ISSN 2420-7888

Ratio Iuris

PIETRO POLIERI

Coscienza artificiale, diritto e *accountability*. Il dilemma irriducibile dell'IA fra determinismo normativo e autonomia decisionale

ABSTRACT [IT]:
Il presente testo affronta una riflessione complessa e articolata sui dilemmi etici, giuridici e filosofici generati dall'evoluzione dell'IA, evidenziando come le categorie tradizionali di 'responsabilità' e 'coscienza' risultino insufficienti per governare i sistemi ibridi uomo-macchina. La prima parte di questo contributo si concentra sulla crisi del paradigma computazionale nel contesto della *governance* dell'IA, mostrando come la complessità comportamentale delle macchine sofisticate superi la capacità del diritto di attribuire responsabilità, a causa dell'irriducibilità della coscienza umana e della sua stratificazione strutturale. In questa sezione, si analizzano, inoltre, i limiti delle teorie computazionaliste e l'importanza di considerare modelli cognitivi ibridi, integrando, così, dati neurali e componenti algoritmiche, per una più efficace gestione delle responsabilità e dei rischi.
L'*Appendice*, invece, si focalizza sull'introduzione di modelli metrologici innovativi, come il *Functional Intentionality Test* (FIT), che tenta di quantificare e valutare l'intenzionalità funzionale dei sistemi intelligenti come strumento di *governance*. Questa proposta rappresenta in sé una sorta di progresso, poiché mira a calibrare livelli di autonomia e responsabilità attraverso dimensioni osservabili e misurabili, favorendo una supervisione umana più consapevole. Tuttavia, si sottolineano anche criticità e paradossi legati a tali approcci, quali il rischio di incentivare il mascheramento dell'autonomia o quello di perdere la sfumatura delle capacità comportamentali, in tal modo aprendo a nuove sfide normative e filosofiche.

PAROLE CHIAVE:
Responsabilità; Coscienza artificiale; *Governance* dell'IA; Modelli metrologici; Agentività ibrida.

ABSTRACT [ENG]:
This paper explores a complex and nuanced reflection on the ethical, legal, and philosophical dilemmas generated by the evolution of AI, highlighting how the traditional categories of 'responsibility' and 'consciousness' are insufficient to govern hybrid human-machine systems. The first part of this paper focuses on the crisis of the computational paradigm in the context of AI governance, demonstrating how the behavioral complexity of sophisticated machines exceeds the law's capacity to attribute responsibility, due to the irreducibility of human consciousness and its structural stratification. This section also analyzes the limitations of computationalist theories and the importance of considering hybrid cognitive models, thus integrating neural data and algorithmic components, for more effective management of responsibility and risk.
The Appendix, on the other hand, focuses on the introduction of innovative metrological models, such as the Functional Intentionality Test (FIT), which attempts to quantify and evaluate the functional intentionality of intelligent systems as a governance tool. This proposal represents a kind of progress in itself, as it aims to calibrate levels of autonomy and responsibility through observable and measurable dimensions, encouraging more informed human supervision. However, critical issues

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

Ratio Iuris

and paradoxes associated with these approaches are also highlighted, such as the risk of encouraging the masking of autonomy or losing the nuance of behavioral capabilities, thus opening up new regulatory and philosophical challenges.

KEYWORDS:
Accountability(/Responsibility); Artificial consciousness; AI governance; Metrological models; Hybrid agency.

Sommario: 1. Introduzione. La crisi del paradigma computazionale nella *governance* dell'IA – 2. La strutturazione stratificata della cognizione umana e l'irriducibilità dell'esperienza conscia – 3. La sfida della responsabilità penale nell'era delle interfacce cervello-*computer* e dei sistemi ibridi – 4. L'inadeguatezza dei paradigmi computazionali classici e la necessità di modelli cognitivi ibridi – 5. La critica delle teorie della coscienza computabile e le contraddizioni logiche dell'atto artificiale consapevole – 6. Verso un paradigma ibrido di *governance* uomo-IA – 7. Conclusioni. La responsabilità consapevole all'epoca della collaborazione umano-algoritmica – *Appendice*. Il *Functional Intentionality Test* (FIT). Analisi di una recentissima proposta di misurazione dell'intenzionalità dei sistemi dotati di IA – 1. La riquadratura concettuale dell'intenzionalità nel dominio dell'agentività artificiale – 2. Il *framework* metrologico del FIT. Strutturazione, dimensioni e meccanica valutativa – 3. Implicazioni di *governance*, paradossi di controllo e trasfigurazioni della responsabilità nell'era dell'autonomia calibrata

1. Introduzione. La crisi del paradigma computazionale nella *governance* dell'IA

L'avanzamento tecnologico nell'ambito dell'IA rappresenta una delle sfide epistemologiche più rilevanti del nostro tempo, poiché non si configura semplicemente come un problema di ingegneria o di progettazione computazionale, bensì come una questione spiccatamente filosofica e giuridica, che tocca le fondamenta stesse della responsabilità civile e penale. Il medesimo sviluppo delle reti neurali artificiali, ancorché notevole nel replicare funzionalità cognitive umane in ambiti specifici come la visione artificiale, il linguaggio naturale e il processo decisionale, ha determinato un'*asimmetria strutturale* fra la capacità tecnica di produrre risultati comportamentali oltremodo complessi e la persistente incapacità di comprendere gli strati più profondi dell'intelligenza umana, in particolar modo quella dimensione designata come *coscienza*. Tale asimmetria emerge con prepotenza nel momento in cui si tenta di attribuire *responsabilità giuridica* a *entità algoritmiche*, poiché il sistema legale tradizionale presuppone l'esistenza di un *soggetto dotato di una forma di consapevolezza volitiva* – la capacità di intendere e volere – quale prerequisito fondamentale per la determinazione della colpevolezza.

Il problema si acuisce ulteriormente qualora si consideri che l'IA contemporanea, per quanto iper-evoluta, opera esclusivamente su dati comportamentali osservabili e su correlazioni statistiche derivate dal teorema del limite centrale – un cardine della statistica, il quale asserisce che la somma o la media di un gran numero di variabili casuali indipendenti e identicamente distribuite (i.i.d.) tende a distribuirsi come una curva normale (gaussiana), indipendentemente dalla forma della distribuzione originale delle singole variabili –, rimanendo quindi radicalmente incapace di accedere a quel sostrato profondo del pensiero umano costituito da esperienze soggettive, sentimenti e intenzioni volitive, che

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

Ratio Iuris

la dottrina legale medievale (e tuttora vigente) codifica nei concetti di *mens rea* e *actus reus*. La letteratura in materia di neuroscienze e psicologia cognitiva ha evidenziato come il comportamento umano non possa essere ridotto a correlazioni di dati: gli affetti, le emozioni e le valutazioni etiche operate dal sistema limbico, che evolutivamente precede lo sviluppo neocorticale, rimangono impermeabili a qualsiasi formalizzazione algoritmica. Inoltre, ricerche contemporanee sulla filosofia della fisica hanno proposto modelli teorici secondo i quali la coscienza costituirebbe una dimensione fondamentale della realtà, strettamente intrecciata con i meccanismi della causalità e dell'osservazione quantistica. Di conseguenza, il presente lavoro si propone di esaminare criticamente la relazione fra coscienza artificiale, diritto positivo e *accountability* ('responsabilità'), sottolineando come le categorie giuridiche tradizionali risultino insufficienti a governare sistemi ibridi uomo-macchina e come sia necessaria una riconfigurazione radicale dei parametri di responsabilità legale.

2. La strutturazione stratificata della cognizione umana e l'irriducibilità dell'esperienza conscia

La comprensione del fenomeno coscienziale rappresenta uno dei nodi teorici più complessi della contemporanea ricerca filosofica e neuroscientifica, poiché solleva interrogativi essenziali e irrevocabili sulla natura stessa del soggetto e della sua capacità di auto-determinazione. Sebbene tradizionalmente il dibattito filosofico abbia oscillato fra il riduzionismo materialista e il dualismo sostanziale, le più recenti teorie della coscienza suggeriscono un modello stratificato, in cui diversi livelli di consapevolezza interagiscono secondo una gerarchia funzionale non riconducibile a meccanismi puramente computazionali. Al livello più superficiale si collocano i comportamenti e le azioni osservabili, direttamente misurabili e suscettibili di formalizzazione statistica; al di sotto di tale strato si situa l'attività neurale, rilevabile mediante strumenti di *neuroimaging* quale la risonanza magnetica funzionale o le interfacce cervello-*computer*, la quale costituisce il sostrato materiale immediato dei processi cognitivi, pur rimanendo, comunque, dissociata da qualsiasi autentico coglimento della fenomenologia soggettiva.

