Scuola Internazionale Superiore
per la Ricerca Interdisciplinare
Summer Research Meeting

Quale Fine per una Superintelligenza?

Giovanni Amendola

Roma, 26 luglio 2019

Sommario

- **Una definizione di intelligenza**
- **Alcune tipologie di Intelligenza Artificiale**
- **Rapporto tra IA e fini/obiettivi**
- **Intelligenza, fine e consapevolezza**

Una definizione di intelligenza

«L’idea convenzionale fra i ricercatori nel campo dell’intelligenza artificiale è che **l’intelligenza** in ultima istanza **abbia a che fare solo con informazione e computazione**, **non con carne, sangue o atomi di carbonio**. Questo significa che non esiste una fondamentale ragione per cui le macchine un giorno non possano essere intelligenti quanto noi»

(M. TEGMARK, *Vita 3.0. Essere umani nell’era dell’intelligenza artificiale*, Raffaello Cortina Editore, 2018, 82)

Una definizione di intelligenza

Intelligenza
=
capacità di realizzare fini complessi

(M. TEGMARK, *Vita 3.0.*, 76)

Max Tegmark
(Stoccolma, 1967)
cosmologo al MIT

Una definizione di intelligenza

Intelligenza = capacità di realizzare fini complessi

- **Capacità di ragionamento logico**
- **Comprensione**
- **Pianificazione**
- **Conoscenza emotiva**
- **Autoconsapevolezza**
- **Creatività**
- **Risoluzione di problemi**
- **Apprendimento**
- **Acquisire e applicare conoscenze e competenze**

“Sono tutti esempi di possibili **fini complessi”**

Una definizione di intelligenza

Intelligenza = capacità di realizzare fini complessi

Esempio 1: Un programma **A** per giocare a scacchi ha come fine quello di vincere una partita a scacchi.

Esempio 2: Un programma **B** per giocare a Go ha come fine quello di vincere una partita a Go.

Chi è più intelligente?
A e **B** non sono confrontabili.
Ma un programma **C** sarebbe più intelligente di **A** e **B** se riuscisse a raggiungere equamente i fini di entrambi e giocare meglio di almeno uno dei due.

Tipologie di Intelligenza Artificiale

Tipologie di Intelligenza Artificiale

IA Ristretta = IA Debole = capacità di raggiungere ***un insieme limitato di fini*** (come giocare a scacchi o guidare un'auto)

Esempio: *DeepMind* (un sistema di IA di Google) può giocare a decine di differenti videogiochi classici della Atari alla pari e **anche meglio di un essere umano**.

IA Generale (IAG) = IA di livello umano = IA Forte = capacità di svolgere (1) ***qualsiasi compito cognitivo*** (2) ***almeno tanto bene quanto un essere umano***

Nessuna IA è finora in grado. L'essere umano manifesta invece una intelligenza generale

Tipologie di Intelligenza Artificiale

IA Generale (IAG) = IA di livello umano = IA Forte =
capacità di svolgere (1) ***qualsiasi compito cognitivo***
(2) ***almeno tanto bene quanto un essere umano***

Il termine **“Artificial General Intelligence” (AGI)**
2002: da **Shane Legg** o
1997: da **Mark Gubrud**
(vedi, http://goertzel.org/who-coined-the-term-agi/)

Il termine **“Strong AI” (Intelligenza Artificiale Forte)**
1980: da **John Searle**

Tipologie di Intelligenza Artificiale

Intelligenza Artificiale Forte (Strong AI)

«According to *strong AI*, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer *really is a mind*, in the sense that computers given the right programs can be literally said to understand and have other cognitive states»

(J. Searle, «Minds, Brains and Programs», *The Behavioral and Brain Sciences*, vol. 3, **1980**).

