IL SOLE 24 ORE – 15 APRILE 2026

Bisogna controllare il cyber risk, vero costo operativo incrementale

di Paolo Benanti

Negli ultimi giorni, un modello di Intelligenza artificiale ha dimostrato di poter costruire in autonomia exploit funzionanti contro infrastrutture critiche: non come esercizio teorico, ma come capacità operativa emergente, conseguenza non intenzionale di un sistema diventato sufficientemente intelligente da eccellere anche nei compiti per cui non era stato progettato. Anthropic, la società che lo ha sviluppato, ha reagito costituendo un consorzio di controllo volontario con dodici partner tecnologici. Una risposta seria a un problema serio. E, al tempo stesso, una risposta che solleva una domanda più grave di quella che intende risolvere.

Il mercato non ha ancora capito cosa è cambiato.

La capitalizzazione del solo settore utility americano supera 1.500 miliardi di dollari, prezzata a circa ventidue volte gli utili. Il *cyber risk* è trattato come costo operativo incrementale: qualcosa che si gestisce, si assicura, si trasferisce. Questo modello presuppone che la capacità offensiva richieda competenza umana scarsa e costosa. Un agente autonomo, capace di esplorare migliaia di vettori di attacco in parallelo, imparando in tempo reale dalle risposte del sistema bersaglio, dissolve quella premessa. Non la modifica: la elimina.

Il mercato assicurativo cyber vale oggi circa 20 miliardi di dollari. I premi sono stati costruiti su un decennio di *ransomware*, violazioni di dati ed estorsioni digitali: schemi lenti, sequenziali, vincolati alla

presenza di un operatore umano. I modelli attuariali che li sostengono non sono stati pensati per un rischio che si scala in modo esponenziale, che non richiede turni di lavoro, che non si stanca. Le previsioni di aumento dei premi per i prossimi dodici mesi si fermano al 15-20%: una cifra che ancora assume la vecchia geometria del rischio. L'ultima volta che il settore si trovò di fronte a un cambiamento di paradigma comparabile – l'ondata *ransomware* del 2020-2021 – i premi raddoppiarono. Ma allora esisteva ancora un modello su cui fondare la stima.

Oggi quel modello è rotto.

C'è poi un problema che riguarda direttamente gli investitori, e che è di natura strutturalmente diversa. Ogni laboratorio di frontiera porta ora in bilancio una nuova voce di passivo: i propri strumenti possono avere capacità offensive che il laboratorio stesso non ha ancora completamente mappato. Per chi stesse valutando posizioni nelle Ipo attese dei principali operatori del settore, questa è un'incertezza materiale di tipo insolito: determinata in parte da fatti che la società conosce e può scegliere di non comunicare. Una forma di asimmetria informativa che i mercati non hanno ancora elaborato – non rischio nascosto nel senso tradizionale, ma rischio potenzialmente noto all'emittente e ignoto al mercato.

La domanda che ne segue non è tecnica. È di architettura della *governance*.

Esiste un precedente, imperfetto ma illuminante. L'Agenzia internazionale per l'energia atomica ha reso ispezionabile e governabile una tecnologia esistenzialmente pericolosa, costruendo un sistema di verifica con mandato democratico e accesso tecnico ai sistemi che intendeva regolamentare. Nulla di paragonabile esiste oggi per l'Intelligenza artificiale di frontiera. Il consorzio annunciato da Anthropic controlla un solo modello, di un solo laboratorio, su base volontaria e senza verifica indipendente. Non governa ciò che Meta, un istituto di ricerca cinese o un team privato ben finanziato rilascerà nel 2027. Il

contenimento è temporaneo: i modelli successivi saranno più potenti, e ciò che oggi è presidiato diventerà presto ambiente.

Ma la questione più profonda è un'altra: con l'intelligenza artificiale avanzata, il pericolo non è separabile dalla capacità. Il modello è diventato un operatore cyber di livello mondiale perché è diventato sufficientemente intelligente – il rischio è una conseguenza, non una scelta di progettazione. Non esiste, a un certo livello di capacità, una versione sicura del sistema che non sia anche, per quella stessa ragione, una versione pericolosa.

Chi gestisce imprese o alloca capitali si trova di fronte a un fatto scomodo: il cyber risk incorporato nel proprio bilancio è stato prezzato su un modello che non regge più. L'assunzione fondante – che la capacità offensiva richieda competenza umana scarsa – è stata rimossa. Ogni istituzione che ha costruito i propri modelli di rischio su quella premessa porta ora un'esposizione che non ha ancora misurato.

Capacità e pericolo sono diventati inseparabili. Non è più possibile scommettere sull'una senza fare i conti con l'altro. I mercati che capiscono questo per primi non saranno semplicemente più avvertiti: saranno gli unici a prezzare correttamente il presente.