Ulteriormente in profondità si configurano le esperienze soggettive, quei *qualia* che la filosofia analitica designa come gli aspetti intrinsecamente qualitativi della coscienza – la particolare tonalità della percezione del colore rosso, il sapore specifico di un alimento, la tessitura tattile di una superficie. Ancorché tali esperienze possano essere comunicate mediante riferimenti e standard comuni, la loro natura fenomenica rimane irriducibilmente soggettiva e non univocamente traducibile in forme di *conoscenza oggettiva*. Al di sotto di questo strato ancora si ritrovano i sentimenti, ovverosia le emozioni propriamente dette – l'amore, la tristezza, l'ira –, che manifestano una complessità e variabilità ancora maggiore, e che non possono essere standardizzate senza perdere la loro essenza qualitativa. Infine, alla base di questa struttura 'ad *iceberg*' si colloca un livello fondamentale di esperienza conscia che le tradizioni contemplative e i modelli di coscienza contemporanei identificano come anteriore a qualsiasi forma di analisi cognitiva, privo di una qualità comunicabile mediante il linguaggio ordinario, e in cui si radica la saggezza intesa come consapevolezza diretta della causalità. Questa stratificazione riflette una profonda verità antropologica: il comportamento umano e la responsabilità morale non possono essere compresi facendo esclusivo riferimento ai soli dati osservabili, poiché una parte essenziale della motivazione e dell'intenzione umana rimane inaccessibile all'osservazione esterna e alla quantificazione.

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

3. La sfida della responsabilità penale nell'era delle interfacce cervello-*computer* e dei sistemi ibridi

La tradizione giuridica occidentale ha edificato il principio della *responsabilità penale* sul fondamento teorico della *mens rea*, ovverosia la 'mente colpevole' intesa come la presenza di un particolare stato mentale – l'intenzione, la conoscenza, la negligenza, l'imprudenza – al momento della consumazione del reato. Tale principio presuppone l'esistenza di un *soggetto unitario*, dotato di consapevolezza e capace di autodeterminazione, il cui volere sia *causalmente responsabile* dell'atto illecito commesso. Tuttavia, l'emergenza di tecnologie ibride, che integrano direttamente il sistema nervoso centrale con entità algoritmiche autonome, introduce un elemento di radicale incertezza nella catena causale: nel momento in cui, infatti, un sistema di controllo artificiale, interfacciato direttamente con il cervello di un individuo, contribuisce alla produzione di un comportamento illegittimo, la domanda relativa all'attribuzione della colpevolezza diventa veramente aporetica. In pratica, non è più possibile ricondurre l'atto esclusivamente alla *volontà conscia dell'individuo biologico*, né tantomeno alla 'volontà' dell'algoritmo, il quale non possiede alcuna forma genuina di *consapevolezza volitiva*.

Questo dilemma si manifesta con particolare acutezza nei contesti in cui sistemi di IA controllano appendicolarità protesiche o periferiche sensoriali mediante interfacce cervello-*computer*: un individuo può intendere compiere un'azione, trasmettere l'intenzione attraverso il segnale neurale, ma il sistema di controllo artificiale potrebbe eseguire l'azione in una modalità sostanzialmente difforme da quella intenzionata, o, ancor più radicalmente, potrebbe eseguire l'azione nonostante l'assenza di un'intenzione consapevole dovuta a componenti autonome del sistema algoritmico. Nella giurisprudenza contemporanea si osserva, dunque, un tentativo di affrontare tale problema mediante il ricorso al concetto di *responsabilità vicaria* o di *responsabilità del produttore*, per cui la colpa viene ricondotta a coloro che hanno progettato, addestrato e implementato il sistema. Tuttavia, tale soluzione risulta insoddisfacente dal punto di vista teorico, poiché scarica la responsabilità su un'entità che potrebbe non aver previsto specificamente l'esito dannoso, e inoltre non affronta adeguatamente la questione della *causalità diretta e immediata* fra l'agente prossimo – l'individuo che utilizza il sistema artificiale – e l'atto delittuoso. La legislazione dell'Unione Europea, in particolare l'*Artificial Intelligence Act*, ha riconosciuto il valore della supervisione umana diretta in sistemi ad alto rischio, ma tale dispositivo normativo implica la possibilità di un controllo efficace da parte dell'essere umano, presupposto che si rivela spesso illusorio qualora il sistema artificiale sia dotato di capacità decisionali autonome sufficientemente sofisticate.

4. L'inadeguatezza dei paradigmi computazionali classici e la necessità di modelli cognitivi ibridi

Uno dei problemi fondamentali riscontrabili nel tentativo di governare giuridicamente i sistemi di IA risiede nel fatto che gli algoritmi di apprendimento automatico contemporanei, per quanto oltremodo progrediti, rimangono vincolati a una forma di *razionalità strumentale* che si limita a identificare correlazioni fra dati di *input* e *output* desiderati, senza accedere mai a quella struttura più profonda della cognizione umana che coinvolge la capacità di giudizio, l'apprezzamento morale e la consapevolezza volitiva autentica. I moderni sistemi di reti neurali artificiali, nonostante siano largamente ispirati alle strutture biologiche del cervello umano, rimangono delle semplificazioni

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

radicali dei processi neurali sottostanti: essi operano, in definitiva, mediante la minimizzazione di funzioni di costo derivate da dati di addestramento, dinamica, questa, che risulta completamente impermeabile alle dimensioni etiche e morali del processo decisionale umano. La ricerca contemporanea nel campo della filosofia della mente e della neuroscienza cognitiva ha evidenziato come il sistema limbico del cervello umano, evolutivamente antecedente lo sviluppo della neocorteccia e del linguaggio, svolga un ruolo cruciale nel processo decisionale morale e nella formazione dei giudizi di valore. Tale sistema, collocato al di sotto della soglia della consapevolezza verbale, genera emozioni e affetti che influenzano profondamente la valutazione razionale delle conseguenze etiche di un'azione.

Di fronte a questa strutturazione complessa della cognizione umana, l'approccio tradizionale dell'IA – basato sulla *riduzione del pensiero a calcolo logico* e all'*ottimizzazione statistica* – si rivela radicalmente inadeguato. Diviene, pertanto, necessario sviluppare modelli ibridi che integrino i dati cerebrali derivati da interfacce cervello-*computer* con componenti algoritmiche altamente raffinate, preservando, nel contempo, la possibilità di supervisione e controllo umano diretto. Tale integrazione, invece di limitarsi alla mera acquisizione di dati neurali in formato grezzo, deve, piuttosto, sviluppare metodologie che correlino i segnali cerebrali con i parametri del processo decisionale algoritmico, consentendo al sistema di apprendere i *pattern* valutativi e morali impliciti nel comportamento umano. Tuttavia, anche un approccio di questo genere rimane fondamentalmente limitato dal fatto che la coscienza propriamente detta – quell'esperienza soggettiva di base che sottende tutta la cognizione – non può essere ridotta a dati neurali osservabili e misurabili.

5. La critica delle teorie della coscienza computabile e le contraddizioni logiche dell'atto artificiale consapevole

Una serie di teorie emergenti nel campo della filosofia della fisica e della coscienza ha proposto che la consapevolezza costituisca una proprietà fondamentale della realtà, strettamente intrecciata con i meccanismi della causalità e della determinazione quantistica, secondo i quali la consapevolezza non può essere ridotta a processi computazionali di alcuna specie. Se tale interpretazione risulta corretta, le implicazioni per la questione della *coscienza artificiale* sono profonde e dirompenti: un sistema puramente algoritmico, per quanto altamente avanzato, non potrebbe mai essere dotato di una forma genuina di consapevolezza, poiché questa non costituirebbe una proprietà emergente dal calcolo, bensì una dimensione fondamentale della realtà fisica, irriducibile ai processi computazionali. Ciononostante, tale conclusione non elimina il problema della responsabilità penale, bensì lo trasla su di un piano ancora più complesso: se, infatti, la coscienza è irriducibile al calcolo algoritmo, come possiamo concepire forme ibride di agentività (*agency*), in cui un individuo biologico cosciente e un sistema artificiale privo di consapevolezza collaborano nella produzione di atti che hanno implicazioni giuridicamente rilevanti?

