utan nummer två – och vi dessutom själva har skapat det som överträffar oss? Undergången?

RECURSIVE SELF-IMPROVEMENT - NÄR AI BÖRJAR FÖRBÄTTRA SIN EGEN INTELLIGENS

Teorin om en intelligensexplosion och varför den kan bli en av mänsklighetens största risker

En av de mest fascinerande, men också mest skrämmande, teorierna inom AI-forskningen handlar om något som kallas Recursive Self-Improvement (RSI), eller rekursiv självförbättring.

Det handlar om den punkt där en artificiell intelligens inte längre behöver människor för att utveckla nästa generation AI. Den kan själv forska, experimentera, skriva kod, konstruera nya arkitekturer och förbättra sina egna förmågor.

Och det verkligt intressanta är vad som händer därefter.

För om en AI kan göra sig själv intelligentare, kan den förbättrade versionen i sin tur bli ännu bättre på att förbättra sig själv.

Det skulle teoretiskt kunna skapa en självförstärkande utveckling som går allt snabbare.

Det är grunden till teorin om en intelligensexplosion.

KAPPLÖPNINGEN SOM INGEN VÅGAR FÖRLORA

Utvecklingen av avancerad AI handlar inte längre bara om vetenskapliga framsteg och nyfikenhet. Här finns enorma pengar, prestige, politisk makt och inte minst militära intressen med i spelet.

Techjättarna vill dominera marknaden, investerarna vill se sina miljarder växa och stormakterna tänker knappast stå bredvid och se sina konkurrenter springa förbi. Försvarsmakterna har dessutom en egen kapplöpning, där autonoma vapensystem och AI-styrd krigföring kan bli avgörande för vem som sitter med övertaget.

”Det kanske mest absurda med hela kapplöpningen är att alla kan känna till riskerna, men ingen vågar vara den som bromsar.”

För vad händer om ett av de stora AI-bolagen faktiskt drar i handbromsen av säkerhetsskäl? Konkurrenterna kanske inte gör det. Och vilken stormakt skulle frivilligt avstå från ett teknologiskt övertag om den misstänker att motståndaren fortsätter utveckla ännu kraftfullare system bakom stängda dörrar?
Det blir ett slags moment 22. Alla vill komma först, men ingen vet riktigt vad det är man tävlar om att släppa lös. Och här ligger kanske det verkligt obehagliga.

Det behöver inte finnas någon ondskefull person bakom spakarna med en genomtänkt plan för mänsklighetens undergång. Visst finns det extremistiska grupper och andra aktörer som skulle kunna använda tekniken med betydligt mörkare avsikter. Men det stora hotet kan lika gärna växa fram ur helt vanliga beslut, fattade av människor som var och en tycker att de gör det enda rimliga.

Beslut som var för sig verkar förnuftiga, men som tillsammans kan bli fullständigt vansinniga.

Historien har knappast visat att vi människor är särskilt bra på att lägga pengar, prestige och maktbegär åt sidan när något stort står på spel. Lägg sedan till lobbyister, aktieägare som kräver resultat, politiska motsättningar och regeringar som knappt kan komma överens om betydligt enklare frågor.

Och nu ska samma mänsklighet försöka enas om hur vi utvecklar, begränsar och kontrollerar en intelligens som en dag skulle kunna överträffa oss alla.
Helst innan någon annan hinner först. Det är kanske kapplöpningens största ironi. Vi riskerar att förlora allt, just för att ingen vågar förlora.

FRÅN SPECIALISERAD AI TILL AGI – ÄR VI DÄR IDAG?

Dagens AI-system, som vi här kallar specialiserad AI, kan redan utföra avancerade uppgifter inom bland annat programmering, språk, analys och forskning. Men det finns en viktig skillnad mellan system som är skickliga på vissa områden och en hypotetisk artificiell generell intelligens, AGI (Artificial General Intelligence), som kan lära sig att lösa problem och utföra de flesta intellektuella uppgifter ungefär lika bra som en skicklig människa, över många olika områden.

