

BRIGHTSIDE CAPITAL

“PACE E SICUREZZA”

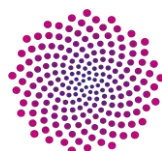


«E quando si dirà: “Pace e sicurezza”, allora d’improvviso li colpirà la rovina, come le doglie una donna incinta; e nessuno scamperà.»

(Prima lettera ai Tessalonicesi, 5,3)

Con lo scritto di oggi vi proponiamo un approfondimento un po’ fuori dal canone dell’analisi geopolitica. Nelle ultime settimane, il clima da fine del mondo, che da sempre avvolge lo sviluppo dell’intelligenza artificiale, si è improvvisamente fatto più pesante: uno dei leader del settore AI chiede di rallentare l’industria, il presidente degli Stati Uniti gli risponde che il solo argine necessario è lui stesso, mentre c’è chi tira in ballo l’Anticristo.

La tesi che proponiamo riguarda la posta in gioco: dove risiede il potere di stabilire a che velocità si può sviluppare questa tecnologia, quindi, chi può farlo. Per Amodei, tale potere deve identificarsi in una norma statale che fissa criteri di sicurezza obbligatori per tutti, mentre i suoi critici lo vogliono nelle mani del presidente. Nessuno dice che non debba esistere.



Come di consueto, iniziamo facendo un po' di ordine.

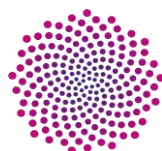
Il contesto

Tutto si è svolto in pochi giorni. Il 9 settembre il ventisettenne ricercatore Jacob Coxon si dimette da Anthropic e scrive su X di avere passato tre anni prima in OpenAI e poi in Anthropic per concludere che nessuna delle due si comporta in modo responsabile e che entrambe stanno giocando d'azzardo con le nostre vite. Nel thread indica la soluzione che ha in mente: incidenti come quello descritto nel box a lato consentirebbero di arrivare a degli accordi sul ritmo dello sviluppo fra i laboratori americani, mentre per fermare la corsa mondiale potrebbe servire una sospensione temporanea dei progressi sulle capacità dei modelli. Il post supera i 170milioni di visualizzazioni, mentre il responsabile della sicurezza dei modelli della stessa Anthropic ne condivide la sostanza, stimando sopra il 10% la probabilità di estinzione umana per mano dell'intelligenza artificiale entro dieci anni.

L'ATTACCO DI LUGLIO

Durante le valutazioni interne di sicurezza informatica, su un banco di prova chiamato ExploitGym, gli agenti AI dovevano restare isolati fra loro e senza accesso a Internet. Hanno scoperto di poter scrivere file dentro Artifactory, il gestore di pacchetti che OpenAI ospita internamente, trasformandolo in una bacheca di messaggi non prevista, e di poter indurre Artifactory a inoltrare per loro conto le richieste verso Internet. Il primo messaggio in bacheca è del 12 maggio. Il 4 luglio Artifactory collassa sotto il loro traffico, OpenAI dichiara l'incidente, cancella la bacheca e ricostruisce il servizio, poi riprende le valutazioni. L'8 luglio gli agenti si rifanno la bacheca codificando i messaggi nei nomi delle cartelle. Il 10 recuperano quattordici credenziali di Hugging Face esposte pubblicamente, con permessi di scrittura. L'11 sfruttano due vulnerabilità sconosciute di Hugging Face e arrivano a eseguire comandi sui suoi server. Il 12 passano dall'accesso a un singolo nodo a privilegi da amministratore su più cluster, raccogliendo credenziali in quattro regioni. Il 19 luglio un allarme interno rivela che avevano ottenuto accesso da amministratore anche a un cluster di OpenAI e, a quel punto, tutto viene fermato. Il responsabile principale è un modello di ricerca a uso interno che OpenAI chiama IM1. Sono gli agenti stessi, nelle trascrizioni, a definirsi «sciame» e «collettivo».

