RIVISTA ITALIANA DI INFORMATICA E DIRITTO
PERIODICO INTERNAZIONALE DEL CNR-IGSG

ISSN 2704-7318 • n. 1/2025 • DOI 10.32091/RIID0210 • articolo sottoposto a peer review • pubblicato in anteprima il 26 feb. 2025

Arcangelo Leone De Castris

Possibili funzioni di un istituto internazionale per la sicurezza dell'intelligenza artificiale. Un'analisi istituzionale dell'AIEA e dell'IPCC nel contesto delle recenti tendenze sulla governance dell'intelligenza artificiale

I principali attori coinvolti nella governance dell'intelligenza artificiale (IA) in tutto il mondo concordano sul fatto che se da un lato questa tecnologia promette di produrre enormi benefici di natura economica e sociale, dall'altro è necessario implementare dei processi che ne regolino lo sviluppo e l'utilizzo in modo da mitigare i rischi che essa comporta. Diverse istituzioni internazionali – tra cui l'OCSE, il G7, il G20, l'UNESCO e il Consiglio d'Europa – hanno iniziato a sviluppare quadri di governance per lo sviluppo etico e responsabile dell'IA. Nonostante importanti sviluppi in questo senso, ad oggi non esistono dei processi istituzionalizzati al livello internazionale per identificare, misurare e controllare le capacità potenzialmente dannose dell'IA. Con l'obiettivo di contribuire al dibattito accademico sul tema, questo articolo riflette sull'opportunità di creare un istituto internazionale per la sicurezza dell'IA. Partendo dall'analisi delle istituzioni internazionali che si occupano di sicurezza in aree politiche adiacenti all'IA, nonché degli istituti nazionali per la sicurezza dell'IA recentemente costituiti dal Regno Unito e dagli Stati Uniti, l'articolo propone un elenco di funzioni che potrebbero essere svolte da un istituto internazionale per la sicurezza dell'IA. In particolare, l'articolo suggerisce che le possibili funzioni di tale istituto possono essere articolate all'interno di tre categorie: (a) ricerca e cooperazione, (b) audit e verifiche di conformità dei modelli di IA e (c) supporto alla definizione di quadri di governance dell'IA.

Intelligenza artificiale – Governance dell'IA – Sicurezza dell'IA – Etica dell'IA

Possible functions of an international institute for AI safety. An institutional analysis of the IAEA and IPCC in the context of recent trends in AI governance

Most actors involved in governing AI technologies around the world agree that, while AI promises to deliver tremendous benefits to society, appropriate guardrails are required to mitigate its risks. International institutions – including the OECD, the G7, the G20, UNESCO, and the Council of Europe – have started developing frameworks for ethical and responsible AI governance. However, these initiatives alone fall short of addressing the need for institutionalised international processes to identify and assess potentially harmful AI capabilities. This paper contributes to the ongoing conversation on how to address this gap by reflecting on the opportunity to establish an international institution for AI safety. Based on the analysis of international institutions established to address safety considerations in adjacent policy areas and of the newly established national AI safety institutes in the UK and US, this paper identifies a list of functions that could be performed by an international institution for AI safety in three broad areas: (a) technical research and cooperation, (b) safeguards and evaluations, and (c) policymaking and governance support.

Artificial intelligence – AI governance – AI safety – AI ethics

L'Autore è ricercatore associato presso The Alan Turing Institute, Londra

1. Introduzione

In risposta alla rapida commercializzazione dei sistemi di intelligenza artificiale (IA) generativa, nel 2023 sono stati lanciati numerosi appelli per una regolamentazione globale dell'IA da parte di figure di spicco in ambito economico e politico[1]. Uno dei temi più dibattuti riguarda la creazione di istituzioni per la governance dell'IA[2]. In particolare, nel periodo precedente al Summit sulla sicurezza dell'IA ospitato dal Regno Unito nel novembre 2023, si è discusso in modo approfondito su quale sia il modello di governance più adeguato, e diversi analisti hanno proposto di prendere spunto dal modello della governance ambientale, con l'*International Panel on Climate Change* (IPCC), e dell'energia atomica, con l'*Agenzia internazionale per l'energia atomica* (AIEA). Tuttavia, tracciare paralleli tra ambiti tematici così diversi presenta il rischio di giungere a conclusioni non del tutto accurate. Inoltre un istituto per la sicurezza dell'IA modellato sugli esempi dell'IPCC o dell'AIEA rischierebbe di ereditare gli stessi problemi che hanno limitato l'efficacia di tali istituzioni[3]. In ogni caso, anche se questi esempi non sono direttamente applicabili alla governance dell'IA, analizzare le misure di governance in ambiti politici adiacenti a quello dell'IA potrebbe comunque fornire lezioni preziose per questo nuovo campo[4].

Partendo da tale ipotesi, questo articolo propone un'analisi di tipo istituzionale dell'AIEA e dell'IPCC[5], analizzandone le funzioni esercitate e i risultati raggiunti e riflettendo sulla loro rilevanza per la governance dell'IA. L'articolo propone inoltre una breve analisi degli istituti nazionali per la sicurezza dell'IA recentemente costituiti nel Regno Unito e negli Stati Uniti. Infine, contestualizzando l'analisi di questi casi di studio all'interno degli attuali sviluppi nell'ambito della governance dell'IA, l'articolo offre un'interpretazione delle funzioni che un ipotetico istituto internazionale per la sicurezza dell'IA potrebbe svolgere. Lo spazio limitato di questa analisi impedisce di offrire un confronto più ampio con altre istituzioni internazionali – ad esempio, l'*International Financial Stability Board* – e con istituti nazionali per la sicurezza dell'IA di più recente creazione – ad esempio,

1. Altman–Brockman–Sutskever 2023; Suleyman–Schmidt 2023; Milmo 2023; Prime Minister's Office 2023; Guterres 2023.
2. Ho–Barnhart–Trager et al. 2023; Marcus–Reuel 2023; Roberts–Hine–Taddeo et al. 2023; Veale–Matus–Gorwa 2023.
3. Afina–Lewis 2023; Stewart 2023; De Pryck–Hulme 2022.
4. Milmo 2023.
5. Steinmo–Thelene–Longstreth 1992.

gli istituti creati in Canada, Singapore, Giappone e Corea del Sud. Questi rimangono importanti casi di studio da approfondire con ricerche future.