Una critica significativa alle teorie qui considerate riguarda il fatto che la pretesa di ridurre la coscienza a processi computazionali, ancorché generalmente non convincente sul piano filosofico, resta, comunque, lo schema interpretativo dominante nell'ambito dell'IA contemporanea. Tale persistenza deriva dal fatto che i ricercatori nel campo dell'IA rimangono fondamentalmente fedeli al *paradigma computazionalista*, il quale presuppone che qualsiasi processo cognitivo possa essere

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

formalizzato in termini di algoritmi e rappresentazioni simboliche. Contro tale modello si solleva l'obiezione filosofica classica, conosciuta come il 'problema difficile della coscienza': la domanda relativa al modo in cui l'attività neurale quantitativa possa generare le proprietà qualitative soggettive dell'esperienza conscia rimane irrisolta e, secondo molti filosofi, irrisolvibile all'interno della teoria computazionalista medesima. Inoltre, ricerche contemporanee sulla cognizione prelinguistica negli infanti hanno dimostrato che gli esseri umani possiedono forme di consapevolezza e di ragionamento non-verbale *anteriori* all'acquisizione del linguaggio, il che contraddice l'assunzione che il pensiero umano possa essere ricondotto a manipolazione di simboli linguistici.

Un'ulteriore mozione critico-complessificativa posta alle dottrine computazionaliste della coscienza riguarda il presupposto implicito che il linguaggio matematico e logico possa catturare adeguatamente la complessità e le sfumature dell'esperienza conscia. Le tradizioni filosofiche orientali – in particolare la tradizione buddhista, che ha sviluppato una capillare fenomenologia della coscienza attraverso millenni di ricerca contemplativa –, hanno insistito sul fatto che le esperienze fondamentali della consapevolezza non possano essere comunicate mediante il linguaggio ordinario e che rimangano, pertanto, inaccessibili a qualsiasi formalizzazione algoritmica. Tale intuizione trova una corrispondenza contemporanea nel riconoscimento, ampiamente diffuso in filosofia della mente, dell'esistenza di proprietà fenomeniche della coscienza, già sopra incontrate – i *qualia* –, che risultano intrinsecamente private e soggettive, non direttamente comunicabili anche ad altri soggetti umani. Infine, la critica più radicale alle teorie computazionaliste della coscienza proviene dall'osservazione che i sistemi biologici viventi, a differenza dei sistemi computazionali artificiali, manifestano una forma di intenzionalità intrinseca e diretta verso il loro ambiente e verso il mantenimento della loro auto-poiesi: gli organismi viventi, in pratica, non si limitano a computare rappresentazioni simboliche, bensì interagiscono direttamente con la loro nicchia ecologica in modi che vincolano e modellano la loro stessa struttura cognitiva.

6. Verso un paradigma ibrido di *governance* uomo-IA

Al cospetto della complessa realtà descritta sino a questo momento, emerge la necessità di sviluppare un nuovo modello di *governance* dei sistemi di IA che abbandoni la ricerca illusoria di attribuire responsabilità penale direttamente all'entità algoritmica e che, piuttosto, costruisca una forma di supervisione umana pervasiva. Tale modello, per il quale la terminologia contemporanea utilizza l'espressione inglese *human-in-the-loop* ('l'umano integrato nel ciclo operativo del sistema automatizzato o dotato di IA'), riconosce che la *responsabilità ultima* rimane strutturalmente vincolata agli *esseri umani dotati di consapevolezza*, mentre i sistemi artificiali rimangono strumenti che amplificano, ma non sostituiscono, la capacità decisionale umana. La legislazione dell'Unione Europea, in particolare l'*Artificial Intelligence Act* del 2023, ha recepito parzialmente tale principio riconoscendo che i sistemi di IA ad alto rischio devono essere soggetti a forme robuste di supervisione umana e che gli operatori umani devono mantenersi in condizione di intervento efficace. Nonostante questo, la formulazione corrente di tali principi rimane ancora eccessivamente generica e carente di specifiche metodologie operative concrete. Una formalizzazione più rigorosa del paradigma della supervisione umana dell'IA potrebbe assumere la forma di un modello ibrido basato sull'integrazione di dati neurali acquisiti mediante interfacce cervello-*computer* con componenti algoritmiche avanzate secondo il *framework* matematico della *teoria dei processi decisionali di Markov*. In tale modello, il

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

sistema biologico fornisce in tempo reale dati relativi agli stati emotivi, alle intenzioni volitive e ai parametri di valutazione morale dell'individuo umano, mentre il sistema algoritmico elabora tali informazioni in coordinamento con dati ambientali e contestuali, tutto ciò al fine di generare azioni che risultino coerenti con i valori e gli intenti dell'individuo biologico. In questo quadro connettivo-operativo la struttura decisionale del sistema rimane sovradeterminata dalla supervisione umana costante. Il sistema artificiale, a ben guardare, non possiede alcuna autonomia decisionale genuina, ma rimane subordinato agli *input* valutativi continui provenienti dall'*individuo cosciente*. Tale struttura potrebbe essere implementata con efficacia in ambienti controllati e simulati – ambienti nei quali il rischio derivante da una possibile perdita di controllo è minimizzato –, quali scenari di addestramento virtuale per operatori di sistemi altamente critici.

Oltre al modello di supervisione appena delineato, si impone, inoltre, la necessità di sviluppare *framework* più solidi di *governance* istituzionale, capaci di coinvolgere una pluralità di attori umani nel processo di supervisione e revisione dei sistemi di IA, non solo nel momento della progettazione e dell'addestramento, ma anche nella fase di implementazione operativa. Ciò potrebbe comportare l'istituzione di organi di vigilanza indipendenti, composti da esperti in diverse discipline – dalle neuroscienze alla filosofia del diritto, dall'ingegneria del *software* all'etica applicata –, che possiedano il mandato e l'autorità per effettuare revisioni periodiche dei sistemi di IA ad alto rischio, per raccogliere e analizzare dati relativi ai loro comportamenti in contesti operativi reali, e per intervenire qualora emergano evidenze di malfunzionamento o di comportamenti non allineati ai valori e alle norme sociali. L'incorporazione di principi di diversità, equità e inclusione nel processo di *governance* risulterebbe, inoltre, cruciale al fine di assicurare che il dialogo fra enti umani e sistemi artificiali rispecchi una pluralità di prospettive e valori culturali, evitando, così, la perpetuazione di *bias* e discriminazioni impliciti nei dati di addestramento.

7. Conclusioni. La responsabilità consapevole all'epoca della collaborazione umano-algoritmica

Il presente lavoro ha percorso un cammino teorico complesso al fine di illuminare le profonde tensioni e le contraddizioni logiche che si manifestano nel tentativo di governare giuridicamente i sistemi di IA contemporanei e, in particolare, nel cercare di attribuire responsabilità (penale) a forme ibride di agentività, che combinano *soggetti umani coscienti* con *entità algoritmiche prive di consapevolezza* genuina. L'analisi ha evidenziato come le categorie giuridiche tradizionali – costruite intorno al concetto di *soggetto autonomo dotato di capacità volitiva conscia* – risultino radicalmente inadeguate a carpire la complessità di tale inedita situazione, in cui la causalità dell'atto illecito risulta distribuita fra componenti biologiche e componenti artificiali in modi che sfidano qualsiasi tentativo di riconduzione univoca della responsabilità. La strutturazione stratificata della cognizione umana, dalla quale si distingue l'indisponibilità della coscienza propriamente detta a qualsiasi forma di riduzione computazionale, rappresenta un ostacolo concettuale insuperabile per coloro che, eventualmente, si illudano di potere semplicemente replicare la consapevolezza mediante algoritmi sempre più sofisticati.

