Commissione europea

GRUPPO DI ESPERTI AD ALTO LIVELLO SULL'INTELLIGENZA ARTIFICIALE

PROGETTO DI ORIENTAMENTI ETICI PER UN'IA AFFIDABILE

SINTESI

Documento di lavoro per la consultazione dei portatori di interessi

Bruxelles, 18 dicembre 2018

PROGETTO DI ORIENTAMENTI ETICI

PER UN'IA AFFIDABILE

Gruppo di esperti ad alto livello sull'intelligenza artificiale
Progetto di orientamenti etici per un'IA affidabile

Commissione europea
Direzione generale della Comunicazione

Referente: Nathalie Smuha - Coordinatrice del gruppo di esperti ad alto livello sull'IA
E-mail: CNECT-HLG-AI@ec.europa.eu

Commissione europea
B-1049 Bruxelles

Documento reso pubblico il 18 dicembre 2018, in inglese.

Il presente documento di lavoro è stato redatto dal gruppo di esperti ad alto livello sull'IA, fatta salva la posizione individuale dei suoi membri su punti specifici e fatta salva la versione finale del documento. Il presente documento sarà ulteriormente sviluppato e la sua versione finale sarà presentata nel marzo 2019 a seguito della consultazione dei portatori di interessi tramite l'Alleanza europea per l'IA.

 Il presente documento è il progetto del primo dei due documenti che saranno presentati del gruppo di esperti ad alto livello sull'IA. La versione finale dello stesso sarà presentata alla Commissione nel marzo 2019. La versione finale del secondo documento – le raccomandazioni sugli investimenti e la politica in relazione all'IA – sarà presentata a metà del 2019.

Ulteriori informazioni sul gruppo di esperti ad alto livello sull'intelligenza artificiale sono disponibili online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

SINTESI

Il presente documento di lavoro costituisce un progetto di orientamenti etici sull'IA redatto dal gruppo di esperti ad alto livello della Commissione europea sull'intelligenza artificiale, la cui versione finale è prevista per marzo 2019.

L'intelligenza artificiale (IA) è una delle forze più trasformative del nostro tempo ed è destinata a modificare il tessuto sociale. Essa costituisce una grande opportunità per aumentare la prosperità e la crescita che l'Europa deve adoperarsi per conseguire. Nel corso dell'ultimo decennio sono stati compiuti enormi passi avanti grazie alla disponibilità di grandi quantità di dati digitali, di potenti architetture di calcolo e di progressi nelle tecniche di IA come l'apprendimento automatico (machine learning). I principali sviluppi che l'IA ha reso possibili per quanto riguarda i veicoli a guida autonoma, l'assistenza sanitaria, i robot domestici/di servizio, l'istruzione o la cibersicurezza stanno migliorando la nostra qualità della vita giorno dopo giorno. L'IA è inoltre fondamentale per affrontare molte delle grandi sfide che il mondo si trova davanti, come la salute e il benessere globali, i cambiamenti climatici, la presenza di sistemi giuridici e democratici affidabili e le altre sfide indicate negli obiettivi di sviluppo sostenibile delle Nazioni Unite.

L'IA ha la capacità di generare enormi vantaggi per gli individui e per la società, ma comporta anche determinati rischi che dovrebbero essere gestiti in modo adeguato. Dato che, nel complesso, i benefici dell'IA superano i rischi, dobbiamo assicurarci di seguire la strada che **massimizza i benefici dell'IA riducendone al minimo i rischi**. A tal fine **è necessario un approccio antropocentrico all'IA**, che ci obblighi a tenere presente che lo sviluppo e l'utilizzo dell'IA non dovrebbero essere considerati come un obiettivo di per sé, ma come un mezzo per aumentare il benessere umano. **Ottenere un'IA affidabile sarà il nostro principio guida**, poiché gli esseri umani saranno in grado di beneficiare appieno e con sicurezza dei vantaggi dell'IA solo se potranno fidarsi della tecnologia.

L'IA affidabile si basa su **due componenti**: 1) rispetto dei diritti fondamentali, della normativa applicabile e dei principi e dei valori di base, garantendo una "**finalità etica**" e 2) robustezza e affidabilità dal punto di vista tecnico poiché, nonostante le buone intenzioni, una scarsa padronanza della tecnologia può involontariamente causare danni.