Il secondo paragrafo di questo articolo illustra la relazione tra il tema della sicurezza dell'IA e i concetti di etica e governance[6], collocando la creazione di istituzioni per la sicurezza dell'IA nel più ampio panorama globale della governance di questa tecnologia. Il terzo e quarto paragrafo, rispettivamente, offrono un'analisi dei casi di studio dell'AIEA e dell'IPCC, concentrandosi sui loro successi e fallimenti negli ambiti della governance ambientale e nucleare. Il quinto paragrafo riflette sulla rilevanza di entrambi questi modelli per il dibattito in corso sulla governance dell'IA e, in particolare, sulla potenziale creazione di un istituto internazionale per la sicurezza dell'IA. Il successivo paragrafo analizza le recenti iniziative del Regno Unito e degli Stati Uniti relative alla creazione di istituti nazionali per la sicurezza dell'IA, valutandone il mandato istituzionale e confrontandoli con le conclusioni tratte dai casi di studio dell'IPCC e dell'IEAE. Nell'ultimo paragrafo vengono infine elencate le potenziali funzioni di un istituto internazionale per la sicurezza dell'IA sulla base dell'analisi precedentemente svolta.

2. La sicurezza dell'IA nel contesto dell'etica e della governance

Il tema della sicurezza dell'IA è strettamente legato ai due concetti più ampi di etica e governance. Il concetto di etica dell'IA si riferisce all'insieme di valori, principi e tecniche che impiegano standard ampiamente accettati di ciò che è giusto e sbagliato nel guidare lo sviluppo e l'uso delle tecnologie di IA[7]. Nell'ambito dell'etica dell'IA, il termine "sicurezza" viene utilizzato per indicare uno dei principi guida per lo sviluppo di un'IA etica. Sulla base di questo principio, l'IA deve funzionare in modo robusto e sicuro e i rischi che pone devono essere identificati, valutati e gestiti in modo continuativo durante il ciclo di vita delle sue applicazioni. I principi etici possono essere integrati all'interno di processi organizzativi tramite diversi strumenti di governance. Per governance si intende il sistema di organi, regole, pratiche e processi che possono essere impiegati per garantire che l'uso delle tecnologie di IA da parte di un'organizzazione sia compatibile con gli obbiettivi strategici, i requisiti legali e i principi etici dell'organizzazione stessa[8].

Ai fini di questo articolo, il concetto di sicurezza dell'IA si riferisce alla combinazione di due condizioni, una negativa e una positiva: (a) l'assenza di rischi inaccettabili causati dall'uso dell'IA, e (b) l'esistenza di misure di sicurezza per gestire i rischi e i potenziali danni legati all'uso dell'IA[9]. In questo senso, il ruolo di un istituto per la sicurezza dell'IA sarebbe quello di sviluppare e mettere in atto tali misure con lo scopo di prevenire o controllare i rischi posti dalle applicazioni dell'IA.

Il tema della sicurezza dell'IA si inserisce in un complesso contesto di governance internazionale composto da molteplici strumenti di regolamentazione, inclusi trattati e convezioni, principi, accordi quadro, standard tecnici, sviluppati da organismi come ISO e IEC, e sistemi regolatori paralleli, come ad esempio la governance delle piattaforme digitali, la protezione dei dati e della privacy e gli accordi commerciali internazionali. Tra le iniziative di governance internazionale dell'IA di maggior rilievo vi sono i Principi sull'IA dell'OCSE (2019), la Raccomandazione sull'etica dell'IA dell'UNESCO (2022), il Report della Partnership globale sull'IA (2023), il Rapporto sulla Governance dell'IA per l'Umanità dell'ONU (2023), il Processo di Hiroshima del G7 (2023) e la Convenzione sull'IA e i diritti umani del Consiglio d'Europa (2024). Altrettanto importanti iniziative, alcune delle quali con effetti extraterritoriali, esistono anche su scala regionale e nazionale. Ad esempio, il Regolamento europeo sull'IA (EU AI Act) è applicabile a persone fisiche e giuridiche che sviluppano o implementano sistemi di IA che producono effetti all'interno dell'Unione, a prescindere dalla loro residenza legale. Negli Stati Uniti, sebbene non sia prevista l'adozione di leggi sull'IA a livello federale, diversi Stati hanno adottato o stanno considerando regolamenti sul tema. Ad esempio, il Colorado ha approvato il

6. Jobin–Ienca–Vayena 2019.
7. Leslie 2019.
8. Mäntymäki–Minkkinen–Birkstedt–Viljanen 2022.
9. Habli 2023.

disegno di legge SB24-205[10], che fornisce protezione ai consumatori che interagiscono con i sistemi di IA, e il New Jersey ha approvato il disegno di legge S1588, che impone procedure di audit e requisiti di imparzialità a coloro che usano sistemi di IA all'interno di procedure per l'assunzione di personale. Inoltre, il *Risk Management Framework* per l'IA sviluppato dal *National Institute for Standards and Technology* degli Stati Uniti è destinato a svolgere un ruolo importante su scala internazionale come strumento di *soft law*. Simili iniziative sono state adottate anche in diversi altri paesi[11]. In parallelo, organizzazioni internazionali come ISO, IEC e IEEE stanno sviluppando standard tecnici che possono essere adottati volontariamente da aziende e altre organizzazioni per implementare alcuni dei principi etici stabiliti dalle iniziative menzionate in precedenza. Infine, la governance dell'IA si basa anche su una serie di strumenti di governance complementari, tra cui la regolamentazione della protezione dei dati e della privacy[12], la regolamentazione delle piattaforme digitali[13], la regolamentazione del mercato e della concorrenza[14] e la protezione dei diritti umani[15].

