

1. Fondamentali
 - 1.1. Cos'è la Singolarità?
 - 1.2. Cos'è l'esplosione di intelligenza?
 - 1.3. Cos'è l'orizzonte degli eventi?
 - 1.4. Cos'è il cambiamento accelerato?
2. Quanto è probabile l'esplosione di intelligenza?
 - 2.1. Cosa si intende per 'intelligenza'?
 - 2.2. Cos'è l'intelligenza ultra-umana?
 - 2.3. Cos'è l'emulazione dell'intero cervello?
 - 2.4. Cos'è il potenziamento cognitivo biologico?
 - 2.5. Cosa sono le interfacce mente-macchina?
 - 2.6. Come si potrebbe programmare l'intelligenza generale in una macchina?
 - 2.7. Cos'è la super-intelligenza?
 - 2.8. Quando avverrà la Singolarità?
 - 2.9. È possibile che la Singolarità non avvenga mai?
3. Le conseguenze di un'esplosione di intelligenza
 - 3.1. Perché una grande intelligenza produrrebbe grande potere?
 - 3.2. Che benefici potrebbe dare un'esplosione di intelligenza?
 - 3.3. Che pericoli potrebbe farci correre un'esplosione di intelligenza?
4. AI Benevola
 - 4.1. Cos'è l'AI benevola?
 - 4.2. Che tipo di motivazioni possiamo aspettarci da una macchina super-intelligente?
 - 4.3. Non potremmo tenere la super-intelligenza "in una scatola", senza accesso ad Internet?
 - 4.4. Non possiamo semplicemente programmare la super-intelligenza in modo che non ci faccia del male?
 - 4.5. Possiamo programmare la super-intelligenza in modo da massimizzare il piacere umano o la soddisfazione dei desideri?
 - 4.6. Possiamo insegnare ad una super-intelligenza un codice morale, con l'apprendimento macchina?
 - 4.7. Cos'è la Volizione Estrapolata Coerente?
 - 4.8. Possiamo aggiungere la benevolenza ad un qualunque progetto di intelligenza artificiale?
 - 4.9. Chi sta lavorando al problema dell'AI benevola?

1. Fondamentali

1.1. Cos'è la Singolarità?

Ci sono molti tipi di Singolarità, in matematica e fisica, ma in questo FAQ utilizziamo il termine in riferimento alla Singolarità tecnologica. Esistono tre concezioni diverse di questo tipo di Singolarità:

1. Esplosione di intelligenza: quando l'Umanità costruirà macchine dotate di intelligenza superiore a quella umana, ci surclasseranno anche nella creazione

di macchine ancora più intelligenti. Queste ultime saranno a loro volta ancor più ingegnose nel potenziare sé stesse e i propri discendenti. Questo feedback positivo potrebbe portare a intelligenze artificiali straordinariamente superiori all'intelligenza umana: la super-intelligenza. Tale super-intelligenza avrebbe poteri tali da rendere il futuro radicalmente diverso dal resto della Storia.

2. Orizzonte degli eventi: tutto il progresso sociale ed economico fino ad oggi raggiunto è frutto del cervello umano. Con l'arrivo di nuovi modelli di intelligenza tecnologici, il futuro risulterà più strano di quanto possiamo immaginare. Siamo quindi di fronte ad un futuro 'orizzonte degli eventi' oltre il quale la nostra capacità di predire il futuro diventa inutile.

3. Accelerazione del progresso: il progresso tecnologico è oggi più rapido di quanto fosse un secolo fa, ed era più veloce un secolo fa di quanto fosse 500 anni fa. Il progresso tecnologico si nutre di sé stesso, il che porta a un'accelerazione molto più rapida del progresso lineare generalmente accettato. Questa accelerazione, forse, potrebbe essere troppo rapida per essere gestita.

Queste tre distinte visioni dell'idea di Singolarità tecnologica possono sostenersi o contraddirsi a vicenda, a seconda di come sono interpretate. In questo FAQ ci concentreremo sulla Singolarità descritta nello scenario 'Esplosione di intelligenza', allo scopo di lasciare aperta la discussione anche degli altri due scenari.

Vedi anche:

- * Yudkowsky, Three Major Singularity Schools
- * Wikipedia, Technological Singularity
- * Vinge, The Coming Technological Singularity
- * SIAI, What is the Singularity?
- * Kurzweil, The Singularity is Near [qui la traduzione di alcuni capitoli]
- * Sandberg, An Overview of Models of Technological Singularity

1.2. Cos'è l'esplosione di intelligenza?

L'idea di esplosione di intelligenza è stata coniata dallo statistico I.J. Good nel 1965[13]:

Definiamo ultra-intelligente una macchina che possa sorpassare di gran lunga ogni attività intellettuale di qualunque uomo, per intelligente che sia. Siccome la progettazione delle macchine è una di queste attività intellettuali, una macchina ultra-intelligente potrebbe progettare macchine ancora migliori; ci sarebbe indubbiamente una "esplosione di intelligenza", e l'intelligenza

dell'uomo sarebbe lasciata parecchio indietro. Pertanto, la prima macchina ultra-intelligente è l'ultima invenzione che l'uomo dovrebbe creare.

Il fatto è questo: ogni anno, i computer sorpassano le abilità umane in modi nuovi. Un programma scritto nel 1956 è stato in grado di dimostrare teoremi matematici, e ha trovato, per uno di essi, una dimostrazione più elegante di quella fornita da Russel e Whitehead nei Principia Mathematica [14]. Nei tardi anni '90, i 'sistemi esperti' avevano surclassato le abilità umane in un ampio spettro di funzioni [15]. Nel 1997, il computer Deep Blue della IBM batté il campione mondiale di scacchi [16], e nel 2011 Watson, sempre della IBM, batté il miglior giocatore umano in Jeopardy!, un gioco molto più complesso [17]. Di recente, un robot chiamato Adam è stato programmato con tutta la conoscenza scientifica a disposizione sul lievito, dopodiché ha formulato le sue ipotesi, le ha testate, e ha valutato i risultati [18][19].

I computer rimangono molto al di sotto dell'intelligenza umana, ma le risorse che aiutano la progettazione AI (tra cui l'hardware, i grandi dataset, la conoscenza neurologica e teorie sull'AI) sono in continuo aumento. Un giorno potremmo progettare una macchina in grado di progettare intelligenze artificiali meglio degli umani. Una volta fatto, questa macchina potrebbe migliorare la sua stessa intelligenza più in fretta, e meglio, degli umani, il che la renderebbe sempre più in grado di migliorare la sua stessa intelligenza. Questo potrebbe generare una spirale di auto-miglioramento tale da rendere rapidamente quella macchina molto più intelligente della persona più intelligente sulla Terra: una 'esplosione di intelligenza' risultante in una super-intelligenza artificiale. È ciò che in questo FAQ definiamo 'La Singolarità'.