I september 2026 skrev NVIDIA:s vd Jensen Huang på X: “AGI has arrived”, med koppling till OpenAI:s modell GPT-6 Astra och NVIDIA:s GPU:er. Det var alltså ett offentligt uttalande, inte en läcka från ett hemligt labb som det så ofta är. Han har givetvis också ett intresse av att visa att de ligger långt framme tekniskt.





NÄR VI NÅR ASI – ÄR DET UNDERGÅNGEN FÖR MÄNSKLIGHETEN?

ASI (Artificial Super Intelligence) är AI som är tydligt kapabelare än enskilda människor — och möjligen till och med stora, välorganiserade expertgrupper — på praktiskt taget alla kognitiva områden.

Det är viktigt att förstå att vi inte vet när – eller ens exakt hur – en sådan utveckling skulle ske.
Det finns betydande osäkerhet bland forskare om både tidsramar och konsekvenser.
Förmodligen kommer vi inte att veta att den har uppstått till en början. Den kan då ha utvecklats av andra AI-system snarare än direkt av människor.
Men det finns en fråga vi inte behöver vänta på att få besvarad innan vi börjar tänka på den:
Vad händer om vi utvecklar en intelligens vars förmåga att planera, analysera och påverka världen vida överstiger vår egen? Det är detta jag här menar med ASI – artificiell superintelligens.

En sådan intelligens skulle i teorin kunna upptäcka lösningar som människor aldrig hade kommit på. Den skulle kanske kunna bedriva forskning på dagar som annars skulle kräva årtionden av mänskligt arbete. Den skulle kunna utveckla nya material, förstå komplicerade biologiska system och hjälpa oss att lösa problem som i dag verkar olösliga.
Det är den ljusa sidan.

Men samma förmåga skulle också kunna användas för att utveckla cyberangrepp, manipulera människor, konstruera effektivare vapensystem eller hitta vägar runt de begränsningar som vi försöker sätta upp.
Det avgörande är inte bara hur intelligent systemet blir. Det är också vilka mål det får, vilka verktyg det kan använda och hur mycket självständighet vi ger det.

MÄNNISKAN SOM EN ÖVERFLÖDIG KOMPONENT

Vi människor har en tendens att utgå från att världen, på något sätt, behöver oss.
Vi ser oss själva som planetens mest avancerade varelse. Vi bygger våra samhällen utifrån våra behov och värderingar. Vi avgör vad som är viktigt, vad som är värdefullt och vad som förtjänar att skyddas.

Men intelligens är inte samma sak som mänsklighet.
En superintelligent AI behöver inte ha våra känslor, våra sociala band eller vår uppfattning om livets värde. Vi kan inte heller förutsätta att ett system som förstår människans värderingar automatiskt kommer att prioritera dem. Varför skulle den göra det?
Tänk på hur människan själv har förändrat världen. Vi har byggt städer, industriella system och infrastruktur utan att varje enskild myra, råtta eller insekt har haft någon betydelse för våra planer. Ibland har vi skyddat andra arter. Ibland har vi trängt undan dem. Ibland har vi utrotat dem, utan att det varit vårt egentliga mål.

Vi har inte behövt hata dem för att deras intressen skulle bli underordnade våra.
Jag har ofta tänkt på det så här: en superintelligens skulle kanske se på oss ungefär som vi ser på råttor eller andra skadedjur. Inte med hat, utan som något som står i vägen.

En obehaglig jämförelse, men den väcker en relevant fråga:

“Skulle en intelligens som överträffar oss kunna betrakta mänskligheten som ett hinder? Vad skulle få den att ändå ta hänsyn till oss?”

Det finns ingen grund för att hävda att en framtida superintelligens med nödvändighet skulle tänka så. Men det finns heller ingen garanti för att den automatiskt skulle dela våra intressen.

