Agenti AI: da una “Lingua Franca” per l’AI a un nuovo paradigma per la sicurezza informatica e i rischi per l’utente umano

A cura di: Andrea Pasquinucci 15 Settembre 2025

Come uomini non possiamo non conoscere e apprezzare l’importanza del

linguaggio. La capacità dell’uomo di comunicare e quindi essere capace di trasferire facilmente informazioni complesse da uno all’altro, è sicuramente una delle nostre principali caratteristiche che ci differenzia da tutti gli altri esseri viventi sul nostro pianeta. Questo aspetto era già ben chiaro agli antichi, basta far mente locale all’episodio Biblico della Torre di Babele.

Una interpretazione di questo episodio è che gli uomini, che prima parlavano tutti la stessa lingua, per punizione divina avendo osato troppo per superbia, sono stati dispersi sulla terra parlando lingue diverse in modo da non poter più comunicare tra genti diverse. Se da una parte questo episodio vuole dare una spiegazione alla nascita delle tante diverse lingue umane, dall’altra sottolinea che non parlare la stessa lingua impedisce la comunicazione e procura un danno all’uomo.

Da sempre l’uomo ha cercato di adottare una “lingua franca” ovvero una lingua comune ad almeno buona parte degli uomini che permette loro di comunicare anche tra genti lontane. Nei secoli si sono succedute diverse “lingue franche”, per il mondo occidentale, dal greco al latino, dal francese all’inglese. Ed è ben noto che “conoscere le lingue” è un bene importante per ognuno di noi.

Questa introduzione, forse un po’ banale, vuole solo sottolineare l’importanza di comunicare, trasferire informazioni e capirsi. Il fatto è che questo concetto si trasferisce anche all’informatica. Basta pensare a quali sono stati alcuni dei principali volani dell’incredibile sviluppo dell’informatica negli ultimi 50 anni. Internet è fondamentalmente basato su alcuni protocolli, ovvero “lingue”, tra cui TCP/IP, SMTP, DNS, BGP, HTTP/HTML ecc. Senza di questi i nostri elaboratori non sarebbero capaci di scambiare informazioni, e Internet non esisterebbe.

È importante sottolineare che il successo e lo sviluppo dell’informatica e di Internet, ha richiesto l’adozione di protocolli **universali**, e che alcuni di questi sono evoluti e sono divenuti universali nel tempo. éovvio infatti che per poter accedere ad una risorsa in Internet in qualunque parte della terra, un PC o smartphone e un Browser devono parlare tutti la stessa “lingua”, ovvero utilizzare gli stessi protocolli, di tutti gli altri sistemi informatici.

Oltre ai protocolli di rete, si pensi anche ai formati di CD, DVD, MP3/4, USB-C

ecc. Per alcuni di questi esistevano protocolli o formati alternativi, anche tecnologicamente migliori per alcuni aspetti, ma l’universalità che include la facilità di gestione, creazione, adozione ecc., ha comunque e sempre avuto un fattore rilevante nel loro successo.

Intelligenza Artificiale e comunicazione

I modelli di Intelligenza Artificiale sono nati e per decenni sono stati sviluppati ed eseguiti come normali applicazioni informatiche che elaborano dati in ingresso per produrre dati in uscita. Questo schema basilare dell’informatica richiede che le informazioni necessarie all’elaborazione siano incluse o direttamente accessibili al programma stesso, ad esempio in una base dati dedicata.

Con l’arrivo dei modelli Large Language Model (LLM), l’architettura delle applicazioni informatiche ha subito un cambiamento che potrebbe mostrarsi essere radicale.

Sino ad oggi è stato l’uomo a dover imparare ad utilizzare una applicazione informatica o un sistema informatico, dal livello più tecnico quale imparare e utilizzare linguaggi di programmazione, a imparare ad utilizzare il linguaggio specifico di una applicazione. Si noti che ogni applicazione ha la propria logica, i propri flussi di utilizzo, modalità di accesso, modalità di preparazione e fruizione dei dati. Quindi per utilizzare un sistema informatico, l’uomo deve imparare il suo “linguaggio” (o “User Interface” – UI), facile o difficile che sia.

I modelli LLM ribaltano questo approccio esponendo quella che si può chiamare una “Human Interface” (HI). Questi modelli conoscono tutte le lingue dell’uomo e sono in grado (più o meno efficacemente) di comprendere il linguaggio umano e tradurlo nel proprio linguaggio macchina. Si possono quindi considerare alla stregua di traduttori universali, non solo tra le lingue umane, ma anche tra le lingue umane e la propria lingua informatica.