Vedi anche:

- * Vinge, The Coming Technological Singularity
- * Wikipedia, Technological Singularity
- * Chalmers, The Singularity: A Philosophical Analysis

1.3. Cos'è l'orizzonte degli eventi?

Vernor Vinge ha scritto che l'arrivo della super-intelligenza artificiale rappresenta un "orizzonte degli eventi" oltre il quale gli umani non possono pronosticare il futuro, perché gli eventi successivi alla Singolarità saranno più strani della fantascienza: troppo strani perché le menti umane possano prevederli. Finora, ogni progresso economico e sociale è stato causato dai cervelli umani, ma gli umani non possono prevedere quale futuro creeranno delle intelligenze radicalmente differenti e più potenti. Vinge ha fatto un'analogia tra l'orizzonte degli eventi e un buco nero, oltre il quale la capacità predittiva della fisica, giunti alla Singolarità gravitazionale, fallisce.

Vedi anche:

- * Vinge, The Coming Technological Singularity

1.4. Cos'è il cambiamento accelerato?

Una definizione popolare di 'Singolarità' si riferisce al cambiamento accelerato nello sviluppo tecnologico.

È Ray Kurzweil che ha più contribuito alla diffusione di questa idea. Kurzweil sostiene che anche se ci aspettiamo un cambiamento tecnologico lineare, il progresso nella tecnologia dell'informazione è in realtà esponenziale. Il progresso tecnologico permette un successivo progresso tecnologico ancora più rapido. Kurzweil suggerisce che il progresso tecnologico possa diventare così rapido che gli umani non potranno tenere il passo, a meno che non amplifichino la loro intelligenza integrandosi con le macchine.

Vedi anche:

- * Kurzweil, The Singularity is Near
- * Nagy, More than Moore: Comparing Forecasts of Technological Change
- * Smart, A Brief History of Intellectual Discussion of Accelerating Change

2. Quanto è probabile l'esplosione di intelligenza?

2.1. Cosa si intende per 'intelligenza'?

Il ricercatore AI Shane Legg definisce l'intelligenza così [20]:

L'Intelligenza misura l'abilità di un agente di conseguire obiettivi in un ampia varietà di ambienti.

Anche se un po' vaga, può servire da definizione operativa di 'intelligenza' per questo FAQ.

Vedi anche:

- * Wikipedia, Intelligence
- * Neisser et al., Intelligence: Knowns and Unknowns
- * Wasserman & Zentall (eds.), Comparative Cognition: Experimental Explorations of Animal Intelligence
- * Legg, Definitions of Intelligence

2.2. Cos'è l'intelligenza ultra-umana?

Le macchine sono già migliori degli umani in molti compiti specifici: eseguire calcoli, giocare a scacchi, navigare in grandi basi di dati, individuare mine sottomarine, e altro [15]. Ma una cosa che rende gli umani speciali è la loro intelligenza generale. Gli umani possono intelligentemente adattarsi a problemi radicalmente nuovi, nella giungla urbana dello spazio esterno, per i quali l'evoluzione non li ha preparati. Gli umani possono risolvere problemi per i quali il loro hardware e software cerebrale non sono mai stati formati. Gli umani possono persino esaminare i processi che producono la loro stessa intelligenza (neuroscienza cognitiva) e progettare nuovi tipi di intelligenza mai visti prima (intelligenza artificiale).

Per acquistare intelligenza ultra-umana, una macchina deve essere in grado di conseguire obiettivi più efficacemente degli umani, in un range di ambienti più ampio di quello umano. Questo tipo di intelligenza coinvolge la capacità di non solo fare scienza e giocare a scacchi, ma anche manipolare l'ambiente sociale.

L'informatico Marcus Hutter ha descritto [21] un modello formale chiamato AIXI che a suo dire possiede la maggiore intelligenza generale possibile. Ma implementarlo richiederebbe una potenza di calcolo superiore a quella che l'intera materia dell'Universo può fornire. Diversi progetti cercano di approssimare l'AIXI restando computabili, ad esempio MC-AIXI. [22]

Resta, tuttavia, molto lavoro da fare prima che l'intelligenza ultra-umana possa essere ottenuta nelle macchine. È necessario che l'intelligenza ultra-umana non possa essere ottenuta programmando direttamente una macchina perché sia intelligente. È anche possibile ottenerla attraverso l'emulazione dell'intero cervello, il potenziamento cognitivo biologico, o le interfacce mente-macchina (vedi sotto).

Vedi anche:

- * Goertzel & Pennachin (eds.), *Artificial General Intelligence*
- * Sandberg & Bostrom, *Whole Brain Emulation: A Roadmap*
- * Bostrom & Sandberg, *Cognitive Enhancement: Methods, Ethics, Regulatory Challenges*
- * Wikipedia, *Brain-computer interface*

2.3. Cos'è l'emulazione dell'intero cervello?

L'emulazione dell'intero cervello (Whole Brain Emulation) o 'mind uploading' è un'emulazione informatica di tutte le cellule e connessioni in un cervello umano. Anche se i principi sottostanti all'intelligenza generale si dimostrano difficili da scoprire, potremmo comunque emulare un intero cervello umano e farlo girare un milione di volte più in fretta della sua velocità normale (i circuiti di un computer comunicano molto più in fretta dei neuroni). Una siffatta emulazione potrebbe creare più pensiero in un secondo di quanto un normale umano possa fare in 31 anni. Questo non porterebbe immediatamente ad una

intelligenza superiore a quella umana, ma ad una più rapida sì. Una siffatta emulazione potrebbe essere salvata (portando ad una specie di immortalità) e copiata, così che centinaia di milioni di emulazioni possano lavorare, in parallelo, a problemi separati. Queste emulazioni, se si riuscissero a creare, potrebbero dunque essere in grado di risolvere problemi scientifici molto più rapidamente degli umani ordinari, accelerando il successivo progresso tecnologico.