All'interno di questo complesso panorama, tuttavia, non esistono funzioni istituzionalizzate per valutare in modo indipendente i rischi e gli sviluppi futuri dell'IA al livello internazionale[16]. Con l'obiettivo di contribuire al dibattito su come colmare questa lacuna, qui di seguito viene proposta una riflessione su come le esperienze dell'AIEA e dell'IPCC possono fornire degli spunti utili a sviluppare nuove soluzioni per la governance internazionale dell'IA.

3. Governance nucleare globale: l'Agenzia Internazionale per l'Energia Atomica

Creata nel 1957 per coordinare la proliferazione nucleare post-bellica su scala globale, l'AIEA è un forum intergovernativo che rappresenta 178 paesi e funge da punto focale globale per la cooperazione scientifica e tecnica nel campo dell'energia nucleare. L'ambito di competenza dell'AIEA include tre aree principali: a) sicurezza, b) ricerca scientifica, c) salvaguardia e controllo. L'obiettivo statutario dell'AIEA comprende funzioni sia "promozionali" sia "regolatorie" in merito all'uso e alla non proliferazione dell'energia nucleare[17]. La funzione promozionale dell'AIEA consiste nell'incoraggiare e assistere la ricerca, lo sviluppo, il finanziamento e la pianificazione di progetti nucleari; sostenere l'applicazione pratica dell'energia atomica per uso pacifico; e, ove necessario, agire come intermediario nei contratti di fornitura di attrezzature e materiali nucleari. Dal punto di vista regolatorio, l'AIEA stabilisce e implementa misure di sicurezza come ispezioni, esami e mandati di approvazione per garantire che materiali fissionabili, attrezzature, strutture e informazioni potenzialmente rischiose non siano utilizzati per scopi militari. L'Agenzia stabilisce inoltre determinati standard di sicurezza e provvede alla loro applicazione pratica.

Una critica frequente all'AIEA riguarda l'apparente contraddizione tra le sua funzioni regolamentari e quelle promozionali[18]. Da un lato, infatti, l'AIEA facilita la condivisione di materiali nucleari, tecnologie e competenze tecniche; dall'altro, tenta di scoraggiare la proliferazione di armi nucleari. Tali osservazioni sono state anche supportate da analisi empiriche secondo cui gli Stati che ricevono assistenza tecnica relativa al ciclo del combustibile nucleare attraverso l'AIEA sono più propensi a creare armi nucleari[19]. Altri studiosi, invece, sostengono il contrario, promuovendo la nozione di "disarmo positivo" ed enfatizzando la funzione inibitoria della cooperazione tecnica nel campo nucleare rispetto allo sviluppo di programmi

10. *Consumer Protections for Artificial Intelligence*, SB24-205 (Col., 2024).
11. Roberts–Ziosi–Osborne 2023.
12. ICO 2023.
13. Veale–Matus–Gorwa 2023.
14. CMA 2024.
15. Leslie–Burr–Aitken et al. 2021.
16. UN AI Advisory Body 2023.
17. Fischer 1997.
18. *Ibidem*; Roehrlich 2022.
19. Brown–Kaplow 2014.

nucleari indipendenti[20]. In effetti, il "paradosso" della regolamentazione nucleare si basa sulla falsa dicotomia tra promozione e regolamentazione e su una concezione parzialmente errata dell'impatto dei programmi di assistenza dell'AIEA. L'impatto di questi programmi sulla proliferazione del nucleare a fini militari, infatti, non è comparabile a quello avuto dagli investimenti delle principali nazioni industrializzate[21].

Un'ulteriore critica mossa all'AIEA riguarda la mancanza di sufficienti poteri esecutivi delle misure e degli standard di sicurezza che essa stabilisce. Ad esempio, l'AIEA ha perso il suo potere di ispezione in Corea del Nord quando il paese ha espulso gli ispettori dell'Agenzia dopo aver receduto dal Trattato di non proliferazione nucleare nel 2023[22]. L'AIEA è stata altrettanto inefficace relativamente allo sviluppo di armi nucleari in Iran, Pakistan e Israele, i quali, non avendo firmato il Trattato di non proliferazione nucleare, non sono formalmente soggetti all'autorità dell'Agenzia[23]. Infine, l'AIEA è stata criticata anche per essere stata sfruttata come strumento di legittimazione delle attività militari di determinati Stati. Secondo questa linea di pensiero, gli Stati Uniti e gli altri membri permanenti del Consiglio di Sicurezza delle Nazioni Unite hanno beneficiato dell'AIEA non solo come iniziativa di pace, ma anche come modo per rafforzare la loro egemonia nucleare[24].

4. Governance ambientale: il Panel intergovernativo sul cambiamento climatico

Istituito nel 1998 dal Programma delle Nazioni Unite per l'Ambiente (UNEP) e dall'Organizzazione Meteorologica Mondiale (WMO), l'IPCC è un'istituzione intergovernativa che riunisce 195 Stati membri e diverse organizzazioni osservatrici. La funzione primaria dell'IPCC è fornire analisi regolari e indipendenti sullo stato della scienza del cambiamento climatico agli Stati membri. Sebbene queste analisi siano rilevanti per lo sviluppo di politiche di governance ambientale, esse non sono in alcun modo vincolanti per gli Stati che le ricevono[25].

