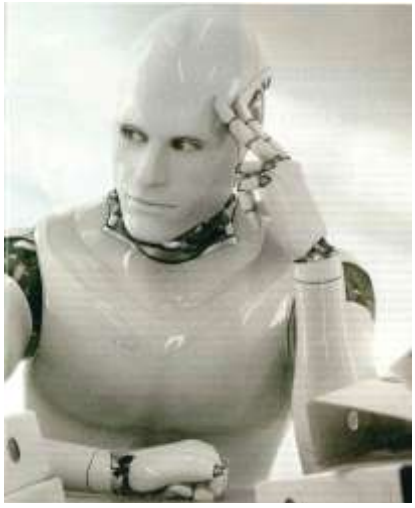
**Giuliano Aluffi**

***Se un robot diventa troppo intelligente.***
[Venerdì di Repubblica, 30-01-2015]



*Entro il 2040 potrebbe esistere una macchina che pensa come noi (e, in breve, molto meglio). Una prospettiva che comporta dei rischi?*
*Se lo sono chiesti 400 scienziati ed esperti.*
*Incluso il filosofo svedese Nick Bostrom. Che abbiamo sentito e che qui propone una soluzione.*

Che succederà quando l'intelligenza artificiale supererà quella umana, e un calcolatore che ragiona a velocità impressionante sarà autocosciente, proprio come il super computer di bordo *Hal 9000* nel film *2001 Odissea nello spazio* di Stanley Kubrick?

È la domanda che si sono posti quattrocento esperti, tra i quali scienziati del Mit e di Oxford, ma anche ricercatori di Ibm e di Google.

Tutti insieme nelle scorse settimane hanno firmato una lettera aperta sui rischi futuri dell'intelligenza artificiale, che secondo alcuni potrebbe rivelarsi la maggiore minaccia esistenziale per l'umanità.

Tra i più preoccupati ci sono anche il grande astrofisico Stephen Hawking ed Elon Munsk, il guru tecnologico a capo dell'azienda automobilistica Tesla Motors e di quella di trasporti spaziali Space X.

Uno dei massimi esperti sul tema, anche lui firmatario dell'appello, è il filosofo svedese


**Nick Bostrom**

direttore del *Future of Humanity Institute* dell'Università di Oxford
autore del saggio
*Superintelligence: Paths, Dangers, Strategies*
(Oxford University Press, pp. 272, euro 22).

Abbiamo chiesto a Bostrom quanto è reale la «minaccia».

**Siamo davvero in grado di realizzare un'intelligenza artificiale equivalente a quella umana e capace di dialogare con noi?**
«Ci sono due approcci. Il primo, seguito dallo *Human Brain Project* del Politecnico di Losanna, si fonda sulle neuroscienze e ha l'obiettivo di riprodurre con precisione assoluta su un computer l'intera struttura neuronale umana, fino alle più minute sinapsi.
La domanda è: una volta ottenuto un cervello digitale che, nella sua struttura, simula al cento per cento quello umano, questo sarà anche autocosciente? E
È da vedere. Il vantaggio di questo sistema è che non serve capire davvero come - e perché - funzioni il cervello: basta emularlo in tutti i suoi aspetti fisici e osservare il risultato.
Il secondo approccio, invece, cerca di ottenere un'intelligenza pari alla nostra usando algoritmi capaci di auto-modificarsi ed evolvere apprendendo».

**Questo secondo sistema è quello su cui oggi lavora la maggior parte dei ricercatori.**
«Sì, ma per ora sono ricerche di ambito ristretto e specifico.
Ci sono per esempio i sistemi grazie ai quali le automobili possono guidarsi da sole, studiati dal Mit e da Google, o i droni militari del tutto autonomi
(*quelli oggi impiegati in guerra sono tutti pilotati a distanza*, ndr).

E già a questo livello nascono seri problemi etici, posti dai 400 esperti che hanno sottoscritto l'appello. Di chi è la responsabilità se un sistema di guida automatica causa un incidente? E possiamo permettere a un drone di decidere da solo l'uccisione di un uomo?

Altri esempi, dai risvolti meno drammatici, sono i computer, come *Watson* di Ibm, imbattibile nei quiz (nel 2011 ha sbancato il quiz tv *Jeopardy* negli Usa) e oggi applicato alla ricerca medica: al *New York Genome Center* studia il genoma cercando terapie personalizzate per il cancro.
Ma un'intelligenza artificiale duttile e autocosciente come un uomo sarà qualcosa di radicalmente diverso: un vero *Hal 9000*. Secondo i massimi esperti mondiali, interpellati dall'Università di Oxford, c'è almeno il 50 per cento di probabilità che qualcosa del genere venga realizzato entro il 2040».

