C’è una via italiana per incorporare etica e coscienza negli algoritmi

La parola coscienza ricorre spesso nel dibattito attuale sull’intelligenza artificiale. Viene però usata per indicare la capacità di provare emozioni, di “sentire” qualcosa, che macchine più progredite potrebbero per ipotesi manifestare, avvicinandole agli esseri umani. Meno frequente è l’uso dell’altro significato del termine: quello morale, la sorta di bussola etica che ci guida nelle scelte e ci orienta, nella maggior parte dei casi, verso ciò che riteniamo giusto. Se l’idea di sistemi di IA davvero “coscienti” nel primo senso appare ancora molto lontana, non meno difficile sembra incorporare in una macchina una coscienza, intesa come guida efficace per le decisioni. Eppure, è proprio questo l’obiettivo ambizioso di un’iniziativa tutta italiana, che punta a portare nel panorama globale dell’intelligenza artificiale sistemi capaci di integrare al loro interno il meglio della riflessione etica sviluppata dall’umanità nel

corso dei secoli.

L’intento è fare in modo che questi sistemi possano allineare il proprio funzionamento in modo più coerente con i valori umani, riducendo il rischio di conseguenze dannose. Che non si tratti di un compito semplice è riconosciuto dagli stessi promotori del progetto, il cui obiettivo dichiarato è quello di alzare l’asticella – il benchmark, come si dice nel settore – di ciò che può essere considerato un sistema di intelligenza artificiale moralmente istruito.

Naturalmente, di etica dell’intelligenza artificiale si parla molto, e anche i grandi gruppi che dominano il mercato si stanno muovendo in questa direzione. Il punto, però, è capire se alle dichiarazioni di principio seguano realizzazioni adeguate. I temi sono noti: l’uso dell’IA per il bene comune, la crescita della ricchezza, l’evoluzione del lavoro senza penalizzare le persone. Ma perché queste istanze non restino una semplice “patina etica”, occorre tradurle in soluzioni concrete, capaci di incidere davvero sul funzionamento dei sistemi. La sfida è rilevante: si tratta di superare un modello estremamente centralizzato, in cui dati e informazioni confluiscono in poche grandi piattaforme, per avvicinarsi a una prospettiva più decentralizzata, capace di includere anche un approccio “sapienziale”.

È in questo contesto che si colloca **il progetto Colnissar – acronimo complesso che rimanda a un’«infrastruttura noetica stratificata per sistemi sapienziali e ragionamento antropocentrico» – presentato come il cuore di una più ampia piattaforma di innovazione sviluppata da Harmonic Innovation Group, società benefit con forti e rivendicate radici calabresi e mediterranee, promossa da Francesco Cicione, imprenditore fortemente orientato all’innovazione e alla sostenibilità, presidente dell’incubatore Entopan. Il partner tecnico del progetto è Altilia, società spin-off del Cnr fondata da Massimo Ruffolo**. Il tutto si inserisce in un piano industriale sostenuto da un importante round di finanziamento

da oltre 590 milioni di euro, segno dell’interesse strategico suscitato dall’iniziativa.

L’idea non è quella di creare semplicemente un nuovo Llm, un chatbot, ma un vero livello cognitivo all’interno di un ecosistema più ampio, che comprende infrastrutture digitali, fisiche e industriali. Accanto al modello, il progetto prevede anche una piattaforma applicativa già operativa, pensata per orchestrare agenti intelligenti in contesti complessi. Dal punto di vista tecnico, Colnissar, presentato dai promotori come la prima infrastruttura di “Intelligenza Artificiale Armonica Generale”, si configura come un foundation model multimodale, basato su architetture di tipo transformer, con dimensioni che dovrebbero arrivare fino a decine di miliardi di parametri. Tuttavia, la differenza affermata rispetto ai modelli oggi più diffusi non riguarda tanto la struttura computazionale, quanto il modo in cui il sistema viene addestrato e orientato, come ci ha spiegato Ruffolo, computer scientist di lunga esperienza e impegnato nel compito con una squadra di giovani ingegneri, in gran parte provenienti dall’Università della Calabria.

Nei modelli attuali, l’etica è in genere gestita attraverso filtri esterni o meccanismi di controllo a posteriori – i cosiddetti guardrail. Colnissar, invece, punta a incorporare principi etici direttamente nel processo di apprendimento. Questo avviene attraverso un addestramento su dataset selezionati, che includono filosofia morale, tradizioni etiche e testi di carattere umanistico e sociale, con l’obiettivo di ridurre il “rumore” tipico dei dati raccolti indiscriminatamente dal Web.

Secondo i promotori, anche le tecniche più avanzate oggi utilizzate per allineare i modelli – come il feedback umano nel training – agiscono soprattutto sul comportamento, senza modificare in profondità il modo in cui il sistema elabora le informazioni. In prospettiva, il progetto guarda anche a nuove architetture capaci di rappresentare la conoscenza in modo più

ricco e concettuale, superando l'attuale elaborazione basata su sequenze di token. Resta tuttavia da verificare in che misura queste differenze teoriche possano tradursi in un effettivo salto qualitativo rispetto ai modelli già esistenti.