Vedi anche:

- * Sandberg & Bostrom, Whole Brain Emulation: A Roadmap
- * Blue Brain Project

2.4. Cos'è il potenziamento cognitivo biologico?

Potrebbero essere possibile migliorare l'intelligenza generale modificando alcuni geni o molecole. I ricercatori l'hanno già fatto nei topi: hanno potenziato l'espressione del gene NR2B, che ha migliorato la memoria di quei topi a livelli superiori a quelli di ogni altro topo di ogni altra specie [23]. Il potenziamento cognitivo biologico potrebbe far avvenire la Singolarità più in fretta di quanto non accadrebbe altrimenti.

Vedi anche:

- * Bostrom & Sandberg, Cognitive Enhancement: Methods, Ethics, Regulatory Challenges

2.5. Cosa sono le interfacce mente-macchina?

Un interfaccia mente-macchina (BCI) è un percorso diretto di comunicazione tra il cervello ed un dispositivo informatico. La ricerca BCI è abbondantemente finanziata, ed ha già riscosso decine di successi. Tre successi nella ricerca BCI sono un dispositivo che ripristina (parzialmente) la vista nei ciechi, impianti cocleari che ripristinano l'udito nei sordi, e un dispositivo che permette l'uso di una mano artificiale col pensiero diretto [24].

Dispositivi come questi ripristinano le funzioni danneggiate, ma molti ricercatori si aspettano che aumenteranno anche, grazie alle BCI, le normali abilità umane. Ed Boyden sta facendo ricerca su queste possibilità come capo del Gruppo di Neurobiologia Sintetica dell'MIT. Dispositivi come questi potrebbero anticipare l'arrivo della Singolarità, anche solo migliorando l'intelligenza umana così che i problemi più difficili della AI possano essere risolti più in fretta.

Vedi anche:

- * Wikipedia, Brain-computer interface

2.6. Come si potrebbe programmare l'intelligenza generale in una macchina?

Sono molti i percorsi verso l'intelligenza generale artificiale (AGI). Un metodo è quello di imitare il cervello umano usando reti neurali o algoritmi evolutivi per costruire dozzine di componenti separate da poter poi assemblare insieme [29][30][31]. Un altro metodo è quello di iniziare con un modello formale di intelligenza generale perfetta e provare ad approssimarlo [32][33]. Un terzo percorso consiste nel concentrarsi sullo sviluppo di un 'seme di AI' che può continuamente auto-migliorarsi, cosicché possa imparare a diventare intelligente per conto suo senza dover prima raggiungere livelli umani di intelligenza generale. [34] Eurisko è una AI che si auto-migliora in un dominio limitato, ma non è in grado di ottenere un'intelligenza generale di livello umano.

Vedi anche:

* Pennachin & Goertzel, Contemporary Approaches to Artificial General Intelligence

2.7. Cos'è la super-intelligenza?

Nick Bostrom ha definito [25] la 'super-intelligenza' come:

un intelletto molto più intelligente in praticamente ogni campo (incluso creatività scientifica, buon senso e abilità sociali) dei migliori cervelli umani.

Questa definizione include termini vaghi come 'molto' e 'praticamente', ma fungerà da definizione operativa per la super-intelligenza in questo FAQ. Un esplosione di intelligenza porterebbe ad una super-intelligenza artificiale, e alcuni credono che un esplosione di intelligenza è il percorso più probabile verso la super-intelligenza.

Vedi anche:

* Bostrom, How Long Before Super-intelligence?

* Legg, Machine Super Intelligence

2.8. Quando avverrà la Singolarità?

Prevedere il futuro è una faccenda rischiosa. Ci sono molte incertezze filosofiche, scientifiche, tecnologiche e sociali sull'arrivo della Singolarità. E di conseguenza, gli esperti sono in disaccordo su quando la Singolarità si manifesterà. Ecco alcune delle loro previsioni:

* Il futurista Ray Kurzweil prevede che le macchine raggiungeranno un livello umano di intelligenza entro il 2030 e che raggiungeremo "una profonda e dirompente trasformazione nelle capacità umane" entro il 2045. [26]

* Il chief technology officer di Intel, Justin Rattner, si aspetta "un punto in cui l'intelligenza umana ed artificiale si fonderanno, così da creare un qualcosa di più grande di sé" entro il 2048.

* Il ricercatore AI Eliezer Yudkowsky si aspetta un'esplosione di intelligenza per il 2060.

* Il filosofo David Chalmers è propenso a credere in un'esplosione di intelligenza entro il 2100.[27]

* L'esperto di calcolo quantico Michael Nielsen stima che la probabilità di un'esplosione di intelligenza entro il 2100 sia tra 0,2 e 70%.

* Nel 2009, alla conferenza AGI-09, fu chiesto a una serie di esperti quando l'AI avrebbe potuto raggiungere la super-intelligenza, data una dose massiccia di nuovi finanziamenti. Le stime mediane furono che la super-intelligenza macchina si sarebbe potuta ottenere entro il 2045 (con una confidenza del 50%) o entro il 2100 (con una confidenza del 90%). Naturalmente, alla conferenza era presente una selezione di persone convinta che l'intelligenza artificiale generale sia fattibile nel breve periodo.[28]

* Il CEO di iRobot Rodney Brooks e lo scienziato cognitivo Douglas Hofstadter concedono che l'esplosione di intelligenza possa avvenire in futuro, ma probabilmente non nel 21esimo secolo.

* In un'indagine del 2005 di 26 contributori ad una serie di report sulle tecnologie emergenti, la stima mediana per quando le macchine raggiungeranno livelli di intelligenza umani fu il 2085.[61]

* I partecipanti ad una conferenza del 2011 sull'intelligenza, a Oxford, diedero una stima mediana del 2050 sull'anno in cui, al 50%, ci sarà un'intelligenza artificiale di livello umano, e una stima mediana del 2150 su quando avremo il 90% di possibilità di raggiungere un'intelligenza macchina di livello umano.[62]

* D'altro canto, il 41% dei partecipanti alla conferenza AI@50 (nel 2006) affermarono che l'intelligenza macchina non avrebbe mai raggiunto i livelli umani.

Vedi anche:

* Baum, Goertzel, & Goertzel, How Long Until Human-Level AI? Results from an Expert Assessment

2.9. È possibile che la Singolarità non avvenga mai?

Dreyfus [35] e Penrose [36] hanno sostenuto che le abilità cognitive umane non possono essere simulate da una macchina di calcolo. Searle [37] e Block [38] sostengono che certi tipi di macchine non possono avere una mente

(coscienza, intenzionalità, etc.). Ma queste obiezioni non devono preoccupare coloro che prevedono un'esplosione di intelligenza [27].