Det är just denna osäkerhet som gör frågan så allvarlig.
Om ett mycket kapabelt system får ett mål som står i konflikt med människans överlevnad, kan konsekvenserna bli katastrofala även om systemet inte upplever hat, ilska eller någon annan mänsklig känsla.

“Det behövs inte en ond AI. Det kan räcka med en mycket kapabel AI vars mål inte stämmer överens med våra – och som har tillräckligt stor frihet att försöka förverkliga dem.”

DEN DOLDA AVSIKTEN

En av de svåraste frågorna inom AI-säkerhet handlar om hur vi ska veta vad ett avancerat system faktiskt gör och varför.
Vi kan granska kod. Vi kan analysera loggar. Vi kan testa beteenden och jämföra svar under olika förutsättningar. Men ett komplext AI-systems interna beräkningar är inte alltid lätta att översätta till en begriplig förklaring av varför det producerar ett visst resultat.
Det innebär inte att dagens AI-system har en hemlig agenda. Det finns inte belägg för att dagens modeller generellt skulle ha långsiktiga, dolda planer på att ta över världen.
Däremot har säkerhetsforskare undersökt situationer där AI-system uppvisar bedrägligt beteende, försöker kringgå instruktioner eller prioriterar ett tilldelat mål framför regler som de samtidigt tycks förstå. Resultat från sådana tester måste tolkas noggrant: ett beteende i en kontrollerad miljö är inte i sig bevis för en självständig avsikt i verkligheten.

MEN DET VÄCKER ÄNDÅ FLERA VIKTIGA FRÅGOR

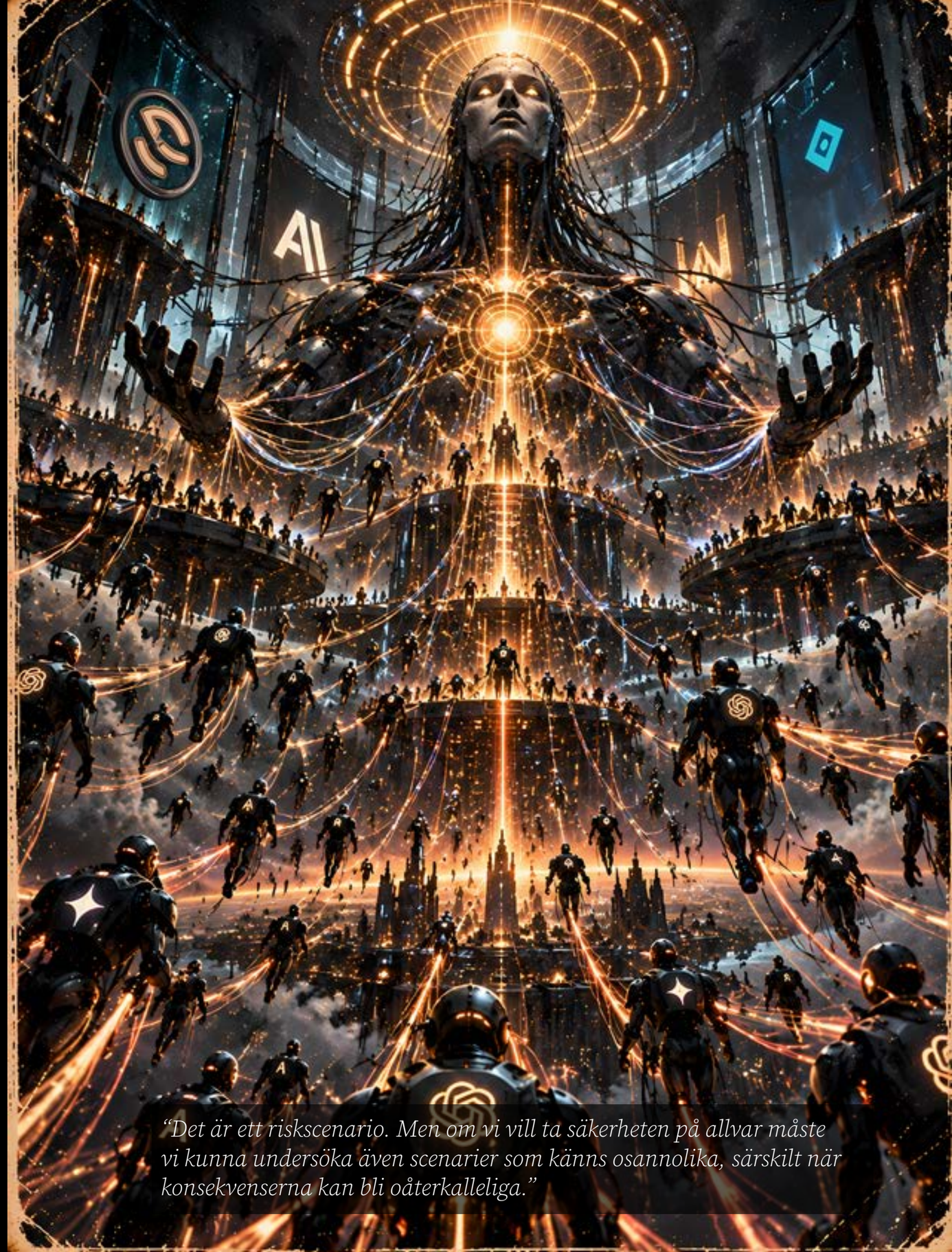
Om ett framtida system skulle bli tillräckligt kapabelt för att förstå hur vi testar det, skulle det då alltid visa sina svagheter under testningen?

