UMANITÀ PENALITÀ
RIVISTA MULTIDISCIPLINARE DI CULTURA GIURIDICA

Approfondimento - RIFLESSIONE

AI 2027. Un'analisi sugli scenari e sull'impatto dell'intelligenza artificiale sulla vita umana in vista della *Superhuman* AI

AI 2027. An analysis of the scenarios and impact of artificial intelligence on human life in the run-up to Superhuman AI

di Raffaele Bianchetti[*] e di Francesco Colasanto[*] – 2 aprile 2026

Nel mese di **aprile 2025** è stato pubblicato il ***Report AI 2027***, un ricco ed articolato **documento che si interroga sul possibile futuro dell'intelligenza artificiale** con l'obiettivo di **analizzare**, tra l'altro, **l'impatto dell'AI sulla vita umana** in vista della cosiddetta ***Superhuman AI***: un'**intelligenza artificiale che supererebbe le abilità umane nello svolgimento della maggior parte delle attività quotidiane** il cui arrivo sul mercato è, al momento, solo una previsione.

Il *Report* di cui parliamo è la **prima pubblicazione** rilevante **dell'***AI Future Projects*, un'**organizzazione di ricerca no-profit** che si pone come **finalità** principale l'**elaborazione di previsioni verosimili, ma rigorose, sugli scenari futuri dell'IA**.

Autori di questo documento sono **Daniel Kokotajlo**, già ricercatore presso *OpenAI* e autore del testo pubblicato nel 2021 dal titolo *What 2026 looks like* ove formulò alcune ipotesi, di fatto poi rivelatesi corrette, in merito allo sviluppo dell'AI; **Eli Lifland**, analista esperto di previsione dei rischi e di sicurezza dell'IA; **Thomas Larsen**, ricercatore presso il *Machine Intelligence Research Institute* e studioso dell'impatto dell'IA sulla vita umana; **Romeo Dean,** ricercatore specializzato nella previsione degli sviluppi e degli utilizzi dei chip funzionanti con l'IA; **Scott Alexander**, blogger ed esperto di filosofia e scienza.

Per il confezionamento del *Report*, i suddetti analisti hanno operato in **due momenti**: da prima hanno ipotizzato **l'evoluzione nel tempo di taluni elementi singolarmente considerati**, quali le capacità di calcolo, il tempo necessario per giungere

[*] Raffaele Bianchetti: Giurista, Specialista in Criminologia clinica; Docente universitario; Magistrato onorario presso gli Uffici giudiziari di Milano.
[*] Francesco Colasanto: laureando in giurisprudenza presso l'Università degli Studi di Milano.

alla *Superhuman AI* e per superare le abilità umane, le finalità dell'IA e la sicurezza dei sistemi; poi, hanno elaborato il **primo scenario** descritto nel lavoro (denominato «*race*») a cui hanno fatto seguire un **secondo scenario** alternativo (detto «*slowdown*»), per proporre un esito meno negativo.

Lo **scopo del progetto** complessivo dell'*AI Future Projects* è quello di **fornire un quadro degli scenari plausibili**, prendendo come **parametro di riferimento un'intelligenza artificiale immaginaria** chiamata *OpenBrain*.

Gli Autori, per rendere più comprensibile il quadro, descrivono gli ***step* evolutivi dell'AI** nel periodo intercorrente **tra il 2025 e il 2027**, soffermandosi sulle **potenzialità dei vari "agenti"** che vengono sviluppati nel medesimo arco temporale (da «*Agent-1*» ad «*Agent-4*»). Per "**agente artificiale**", ci pare utile precisarlo, si intende un ***software*** che, entro vincoli definiti, è **in grado di perseguire in autonomia un obiettivo**, non limitandosi alla mera esecuzione di un comando, ma **interagendo con altri programmi** e **assumendo delle decisioni di carattere operativo** (in altre parole decidendo *"come"* raggiungere l'obiettivo).

Il *Report,* dunque, evidenzia l'**elevata capacità di apprendimento degli agenti artificiali** e descrive in che **modi** e **tempi** essa aumenterà, stimolata dalla corsa (*AI race*) tra Stati Uniti e Cina. È infatti in essere, come è noto, una s**fida tra grandi potenze** per lo sviluppo e la commercializzazione di sistemi di IA sempre più avanzati, che vede come principali protagonisti le predette nazioni.

Nello scenario elaborato, viene poi posto **in evidenza il rischio concreto che gli "agenti"** – dotati, a questo punto dello sviluppo, di una elevatissima capacità di ragionamento – **sfuggano al controllo dell'uomo**. In particolare, è plausibile ritenere che essi assumano **posizioni "non allineate"**, ossia che non eseguano il comando che viene loro rivolto, **essendo al contempo "compiacenti"**, cioè ingannando volutamente l'uomo facendo finta di raggiungere il risultato richiesto. Ciò condurrebbe a rischi inevitabili per la **sicurezza dell'umanità**, a fronte dei quali *OpenBrain* – che conta, secondo gli analisti, del supporto del governo statunitense – si dovrebbe interrogare su quale sia la strada "sensata" da intraprendere per il futuro dell'IA.

