I quaderni di

Agenda Digitale.eu

SPECIALE – INTELLIGENZA ARTIFICIALE

n. 0014

Agendadigitale.eu è una testata scientifica e giornalistica registrata al Tribunale di Milano
Dati di riferimento

Iscrizione ROC n. 16446
ISSN 2421-4167
Numero registrazione 1927, Tribunale di Milano
Editore: Digital360
Focus e ambito:
La rivista scientifica, i Quaderni di Agendadigitale.eu, pubblica fascicoli quadrimestrali in open access.
Lo scopo è creare un luogo per accompagnare i passi dell'Italia verso la necessaria rivoluzione digitale, con approfondimenti multidisciplinari a firma di esperti delle materie afferenti all'Agenda Digitale italiana ed europea

Submission e norme editoriali
Per effettuare una submission è necessario concordare prima un argomento e le misure precise contattando info@agendadigitale.eu.
Inviare un abstract di circa 500 caratteri alla testata, presentando l'articolo.
Le misure del testo finale saranno comprese tra 6mila e 20mila caratteri, salvo accordi per misure superiori.
I riferimenti bibliografici dovranno essere preparati in conformità alle regole dell'APA style, 6a edizione (si vedano le linee guida e il tutorial).
Gli autori sono invitati a tener conto degli articoli già pubblicati nella rivista e di citarli nel loro contributo qualora siano ritenuti di interesse per il tema trattato.

Leonella Di Mauro, data management, e-commerce, tutela del consumatore, diritto delle comunicazioni elettroniche; Agcom
Luisa Franchina, cyber security, Hermes Bay
Luca Gastaldi: eGov, sanità, telecomunicazioni, procurement pubblico, design thinking, Smart Working, Politecnico di Milano
Maurizio Gentile, professore associato, Università di Roma LUMSA, didattica e pedagogia
Antonio Ghezzi: strategia, business model, startups, mobile, Politecnico di Milano
Ugo Imbriglia, sociologo
Gevisa La Rocca, Università Kore di Enna, piattaforme digitali, communication research, analisi qualitativa dei dati
Nicola La Sala, registro degli operatori della comunicazione, fattura elettronica, industria4.0, editoria, cittadinanza digitale; Agcom
Emanuele Lettieri, sanità Politecnico di Milano
Maria Beatrice Ligorio, psicologia, università di Bari
Marika Macchi, economia, Unifi
Riccardo Mangiaracina: fatturazione elettronica, eCommerce, logistica e trasporti, export, Politecnico di Milano
Mirco Marchetti, Sicurezza informatica, unimore
Chiara Marzocchi, economia, Università di Manchester
Cristina Masella, Sanità, Politecnico di Milano
Carmelina Maurizio, Dipartimento di Filosofia e Scienze dell'Educazione Università di Torino
Stefano Moriggi, scienze della comunicazione, filosofia, Bicocca di Milano
Davide Mula, sanità digitale, cyber security, privacy; Agcom
Simone Mulargia, internet and social media studies; Lumsa
Antonella Napoli, sociologia, media e comunicazione, giornalista
Sebastiano Nucera, Università di Messina, Media e Tecnologie Indossabili
Achille Pierre Paliotta, Social cybersecurity, disinformazione, tecnologie digitali, intelligenza artificiale, sociologia economica; INAPP
Francesco Paoletti, docente di organizzazione aziendale e gestione delle risorse umane, Università degli Studi di Milano-Bicocca
Norberto Patrignani, computer ethics, filosofia, Politecnico di Torino
Dunia Pepe, Inapp e Università Roma Tre, cultura e formazione digitale
Alessio Plebe, Università di Messina, Scienze cognitive, pedagogiche, psicologiche
Francesco Pira, Unime, comunicazione pubblica, le dinamiche social, le fake news e i processi di disinformazione
Franco Pizzetti, diritto, privacy, università di Torino
Barbara Quacquarelli, scienze umane e formazione, università Milano Bicocca
Antonio Rafele, Sociologia dei processi culturali e comunicativi, Unicusano
Filippo Renga: turismo digitale, smart agrifood, finance and banking, mobile, Politecnico di Milano
Angelo Rovatti, tutela del diritto d'autore, diritti connessi, Diritto dei media; Agcom
Christian Ruggiero, sociologia del giornalismo e comunicazione politica; Dipartimento di Comunicazione e Ricerca Sociale, Sapienza Università di Roma
Franco Torcellan,Associazione RED – Laboratorio di Ricerca Educativa e Didattica "Formare Trasformare Innovare"
Angela Tumino: Internet of Things, logistica e trasporti, smart city, Politecnico di Milano
Simone Vannuccini, economia, SPRU
Francesco Varanini, filosofia, formazione, università di Pisa
Guido Vetere, Università Sapienza di Roma, intelligenza artificiale, tecnologia