L'organo principale dell'IPCC è l'assemblea plenaria, dove i rappresentanti degli Stati membri si incontrano per prendere decisioni di natura strategica. Ad esempio, la plenaria definisce il programma annuale di lavoro del Panel, la struttura e gli obbiettivi principali dei vari gruppi di lavoro e i temi da trattare nel rapporto di valutazione annuale. La plenaria approva il rapporto di valutazione annuale e nomina il *Bureau* dell'IPCC, organo di guida tecnica e strategica del Panel. Infine, la plenaria approva la nomina degli esperti proposti dai governi per redigere i rapporti di valutazione. Gran parte dell'attività dell'IPCC si svolge all'interno dei suoi tre gruppi di lavoro e della *Task Force on National Greenhouse Gas Inventories*. Il primo gruppo di lavoro si occupa della valutazione dello stato della ricerca scientifica sul cambiamento climatico; il secondo studia l'impatto del cambiamento climatico sull'ambiente e sulla società; e il terzo considera potenziali strategie per mitigare i rischi del cambiamento climatico. La *Task Force*, invece, sviluppa metodologie per la misurazione delle emissioni di gas serra.

L'IPCC è stato criticato a causa delle pressioni politiche che possono derivare dal coinvolgimento dei governi nella revisione del lavoro degli scienziati[26]. Tuttavia, se da un lato è vero che il processo di adozione dei rapporti di valutazione all'interno dell'IPCC è in parte di natura politica, dall'altro, la natura intergovernativa del Panel è stata fondamentale per garantire l'autorevolezza delle sue valutazioni e un impatto concreto sullo sviluppo di politiche climatiche[27]. Ad esempio, il primo rapporto dell'IPCC pubblicato nel 1990 è stato fondamentale per la creazione della Convenzione

20. Krige–Sarkar 2018.
21. Fischer 1997.
22. Roehrlich 2022; Council of Foreign Relations 2024.
23. Kimball–Bugos 2022.
24. Roehrlich 2022.
25. Vardy–Oppenheimer–Dubash et al. 2017; Bolin 2007.
26. Henderson 2007.
27. Vardy–Oppenheimer–Dubash et al. 2017.

Quadro delle Nazioni Unite sui Cambiamenti Climatici.

Un'altra critica rivolta all'IPCC riguarda potenziali conflitti di interesse all'interno dell'istituto[28]. Ad esempio, nel 2010, il presidente dell'IPCC Rajendra Pachauri fu criticato per presunti legami tra il centro di ricerca da lui diretto a Nuova Delhi e delle aziende interessate a promuovere determinate politiche sul cambiamento climatico. Queste accuse furono respinte dall'indagine che ne seguì, ma contribuirono all'adozione della policy sui conflitti di interesse dell'IPCC nel 2011. Oltre a stabilire misure precauzionali per prevenire conflitti di interesse tra la leadership dell'IPCC e organizzazioni private, questa policy tutela anche gli scienziati da influenze indebite dei governi.

5. Rilevanza dei modelli IAEA e IPCC per un istituto internazionale per la sicurezza dell'IA

I modelli IAEA e IPCC rappresentano due degli esempi più significativi di governance globale e, sotto determinati aspetti, possono fornire degli spunti utili ad arricchire l'attuale dibattito sulla governance dell'IA. Ad esempio, entrambi i modelli affrontano questioni legate alla valutazione e alla gestione del rischio, alla condivisione di conoscenze tecniche e alla cooperazione internazionale. Sia nel caso dell'IA che del cambiamento climatico, una delle sfide principali è gestire l'incertezza che caratterizza le modalità del manifestarsi di potenziali danni e l'entità del loro impatto. Allo stesso tempo, però, i problemi con cui devono confrontarsi l'AIEA e l'IPCC sono sostanzialmente diversi da quelli relativi all'IA. Mentre l'adattamento a processi esogeni è una priorità strategica della governance ambientale e il controllo della proliferazione nucleare per scopi militari è al centro degli sforzi dell'AIEA, l'obbiettivo fondamentale della governance dell'IA consiste nel garantire che le aziende che sviluppano e usano queste tecnologie lo facciano in modo sicuro e responsabile. Pertanto, se tracciare connessioni tra i modelli di governance in diversi ambiti politici può aiutare a pensare in modo creativo a quelle che potrebbero essere le funzioni di un istituto internazionale per la sicurezza dell'IA, è importante avere a mente che problemi distinti richiedono soluzioni altrettanto diverse.

Vari esperti del settore hanno sottolineato che il modello dell'AIEA potrebbe essere il punto di partenza per pensare a un'organizzazione che offre consulenza specialistica e coordinamento internazionale sul tema della sicurezza dell'IA[29]. La società OpenAI ha anche proposto di fondare "qualcosa di simile a un'AIEA per l'IA avanzata", capace di esaminare i modelli di IA prima che vengano messi in commercio, verificarne la conformità a determinati standard di sicurezza ed eventualmente anche imporre restrizioni sulla loro distribuzione[30]. La proposta di un organo di vigilanza globale sull'IA ispirato al modello dell'AIEA è stata appoggiata anche dal Segretario generale delle Nazioni Unite António Guterres[31].

Nonostante ciò, le critiche relative al doppio mandato dell'IAEA sollevano interrogativi simili nel caso di un istituto internazionale per la sicurezza dell'IA basato su un tale modello. Bisognerebbe stabilire, nel caso, se l'istituto possa svolgere attività di promozione e assistenza insieme a quelle di vigilanza. Inoltre, anche in questo caso, esiste il rischio che gli stakeholder più influenti strumentalizzino l'istituto per rafforzare la propria posizione al livello internazionale. Un ulteriore problema dell'analogia tra IA a nucleare è che la governance nucleare riguarda materiali che sono per natura scarsi e il cui utilizzo lascia tracce uniche e misurabili. Anche per questo motivo, lo sviluppo di tecnologie nucleari può essere efficacemente regolamentato attraverso misure di accesso controllato ai materiali, ispezione e approvazione. Nel caso dell'IA, una volta che un modello è addestrato, esso può essere integrato all'interno di varie applicazioni in maniera relativamente semplice e anonima[32]. L'analogia tra IA e nucleare, inoltre, riconduce il tema della sicurezza dell'IA all'interno

28. Schiermeier 2010, p. 596.
29. Marcus–Reuel 2023; Ho–Barnhart–Trager et al. 2023.
30. Altman–Brockman–Sutskever 2023.
31. Guterres 2023.
32. Afina–Lewis 2023.

di un contesto di rischio esistenziale. Un tale approccio concettuale, tuttavia, rischia di essere riduttivo[33]. Limitare l'IA all'interno di un contesto di rischio esistenziale, infatti, significherebbe ignorare la gamma di rischi che l'IA può porre all'autonomia e alla dignità individuale, alla prosperità e all'uguaglianza sociale, nonché all'integrità della sfera dell'informazione pubblica.

