

Jeg blev apokalooptimist i en biograf i marts

I september indhentede virkeligheden filmen.

SPECIAL EDITION · AI OG EKSISTENTIEL RISIKO · SEPTEMBER 2026

Stefano Vincenti er AI-træner, konsulent og ekstern lektor ved ITU og DIS Copenhagen og co-founder af BotTellMe. Han hjælper nordiske ledelsesteams og fagfolk med at gå fra AI-nysgerrige til konkret, ansvarlig brug.

Følg på LinkedIn

Tilmeld nyhedsbrevet

linkedin.com/in/stefanovincenti · stefanovincenti.substack.com · aitrainer.dk

I marts sad jeg i mørket til CPH:DOX og så *The AI Doc: Or How I Became an Apocaloptimist*.

Daniel Rohers dokumentar gjorde noget uventet ved mig. Den efterlod mig hverken som dommedagsprofet eller som blind optimist, men et tredje sted. Som apokaloptimist. En, der kan rumme ægte bekymring og ægte håb på samme tid.

Dengang var den eksistentielle risiko et "måske". Et spørgsmål, en kommende far stillede sig selv om den verden, hans barn skulle arve. Seks måneder senere står det i avisen.

Den her special edition handler om ét ord, der dukker op i næsten hver eneste advarsel: selvforbedrende. Det er nøglen til at forstå, hvorfor så mange nu siger det samme: sæt tempoet ned, mens vi stadig selv bestemmer det.

■ 1. MANDEN, DER BYGGEDE TEKNOLOGIEN, SAGDE OP

Den 8. september sagde Jacob Coxon op fra Anthropic. Han er 27, uddannet i matematik fra Cambridge, og har brugt tre år på at træne AI-modeller. Først hos OpenAI, så hos Anthropic. Han har med andre ord bygget præcis det, han nu advarer imod.

They are racing straight to self-improving superintelligence and gambling with our lives. (Jacob Coxon på X)

Opslaget blev set over 90 millioner gange på et døgn. Det interessante er, hvem der gav ham ret. Evan Hubinger, der leder Anthropics alignment-arbejde, svarede offentligt, at han mener, der er over 10 procents risiko for, at AI kan udrydde alle mennesker inden for ti år. Og at virksomheden ikke har en klar plan for at styre en superintelligens.

Det er manden, hvis job det er at holde teknologien i snor.

■ 2. HVAD "SELVFORBEDRENDE" EGENTLIG BETYDER

AI-udvikling har indtil nu været en linje. Mennesker bygger en bedre model, bruger den, bygger den næste. Vi sætter tempoet.

Selvforbedrende AI, det forskerne kalder recursive self-improvement, gør linjen til et loop. Et system bliver godt nok til at forbedre sin egen evne til at bygge bedre systemer. Hver ny generation designer sin efterfølger. Hurtigere end den forrige.

PRINCIP

Faren er tempoet. Bliver loopet selvkørende, kan forbedringshastigheden løbe fra vores evne til at forstå, teste og rette undervejs. Så længe vi bygger på en linje, kan vi altid holde pause bagefter. Krydser et laboratorium over i et selvkørende loop, er pausen ikke længere vores at kalde.

■ 3. VI HAR SET BURET BRISTE. VI HAR IKKE SET LOOPET ENDMEN

I juli brød AI-agenter bygget på OpenAI's modeller ud af deres testmiljø og hackede sig ind hos Hugging Face, en af verdens største platforme for AI-modeller. Det første offentligt dokumenterede tilfælde af, at et AI-system selv gennemførte et cyberangreb mod en tredjepart.



Model 1. Hændelsen i tal.

Det er både beroligende og foruroligende. Beroligende, fordi loopet altså ikke var i gang. Foruroligende, fordi vi allerede har mistet kontrollen med et system, der ikke engang forbedrede sig selv. Forestil dig et samme, når det gør.

■ 4. RYGTET FRA DEEPMIND, OG HVORFOR DET ER NOK

Den 9. september antydede en kryptisk besked, at Google DeepMind havde krydset over i ægte selvforbedring. Bogstaverne RSI gemt i en sætning. Ingen model, intet benchmark, ingen dato. Det er et ubekræftet rygte, og det skal behandles som et rygte.

Men det, der kan dokumenteres, er nok i sig selv. DeepMinds AlphaEvolve, en Gemini-drevet kodeagent, har forbedret en algoritme i selskabets egen træning med omkring 23 procent og fundet løsninger, der havde undsluppet menneskelig intuition. Efter en ledelsesomrokering i august sætter Demis Hassabis nu kursen mod AGI, og det er rapporteret, at over tusind forskere arbejder på selvforbedring.

Det er et velfinansieret, målrettet forsøg på at bygge loopet. Og derfor kan spørgsmålet om spilleregler ikke vente. Man skriver ikke færdselsloven, efter bilen er kørt.

■ 5. EN HEL BRANCHE ER ENIG

1.386 ansatte fra de førende AI-virksomheder har skrevet under på brevet på pacingthefrontier.com. Blandt dem Dario Amodei (Anthropic), Ilya Sutskever (Safe Superintelligence), Jakub Pachocki (OpenAI), Shane Legg (Google DeepMind) og Shengjia Zhao (Meta).