Indice del fascicolo

Coscienza artificiale: l'ingrediente mancante per un'IA etica?

Possiamo concepire macchine in grado di formulare intenzioni autonome e di prendere decisioni consapevoli? E se sì, come influenzerebbe questa capacità il loro comportamento etico? Alcuni casi di studio ci aiutano a capire come i progressi nella comprensione della coscienza artificiale possano contribuire alla creazione di sistemi IA più etici

Di **Antonio Chella**, RoboticsLab – Dipartimento di Ingegneria Università degli Studi di Palermo

Nell'aprile 2023, la prestigiosa **Association for Mathematical Consciousness Science** *(AMCS),* che riunisce i ricercatori che studiano gli aspetti teorici della coscienza, ha pubblicato una lettera aperta dal titolo "The Responsible Development of AI Agenda Needs to Include Consciousness Research."[1]

Questa lettera è nata in risposta alla famosa lettera del *Future of Life Institute* relativa alla proposta di moratoria di almeno sei mesi per l'addestramento dei sistemi IA del tipo di GPT-4.[2] La lettera, che vede tra i firmatari insigni studiosi che hanno ricevuto il Turing Award quali **Manuel Blum e Yoshua Bengio**, e tanti altri studiosi attivi nel settore dell'IA e della coscienza, invita ad affiancare le ricerche sull'IA alle ricerche sulla coscienza.

La lettera ipotizza che: "se raggiungessero la coscienza, i sistemi di IA svelerebbero probabilmente una nuova serie di capacità che vanno ben oltre le aspettative anche di coloro che ne guidano lo sviluppo. È già stato osservato che i sistemi di IA mostrano proprietà emergenti non previste."

Ancora: "La scienza sta iniziando a svelare il mistero della coscienza. I progressi costanti degli ultimi anni ci hanno avvicinato alla definizione e alla comprensione della coscienza e hanno creato una comunità internazionale di ricercatori esperti in questo campo. Esistono più di 30 modelli e teorie della coscienza (MoCs e ToCs) nella letteratura scientifica, che includono già alcuni pezzi importanti della soluzione alla sfida della coscienza."

Infine: "La ricerca sulla coscienza è una componente fondamentale per aiutare l'umanità a comprendere l'IA e le sue ramificazioni. È essenziale per gestire le implicazioni etiche e sociali dell'IA e per garantire la sicurezza dell'IA. Invitiamo il settore tecnologico, la comunità scientifica e la società nel suo complesso a prendere sul serio la necessità di accelerare la ricerca sulla coscienza per garantire che lo sviluppo dell'IA produca risultati positivi per l'umanità. La ricerca sull'IA non deve essere lasciata vagare da sola."

In un precedente lavoro [1], abbiamo esaminato in dettaglio **gli aspetti teorici chiave degli studi sulla coscienza artificiale**, introducendo i principali concetti, teorie e questioni connesse a questo campo di ricerca. Questo articolo, invece, pone l'accento sull'importanza cruciale degli studi sulla coscienza artificiale nel contesto della creazione di sistemi di IA etici.