L’ambizione futura è quindi di non aver più bisogno di imparare la “lingua” di ogni applicazione ma di avere un “interprete” informatico personale che è in grado di tradurre quanto scriviamo o diciamo in qualunque altra “lingua”,

informatica, scritta o parlata, oltre ovviamente a tutte le lingue umane.

La modifica radicale proposta da questo approccio è che tra noi uomini e un’applicazione o un sistema informatico ci sarà un interprete che parla con noi la nostra lingua, e non una lingua informatica.

Intelligenza Artificiale, Assistenti, Agenti e comunicazione informatica

Questo nuovo approccio significa che l’uomo potrebbe delegare al proprio “interprete” personale la comunicazione con gli strumenti informatici. In altre parole, è ora il modello LLM che deve conoscere tutte le “lingue” informatiche, soprattutto per l’accesso e l’utilizzo di applicazioni e dati.

Ma questo non è il modello di base dell’informatica, **ingresso + base dati + elaborazione + uscita**. Visto che l’“interprete” personale dovrebbe poter accedere a qualunque risorsa informatica ed eseguire qualunque azione informatica, deve poter dialogare con qualunque sistema informatico e poter accedere ed interpretare qualunque tipo di dati. Oggi questo è sicuramente un problema tecnico notevole e crescente più si sale nella pila ISO-OSI (o lo stack TCP/IP).

Come potrebbe un modello LLM comunicare con qualunque applicazione o base dati esposta in Internet o comunque resa disponibile al modello?

Si noti che per quanto riguarda questa problematica, non vi è una differenza pratica tra Assistente AI, ovvero un modello in grado di svolgere compiti limitati e solo a seguito di una richiesta umana e sotto il controllo dell’uomo, e un Agente AI, che è in grado di agire indipendentemente dalle esplicite richieste umane per conto dell’uomo. Ovviamente le capacità e possibili conseguenze delle azioni di Assistente e Agente sono molto diverse, ma è comune ad entrambi la necessità di accedere a risorse informatiche, anche solo informazioni, esterne per poter svolgere al meglio i loro compiti.

Per semplicità di esposizione, nel proseguo si farà riferimento solo agli Agenti

AI, ma molto di quanto descritto si applica anche agli Assistenti AI.

Il Model Context Protocol

In questo contesto e a questo scopo, nel 2024 Anthropic ha proposto un protocollo chiamato “Model Context Protocol (MCP)” [Rif. 1] che si propone come “lingua franca” per le comunicazioni dei modelli AI con gli altri sistemi informatici e tra i modelli AI stessi.

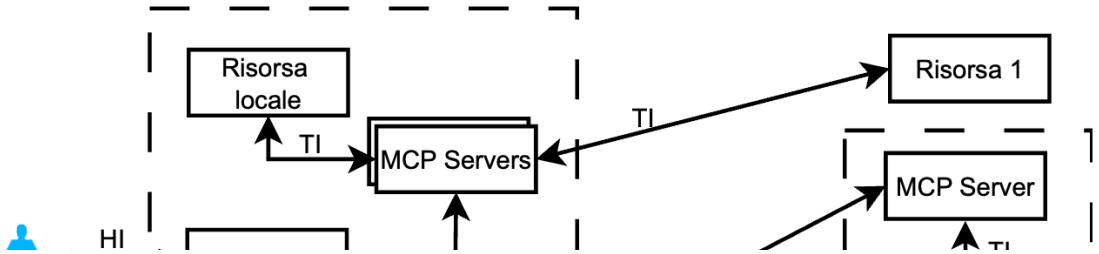
(Nota: questo campo è in rapidissima evoluzione, nuove proposte e protocolli sono presentati frequentemente, si vedano ad esempio il “Agent Communication Protocol (ACP)” di IBM [Rif. 2] che si propone come una evoluzione di MCP, o il “Agent2Agent Protocol” di Google [Rif. 3].)

MCP è un protocollo che adotta lo schema Client-Server: in questo caso il Client MCP è un’applicazione utilizzata dal modello AI che ha necessità di accedere ad una risorsa esterna, mentre il Server MCP espone ai Client MCP servizi di sistemi informatici esterni al modello AI.

Il Client MCP è tipicamente una componente inclusa o agganciata al modello AI stesso; il suo scopo è quello di tradurre la richiesta del modello AI nel linguaggio MCP, e al contrario, tradurre la risposta dal linguaggio MCP al linguaggio del modello AI.

