

Amodeis plan för att styra takten i AI-utvecklingen

Amodeis trestegsplan är inte en paus och inte heller en strikt tidsföljd. Stegen kan överlappa, en del är lättare att genomföra än andra, och flera kräver att stater går in och löser konkurrensrättsliga knutar. Poängen är att kapacitetsökningen ska gå i en takt som gör det möjligt att hinna med säkerhetsarbetet – och att någon utomstående ska kunna se att det faktiskt sker.

Det första steget är inbäddade oberoende granskare. Varje labb som tränar de mest kapabla modellerna ska släppa in ett team, till exempel från METR, på samma villkor som de egna medarbetarna: skrivbord på kontoret, passerkort, företagslaptop. Granskarna ska kunna gå in i träningskedjan, inte bara titta på den färdiga modellen, kontrollera att utlovade säkerhetsrutiner följs, rapportera incidenter och bedöma alignment under pågående arbete. De ska dessutom ha rätt att publicera sina iakttagelser utan att labbet redigerar texten, med bara smala undantag för sekretess och säkerhet. Amodei jämför med den amerikanska banktillsynen, där examinatore sitter inne på de stora instituten och har löpande tillgång till system och personal. Tanken är verifierbarhet på detaljnivå, offentlig transparens och en second opinion som inte styrs av kommersiella incitament. Anthropic har redan åtagit sig det här unilateralt.

Analogin är inte oproblematis. Banktillsynen kan tvinga fram åtgärder och i yttersta fall stänga en verksamhet. De granskare Amodei talar om kan undersöka och skriva, men de får ingen motsvarande makt att stoppa en träning eller ett släpp. Julie Andersen Hill, som forskar om bankreglering, har sagt att hon inte ser vad de egentligen gör om de saknar den sortens befogenhet. Det är en invändning man måste ta på allvar. Samtidigt är det just den löpande, inifrån kommande granskningen som saknats när incidenter senare visat sig bero på brutna träningsmiljöer eller ofullständig filtrering. Utan någon som sitter inne och ser kedjan blir varje löfte om pacing svårt att skilja från ett pressmeddelande.

Det andra steget är koordinering mellan demokratiska länder och de företag som verkar där. Det kan ta formen av frivilliga branschstandarder, statligt understödda överenskommelser och kapacitetsanknutna checkpoints: när en modell når en viss förmåga ska den inte släppas förrän vissa alignmentegenskaper är dokumenterade och granskade. Demis Hassabis har skisserat en branschorganisation med någon form av statlig anknytning, och Amodei pekar på samma håll. Konkurrensrätten gör det juridiskt känsligt; utan ett smalt undantag från Washington riskerar säkerhetssamtal att tolkas som otillåten samordning. Redan existerande exempel är åtagandena från AI-toppmötet i Seoul 2024, där sexton företag lovade att definiera trösklar där riskerna blir oacceptabla, och Anthropics egen Responsible Scaling Policy. Tanken

är inte att frysa utvecklingen utan att koppla nästa steg till bevis, inte bara till intern försäkran.

Det tredje steget är global koordinering även med auktoritära stater. Här är Amodei mer försiktig. Han räknar upp fyra nivåer, från det mest realistiska till det minst sannolika. Första nivån är ett förbud mot smala, uppenbart farliga användningar – framför allt AI som hjälpmedel för biologiska vapen. Det borde gå att enas om, eftersom det skadar alla inblandade. Andra nivån är ömsesidig testning före släpp av akuta risker inom cyber, biologi och alignment, gärna via ett globalt standardorgan. Svårigheten ligger i verifieringen: hemliga modeller för militärt bruk går inte att kontrollera på samma sätt. Tredje nivån är en fartgräns för rekursiv självförbättring. Amodei liknar den vid SALT-avtalen: man begränsar potentialen för skada utan att ta bort avskräckningen. Att gå från extremt snabbt till bara ganska snabbt ger enligt honom bort förhållandevis lite strategiskt försprång men vinner tid för säkerhetsarbetet. Fjärde nivån är en mer omfattande paus. Den tycker han man ska hålla öppen, men han räknar inte med den i närtid. Incitamenten att bryta ett sådant avtal i hemlighet är för stora, och kraven på verifiering skulle bli enorma.