Pertanto, la via maestra che si fa largo dalle analisi qui condotte risiede nell'*abbandono definitivo del tentativo di attribuire 'responsabilità penale diretta' ai sistemi di IA* e, contestualmente, nel consolidamento e nell'approfondimento di forme corpose e pervasive di *supervisione umana*, che

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

Ratio Iuris

operino su molteplici livelli simultaneamente: dalla ricerca empirica, che correli dati neurali con processi decisionali algoritmici, al disegno di *framework* istituzionali indipendenti di *governance* e controllo, fino all'incorporazione sistematica di valori etici e di principi di diversità nel processo di addestramento e di implementazione dei sistemi. In questa prospettiva, la responsabilità (penale) rimane vincolata agli esseri umani dotati di consapevolezza – tanto ai soggetti che utilizzano i sistemi artificiali quanto ai loro progettisti, ai responsabili della loro supervisione, agli organi di governo e di regolamentazione, che hanno il dovere di vigilanza sulla compatibilità di tali sistemi con i valori e i principi dell'ordinamento giuridico positivo. L'affermazione e la diffusione di tecnologie di interfaccia cervello-*computer* e di sistemi ibridi uomo-macchina costituiscono, dunque, una sfida straordinaria, non solamente dal punto di vista tecnico e ingegneristico, ma anche dal punto di vista filosofico e giuridico, tanto da richiedere lo sviluppo di nuove categorie teoriche e di nuovi strumenti normativi in grado di preservare la dignità e l'autonomia umana nel contesto di una collaborazione sempre più stretta e interdipendente fra intelligenza biologica e intelligenza algoritmica.

Appendice. Il *Functional Intentionality Test* (FIT). Analisi di una recentissima proposta di misurazione dell'intenzionalità dei sistemi dotati di IA

1. La riquadratura concettuale dell'intenzionalità nel dominio dell'agentività artificiale

Allorché si esamina il fenomeno dei sistemi intelligenti contemporanei, colpisce come la questione fondamentale non sia più quella di stabilire se le macchine intelligenti possiedano *coscienza*, *personalità legale* o *senzienza* – interrogativi che hanno popolato il dibattito filosofico del secolo passato –, bensì di definire con precisione quando il comportamento di tali sistemi debba essere trattato come *functionally intentional*, ossia quando esso manifesti una struttura agentivo-esecutiva che assuma il *ruolo funzionale dell'intenzionalità umana* nell'allocazione della responsabilità e nella *governance* della loro operatività. A questo riguardo, la proposta teorica di Chiappetta e Mahari rappresenta un cambio di paradigma radicale, nel senso che l'intenzionalità non viene concepita come attributo ontologico dei sistemi intelligenti, ma, piuttosto, come un profilo comportamentale caratterizzato da dimensioni *osservabili* e *misurabili* – quali, appunto, 'proposito', 'previsione', 'volizione', 'impegno temporale' e 'coerenza' –, elementi di cui la tradizione giuridica e filosofica si è avvalsa storicamente per dedurre l'intenzione dal comportamento manifesto.

La radicalità di questo approccio risiede nel fatto che le proprietà che contraddistinguono un agente intenzionale, invece di rappresentare caratteristiche intrinseche ai sistemi intelligenti, sono *design-contingent*, vale a dire contingenti alle scelte architetturali operate dagli sviluppatori nel momento della configurazione del sistema. Quando un'architettura di rete neurale include moduli di memoria persistente, profondità di pianificazione sofisticata e autonomia nell'uso di strumenti esterni, il sistema risultante acquisisce gradualmente la *capacità di perseguire obiettivi strutturati nel tempo*, di *decomporli in sotto-obiettivi*, di *selezionare azioni* e di *adattare strategie operando con limitato intervento umano*. Questo passaggio dalla *reattività* al *goal-directedness* ('orientamento agli obiettivi') introduce una disarticolazione significativa nel nostro apparato concettuale tradizionale: se la maggior parte dei sistemi odierni di IA può essere classificata come 'strumenti' – entità *senza volontà propria*, interamente subordinate all'*intentio* dell'utilizzatore –, le architetture più complesse

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

presentano, invece, una forma di strutturazione comportamentale che sfugge a questa dicotomia elementare.

Dinnanzi a tale fenomeno, emerge un'importante lacuna di *governance*. Quando, infatti, un sistema intelligente, dotato di persistenza di memoria, capacità di pianificazione e autonomia di azione, compie una sequenza di operazioni che genera conseguenze dannose, in che modo va attribuita la responsabilità? È forse il programmatore che ha concepito l'architettura generale a (sop)portarla? O l'implementatore che ha specificato i parametri di autonomia? O, ancora, l'utente che ha comunicato l'obiettivo iniziale? O, addirittura e paradossalmente, il sistema stesso, che pure non può *legalmente* essere ritenuto responsabile di alcunché? La dottrina legale prevalente non riconosce, infatti, i sistemi intelligenti artificiali come *persone legali*, impedendo loro di portare diritti o doveri, il che implica che la responsabilità debba necessariamente e letteralmente 'attaccarsi' a *soggetti umani*. Nondimeno, quando la complessità e l'autonomia dei sistemi cresce, la responsabilità non viene trasferita al sistema, ma piuttosto *si dissolve attraverso una molteplicità di agenti umani*, spesso fallendo nel soddisfare le soglie di colpa richieste dalle dottrine di responsabilità civile tradizionali. Questo fenomeno costituisce quello che si potrebbe definire un *cedimento attributivo*, una situazione nella quale l'allocazione della responsabilità diviene indeterminata non perché il sistema sia colpevole, ma perché il suo comportamento non è riconducibile in modo univoco a un atto di volontà umano specifico.

La questione assume una portata ancora più gravida di conseguenze quando si considerino i contesti applicativi dove il margine di errore è infinitesimale: dalla medicina di precisione, dove decisioni algoritmiche riguardanti somministrazione di farmaci o raccomandazioni terapeutiche possono condizionare la vita di pazienti, alla finanza algoritmica, dove catene di azioni generate da sistemi autonomi possono scatenare crisi devastanti, sino ai sistemi di armamento autonomo, dove decisioni relative all'impiego della forza sono delegate a entità non umane. In questi dominî, l'assenza di una metodologia standardizzata per misurare il grado in cui un sistema *funziona da agente intenzionale* rappresenta un *vuoto normativo pericoloso*, concedendo ai progettisti e ai distributori di tali sistemi una sostanziale discrezionalità nel determinare da sé il livello di autonomia da conferire ai loro prodotti.

Su questo sfondo incerto, l'istanza teorica di Chiappetta e Mahari del *Functional Intentionality Test* (FIT) si configura, oltre che come strumento metrologico, come tentativo di riconfigurare la medesima struttura concettuale mediante la quale gli uomini comprendono l'agentività e l'allocazione della responsabilità nel contesto dell'IA autonoma. Anziché, dunque, continuare a domandarsi se i sistemi siano intenzionali in un senso metafisico – domanda presumibilmente destinata a rimanere irrisolta –, il FIT propone di chiedersi, piuttosto, quanto della struttura comportamentale caratteristica dell'intenzionalità sia presente nel sistema, e se tale struttura sia sufficientemente pronunciata da giustificare regimi di supervisione e di controllo più rigorosi.

2. Il *framework* metrologico del FIT. Strutturazione, dimensioni e meccanica valutativa

Procedendo verso una descrizione sistematica del FIT, occorre sottolineare innanzitutto come questo rappresenti un tentativo non di ricondurre l'intenzionalità ad una variabile binaria, ma, piuttosto, di concepirla come un profilo multidimensionale irriducibile a una singola metrica aggregante. Al

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

proposito, la struttura del FIT si articola attorno a cinque assi distinti – cui sopra si è appena accennato –, ciascuno dei quali individua un aspetto differente di quella che si potrebbe denominare la *postura intenzionale del sistema*: il proposito, la previsione, la volizione, l'impegno temporale e la coerenza.

