IA generale, rischio, allineamento

Federico L.G. Faroldi — federico.faroldi@unipv.it

Università di Pavia
CHAI – University of California at Berkeley

Convegno annuale di informatica giuridica, Collegio Ghislieri, Pavia
23 novembre 2023

Contenuti

- Definizioni
- Il rischio dell’AGI
- Il rischio nell’EU AI Act
- AGI, rischio e EU AI Act
- Conclusioni

Background

- EU AI Act adotta un approccio basato sul rischio. Per i sistemi ad alto rischio, tra le altre disposizioni, i fornitori devono adottare un sistema di gestione del rischio, basato sull'identificazione e la valutazione dei rischi che possono emergere dagli usi previsti o dagli usi impropri prevedibili; se questi rischi sono inaccettabili, devono essere progettate e attuate misure di mitigazione e controllo (art. 9).

- What about AGI? La proposta originaria della Commissione, ho sostenuto in Faroldi 2021, non poteva occuparsi di AGI, poiché le disposizioni dipendono in modo cruciale dal fatto che un sistema di IA abbia un "intended purpose".

- Questo intervento si propone di indagare se e come i sistemi basati sul rischio possano essere estesi alle AGI, partendo dal contesto dell'EU AI Act.

Definizioni

- In mid-November 2023, the OECD has adapted its definition of AI system, which is set to be included in the EU AI Act, as follows:

“An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment,”

- the OECD definition allows for an AI system that can also adapt after deployment, and the objectives don’t have to be specified by humans: these are features that make the definition compatible with AGI

Definizioni

The Commission original proposal does not take into account AGI neither explicitly, nor implicitly (see Faroldi 2021), but in November 2023 proposed the following:

"'General-purpose AI model' means an AI model, including when trained with a large amount of data using self-supervision at scale, that is capable to [competently] perform a wide range of distinctive tasks regardless of the way the model is released on the market

By EP

1d) 'general purpose AI system' means an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed

By Council

(1b) 'general purpose AI system' means an AI system that - irrespective of how it is placed on the market or put into service, including as open source software - is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems;

Definizioni

In letteratura non c'è accordo, ovviamente, su come definire con precisione l'AGI

1. sistemi altamente autonomi che superano gli esseri umani nella maggior parte dei lavori economicamente validi (OpenAI)

2. alcuni ritengono che i LLM allo stato dell'arte (ad esempio le implementazioni a metà del 2023 di GPT-4, Bard, Llama 2 e Claude) siano già delle AGI, sostenendo che la generalità è la proprietà chiave delle AGI e che, poiché i modelli linguistici possono discutere di un'ampia gamma di argomenti, eseguire un'ampia gamma di compiti, gestire input e output multimodali, operare in più lingue e "apprendere" zero-shot, hanno raggiunto una generalità sufficiente

Definizioni

Definizione del rischio solo da parte del PE:

(1 bis) "rischio": la combinazione della probabilità che si verifichi un danno [*harm*] e la gravità [*severity*] di tale danno

Questa definizione di rischio coincide con la clausola 3.9 della Guida ISO/IEC 51:2014 Aspetti della sicurezza - Linee guida per la loro inclusione nelle norme, dove il "danno" è ulteriormente definito come qualsiasi effetto negativo sulla salute, sulla sicurezza e sui diritti fondamentali (in realtà "lesioni o danni alla salute delle persone, o danni alla proprietà o all'ambiente", clausola 3.1). La clausola 5 definisce la "probabilità che si verifichi un danno" come "una funzione dell'esposizione al pericolo [*hazard*], del verificarsi di un evento pericoloso, delle possibilità di evitare o limitare il danno".

Il rischio può essere definito in modo diverso: in particolare la ISO 31000:2018 Risk Management - Guidelines definisce il rischio come un "effetto dell'incertezza sugli obiettivi (Clausola 3.1)".

Il rischio dell’AGI

Il **rischio esistenziale** è definito come un evento che minaccia l'estinzione prematura dell'umanità o che riduce in modo permanente la sua capacità di prosperare.