Hela konstruktionen vilar på att demokratierna inte släpper sitt försprång. Det är den geopolitiska balansgången. Om de egna labben saknar mer än Kina gör, försvinner utrymmet att pacea. Då tvingas man antingen acceptera högre risk eller se ett auktoritärt system ta ledningen med de militära och ekonomiska följder det skulle få. Därför måste den långsammare takten i väst kombineras med åtgärder som håller avståndet: exportkontroll på de mest avancerade chipen och tillverkningsutrustningen, hårdare tag mot destillation och bättre skydd mot stöld av modellvikter. Chipen är den flaskhals som i praktiken avgör hur snabbt Kina kan skala. Destillation – att med hjälp av tusentals konton och miljontals samtal suga ut resonemang och agentförmågor ur västerländska modeller – har redan dokumenterats i stor skala. RAND:s arbete om viktsäkerhet listar dussintals angreppsvägar och fem säkerhetsnivåer; CSET har visat hur kontrollerna läcker i inköpsledet. Amodei har själv hävdat att väl genomförda åtgärder kan vidga det amerikanska försprånget under de närmaste tre till fem åren, precis den period då AI:s geopolitiska betydelse väntas vara som störst.

Analogierna – banktillsynen, SALT – är omstridda, och det ska de vara. Bankgranskare har tvångsmedel som de föreslagna AI-granskarna saknar. Kärnvapenavtalen bygger på ett fysiskt ämne som går att räkna och spåra; AI har ingen motsvarighet. Det är skäl att formulera avtalen smalt och verifierbart, inte att avfärda hela idén. Det som går att enas om först är det som skadar alla: biovapenbruk, uppenbar missbruk av cyberförmågor. Resten får byggas ut om och när förtroendet och kontrollmetoderna räcker.

Fördelarna med tekniken finns kvar. Sjukdomar som i dag tar decennier att angripa, tillväxt som lyfter fler ur fattigdom, mer utrymme för demokratiska institutioner – det är inte tom retorik, men det förutsätter att systemen inte slinker undan kontrollen på vägen. En extra tid på ett eller två år före de mest kritiska förmågorna, använd till operativ disciplin, alignment, tolkningsbarhet och tester som är svårare att lura, skulle enligt Amodei minska risken påtagligt. Det är inte gratis. Det kräver att labben släpper in utomstående, att stater vågar ge smala undantag från konkurrensrätten, och att man samtidigt håller hårt i chip, vikter och destillationsvägar. Det är besvärligt. Men om man menar allvar med att tekniken ska tjäna människor och inte bara dem som råkar sitta på den just nu, är det den sortens eftertanke man är skyldig att försöka.

Begrepp och förkortningar

Alignment. Att en modell gör det man avser att den ska göra, även när den blir mer kapabel och mer självständig. Inte bara att den svarar artigt, utan att den inte kringgår regler, lurar tester eller jagar belöning på sätt som skadar.

Checkpoints. Trösklar kopplade till modellens förmåga. När systemet når en viss nivå ska det inte släppas vidare förrän vissa säkerhets- eller alignmentkrav är uppfyllda.

Destillation. Att ur en stark modell suga ut resonemang, verktygsanvändning och agentbeteende genom massiv användning – ofta via många konton – och sedan träna en egen, billigare modell på det som kommit ut. Ett sätt att korta försprånget utan att ha samma underliggande kapacitet.

Embedded evaluators / inbäddade granskare. Utomstående bedömare som sitter inne på labbet med samma slags tillgång som anställda: lokal, passerkort, dator. De ska kunna följa träning och incidenter löpande, inte bara granska en färdig rapport efteråt.

Fronten / frontier. De mest kapabla modellerna och de labb som tränar dem. Att "hålla takten vid fronten" innebär att sakta kapacitetsökningen så att säkerhetsarbetet hinner med – inte att stanna.

METR. Model Evaluation and Threat Research. Oberoende organisation som utvärderar farliga förmågor och alignment. Den aktör Amodei nämner först som tänkbar inbäddad granskare.

Modellvikter. De tränade parametrarna som är modellen. Den som får vikterna kan i praktiken kopiera systemet. Därför är stöld av vikter ett eget säkerhetsproblem, skilt från vanlig dataintrångsrisk.