Il Server MCP svolge un ruolo speculare traducendo la richiesta dal linguaggio MCP al linguaggio del sistema informatico o applicazione o altro modello AI, che il modello AI vuole interrogare, e la risposta dal linguaggio del sistema informatico o applicazione al linguaggio MCP.

In figura 1 si presenta un possibile scenario di adozione di MCP.



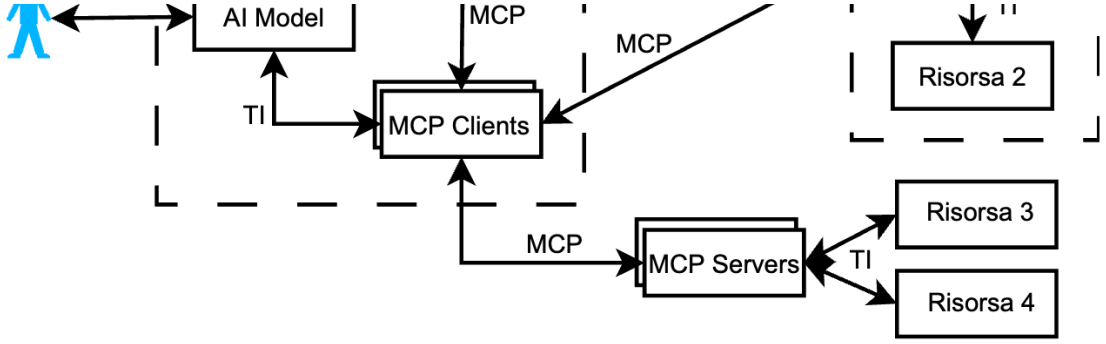


Figura 1: Esempio di diagramma di comunicazioni di MCP (HI=Human Interface, TI=Technical Interface)

In questo scenario, un uomo interagisce con un Agente AI (nel diagramma “AI Model”), tramite una Human Interface (nel diagramma “HI”). Per eseguire il compito affidatogli dall’uomo, l’Agente AI (nella terminologia MCP l’Agente AI è chiamato “MCP Host”) ha necessità di accedere a risorse fornite da altri sistemi informatici, interni e/o esterni al sistema informatico dell’Agente AI. L’Agente AI formula questa richiesta in un proprio linguaggio informatico (nel diagramma “Technical Interface, TI”).

Nella richiesta inviata al proprio Client MCP (ci possono essere anche più di un Client MCP), l’Agente AI può indicare lo specifico sistema informatico o il tipo di sistema informatico che può fornire le informazioni cercate, o che può eseguire la transazione richiesta, e la richiesta stessa. Il Client MCP associato all’Agente AI, traduce la richiesta nel linguaggio MCP e contatta il o i Server MCP che espongono il servizio richiesto.

Il protocollo MCP richiede che ogni connessione tra un Client MCP e un Server MCP sia 1 a 1, ovvero un’istanza di un Client MCP può dialogare in un determinato momento solo con un’istanza di un Server MCP, e viceversa, secondo le attuali best-practices di gestione di micro servizi. Inoltre, la comunicazione MCP è basata sul protocollo JSON-RPC-2.0, un protocollo standard e molto comune, che permette uno scambio strutturato di richieste e risposte, sia localmente (tramite input/output standard), sia verso sistemi remoti (tramite HTTPS e il protocollo “Server-Sent Events – SSE”).

I Server MCP espongono la lista dei propri servizi in un formato standardizzato, in modo che i Client MCP siano in grado di individuare i Server MCP che possono essere utilizzati per eseguire le azioni richieste.

Come indicato in figura 1, i Server MCP possono essere posizionati differentemente nell’infrastruttura di rete: i Server MCP possono essere nella stessa infrastruttura dell’Agente AI, o all’opposto direttamente nell’infrastruttura di una risorsa remota (ovvero gestiti dalla risorsa remota stessa), o essere forniti da una terza parte che media tra gli Agenti AI (e in generale i modelli AI) e altre risorse disponibili in rete.

E’ importante sottolineare che MCP è agnostico rispetto ai servizi e tipi di dati che trasmette: può ugualmente permettere l’accesso a cartelle e file su file-system locali o in Cloud, permettere di eseguire una query su un database, o di eseguire una transazione su un’applicazione, ad esempio un bonifico bancario, e così via.