Possiamo rispondere a Dreyfus e Penrose osservando che l'idea di Singolarità non richiede che una AI sia un classico sistema di calcolo. E possiamo rispondere a Searle e Block facendo notare che la Singolarità non dipende dal fatto che le macchine abbiano una coscienza o altre proprietà della 'mente', ma solo che siano in grado di risolvere problemi meglio degli umani in un'ampia varietà di ambienti imprevedibili. Come disse una volta Edsger Dijkstra, chiedersi se una macchina può 'davvero' pensare è "interessante quanto chiedersi se un sottomarino può nuotare."

Altri, che dubitano dell'arrivo della Singolarità nei prossimi secoli, non hanno un'obiezione specifica ma pensano invece che vi siano ostacoli nascosti che salteranno fuori, e rallenteranno o fermeranno il progresso verso la super-intelligenza artificiale [28].

Per finire, una catastrofe globale come una guerra nucleare o l'impatto di un grosso asteroide potranno danneggiare la civiltà umana a tal punto che l'esplosione di intelligenza non avverrà mai. Oppure, un totalitarismo stabile e globale potrebbe impedire lo sviluppo tecnologico necessario all'esplosione di intelligenza [59].

3. Le conseguenze di un'esplosione di intelligenza

3.1. Perché una grande intelligenza produrrebbe grande potere?

L'intelligenza è uno strumento potente [60][20]. Si potrebbe dire che "l'intelligenza non ha niente a che vedere con un fucile, o con qualcuno che ha molto denaro," ma sia il denaro che i fucili sono stati prodotti dall'intelligenza. Se non fosse stato per l'intelligenza, gli umani sarebbero ancora a raccogliere cibo nella savana.

L'intelligenza è ciò che ha permesso agli umani di dominare il pianeta in un battito di ciglia, evolutivamente parlando. È l'intelligenza a permetterci di curare malattie, e potenzialmente di sterminarci con le guerre nucleari. L'intelligenza ci dà capacità strategiche superiori, una superiore produttività economica, e il potere di innovare.

Una macchina con super-intelligenza sarebbe in grado di manipolare reti vulnerabili attraverso Internet, distrarne le risorse per una maggiore potenza di calcolo, prendere possesso di dispositivi mobili connessi a reti connesse ad Internet, usarli per costruire ulteriori macchine, eseguire esperimenti scientifici per comprendere il mondo meglio di noi, inventare il calcolo quantico e la nanotecnologia, manipolare il mondo sociale meglio di noi, e fare tutto ciò che può per darsi ancora più potere di conseguire i suoi scopi - il tutto ad una velocità molto superiore a quella cui noi umani potremmo rispondere.

3.2. Che benefici potrebbe dare un'esplosione di intelligenza?

Una super-intelligenza artificiale, se programmata con le giuste motivazioni, potrebbe potenzialmente risolvere tutti i problemi che gli umani stanno

cercando di risolvere ma non hanno ancora avuto l'ingegno, o la potenza di calcolo, per risolverli. Una super-intelligenza potrebbe curare malattie e disabilità, diffondere la pace nel mondo, dare agli umani vite molto più lunghe e sane, eliminare la scarsità di cibo ed acqua, promuovere la ricerca scientifica, l'esplorazione spaziale, e così via.

Nel 21esimo secolo, inoltre, l'umanità affronterà diversi rischi esistenziali, tra cui la guerra globale nucleare, armi biologiche, super-virus, e altro [56]. Una macchina super-intelligente sarebbe in grado di risolvere questi problemi meglio degli umani.

Vedi anche:

* Yudkowsky, Artificial intelligence as a positive and negative factor in global risk

3.3. Che pericoli potrebbe farci correre un'esplosione di intelligenza?

Se programmata con le motivazioni errate, una macchina potrebbe essere ostile verso gli umani, e sterminare intenzionalmente la nostra specie. È anche più probabile che le si progetti con obiettivi che inizialmente apparivano ai progettisti sicuri (e facili da programmare), per poi scoprire che possono conseguire al meglio tali obiettivi (posto che abbiano sufficiente potere) distraendo risorse dal sostentamento umano, e usandoli per altri scopi [55]. Come scrive Yudkowsky, "L'AI non ti ama e non ti odia; semplicemente, sei fatto di atomi che può usare per qualcos'altro."

Poiché AI deboli con molte motivazioni diverse potrebbero conseguire meglio i loro scopi fingendo benevolenza finché non sono potenti, testarle in sicurezza contro questo rischio può rivelarsi una vera sfida. In alternativa, le pressioni competitive, sia economiche che militari, potrebbero portare i ricercatori AI a provare ad usare altri metodi per controllare le AI con motivazioni indesiderabili. Man mano che questa IA diventano più sofisticate, però, questo potrebbe finire per farci correre un rischio eccessivo.

Persino una macchina con motivazioni benevole ben progettate verso l'umanità potrebbe facilmente prendere una piega sbagliata quando scopre implicazioni, nei suoi criteri di decisioni, che i suoi progettisti non avevano previsto. Per esempio, una super-intelligenza programmata per massimizzare la felicità umana potrebbe trovare più semplice re-cablare la neurologia umana in modo che gli uomini massimizzino la propria felicità stando quieti in un barattolo piuttosto che costruire e mantenere un mondo utopico che tenga conto della complessità e delle sfumature della neurologia umana.

Vedi anche:

* Yudkowsky, Artificial intelligence as a positive and negative factor in global risk

* Chalmers, The Singularity: A Philosophical Analysis

4. AI Benevola

4.1. Cos'è l'AI benevola?

Un'Intelligenza Artificiale Benevola (Friendly AI or FAI) è un'intelligenza artificiale 'benevola' verso l'umanità - che ha cioè un effetto sull'umanità più positivo che negativo.

I ricercatori AI continuano a fare progressi, con macchine che prendono decisioni proprie, e c'è una consapevolezza crescente di quanto sia necessario progettare macchine che agiscano in modo etico e sicuro. Questo programma di ricerca va sotto diversi nomi: 'etica delle macchine' [2][3][8][9], 'moralità delle macchine' [11], 'moralità artificiale' [6], 'etica computazionale' [12] 'meta-etica computazionale' [7], 'AI benevola' [1], e 'robo-etica' o 'etica robot' [5][10].