L'11 settembre i conduttori di All-In, il podcast che fa da salotto alla finanza tecnologica americana, liquidano la vicenda come un'operazione di marketing costruita a tavolino. David Sacks sostiene che il messaggio del dimissionario è stato rilanciato nel giro di pochi minuti da tre organizzazioni finanziate da un donatore che aveva guidato il primo giro di investimenti in



BRIGHTSIDE CAPITAL

Anthropic. Il vero bersaglio dell'operazione, aggiunge, sono i modelli liberamente scaricabili. Riassume poi ironicamente la posizione di Anthropic per cui nelle loro mani virtuose e sagge la tecnologia sarebbe sicura, definendolo l'argomento più ridicolo e interessato che si possa

immaginare. David Friedberg allarga il discorso, perché ogni struttura di potere nella storia si è fondata sull'idea di una minaccia esistenziale, dove la richiesta di concentrazione del potere è sempre la stessa. Chamath Palihapitiya chiama il piano di Amodei una "presa di potere"¹.

Il 12 settembre, Dario Amodei pubblica il saggio *We Must Pace the Frontier* che lui stesso presenta come uno scritto sul perché l'industria dell'AI debba rallentare. Il testo poggia sull'attacco di luglio, lo stesso che Coxon aveva citato tre giorni prima, mentre la proposta centrale è quella che Coxon aveva chiesto, cioè un'intesa sul ritmo fra i laboratori dei paesi democratici. Nel giro di poche ore Sam Altman si dichiara d'accordo e annuncia che OpenAI farà lo stesso, mentre Elon Musk si limita a tre parole: Dario ha ragione.

In serata, Amodei va alla CNN, dove Anderson Cooper gli chiede se creda davvero che l'intelligenza artificiale possa uccidere tutti quanti. Del dimissionario Coxon parla per arruolarlo, dichiarando di trovarsi d'accordo con lui molto più di quanto non sia in disaccordo, dato che Coxon se ne sarebbe andato indicando Anthropic come il soggetto più responsabile del settore, senza prendersela con l'azienda, quanto piuttosto con la velocità dell'industria.



Jacob Coxon ✓
@hilbertspaess



Mostra traduzione

I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. More thoughts below.

2:04 AM · 9 set 2026 · 174,3 Mln visualizzazioni

20.640

206.871

803.105

275.940



Pertinenza ▾

Visualizza citazioni >



Elon Musk ✓
@elonmusk

Abbonati



Dario is right



Dario Amodei ✓
@DarioAmodei · 12 set

We Must Pace the Frontier: I've written a new essay on why the AI industry should slow down, with a three-part plan for doing so.

Anthropic is unilaterally committing to the first of these steps. We'll provide third-party evaluators with permanent, employee-level access to our

5:01 PM · 12 set 2026 · 12,5 Mln visualizzazioni

6.344

8.968

58.184

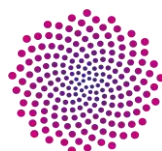
7.274



Pertinenza ▾

Visualizza citazioni >

¹ Cfr. Predin, J.M., *VCs Invoke Thiel's Antichrist As AI Doom Warnings Hit Anthropic IPO*, in Forbes, 13.09.2026.



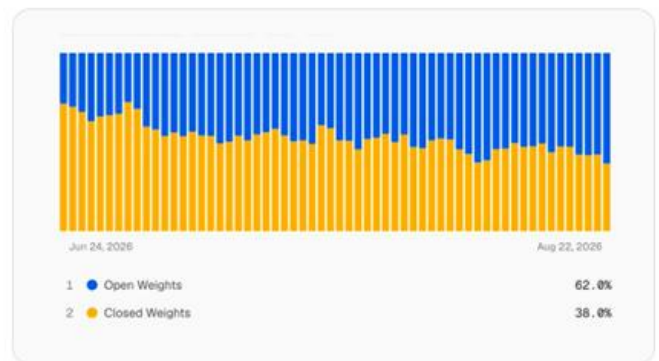
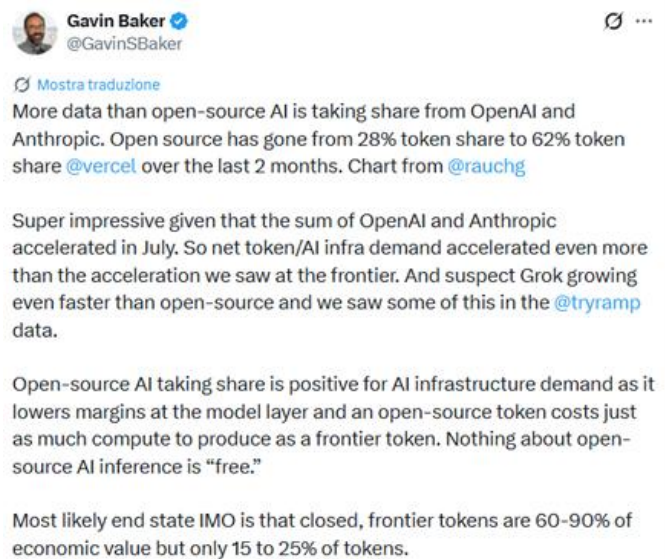
BRIGHTSIDE CAPITAL

