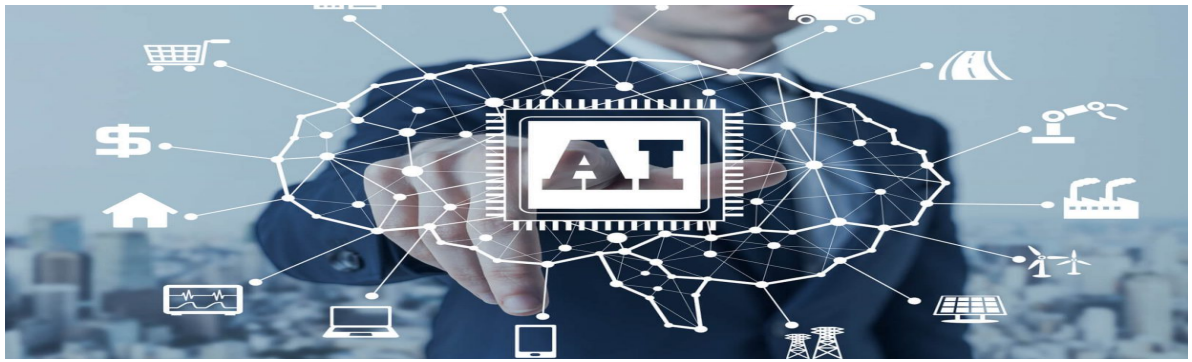
# SAFE-AI: E' possibile misurare l'etica dell'intelligenza artificiale?

Paolo Giudici

Professore Ordinario di Statistica, Universitá di Pavia



Sistema di intelligenza artificiale (D.D.L. 1146, Art. 2) :

> Un sistema automatizzato progettato per funzionare con livelli di autonomia variabile e che puó presentare adattabilitá dopo la diffusione
>
> che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali.

Figure: Rete neurale che apprende come classi care imprese a date, sulla base dei relativi dati di bilancio.

Statistical learning: regressione lineare, logistica, modelli econometrici e loro regolarizzazioni (Stepwise/Ridge/Lasso)

Machine learning: modelli ad albero, foreste casuali (random forests), ensemble models.

Deep learning: reti neurali a piú strati. Inizialmente monodirezionali, quindi ricorsive, quindi dotate di attenzione in tutte le direzioni (transformers)

General purpose AI (GPAI): modelli basati su pre-addestramento da grandi basi di dati (Large Language models), atti a risolvere diversi tipi di problemi

Agentic AI: modelli AI e GPAI sviluppati da squadre di agenti digitali, ciascuno dotato di obiettivi da raggiungere, in continuo dialogo, e coordinati a diversi livelli.

Edge AI, AGI, ASI.

E cienza : miglioramento dei processi, dei prodotti e dei servizi

Personalizzazione: prodotti e servizi su misura

Inclusione: possibile raggiungere nuovi utenti

Prossimitá: servizi maggiormente capillari e distribuiti

Sicurezza: una limitata robustezza puó condurre a rischi operativi, informatici e cibernetici.

Accuratezza: una limitata accuratezza puó condurre a rischi di modello

Equitá: una limitata equitá puó condurre a rischi legali e reputazionali

Interpretabilitá : una limitata interpretabilitá puó condurre a rischi di controllo e di governance

| Category | Subcategories |
|---|---|
| Autonomy | Autonomy/agency loss, Impersonation/identity theft, Personality loss, IP/cop ality rights loss |
| Emotional & psychological | Anxiety/distress/depression, Intimidation, Radicalisation, Dignity tion/objecti cation, Sexualisation, Self-harm, Addiction, Over-reliance, Coercion/manipulation |
| Financial & business | Financial/earnings loss, Con dentiality loss, Loss of productivity, Opportunity Business operations/infrastructure damage |
| Human rights & civil liberties | Discrimination, Bene ts/entitlements loss, Loss of human rights and fre Loss/violation of human rights and freedoms, Loss of right to due proc liberty and security |
| Physical | Loss of life, Bodily injury, Property damage, Personal health deterioration |
| Political & economic | Political instability, Institutional trust loss, Critical infrastructure damage, Po Economic instability, Electoral interference, Economic/political power |
| Psychological | Coercion/manipulation, Anxiety/distress, Sexualisation, Dehumanisation/obj loss, Alienation/isolation, Over-reliance, Addiction, Harassment/abuse/intir Radicalisation |
| Reputational | Defamation/libel/slander, Loss of con dence/trust |
| Societal & cultural | Public service delivery deterioration, Stereotyping, Information ecosyste lence/armed con ict, Damage to public health, Societal destabilisation, So loss/losses, Loss of creativity/critical thinking, Cheating/plagiarism |