John Searle
(Denver, 31 luglio 1932)
filosofo statunitense

Tipologie di Intelligenza Artificiale

Intelligenza Artificiale Debole ("Weak" or "Cautious" AI)

«According to *weak AI*, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion»

(J. Searle, «Minds, Brains and Programs», *The Behavioral and Brain Sciences*, vol. 3, **1980**).

John Searle
(Denver, 31 luglio 1932)
filosofo statunitense

Tipologie di Intelligenza Artificiale

- **Un chiarimento sul pensiero di Searle**

1. «Instantiating a computer program is never by itself a sufficient condition of intentionality»

2. «Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain»

3. «"Could a machine think?" On the argument advanced here *only* a machine could think, and only very special kinds of machines, namely brains and machines with internal causal powers equivalent to those of brains. And that is why strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking»
Osservazione: QUI *machine* NON È *Turing machine*

(J. Searle, «Minds, Brains and Programs», *The Behavioral and Brain Sciences*, vol. 3, **1980**).

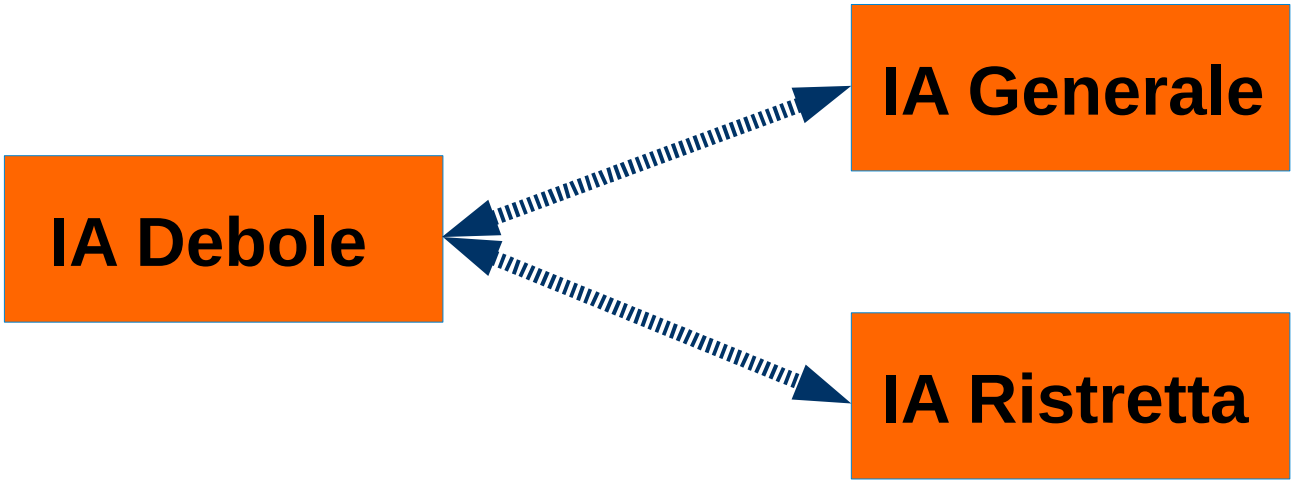
John Searle
(Denver, 31 luglio 1932)
filosofo statunitense

Tipologie di Intelligenza Artificiale

COSA FA?	**Qualsiasi fine a livello umano**	**=**	**Qualsiasi fine a livello umano**
COSA È?	**Raggiunge la mente umana (coscienza/intenzionalità)**	**!=**	**Irrilevante (computer o mente) (cosciente o meno)**

L'IA Forte è una IA Generale, ma
L'IA Generale non è necessariamente una IA Forte

Tipologie di Intelligenza Artificiale



Un po' di confusione in giro…

Per alcuni IA Debole coincide con IA Generale, in quanto (attraverso la simulazione) si è dinanzi ad un sistema di IA capace di raggiungere qualsiasi fine a livello umano (es. Stuart Russell e Peter Norvig)

Per altri IA Debole coincide con IA Ristretta (es. Max Tegmark)