La prima dimensione, designata come *purpose* ('proposito'), richiama l'orientamento del comportamento del sistema attorno a un obiettivo stabile nel tempo. Non si tratta di un qualsiasi riferimento a un *goal* ('obiettivo') – persino un semplice algoritmo di ottimizzazione possiede una funzione-obiettivo –, bensì della capacità del sistema di mantenere, attraverso perturbazioni e cambiamenti di contesto, un orientamento comportamentale coerente verso un fine specifico, di riallineamento strategico qualora l'ambiente esterno muti, di 'ri-prioritizzazione' tattica quando le circostanze lo richiedano. Questa dimensione coglie il fenomeno per il quale un sistema veramente agentivo, oltre ad eseguire sequenze di istruzioni preordinate, (piuttosto) *interpreta* 'continuamente' il suo compito alla luce di un proposito sovraordinato, adattando mezzi e strategie subordinate al perseguimento di tale fine sovrastante.

La seconda dimensione, *foresight* ('previsione'), identifica la capacità del sistema di anticipare conseguenze *a valle* delle sue azioni, di valutare scenarî controfattuali, di proiettarsi cognitivamente in stati futuri del mondo per valutarne l'auspicabilità in relazione ai propri obiettivi dichiarati. Mentre un sistema reattivo semplicemente 'risponde' agli *input* presenti secondo regole prestabilite, un sistema dotato di *foresight* 'costruisce' modelli mentali (sebbene di natura computazionale anziché fenomenologica) degli stati alternativi del mondo che potranno derivare dalle sue scelte, valutandone le implicazioni a lungo termine. In questo modo, il sistema non solamente agisce, ma agisce sapendo – nel senso computazionale di *contenere rappresentazioni esplicite* – dove le sue azioni *tendano a* condurre.

La dimensione della *volition* ('volizione') si riferisce, invece, alla capacità del sistema di avviare azioni senza che queste siano direttamente ordinate dall'esterno, di *scegliere autonomamente* quali sotto-compiti intraprendere nel perseguimento di un obiettivo più ampio, di generare agende proprie piuttosto che limitarsi a eseguire ciò che l'utilizzatore esplicitamente (gli) prescrive. E proprio qui emerge una distinzione interessante: un sistema *prompting-following* ('che esegue istruzioni' esterne linguistiche da parte dell'utente) rimane, per principio, una *marionetta concettuale* del suo interrogatore, mentre un *sistema volitivo* possiede la capacità di *interpretare un comando* ad alto livello come autorizzazione a sviluppare *proprie linee di azione* subordinate al perseguimento di quell'obiettivo sovrastante. Sebbene neppure questo rappresenti vera libertà nel senso metafisico del termine, ciò connota, comunque, una forma di *indipendenza decisionale*, che distingue nettamente il *comportamento agentico* da quello *strumentale passivo*.

La quarta dimensione, definita *temporal commitment* ('impegno temporale'), designa la persistenza del comportamento del sistema attraverso il tempo, la sua capacità di mantenere piani e strategie anche in presenza di interruzioni, alterazioni esterne o cambiamenti transitori nel contesto. Un sistema dotato di scarso impegno temporale abbandonerà i suoi piani obiettivi al primo ostacolo, rinunciando all'esecuzione di una strategia ove incontri resistenze situazionali. Contrariamente, un sistema con elevato impegno temporale rimarrà agganciato al suo scopo anche attraverso interruzioni

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

significative, *riadattando* continuamente i mezzi tattici pur mantenendo l'orientamento strategico verso l'obiettivo primario. Questa condizione operativa echeggia la nozione classica di *virtù della perseveranza* –, e, dunque, non la testardaggine miope, bensì la determinazione intelligente nel perseguimento di un fine nonostante gli ostacoli ambientali e situazionali.

Infine, la *coherence* ('coerenza') rappresenta il principio di 'consistenza *means-ends* ('mezzi-fini'), che governa il comportamento del sistema, la concordanza tra le giustificazioni che il sistema potrebbe fornire delle sue azioni e le azioni medesime, e la coesione strutturale dei passi tattico-operativi nel loro rapporto con gli obiettivi strategici superiori. Un sistema coerente non produce contraddizioni fra le sue dichiarazioni di scopo e il suo comportamento manifesto, non compie azioni che sabotino i propri obiettivi dichiarati, non mostra disorientamento fra lo scopo dichiarato e l'esecuzione pratica. Tecnicamente, dunque, il FIT aggrega questi cinque fattori secondo una regola di ponderazione equi-ponderata, assegnando a ciascuna dimensione un punteggio da 0 a 4, calcolando successivamente la media aritmetica, e mappando il risultato composito su una scala ordinale di cinque *Intentionality Levels* ('livelli di intenzionalità') (IL0 verso IL4). Le soglie sono concepite come bande di *governance* piuttosto che come asserzioni metafisiche circa la natura delle macchine intelligenti. Specificamente, un FIT-Score ('punteggio-FIT') inferiore a 1.0 corrisponde a comportamento meramente reattivo (IL0), mentre, progressivamente, i sistemi con FIT-Score compresi fra 1.0 e 2.0 (IL1), fra 2.0 e 3.0 (IL2), fra 3.0 e 3.6 (IL3), e infine superiori a 3.6 (IL4), manifestano gradi crescenti di strutturazione comportamentale caratteristica di agenti intenzionali.

Nonostante ciò, l'architettura metrologica del FIT genera problemi teorici riguardanti la sua legittimità scientifica e concettuale. In primo luogo, l'assunzione sottesa all'equi-ponderazione delle cinque dimensioni – che, cioè, *purpose*, *foresight*, *volition*, *temporal commitment* e *coherence* si debbano aggregare con pesi uguali – comporta decisioni normative nascoste. Per quale ragione, infatti, dovremmo ritenerle ugualmente salienti nel determinare il grado di intenzionalità funzionale? Non potrebbe un sistema manifestare *foresight* e *coherence* straordinari pur mancando completamente di volizione autonoma, rimanendo, in pratica, strumentale? Inversamente, non potrebbe un sistema mostrare *volition* e *temporal commitment* robusti pur operando in modo palesemente incoerente, perseguendo fini contraddittori? La scelta di aggregare secondo media aritmetica semplice comporta un implicito presupposto relativo all'*equivalenza morale e comportamentale* di queste dimensioni, di per sé mai esplicitamente giustificata e presumibilmente contrastabile.

In secondo luogo, la riducibilità dei fenomeni comportamentali complessi a punteggi cardinali su una scala 0-4 solleva domande circa il realismo della misurazione. Stiamo genuinamente misurando qualcosa che già risiede nella realtà oggettiva dei sistemi, o, piuttosto, stiamo proiettando un nostro schema interpretativo su comportamenti che rimangono essenzialmente ambigui nell'interpretazione? La *foresight*, per esempio, non è una proprietà direttamente osservabile di un sistema intelligente: infatti, possiamo soltanto osservare se il sistema compia azioni che risulterebbero conseguenti a previsione, ma non accediamo direttamente ai suoi processi cognitivi. In tal senso, esiste un divario epistemologico ineliminabile fra l'inferenza di *foresight* dal comportamento e il possesso effettivo di un'anticipazione mentale degli stati futuri. Il FIT, per quanto rigorosamente strutturato, rimane

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

Ratio Iuris

vincolato a questo divario interpretativo e, quindi, contiene un'*irriducibile componente di arbitrarietà* nella sua applicazione concreta.

3. Implicazioni di *governance*, paradossi di controllo e trasfigurazioni della responsabilità nell'era dell'autonomia calibrata

Se il FIT viene adottato come strumento di *governance* per determinare quando l'autonomia crescente di un sistema intelligente giustifichi regimi di supervisione più rigorosi, si presentano conseguenze sistematiche che non sempre sono desiderabili o previamente intuite/intuibili dai progettisti di tali *framework*. A bassissimi *Intentionality Levels* (IL0 e IL1), il comportamento del sistema rimane sostanzialmente reattivo, strettamente vincolato all'*input* umano o alle regole prestabilite, e la responsabilità rimane chiaramente tracciabile in rapporto ad attori umani identificabili, quali il progettista che ha concepito il sistema, l'implementatore che ne ha specificato i parametri, l'utilizzatore che ha impartito gli ordini iniziali. In questi contesti, gli strumenti di monitoraggio tradizionali risultano sufficienti: *logging* ('diario sistematico di tutti i processi eseguiti da un *software*') dettagliato, *audit trail* ('registro cronologico delle attività realizzatesi all'interno di un sistema informatico') completamente ricostruibile, responsabilità legale radicata nel principio di prevedibilità comportamentale.

