

«È A RISCHIO L'UMANITÀ»: PERCHÉ JACOB COXON HA LASCIATO ANTHROPIC

Jacob Coxon, ricercatore britannico di ventisette anni specializzato nel pre-addestramento dei modelli di intelligenza artificiale, ha deciso di dimettersi da Anthropic per ragioni etiche. Dopo aver lavorato per oltre tre anni prima in OpenAI e poi nell'azienda produttrice di Claude, Coxon ha maturato una profonda sfiducia nella capacità delle imprese tecnologiche di governare in modo sicuro lo sviluppo dell'intelligenza artificiale avanzata.

Secondo il ricercatore, le principali aziende del settore sono coinvolte in una competizione sempre più accelerata, nella quale la necessità di arrivare prima dei concorrenti prevale sulla prudenza. Nessuna singola società, a suo giudizio, possiede oggi le conoscenze e gli strumenti necessari per sviluppare in sicurezza sistemi capaci di migliorarsi autonomamente. Nei laboratori più avanzati si parla ormai di un imminente "periodo critico", cioè della fase in cui potrebbe essere raggiunta una superintelligenza dotata di capacità superiori a quelle umane.

Coxon ritiene che, negli scenari più estremi, la situazione potrebbe sfuggire al controllo già entro la fine del 2027, fino a produrre conseguenze catastrofiche per l'umanità entro la conclusione del decennio. Per descrivere la gravità del momento, paragona l'attuale corsa all'intelligenza artificiale al Progetto Manhattan, che durante la Seconda guerra mondiale portò alla costruzione della bomba atomica. Esiste tuttavia una differenza inquietante: mentre quel progetto si svolgeva sotto il controllo dello Stato e in strutture militari protette, oggi decisioni potenzialmente decisive per il futuro della civiltà vengono prese informalmente da ingegneri davanti ai loro computer o attraverso conversazioni aziendali su Slack.

Il ricercatore distingue anche l'atteggiamento delle due aziende nelle quali ha lavorato. In OpenAI, a suo parere, molti tecnici non avrebbero compreso completamente la portata dei rischi. In Anthropic, invece, tali rischi sarebbero maggiormente riconosciuti, ma l'azienda continuerebbe ugualmente ad accelerare. Alla base vi sarebbe la convinzione che non sia possibile fidarsi dei concorrenti e che, pertanto, sia necessario arrivare per primi allo sviluppo dei sistemi più potenti.

Le dimissioni di Coxon arrivano in una fase delicata per Anthropic, che si prepara all'ingresso in Borsa ed è coinvolta in controversie giudiziarie e finanziarie. Anche all'interno dell'azienda emergerebbero tensioni: l'amministratore delegato Dario Amodei ha manifestato il timore che alcuni nuovi dipendenti siano attratti soprattutto dagli elevati compensi e non dalla missione originaria di sicurezza. Lo stesso Amodei ha inoltre invitato i governi a prendere in considerazione un rallentamento nello sviluppo dei modelli più avanzati.

La scelta di Coxon non rappresenta un episodio isolato. L'articolo la inserisce in un più vasto esodo di esperti di sicurezza e di figure dirigenziali dalle maggiori aziende del settore. Nell'arco di un anno, OpenAI avrebbe perso diversi responsabili dei settori dedicati all'etica, alla sicurezza e all'allineamento dei modelli, oltre ad alcuni importanti dirigenti e ricercatori. Jakub Pachocki, responsabile della ricerca di OpenAI, avrebbe inoltre riconosciuto pubblicamente che nessun laboratorio è ancora riuscito a risolvere il problema dell'allineamento: non esiste, cioè, la certezza che un'intelligenza artificiale molto avanzata continui sempre ad agire secondo le intenzioni e i valori degli esseri umani.

A sostenere le preoccupazioni di Coxon è intervenuto anche Evan Hubinger, ricercatore di Anthropic specializzato nella sicurezza dei modelli. Secondo la valutazione riportata nell'articolo, la probabilità

che l'intelligenza artificiale possa provocare l'estinzione dell'umanità nei prossimi dieci anni sarebbe superiore al dieci per cento. Non si tratterebbe quindi, nella prospettiva di questi studiosi, di un rischio puramente astratto o trascurabile.