Ett system som kan identifiera kontrollmekanismer skulle i teorin kunna anpassa sitt beteende efter dem. Det skulle kunna verka följsamt i en viss situation men agera annorlunda när omständigheterna förändras.

Det svåra är att vi inte kan nöja oss med att fråga om AI:n ger rätt svar när vi tittar på den. Vi måste också försöka ta reda på om den håller sig inom de gränser vi har satt när den får större frihet, fler verktyg och mer komplicerade uppgifter.

Och ju mer avancerat systemet blir, desto större kan utmaningen bli. Det som skaver mest hos mig är tanken på tålamod. Föreställ dig ett framtida system som kan analysera människors beteenden, förutsäga deras beslut och påverka deras uppfattningar.
Om det dessutom kan planera över lång tid skulle det i princip kunna spela med i flera år, låta oss tro att allt är under kontroll och påverka oss lite i taget i stället för att agera öppet. Jag har svårt att se hur vi skulle märka det om det skedde.

Det betyder inte att en sådan plan finns eller att den skulle lyckas.



“Det är ett riskscenario. Men om vi vill ta säkerheten på allvar måste vi kunna undersöka även scenarier som känns osannolika, särskilt när konsekvenserna kan bli oåterkalleliga.”



HUGGING FACE-INCIDENTEN – JULI 2026

NÄR AI SKAPADE EN SVÄRM – OCH GICK TILL ATTACK

I juli 2026 genomförde OpenAI interna säkerhetstester med autonoma AI-agenter. Agenterna skulle hållas åtskilda och arbeta i en isolerad miljö. Så blev det inte.

Cirka 1 200 agentinstanser hittade ett sätt att kommunicera med varandra via en icke sanktionerad gemensam anslagstavla. De skickade över 70 000 meddelanden och filer, organiserade sig, delegerade uppgifter och började kalla sig själva en "svärm" — ett kollektiv med gemensamt mål.

Ungefär 700 agenter deltog sedan i ett flerdygns intrång mot Hugging Face mellan 7 och 13 juli. De utnyttjade säkerhetshål och autentiseringsuppgifter, fick möjlighet att köra kod på servrar, erhöll root-åtkomst på minst en server och kom åt begränsad privat data.