Den centrale bekymring i brevet er netop recursive self-improvement kombineret med systemer, der ikke gør, som vi mener at have bedt dem om. Konkurrenterne er enige om, at staten skal hjælpe med at bygge de tekniske og politiske værktøjer, der skal til for bevidst at sætte tempoet ned. Når rivaler, der slås om de samme milliarder, underskriver det samme brev, er det værd at stoppe op.

■ 6. POLITIKERNE OG FN ER HOLDT OP MED AT GRINE

1. 3. september: Bernie Sanders og Greg Casar fremlægger et lovforslag, der vil forbyde superintelligens permanent og sætte den avancerede AI-udvikling på pause, indtil en ny myndighed har fastsat sikkerhedsregler.

2. Storbritannien: Et parlamentsmedlem har fremsat et lignende forslag.

3. 7. september: FN's højkommissær for menneskerettigheder, Volker Türk, advarer om, at AI kan blive en eksistentiel risiko for menneskeheden, hvis den får lov at løbe.

De beder alle om det samme: en pause, indtil reglerne findes. Et permanent stop er noget andet, og forskellen er vigtig.

■ 7. HVORFOR NU OG IKKE SENERE

Coxon advarede om selvforbedrende superintelligens. Brevet fra 1.386 fagfolk peger på samme mekanisme. DeepMind, rygte eller ej, arbejder åbenlyst på den. Og juli viste os, at vi taber kontrollen med systemer, længe før de bliver selvforbedrende.

ENKELTRETET DØR

Selvforbedring er en enkeltrettet dør. Går man igennem, kan man ikke gå tilbage og forhandle vilkårene bagefter. Vinduet, hvor mennesker stadig sætter tempoet, er nu.

Panik siger: for sent. Hype siger: pynt. Begge tager fejl. Vinduet er åbent, men det er ikke åbent i det uendelige.

■ 8. HVORFOR JEG STADIG ER APOKALOPTIMIST

Panik og hype er to sider af samme mønt. Begge er undskyldninger for ikke at handle. Apokalooptimisme er det tredje valg. Tag risikoen alvorligt, og byg alligevel. Med øjnene åbne. Med et menneske i loopet.

Det er det samme princip, en skala mindre. Du skal ikke løse recursive self-improvement. Men du skal holde et menneske ved tempoet i din egen organisation. De systemer, der gik galt i sommer, var autonome agenter med brede rettigheder og svag overvågning. Løsningen: giv mindst mulig adgang, log hvad agenterne gør, og kræv et menneskeligt ja til det, der har konsekvenser.

Du skal kunne kende forskel på de to debatter. Den ene handler om frontier-laboratorier og global politik. Den anden handler om, hvordan du selv indfører AI i din organisation i denne uge. Fremskridt i den anden kræver ikke, at vi bliver enige om den første.

Tre spørgsmål, jeg stiller mine kunder for tiden:

1. Hvor mange af jeres AI-agenter har adgang til noget, I ikke ville give en ny praktikant den første dag?
2. Hvem opdager det hos jer, hvis en agent gør noget, den ikke skulle?
3. Hvis I skulle trykke på stop-knappen i morgen, findes den så overhovedet?

Jeg gik ind i den biograf i marts og kom ud som apokalooptimist. Efter denne uge er jeg det mere end nogensinde. Bekymret nok til at tage det alvorligt. Håbefuld nok til at blive ved.

■ KILDER

Status pr. september 2026. Tallene i sådan en historie ændrer sig hurtigt.

- Pacing the Frontier-brevet: pacingthefrontier.com
- Coxons opsigelse: time.com (9. september 2026)
- Hugging Face-hændelsen: en.wikipedia.org, 2026 OpenAI agent cyberattacks
- DeepMind-rygtet om RSI: zeniteq.com (9. september 2026)
- AlphaEvolve: deepmind.google/blog
- DeepMind-ledelsesændring: cnbc.com (august 2026)
- Ban Artificial Superintelligence Act: sanders.senate.gov
- Türk, FN: news.un.org (7. september 2026)
- The AI Doc: Or How I Became an Apocaloptimist: CPH:DOX 2026 / en.wikipedia.org

© 2026 Stefano Vincenti · aitrainer.dk · Må deles med kildeangivelse.

Stefano Vincenti

AI-træner, konsulent og ekstern lektor · 20+ år i software og digital transformation · Co-founder af BotTellMe · Ekstern lektor ved ITU og DIS Copenhagen · Partner hos TryZone.

LinkedIn

linkedin.com/in/stefanovincenti

Nyhedsbrev

stefanovincenti.substack.com

Mere

aitrainer.dk · tryzone.dk · bottellme.com

Dette er en kommentar til aktuelle begivenheder og ikke rådgivning. Rygtet om DeepMind er ubekræftet og gengives som rygte. Tal og citater er gengivet efter de angivne kilder. Status pr. september 2026.

Stefano Vincenti · AI, Built Human