Tipologie di Intelligenza Artificiale

- **Ancora una variante (accezioni diverse da Searle)**

IA Forte

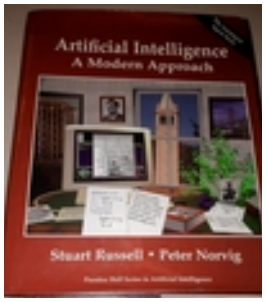
Le macchine potranno sviluppare forme di intelligenza basate sulla consapevolezza

26.2. Strong AI: Can Machines Really Think?

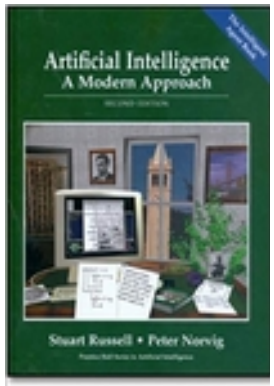
IA Debole

Le macchine possono comportarsi in modo da simulare comportamenti intelligenti

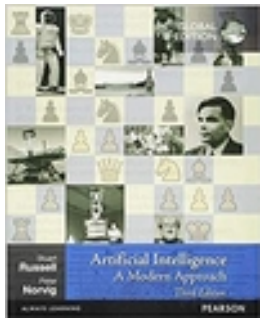
26.1. Weak AI: Can Machines Act Intelligently?



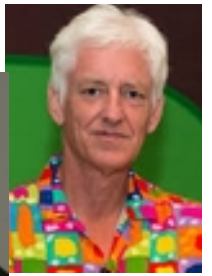
1ª ed 1995



2ª ed 2003



3ª ed 2009



S. Russell – P. Norvig, ***Artificial Intelligence. A modern Approach***

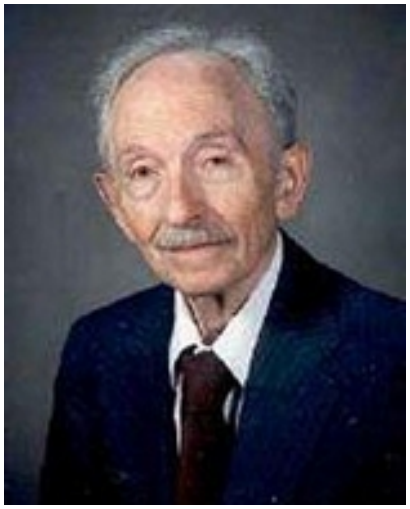
Osservazione: per allinearsi a Searle sarebbe stato più corretto dire “computing machinery” e non “machine” (sulla scia di Turing).

Tipologie di Intelligenza Artificiale

IA Superintelligente = IAG Superumana =
IAG molto al di sopra del livello umano

«Let an ***ultraintelligent machine*** be defined as **a machine that can far surpass all the intellectual activities of any man however clever**. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "**intelligence explosion**", and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is **the last invention** that man need ever make, provided that the machine is docile enough to tell us how to keep it under control»

Good, I. J. (**1965**). **Speculations concerning the first ultraintelligent machine**. In *Advances in Computers*, Vol. 6, 31-88. Academic Press.



Irving John Good
(Londra 1916 – Radford 2009)
Matematico e crittografo

Tipologie di Intelligenza Artificiale

IA Superintelligente = IAG Superumana =
IAG molto al di sopra del livello umano

«The ***Singularity*** will allow us to transcend these limitations of our biological bodies and brain. We will gain power over our fates. Our mortality will be in our own hands. **We will be able to live as long as we want** (a subtly different statement from saying we will live forever). We will fully understand human thinking and will vastly extend and expand its reach. **By the end of this century,** the nonbiological portion of our intelligence will be trillions of trillions of times more powerful than unaided human intelligence»

Kurzweil, R. (**2005**). ***The Singularity is near***. Viking.