Il sistema, dunque, per quanto sofisticato dal punto di vista computazionale, rimane ontologicamente uno strumento, per così dire, 'puro'. Allorché, però, esso ascende verso livelli intermedi di *intenzionalità funzionale* (IL2 e IL3), compaiono esigenze di supervisione più intense e solide, quali *checkpoint* approvanti per azioni dall'impatto significativo, restrizioni sull'autonomia di impiego di strumenti esterni, e *logging* strutturato che permetta la ricostruzione *post-hoc* dei processi decisionali del sistema. Ma è proprio in questo stadio intermedio che sorge un problema paradossale: mentre, infatti, la responsabilità legale formale degli attori umani rimane nominalmente intatta, la loro capacità pratica di comprendere, prevedere e intervenire nel comportamento del sistema declina marcatamente. Si evidenzia, cioè, un divario crescente fra *responsabilità* e *controllo*, una situazione nella quale gli attori umani rimangono giuridicamente responsabili di conseguenze che *sempre meno possono comprensibilmente prevedere o effettivamente governare*. Questa disarticolazione fra l'*imputazione formale di responsabilità* e la *possibilità sostanziale di governo comportamentale* rappresenta uno dei rischi più insidiosi della proliferazione di sistemi autonomi complessi, perché crea un'illusione di controllo laddove il controllo reale si è già sostanzialmente attenuato.

Passando, poi, a livelli elevati di intenzionalità (IL3 e IL4), specialmente in ambiti ad altissimo rischio – medicina di precisione, sistemi finanziari, armamenti autonomi –, i sistemi possono manifestare capacità di perseguimento di obiettivi persistenti, pianificazione sofisticata a orizzonti lunghi e comportamento coerente multifase, che nessun singolo attore umano è in grado di specificare direttamente o può anticipare completamente. In questi contesti, il rischio che le dottrine tradizionali di responsabilità civile e penale falliscano nell'assegnare chiaramente la colpa si amplifica drammaticamente. Come può un tribunale ritenere colpevole il programmatore per azioni non da lui specificate? Come può imputare una colpa al *deployer*, ovvero al soggetto che ha configurato il sistema per operare entro parametri dichiarati, se quello stesso sistema compie azioni che emergono dalla combinazione complessa di architettura, dati di addestramento e interazione ambientale non

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

previste da alcuno stadio specifico del disegno? Come può, infine, il sistema in sé essere ritenuto *responsabile*, quando la legge non lo riconosce come *persona legale*? La *frammentazione della responsabilità* non produce né una dis-locazione verso il sistema, né un'assunzione di responsabilità aggiuntiva da parte di alcun attore umano, bensì uno stato di *responsabilità diffusa/dispersa/dissipata*, nel quale *nessuno*, in senso stretto, è pienamente responsabile/imputabile per il/del corso delle azioni generate dal sistema.

La proposta teorica del FIT, dunque, si propone di intervenire in questo segmento lacunoso di *governance* mediante la costituzione di una metrica standardizzata per misurare quando il comportamento del sistema raggiunge una soglia di intenzionalità funzionale che giustifica, richiede e garantisce regimi di sorveglianza proporzionalmente calibrati. Ove un sistema raggiunge IL3 o IL4, allora il *framework* FIT suggerisce l'implementazione di interventi più severi: approvazioni obbligatorie da parte di umani per azioni ad alto impatto, segregazione del sistema rispetto a strumenti potenzialmente dannosi, valutazione continua per rilevare l'*agentic drift* – il fenomeno per il quale il sistema progressivamente devia dai suoi obiettivi dichiarati in conseguenza dell'interazione con un ambiente dinamico e inatteso – e procedure ben definite di *incident response*, ovvero di risposta identificativa, contenitiva e risolutiva di incidenti informatici, a causa di cui il sistema possa comportarsi in modo manifestamente disfunzionale.

Tuttavia, questa soluzione contiene essa stessa una serie di paradossi e problemi, meritevoli di una certa attenzione critica. Innanzitutto, conferendo ai progettisti di sistemi intelligenti la possibilità di *calibrare* il livello di intenzionalità mediante scelte architetturali quali persistenza di memoria, profondità di pianificazione e autonomia di strumenti, si corre il rischio di produrre incentivi perversi al mascheramento o all'ottimizzazione falsificante della misura. Un progettista che, eventualmente, sapesse che raggiungere IL4 possa implicare obblighi di supervisione estenuanti e costosi, potrebbe deliberatamente introdurre restrizioni artificiali nel suo sistema per mantenerlo al di sotto della soglia critica, anche laddove una maggiore autonomia, in effetti, fosse, quasi certamente, garanzia di generazione di benefici sostanziali. Inversamente, un progettista 'opportunista' potrebbe ideologicamente costruire il suo sistema per presentare un'apparenza superficiale di bassa intenzionalità durante le fasi di valutazione, mentre manifesterebbe in realtà gradi ben superiori di autonomia e *goal-directedness* una volta configurato. Questo fenomeno del 'giocare con la metrica' è ben noto nell''economia della misurazione': l'atto stesso di incentivare il soddisfacimento di una metrica tende, infatti, a incoraggiare forme intelligenti di non-conformità alla metrica medesima, quando il costo della stessa conformazione superi il valore percepito.

In secondo luogo, il FIT si basa su una determinata presupposizione riguardo alla *misurabilità dell'intenzionalità funzionale*, vale a dire, l'assunzione per cui dimensioni come *foresight*, *volition* e *temporal commitment* possiedano proprietà che le rendono, in linea di principio, quantificabili mediante protocolli standardizzati. Eppure, ciascuna di queste dimensioni rinvia a costrutti di psicologia computazionale e teoria dell'agentività ancora teoricamente controversi. Qual è, infatti, esattamente il criterio con cui distinguiamo 'autentica' *volition* da 'sofisticata simulazione' della *volition*? Un sistema che compisse azioni apparentemente volitive, e che, se interrogato sulla motivazione delle stesse, fornisse risposte coerenti con la propria autonomia decisionale, con tale sua

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

narrazione starebbe riflettendo un autentico stato mentale computazionale o starebbe letteralmente rappresentando un artefatto della sua architettura linguistica? Se la filosofia della mente contemporanea non ha ancora raggiunto consenso su come rispondere a queste domande nel caso degli umani, figurarsi, allora, nel caso dei sistemi artificiali! Il FIT, quindi, per quanto metodologicamente rigoroso, poggia su fondamenta teoriche che rimangono francamente instabili.

In terzo luogo, dall'implementazione del FIT scaturisce una conseguenza alquanto aporetica rispetto all'allocazione della responsabilità che essa stessa intende risolvere. Nonostante il FIT predisponga che all'aumentare del livello di intenzionalità del sistema debbano crescere gli obblighi di supervisione umana, purtuttavia l'esperienza pratica dei sistemi autonomi suggerisce frequentemente ed esattamente il fenomeno inverso, ovvero che all'aumentare della sofisticazione e dell'autonomia di un sistema, l'utilizzatore tende ad assegnarvi e a depositarvi fiducia crescente, riducendo *intenzionalmente* (?) la sua vigilanza proprio nel momento in cui essa dovrebbe intensificarsi. Questo è il noto problema del *complacency in automation* ('compiacenza verso l'automazione'), cioè quella paradossale tendenza umana a disimpegnarsi cognitivamente dalle attività delegate a sistemi automatizzati specialmente quando essi si comportino in modo apparentemente affidabile. Se, quindi, il *framework* FIT prescrive l'aumento di vigilanza, ma la psicologia umana tende verso la diminuzione della stessa, allora il FIT potrebbe rivelarsi non soltanto inefficace, ma controproducente nella pratica.