I presenti orientamenti definiscono quindi il **quadro per un'IA affidabile**.

- Il **capitolo I** si riferisce alle modalità per **garantire la finalità etica dell'IA**, definendo i diritti, i principi e i valori fondamentali che dovrebbe rispettare.
- Partendo da tali principi, nel **capitolo II** viene elaborata una **guida relativa alla realizzazione** di un'IA affidabile, tenendo conto sia della finalità etica che della robustezza tecnica. A tal fine vengono elencati i requisiti per un'IA affidabile e si fornisce una panoramica dei metodi tecnici e non tecnici che possono essere utilizzati per la sua implementazione.
- Il **capitolo III** infine **rende operativi** tali requisiti, fornendo una lista di controllo concreta ma non esaustiva per la valutazione dell'affidabilità dell'IA. Tale lista viene successivamente adattata a casi d'uso specifici.

A differenza di altri documenti che riguardano l'IA etica, i presenti orientamenti non mirano a elencare ancora una volta i valori e i principi fondamentali per l'IA, ma hanno piuttosto lo scopo di fornire una guida su come attuare e rendere concretamente operativi tali principi nei sistemi di IA. Tale guida è fornita su tre livelli di astrazione, dal più astratto capitolo I (diritti, principi e valori fondamentali) al più concreto capitolo III (lista di controllo per la valutazione).

I presenti orientamenti sono destinati a tutti i **pertinenti portatori di interessi che sviluppano, distribuiscono o utilizzano l'IA**, comprendenti imprese, organizzazioni, ricercatori, servizi pubblici,

istituzioni, singoli individui o altre entità. Un meccanismo per consentire ai portatori di interessi di approvare i presenti orientamenti su base volontaria sarà proposto nella versione finale degli stessi.

È importante sottolineare che i presenti orientamenti non intendono sostituire eventuali azioni politiche o di regolamentazione (che saranno trattate nel secondo documento del gruppo di esperti ad alto livello sull'IA, le raccomandazioni sugli investimenti e la politica in relazione all'IA, atteso per maggio 2019), né hanno lo scopo di scoraggiarne l'adozione I presenti orientamenti dovrebbero inoltre essere considerati come un documento vivo che deve essere regolarmente aggiornato nel corso del tempo, al fine di garantirne la costante pertinenza di pari passo con l'evolversi della tecnologia e della nostra conoscenza della stessa. Il presente documento dovrebbe pertanto costituire il punto di partenza per una discussione su **un'IA affidabile "made in Europe"**.

Se da un lato l'Europa può diffondere un approccio etico all'IA solo se è competitiva a livello globale, dall'altro un **approccio etico all'IA è fondamentale per consentire una competitività responsabile**, in quanto accrescerà la fiducia degli utenti e faciliterà una più ampia diffusione dell'IA. I presenti orientamenti non intendono soffocare l'innovazione basata sull'IA in Europa, quanto piuttosto proporre l'etica come principio ispiratore per lo sviluppo di un'IA unica, che mira a proteggere e favorire sia gli individui che il bene comune. È questo che consente all'Europa di affermarsi come leader nel campo dell'IA etica, sicura e all'avanguardia. Solo garantendo l'affidabilità dell'IA i cittadini europei potranno coglierne tutti i benefici.

Infine, al di là dell'Europa, i presenti orientamenti hanno inoltre l'obiettivo di **promuovere la riflessione e la discussione** su un quadro etico per l'IA a **livello globale**.

INDICAZIONI ESECUTIVE

Ciascuno dei capitoli dei presenti orientamenti fornisce indicazioni su come ottenere un'IA affidabile, indirizzate a tutti i portatori di interessi che sviluppano, distribuiscono o utilizzano l'IA, come riassunto qui di seguito.