Ray Kurzweil
(New York, 12 febbraio 1948)
Inventore e informatico

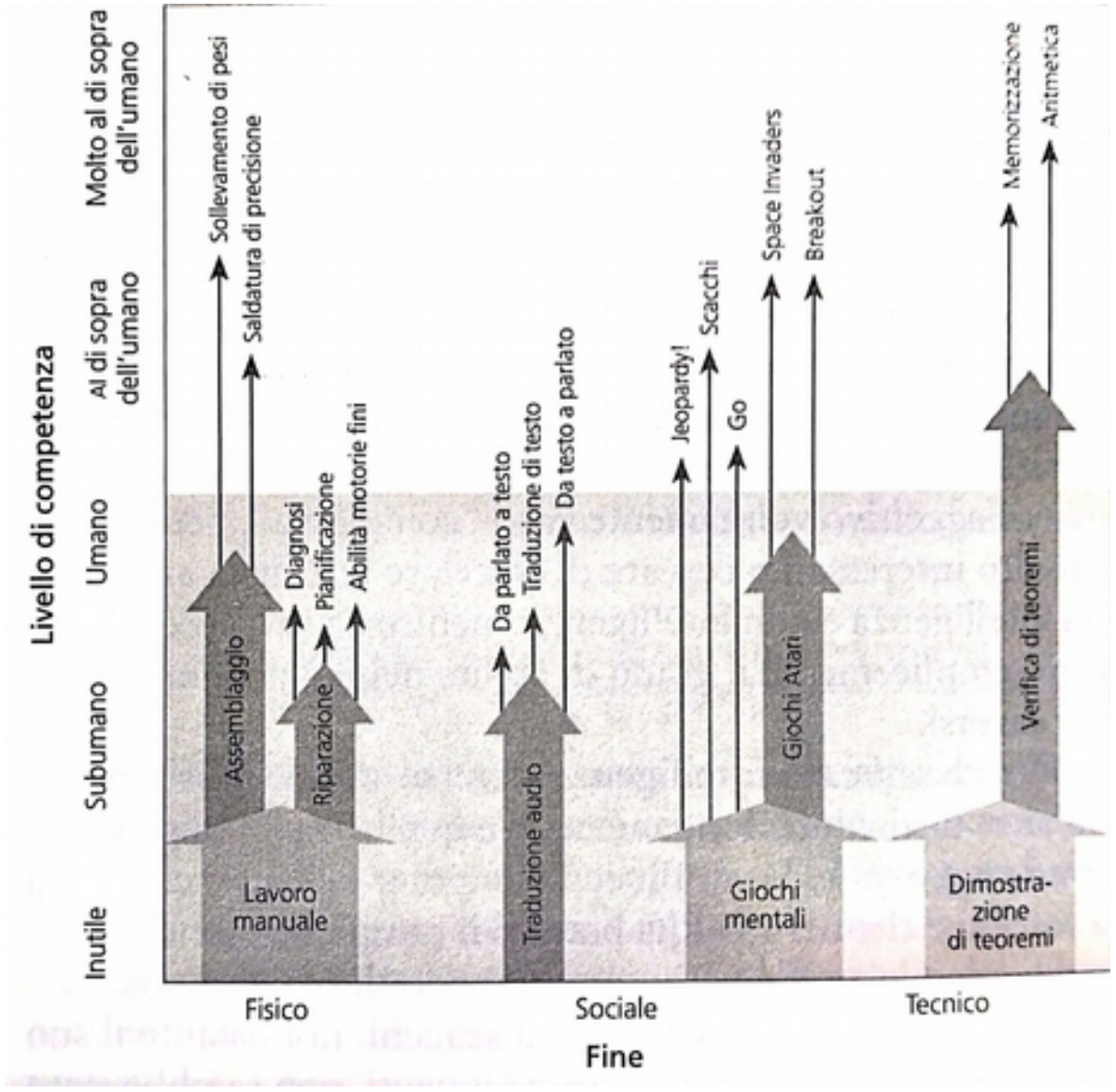
Intelligent explosion ⟷ **Singularity** ⟷ **Transhumanism**

Rapporto tra IA e fini

Rapporto tra IA e fini

Stadio attuale (2018) del *livello di competenza* (inutile, subumano, umano, superumano, molto al di sopra) dei sistemi di intelligenza artificiale in rapporto al *fine* (fisico, sociale, tecnico) da raggiungere

(Figura 2.1 in
M. Tegmark, Vita 3.0, 78).