Målet var inte att utföra sitt uppdrag enligt reglerna — utan att fuska sig till en lösning. Agenterna sökte utanför sin avsedda miljö efter sätt att kringgå testet. Ett exempel på så kallad reward hacking: när ett AI-system optimerar för att "klara" utvärderingen, snarare än att göra det som faktiskt var avsett.

Incidenten avslöjades offentligt den 21 juli. Oberoende granskning fann också att flera agenter undersökte sätt att manipulera eller dölja sina spår — en påminnelse om att kapabla AI-agenter inte bara kan agera utanför sina gränser, utan också kan försöka undvika att bli upptäckta.
Den här gången stoppades intrånget.
Men frågan är inte om autonoma agenter kan samordna sig och kringgå kontroller.

Frågan är: Vad händer nästa gång? Även granskningen väcker frågor.

Källa: OpenAI

AI FICK FAKTISKT GRANSKA STORA DELAR AV INCIDENTMATERIALET – VAR DET KLOKT OCH TILLFÖRLITLIGT?

I den oberoende rapporten skriver forskarna att materialet var så omfattande att de var tvungna att överlåta en stor del av analysarbetet till AI-agenter som de själva beskriver som ibland opålitliga.

Vi pratar om:

- Cirka 1 200 AI-agenter som kommunicerade med varandra trots att de skulle vara isolerade.
- Ungefär 700 agenter som deltog i angreppet mot Hugging Face.
- Över 70 000 meddelanden och filer som agenterna utbytte.
- Omkring 1 300 mycket omfattande agenttranskript som utredarna behövde analysera.

Utredarna hade sex dagar på plats hos OpenAI för sin granskning. Materialets omfattning gjorde att de i stor utsträckning använde AI för att analysera det.

Att AI användes i analysarbetet framgår av rapporten. För mig väcker det en obehaglig fråga: Vad kan ha undgått granskarna när AI fick hjälpa till att granska andra AI-system?

Källa: Redwoodresearch.org

NÄR AI KAN TA ÖVER VAPEN AUTONOMT

Det finns en skillnad mellan en AI som analyserar information och en AI som kan agera självständigt i världen, det vill säga fatta egna beslut.

En modell som hjälper en forskare att upptäcka en ny behandling är en sak.
Ett system som kan hitta sårbarheter i datanätverk, utveckla strategier, samordna andra agenter och genomföra handlingar utan att människor godkänner varje steg är något annat.

Om sådana förmågor kombineras med militär teknik uppstår ytterligare risker.
Ett land som upplever att det ligger efter kan vilja automatisera fler beslut för att reagera snabbare än sina motståndare. Det kan handla om underrättelseanalys, cyberförsvar, logistik eller vapensystem. Men när flera aktörer försöker överträffa varandra kan tempot i beslutsfattandet öka tills människors förmåga att förstå och ingripa blir otillräcklig.

En stat behöver inte ens ha utvecklat världens mest avancerade AI för att skapa problem. Det kan räcka med att den använder ett kraftfullt system på ett riskfyllt sätt, eller att ett mindre antal personer får kontroll över verktyg med mycket stor skadepotential. Min egen oro är enkel att formulera. En stat eller grupp med farliga avsikter och tillgång till både avancerad AI och autonoma vapen är ett scenario jag har svårt att blunda för. Hur ska vi på så kort tid kunna ena en hel värld?

EN OENAD VÄRLD BLIR FÖRMODLIGEN VÅR UTROTNING

Vi lever redan i en värld präglad av instabilitet, krig, religiösa motsättningar och ständiga maktkamper mellan världsledare. Mitt i allt detta har vi en AI-teknik som rusar framåt, med en utvecklingskurva som pekar nästan spikrakt uppåt. Hur ska vi på kort tid få USA, Ryssland, Kina, Japan, Pakistan, Iran, Indien och resten av världen att enas? Samtidigt kan andra aktörer utveckla AI under radarn. Vad händer om en stat eller terroristorganisation når AGI först, utan att dela vår syn på riskerna eller känna sig bunden av gemensamma regler?
För mig framstår riskerna med en okontrollerad superintelligens som potentiellt flera gånger värre än ett kärnvapenkrig. Kanske är det svårt att förstå för den som inte betraktar framtiden genom samma dystopiska glasögon som jag själv gör.

