Euristiche cognitive ed etica dell'IA

Guglielmo Tamburrini
Università di Napoli Federico II

166° Mercoledì di Nexa
Politecnico di Torino, 8 novembre 2023

Il problema e la sua fonte

- **Etica dell'IA:**
    1. analisi delle implicazioni etiche dell'IA
    2. disegno e giustificazione di politiche etiche
    3. attuazione efficace delle politiche etiche
- **Problema:** Le *euristiche cognitive* ostacolano l'etica dell'IA?
    - Quali implicazioni per ciascuno dei passi 1-3?
- **Fonte:** processi decisionali euristici e analitici; bias dei processi euristici

processi decisionali

- **euristici:** scorciatoie rapide e cognitivamente meno onerose
  - Spesso automatici, carichi emotivamente, default…
- **analitici:** più lenti e cognitivamente costosi
  - Spesso consapevoli, riflessivi, distaccati emotivamente….
- pregiudizi ed errori sistematici (bias) nelle euristiche

- Sistema1/Sistema2
  - Bias da Sistema1: fallacia della congiunzione, ancoraggio, disponibilità cognitiva (Kahnemann & Tversky)

IA rossa e IA verde

un caso esemplare

Tamburrini G. (2022). *The AI carbon footprint and responsibilities of AI scientists*, in «Philosophies», vol. 7, doi: 10.3390/philosophies7010004

CO2 Equivalent Emissions (Tonnes) by Selected Machine Learning Models and Real Life Examples, 2022

Source: Luccioni et al., 2022; Strubell et al., 2019 | Chart: 2023 AI Index Report