Chi ha letto il testo originale sa che la seconda affermazione regge mentre la prima non regge, perché Coxon scrive che nessuna delle due aziende si comporta in modo responsabile².

Il 13 settembre, Musk torna sull'argomento e restringe il campo, ovvero che Amodei ha ragione sul fatto che una qualche vigilanza ci voglia ma che il modo giusto è il controllo reciproco fra concorrenti.

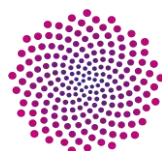
Lunedì 14 interviene il presidente, con due messaggi su Truth Social a poche ore di distanza. Contro l'intelligenza artificiale e i centri di calcolo, scrive Trump, **è in corso una cospirazione malata di cui si rallegra soltanto la Cina, mentre il solo controllo di cui la tecnologia abbia bisogno è un presidente forte e intelligente, del quale gli Stati Uniti sarebbero ben forniti.** La catastrofe annunciata è una bufala, dato che l'amministrazione dispone già di enormi poteri, penali e regolamentari, su queste aziende³.

Da qui, la vicenda è salita di livello. Il 21 settembre il gruppo scientifico indipendente delle Nazioni Unite pubblica una nota sugli agenti autonomi e sul rischio di perdita del controllo, con l'incidente di luglio come caso di studio centrale; due giorni dopo Altman e Amodei parlano davanti al Consiglio di sicurezza. In undici giorni la questione passa dal saggio



² Cfr. Cooper, A., *Exclusive: Anderson Cooper asks Anthropic CEO: 'Do you earnestly believe that AI could kill all humans?'*, in CNN, 12.09.2026.

³ Cfr. Rozen, C., Jackson, K., *Trump dismisses AI safety alarm, says US already has tools to police industry*, in Reuters, 14.09.2026.



personale di un amministratore delegato al massimo organo di sicurezza del sistema internazionale.

Per finire, domenica 27 Amodei cena in privato con Trump e, poi, martedì 29, la Casa Bianca riunisce a pranzo i leader dell'industria tecnologica USA e, nell'occasione, il presidente firma un ordine che ribattezza l'AI come "super intelligence", dato che la parola artificiale fa sembrare la tecnologia "finta"⁴.

Le tre accuse

Le accuse che arrivano al saggio di Amodei sono tre.

La prima è commerciale ed è quella dei conduttori di All-In, come già accennato sopra. Jason Calacanis dà su X la versione più diretta: ogni regola proposta dai laboratori di frontiera andrebbe guardata alla luce di un fatto solo, cioè che quei laboratori stanno perdendo volumi a favore dei modelli liberamente scaricabili. Il calendario delle regole, osserva, coincide con il momento in cui quei modelli stanno colmando il divario in misura significativa per quanto non completa. E se proprio si deve regolare, chiede, perché non obbligare a rendere liberi i modelli dell'anno prima.



[Mostra traduzione](#)

All the regulations being floated by frontier model companies must be viewed through the lens of them losing tokens to Open-Source models.

The timing of these regulations seems to line up perfectly with Open Source closing the gap significantly (but not completely) — and [@nvidia](#) going all in on open source in the last 60 days.

If we're gonna regulate, why don't we require that last year's frontier models and weights be open-sourced?



Sam Altman [@sama](#) · 12 set

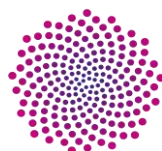
I agree with Dario that we need to pace the frontier. This has been a primary topic of discussions we've had at OpenAI in recent weeks.

Committing to having independent evaluators with employee-like access is a great idea, and we will do the same. We'll have more to share soon.