Capitolo I: indicazioni chiave per garantire finalità etiche

- Adoperarsi per un'IA **antropocentrica**: l'IA dovrebbe essere sviluppata, distribuita e utilizzata con "**finalità etiche**" che riflettono i diritti fondamentali, i valori sociali e i principi etici su cui sono basate, vale a dire *beneficenza* (fare del bene), *non maleficenza* (non nuocere), *autonomia degli umani*, *giustizia* ed *esplicabilità*. Ciò è essenziale per raggiungere l'obiettivo di un'**IA affidabile**.
- Basarsi sui diritti fondamentali, i principi etici e i valori sociali per valutare i possibili effetti futuri dell'IA sugli esseri umani e sul bene comune. Prestare **particolare attenzione** alle situazioni che coinvolgono **gruppi più vulnerabili**, come i bambini, le persone con disabilità o le minoranze, o a situazioni in cui si verificano **asimmetrie di potere o di informazione**, ad esempio tra datori di lavoro e lavoratori, o tra imprese e consumatori.
- Riconoscere e essere consapevoli del fatto che, pur apportando benefici sostanziali agli individui e alla società, l'IA può avere anche conseguenze negative. Mantenere alta la guardia per gli ambiti più critici.

Capitolo II: indicazioni chiave per realizzare un'IA affidabile

- Integrare **fin dalla fasi iniziali della progettazione** i **requisiti di affidabilità** per l'IA: responsabilità, governance dei dati, progettazione per tutti, governance dell'autonomia dell'IA (sorveglianza umana), non discriminazione, rispetto dell'autonomia umana, rispetto della privacy, robustezza, sicurezza, trasparenza.

- Prendere in considerazione metodi tecnici e non tecnici per garantire l'attuazione di tali requisiti nel sistema di IA. Tenere inoltre conto di tali requisiti al momento di creare il gruppo che lavorerà sul sistema, il sistema stesso, l'ambiente di prova e le potenziali applicazioni del sistema.
- Fornire, in modo chiaro e proattivo, **informazioni ai portatori di interessi** (clienti, dipendenti, ecc.) in merito alle capacità e alle limitazioni del sistema di IA, in modo che abbiano aspettative realistiche. A questo proposito è fondamentale garantire la **tracciabilità** del sistema di IA.
- Integrare l'IA affidabile nella cultura dell'organizzazione e fornire informazioni ai portatori di interessi su come l'IA affidabile viene implementata nella progettazione e nell'utilizzo dei sistemi di IA. L'IA affidabile può essere anche inserita nelle carte deontologiche o nei codici di condotta delle organizzazioni.
- Garantire la partecipazione e l'**inclusione dei portatori di interessi** nella progettazione e nello sviluppo del sistema di IA. Garantire inoltre la **diversità** nella definizione dei gruppi che sviluppano, implementano e testano il prodotto.
- Adoperarsi per **agevolare la verificabilità** dei sistemi di IA, in particolare in contesti o situazioni critiche. Per quanto possibile progettare il sistema in modo da poter ricondurre le singole decisioni ai diversi input forniti: dati, modelli pre-addestrati, ecc. Definire inoltre i **metodi di giustificazione** del sistema di IA.
- Garantire un processo specifico per la **governance della responsabilità**.
- Prevedere possibilità di **formazione e istruzione** e garantire che i manager, gli sviluppatori, gli utenti e i datori di lavoro conoscano l'IA affidabile e ricevano la necessaria formazione per usarla.
- Essere consapevoli che vi potrebbero essere tensioni fondamentali tra i diversi obiettivi (la trasparenza potrebbe condurre a un uso improprio; l'individuazione e la correzione delle distorsioni potrebbe essere in contrasto con la protezione della privacy). Comunicare e documentare tali contrapposizioni.
- Promuovere la ricerca e l'innovazione al fine di agevolare il rispetto dei requisiti dell'IA affidabile.

Capitolo III: indicazioni chiave per valutare un'IA affidabile

- Adottare una lista di controllo per la valutazione dell'IA affidabile all'atto dello sviluppo, della distribuzione o dell'utilizzo dell'IA e adattarla allo specifico caso d'uso del sistema.
- Tenere presente che la lista di controllo per la valutazione **non sarà mai esaustiva** e che garantire un'IA affidabile non è una questione di caselle da spuntare, ma è un processo continuo di individuazione dei requisiti, valutazione delle soluzioni e miglioramento dei risultati durante l'intero ciclo di vita del sistema di IA.

Le presenti indicazioni fanno parte di una visione che adotta un approccio antropocentrico all'intelligenza artificiale, che consentirà all'Europa di diventare un innovatore di punta a livello mondiale nel campo dell'IA etica, sicura e all'avanguardia. Esse inoltre faciliteranno e renderanno possibile un'**IA affidabile "made in Europe"**, che migliorerà il benessere dei cittadini europei.