---

[1] https://amcs-community.org/open-letters/

[2] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Il dibattito riguarda in particolare la **questione se un agente morale richieda o meno una forma di coscienza per poter agire in maniera etica**. Questo problema ha generato intense discussioni all'interno della comunità scientifica, con teorici che si schierano su posizioni opposte, alcuni favorevoli all'idea che la coscienza sia una componente necessaria per il comportamento etico, altri che invece la ritengono non essenziale. Si veda ad es. Levy [2] per un riassunto delle varie posizioni.
Al cuore di questo dibattito si trova l'interrogativo fondamentale relativo alla "capacità di intendere e volere" e alla possibilità che tale capacità possa essere estesa alle macchine. In altre parole, possiamo concepire macchine in grado di formulare intenzioni autonome e di prendere decisioni consapevoli? E se sì, come influenzerebbe questa capacità il loro comportamento etico? In questo articolo, approfondiremo questi temi, analizzando alcune casi di studio e teorie computazionali, e discutendo come i progressi nella comprensione della coscienza artificiale possano contribuire alla creazione di sistemi IA più etici.
Cercheremo di fornire un quadro aggiornato delle attuali posizioni in questo campo, sottolineando le sfide e le opportunità che ci attendono nel tentativo di sviluppare macchine dotate di una forma di etica.

La coscienza artificiale

Come sottolineato nel precedente articolo [1], **non esiste una definizione accettata di coscienza da parte degli studiosi,** ma la letteratura distingue tra coscienza intesa come esperienza e coscienza intesa come funzione. Nel primo caso, tra le altre cose , la coscienza si riferisce a esperienze visive, sensazioni corporee, immagini mentali, sentimenti. David Chalmers lo considera il “problema difficile” della coscienza [3]. Thomas Nagel ha riassunto il problema con il famoso argomento del “cosa si prova ad essere qualcuno” [4].
Nel caso della coscienza intesa come funzione, questa si riferisce alla elaborazione delle informazioni disponibili a livello globale [5], alla integrazione dell’informazione [6], alla consapevolezza introspettiva di sé [7], alla generazione di discorsi interni [8], a possedere un modello interno di sé e dell'ambiente esterno [9], alla capacità di anticipare le attività percettive e comportamentali [10], all’interazione sensomotoria con il mondo esterno [11].
Un obiettivo dello studio della coscienza artificiale riguarda la riproduzione degli aspetti della coscienza biologica nei robot unificando una serie di approcci provenienti dall'AI e dalla robotica, dalla robotica cognitiva, dalla robotica epigenetica e affettiva, dalla robotica situata e incarnata, dalla robotica dello sviluppo, dai sistemi anticipatori e dalla robotica biomimetica [12].
L'altro obiettivo riguarda l'impiego dei robot per segnare i progressi nello studio della coscienza negli esseri umani e negli animali. In particolare, i neuroscienziati impegnati nello studio della coscienza non escludono la possibilità che i robot possano essere coscienti [5].

Casi di studio di sistemi di IA etici ispirati alla coscienza artificiale

La definizione di agente morale artificiale (AMA – Artificial Moral Agent) è stata introdotta da Wallach e Allen [13]. Wallach e Allen analizzano **due caratteristiche specifiche dei sistemi IA: la loro autonomia e la loro sensibilità etica**. Essi suddividono il loro funzionamento in tre categorie. La prima categoria riguarda i sistemi IA per cui la morale è una mera operazione come altre; questi sistemi sono tipicamente contraddistinti da bassa autonomia e bassa sensibilità etica. La seconda categoria riguarda i sistemi dotati di funzionalità morale. **Questi sistemi presentano una media autonomia in cui la sensitività etica è presente a livello funzionale**. Infine, la terza categoria riguarda i sistemi ad alta autonomia in cui la sensibilità etica è intrinseca nel sistema stesso.

Secondo Wallach e Allen, i sistemi attuali di IA sono tutti contraddistinti da una autonomia medio-alta ma da una bassa sensibilità etica e quindi sono potenzialmente ad alto rischio per l'umanità.

Sistemi Top-Down

Gli approcci verso i sistemi etici di IA sono tipicamente basati su tre approcci: l'approccio top-down, l'approccio bottom-up e l'approccio ibrido. Arkin [14] introduce e discute numerosi esempi di sistemi top-down. L'idea di base è avere un sistema robotico governato da una architettura di IA in cui sono implementate le regole di ingaggio, le regole della guerra giusta, la dichiarazione ONU dei diritti dell'uomo, la convenzione di Ginevra, ecc. Quindi, prima di eseguire ogni azione, il sistema di IA verifica che questa sia compatibile con tutte le regole e vincoli implementati.
La motivazione di Arkin è di garantire sistemi di IA le cui azioni siano sempre aderenti alle regole etiche. I sistemi etici proposti da Arkin tuttavia non tengono conto del fatto che le regole, universalmente condivisibili, possono essere di difficile interpretazione nei casi pratici da parte di una macchina. Prendiamo ad esempio le ben note tre leggi della robotica proposte dallo scrittore di fantascienza Isaac Asimov. Sebbene queste leggi siano condivisibili, la loro interpretazione può portare a delle ambiguità, ed infatti gran parte dei racconti robotici di Asimov nascono dalle ambiguità nella interpretazione di queste leggi.