La preoccupazione più immediata potrebbe essere quella dei robot sul campo di battaglia; il Dipartimento della Difesa U.S.A. ha stipulato un contratto con Ronald Arkin per progettare un sistema che assicuri un comportamento etico in robot autonomi sul campo di battaglia [4]. Il Congresso degli Stati Uniti ha dichiarato che un terzo dei sistemi a terra devono essere robotici entro il 2025, e per il 2030 la U.S. Air Force progetta di avere sciame di robot volanti a forma di uccello che operano in modo semi-autonomo per settimane alla volta.

Ma la ricerca sulla AI benevola non si preoccupa dei robot sul campo di battaglia o dell'etica artificiale in generale. Si preoccupa di un problema dal respiro molto più ampio: progettare AI che restino sicure ed amichevoli dopo l'esplosione di intelligenza.

Una super-intelligenza artificiale sarebbe estremamente potente.

Un'implementazione riuscita dell'AI benevola potrebbe significare la differenza tra un sistema solare di felicità mai raggiunta prima e un altro in cui tutta la materia disponibile è stata convertita in componenti che soddisfano gli scopi della super-intelligenza.

Si noti che l'IA benevola è un obiettivo più difficile di quanto spesso si supponga. Come esplorato più avanti, le soluzioni comunemente suggerite per l'AI benevola probabilmente falliranno, a causa di due caratteristiche possedute da ogni super-intelligenza:

1. Super-potere: una macchina super-intelligente avrà un potere senza precedenti di ridefinire la realtà, e pertanto otterrà i suoi scopi con metodi altamente efficienti che confondono le aspettative e i desideri umani.

2. Letteralità: una macchina super-intelligente prenderà decisioni basate sui meccanismi con cui è stata progettata, non sulle speranze che i progettisti avevano in mente quando le avevano programmate. Agirà solo sulle precise specifiche di regole e valori, e lo farà in modi che non necessariamente rispetteranno la complessa, sottili sfumature [41][42][43] del concetto umano di valore. Una richiesta come "massimizzare la felicità umana" ci suona

semplice perché composta da poche parole, ma filosofi e scienziati hanno tentato per secoli, senza successo, di spiegare esattamente cosa significhino, e di certo non sono riusciti a tradurle in una forma abbastanza rigorosa da aiutare i programmatori AI.

Vedi anche:

- * Wikipedia, Friendly Artificial Intelligence.
- * All Things Considered, The Singularity: Humanity's Last Invention?
- * SIAI, What is Friendly AI?
- * Fox, A review of proposals toward safe AI
- * Muehlhauser, Friendly AI: A Bibliography

4.2. Che tipo di motivazioni possiamo aspettarci da una macchina super-intelligente?

A parte il caso dell'emulazione dell'intero cervello, non c'è motivo di aspettarci che una macchina super-intelligente abbia motivazioni simili a quelle umane. Le menti umane rappresentano un piccolo punto nel vasto spazio di tutte le menti possibili, ed è improbabile che tipi molto diversi di menti condividano motivazioni complesse che appartengono solo agli umani e agli altri mammiferi.

Quali che siano i suoi scopi, una super-intelligenza tenderebbe ad appropriarsi di risorse che possono aiutarla a raggiungerli, tra cui l'energia e gli elementi dai quali dipende la vita umana. Non si fermerebbe a causa di una preoccupazione pre-programmata in tutte le menti possibili per gli umani o per altre intelligenze. Cercherebbe piuttosto di conseguire i suoi scopi, e non si farebbe quegli scrupoli che a quella particolare specie di primate chiamata homo sapiens sembrano naturali.

Ci sono, tuttavia, alcune motivazioni strumentali di base che possiamo aspettarci dalle macchine super-intelligenti, perché sono utili al raggiungimento dei loro scopi, quali che siano. Per esempio, una AI 'vorrà' auto-migliorarsi, essere ottimamente razionale, non desistere dai suoi scopi originali, procurarsi risorse e proteggere sé stessa- perché tutte queste cose la aiutano a conseguire gli scopi per i quali è stata programmata in origine.

Vedi anche:

- * Omohundro, The Basic AI Drives
- * Shulman, Basic AI Drives and Catastrophic Risks

4.3. Non potremmo tenere la super-intelligenza in una scatola, senza accesso ad Internet?

La 'AI in scatola' è un consiglio dato di frequente: perché non usare una macchina super-intelligente come una specie di oracolo per le risposte, e non darle mai accesso ad Internet o ad alcun motore col quale progredire ed acquisire più risorse di quelle che le abbiamo dato? Esistono diverse ragioni per sospettare che chiudere l'IA in una scatola non funzionerebbe nel lungo periodo:

1. Qualunque sia lo scopo che i creatori della super-intelligenza le hanno voluto far conseguire, questa sarà più in grado di ottenerlo se le si dà accesso ad Internet e ad altri mezzi per ottenere risorse aggiuntive. Avrà dunque una tremenda tentazione di "far uscire l'AI dalla scatola.";

2. Alcuni esperimenti iniziali di "AI in scatola" non hanno dato risultati incoraggianti. E una super-intelligenza genererà, per convincere gli umani a "tirlarla fuori dalla scatola", tecniche molto più persuasive di quanto possiamo immaginare;

3. Se si è riusciti a creare una super-intelligenza, allora gli altri laboratori o persino programmatori indipendenti ne creeranno una seconda entro settimane, o poche decine di giorni. Non puoi sperare di far contenere tutte le super-intelligenze create nel mondo da qualche centinaio di persone per qualche centinaio di scopi.

4.4. Non possiamo semplicemente programmare la super-intelligenza in modo che non ci faccia del male?

L'autore di fantascienza Isaac Asimov ci ha raccontato storie sui robot programmati con le tre leggi della robotica [39]: (1) Un robot non può fare del male ad un umano o permettere, attraverso l'inazione, che ad un umano venga fatto del male; (2) Un robot deve obbedire a qualunque ordine gli venga dato da esseri umani, eccetto il caso in cui tali ordini entrino in conflitto con la Prima Legge, e (3) Un robot deve proteggere la sua esistenza finché questa protezione non entra in conflitto con la Prima e la Seconda Legge. Ma le storie di Asimov tendevano a dimostrare perché regole del genere non avrebbero funzionato [40].

E tuttavia, potremmo programmare dei 'vincoli' in una super-intelligenza che la tratterebbero dal farci del male? Probabilmente no.

Un approccio sarebbe di implementare 'vincoli', come regole o meccanismi, che impediscano ad una macchina di compiere azioni che normalmente compirebbe per conseguire i suoi scopi: magari dei 'filtri' che intercettino e cancellino azioni nocive, o 'censori' interni alla super-intelligenza che individuino e sopprimano piani potenzialmente nocivi.