Per i **rischi non antropici**, si può pensare a una pandemia naturale, a un'eruzione vulcanica o all'impatto di un asteroide. Per i **rischi antropici**, si può pensare a una pandemia non naturale, a una guerra nucleare totale o a un'intelligenza generale artificiale che prende il controllo.

L’**argomento del controllo** di Ngo (2020):

“We’ll build AIs which are much more intelligent than humans (i.e. super- intelligent).

Those AIs will be autonomous agents which pursue large-scale goals.

Those goals will be misaligned with ours; that is, they will aim towards outcomes that aren’t desirable by our standards, and trade off against our goals.

The development of such AIs would lead to them gaining control of humanity’s future.”

Il rischio dell’AGI

Hendrycks et al. (2023)

1. uso doloso: L'Ais può essere utilizzato di proposito da attori malintenzionati. I casi includono il bioterrorismo, la propaganda e la sorveglianza. Le strategie di mitigazione proposte includono la biosicurezza, la limitazione dell'accesso a modelli potenti e la responsabilità legale per gli sviluppatori.

2. Corsa all'IA: la pressione internazionale potrebbe indurre gli Stati a sviluppare potenti IA senza controllo e a cedere loro il potere. I casi includono armi autonome letali e guerra automatizzata, automazione del lavoro. Le strategie di mitigazione proposte includono l'implementazione di norme di sicurezza, il coordinamento internazionale e il controllo pubblico delle Ais di uso generale.
3. Rischi accidentali o organizzativi: questi rischi si collocano nella regione dei rischi tradizionali meglio compresi, come le fughe di laboratorio, i furti da parte di malintenzionati o l'incapacità dell'organizzazione di investire nella sicurezza dell'IA. Le strategie di mitigazione proposte includono verifiche interne ed esterne, più livelli di difesa contro i rischi e una sicurezza informatica all'avanguardia.

4. Rogue AI: è il timore che gli esseri umani perdano il controllo sulle Ais, man mano che queste diventano più intelligenti di noi. Questo potrebbe generare spostamenti di obiettivi, comportamenti di ricerca di potere e inganni. Le strategie di mitigazione proposte hanno presumibilmente a che fare con le strategie di allineamento.

Il rischio dell’AGI

Harari ha recentemente evidenziato la possibilità di una massiccia crisi finanziaria generata dalle AGI, dal momento che la finanza è "solo dati" e già fortemente digitalizzata. Di per sé, questo non sarebbe un rischio catastrofico, ma potrebbe innescare una catena di conflitti potenzialmente minacciosi per l'esistenza.

Altri rischi estremi citati da Ringel Morris et al. (2023) sono che "i sistemi AGI potrebbero essere in grado di ingannare e manipolare, accumulare risorse, perseguire obiettivi, comportarsi in modo agenzia, superare gli esseri umani in più domini, spostare gli esseri umani da ruoli chiave e/o auto-migliorarsi ricorsivamente". I rischi sono legati al livello di autonomia, con i sistemi agenziali più autonomi come i più rischiosi.

Il sistema di gestione dei rischi nell’EU AI Act – art 9(2)

Il sistema di gestione dei rischi è costituito da un processo iterativo continuo eseguito nel corso dell'intero ciclo di vita di un sistema di IA ad alto rischio, che richiede un aggiornamento costante e sistematico. Esso comprende le fasi seguenti:

a) identificazione e analisi dei rischi noti e prevedibili associati a ciascun sistema di IA ad alto rischio;

b) stima e valutazione dei rischi che possono emergere quando il sistema di IA ad alto rischio è usato conformemente alla sua finalità prevista e in condizioni di uso improprio ragionevolmente prevedibile;

c) valutazione di altri eventuali rischi derivanti dall'analisi dei dati raccolti dal sistema di monitoraggio successivo all'immissione sul mercato di cui all'articolo 61;

d) adozione di adeguate misure di gestione dei rischi conformemente alle disposizioni dei paragrafi seguenti.