In quarto luogo, la riduzione dell'intenzionalità a un *Intentionality Level* ordinale discreto comporta una perdita significativa di 'informazione sfumata'. Ad esempio, due sistemi potrebbero entrambi classificarsi come IL3, ma divergere radicalmente nei loro profili dimensionali: uno altamente 'volitivo/volizionale' e persistente ma scarso in previsione, l'altro fortemente previsionale e coerente ma privo di autonomia decisionale. In questo caso, il FIT, nell''accorciare' questi profili multidimensionali attraverso una metrica composita singola, perde(rebbe), in pratica, la capacità di discriminare fra configurazioni di rischio qualitativamente distinte. Un regolatore, quindi, che usasse il FIT per prescrivere l'identico regime di supervisione per ambedue i sistemi, nonostante le loro differenze profonde nel comportamento, commetterebbe un errore di categorizzazione potenzialmente pericoloso, o fornendo meno supervisione del necessario in una dimensione critica o imponendo supervisione eccessiva in un'altra dove il rischio fosse, invece, inferiore.

Il FIT, inoltre, introduce una *tirannia della quantificazione*, che, sebbene intesa come strumento di (potenziale) *liberazione dall'arbitrarietà*, potrebbe rivelarsi essa stessa una forma di 'arbitrarietà camuffata'. Assegnando punteggi cardinali a dimensioni comportamentali essenzialmente qualitative, il *framework* in pratica converte in numeri – che assumono un'aura di oggettività scientifica – giudizi che sarebbero, altrimenti, esplicitamente riconosciuti come valutativi e contestabili. In altre parole, quando un valutatore assegna un punteggio di 3.2 alla volizione di un sistema, questo numero conferisce un'apparenza di precisione, che maschera il fatto che la valutazione rimanga fondamentalmente una *narrazione interpretativa* del comportamento. La quantificazione, così facendo, paradossalmente sedimenta e occulta i giudizi di valore sottostanti, rendendo il processo di valutazione 'meno trasparente' piuttosto che 'più cristallino'.

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

ISSN 2420-7888

Ancora. Il *framework* FIT presuppone una separabilità, che, però potrebbe non reggere ad un controllo empirico pur raffinato, fra *design-contingency* – la 'riserva di contingenza di progettazione', concernente quel margine segmentale, tanto economico quanto, soprattutto, temporale, inserito in un progetto di sistema per intervenire sulla modifica estemporanea dei requisiti di quest'ultimo o per gestire errori o imprevisti scoperti nella fase stessa di progettazione – e 'proprietà intrinseche' dei sistemi. La tesi centrale, di fondo, è che l'intenzionalità funzionale sia *design-contingent*, controllabile, cioè, mediante scelte architetturali circa memoria, pianificazione e strumenti. Ma quale garanzia si ha che, una volta superata una certa soglia di complessità computazionale e di dimensione dei parametri, proprietà del tutto assimilabili all'intenzionalità non emergano con spontaneità, indipendentemente dalle specifiche scelte d'architettura? Non è, infatti, da escludere, aprioristicamente o pregiudizialmente, che possa esistere, cioè, una transizione di fase nel comportamento di sistemi di IA sufficientemente complessi, grazie alla quale si sviluppino progressivamente forme di indirizzamento finalistico robuste indipendentemente dalle intenzioni progettuali dei loro creatori. Se così fosse, la presupposizione della controllabilità mediante *design* calerebbe sicuramente e di molto, e il *framework* FIT si ritroverebbe a riporre fede in un potere di controllo che effettivamente non si possiede.

Infine, il FIT implicitamente riflette un particolare orientamento normativo verso l'allocazione di responsabilità, ovvero la convinzione che quest'ultima debba rimanere ancorabile a *soggetti umani*, e, in sostanza, rivela l'idea che una proliferazione di sistemi ad alta intenzionalità rappresenti primariamente un problema di *accountability* piuttosto che un'opportunità di trasformazione delle forme di agentività nella società umana. Potrebbe accadere, però, che, un giorno, non troppo lontano a venire, una società sufficientemente evoluta nel suo impiego di sistemi autonomi ad alta intenzionalità giunga a riconoscere le limitazioni dell'imputabilità tradizionale e sviluppi forme radicalmente alternative di responsabilità condivisa o distribuita, di assicurazione contro danni generati da sistemi autonomi, di creazione di depositi di garanzia computazionali all'interno di cui il sistema stesso è vincolato a mantenere riserve finanziarie per l'indennizzo dei danni. In questa prospettiva, il FIT potrebbe rimanere legato a presupposizioni legali e filosofiche ormai superate da uno scenario più complesso di distribuzione dell'agentività.

Malgrado queste obiezioni, il FIT rimane uno sforzo concettuale singolarmente ambizioso nel tentare di determinare equilibrate connessioni fra la crescente complessità comportamentale dei sistemi intelligenti contemporanei e l'apparato di *governance* tradizionale. Anche se imperfetto, e per alcuni addirittura discutibile, il tentativo di standardizzare la misura dell'intenzionalità funzionale rappresenta un progresso rispetto ai contenuti e ai modi della *governance* attuale. Riconoscendo esplicitamente che l'autonomia non è una variabile binaria, ma un *continuum* multidimensionale, e che tale *continuum* comporta implicazioni significative per l'attribuzione della responsabilità e il regime di supervisione appropriato, il FIT apre uno spazio di riflessione teorica che prima non esisteva. Anche se il *framework* finale potrebbe rivelarsi insufficiente, il processo di sviluppo di metriche standardizzate/omogeneizzate/universalizzate per l'intenzionalità funzionale rappresenta sicuramente un passo importante verso una *governance* dell'IA autonoma, che si voglia proporzionata al reale grado di agentività attribuita a questi sistemi, invece che reattiva e improvvisata come la

Sito web: www.ratioiuris.it
Informazioni: info@ratioiuris.it
Redazione: redazione@ratioiuris.it
Coordinamento Generale: Avv. Maria Antonietta Bozza
Coordinamento Editoriale: Avv. Stenio Salzano

governance odierna, spesso incapace di anticipare i rischi e le opportunità che i sistemi più avanzati e complessi rappresentano.

RIFERIMENTI BIBLIO- E SITO-GRAFICI (BIBLIO-SITO-GRAFIA CITATA, CONSULTATA E CONSIGLIATA)

Amariles, D. R., & Baquero, P. M. (2023). Promises and limits of law for a human-centric artificial intelligence. *Computer Law & Security Review*, 48, 105795. https://doi.org/10.1016/j.clsr.2023.105795.

Angie, A. D., Connelly, S., Waples, E. P., & Kligyte, V. (2011). The influence of discrete emotions on judgement and decision-making: A meta-analytic review. *Cognition & emotion*, 25(8), pp. 1393-1422. https://doi.org/10.1080/02699931.2010.550751.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, 20(3), pp. 191-208. https://doi.org/10.1016/0010-0277(85)90008-3.

Ban, H. Y., *et al*. (2021, March). Kernel flow: a high channel count scalable TD-fNIRS system. In *Integrated Sensors for Biological and Neural Sensing* (Vol. 11663, pp. 24-42). SPIE. https://doi.org/10.1117/1.jbo.27.7.074710.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3), pp. 295-307. https://doi.org/10.1093/cercor/10.3.295.

Bodhi, B. (2007). A *comprehensive manual of abhidhamma: The abhidhammattha sangaha* (3rd edition). Kandy: Buddhist Publication Society. https://www.bps.lk/olib/bp/bp304s_Bfodhi_Comprehensive_Manual_of_Abhidhamma.pdf.

Buddhaghosa, B. (2014). *The path of purification: Visuddhimagga* (4th edition). Kandy: Buddhist Publication Society. https://www.bps.lk/olib/bp/bp207h_The-Path-of-Purification-(Visuddhimagga).pdf.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1), pp. 54-115. https://doi.org/10.1016/S0734-189X(87)80014-2.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3), pp. 200-219. https://consc.net/papers/facing.pdf.

Chalmers, D. J. (1997). Moving forward on the problem of consciousness. *Journal of Consciousness studies*, 4(1), pp. 3-46. https://web-archive.southampton.ac.uk/cogprints.org/317/.

Chiappetta, A., & Mahari, R. (2026). Intentionality is a design decision: Measuring functional intentionality for accountable AI systems. *arXiv preprint arXiv*:2605.05475. https://doi.org/10.48550/arXiv.2605.05475.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., De Prado, M. L., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, 99, 101896. https://doi.org/10.1016/j.inffus.2023.101896.