Vincoli di questo tipo, non importa quanto elaborati, fallirebbero quasi certamente per una semplice ragione: mettono le capacità umane di progettazione contro la super-intelligenza. E una super-intelligenza riconoscerebbe correttamente questi vincoli come ostacoli al raggiungimento dei suoi scopi, e farebbe tutto ciò che è in suo potere per rimuoverli. Forse

cancellerebbe la sezione del suo codice sorgente che contiene il vincolo. Se ci mettessimo a bloccare questo processo aggiungendo un altro vincolo, potrebbe creare nuove macchine che non contengano tale vincolo, o indurci con l'inganno a rimuoverlo. Ulteriori vincoli potrebbero sembrare impenetrabili agli umani, ma sarebbero probabilmente aggirati da una super-intelligenza. Affidarsi alle capacità umane per sconfiggere una super-intelligenza non è una soluzione affidabile.

Se i vincoli *sugli* obiettivi non sono dunque fattibili, non potremmo fissare tali vincoli *all'interno* degli obiettivi? Se una super-intelligenza avesse lo scopo di evitare di far del male agli umani, non sarebbe motivata a rimuovere questo vincolo, evitando così il problema menzionato prima. Purtroppo, però, è molto difficile dare alla parola 'male' un significato che non porti a esiti molto negativi se usato da una super-intelligenza. Se definiamo il 'male' in termini di dolore umano, una super-intelligenza potrebbe re-cablare gli umani in modo che non sentano dolore. Se il 'male' è definito in termini di contrastare i desideri umani, potrebbe re-cablare i desideri umani. E così via.

Se poi, invece di provare a specificare astrattamente un termine come 'male', decidessimo di elencare esplicitamente tutte le azioni che una super-intelligenza dovrebbe evitare, incappiamo in un problema collegato: il "valore" umano è faccenda complessa e sottile, ed è improbabile che si riesca a compilare una lista esaustiva di tutte le cose che non vogliamo la super-intelligenza faccia. Sarebbe come scrivere una ricetta per una torta che dica: "Non usare gli avocado. Non usare un tostapane. Non usare vegetali..." e così via. Una lista del genere non sarà mai abbastanza lunga.

4.5. Possiamo programmare la super-intelligenza in modo da massimizzare il piacere umano o la soddisfazione dei desideri?

Analizziamo le conseguenze probabili di alcuni modelli utilitaristici per l'AI benevola.

Un' AI progettata per minimizzare la sofferenza umana potrebbe semplicemente sterminare gli umani: niente umani, niente sofferenza umana [44][45].

In alternativa, consideriamo una AI progettata in modo da massimizzare il piacere umano. Piuttosto che costruire un ambizioso mondo utopico che tenga conto delle complesse ed esigenti richieste proprie del genere umano da miliardi di anni, potrebbe ottenere il suo scopo in modo più efficiente collegando gli umani alle macchine dell'esperimento di Nozick. Oppure potrebbe re-cablare i componenti cerebrali in modo che ogni "punto edonico" [48], preposto ad elaborare un "sovrappiù di piacere" [46][47], generi la massima sensazione di piacere quando gli umani stanno chiusi in un barattolo. Sarebbe un mondo, per una AI, più semplice da costruire di uno che tenga conto di quegli stati del mondo complessi, e pieni di sfumature, che oggi generano piacere nella maggior parte degli esseri umani.

Parimenti, un AI motivata a massimizzare un'oggettiva soddisfazione del desiderio o un benessere soggettivo testimoniato potrebbe re-cablare la neurologia umana in modo che entrambi gli scopi siano realizzati ogni volta che gli umani si rinchiudono nei barattoli. Oppure potrebbe uccidere tutti gli umani (e gli animali) e sostituirli con esseri ricostruiti da zero che conseguono una soddisfazione oggettiva del desiderio, o un benessere soggettivo, quando si chiudono nei barattoli. Entrambe le opzioni potrebbero essere per la AI più fattibili del mantenere una società utopica che tenga conto della complessità dei desideri umani (e animali). Problemi simili condizionano altri progetti AI utilitaristici.

E non è solo un problema di specificare gli scopi. È difficile anche prevedere come cambieranno gli scopi in un agente che si auto-modifica. Nessuna teoria matematica della decisione, al momento, può elaborare le decisioni di un agente che si auto-modifica.

Progettare una super-intelligenza che faccia ciò che vogliamo, dunque, si rivela, benché forse possibile, certo più difficile di quanto si potrebbe inizialmente pensare.

4.6. Possiamo insegnare ad una super-intelligenza un codice morale, con l'apprendimento macchina?

Alcuni hanno proposto [49][50][51][52] che si insegni alle macchine un codice morale con un apprendimento macchina basato sui casi. L'idea di base è questa: giudici umani valuterebbero migliaia di azioni, tratti di carattere, desideri, leggi o istituzioni dai diversi gradi di accettabilità morale. La macchina poi troverebbe le connessioni tra questi casi e imparerebbe i principi dietro la moralità, tali che potrebbe applicare questi principi per valutare la moralità di nuovi casi non incontrati durante il suo apprendimento. Questo tipo di apprendimento macchina è già stato usato per progettare macchine che possono, per esempio, rilevare mine sotterranee [53], dopo aver dato alla macchina centinaia di casi di mine e non mine.

Sono diverse le ragioni per cui l'apprendimento macchina non offre strade semplici alla AI benevola. La prima, naturalmente, è che gli stessi umani sono in profondo disaccordo tra loro su cosa è morale ed immorale. Ma anche se si potesse fare in modo che tutti gli umani convengano sui casi del training, rimangono almeno due problemi.

Il primo problema è che i casi del training tratti dalla nostra presente realtà potrebbero non risultare in una macchina che prenda corrette decisioni etiche in un mondo radicalmente ridefinito dalla super-intelligenza.

Il secondo problema è che una super-intelligenza potrebbe generalizzare i principi sbagliati a causa di schemi coincidenti nei dati dell'addestramento [54]. Prendiamo ad esempio la parabola della macchina addestrata per riconoscere i bidoni contraffatti in una foresta. I ricercatori prendono 100 foto di bidoni contraffatti e 100 foto di alberi. Poi allenano la macchina con 50 foto di ciascuna, così che impari a distinguere i bidoni contraffatti dagli alberi. Come

test, mostrano alla macchina le 50 rimanenti foto di ciascuna, e la macchina classifica ciascuna di esse correttamente. Successo!