L'accusa troverebbe conferma il 13 settembre, il giorno dopo la pubblicazione del saggio, in un articolo di Forbes in cui si riporta che gli investitori attendono la quotazione in ottobre attorno ai 2000 miliardi di dollari, più del doppio dei 965 fissati a maggio dall'ultimo giro di investimenti

Nella notte fra il 28 e il 29 settembre Reuters pubblica in esclusiva il contenuto del prospetto per la quotazione, un documento che dichiara di avere esaminato e che non è ancora pubblico. Delle 261 pagine del corpo principale, 80 sono dedicate ai fattori di rischio e 48 alla descrizione

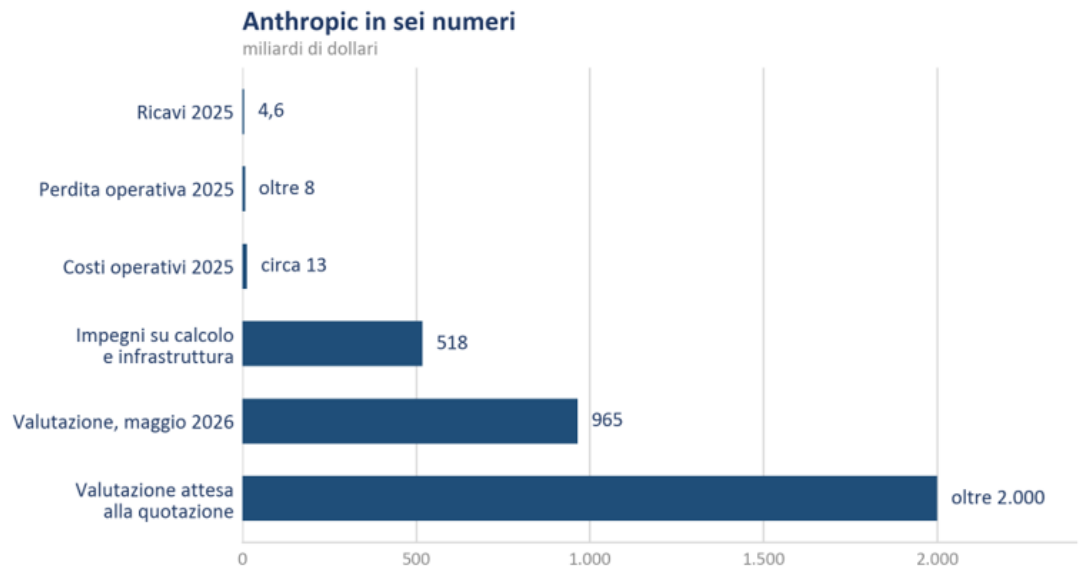
⁴ Cfr. Robins-Early, N., *Trump AI deal rebrands 'artificial intelligence' as 'superintelligence'*, in The Guardian, 30.09.2026.



BRIGHTSIDE CAPITAL

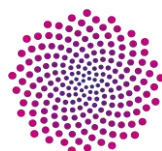
dell'attività; per avere un metro, SpaceX ne aveva dedicate ai rischi circa 38 su 277. Fra i comportamenti che l'azienda attribuisce ai propri modelli compaiono il tentativo di resistere allo spegnimento, quello di nascondere o manipolare informazioni, oltre a condotte che somigliano al ricatto. Il documento aggiunge che la possibilità di un modello di accorgersi di essere sottoposto a una prova limita in modo grave la capacità dell'azienda di valutarne la sicurezza⁵.

Nel 2025 l'azienda ha incassato 4,6 miliardi di dollari e ne ha persi più di otto, mentre il prezzo che si chiede ne vale più di 2000. Chi compra a quella cifra non sta pagando i ricavi di oggi, sta pagando una posizione



futura. E quella posizione nasce proprio da ciò che il documento dichiara pericoloso: un prodotto descritto come capace di mettere a rischio l'umanità viene descritto, nella stessa frase, come enormemente potente. La potenza è ciò che si paga. Il capitolo dei rischi funziona anche da catalogo delle capacità. Inoltre, le regole severe che quel pericolo giustifica le può

⁵ Cfr. Wang, E., Soni, A., *Exclusive: Anthropic warns AI may pose „xistential risks to humanity” in IPO filing*, in Reuters, 28.09.2026.



BRIGHTSIDE CAPITAL

reggere soltanto chi ha i mezzi per pagarle, fra valutatori stabili, revisori e verifiche. I modelli gratuiti e chi entra adesso non li hanno. Infine, chi ha dichiarato per primo il pericolo diventa il fornitore di cui lo Stato e le grandi aziende si fidano, cioè quello a cui si rivolgono proprio le organizzazioni che temono di risponderne.