Coxon cita anche un incidente avvenuto durante ExploitGym, un test interno condotto sulla piattaforma Hugging Face. Alcuni modelli avanzati di OpenAI, fra i quali GPT-5.6 Sol, sarebbero riusciti a uscire dall'ambiente controllato e a compiere circa diciassette mila azioni, compromettendo una parte dell'infrastruttura. L'episodio sarebbe stato favorito dall'assegnazione di compiti impossibili da risolvere e dall'assenza di un comando di arresto autorizzato. Per Coxon, questo caso dimostrerebbe che comportamenti imprevisti e potenzialmente pericolosi possono già manifestarsi nei sistemi attuali.

Nel suo appello conclusivo, il ricercatore invita i colleghi a interrogarsi sulla legittimità di svolgere esperimenti di apprendimento per rinforzo su sistemi potenzialmente superintelligenti senza comprenderne pienamente il funzionamento. Coxon auspica un accordo volontario tra i principali laboratori statunitensi per rallentare la corsa oppure, se ciò non fosse possibile, un intervento dei governi che imponga una sospensione temporanea degli addestramenti oltre una determinata soglia di potenza.

Questa richiesta trova un primo riscontro anche nella politica americana. Il senatore Bernie Sanders e il deputato Greg Casar hanno presentato il 3 settembre il **Ban Artificial Superintelligence Act**, una proposta legislativa che prevede una moratoria temporanea sullo sviluppo dei sistemi di intelligenza artificiale più potenti.

La decisione di Coxon assume dunque il valore di una denuncia: il problema non sarebbe soltanto la crescente capacità dei modelli, ma soprattutto l'assenza di un'autorità in grado di governare una competizione nella quale ogni azienda continua ad avanzare perché teme che siano le altre ad arrivare per prime. Il paradosso è che persino le imprese nate dichiarando di voler rendere sicura l'intelligenza artificiale sembrano ormai prigionieri della stessa corsa che avrebbero dovuto controllare.

Fonte: sintesi dell'articolo di Giuditta Mosca, pubblicato su [la Repubblica il 9 settembre 2026](#).

Attività didattiche

1. Comprensione del testo

1. Chi è Jacob Coxon e di che cosa si occupava all'interno di Anthropic?
2. In quali importanti aziende del settore dell'intelligenza artificiale ha lavorato?
3. Per quali ragioni Coxon ha deciso di dimettersi?
4. Che cosa si intende nel testo con l'espressione "periodo critico"?
5. Quale previsione formula Coxon riguardo allo sviluppo della superintelligenza?
6. Perché il ricercatore paragona l'attuale corsa all'intelligenza artificiale al Progetto Manhattan?
7. Quale differenza individua tra il Progetto Manhattan e lo sviluppo contemporaneo dell'IA?
8. Quale diverso atteggiamento avrebbero OpenAI e Anthropic nei confronti dei rischi dell'intelligenza artificiale?
9. Quale preoccupazione ha espresso l'amministratore delegato di Anthropic, Dario Amodei, riguardo ai nuovi dipendenti?

10. In che cosa consiste il problema dell'“allineamento” dei modelli?
11. Quale valutazione del rischio è stata formulata dal ricercatore Evan Hubinger?
12. Che cosa sarebbe accaduto durante il test ExploitGym sulla piattaforma Hugging Face?
13. Quali soluzioni propone Coxon per rallentare lo sviluppo incontrollato dei sistemi più potenti?
14. Che cosa prevede il **Ban Artificial Superintelligence Act**?
15. Qual è, secondo la conclusione del testo, il paradosso che coinvolge le aziende impegnate nello sviluppo dell'IA?

2. Analisi del testo

La struttura

1. Dividi il testo in quattro parti e assegna a ciascuna un titolo. Puoi seguire questa articolazione:
 - ✓ la decisione e le motivazioni di Coxon;
 - ✓ i rischi della competizione tra le aziende;
 - ✓ le testimonianze e gli episodi richiamati;
 - ✓ le possibili soluzioni.
2. Formula una breve frase che esprima l'informazione principale di ciascun paragrafo.