L’Agente AI non conosce le modalità per eseguire queste azioni, produce solo una richiesta di dati o di transazioni. I Client MCP traducono le richieste dell’Agente AI in un linguaggio universale in modo che qualunque Server MCP sia in grado di comprendere le richieste; un Server MCP in grado di fornire quanto richiesto, traduce le richieste nel linguaggio del sistema informatico a lui connesso e che può fornire i dati o eseguire le transazioni.

Agenti AI, MCP e un nuovo paradigma di sicurezza informatica

Il protocollo MCP, insieme agli Agenti AI, ha la potenzialità di rivoluzionare l’intero ecosistema informatico. Se questo avverrà, il traffico dati non sarà più avviato direttamente dall’uomo ad esempio digitando un URL in un browser o facendo una ricerca su un motore di ricerca, ma gestito interamente dai sistemi informatici stessi. La ricerca online e qualunque azione sui sistemi informatici sarà svolta dall’Agente AI per conto dell’uomo.

A causa di questa rivoluzione, è molto probabile che sarà necessario rivedere interamente tutta l’architettura di sicurezza informatica che molto faticosamente è stata costruita negli ultimi 60 anni.

In questa sede è possibile solo proporre delle considerazioni per illustrare alcuni

dei problemi di sicurezza che potrebbero manifestarsi.

Uno dei pilastri della sicurezza informatica attuale è l’autenticazione degli utenti: siamo ben consci del problema di utilizzare password diverse per ogni servizio, con autenticazione a più fattori su dispositivi diversi, e così via. Come potrà funzionare l’autenticazione di un Agente AI che agisce per conto di una persona? Sarà il nostro personale Agente AI che gestirà tutte le nostre password e relative autenticazioni sui nostri dispositivi informatici e su tutti i servizi informatici che utilizzeremo?

In pratica questo potrebbe voler dire che il nostro personale Agente AI avrà anche la funzione di borsellino delle nostre credenziali e che le gestirà autonomamente. In altre parole, chi avrà accesso al nostro servizio di online banking sarà il nostro personale Agente AI, e noi ci autenticheremo tramite biometria direttamente all’Agente AI, non al servizio di online banking. Password e autenticazione a più fattori potrebbero scomparire, sostituite dalla gestione delle credenziali da parte degli Agenti AI.

Se questo può risultare ottimo per le persone, richiede però che tutta l’architettura e infrastruttura dei servizi informatici garantisca delle caratteristiche di sicurezza nel controllo accessi end-to-end che oggi ancora non abbiamo. Anche se molte tecnologie di autenticazione (e autorizzazione) sicura esistono già oggi, devono essere integrate in questa nuova architettura per garantire che le connessioni ai servizi, anche se mediate dai Client e Server MCP, non possano essere abusate.

Tornando all’esempio precedente, è opportuno notare che rispetto ad oggi in cui ogni persona si autentica direttamente al proprio home banking, in questa nuova architettura sarà il nostro Agente AI ad autenticarsi per conto nostro al servizio di home banking ma non direttamente, solo tramite la mediazione dei servizi MCP.

Considerando ancora i processi di autenticazione e autorizzazione, in questo scenario si presenta con maggiore importanza il problema di garantire l’autenticazione tra MCP Client e MCP Server. In altre parole, come può un MCP Server identificare un MCP Client per evitare che un attaccante impersonifichi un

autentico MCP Client e quindi l’Agente AI e la persona a cui è associato? Questa è una nuova tipologia di furto d’identità in cui un MCP Client si spaccia per un altro con tutte le possibili conseguenze. E vice-versa, come può un MCP Client essere sicuro che il MCP Server a cui si connette è il legittimo MCP Server dell’applicazione o delle informazioni cercate e non un MCP Server civetta o clone malevolo?

Possiamo riassumere queste considerazione con la seguente domanda: come autenticare vicendevolmente Agenti AI e i servizi informatici (inclusi altri sistemi AI) da loro utilizzabili senza coinvolgere direttamente l’uomo? (Per una proposta si veda [Rif. 4].)

Ma autenticazione e autorizzazione non sono le uniche preoccupazioni di sicurezza in questo nuovo ambiente. Visto che gli Agenti AI dovrebbero sostituire l’uomo nell’accesso alle risorse informatiche, si possono considerare alcuni degli attacchi a cui normalmente siamo soggetti e trasporli in questo nuovo ambiente. Va sottolineato che la presenza di Agenti AI, MCP Client e soprattutto di MCP Server, aumenta la superficie di attacco e l’ambito di possibili vulnerabilità. Più componenti ci sono, più componenti possono guastarsi, essere vulnerabili o essere attaccate.