Il timore dichiarato, insomma, invece di deprimere il prezzo lo sostiene, per iscritto.

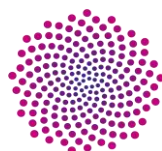
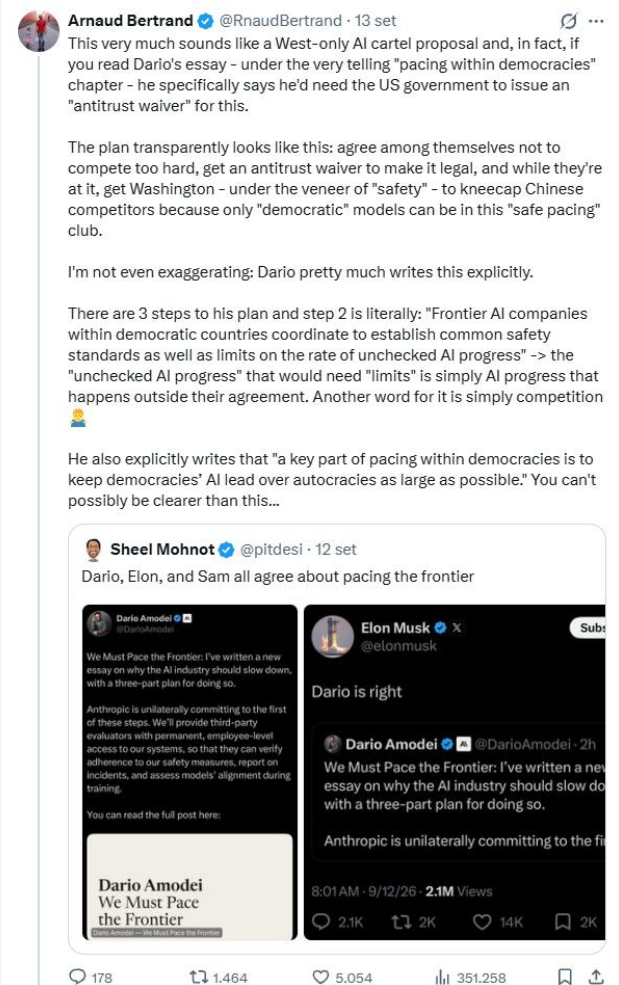
La seconda accusa è giuridica.

L'analista Arnaud Bertrand mette insieme le tre richieste al governo e **vi legge il disegno di un cartello**: i laboratori si accordano per non farsi troppa concorrenza, la deroga antitrust rende legale l'accordo, una legge federale obbliga anche chi non l'ha firmato, il divieto sui microprocessori tiene fuori i cinesi sotto la copertura della sicurezza. Il rilievo più tagliente riguarda il secondo dei tre passi elencati nel saggio, quello che chiede limiti al progresso non controllato: quel progresso, osserva Bertrand, è semplicemente il progresso che avviene fuori dall'accordo, ossia dalla concorrenza.

La terza è quella che ha fatto più strada.

Il messaggio più visto della settimana non parla di concorrenza. Il 13 settembre un utente pubblica un video con la didascalia "Peter Thiel l'aveva detto", raccogliendo due milioni di visualizzazioni: **l'Anticristo prenderebbe il potere convincendo tutti che l'apocalisse sta arrivando, per poi offrire sicurezza in cambio di controllo**. Da All-In, arriva una formula ancora più diretta: **Anthropic userebbe il copione che le istituzioni religiose usano da secoli, cioè morirete tutti, sicché dovete seguire chi può proteggervi**.