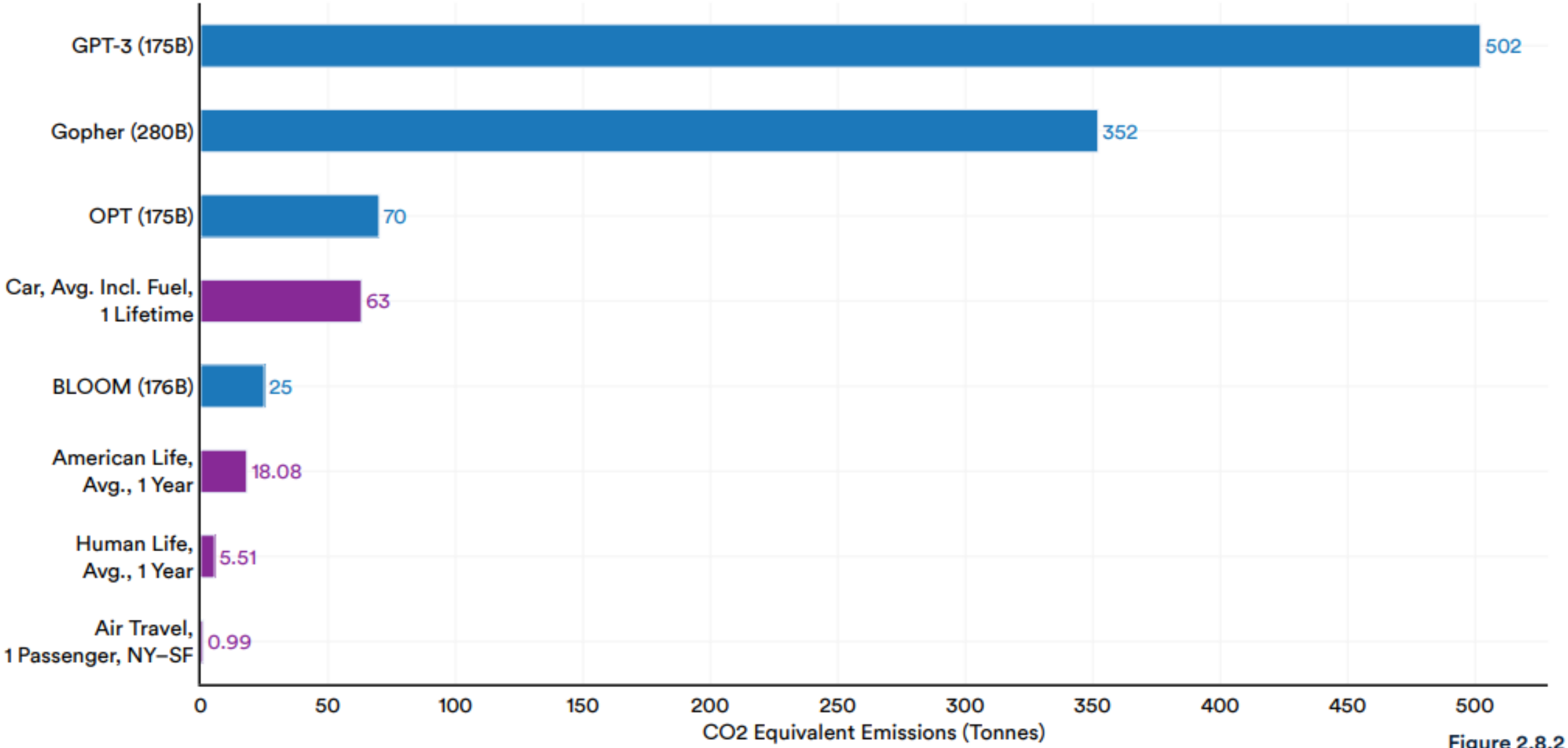

Figure 2.8.2

IA rossa: reazioni di sorpresa
Perché?

- euristica WYSIATI (What You See Is All There Is): Ci focalizziamo su ciò che è chiaramente visibile, già sappiamo o ci appare vero. Non facciamo attenzione a informazioni rilevanti ma poco note, difficili da reperire, da comprendere.
- Bias (per tutto il digitale)
  - Impercettibile la materialità dei processi di elaborazione simbolica
  - Ulteriore schermatura di metafore come *cloud computing*

IA rossa: reazioni di sorpresa
Perché 2

- ***Euristica dell'atteggiamento intenzionale*** (D. Dennett: *intentional stance*)
  - Attribuzione di razionalità ai sistemi dell'IA: convinzioni e desideri
  - Euristica utile per prevedere, spiegare, interagire
- Bias: vela sistematicamente l'impronta di carbonio
- da confrontare con
- Atteggiamento del progetto (*design stance*)
- Atteggiamento fisico (*physical stance*)

IA verde: buone pratiche per ricercatori
(e altre parti interessate)

- **Sviluppare** metodi, algoritmi e architetture di sistema energeticamente più efficienti
  - per addestramento e funzionamento a regime di modelli e applicazioni.
- **Utilizzare** processori hardware ottimizzati per IA
  - unità di elaborazione grafica (GPU), tensoriale (TPU), calcolatori neuromorfi (?), …
- **Eseguire** i calcoli presso CED alimentati da fonti rinnovabili o energeticamente più efficienti

- Patterson D., Gonzales J., Hölzle U., Le Q., Liang C., Mungia L. M., Rotchchild D., So D., Texier M., e Dean J. (2022). *The carbon footprint of machine learning will first plateau, and then shrink*, in «Computer», 55, pp. 18-28.

Il pacchetto di buone pratiche e le euristiche

- **Euristica dell'azione isolata**: chi compie un'azione per ridurre un rischio, ne rimane appagato e meno facilmente intraprende azioni aggiuntive per ottenere ulteriore protezione.
  - Weber E. U. (2006). *Experience-based and description-based perceptions of long-term risks: why global warming does not scare us (yet)*, in «Climatic Change», 77, pp. 103–120, doi: 10.1007/s10584-006-9060-3.
- **Euristica del gregge**: induce a conformarsi al comportamento dei più. non si abbandona l'IA rossa praticata dai più.

Buone pratiche come politiche etiche
Nuovi bias in campo

- giustificazione di obblighi imperfetti :
  - **Etica dei doveri**. Preservare l'agentività morale sul pianeta (Jonas)
  - **Contrattualismo**. Principio del giusto risparmio delle risorse per sostenere una società giusta (Rawls)
  - **Etica delle conseguenze**. Benessere aggregato per le generazioni future, per altre specie animali
- argomenti di tipo analitico e riflessivo, assenza di reciprocità, lontananza temporale o spaziale, carenza di connotazioni affettive

Carenza di connotazioni affettive
L'euristica del marcatore somatico (A. Damasio)