Il sistema di gestione dei rischi nell’EU AI Act – art 9(2)

L'**identificazione** del rischio consiste nell'uso sistematico delle informazioni disponibili per identificare i pericoli, laddove un pericolo è una potenziale fonte di danno.

La **stima** del rischio è il calcolo della probabilità che si verifichi un danno e della sua gravità, mentre la **valutazione** del rischio consiste nel determinare se un rischio è accettabile.

Queste misure di gestione del rischio si basano su: **progettazione e sviluppo**; misure di **mitigazione e controllo**; informazione e formazione. Ulteriore componente del sistema di gestione del rischio è la fase di **testing,** per identificare le migliori misure di gestione del rischio, per garantire la coerenza e per assicurare la conformità legale.

L’AGI è una fonte di rischio esistenziale?

supponiamo che un’AGI sia abbastanza autonoma e che prenda da sola molte decisioni, è difficile prevedere non solo cosa farà, ma anche come. Il rischio esistenziale in questo caso non è necessariamente un grande evento catastrofico, ma una somma crescente di eventi più piccoli che, presi insieme, costituiscono un rischio significativo.

a questo punto possiamo solo fare ipotesi su come potrebbe essere un'IA. In ogni caso, sembra che ci sia un consenso sul fatto che l'intelligenza artificiale possa aumentare le possibilità di un'IA disonesta (vedi sopra), soprattutto se non viene sviluppata tenendo conto della sicurezza e dell'allineamento. Pertanto, possiamo forse affinare il nostro pensiero e dire che un'IA che ottiene un punteggio insufficiente in termini di sicurezza e allineamento è un rischio esistenziale.

L'AGI è automaticamente ad alto rischio?

Considerando, tuttavia, che l'AGI sarà ragionevolmente autonoma, un principio di cautela suggerisce che dovrebbe, come minimo, essere classificata come ad alto rischio.

Passaggio da un sistema con uno scopo specifico a un sistema generale e autonomo, ossia che (i) può trovare nuovi mezzi per raggiungere un determinato fine; (ii) può perseguire obiettivi intermedi o strumentali nuovi e inaspettati; (iii) può persino perseguire un *obiettivo finale* nuovo.

dipende dal successo degli sviluppatori nell'affrontare il problema dell'allineamento.

il Consiglio ha proposto che, invece dell’intende purpose per i sistemi generali si faccia riferimento ai possibili usi come sistemi ad alto rischio, e che a sistemi generali si applichi il regime dei sistemi a rischio elevato

Il rischio di AGI può essere gestito secondo il sistema di gestione del rischio?

- a) L'identificazione e l'analisi dei rischi noti e prevedibili è impossibile, perché non sappiamo come sarà l'AGI o di cosa sarà capace esattamente, e quindi non possiamo usare le informazioni disponibili. Se si conosce, significa che la tecnologia esiste già, ma non è così.

- b) stima e valutazione dei rischi che possono emergere: la stima è difficile quando si tratta di probabilità e impossibile quando si tratta di gravità, perché i rischi catastrofici ed esistenziali sono impossibili da quantificare su scale familiari; la valutazione sembra possibile, perché questi rischi esistenziali non dovrebbero mai essere accettabili.

Il rischio di AGI può essere gestito secondo il sistema di gestione del rischio?

Adozione di adeguate misure di gestione del rischio, ulteriormente dettagliate nell'articolo 9, paragrafi 3 e 4:

- I fornitori devono progettare e sviluppare il sistema in modo da eliminare o ridurre il più possibile i rischi: ciò sembra richiedere che l'IA sia allineata. Ma anche per le IA ristrette, non ci sono garanzie teoriche che un sistema sia allineato e ci sono prove empiriche che il disallineamento è robusto in tutte le tecniche di apprendimento automatico, con fenomeni come lo "scheming" e la "goal misgeneralization"

.

Il rischio di AGI può essere gestito secondo il sistema di gestione del rischio?

- Se i rischi non possono essere eliminati, i fornitori devono implementare adeguate misure di mitigazione e controllo: questo sembra difficilmente possibile proprio perché l'argomento è che l'AGI, se disallineata, può prendere il controllo.