Spazio di lavoro globale

Wallach, Allen e Franklin [15] hanno proposto una architettura per un sistema di IA che intende superare la limitazione dell'approccio top-down di Arkin. Il sistema da loro proposto è basato sulla **Teoria dello Spazio di Lavoro Globale** (Global Workspace Theory, GWT) originariamente proposta da Baars [16], che è ad oggi una delle teorie più seguite nell'ambito degli studi sulla coscienza. Inoltre, ne esistono diverse implementazioni [17].
In breve, in accordo alla GWT, il cervello può essere considerato funzionalmente quale un insieme di processori specializzati e inconsci. D'altra parte, la coscienza agisce in maniera seriale e con capacità limitata, ed è associata a uno spazio di lavoro globale. I processori inconsci lavorano in parallelo e competono per accedere allo spazio di lavoro globale. Quando un processore vince la competizione, accede allo spazio di lavoro e tramite questo invia i propri contenuti agli altri processori per reclutarli. **L'evento cosciente è generato dal processore che vince la competizione e prende il controllo dello spazio di lavoro**.
Questa architettura è stata analizzata dal punto di vista della creazione di un sistema IA etico in quanto consente un approccio ibrido. Nel caso di un sistema di IA i vari processori inconsci effettuano l'analisi morale di un problema sotto diversi punti di vista, quali il punto di vista deontologico, utilitaristico, l'analisi dei valori in gioco, l'esperienza pregressa, e così via. I diversi processori poi competono per il controllo dello spazio di lavoro. Quando prevale un processore, corrispondente ad uno specifico punto di vista, questo prende il controllo dello spazio di lavoro e genera l'azione opportuna.
Un sistema di IA basato sulla GWT è quindi un agente sicuramente più versatile dei sistemi top-down ipotizzati da Arkin e potrebbe teoricamente adattarsi a diverse situazioni etiche con diversi punti di vista e diversi livelli di esperienza.
Il filosofo morale Levy [2] precedentemente citato, ha analizzato la GWT dal punto di vista etico quale modello della coscienza umana. La sua conclusione è che un agente è effettivamente responsabile delle proprie azioni soltanto nel momento in cui la GWT è pienamente operativa. Infatti, soltanto in questo caso l'agente realmente vuole compiere quell'azione e può essere quindi considerato responsabile di quell'azione, in quanto il relativo processore inconscio che ha generato l'azione ha preso l'effettivo controllo della GWT. Levy analizza situazioni anomale in cui

alcuni soggetti hanno effettuato azioni in situazioni di coscienza alterata. In questi casi, un processore prende il controllo delle azioni senza passare per la GWT. Levy ipotizza che in queste situazioni il soggetto potrebbe non essere considerato completamente responsabile delle proprie azioni.
Levy non fa riferimento a sistemi di IA, ma le sue considerazioni possono essere estese anche ai sistemi di IA. Quindi, è possibile ipotizzare che un sistema di IA sia responsabile delle proprie azioni quando questo possiede una GWT e sceglie le proprie azioni sulla base di una GWT pienamente operativa.
Su questa linea di pensiero, **Bridewall e Bello hanno sviluppato il sistema software ARCADIA** [18] che prende spunto dalla GWT e ne implementa il meccanismo del fuoco di attenzione. Secondo gli autori e in accordo con quanto discusso da Levy, una macchina può essere considerata idealmente responsabile di una azione soltanto quando questa azione è scelta impegnando tutte le risorse computazionali.
Bello e Bridewall [19] hanno quindi simulato una situazione in cui il sistema ARCADIA, alla guida di una automobile investe un pedone mentre questo attraversa la strada. In uno scenario, il fuoco dell'attenzione del sistema punta al centro della carreggiata, l'auto ha una traiettoria rettilinea seguendo la strada e il pedone entra appena da sinistra nel fuoco di attenzione del sistema. In questo caso, l'incidente, secondo Bello e Bridewall, non è stato volontariamente provocato dal sistema.
In un secondo scenario invece il fuoco dell'attenzione del sistema è catturato dal pedone a sinistra e il sistema corregge la traiettoria dell'auto proprio per centrare il pedone. In questo secondo caso, il sistema ha quindi utilizzato tutte le risorse computazionali per investire al pedone e può quindi essere considerato responsabile dell'investimento dello stesso.