- Nessuna esperienza diretta e viscerale delle conseguenze climatiche negative dell'IA rossa per le generazioni attuali o future
- Nessuna riattivazione dei marcatori somatici nel processo decisionale
  - Damasio A. R. (1995). *L'errore di Cartesio. Emozione, ragione e cervello umano*. Milano, Adelphi.
- Bias: l'euristica del marcatore somatico non veicola segnali d'allarme quando si decide di procedere nel solco consueto dell'IA rossa

distanza spaziale, temporale e sociale

- Rappresentazione astratta e senza connotazioni affettive di eventi distanti temporalmente, spazialmente, ipotetici
    - Trope Y., Liberman N. (2010). *Construal-level theory of psychological distance*, in «Psychological Review», 117, pp. 440-463.
- Euristica di benefici psicologicamente distanti: sconti iperbolici
- Euristica dello status quo (avversione alla perdita): peso maggiore sui costi immediati dell'IA verde che sui benefici lontani
- Euristica del marcatore somatico: rappresentazioni astratte ed emotivamente povere

Contromisure: incentivi e spinte gentili

- **Che fare**
  - ridurre la distanza psicologica, indurre affetti negativi per l'IA rossa e positivi per l'IA verde, offrire ricompense immediate a chi si impegna per l'IA verde, in funzione della varietà e dell'efficacia delle azioni intraprese
- **Chi**
  - associazioni professionali e scientifiche meno soggette ai bias da euristiche cognitive
  - possono incidere sui comportamenti individuali e collettivi
  - il codice etico dell'IEEE assevera "l'impegno ad aderire alle pratiche per la progettazione etica e lo sviluppo sostenibile"

Incentivi
ricompense sul breve periodo, affetti positivi

- **Definire** una nuova idea di “buon” risultato in IA
  - accuratezza + efficienza energetica
- **Promuovere** carriere, fondi di ricerca, pubblicazioni di “nuovo tipo”
  - Affetti positivi e ricompense immediate per l’IA verde
- **Introdurre** competizioni sull’efficienza energetica (nella scia di scacchi, Go, Robocup, Darpa Grand Challenge,…)
  - Contrasto agli sconti iperbolici e al bias da azione singola
  - leva sul bias da gregge (conformity bias)

- tensioni etiche? come nel caso di altre spinte gentili in IA?
  - benessere ambientale/doveri di indirizzo vs libertà di ricerca?

Estensioni oltre l’IA rossa/verde

- principi etici generali per lo sviluppo responsabile dell’IA
- controllo umano significativo (MHC) sull’autonomia operativa dei sistemi dell’IA

- Tamburrini, G. (2020). *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*. Roma, Carocci Editore.

Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

The landscape of AI ethics principles

Anna Jobin, Marcello Ienca, Effy Vayena (2019), Nature Machine Intelligence 1, 389–399.

EU principles for thrustworthy AI (7)

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and Data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability

Troppi, troppo astratti e ramificati

- difficile rilevare tensioni tra principi vaghi e ciò che le euristiche mentali suggeriscono qui e ora
- bilanciamento analitico dei principi e bias da azione isolata

Autonomia operativa dell'IA e controllo umano significativo (MHC)

- MHC ≠ supervisione puramente nominale
  - implica *situational awareness* - consapevolezza della situazione
  - implica valutazione analitica e ponderata delle opzioni in campo
- presuppone
  - Capacità cognitive, disposizione al pensiero razionale, protocolli, addestramento e **tempo** adeguati

- Armi autonome, tempo limitato ed MHC

- Amoroso D. e Tamburrini G. (2021). *Toward a Normative Model of Meaningful Human Control over Weapons Systems*, in «Ethics and International Affairs», 35, pp. 245-272.

Integrare l'IA in NC3?

Tamburrini G. (2023). *Nuclear weapons and the militarization of AI*, in P. Cotta Ramusino et al. (a cura di), *Nuclear risks and arms control. Proceedings of the XXII Amaldi Conference*, Cham, Springer, pp. 147-158, doi: 10.1007/978-3-031-29708-3.

- "**AI should assist in** some aspects of nuclear command and control: **early warning, early launch detection**, and multi-sensor fusion…" (US National Security Commission on AI, *Final report* 2021, p. 104)
- AI may "increase reliability, reduce accident risks, shorten processing time, **buy more time for decision-makers"**
- Ma è davvero così?
  - fragilità, vulnerabilità, opacità dell'IA

Conclusioni e prossimi passi

- impegno etico di facciata (*ethical washing*)
- mancanza di informazioni, ignoranza indotta, manipolazione
- mancanza di forza morale (ἀκρασία)
- **Bias ed euristiche cognitive**

- Estendere l'analisi all'attuazione dei
  - principi etici per l'IA responsabile
  - MHC
- Estendere il repertorio di euristiche cognitive da considerare