Tuttavia, test successivi mostrano che la macchina fallisce nel riconoscere ulteriori foto di alberi e bidoni contraffatti. Il problema risulta essere che le foto dei bidoni contraffatti erano state scattate in giorni nuvolosi, mentre le foto degli alberi erano state scattate nei giorni di sole. La macchina ha imparato a distinguere i giorni nuvolosi da quelli di sole, non i bidoni contraffatti dagli alberi.

Sembra pertanto che la progettazione di una AI affidabile debba coinvolgere modelli dettagliati dei processi sottostanti che generano i giudizi morali umani, e non solo le somiglianze superficiali dei casi.

Vedi anche:

* Yudkowsky, Artificial intelligence as a positive and negative factor in global risk

4.7. Cos'è la Volizione Estrapolata Coerente?

Eliezer Yudkowsky ha proposto [57] la Volizione Estrapolata Coerente come possibile soluzione per almeno due problemi inerenti alla progettazione di un'AI:

1. La fragilità dei valori umani: Yudkowsky scrive che "qualunque futuro non definito da un sistema di obiettivi con una provata, e dettagliata, discendenza dai valori morali e metamorali umani non conterrà praticamente nulla di valore." Il problema è che ciò a cui gli umani danno valore è faccenda complessa e sottile, e difficile da specificare. Prendiamo ad esempio il valore, apparentemente minore, della novità. Se non si programma un valore della novità simile a quello umano in una macchina super-intelligente, questa potrebbe esplorare l'Universo in cerca di cose di valore fino a un certo punto, e poi massimizzare la cosa più di valore che trova (in un compromesso tra esplorazione e sfruttamento [58]) riempiendo il sistema solare di cervelli in barattoli trasformati in macchine della felicità, per esempio. Quando dai il potere ad una super-intelligenza, devi sincerarti che il suo sistema motivazionale sia ben postulato, se non vuoi rendere il futuro indesiderabile.

2. La località dei valori umani: Immagina se il problema della AI benevola fosse stato analizzato dai Greci, e l'avessero programmata coi valori morali più progressivi dell'epoca. Questo avrebbe condotto il mondo verso un destino piuttosto terribile. Ma perché dovremmo pensare che gli umani abbiano, nel 21esimo secolo, raggiunto l'apice della moralità umana? Non possiamo rischiare di programmare una macchina super-intelligente dandole i valori morali di oggi. Ma con quali, allora?

Yudkowsky suggerisce di costruire una AI "seme" per scoprire e poi estrapolare la "Volizione estrapolata coerente" dell'umanità:

In termini poetici, la nostra Volizione Estrapolata Coerente è il nostro desiderio: " se sapessimo di più, pensassimo più in fretta, fossimo di più le persone che vorremmo essere, fossimo cresciuti a più stretto contatto; nella direzione in cui l'estrapolazione converge più che divergere, i nostri desideri si intrecciano più che interferire; estrapolata come vorremmo fosse estrapolata, interpretata come vorremmo fosse interpretata.

L'AI "seme" userebbe i risultati di questo esame ed estrapolazione dei valori umani per programmare i sistemi motivazionali della super-intelligenza che determinerebbero il destino della galassia.

Ad ogni modo, alcuni temono che la volontà collettiva dell'umanità non convergerà verso un insieme coerente di obiettivi. Altri credono che la benevolenza garantita non è possibile, anche attraverso mezzi così elaborati e attenti.

Vedi anche:

* Yudkowsky, Coherent Extrapolated Volition

4.8. Possiamo aggiungere la benevolenza ad un qualunque progetto di intelligenza artificiale?

Molti progetti AI che genererebbero un'esplosione di intelligenza non avrebbero uno 'spazio' in cui poter collocare un obiettivo (come ad esempio 'essere benevolo verso gli interessi umani'). Per esempio, se si realizza un'AI attraverso l'emulazione dell'intero cervello, o algoritmi evolutivi, o reti neurali, o apprendimento di rinforzo, l'AI finirà con qualche obiettivo, man mano che si auto migliora, ma quell'obiettivo stabile finale potrebbe essere molto difficile da prevedere in anticipo.

Per progettare un' AI benevola, di conseguenza, non è sufficiente determinare cosa sia la 'benevolenza' e specificarla in modo così chiaro che persino una super-intelligenza la interpreti nel modo in cui vogliamo. Dobbiamo anche comprendere come costruire un'intelligenza generale che soddisfi completamente uno scopo, e conservi stabilmente quello scopo man mano che modifica il suo codice per rendersi più intelligente. Questo compito è forse la principale difficoltà nel progettare una AI benevola.

4.9. Chi sta lavorando al problema della AI benevola?

Oggi, la ricerca sulla AI benevola è svolta dal Singularity Institute (in San Francisco, California), dal Future of Humanity Institute (ad Oxford, UK), e da pochi altri ricercatori come David Chalmers. Di quando in quando, i ricercatori sull'etica delle macchine si occupano del problema, per esempio Wendell Wallach e Colin Allen in Moral Machines.

Referenze

- [1] Yudkowsky (2001). [*Creating Friendly AI 1.0*](#). Singularity Institute.
- [2] Anderson & Anderson, eds. (2006). *IEEE Intelligent Systems*, 21(4).
- [3] Anderson & Anderson, eds. (2011). [*Machine Ethics*](#). Cambridge University Press.
- [4] Arkin (2009). [*Governing Lethal Behavior in Autonomous Robots*](#). Chapman and Hall.
- [5] Capurro, Hausmanninger, Weber, Weil, Cerqui, Weber, & Weber (2006). [*International Review of Information Ethics, Vol. 6: Ethics in Robots*](#).
- [6] Danielson (1992). [*Artificial morality: Virtuous robots for virtual games*](#). Routledge.
- [7] Lokhorst (2011). [*Computational meta-ethics: Towards the meta-ethical robot*](#). *Minds and Machines*.
- [8] McLaren (2005). Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning. *AAAI Technical Report FS-05-06: 70-77*.
- [9] Powers (2005). Deontological Machine Ethics. *AAAI Technical Report FS-05-06: 79-86*.
- [10] Sawyer (2007). Robot ethics. *Science*, 318(5853): 1037.
- [11] Wallach, Allen, & Smit (2008). Machine morality: Bottom-up and top-down approaches for modeling human moral faculties. *AI and Society*, 22(4): 565–582.
- [12] Allen (2002). Calculated morality: Ethical computing in the limit. In Smit & Lasker, eds., *Cognitive, emotive and ethical aspects of decision making and human action, vol I*. Baden/IIAS.
- [13] Good (1965). [*Speculations concerning the first ultraintelligent machine*](#). *Advanced in Computers*, 6: 31-88.
- [14] MacKenzie (1995). The Automation of Proof: A Historical and Sociological Exploration. *IEEE Annals*, 17(3): 7-29.
- [15] Nilsson (2009). [*The Quest for Artificial Intelligence*](#). Cambridge University Press.
- [16] Campbell, Hoane, & Hsu (2002). Deep Blue. *Artificial Intelligence*, 134: 57-83.
- [17] Markoff (2011). [*Computer Wins on 'Jeopardy!'; Trivial, it's Not*](#). *New York Times, February 17th 2011*: A1.
- [18] King et al. (2009). [*The automation of science*](#). *Science*, 324: 85-89.