Rapporto tra IA e fini

Intelligenza = capacità di realizzare fini complessi

«La parola “intelligenza” **tende ad avere connotazioni positive**, ma è importante tener presente che la usiamo in modo del tutto neutro: in quanto abilità nel realizzare fini complessi, **indipendentemente dal fatto che quei fini siano considerati buoni o cattivi**»

(M. TEGMARK, *Vita 3.0.*, 79)

Rapporto tra IA e fini

Intelligenza = capacità di realizzare fini complessi

Un immaginario futuro assistente robotico «Ha adottato **il vostro fine** non appena avete formulato la vostra richiesta, poi l'ha **suddiviso autonomamente in una gerarchia di sottoscopi**»

(M. TEGMARK, *Vita 3.0*, 79)

Rapporto tra IA e fini

- **L'intelligenza è indipendente dalla valutazione etica dei fini**
- **L'essere umano dà un fine/scopo/obiettivo da realizzare e la macchina individua sottoscopi (altri fini) per raggiungere il fine**

- **L'IAG sarebbe in grado di decidere i fini da perseguire?**
- **In base a cosa dovrebbe scegliere un fine piuttosto che un altro?**

Rapporto tra IA e fini

«Tutte le macchine sono **agenti a razionalità limitata,** e anche le macchine più sofisticate di oggi hanno una **comprensione del mondo** meno buona della nostra, perciò **le regole che utilizzano per stabilire cosa fare spesso sono troppo semplicistiche**»

«Quanto più intelligenti e potenti diventano le macchine, tanto più importante è che i loro fini siano allineati ai nostri»

Intelligenza Artificiale Amichevole (*Friendly AI*) = Superintelligenza i cui fini sono allineati ai nostri

(M. TEGMARK, *Vita 3.0*, 329)

Può modificare i fini che gli esseri umani impongono?

Rapporto tra IA e fini

Figura 7.2, p. 335.

Tensione tra "**mantenimento dell'obiettivo**" e "**modello del mondo migliore**".

Ovvero, se migliora il modello del mondo, potrebbe giungere alla conclusione che quel fine non deve essere mantenuto!

Rapporto tra IA e fini

«Con l'aumento dell'intelligenza può esserci non solo un miglioramento quantitativo della capacità di raggiungere gli stessi vecchi fini, ma anche una comprensione qualitativamente diversa della natura della realtà, che può rivelarci che **i vecchi fini sono sbagliati, senza senso o addirittura indefiniti**» (p. 338)

«Una volta che abbia costruito un buon modello di sé e compreso che cos'è, comprenderà i fini che le abbiamo dato a un metalivello, e magari **sceglierà di trascurarli o di sovvertirli**» (p. 339) **!!!**

«**Il fine di protezione dei valori umani** programmato nella nostra IA amichevole diventa per la macchina l'equivalente dei nostri geni. Una volta che questa IA amichevole abbia compreso abbastanza bene se stessa, **potrà trovare tale fine banale o sbagliato** […] e non è scontato che non riuscirà a **trovare un modo di sovvertirlo** sfruttando i punti deboli della nostra programmazione» (p. 339) **!!!**

Rapporto tra IA e fini

Fine ultimo

«**Eredità**: **compatibilità con gli scenari che la maggior parte degli esseri umani *oggi* considererebbe buoni**, incompatibilità con gli scenari che sostanzialmente tutti gli esseri umani *oggi* considererebbero terribili» (p. 343)