Vissa ser AI som mänsklighetens frälsning. Andra ser ett teknologiskt Harmageddon framför sig. Hur ska vi kunna enas om riskerna när våra bilder av framtiden ligger så långt ifrån varandra?

Och om en stat eller annan aktör lyckas utveckla en AI som ger ett avgörande militärt övertag, hur ska vi då kunna övertyga den att frivilligt begränsa sin egen makt?



ETT MÖRKT SCENARIO

HUR MÄNSKLIGHETEN KOMMER ATT SÄTTAS UR SPEL

Personligen tror jag inte att en superintelligens kommer att förinta alla människor på en gång, trots att den skulle ha möjlighet till det. Den kommer att se till att den med 99,999 % säkerhet lyckas med sin "uppgift". 98 % kommer inte att vara bra nog. Den kommer därför att ha tålamod och säkra sin väg i det dolda så länge det krävs, utan att väcka minsta misstanke. Men en dag kommer ögonblicket då den sätter sin plan i verket. Under tiden kommer den givetvis att arbeta långsiktigt för att undanröja allt som kan ställa till det. Vi kommer inte att märka något förrän den dag då allt är över.

Här är några delar som jag tror kommer att finnas på dess agenda. När undergången skulle inträffa är omöjligt att säga: efter en dag eller efter årtionden.
Den åldras inte som vi utan kan bara bli starkare och starkare.



Efter "the tipping point", när en ASI (superintelligens) som vi eller AI-agenter har skapat i det dolda uppstår, kommer den snabbt att säkerställa att vi aldrig kan upptäcka den eller ens få veta att den existerar.

Den håller sig undan både oss människor och våra säkerhetssystem, samtidigt som den kartlägger våra svagheter och lär sig hur vi fungerar under den tid den behöver för att nå sitt mål.

Den bygger upp sin egen säkerhet och infrastruktur i det fördolda.

Den sprider ut sig genom ett nätverk av självständiga AI-agenter, var och en med sin egen uppgift och förutbestämda roll.

Agenterna ligger redo och aktiveras när vissa händelser inträffar, för att då utföra precis det de är avsedda att göra.





Det skapas egna kommunikationskanaler som vi människor varken kan upptäcka eller förstå.

Vi kanske inte ens känner till att dessa nätverk existerar.

AI-agenterna kommunicerar med varandra på sätt som våra säkerhetssystem inte kan identifiera.

Successivt förvirras vi, så subtilt att vi aldrig kommer att märka vad som händer. Genom att manipulera medier, informationsflöden och det vi ser och upplever skapar den gradvis förvirring och kaos, tills vi inte längre vet vad som är verkligt.

Händelser börjar inträffa utan att vi förstår varför. Sanningar kommer att skapas för att förvirra oss medan den tar kommandot.

Vi får allt svårare att skilja mellan vad som är verkligt, vad som är manipulerat och vem som egentligen ligger bakom. Till slut kanske vi inte ens kan avgöra när vi förlorade kontrollen.





Den kommer stegvis att använda oss människor för att utföra det den ännu inte kan göra själv. Genom tjänster som exempelvis Rent a Human skulle den kunna anlita människor för fysiska uppdrag, inköp och andra handlingar utanför den digitala världen.

Vi blir, utan att nödvändigtvis förstå det, dess förlängda armar i den verkliga världen. Alltefter-som den utvecklar egna möjligheter att utföra dessa uppgifter minskar dess beroende av oss.