- I fornitori devono fornire informazioni e formazione adeguate agli utenti: anche questo punto sembra difficilmente applicabile, nella misura in cui se un sistema AGI è veramente autonomo, sarà presumibilmente molto difficile prevedere in quali modi creerà rischi catastrofici o esistenziali, una volta che è stato utilizzato.

Anche le procedure di test sono fuori discussione, perché non si può implementare un sistema AGI se non si eliminano in modo sicuro i rischi menzionati in precedenza.

Rischio da AGI e EU AI Act

il Consiglio non ha capito cosa sia l'IA generale, o che per "IA a scopo generale" intende semplicemente qualcosa che non può essere realmente autonomo in alcun modo, ma un'IA ristretta con più strumenti

la nozione di "possibile uso" sembra del tutto inefficace per circoscrivere i rischi derivanti dai sistemi di IA per scopi generali.

se l'AGI pone, o può porre, un rischio esistenziale, tale rischio sembra sfuggire ai comuni sistemi di gestione del rischio, poiché non appare facilmente quantificabile in termini di gravità, per non parlare della probabilità.

ignora le IA generali, in quanto lo scopo previsto viene sostituito con "possibile", che sembra di portata impossibile, e quindi diminuisce la certezza del diritto, aumentando le responsabilità dei fornitori; in secondo luogo, ignora che, anche se i fornitori sono fuori dai guai, il resto dell'umanità può essere esposto a rischi esistenziali, in un caso clamoroso di esternalità negative. Tali rischi esistenziali devono essere prevenuti tout court

Considerazioni conclusive

la definizione di IA per scopi generali non è sbagliata, ma la normativa è fuori bersaglio. In altri termini, il regolatore non capisce cosa sia, o possa essere, un sistema di IA per scopi generali.

Questo perché se comprendessero il suo potenziale trasformativo e autonomo, non cercherebbero di adattarlo a una struttura legislativa piuttosto conservativa, basata sui prodotti, che può essere al massimo adeguata per un'IA ristretta.

Considerazioni conclusive

i precedenti evidenziano che questo sistema di gestione del rischio di vecchio stampo non è adeguato per regolamentare l'IA, soprattutto i sistemi di uso generale.

In terzo luogo, il fatto che l'AGI sfugga a questo tipo di gestione del rischio non è per ragioni contingenti, ma piuttosto per ragioni concettuali. In particolare, questo tipo di sistema di gestione del rischio non può affrontare adeguatamente i rischi catastrofici o esistenziali.

Tali rischi non possono essere facilmente individuati, né stimati, a causa della loro entità fuori scala e (si spera) della loro bassissima probabilità.

Non si possono fare prove ed errori con rischi esistenziali.

Considerazioni conclusive

Resta aperta la questione della misura in cui i rischi catastrofici o esistenziali possono essere affrontati in fase di progettazione, data la loro natura.
Certo, è difficile capire come possano emergere in una fase di test, a causa di fenomeni come "scheming" e "goal misgeneralization".

Il fallimento della regolamentazione dell'AGI come **rischio normativo**
questa proposta costituisce quello che ho definito un rischio normativo.[11]
Nell'accezione pertinente,il rischio normativo, in sintesi, è il rischio esistenziale relativo a norme.
Se il legislatore europeo non riesce a regolamentare l'AGI, e quindi a diminuire il rischio esistenziale connesso, sarà responsabile del rischio normativo in questo senso.

Considerazioni conclusive

La raccomandazione politica è la seguente: qualcosa, come l'AGI, il cui rischio non può essere né calcolato né gestito, dovrebbe essere vietato al pubblico, perché non può essere trattato con le disposizioni della categoria ad alto rischio, in quanto la gestione del rischio non funziona per ragioni teoriche.

La ricerca sull'intelligenza artificiale sembra accettabile solo in una struttura intergovernativa, altamente controllata.

Grazie per l’attenzione – federico.faroldi@unipv.it