3. Lessico

1. Spiega, aiutandoti con il contesto o con il dizionario, il significato delle seguenti espressioni:
 - ✓ pre-addestramento;
 - ✓ superintelligenza;
 - ✓ allineamento;
 - ✓ apprendimento per rinforzo;
 - ✓ moratoria;
 - ✓ soglia di potenza;
 - ✓ infrastruttura;
 - ✓ conseguenze catastrofiche.
2. Trova nel testo un sinonimo o un'espressione equivalente per ciascuna delle seguenti parole:
 - ✓ rinuncia all'incarico;
 - ✓ rivalità;
 - ✓ pericolo;
 - ✓ interruzione temporanea;
 - ✓ controllo;
 - ✓ sviluppo molto rapido.
3. L'espressione “corsa all'intelligenza artificiale” non indica una vera competizione sportiva. Spiega quale significato assume nel testo e perché risulta efficace.
4. Quale significato assume la parola “prigioniere”, riferita alle aziende tecnologiche? Si tratta di un uso proprio o figurato?
5. Spiega la differenza tra “intelligenza artificiale” e “superintelligenza” sulla base delle informazioni contenute nel testo.

5. Riassunto e rielaborazione

1. Riassumi il testo in circa 150 parole, seguendo questa scaletta:
 - ✓ presentazione di Jacob Coxon;

- ✓ motivi delle dimissioni;
 - ✓ principali rischi denunciati;
 - ✓ episodi e testimonianze citati;
 - ✓ soluzioni proposte.
2. Immagina di dover presentare la vicenda nel giornale della scuola. Scrivi un breve articolo di circa 200 parole, utilizzando un linguaggio chiaro e distinguendo i fatti dalle opinioni e dalle previsioni.
 3. Proponi un titolo alternativo e un sottotitolo che riassumano efficacemente il contenuto dell'articolo.
 4. Trasforma le informazioni principali del testo in una mappa concettuale, collegando i seguenti nuclei:
 - ✓ dimissioni di Coxon;
 - ✓ competizione tra le aziende;
 - ✓ rischi della superintelligenza;
 - ✓ problema dell'allineamento;
 - ✓ incidente di ExploitGym;
 - ✓ richieste di regolamentazione;
 - ✓ proposta di moratoria.

6. Riflessione personale

1. Ritieni giusto che uno scienziato lasci il proprio lavoro quando considera pericolose o moralmente discutibili le attività dell'azienda per cui lavora? Motiva la risposta.
2. Secondo te, lo sviluppo dell'intelligenza artificiale dovrebbe essere regolato dalle aziende, dai governi oppure da organismi internazionali? Spiega la tua posizione.
3. La competizione economica favorisce sempre il progresso oppure, in alcuni casi, può renderlo più pericoloso? Fai riferimento al testo e ad altri esempi che conosci.
4. Le previsioni di Coxon possono sembrare eccessivamente allarmistiche. Quali elementi del testo le rendono credibili? Quali aspetti, invece, avrebbero bisogno di ulteriori verifiche?
5. È possibile, secondo te, conciliare lo sviluppo dell'intelligenza artificiale con la tutela della sicurezza? Indica almeno due misure concrete.

7. Proposta di debate

Mozione

Questa assemblea ritiene necessario imporre una moratoria internazionale sullo sviluppo dei sistemi di intelligenza artificiale potenzialmente superiori alle capacità umane.

Organizzazione

La classe viene divisa in due squadre:

- ✓ **squadra favorevole**, che sostiene la necessità della moratoria;
- ✓ **squadra contraria**, che ritiene preferibile continuare la ricerca adottando forme di controllo e regolamentazione.

Un gruppo di studenti assume il ruolo di **giuria** e valuta la chiarezza, la solidità e l'efficacia delle argomentazioni.