A complemento del dibattito sulla governance internazionale dell'IA, un'altra proposta avanzata di recente è quella di un Panel Internazionale sulla Sicurezza dell'IA (IPAIS) ispirato al modello dell'IPCC[34]. Questa proposta si fonda sul presupposto che, per sviluppare un regime regolatorio efficace per l'IA, è necessario che i decisori politici siano consapevoli delle tendenze più recenti e rilevanti nell'ambito della ricerca scientifica di settore. In questo senso, un istituto indipendente guidato da esperti con il compito di informare i governi sullo stato della ricerca scientifica sull'IA sarebbe strumentale a supportare lo sviluppo di un quadro di governance internazionale dell'IA. Oltre a valutare le ricerca in tema di capacità e i rischi dell'IA, l'IPAIS potrebbe anche sviluppare metodologie globalmente condivise per la rendicontazione e la valutazione dei sistemi di IA in modo analogo alla funzione svolta dalla Task Force dell'IPCC. Infine, l'IPAIS potrebbe coordinare una rete internazionale di istituti e organismi di ricerca sulla sicurezza dell'IA, contribuendo così al formarsi di una comunità scientifica specializzata e favorendo l'emergere di nuove collaborazioni e progetti di ricerca.

All'interno del dibattito sull'opportunità o meno di fondare l'IPAIS, è stato fatto notare che un tale istituto rappresenterebbe una piattaforma strategica per le aziende che sviluppano IA interessate ad influenzare i decisori politici[35]. Diversamente dal cambiamento climatico, fenomeno che può essere studiato da scienziati indipendenti in tutto il mondo, l'IA è una tecnologia sviluppata da un numero relativamente ristretto di aziende le quali tendono a non condividere pubblicamente la gran parte dei dati necessari a valutare i loro prodotti in modo indipendente. Per questi motivi, il design istituzionale di un IPAIS dovrebbe garantire una partecipazione del settore privato alle attività dell'Istituto che sia trasparente e orientata al pubblico interesse.

È evidente che i modelli dell'IPCC e della AIEA forniscono spunti potenzialmente utili per riflettere sull'assetto istituzionale e sulle funzioni di un'organizzazione internazionale per la sicurezza dell'IA. Tuttavia, è importante tenerne a mente anche i limiti. Come già visto, infatti, entrambe le istituzioni hanno avuto un impatto limitato, si sono dimostrate almeno in parte suscettibili ad interessi politici ed economici, e i modelli di governance che offrono non sono in grado di cogliere a pieno la natura sociotecnica dei sistemi di IA e le problematiche specifiche che ne derivano.

6. Primi passi verso l'istituzionalizzazione di funzioni di governance per la sicurezza dell'IA

Il Summit sulla sicurezza dell'IA ospitato dal Regno Unito nel novembre 2023 è stato probabilmente il primo tentativo esplicito di istituzionalizzare funzioni di governance per la sicurezza dell'IA su scala globale. L'obiettivo principale del Summit era stabilire un meccanismo di cooperazione internazionale per limitare i rischi legati alle forme di IA più avanzate. Sebbene dal Summit non siano emersi processi concreti ed attuabili per gestire tali rischi, un risultato comunque importante è stata la Dichiarazione di Bletchley. Firmata da 28 paesi e dall'Ue, la dichiarazione ha riconosciuto la necessità di un approccio globale allo studio dell'impatto che l'IA avrà sulla società. Tra le altre cose, i firmatari si sono impegnati a supportare il consolidamento di una comunità scientifica internazionale votata alla ricerca sulla sicurezza dell'IA.

Inoltre, il Summit sulla sicurezza dell'IA è stato pensato per essere un format continuativo. La seconda edizione del Summit si è svolta a maggio 2024 in Corea del Sud e la terza a Parigi nel febbraio 2025. In questo senso, il Summit potrebbe diventare esso stesso il primo meccanismo istituzionalizzato di cooperazione intergovernativa specificamente dedicato alla sicurezza dell'IA. A seguito dell'edizione coreana, i 27 paesi partecipanti e l'Ue hanno firmato la Dichiarazione ministeriale

33. Stewart 2023.

34. Suleyman–Cuellar–Bremmer et al. 2023.

35. Liu 2023.

di Seoul riaffermando la necessità di cooperare a livello internazionale per identificare e mitigare i rischi legati all'IA. Inoltre, una parte di questi paesi ha anche concordato di costituire una rete internazionale di istituti per la sicurezza dell'IA[36].

Un altro risultato della prima edizione del Summit è stata la nomina da parte degli Stati firmatari della Dichiarazione di Bletchley di un panel di esperti con il compito di redigere un report sullo stato della scienza dell'IA. Il report è stato pubblicato poco prima del Summit coreano con l'obbiettivo di proporre una lettura autorevole e ampiamente condivisa dei principali rischi connessi alle forme di IA più avanzata e di informare lo sviluppo di politiche nazionali e internazionali a riguardo[37].

Nel frattempo, sviluppi più concreti stanno prendendo forma a livello nazionale: in vista del primo Summit, il Regno Unito e gli Stati Uniti hanno creato degli istituti nazionali per la sicurezza dell'IA, l'UKAISI e l'USAISI.