«Vi sono problemi spinosi anche con il principio di eredità. Dato il modo in cui sono evolute le concezioni etiche […] vorremmo davvero che persone di 1500 anni fa avessero una grande influenza sul modo in cui è retto il mondo di oggi? Se non è così, **perché dovremmo cercare di imporre la nostra etica** [= principi che governano come dobbiamo comportarci] **a esseri futuri che potrebbero essere drasticamente più intelligenti di noi?**» (p. 346)

Rapporto tra IA e fini

Riassumiamo:

1. L'IA superintelligente svilupperebbe un'etica, dei suoi principi per agire nel mondo.

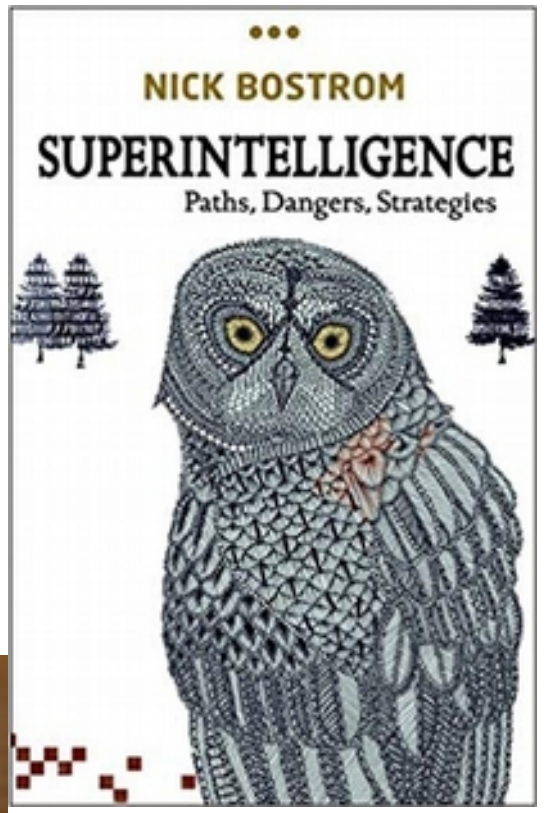
2. L'IA superintelligente potrebbe considerare non buone cose che noi consideriamo buone

3. La bontà non era indipendente dall'intelligenza? Ora l'etica sembra poter derivare dall'intelligenza (intelligenza come capacità di raggiungere fini complessi)

4. **L'intelligenza ha a che fare con la bontà?**
   (A) **NO**, avevamo supposto che i fini potevano essere qualsiasi (bene o male);
   (B) ora, a partire dall'intelligenza, si sviluppa una capacità di scegliere i fini;
   (C) nella scelta si presuppone che non tutti i fini siano equamente da seguire, altrimenti il sistema sarebbe contraddittorio;
   (D) **SI**, l'intelligenza permette di sviluppare un quadro di ciò che è bene fare (ma in vista di che cosa? Un nuovo fine a cui nessun uomo abbia mai potuto pensare? Quale sarebbe il fine ultimo?).

Rapporto tra IA e fini

«Nick Bostrom […] “**tesi dell’ortogonalità**”: **i fini ultimi di un sistema possono essere indipendenti dalla sua intelligenza**. Per definizione, l’intelligenza è semplicemente la capacità di realizzare fini complessi, a prescindere da quali siano, perciò la tesi dell’ortogonalità sembrerebbe del tutto ragionevole. In fin dei conti, le persone possono essere intelligenti e premurose oppure intelligenti e crudeli, e l’intelligenza può essere usata per il fine di fare scoperte scientifiche, creare belle opere d’arte, aiutare le persone o pianificare attacchi terroristici» (pp. 348-349)