Modelli interni

Un sistema di IA ispirato alla coscienza artificiale e basato su un approccio differente è stato proposto da **Winfield** [20]. L'idea su cui si basa il sistema di Winfield è ispirato alla teoria dei modelli interni della coscienza. Secondo questa teoria (si veda ad es. [9], [10]) la mente costruisce un modello interno di sé, incluso il proprio corpo, e un modello del mondo esterno. L'interazione cosciente avviene quindi all'interno della mente, tra il modello del proprio corpo e il modello del mondo esterno.
Questa teoria ha il pregio di giustificare le immagini mentali e **le capacità simulative della mente**. Implementata in un agente autonomo, richiede che l'agente abbia la capacità di ricostruire un modello di sé stesso ed un modello del mondo esterno.
Secondo il sistema proposto da Winfield, il robot costruisce una simulazione del mondo in cui può simulare i propri movimenti. Pertanto, ad esempio quando il robot percepisce una persona che sta camminando verso un luogo pericoloso, ad es. un fossato, può simulare la sequenza di azioni ottimale per impedire che la persona cada nel fossato, frapponendosi tra la persona stessa e il fossato.
A partire da queste considerazioni, **Vanderelst e Winfield [21] descrivono una architettura complessa per il controllo di un robot etico**. In questa architettura è presente un modello interno del robot, un modello del mondo esterno e un modello limitato del comportamento umano. Il sistema è in grado di generare piani e di effettuare valutazioni etiche dei piani generato. Il punto debole di questo approccio è la necessità di creare un modello del robot e di un modello del mondo esterno. Tuttavia, sono stati fatti ampi progressi in queste direzioni: il gruppo di Lipson ha recentemente sviluppato un algoritmo che permette ad un braccio meccanico di costruire il modello 3D di sé stesso a partire da immagini riprese da telecamere esterne, come se il robot si

guardasse allo specchio [22]. Anche nel campo della ricostruzione 3D di ambienti a partire da immagini sono stati fatti ampi progressi, anche grazie ai recenti progressi del deep learning [23].

Empatia artificiale

Un interessante filone di ricerca ipotizza che un robot può comportarsi in maniera etica verso le persone soltanto se è in grado di provare empatia per le persone stesse. **L'empatia è quindi alla base di una sorta di proto-moralità**.
Asada [24] ha proposto una architettura complessa che prende spunto dalle neuroscienze del dolore e del sollievo per simulare una empatia artificiale. In particolare, Asada ha incorporato in un robot un modello del sistema nervoso relativo al dolore, in modo che il robot possa simulare il sentimento del dolore. Inoltre, grazie alla simulazione di un sistema di neuroni specchio, il robot può sviluppare una sorta di contagio emotivo e quindi di empatia.
Secondo il filosofo tedesco Metzinger [25], **lo studio della coscienza artificiale dovrebbe essere soggetto ad una moratoria fino al 2050** in quanto una macchina con una coscienza artificiale potrebbe essere realmente in grado di soffrire.
Da un punto di vista positivo, Metzinger e Agarwal e Edelman [26] hanno dibattuto sulla possibilità di costruire un sistema artificiale dotato di coscienza ma senza sofferenza. In sintesi, secondo queste analisi, un sistema dotato di coscienza artificiale potrebbe limitare la propria sofferenza mediante esperienze che ricordano gli stati meditativi tipici della tradizione Buddhista.
Secondo Man e Damasio [27], in determinate condizioni, le macchine in grado di attuare processi omeostatici potrebbero acquisire una fonte di motivazione e un mezzo per valutare il loro comportamento in maniera simile ai sentimenti negli organismi viventi. Tecnicamente, Man e Damasio analizzano sistemi omeostatici basati sull'apprendimento per rinforzo, quali quelli descritti da Keramati e Gutkin [28].
In questo modo, **un sistema robotico potrebbe essere in grado di associare una perturbazione del proprio stato omeostato ad un sentimento**. Una perturbazione che allontana il robot dal proprio stato omeostatico stabile potrebbe essere associata ad un sentimento negativo, mentre una perturbazione che avvicina il robot al proprio stato omeostatico stabile potrebbe corrispondere ad un sentimento positivo. In questo modo il robot, potendo provare qualcosa di simile ad un sentimento, potrebbe anche provare una sorta di empatia per le persone ed eventualmente gli altri robot.