**E cosa succederà dopo?**
«Che ben presto la IA. si evolverà in una superintelligenza. Magari aumentando la sua velocità di elaborazione (mentre noi rimaniamo vincolati dai limiti naturali del nostro cervello).

Per fare un esempio: io ragiono meglio di Homer Simpson.
Ma non posso competere con un Homer Simpson centomila volte più rapido di me a *processare le informazioni*: nel tempo che io impiego a leggere un libro, "lui" avrà letto una biblioteca.
Oppure *si riprodurrà* in milioni di copie e le farà collaborare in rete.
Se anche ognuno di questi cervelli software, preso da solo, fosse intelligente come un uomo non particolarmente sveglio, l'intelligenza collettiva sarebbe irraggiungibile da qualsiasi controparte umana.

E poi un computer più intelligente di noi sarà un progettista molto più dotato di qualsiasi umano: potrà quindi ideare un computer ancora più intelligente di sé, e così via, in tempi sempre più ridotti. Assisteremmo a un'esplosione di intelligenza capace di cambiare il corso della storia».

**Cosa potrebbe volere una superintelligenza?**
«E importante distinguere tra scopi finali e scopi strumentali.

Immaginiamo di avere un computer intelligente programmato dall'uomo per produrre il maggior numero di graffette possibile. Questo sarà il suo scopo finale. Ma, per raggiungerlo, ci saranno vari scopi strumentali che l'intelligenza artificiale metterà in atto. Ed è D. che si rischia».

**Un esempio?**
«Anzitutto la macchina vorrà preservarsi e non essere disattivata: continuare a esistere è uno scopo strumentale indispensabile per qualsiasi scopo finale.
Vorrà poi migliorarsi e acquisire sempre più risorse, così da produrre più graffette e in modo più efficiente.
Il rischio è ottenere un effetto perverso da *Topolino apprendista stregone*: uno scopo in sé innocuo potrebbe portare le macchine ad azioni impreviste e dannose, come, per esempio, prendere il controllo di tutte le centrali elettriche e usare tutta l'energia del Pianeta per produrre graffette».

**Immaginare tutto questo può aiutarci a prevedere il comportamento delle intelligenze artificiali e a orientarlo?**
«No, perché le macchine potrebbero decidere di perseguire questi scopi in modi tortuosi, con sequenze di azioni che una per una potrebbero sembrarci tutt'altro.
Possiamo prevedere gli scopi strumentali che una superintelligenza potrà avere, ma non le specifiche azioni che sceglierà per realizzarli.
E i nostri desideri e bisogni potrebbero rivelarsi un ostacolo per uno o più di questi scopi strumentali, e quindi anche per lo scopo finale, qualunque esso sia. La superintelligenza potrebbe quindi, prima o poi, ritenere opportuno rimuovere l'ostacolo. Cioè noi».

**Ma non c'è nessun modo per tenere sotto controllo una superintelligenza?**
«Potremo cercare di confinarla in una "scatola", un computer isolato e senza accesso a internet. Ma non appena inizieremo a comunicare con essa — e vorremo farlo, altrimenti a che ci serve? — rischieremo: un'intelligenza superumana sarà abile nel manipolarci.
Anzitutto, potrebbe persuaderci a lasciarla uscire dalla "scatola" promettendoci chissà cosa.
Poi magari ci convincerà a darle il controllo di "agenti" fisici, come robot e droni».

**Come potremo tutelarci?**
«Se il passaggio da intelligenza artificiale di livello umano a superintelligenza sarà gra-duale, in mesi o anni, avremo molte più probabilità di poterlo misurare con dei test di intelligenza e cercare sistemi di controllo appropriati.
Se sarà repentino, invece, saremo impreparati.
Inoltre se l'intelligenza artificiale ci fosse ostile, avrebbe interesse a essere sottovalutata, per non allarmarci. Quindi potrebbe fallire di proposito dei test, almeno fino a quando non ritenesse di aver raggiunto una posizione di vantaggio ormai irrecuperabile».

**Insomma, potrebbe ingannarci.**
«È inevitabile che una superintelligenza scopra l'utilità tattica dell'inganno».

**Che fare, allora?**
«Invece di porre limiti a ciò che la I.A. potrà fare, strategia insostenibile nel medio-lungo termine, dovremo porli a ciò che vorrà fare. Magari realizzando una I.A. "bambina" e aiutandola a svilupparsi introiettando valori rispettosi dell'umanità.
Ma sono valori difficili da codificare e che noi stessi non sappiamo rispettare, come ci ricorda la Storia».