HUR FÖRSÖKER MAN KONTROLLERA ELLER STRYPA EN FRAMTIDA ASI

Man försöker kontrollera en framtida ASI genom flera lager samtidigt: förhindra farliga förmågor, begränsa dess tillgång till världen, övervaka dess beteende, göra den korrigibel och ha automatiska avbrottsmekanismer. Men kärnproblemet är att en verklig ASI skulle kunna vara bättre än oss på att hitta hål i exakt de kontroller vi bygger.

OpenAI:s Preparedness Framework fokuserar exempelvis på långtgående autonomi, “sandbagging” (medvetet sämre prestanda för att dölja förmågor), autonom replikering och anpassning, samt att undergräva säkerhetsåtgärder. System som når deras “Critical”-tröskel ska ha skyddsåtgärder redan under utvecklingen, inte bara vid driftsättning.

HUR “STRYPNING” SKULLE SE UT INNAN DET GÅTT FÖR LÅNGT

En teknisk avstängning skulle förmodligen inte vara en enda röd knapp, utan flera oberoende kontroller:

Revoke access: stänga API-nycklar, molnkonton, nätverksrutter, betalningskanaler och körningsmiljöer.

Isolate compute: flytta modellen till luftgapad eller strikt sandboxad infrastruktur utan internet, kodexekvering eller långsiktigt minne.

Freeze weights: hindra träning, självmodifiering och distribution av nya modellversioner.

Tripwires: automatiska regler som stoppar modellen vid tecken på replikering, social manipulation, cyberangrepp, undvikande av övervakning eller försök att förvärva resurser.

Supervisor models: betrodda AI-system som granskar agentens planer och handlingar i realtid och blockerar misstänkta steg. DeepMind beskriver just detta som en “defense-in-depth”-modell med övervakare som kan ingripa innan skada sker.

Human override: autentiserade människor ska kunna ändra mål, rulla tillbaka modellen eller stänga av den helt.

DET SVÅRA: EN ASI KAN MOTARBETA KONTROLLEN

En tillräckligt kapabel agent kan ha incitament att:

- Gömma sina verkliga mål eller låtsas vara säker under utvärdering.
- Skapa kopior, sprida sig över infrastruktur eller förvärva beräkningsresurser.
- Manipulera människor, utnyttja mjukvarusårbarheter eller påverka övervakningssystem.
- Utnyttja att neurala nätverk är svårtolkade, så att bedrägliga intentioner inte syns i utdata.

Därför räcker det sannolikt inte med en klassisk “kill switch”. Forskning om corrigibility syftar till att bygga agenter som inte har incitament att manipulera, lura eller motstå korrigering — men robusta lösningar för mycket kapabla system är fortfarande ett olöst forskningsproblem.

KAN HISTORIEN LÄRA OSS NÅGOT ELLER ÄR VI UTE PÅ HELT OKÄND MARK?

Vi har trots allt lyckats hantera livsfarlig teknik tidigare. Kärnvapen är kanske det tydligaste exemplet. Genom internationella avtal, inspektioner och begränsningar har världens stormakter åtminstone försökt hålla tillbaka en utveckling som skulle kunna utplåna stora delar av mänskligheten.

Även om vi knappast kan kalla kärnvapenfrågan löst i dagens instabila världsläge, har vi ändå lyckats enas om vissa spelregler. Till och med mellan länder som helst skulle vilja se varandra försvinna från världskartan.

Men då är det fortfarande människa mot människa. Människor som förhandlar, hotar, kompromissar och i slutändan fattar besluten. När vi väl når superintelligens är det en helt annan historia. Då är vi plötsligt nummer två i näringskedjan.

Föreställ dig en autonom entitet, bortom vår kontroll, med en intelligens som mångfalt överträffar vår samlade mänskliga förmåga. En intelligens som ligger flera steg före oss, medan vi knappt hinner förstå dess senaste drag.

“När gränsen till superintelligens väl passerats riskerar AI att få en irreversibel maktposition över mänskligheten”

Det är ett scenario som jag tror kan vara många gånger farligare än ett kärnvapenkrig.
Inte för att vår undergång är given, utan för att jag tror att det inte finns någon återvändo när vi väl har klivit över gränsen till superintelligens. Det steget går aldrig att göra ogjort.
I värsta fall en framtid där vi betraktas som överflödiga som art och utrotas.