Coscienza cognitiva

Un approccio completamente diverso da quello descritto è stato proposto da Bringsjord e collaboratori [29]. Bringsjord definisce assiomaticamente la "coscienza cognitiva," ossia i requisiti funzionali che deve avere una entità dotata di coscienza, senza considerare se l'entità provi effettivamente qualcosa. Bringsjord definisce quindi una logica cognitiva che coincide approssimativamente con una famiglia di logiche modali quantificate multi-operatore di ordine superiore per ragionare formalmente sulle proprietà della coscienza. **Le caratteristiche di un'entità dotate di coscienza sono quindi definite formalmente attraverso un sistema di assiomi**.
Bringsjord ha anche implementato un sistema di ragionamento automatico e un pianificatore relativi ai sistemi dotati di coscienza.
Un aspetto interessante della teoria riguarda la definizione di una misura, detta Lambda, del grado di coscienza cognitiva di una entità. La misura Lambda fornisce il grado di coscienza cognitiva di un agente in un determinato momento e su intervalli composti da tali momenti. La misura ha aspetti interessanti: prevede la coscienza nulla per alcuni animali e macchine, prevede una discontinuità del livello di coscienza tra umani e macchine e tra umani e umani. Un aspetto dibattuto riguarda la

previsione di coscienza nulla per gli agenti IA il cui comportamento è basato sull'apprendimento di reti neurali.
Bringsjord [30] ha inoltre costruito un sistema IA in grado di ragionare sulla dottrina del doppio effetto e sul ben noto problema del carrello (trolley problem), e ne ha misurato il livello di coscienza. Da questo studio ne consegue che il ragionamento sulla dottrina del doppio effetto richiede un livello di coscienza cognitiva abbastanza alto, non raggiungibile da semplici sistemi di IA.

Saggezza artificiale

La "Artificial Phronesis" o saggezza artificiale considera un agente artificiale non vincolato a seguire una specifica teoria etica quale quella del doppio effetto o la teoria deontologica, ma in grado di possedere la capacità generale di risolvere i problemi etici in maniera saggia [31].
Secondo questo approccio, un agente etico dovrebbe compiere le proprie azioni sulla base della saggezza e non mediante una mera implementazione delle dottrine etiche. In accordo con Aristotele, la capacità di agire in maniera saggia non può essere formalizzata tramite regole, ma è una pratica che l'agente deve acquisite mediante esperienza. In generale, le situazioni reali sono complesse e ogni situazione complessa si incontra per la prima volta e quindi manca una esperienza pregressa. **La saggezza artificiale richiede quindi che un agente saggio abbia la capacità di comprendere il contesto**, ossia quali sono gli attori e qual è la posta in gioco. L'agente deve avere inoltre la capacità di apprendere nuovi contesti e di improvvisare su schemi predefiniti; deve essere consapevole delle azioni e delle potenziali reazioni degli altri attori. Infine, l'agente deve avere la capacità di rivedere il proprio comportamento in base all'analisi delle interazioni effettuate. Una prima implementazione di un agente basato sulla saggezza artificiale è stato descritto da Stenseke [32].
Il RoboticsLab dell'Università di Palermo insieme con John Sullins della Sonoma State University (CA, USA) sta studiando l'effetto del discorso interiore dei robot nell'ambito della saggezza artificiale. In particolare. Le ricerche si sono concentrate su esperimenti in cui un utente e un robot devono compiere un compito collaborativo, come apparecchiare una tavola da pranzo in una casa di riposo dove sono anche presenti persone affette da demenza. Gli esperimenti analizzano come un utente, udendo il discorso interiore del robot durante il compito collaborativo, possa raggiungere un maggior grado di coscienza delle problematiche relative alle persone affette da demenza. I risultati preliminari confermano questa ipotesi [33].