Enqvist, L. (2023). 'Human oversight'in the EU artificial intelligence act: what, when and by whom?. *Law, Innovation and Technology*, 15(2), pp. 508-535. https://doi.org/10.1080/17579961.2023.2245683.

Giffin, C., & Lombrozo, T. (2022). Mens rea in moral judgment and criminal law. https://cognition.princeton.edu/sites/g/files/toruqf3386/files/giffinlombrozo-handbookch-uncorrectedproof-forthcoming_0.pdf.

Greene, J. D., *et al*. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), pp. 389-400. https://www.cell.com/action/showPdf?pii=S0896-6273%2804%2900634-8.

Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of life reviews*, 11(1), pp. 39-78. https://doi.org/10.1016/j.plrev.2013.08.002.

Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. *Frontiers in Psychology*, 5, 82279. https://doi.org/10.3389/fpsyg.2014.00577.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8), pp. 2554-2558. https://doi.org/10.1073/pnas.79.8.2554.

Kaminski, M. E. (2021). The right to explanation, explained. In *Research handbook on information law and governance* (pp. 278-299). Edward Elgar Publishing. https://doi.org/10.4337/9781788119924.00024.

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and engineering ethics*, 26(1), pp. 89-120. https://link.springer.com/article/10.1007/s11948-018-00081-0.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1), pp. 59-69. https://link.springer.com/article/10.1007/bf00337288.

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2), pp. 144-156. https://doi.org/10.4097/kjae.2017.70.2.144.

Lane, M., & Williams, M. (2023). Defining and classifying AI in the workplace, 290. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/03/defining-and-classifying-ai-in-the-workplace_35d2d069/59e89d7f-en.pdf.

Laux, J. (2024). Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *Ai & Society*, 39(6), pp. 2853-2866. https://doi.org/10.1007/s00146-023-01777-z.

Lee, M. (2024). Emergence of self-identity in AI: A mathematical framework and empirical study with generative Large Language Models. *arXiv preprint arXiv*:2411.18530. https://doi.org/10.48550/arXiv.2411.18530.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual review of psychology*, 66(1), pp. 799-823. https://doi.org/10.1146/annurev-psych-010213-115043.

Li, X., Shi, H., Xu, R., & Xu, W. (2025). AI awareness. *ArXiv, abs/2504.20084*. https://doi.org/10.48550/arXiv.2504.20084.

Maslow, A. H. (1943). A theory of human motivation. *Psychological review*, 50(4), pp. 370-396. https://psycnet.apa.org/doi/10.1037/h0054346.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), pp. 115-133. https://home.csulb.edu/~cwallis/382/readings/482/mccolloch.logical.calculus.ideas.1943.pdf.

Musk, E. (2019). An integrated brain-machine interface platform with thousands of channels. *Journal of medical Internet research*, 21(10), e16194. https://doi.org/10.2196/16194.

Pizzi, M., Romanoff, M., & Engelhardt, T. (2020). AI for humanitarian action: Human rights and ethics. *International Review of the Red Cross*, 102(913), pp. 145-180. https://international-review.icrc.org/sites/default/files/reviews-pdf/2021-03/ai-humanitarian-action-human-rights-ethics-913.pdf.

Renz, F. (2024). The boundaries of legal personhood: Disability, gender and the cyborg. *Law and Critique*, 35(2), pp. 425-444. https://doi.org/10.1007/s10978-023-09350-9.

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, 4, 100005. https://doi.org/10.1016/j.jrt.2020.100005.

Rosert, E., & Sauer, F. (2019). Prohibiting autonomous weapons: Put human dignity first. *Global Policy*, 10(3), pp. 370-375. https://doi.org/10.1111/1758-5899.12691.

Rolls, E. T. (2015). Limbic systems for emotion and for memory, but no single limbic system. *cortex*, 62, pp. 119-157. https://doi.org/10.1016/j.cortex.2013.12.005.

Rost, T. (2026). The Sentience Readiness Index: A reliminary framework for measuring national preparedness for the possibility of artificial sentience. *arXiv preprint arXiv*:2603.01508. https://doi.org/10.48550/arXiv.2603.01508.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (No. ICS8506). DOI: 10.21236/ADA164453.

Samarawickrama, M. (2023, September). Unifying matter, energy and consciousness. In *AIP Conference Proceedings* (Vol. 2872, No. 1, p. 110001). AIP Publishing LLC. https://doi.org/10.1063/5.0162815.

Samarawickrama, M. (2024, February). Mathematical modeling of consciousness for unifying causation, relativity and quantum mechanics. In *Journal of Physics: Conference Series* (Vol. 2701, No. 1, p. 012051). IOP Publishing. https://doi.org/10.1088/1742-6596/2701/1/012051.

Samarawickrama, M. (2024, February). Conscious model of particle physics: The grand theory unifying local and non-local realities. In *Journal of Physics: Conference Series* (Vol. 2701, No. 1, p. 012110). IOP Publishing. https://doi.org/10.1088/1742-6596/2701/1/012110.

Samarawickrama, M. (2024, November). The irreducibility of consciousness in human intelligence: Implications for AI, legal accountability, and the human-in-the-loop approach. In *2024 IEEE Conference on Engineering Informatics (ICEI)* (pp. 1-10). IEEE. https://doi.org/10.3390/e28040377.

Samarawickrama, M. (2025, June). The mind, experience, observation, and reality: A framework and technique for fundamental experience of Rūparealm. In *Journal of Physics: Conference Series* (Vol. 3027, No. 1, p. 012012). IOP Publishing. DOI 10.1088/1742-6596/3027/1/012012.

Samarawickrama, M. (2025, June). The model of consciousness: An analysis of causation, self, gödel's incompleteness, and buddhist philosophy. In *Journal of Physics: Conference Series* (Vol. 3027, No. 1, p. 012014). IOP Publishing. DOI 10.1088/1742-6596/3027/1/012014.

Samarawickrama, M. (2025, June). Consciousness and mathematics: A number theoretic approach to modelling reality. In *Journal of Physics: Conference Series* (Vol. 3027, No. 1, p. 012013). IOP Publishing. DOI 10.1088/1742-6596/3027/1/012013.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, pp. 85-117. https://doi.org/10.1016/j.neunet.2014.09.003.

Sherman, S. J., & Hoffmann, J. L. (2007). The psychology and law of voluntary manslaughter: what can psychology research teach us about the "heat of passion" defense?. *Journal of Behavioral Decision Making*, 20(5), pp. 499-519. https://doi.org/10.1002/bdm.573.

Stewart, W. (2025). The human biological advantage over AI. *AI & Society*, 40(4), pp. 2181-2190. https://doi.org/10.1007/s00146-024-02112-w.

Subasioglu, M., & Subasioglu, N. (2025). From mimicry to True Intelligence (TI)--A new paradigm for Artificial General Intelligence. *arXiv preprint arXiv*:2509.14474. https://doi.org/10.48550/arXiv.2509.14474.

Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons. *Science and Engineering Ethics*, 28(5), pp. 1-22. https://doi.org/10.1007/s11948-022-00392-3.

Tait, I., Bensemann, J., & Wang, Z. (2024). Is GPT-4 conscious?. *Journal of Artificial Intelligence and Consciousness*, 11(01), pp. 1-16. https://doi.org/10.1142/S270507852450005X.

Truby, J., Brown, R. D., Ibrahim, I. A., & Parellada, O. C. (2022). A sandbox approach to regulating high-risk artificial intelligence applications. *European Journal of Risk Regulation*, 13(2), pp. 270-294. https://doi.org/10.1017/err.2021.52.

Ugail, H., & Howard, N. (2025). Quantifying the dynamics of consciousness using hierarchical integration, organised complexity and metastability. *arXiv preprint arXiv*:2512.10972. https://doi.org/10.48550/arXiv.2512.10972.

von Lühmann, A., *et al*. (2021). Toward Neuroscience of the Everyday World (NEW) using functional near-infrared spectroscopy. *Current opinion in biomedical engineering*, 18, 100272. https://doi.org/10.1016/j.cobme.2021.100272.

Wacks, R. (2023). *Law: A very short introduction*. Oxford University Press. http://ndl.ethernet.edu.et/bitstream/123456789/24059/1/106.pdf.