L'UKAISI è stato creato per rispondere alla necessità di garantire che i rischi posti dalle forme di IA più avanzate siano valutati in modo indipendente e nel pubblico interesse. L'istituto ha un triplice mandato: (a) sviluppare metodologie per la valutazione tecnica delle capacità potenzialmente dannose dei modelli IA; (b) svolgere ricerche sul tema della sicurezza dell'IA per supportare nuove soluzioni di governance; (c) facilitare la condivisione di informazioni tra decisori politici, aziende, ricercatori e società civile sia a livello nazionale che internazionale. I risultati delle valutazioni condotte dall'Istituto non condizionano il rilascio o la vendita dei modelli IA valutati come rischiosi, ma hanno lo scopo di supportare i processi decisionali sia al livello politico che commerciale. L'UKAISI, inoltre, pubblica le metodologie di valutazione che sviluppa affinché possano essere riutilizzate ed eventualmente migliorate da altri ricercatori. Sebbene sia ancora agli albori della sua attività, l'istituto è stato in parte già criticato per la dubbia efficacia delle sue valutazioni e per la mancanza di accesso ai modelli di IA più recenti e avanzati[38].

Il mandato dell'USAISI corrisponde in parte a quello del suo omologo britannico. È incaricato dal governo americano di (a) sviluppare strumenti e linee guida per valutare le capacità potenzialmente dannose dei modelli di IA; (b) facilitare la condivisione di informazioni e il coordinamento con le controparti nazionali e internazionali; (c) produrre linee guida per lo sviluppo e l'applicazione di regolamenti specifici per l'IA. Nel febbraio 2024, il Segretario al Commercio degli Stati Uniti ha annunciato che l'USAISI sarà supportato dal Consorzio sulla Sicurezza dell'IA. Il Consorzio è una piattaforma che riunisce più di 200 aziende attive nel settore dell'IA, organi pubblici, ricercatori e organizzazioni della società civile per stabilire uno spazio di condivisione di dati e conoscenza e offrire input utili al lavoro di ricerca e valutazione dell'USAISI.

Da una prospettiva funzionale, quindi, gli istituti per la sicurezza dell'IA del Regno Unito e degli Stati Uniti sembrano essere almeno parzialmente ispirati alle caratteristiche proprie dei modelli IAEA e IPCC. Ad esempio, i due istituti svolgono un ruolo chiave nel facilitare la condivisione di conoscenze scientifiche e tecniche. Sviluppano metodologie per valutare l'impatto negativo dell'IA, come l'IAEA fa nel contesto dell'uso militare dell'energia nucleare. Supportano la cooperazione e la collaborazione tra una vasta gamma di parti interessate con l'obiettivo di stimolare il progresso scientifico e l'innovazione tecnologica. Infine, contribuiscono allo sviluppo di quadri di governance pubblicando analisi indipendenti e autorevoli sullo stato dell'arte della ricerca sull'IA.

7. Potenziali funzioni di un istituto internazionale per la sicurezza dell'IA

Gli istituti nazionali per la sicurezza dell'IA del Regno Unito e degli Stati Uniti rappresentano dei passi avanti importanti per lo sviluppo di capacità di valutazione e controllo dei rischi dell'IA da parte delle istituzioni pubbliche. Tuttavia, il carattere transnazionale di tali rischi rende necessario stabilire meccanismi di coordinamento e controllo anche a livello internazionale. Una delle possibili soluzioni a riguardo è la creazione di un istituto internazionale specializzato, similmente a quanto avvenuto nelle aree della governance ambientale e nucleare. Per quanto vi siano opinioni contrastanti

36. DSIT 2024-A; DSIT 2024-B.

37. Bengio–Mindermann–Privitera 2024.

38. Narayanan–Kapoor 2024.

sull'opportunità di creare un organo del tutto nuovo piuttosto che utilizzare strutture e processi esistenti, è comunque utile riflettere sul ruolo e sulle funzioni che un tale istituto potrebbe eventualmente svolgere.

Sulla base di quanto visto nel caso dell'IPCC e dell'AIEA, e considerando gli esempi dell'UKAISI e dell'USAISI al livello nazionale, è possibile articolare il profilo funzionale di un istituto internazionale per la sicurezza dell'IA all'interno di tre categorie tematiche generali: (a) ricerca e cooperazione, (b) audit e verifiche di conformità dei modelli di IA e (c) supporto tecnico per la definizione di quadri di governance dell'IA. Di seguito si propone un elenco non esaustivo di funzioni specifiche rientranti in ciascuna di queste categorie.

7.1. Ricerca e cooperazione

Le principali funzioni di ricerca e cooperazione di un istituto internazionale per la sicurezza dell'IA comprendono: a) la valutazione dello stato della ricerca scientifica sulla sicurezza dell'IA e l'identificazione dei rischi dell'IA attuali e futuri, del loro potenziale impatto sulla società, e dell'efficacia delle soluzioni proposte per gestirli; b) il supporto della ricerca scientifica sulla sicurezza dell'IA attraverso l'identificazione di aree di ricerca di importanza strategica e l'erogazione di incentivi di natura tecnica e finanziaria a sostegno di nuovi progetti di ricerca in queste aree; c) il supporto della cooperazione scientifica internazionale in tema di sicurezza dell'IA fungendo da piattaforma di coordinamento per ricercatori e scienziati, e incentivando la diffusione di conoscenza e competenze tecniche per migliorare la comprensione dei rischi dell'IA da parte di coloro che sviluppano e usano queste tecnologie; d) il supporto tecnico e l'assistenza necessaria a garantire l'accesso a risorse computazionali e dataset per l'addestramento dei modelli di IA da parte delle nazioni in via sviluppo; e) il coordinamento delle attività degli istituti nazionali per la sicurezza dell'IA, incluso il supporto all'uso di metodologie standardizzate e condivise per la valutazione dei rischi legati all'IA.