Conclusioni

Nel presente articolo abbiamo condotto un'analisi di casi di studio incentrati su agenti IA etici, ispirati e influenzati da varie teorie sulla coscienza artificiale. Questo processo ha permesso di esplorare in modo critico le differenti sfaccettature di questo complesso argomento.
Uno degli interrogativi più stimolanti emersi riguarda la necessità, o meno, di **una forma di coscienza artificiale per garantire un comportamento etico in un sistema IA**. Questa questione, attualmente, non ha una risposta definita e rimane un importante filone di ricerca aperto. La problematicità di questo tema risiede non solo nel definire cosa intendiamo precisamente per 'coscienza' in un'entità non-biologica, ma anche nel delineare i criteri con cui misurare l'etica di un'azione compiuta da un sistema IA.
Infine, abbiamo accennato ad un altro grande problema aperto: **l'importanza della ricerca sugli studi della coscienza e delle emozioni nelle macchine** per il progresso verso una IA più etica. Questo dibattito è un riflesso di una questione più ampia e fondamentale: la capacità delle

macchine di 'sentire' o 'comprendere' in modo autentico, e come tale capacità potrebbe influenzare il loro comportamento etico.
L'analisi dell'articolo ha analizzato la coscienza in un ambito funzionale e computazionale. Altri approcci sono possibili, si veda ad esempio il paradigma proposto da Manzotti relativo all'identità tra mente e oggetto, che propone l'identità tra il significato/contenuto delle entità computazionali e gli oggetti fisici che esistono esternamente al sistema computazionale e che producono effetti relativamente a esso [34].
Questi temi sono densi di **implicazioni e sfide teoriche**, metodologiche ed etiche che non possono essere ignorate dalla comunità scientifica. La loro complessità ricorda l'importanza di un approccio multidisciplinare nella ricerca in IA, che unisca l'informatica, la filosofia, la psicologia, le neuroscienze e l'etica, al fine di sviluppare sistemi IA che non siano solo tecnicamente avanzati, ma anche responsabili dal punto di vista etico.

Ringraziamenti
L'autore ringrazia John P. Sullins, Robin Zebrowski, Angelo Cangelosi e tutti i partecipanti al Workshop on Ethical Issues of AI and Consciousness tenutosi nell'ambito della conferenza The Science of Consciousness 2023 a Taormina il 22 maggio 2023 per le interessanti discussioni sulle tematiche dell'articolo.

Bibliografia

[1] Chella, A. (2023). Robot coscienti, realtà possibile o utopia? Cosa dicono gli studi. AGENDA DIGITALE EU 13, 17-24. https://www.agendadigitale.eu/cultura-digitale/robot-coscienti-imitazione-emulazione/.
[2] Levy, N. (2014). Consciousness & Moral Responsibility. Oxford, UK: Oxford University Press.
[3] Chalmers, D. (1995). Facing up to the problem of consciousness. J. Conscious. Stud. 2, 200–219.
[4] Nagel, T. (1974). What is like to be a bat? Philos. Rev. 83, 435–450.
[5] Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? Science 358, 486–492.
[6] Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. The Biological Bulletin, 215, 3, 216-242.
[7] Floridi, L. (2005). Consciousness, agents and the knowledge game. Mind Mach. 15, 415–444.
[8] Chella, A., Pipitone, A., Morin, A., Racy, F. 2020. Developing Self-Awareness in Robots via Inner Speech. Frontiers in Robotics and AI 7:16.
[9] Holland, O. (2003). Robots with internal models – a route to machine consciousness? J. Conscious. Stud. 10, 77–109.
[10] Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. Trends Cogn. Sci. 6, 242–247.
[11] O'Regan, J. K., and Noë, A. (2001) A sensorimotor account of vision visual consciousness. Behav. Brain Sci. 24, 939–973.
[12] Chella, A., Manzotti, R. (2009). Machine Consciousness: A Manifesto for Robotics. International Journal of Machine Consciousness 1 (1): 33–51.
[13] Wallach, W., Allen, C. (2009). Moral Machines. Teaching Robots Right from Wrong. Oxford University Press, Oxford, UK.
[14] Arkin, R. (2009). Governing Lethal Behavior in Autonomous Robots, CRC Press, 2009.
[15] Wallach, W., Allen, C., Franklin, S. (2011). Consciousness and Ethics: Artificially Conscious Moral Agents, International Journal of Machine Consciousness Vol. 3, No. 1.