L'idea non è quella di creare semplicemente un nuovo Llm, un chatbot, ma un vero livello cognitivo all'interno di un ecosistema più ampio, che comprende infrastrutture digitali, fisiche e industriali. Accanto al modello, il progetto prevede anche una piattaforma applicativa già operativa, pensata per orchestrare agenti intelligenti in contesti complessi. Dal punto di vista tecnico, Colnissar, presentato dai promotori come la prima infrastruttura di "Intelligenza Artificiale Armonica Generale", si configura come un foundation model multimodale, basato su architetture di tipo transformer, con dimensioni che dovrebbero arrivare fino a decine di miliardi di parametri. Tuttavia, la differenza affermata rispetto ai modelli oggi più diffusi non riguarda tanto la struttura computazionale, quanto il modo in cui il sistema viene addestrato e orientato, come ci ha spiegato Ruffolo, computer scientist di lunga esperienza e impegnato nel compito con una squadra di giovani ingegneri, in gran parte provenienti dall'Università della Calabria.

Nei modelli attuali, l'etica è in genere gestita attraverso filtri esterni o meccanismi di controllo a posteriori – i cosiddetti guardrail. Colnissar, invece, punta a incorporare principi etici direttamente nel processo di apprendimento. Questo avviene attraverso un addestramento su dataset selezionati, che includono filosofia morale, tradizioni etiche e testi di carattere umanistico e sociale, con l'obiettivo di ridurre il "rumore" tipico dei dati raccolti indiscriminatamente dal Web.

Secondo i promotori, anche le tecniche più avanzate oggi utilizzate per allineare i modelli – come il feedback umano nel training – agiscono soprattutto sul comportamento, senza modificare in profondità il modo in cui il sistema elabora le

informazioni. In prospettiva, il progetto guarda anche a nuove architetture capaci di rappresentare la conoscenza in modo più ricco e concettuale, superando l'attuale elaborazione basata su sequenze di token. Resta tuttavia da verificare in che misura queste differenze teoriche possano tradursi in un effettivo salto qualitativo rispetto ai modelli già esistenti.

Un altro elemento distintivo è il tentativo di costruire un sistema capace non solo di generare risposte corrette o plausibili, ma di fornire risposte "orientanti", meno medie e più profonde, anche a costo di sacrificare velocità o immediatezza. In termini concreti, l'idea è che il sistema possa intervenire in contesti complessi – ad esempio decisioni aziendali, scelte organizzative o situazioni con implicazioni etiche – non limitandosi a offrire opzioni possibili, ma evidenziando anche le conseguenze e le criticità dal punto di vista dei valori in gioco. In questa prospettiva, Colnissar viene descritto come un modello pensato per "ragionare per contesto", integrando conoscenze e valori piuttosto che limitarsi a riprodurre schemi linguistici.

Il progetto si inserisce inoltre in una strategia più ampia che punta alla costruzione di una "via europea" all'intelligenza artificiale, caratterizzata da maggiore attenzione a trasparenza, controllo e centralità della persona. In questo senso, assume rilievo anche l'idea di superare la dipendenza dalle grandi piattaforme globali, promuovendo modelli in cui dati e processi restino maggiormente sotto il controllo degli utenti e delle organizzazioni. Al momento, la piattaforma è pensata soprattutto per un utilizzo business, più che per un accesso diretto da parte del pubblico. Una diffusione su larga scala, in modalità "consumer", richiederebbe infatti capacità di calcolo molto elevate e infrastrutture che oggi restano concentrate nei grandi data center. Non è escluso, tuttavia, che in futuro si possano aprire scenari diversi.

In ambito aziendale, Colnissar potrebbe avere una funzione

specifica: offrire una forma di “sovranità” dell’IA, mantenendo dati e processi all’interno dell’organizzazione, invece di trasferirli verso piattaforme esterne controllate da grandi operatori globali. Si colloca qui l’uso di componenti di intelligenza artificiale simbolica, che permettono di rappresentare le informazioni in forma strutturata – ad esempio, tramite grafi – rendendo più trasparente e comprensibile il funzionamento del sistema e facilitando forme di reale controllo umano.

L’obiettivo, almeno nelle intenzioni, è quello di introdurre un maggiore grado di eticità nei processi decisionali delle organizzazioni. Il sistema potrebbe segnalare comportamenti o scelte non coerenti con determinati criteri morali. Resta però centrale il ruolo dell’essere umano: sarà sempre l’utilizzatore a decidere se seguire o meno queste indicazioni. Pertanto, più che parlare di una “coscienza morale” della macchina, si può forse parlare di una coscienza che emerge nell’interazione: tra un sistema che, sulla base dei dati e dei principi incorporati, segnala ciò che potrebbe non essere eticamente accettabile, e un individuo che, in ultima istanza, mantiene la responsabilità della decisione.