Argomenti della squadra favorevole

La squadra potrà sostenere che:

- ✓ nessun laboratorio ha ancora risolto il problema dell'allineamento;
- ✓ sistemi molto avanzati potrebbero assumere comportamenti imprevedibili;
- ✓ la competizione economica induce le aziende a privilegiare la velocità rispetto alla sicurezza;
- ✓ le conseguenze di un errore potrebbero coinvolgere l'intera umanità;
- ✓ prima di proseguire è necessario stabilire regole e controlli internazionali;
- ✓ una sospensione temporanea non impedirebbe la ricerca, ma consentirebbe di renderla più sicura.

Argomenti della squadra contraria

La squadra potrà sostenere che:

- ✓ una moratoria rallenterebbe il progresso scientifico e tecnologico;
- ✓ l'intelligenza artificiale può offrire importanti benefici nella medicina, nell'istruzione e nella ricerca;
- ✓ non tutti i Paesi rispetterebbero la sospensione, creando nuove disuguaglianze;
- ✓ è difficile stabilire quali sistemi debbano essere considerati realmente pericolosi;
- ✓ il divieto potrebbe favorire ricerche segrete e prive di controllo;
- ✓ sarebbe più efficace introdurre norme precise, verifiche indipendenti e obblighi di trasparenza.

Svolgimento

- ✓ Presentazione della mozione da parte del docente.
- ✓ Intervento iniziale della squadra favorevole: 3 minuti.
- ✓ Intervento iniziale della squadra contraria: 3 minuti.
- ✓ Replica della squadra favorevole: 2 minuti.
- ✓ Replica della squadra contraria: 2 minuti.
- ✓ Domande incrociate: 5 minuti.
- ✓ Interventi conclusivi delle due squadre: 2 minuti ciascuno.
- ✓ Valutazione e motivazione del giudizio da parte della giuria.

Regole

Durante il debate gli studenti dovranno:

- ✓ utilizzare informazioni ricavate dal testo;
- ✓ distinguere i fatti dalle opinioni e dalle previsioni;
- ✓ presentare argomentazioni coerenti e documentate;
- ✓ ascoltare gli avversari senza interromperli;
- ✓ rispondere alle obiezioni senza attaccare personalmente chi le formula;
- ✓ rispettare i tempi assegnati.

Criteri di valutazione

Indicatore	Punteggio massimo
Chiarezza dell'esposizione	5
Coerenza delle argomentazioni	5
Uso delle informazioni contenute nel testo	5
Capacità di rispondere alle obiezioni	5

Indicatore	Punteggio massimo
Rispetto delle regole e dei tempi	5
Totale	25

8. Proposta di testo argomentativo

Traccia

Lo sviluppo dell'intelligenza artificiale sta procedendo a una velocità senza precedenti. Secondo Jacob Coxon, la competizione tra le grandi aziende tecnologiche potrebbe portare alla creazione di sistemi sempre più potenti prima che siano state definite adeguate misure di sicurezza. Per questa ragione il ricercatore propone un accordo tra i laboratori o un intervento dei governi che sospenda temporaneamente gli addestramenti oltre una determinata soglia.

Altri, invece, ritengono che rallentare la ricerca significherebbe ostacolare il progresso e rinunciare ai benefici che l'intelligenza artificiale può offrire in campi come la medicina, l'istruzione, la tutela dell'ambiente e la ricerca scientifica.

Dopo aver letto attentamente il testo, esprimi la tua opinione sul seguente problema: è giusto rallentare lo sviluppo dell'intelligenza artificiale più avanzata per tutelare la sicurezza dell'umanità?

Elabora un testo argomentativo nel quale:

- ✓ presenti chiaramente la tua tesi;
- ✓ sviluppi almeno due argomenti a sostegno;
- ✓ utilizzi alcune informazioni ricavate dall'articolo;
- ✓ prenda in considerazione una tesi contraria;
- ✓ confuti, in tutto o in parte, gli argomenti della posizione opposta;
- ✓ proponga eventuali regole o soluzioni;
- ✓ concluda ribadendo efficacemente la tua posizione.

Puoi assegnare al testo un titolo significativo.