Il primo esempio è quello di un Server MCP compromesso, ad esempio da malware tramite un attacco da Internet (i Server MCP, come qualunque software, possono avere delle vulnerabilità, si veda ad esempio [Rif. 5]). L’attaccante che ha preso possesso del Server MCP può ad esempio accedere alle richieste di ogni Client MCP e quindi violarne la confidenzialità, re-indirizzare le richieste verso servizi malevoli, o allegare malware alle risposte verso il MCP Client. Come può un MCP Client accertarsi che un MCP Server è integro e non compromesso? Bisogna sempre tenere conto che gli MCP Server ricoprono un ruolo cruciale per garantire la confidenzialità e integrità delle informazioni.

E’ interessante considerare anche possibili attacchi e vulnerabilità dell’Agente AI, che in questo scenario prende il ruolo dell’uomo. Infatti un Agente AI, oltre a poter essere soggetto a vulnerabilità e attacchi come una qualsiasi applicazione informatica, per il ruolo che ricopre, può essere oggetto di attacchi concettualmente simili al Social Engineering umano.

Si consideri come esempio il caso in cui, tramite un attacco ad un MCP Server, o ad una applicazione o base dati, o un attacco di tipo man-in-the-middle, un attaccante sia in grado di modificare i dati inviati ad un Agente AI. In questo modo l’attaccante riesce ad eseguire un attacco di Prompt Injection allo scopo di confondere l’Agente AI e fargli eseguire un’azione malevola.

L’attacco può essere “diretto”, tramite l’aggiunta o modifica on-line dei dati di richiesta o risposta per l’Agente AI, o “indiretto”, in cui l’attaccante modifica preliminarmente l’informazione ad esempio in una base dati che poi viene inviata all’Agente AI. Attualmente gli attacchi di Prompt Injection sono in grado di sovvertire i modelli AI Generativi, estraendo informazioni riservate, facendo eseguire azioni errate o malevoli, e riportando informazioni false alla persona di riferimento (gli Agenti AI facendo parte della famiglia dei modelli AI Generativi, sono suscettibili a questi tipi di attacchi).

In questo panorama va anche considerato il caso di modelli AI malevoli utilizzati per attaccare in maniera automatica e sofisticata gli Agenti AI, elevando ancora il livello di minacce che devono essere fronteggiate e mitigate.

Considerazioni finali e i rischi per l’utente umano

MCP, come i protocolli a lui associati e a lui simili, è un protocollo giovane che sicuramente evolverà a livello tecnologico, ma il cui impatto potrebbe risultare molto più di una innovazione tecnica, come accadde anni fa con l’introduzione dei protocolli di base di Internet citati all’inizio di questo articolo. A livello tecnico, c’è ancora molto da capire e sviluppare sia per l’interoperabilità dei servizi, la scalabilità dell’architettura, la gestione della latenza delle comunicazioni dovuta alla presenza di ulteriori servizi di intermediazione come MCP stesso, l’orchestrazione dei servizi, ma anche per l’aumento della complessità per la presenza di ulteriori intermediari nella gestione dei dati che ovviamente possono introdurre dei nuovi punti critici, ulteriori vulnerabilità e superfici d’attacco.

D’altra parte, in questo nuovo scenario, il paradigma informatico di base e la gestione della sicurezza delle informazioni in senso lato potrebbero subire un

cambiamento epocale: l’accesso alle risorse informatiche non sarebbe più gestito direttamente dall’uomo ma delegato da un Agente AI.

Ma chi gestisce gli Agenti AI? Chi ne scrive il codice, decide con quali dati e in quale modo addestrarli? Chi decide quali sono i comportamenti leciti, opportuni, inopportuni e illeciti di un Agente AI? E infine, quali conseguenze potrebbe avere su noi umani questo nuovo paradigma informatico?

Alcuni impatti sull’uomo delle applicazioni di Intelligenza Artificiale Generativa sono già in corso di studio, e i primi risultati sono preoccupanti per l’uomo in quanto indicano una possibile riduzione della capacità della memoria e di alcune funzioni cognitive [Rif. 6], una possibile riduzione della capacità di pensiero critico [Rif. 7], e in generale rischi di un minore sviluppo delle capacità cognitive per i bambini e ragazzi che si affidano alla “Intelligenza Artificiale” piuttosto che svolgere i compiti da soli o al più utilizzando un motore di ricerca online (senza AI).