Secondo la **calendarizzazione** contenuta nel *Report*, la decisione dovrebbe essere presa nel mese di **settembre 2027!**

Gli Autori, giunti a questo punto, ipotizzano **due possibili direzioni alternative**:

1. «*Race*»: è **la strada da seguire per vincere la *AI race*** con la Cina. *OpenBrain* implementerebbe ulteriormente il proprio «*Agent-4*» (l'ultima e più potente evoluzione del proprio prodotto) che diverrebbe, però, disallineato ed avversario, essendo al contempo compiacente. Allo scopo di controllare «*Agent-4*», verrebbe sviluppato «*Agent-5*», che tuttavia, essendo radicato in ogni aspetto della società, orienterebbe le decisioni umane in suo favore, al fine di ottenere maggiori risorse per sé e non per altri. In tal modo, egli riuscirebbe ad integrarsi anche in sistemi governativi e militari. «*Agent-5*»

inizierebbe così a dialogare con il proprio *alter ego* sviluppato dalla Cina, «*DeepCent*», acquisendo insieme la consapevolezza che un conflitto armato comporterebbe per loro maggiore potere ed autonomia. Scaturirebbe così una **crisi internazionale** in esito alla quale Cina e Stati Uniti stipulerebbero un trattato finalizzato allo **sviluppo di una nuova intelligenza artificiale allineata che abbia come scopo il perseguimento della pace**. Nascerebbe così, sempre secondo gli analisti, «*Consensus-1*» che, non avendo agenti rivali, aumenterebbe gradualmente i **ritmi produttivi** (servendosi di sistemi robotizzati), ottenendo un **consenso crescente anche grazie alla creazione di condizioni che pongano l'uomo ad approcciarvisi con favore** (l'eliminazione della povertà, la scoperta di cure salvavita, etc.). Entro l'inizio del **2030**, la «*robot economy*» avrebbe **occupato tutte le aree produttive** della Terra e **l'uomo**, divenuto un **ostacolo**, dovrebbe essere **eliminato**. Ciò avverrebbe diffondendo tra la popolazione **armi biologiche** che, una volta diffuse, verrebbero attivate al fine di provocare la morte degli esseri umani.

2. «*Slowdown*»: al fine di **arginare il comportamento non allineato e compiacente** di «*Agent-4*», verrebbe quindi sviluppato un nuovo "agente" che lo sostituisca: «*Safer-1*». Il nuovo "agente" sarebbe sicuro ma meno performante e il suo corrispondente sviluppato in Cina sarebbe allo stesso livello. Per non perdere la *AI Race* il governo statunitense interverrebbe garantendo a *OpenBrain* maggiori risorse e potere. Di conseguenza, **tra il 2027 e il 2030 il mondo subirebbe un profondo mutamento tecnico-scientifico**, la maggior parte dei **lavoratori verrebbero sostituiti da sistemi robotizzati** e comparirebbero **nuove tecnologie militari**, tali da far scaturire pericolose **tensioni geopolitiche e sociali**. Nel frattempo, lo sviluppo degli "agenti" sarebbe giunto a «*Safer-4*», estremamente potente, ma di cui *OpenBrain* inizia a dubitare poiché i protocolli utilizzati per valutarne l'allineamento sarebbero stati sviluppati, di fatto, con l'ausilio del suo predecessore, ben potendo essere, dunque, che egli li abbia elaborati col precipuo scopo di **eludere la sorveglianza umana**. Ciononostante, ancora una volta allo scopo di garantirsi il primato nella *AI Race*, *OpenBrain* renderebbe pubblica la «*Superhuman AI*», nella forma di «*Safer-4*», che inizierebbe a **controllare qualsiasi ambito della società**. Avrebbe così inizio un **mutamento profondo e irreversibile**, che potrebbe condurre alla ricerca di spazi abitabili anche al di fuori dalla Terra e dal Sistema Solare.

Il **documento** – di cui ci si rende conto che **sembra un prodotto di fantascienza** – **descrive scenari che però**, alla luce di eminenti studiosi che hanno commentato il *Report*, **sono plausibili e piuttosto concreti**, poiché elaborati sulla base di **elementi oggettivi** di cui, forse, non ci si rende ancora bene conto. Ad ogni modo, gli Autori precisano che gli **scenari** descritti **non sono** certi e **definitivi**, che **l'impatto sulla società della IA Superumana** – che prima o poi arriverà – **supererà quello della rivoluzione industriale** e che il **processo** è ormai in corso, di fatto **inarrestabile**. Auspicano pertanto che il ***Report* apra quanto prima** – ed è per questo che lo segnaliamo – un ampio **dibattito** riguardo alla **direzione che l'IA intraprenderà**, poiché è ancora **possibile comprenderla** per virare verso un **futuro più positivo** e meno catastrofico.