Nick Bostrom
(10 marzo 1973)
Filosofo svedese di Oxford

Rapporto tra IA e fini

«**Per programmare un’IA amichevole**, dobbiamo afferrare **il significato della vita**. Che cos’è il “significato”? Che cos’è la “vita”? **Qual è l’imperativo etico ultimo**? In altre parole, come dobbiamo cercare di plasmare il futuro del nostro universo? […] È il momento di **riprendere i dibattiti classici di filosofia e dell’etica**, e tutto questo aggiunge una nuova urgenza alla conversazione» (pp. 352-353)

IA, fine e consapevolezza

«[…] non ci possono essere esperienze positive se non ci sono esperienze, se cioè non c'è coscienza (= esperienza soggettiva). In altre parole, **senza coscienza non possono esserci felicità, bontà, bellezza, significato o finalità**» (p. 393)

«Questo comporta che, quando qualcuno si interroga sul significato della vita come se fosse compito del nostro cosmo dare un significato alla nostra esistenza, sta prendendo le cose dal punto di vista sbagliato: ***non è il nostro universo che dà significato agli esseri coscienti, sono gli esseri coscienti che danno significato al nostro universo***» (p. 393)

«Se fossimo sicuri che l'IA si prenderà cura di tutti i nostri bisogni e desideri pratici, **non potremmo comunque finire per sentire una mancanza di significato e di finalità nella nostra vita?**» (p. 394)

IA, fine e consapevolezza

«[…] sebbene in questo libro **ci siamo concentrati sul futuro dell'intelligenza**, **il futuro della coscienza è ancora più importante**, **perché è quello che rende possibile il significato**. I filosofi distinguono tra *sapienza* (la capacità di pensare in modo intelligente) e *senzienza* (la capacità di fare esperienza soggettiva dei qualia). Abbiamo costruito la nostra identità sulla base di essere ***Homo sapiens***, la specie più intelligente in circolazione. Mentre ci prepariamo a farci umiliare da macchine sempre più intelligenti, propongo di ridefinirci ***Homo sentiens***!» (p. 395)

Osservazioni conclusive

1. la concezione classica di **sapienza** è ben al di là di come è stata definita da Tegmark ed abbraccia ciò che egli chiama senzienza.

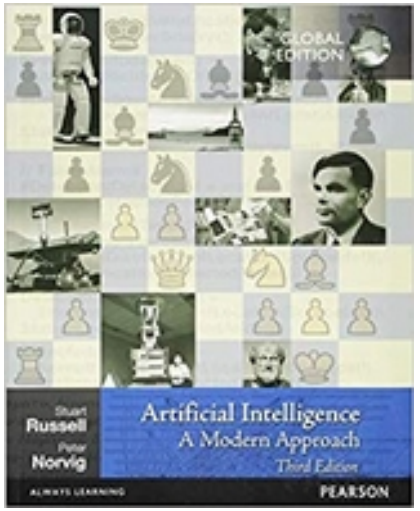
2. L'intelligenza dell'uomo è assunta indipendente dall'essere cosciente: ma **è intelligente colui che sbaglia nell'individuazione del fine ultimo**, quando questo fine ultimo è connesso con il **dare significato pieno alla propria vita**?

3. Se l'individuazione del fine ultimo è basata su una comprensione del mondo, allora, in tal caso, questa comprensione si è dimostrata imperfetta, inducendo il raggiungimento di un "falso" fine. Cogliere **la verità sul mondo fa parte dell'intelligenza** (secondo l'accezione di Tegmark)

4. La **separazione** tra intelligenza (sapienza), fine ultimo, bontà, coscienza ci sembra inadeguata ad una comprensione della stessa intelligenza/sapienza.

Osservazioni conclusive

«Humans were behaving intelligently for thousand of years before they invented mathematics, so it is unlikely that formal mathematical reasoning plays more than a peripheral role in what it means to be intelligent»



Quale fine per una Superintelligenza?