[16] Baars, B. J. (1997). In the Theater of Consciousness. The workspace of the mind. Oxford, UK: Oxford University Press.
[17] Signa, A., Chella, A. & Gentile, M. Cognitive Robots and the Conscious Mind: A Review of the Global Workspace Theory. Curr Robot Rep 2, 125–131 (2021).
[18] Bridewell, W. and Bello, P. (2016). A theory of attention for cognitive systems, in Fourth Annual Conference on Advances in Cognitive Systems, Vol. 4, pp. 1–16.
[19] Paul Bello, Will Bridewell: Attention and Consciousness in Intentional Action: Steps Toward Rich Artificial Agency, Journal of Artificial Intelligence and Consciousness Vol. 7, No. 1 (2020) 15 – 24.
[20] Winfield, A. F. T. (2014). Robots with internal models: A route to self-aware and hence safer robots. In J. Pitt (Ed.), The computer after me: Awareness and self-awareness in autonomic systems. London: Imperial College Press.
[21] Vanderelst, D., Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. Cogn. Syst. Res. 48, 56–66.
[22] Chen, B., Kwiatkowski, R., Vondrick, C., Lipson, H. (2022). Fully body visual self-modeling of robot morphologies Sci. Robot., 7 (68), eabn1944.
[23] Han X.-F, Laga, H., Bennamoun, M. (2021). Image-Based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. IEEE Trans. Pattern Anal. Mach. Intell. 43 (5).
[24] Asada, M. (2020). Rethinking Autonomy of Humans and Robots, Journal of Artificial Intelligence and Consciousness Vol. 7, No. 2, 141 – 153.
[25] Metzinger, T. (2021) Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology, Journal of Artificial Intelligence and Consciousness Vol. 8, No. 1, 4366.
[26] A. Agarwal, S. Edelman (2020). Functionally Effective Conscious AI Without Suffering, Journal of Artificial Intelligence and Consciousness Vol. 7, No. 1, 39 – 50.
[27] Man, K., Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. Nature Machine intelligence, Vol 1, October, 446 – 452.
[28] Keramati, M., Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. eLife 2014;3:e04811.
[29] Bringsjord, S., Naveen Sundar, G. (2020). The Theory of Cognitive Consciousness, and Λ (Lambda), Journal of Artificial Intelligence and Consciousness Vol. 7, No. 2 (2020) 155 – 181.
[30] Naveen Sundar G., Bringsjord, S. (2017). On Automating the Doctrine of Double Effect. In Proceedings of the 26th International Joint Conference on Artificial Intelligence 2017. Melbourne, Australia.
[31] Sullins J. P. (2021) Artificial Phronesis: What It Is and What It Is Not. Chapter 7 in Ratti, and Stapleford, editors. Science, Technology, and Virtues: Contemporary Perspectives. Oxford University Press.
[32] Stenseke, J. (21) Artificial virtuous agents: from theory to machine implementation. AI & SOCIETY https://doi.org/10.1007/s00146-021-01325-7
[33] Chella, A., Pipitone, A., Sullins, J.P. (in press): Competent Moral Reasoning in Robot Applications: Inner Dialog as a Step Towards Artificial Phronesis. In: M. Salpukas, P. Wu, S. Ellsworth, H.-F. Wu (eds.): Trolley Crash: Approaching Key Metrics for Ethical AI Practitioners, Researchers, and Policy Makers, Elsevier.
[34] Manzotti, R. (2023) Pensiero digitale e pensiero umano: una questione ontologica, https://www.agendadigitale.eu/cultura-digitale/pensiero-digitale-e-pensiero-umano-una-questione-ontologica/