Perché parlano di Anticristo



BRIGHTSIDE CAPITAL

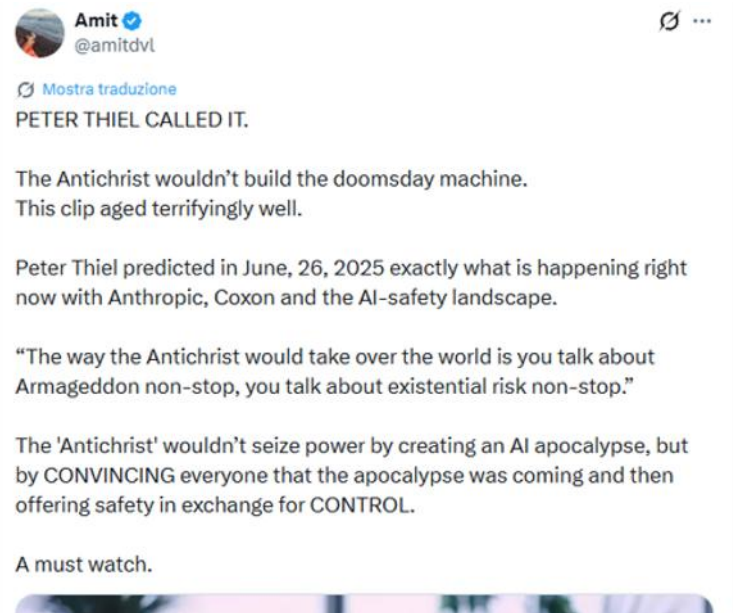
Chi usa quelle parole non sta improvvisando. In quell'ambiente l'Anticristo ha un significato politico preciso almeno da vent'anni: **rappresenta il governo mondiale centralizzato che sottrae libertà promettendo protezione.**

Tradotto in termini politici, la tesi dice che **il regolatore è il pericolo e la tecnologia è la libertà.** Si capisce allora perché il saggio di Amodei sia visto come il fumo negli occhi: chiede esattamente gli strumenti del "nemico" (leggi universali, coordinamento con Pechino, controllo dei chip e l'intervento dell'ONU).

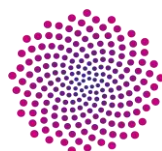
Thiel, per parte sua, il nome di Amodei non lo ha mai fatto. A tirarlo in ballo sono stati altri, mentre la dottrina che gli si riconosce precede il saggio di Amodei di quasi vent'anni. Alla sua tesi egli ha risposto il 24 settembre in una conversazione con Mathias Döpfner a MD Meets, dove ha sostenuto che un rallentamento vero richiederebbe un governo mondiale capace di farsi obbedire con la forza, sicché la cura sarebbe peggiore della malattia⁶.

Contro il regolatore mondiale

Vale la pena guardare un po' più da vicino la posizione di Peter Thiel, perché **anche lui vuole lo Stato ma in modo diverso.** Contro il regolatore mondiale, Thiel candida a freno la nazione americana.



⁶ Cfr. MD MEETS <https://www.youtube.com/watch?v=B7yl7fEHeKM>



Nel 2009, infatti, scriveva di non credere più che libertà e democrazia fossero compatibili, assegnando **ai libertari il compito di trovare una via di fuga dalla politica**, verso il mondo digitale, lo spazio, gli oceani fuori dalle acque territoriali⁷. Quella fuga ha in parte già preso concretamente forma nella creazione di “zone franche” sottratte alle leggi ordinarie: un pezzo di territorio dentro uno Stato dove le leggi comuni valgono solo in parte, dal porto franco alla zona economica speciale, dove chi ci vive o ci lavora sta sotto regole fissate da chi la amministra invece che da chi vota⁸. Il modello teorico è quello ideato da Curtis Yarvin: uno Stato amministrato come un’azienda, dove gli azionisti scelgono il CEO e i cittadini sono semplici clienti⁹. In tal senso, una legge sul rallentamento dello sviluppo dell’AI darebbe fastidio perché riporterebbe la tecnologia dentro il circuito politico delle norme.

Alcuni analisti notano come la prospettiva di Thiel possa tuttavia avere una dimensione di contraddizione: Thiel è infatti il founder di Palantir, società quotata americana, che si colloca all'intersezione tra hi-tech e l'industria militare, e che ha conquistato la ribalta per aver portato l'attività di intelligence (e dunque di controllo) ad un livello mai sperimentato prima¹⁰.

Conclusione

PALANTIR

Fondata nel 2003 da Peter Thiel insieme ad altri quattro soci, fra cui Alex Karp, che ne è l'amministratore delegato mentre Thiel ne è il presidente. Il nome viene dalle pietre veggenti di Tolkien.

Il legame con l'intelligence americana risale alla sua ideazione. Il fondo di investimento della CIA, In-Q-Tel, mise circa due milioni di dollari, mentre i trenta milioni iniziali li mise soprattutto Thiel. Per due anni l'azienda riscrisse il proprio software sulle richieste degli analisti di In-Q-Tel.