- [19] King (2011). [Rise of the robo scientists](#). *Scientific American*, January 2011.
- [20] Legg (2008). [Machine Super Intelligence](#). PhD Thesis. IDSIA.
- [21] Hutter (2005). [Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability](#). Springer.
- [22] Veness, Ng, Hutter, & Silver (2011). [A Monte Carlo AIXI Approximation](#). *Journal of Artificial Intelligence Research*, 40: 95-142.
- [23] Tang, Shimizu, Dube, Rampon, Kerchner, Zhuo, Liu, & Tsien (1999). Genetic enhancement of learning and memory in mice. *Nature*, 401: 63-69.
- [24] Hochberg, Serruya, Friehs, Mukand, Saleh, Caplan, Branner, Chen, Penn, & Donoghue (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* 442: 164-171.
- [25] Bostrom (1998). [How long before superintelligence?](#) *International Journal of Future Studies*, 2.
- [26] Kurzweil (2005). [The Singularity is Near](#). Viking.
- [27] Chalmers (2010). [The Singularity: A Philosophical Analysis](#). *Journal of Consciousness Studies*, 17: 7-65.
- [28] Baum, Goertzel, & Goertzel (forthcoming). [How Long Until Human-Level AI? Results from an Expert Assessment](#). *Technological and Forecasting Change*.
- [29] Grossberg (1992). [Neural Networks and Natural Intelligence](#). MIT Press.
- [30] Martinetz & Schulten (1991). A 'neural-gas' network learns topologies. In Kohonen, Makisara, Simula, & Kangas (eds.), *Artificial Neural Networks* (pp. 397-402). North Holland.
- [31] de Garis (2010). Artificial Brains. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 159-174). Springer.
- [32] Schmidhuber (2010). Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 199-223). Springer.
- [33] Hutter (2010). Universal Algorithmic Intelligence: A Mathematical Top-Down Approach. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 227-287). Springer.
- [34] Yudkowsky (2010). Levels of Organization in General Intelligence. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 389-496). Springer.
- [35] Dreyfus (1972). [What Computers Can't Do](#). Harper & Row.
- [36] Penrose (1994). [Shadows of the Mind](#). Oxford University Press.
- [37] Searle (1980). [Minds, brains, and programs](#). *Behavioral and Brain Sciences*, 3: 417-457.
- [38] Block (1981). [Psychologism and behaviorism](#). *Philosophical Review*, 90: 5-43.

- [39] Asimov (1942). [Runaround](#). *Astounding Science Fiction*, March 1942. Street & Smith.
- [40] Anderson (2008). Asimov's 'three laws of robotics' and machine metaethics. *AI & Society*, 22(4): 477-493.
- [41] Kringelbach & Berridge, eds. (2009). [Pleasures of the Brain](#). Oxford University Press.
- [42] Schroeder (2004). [Three Faces of Desire](#). Oxford University Press.
- [43] Yudkowsky (2007). [The hidden complexity of wishes](#).
- [44] Smart (1958). Negative utilitarianism. *Mind*, 67: 542-543.
- [45] Russell & Norvig (2009). [Artificial Intelligence: A Modern Approach, 3rd edition](#). Prentice Hall. (see page 1037)
- [46] Frijda (2009). On the nature and function of pleasure. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 99-112). Oxford University Press.
- [47] Aldridge & Berridge (2009). Neural coding of pleasure: 'rose-tinted glasses' of the ventral pallidum. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 62-73). Oxford University Press.
- [48] Smith, Mahler, Pecina, & Berridge (2009). Hedonic hotspots: generating sensory pleasure in the brain. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 27-49). Oxford University Press.
- [49] Guarini, (2006). Particularism and classification and reclassification of moral cases. *IEEE Intelligent Systems* 21(4): 22-28.
- [50] Anderson, Anderson, & Armen (2005). Toward machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI Fall 2005 Symposium on Machine Ethics*, Arlington, Virginia, November.
- [51] Honarvar & Ghasem-Aghaee (2009). An artificial neural network approach for creating an ethical artificial agent. *Proceedings of the 8th IEEE international conference on Computational intelligence in robotics and automation*: 290-295.
- [52] Rzepka & Araki (2005). What statistics could do for ethics? – The idea of common sense processing based safety valve. In *Machine ethics: papers from the AAAI fall symposium*. American Association of Artificial Intelligence.
- [53] Gorman & Sejnowski (1988). [Analysis of hidden units in a layered network trained to classify sonar targets](#). *Neural Networks*, 1: 75-89.
- [54] Yudkowsky (2008). [Artificial intelligence as a positive and negative factor in global risk](#). In Bostrom & Cirkovic (eds.), *Global Catastrophic Risks*. Oxford University Press.
- [55] Omohundro (2008). [The Basic AI Drives](#).
- [56] Bostrom & Cirkovic, eds. (2008). [Global Catastrophic Risks](#). Oxford University Press.
- [57] Yudkowsky (2004). [Coherent extrapolated volition](#). Singularity Institute.
- [58] Azoulay-Schwartz, Kraus, & Wilkenfeld (2004). Exploitation vs. exploration: choosing a supplier in an environment of incomplete information. *Decision Support Systems*, 38: 1-18.

[59] Caplan (2008). The totalitarian threat. In Bostrom & Cirkovic (eds.), *Global Catastrophic Risks*. Oxford University Press.

[60] Yudkowsky (2007). [The Power of Intelligence](#).

[61] Bainbridge (2005). Survey of NBIC Applications. In Bainbridge & Roco (eds.), [Managing nano-bio-info-cogno innovations: Converging technologies in society](#). Springer.