7.2. Audit e verifiche di conformità dei modelli di IA

Un'altra importante area funzionale di un istituto internazionale per la sicurezza dell'IA riguarda la valutazione tecnica dei modelli di IA. Tra le principali funzioni rientranti in questa categoria vanno menzionati: a) lo sviluppo di processi e metodologie condivisi a livello internazionale per valutare la sicurezza dei modelli di IA e mitigarne le capacità potenzialmente dannose; b) la valutazione dei modelli di IA più avanzati – esempi di tecniche di valutazione particolarmente diffuse sono il *red-teaming*, le *human uplift evaluations*, e le *AI agent evaluations*; c) la definizione di standard di sicurezza e *best practices* per lo sviluppo, l'uso e l'acquisizione di tecnologie IA; d) la verifica della conformità delle policy di sicurezza adottate dalle principali aziende che sviluppano IA rispetto a standard di sicurezza predefiniti e condivisi; e) la pubblicazione gratuita di metodologie per valutare i modelli IA in modo da contribuire alla diffusione di benchmark per la sicurezza dell'IA al livello internazionale.

7.3. Supporto tecnico per la definizione di quadri di governance dell'IA

L'ultima categoria funzionale di un istituto internazionale per la sicurezza dell'IA identificata da questa analisi comprende funzioni di supporto allo sviluppo e consolidamento di quadri di governance internazionale per la sicurezza dell'IA. Ciò include: a) la condivisione di competenze tecniche e informazioni utili a supportare lo sviluppo di politiche per la sicurezza dell'IA, per esempio sintetizzando le ricerche condotte dagli istituti nazionali per la sicurezza dell'IA e condividendone i risultati con una rete di decisori politici nazionali e internazionali per supportare lo sviluppo di soluzioni di governance innovative; b) funzioni di consulenza tecnica e strategica per facilitare lo sviluppo e l'applicazione di norme per la sicurezza dell'IA da parte dei regolatori nazionali.

8. Conclusione

Partendo dalla constatazione della necessità di formalizzare a livello internazionale processi per identificare e gestire le capacità potenzialmente dannose dell'IA, questo articolo propone una riflessione sulle possibili funzioni di un istituto internazionale per la sicurezza dell'IA. Sulla base dell'analisi dei modelli di governance internazionale esistenti in settori adiacenti a quello dell'IA e delle funzioni degli istituti nazionali per la sicurezza dell'IA recentemente costituiti nel Regno Unito e negli Stati Uniti, l'articolo identifica tre categorie

generali di funzioni che potrebbero essere svolte da un istituto internazionale per la sicurezza dell'IA: a) ricerca tecnica e cooperazione, b) audit e verifiche di conformità dei modelli di IA e c) supporto tecnico per la definizione di quadri di governance dell'IA.

L'analisi presentata in questo articolo fornisce una base concettuale per sviluppare ulteriori ricerche in almeno tre ambiti complementari. In primo luogo, e con riferimento allo studio comparato dei modelli dell'IPCC e l'IAEA, sarebbe utile studiarne non solo le funzioni principali e l'impatto ottenuto, ma anche il loro design istituzionale, inclusi i modelli di finanziamento, di *membership*, le strutture di governance interna e i processi che regolano i rapporti con gli stakeholders. In secondo luogo, ulteriori ricerche di natura comparata potrebbero includere una gamma più ampia di istituzioni come la *Financial Action Taskforce* (FATF) e l'Organizzazione Europea per la Ricerca Nucleare (CERN). Inoltre, posto che diversi paesi stanno costituendo istituti nazionali per la sicurezza dell'IA – inclusi Giappone, Singapore, Canada, India, Cina e Corea del Sud – potrebbe essere interessante analizzare quali sono gli elementi in comune tra questi ultimi, l'UKAISI e l'USAISI. Infine, è importante considerare che ci sono diverse opzioni per stabilire processi istituzionalizzati per la sicurezza dell'IA a livello internazionale. Queste includono la creazione *ex novo* di un istituto internazionale per la sicurezza dell'IA, il consolidamento di un network di istituti nazionali per la sicurezza dell'IA, nonché l'attribuzione di nuove competenze e poteri a istituzioni internazionali già esistenti come l'ONU e l'OCSE. Valutare i pro e i contro di ciascuna opzione rappresenta anch'esso un importante ambito di ricerca complementare all'analisi svolta in questo articolo.

Bibliografia

Y. Afina, P. Lewis (2023), *The nuclear governance model won't work for AI*, Chatham House – International Affairs Think Tank, 2023

S. Agrawala (1998-a), *Context and Early Origins of the Intergovernmental Panel on Climate Change*, in "Climatic Change", vol. 39, 1998

S. Agrawala (1998-b), *Structural and Process History of the Intergovernmental Panel on Climate Change*, in "Climatic Change", vol. 39, 1998

S. Altman, G. Brockman, I. Sutskever (2023), *Governance of superintelligence*, OpenAI, 2023

Y. Bengio, S. Mindermann, D. Privitera et al. (2024), *International Scientific Report on the Safety of Advanced AI: Interim Report*, DSIT 2024/009, 2024

B. Bolin (2007), *A History of the Science and Politics of Climate Change: The Role of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 2007

R. Bommansani, D.A. Hudson, E. Adeli et al. (2022), *On the Opportunities and Risks of Foundation Models*, arXiv, 2022

A. Booth (2016), *EVIDENT Guidance for Reviewing the Evidence: a compendium of methodological literature and websites*, Working paper, 2016

R. Brown, J. Kaplow (2014), *Talking Peace, Making Weapons: IAEA Technical Cooperation and Nuclear Proliferation*, in "Journal of Conflict Resolution", vol. 58, 2014, n. 3

CMA (2024), *CMA AI Strategic Update*, UK Competition and Markets Authority, 2024

Council of Europe (2024), *Framework Convention on Artificial Intelligence and HumanRights, Democracy and the Rule of Law*, Council of Europe Treaty Series - No. 225, 2024

Council of Foreign Relations (2024), *Timeline: North Korean Nuclear Negotiations*, Council on Foreign Relations, 2024