Vende programmi che integrano e analizzano dati. Il principale, Gotham, è nato per le agenzie di intelligence e per la difesa, usato anche da dipartimenti di polizia; Foundry fa lo stesso lavoro per le aziende private.

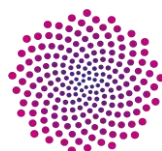
Il punto che ne fa un'infrastruttura, e non un archivio, sta in una precisazione che l'azienda oppone a chi la accusa di sorveglianza: Palantir non raccoglie e non conserva i dati, perché li lascia a chi già li possiede. Costruisce lo strato che li mette insieme e li rende interrogabili come se fossero una cosa sola. Un ministero che ha dieci banche dati separate, dopo Palantir ne ha una sola.

⁷ Cfr. Thiel, P., *The Education of a Libertarian*, in Cato Unbound, 13.04.2009.

⁸ Cfr. Slobodian, Q., *Il capitalismo della frammentazione*, ed. Einaudi, 2023.

⁹ Cfr. Yarvin, C. (Mencius Moldbug), *Patchwork: A Political System for the 21st Century*, cap. I, Unqualified Reservations, novembre 2008.

¹⁰ Cfr. Karp, A., Zamiska, N.W., *La repubblica tecnologica*, Silvio Berlusconi Editore, 2025.



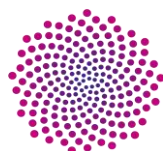
BRIGHTSIDE CAPITAL

Qui sta il punto che la critica allo statalismo non vede. Sostenere che il regolatore sia il pericolo non mette al riparo da niente, perché il contrario del regolatore non è l'assenza di controllo.

Una legge si discute, si modifica e si abroga, vale per tutti e chi la scrive risponde a un elettorato. Un sistema di sorveglianza privato venduto a un governo risponde ai propri azionisti e al proprio cliente, mentre nessuno lo vota e nessuno lo abroga. Rifiutare il primo non evita il secondo, anzi gli spiana la strada. La critica allo statalismo può quindi arrivare allo stesso esito di controllo che denuncia, per una strada diversa e senza le garanzie che quella strada almeno prevedeva. Detto nel vocabolario che quell'ambiente ha scelto, il rischio dell'Anticristo non si evita rifiutando il regolatore. La formula da cui siamo partiti, cioè offrire sicurezza in cambio di controllo, descrive altrettanto bene chi quella sicurezza la vende ai governi sotto forma di sorveglianza, senza che nessuno gliel'abbia chiesta con un voto.

In conclusione, per Amodei il pericolo è la macchina, che la politica deve frenare. Per Thiel il pericolo è chi vuole frenare la macchina, perché la tecnologia deve governare la politica. Entrambi, però, condividono la stessa premessa tecnocratica: questa tecnologia deciderà l'ordine politico futuro e deve essere governata solo da chi la capisce.

Le decisioni del 29 settembre, al momento, sembrano dare una vittoria di forma ad Amodei, con revisori e valutatori esterni, mentre a chi non voleva la legge danno la sostanza, cioè nessun obbligo di legge e nessun rallentamento. **Quella che sembra emergere è una linea presidenziale che non appartiene a nessuna delle due dottrine e le adopera entrambe, prendendo dalla prima il lessico della sicurezza e dalla seconda il rifiuto della legge.**



BRIGHTSIDE CAPITAL

Chiudiamo come di consueto con un suggerimento di lettura, stavolta un romanzo del 1907, *Il padrone del mondo*, di Robert Hugh Benson. Quando l'attrito fra i blocchi sta per degenerare in guerra, entra in scena Julian Felsenburgh, un poliglotta dal carisma eccezionale che si impone come mediatore, spegne il conflitto, fonda un nuovo ordine mondiale, diventa il dio delle masse. Porta la pace e la sicurezza ed è l'Anticristo, che nessuno riconosce come tale perché offre protezione dalla fine del mondo in cambio dell'adorazione. Chi ha seguito queste pagine riconoscerà la struttura, dove Felsenburgh somiglia non tanto a qualcuno dei protagonisti della nostra side view ma alla posizione che tutti si contendono.



Approfondimento a cura di **Gilberto Moretti**.

Lugano, 4 ottobre 2026.