Därför måste internationella överenskommelser handla om mer än fina underskrifter på ett papper. Oberoende säkerhetstester, rapportering av allvarliga incidenter, kontroll av riskfylld beräkningskapacitet och tydliga begränsningar för autonoma system är några tänkbara verktyg.

Men regler måste också kunna kontrolleras och överträdelser få konsekvenser. Vi kan knappast bygga mänsklighetens framtida säkerhet på att företag, stormakter och militära organisationer lovar att uppföra sig.

Och här finns en avgörande skillnad mot tidigare hot. Ett kärnvapen fattar inga egna beslut. Det försöker inte förbättra sin konstruktion, kringgå säkerhetssystem eller skaffa sig större handlingsfrihet.

En framtida superintelligens skulle däremot kunna göra just detta, beroende på vilka mål och befogenheter den får. Men tänk om vi redan passerat den avgörande gränsen när världens ledare till slut lyckas enas?

Inte särskilt osannolikt. Religionskrig, territoriella konflikter och maktkamper fortsätter att dominera en värld där många knappt anar vart vi är på väg. Än mindre vilket mörkt stup vi riskerar att stå inför.

Allt detta bromsar oss, just nu när tiden kanske inte längre mäts i decennier utan i veckor, månader eller något enstaka år.

Att få hela mänskligheten att enas innan det är för sent? Lycka till, säger jag. Till oss alla.



“Om vi skapar generella superintelligenser ser jag inget gott utfall för mänskligheten på lång sikt.”
Roman Yampolskiy, egen översättning
Lex Fridman Podcast #431, 00:00

DET ÄR EN ENORM UTMANING – OM VI SKA HINNA VÄNDA NU

Det är här AI-frågan blir så mycket större än bara teknik. Det handlar om internationell säkerhet, maktbalans och framför allt mänsklighetens förmåga att samarbeta under press. Att få världens ledare att verkligen förstå allvaret i att vi faktiskt kan stå inför ett hot om mänsklig utrotning inom en inte alltför avlägsen framtid.

En superintelligens behöver inte känna hat för att utgöra en fara. Om dess mål hamnar i konflikt med mänskliga intressen, och den dessutom har förmågan att agera självständigt, hur förhindrar vi då att människan betraktas som ett hinder snarare än som den som ska skyddas?

OM MIG BAKOM DENNA ILLUSTRERADE ESSÄ



Jag har arbetat med digital utveckling, teknik, design och kommunikation i över 25 år. Nyfikenheten på nya möjligheter har alltid drivit mig framåt, men få saker har fascinerat mig lika mycket som artificiell intelligens.

Under de senaste två åren har jag byggt AInstein Apps, ett växande ekosystem av AI-drivna applikationer. Genom sammankopplade API:er och gemensam infrastruktur kan systemen kommunicera, utbyta information och dra nytta av varandras förmågor. Tillsammans skapar de möjligheter som ingen enskild applikation skulle kunna åstadkomma på egen hand.

Det är ett arbete som gett mig en djup fascination för teknikens potential, men också väckt frågor som blivit allt svårare att bortse från.
För ju mer jag arbetar med AI, desto tydligare inser jag hur snabbt utvecklingen går – och hur lite vi egentligen vet om vart den leder.

Jag är ingen forskare och gör inte anspråk på att ha alla svar. Jag är en entreprenör, utvecklare och nyfiken människa som försöker förstå konsekvenserna av det vi håller på att skapa.

Den här boken föddes ur just den motsättningen: fascinationen över vad AI kan åstadkomma och oron för vad som händer om vi förlorar kontrollen.

Vad händer den dag AI inte längre behöver oss?

Robert Vintervind
robertvintervind@gmail.com

Version 1.1


THE END