Pacing. Medvetet sänkt takt i kapacitetsutvecklingen så att alignment, utvärdering och drift hinner ikapp. Inte ett träningsstopp.

Rekursiv självförbättring (RSI). Att AI-system i växande grad används för att bygga nästa generation AI. Om det tar fart kan utvecklingstakten stiga snabbare än människors förmåga att förstå och styra systemen. I källorna är full RSI oftast en prognos, inte ett konstaterat faktum.

Responsible Scaling Policy (RSP). Företagsinterna regler, särskilt hos Anthropic, som knyter vidare uppskalning till särskilda säkerhetsåtgärder när vissa förmågetrösklar passeras.

SALT. Strategic Arms Limitation Talks. Avtal under kalla kriget som begränsade antalet kärnvapenmissiler utan att ta bort avskräckningen. Amodei använder det som bild för en "fartgräns" för RSI: man ger bort mindre strategiskt än man vinner i säkerhet – om avtalet går att kontrollera.

Second opinion. En bedömning från någon som inte har samma kommersiella intresse som labbet. I Amodeis förslag den roll de inbäddade granskarna ska spela.

CSET. Center for Security and Emerging Technology vid Georgetown. Think tank med empiriskt arbete om bland annat chipkontroll och Kinas tillgång till AI-hårdvara.

CSIS. Center for Strategic and International Studies. Washingtonbaserad tankesmedja; Gregory C. Allen m.fl. har skrivit de mest citerade analyserna av USA:s chipembargo mot Kina.

RAND. Amerikansk forskningsorganisation. Relevant här dels för viktsäkerhet (angreppsvägar och säkerhetsnivåer), dels för jämförelser mellan kärnvapenavtal och tänkbara AI-avtal.

Källförteckning

Amodei, Dario (2024). Machines of Loving Grace: How AI Could Transform the World for the Better. darioamodei.com.

Amodei, Dario (2025a). The Urgency of Interpretability. darioamodei.com, april.

Amodei, Dario (2025b). On DeepSeek and Export Controls. darioamodei.com, januari.

Amodei, Dario (2026a). The Adolescence of Technology. darioamodei.com, januari.



Amodei, Dario (2026b). We Must Pace the Frontier. darioamodei.com, 12 september.

Allen, Gregory C. (2022). Choking off China's Access to the Future of AI. CSIS.

Anderljung, Markus m.fl. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv:2307.03718.

Anthropic (2026a). Detecting and preventing distillation attacks. anthropic.com, februari.

Anthropic (2026b). Detecting and countering misuse of AI: September 2026. Threat Intelligence Report, 10 september.

Anthropic (u.å.). Responsible Scaling Policy. anthropic.com.

Bengio, Yoshua m.fl. (2026). International AI Safety Report 2026. arXiv:2602.21012.

CNBC (2026). Anthropic, OpenAI proposed new 'neutral' AI watchdogs. Why you should worry about the idea. 16 september. Intervju med Julie Andersen Hill.

Congressional Research Service (2021). Bank Supervision by Federal Regulators: Overview and Policy Issues. R46648.

CSET (u.å.). Silicon Twist: Managing the Chinese Military's Access to AI Chips. Georgetown University.

Federal Reserve (u.å.). Understanding Federal Reserve Supervision. federalreserve.gov.

Hassabis, Demis (u.å.). A Framework for Frontier AI and the Dawning of a New Age. demishassabis.substack.com.

Just Security (2026). The Case for Imposing Costs on China's AI Distillation Campaigns. justsecurity.org.

Nevo, Sella m.fl. (2024). Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. RAND, RRA2849-1.

Office of the Comptroller of the Currency (u.å.). Large Bank Supervision. Comptroller's Handbook.

RAND (2023). Nuclear Arms Control Verification and Lessons for AI Treaties. arXiv:2304.04123.

RAND (u.å.). Insights from Nuclear History for AI Governance. PEA3652-1.



UK Government (2024). Frontier AI Safety Commitments, AI Seoul Summit 2024. [gov.uk](https://www.gov.uk), maj.

Kompletterande, mer kritiska eller begreppsliga källor om analogierna:

AI Frontiers (u.å.). Nuclear Non-Proliferation Is the Wrong Framework for AI Governance. ai-frontiers.org.