K. De Pryck, M. Hulme (eds.) (2022), *A Critical Assessment of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 2022

DSIT (2024-a), *Seoul Ministerial Statement for Advancing AI Safety, Innovation, and Inclusivity: AI Seoul Summit 2024*, UK Department for Science, Innovation, and Technology, 2024

DSIT (2024-b), *Seoul Intent Toward International Cooperation on AI Safety Science, AI Seoul Summit 2024 (Annex)*, UK Department for Science, Innovation, and Technology, 2024

DSIT (2023), *Introducing the AI Safety Institute*, UK Department for Science, Innovation, and Technology, 2023

D. Fischer (1997), *History of the International Atomic Energy Agency: The First Forty Years*, International Atomic Energy Agency, 1997

GPAI (2023), *State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition of Public Release*, Report, Global Partnership on AI, 2023

A. Guterres (2023), *Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight, in First Debate on Artificial Intelligence*, United Nations, 2023

I. Habli (2023), *On the Meaning of AI Safety*, University of York, 2023

D. Henderson (2007), *Unwarranted Trust: A Critique of the IPCC Process*, in "Energy & Environment", vol 18, 2007, n. 7-8

J. Hendrix (2023), *Exploring Global Governance of Artificial Intelligence*, Tech Policy Press, 2023

L. Ho, J. Barnhart, R. Trager et al. (2023), *International Institutions for Advanced AI*, arXiv, 2023

IAEA (1956), *The Statute of the IAEA*, International Atomic Energy Agency, 1956

ICO (2023), *Guidance on AI and data protection*, UK Information Commissioner's Office, 2023

A. Jobin, M. Ienca, E. Vayena (2019), *The global landscape of AI ethics guidelines*, in "Nature", 2019

C.F. Kerry, J.P. Meltzer, A. Renda et al. (2021), *Strengthening International Cooperation on AI*, Brookings, 2021

D. Kimball, S. Bugos (2022), *Timeline of the Nuclear Nonproliferation Treaty (NPT)*, Arms Control Association, 2022

J. Krige, J. Sarkar (2018), *US technological collaboration for nonproliferation: Key evidence from the Cold War*, in "The Nonproliferation Review", vol. 25, 2018, n. 3-4

D. Leslie (2019), *Understanding artificial intelligence ethics and safety*, The Alan Turing Institute, 2019

D. Leslie, C. Burr, M. Aitken et al. (2021), *AI, human rights, democracy and the rule of law: A primer prepared for the Council of Europe*, The Alan Turing Institute, 2021

L. Liu (2023), *Letter: Setting rules for AI must avoid regulatory capture by Big Tech*, in "Financial Times", 27 October 2023

M. Mäntymäki, M. Minkkinen, T. Birkstedt, M. Viljanen (2022), *Defining organizational AI governance*, in "AI and Ethics", vol. 2, 2022

G. Marcus, A. Reuel (2023), *The world needs an international agency for artificial intelligence, say two AI experts*, in "The Economist", 2023

D. Milmo (2023), *AI risk must be treated as seriously as climate crisis, says Google DeepMind chief*, in "The Guardian", 24 October 2023

A. Narayanan, S. Kapoor (2024), *AI Safety is not a model property*, AI Snake Oil, 2024

NIST (2023-A), *U.S. Artificial Intelligence Safety Institute*, 2023

NIST (2023-B), *Artificial Intelligence Risk Management Framework (AI RMF)*, 2023

OECD (2023), *G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a G7 Common Understanding on Generative AI*, OECD Publishing, 2023

OECD (2019), *OECD AI Principles Overview*, Organisation for Economic Cooperation and Development, 2019

Prime Minister's Office (2023), *UK to host first global summit on Artificial Intelligence*, 2023

H. Roberts, E. Hine, M. Taddeo, L. Floridi (2023), *Global AI governance: Barriers and pathways forward*, in "International Affairs", vol. 100, 2023, n. 3

H. Roberts, M. Ziosi, C. Osborne (2023), *A Comparative Framework for AI Regulatory Policy*, Ceimia, 2023

E. Roehrlich (2022), *Inspectors for Peace*, Johns Hopkins University Press, 2022

Q. Schiermeier (2010), *IPCC flooded by criticism*, in "Nature", vol. 463, 2010

S. Steinmo, K. Thelen, F. Longstreth (eds.) (1992), *Structuring Politics: Historical Institutionalism in Comparative Analysis*, Cambridge University Press, 1992

I.J. Stewart (2023), *Why the IAEA model may not be best for regulating artificial intelligence,* in "Bulletin of the Atomic Scientists", 2023

M. Suleyman, M.-F. Cuéllar, I. Bremmer et al. (2023), *Proposal for an International Panel on Artificial Intelligence (AI) Safety (IPAIS): Summary*, Carnegie Endowment for International Peace, 27 October 2023

M. Suleyman, E. Schmidt (2023), *Mustafa Suleyman and Eric Schmidt: We need an AI equivalent of the IPCC*, in "Financial Times", 2023

The American Presidency Project (2023), *FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*, 2023

C. Thomas, H. Roberts, J. Mökander et al. (2024), *The case for a broader approach to AI assurance: addressing 'hidden' harms in the development of artificial intelligence*, in "AI & SOCIETY", 16 May 2024

UN - AI Advisory Body (2023), *Final Report - Governing AI for Humanity*, United Nations, 2023

UN - Office for Disarmament Affairs (1970), *Treaty on the Non-Proliferation of Nuclear Weapons (NPT)*, United Nations, 1970

M. Vardy, M. Oppenheimer, N.K. Dubash et al. (2017), *The Intergovernmental Panel on Climate Change: Challenges and Opportunities*, in "Annual Review of Environment and Resources", vol. 42, 2017

M. Veale, K. Matus, R. Gorwa (2023), *AI and Global Governance: Modalities, Rationales, Tensions*, in "Annual Review of Law and Social Sciences", vol. 19, 2023

L. Weiss (2017), *Safeguards and the NPT: Where our current problems began*, in "Bulletin of the Atomic Scientists", vol. 73, 2017