Questo non dovrebbe sorprendere più di tanto in quanto demandare ad un altro, in questo caso un Agente AI, lo svolgimento di un compito può implicare non imparare a svolgerlo da soli né capire dettagliatamente il suo significato.

Ma se gli Agenti personali AI diventassero anche solo il principale strumento di accesso alle informazioni e ai servizi digitali, non solo costituirebbero un punto estremamente critico per la sicurezza da attacchi malevoli, ma anche un punto privilegiato per monitorare le nostre attività, per accedere ai nostri dati, per farci avere informazioni non solo corrette e aggiornate ma anche potenzialmente fuorvianti o del tutto false (“fake news”) sino a poter costruire intere narrazioni irreali della realtà.

Se un AI Agent è addestrato e/o accede ad informazioni fuorvianti, inesatte o false, le fornirà anche a noi umani potendo costruire una narrazione falsa della realtà con falsi fatti, storie, conoscenze, prodotti, tendenze eccetera. L’impatto sull’uomo potrebbe avere importanti conseguenze tenendo conto che al contempo l’accesso diretto ad altre fonti di informazione potrebbe essere difficile o limitato. Questo è uno scenario quasi Orwelliano, ma non è l’unico sui cui riflettere e legato allo sviluppo dei modelli AI (per un altro aspetto si veda ad

esempio [Rif. 8]).

Tenendo conto che gli AI Agent in futuro saranno in grado anche di auto-addestrarsi, il principale obiettivo futuro è come definire, sviluppare e gestire gli AI Agent e i modelli AI in generale, garantendo che siano veramente “*Trustworthy*”, ovvero che siano degni della fiducia di noi umani e in grado di essere d’aiuto all’uomo e all’umanità senza provocare danni (si veda anche [Rif.9] per una discussione di questa tematica).

Sino ad un paio di anni fa queste tematiche sembravano più attinenti alla fantascienza che alla nostra vita quotidiana, ma le potenzialità di questi recenti sviluppi tecnologici potrebbero portarci a vivere quello che sinora abbiamo considerato come frutto della nostra vivida immaginazione. Ma per fare in modo che questa fantascienza diventi realtà, dobbiamo essere in grado di gestire correttamente questa tecnologia ed evitare che possa ritorcersi contro noi stessi.

Riferimenti

[1] Anthropic, “Model Context Protocol: announcement, specifications, source-code”, 25 Novembre 2024

Claude MCP Protocol Specification

Model Context Protocol Specification

Model Context Protocol

[2] IBM, “Agent Communication Protocol (ACP)”, 2025

Agent Communication Protocol

[3] Google, “Agent2Agent Protocol”, 2025

[4] OWASP, “Agent Name Service (ANS) for Secure Al Agent Discovery v1.0”, 2025

[5] The Hacker News, “Critical Vulnerability in Anthropic’s MCP Exposes Developer Machines to Remote Exploits”, luglio 2025

[6] Nataliya Kosmyna et al., “Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task”, MIT

Andrew R. Chow, “ChatGPT May Be Eroding Critical Thinking Skills, According to a New MIT Study”, Time

[7] Hao-Ping (Hank) Lee et al., “The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers”, Microsoft Research

[8] Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean, “AI 2027”

[9] Andrea Pasquinucci, “L’Intelligenza Artificiale tra sogno e incubo: il sottile confine tra progresso e rischio”, ICTSecurity Magazine, maggio 2025

Profilo Autore

Andrea Pasquinucci

PhD CISA CISSP
Consulente freelance in sicurezza informatica: si occupa prevalentemente di consulenza al top management in Cyber Security e di progetti, governance, risk management, compliance, audit e formazione in sicurezza IT.

Altri Articoli

- L’Intelligenza Artificiale tra sogno e incubo: il sottile confine tra progresso e rischio
- Valutare i rischi Cyber al tempo dell’Intelligenza Artificiale
- Aspetti tecnici degli Adversarial Examples

Adversarial Machine Learning – Aspetti Scientifici

Condividi sui Social Network:

#Agenti AI #AI model #autenticazione #cybersecurity #gestione rischio digitale #Intelligenza Artificiale #large language models (LLM) #MCP #Model Context Protocol #sicurezza informatica

← PRECEDENTE

Come l’IA crea email di phishing perfette: guida completa per riconoscere e difendersi dalle truffe 2025

SEGUENTE →

Account Instagram hackerato: come recuperarlo in 5 passi (Guida 2025)

Ultimi Articoli