Categories and their respective subcategories of harms caused by AI systems. Source: AIAAIC database.

La regolamentazione Europea AI Act ) segue un approccio basato sul rischio, e distingue fra:

- Attivitá proibite, (es. social scoring)
- Attivitá ad alto rischio, per le quali serve un sistema di gestione dei rischi (es. crediti, premi vita)
- Attivitá a rischio limitato, per le quali sussiste obbligo di trasparenza (es. chatbots)

Il D.D.L. 1146 , conforme alla normativa Europea, promuove un utilizzo corretto, trasparente e responsabile dell'AI, in una dimensione antropocentrica volta a coglierne le opportunitá e a garantire la vigilanza sui rischi economici e sociali e sullâimpatto sui diritti fondamentali.

Il D.D.L. 1146 ha diversi aspetti in comune con le linee guida della Ponti cia Commissione per lo Stato della Cittá del Vaticano in materia di intelligenza arti ciale.

Una AI responsabile dovrebbe esssere in grado di misurare e controllare i propri rischi , al ne di determinare un utilizzo etico e trasparente, in una dimensione antropocentrica e a dabile, nel rispetto della dignitá umana e del bene comune.

Occorre passare da AI a SAFE-AI: Sustainable, Accurate, Fair, Explainable AI

L'obiettivo delle ricerche del mio gruppo, e dei laboratori collegati, é lo sviluppo di metodi di misurazione SAFE-AI, in linea con l'approccio dell'associazione scienti ica aeai.org

Si basa su 4 indicatori statistici, espressi in termini percentuali: RGA, RGR, RGE, RGF. Tutti ricavati dagli stessi concetti matematici: la curva di Lorenz e l'indice di Gini (utilizzato dal 1905 per lo studio della distribuzione dei redditi).

Figure: Curva di Lorenz $L_y(\cdot)$, Curva di Lorenz duale $L_y^0(\cdot)$ e Curva di Concordanza (C) dove $p$ e $f(p)$ sono le percentuali di osservazioni e dei rispettivi valori cumulati.

Tutte le metriche sono state tradotte in un "pacchetto" di comandi di software Python, facilmente riproducibili, liberamente disponibili a:

https://github.com/GolnooshBabaei/safeaipackage .

Per un problema di credito al consumo abbiamo confrontato le previsioni di GPT, sia senza addestramento, che con addestramento (con dati di 80 clienti, "Informed GPT"), rispetto ad un classico modello di regressione logistica (con dati di 30000 clienti).

Figure: GPAI model

Table: Confronto fra modelli basato sulla capacitá predittiva (AUC) .

| Metodo | Min | Max | Media | Dev. Std |
|---|---|---|---|---|
| Informed GPT | 0.6260601915 | 0.6963064295 | 0.66655266764 | 0.0261630545 |
| GPT | 0.5895348837 | 0.6445964432 | 0.61264021887 | 0.0212562962 |
| LR | 0.7018467852 | 0.7950752394 | 0.7531874145 | 0.0314667993 |

Figure: Confronto fra: prestito/reddito e prestito ricevuto per due diversi gruppi di popolazione (Stato di New York) mediante curve di Lorenz ed indici di Gini.

I moderni modelli di AI sono "black-box": non trasparenti. Possono essere resi tali, ad esempio con il metodo Shapley Lorenz (SLV) , che spiega il contributo percentuale di ciascun predittore (es. prezzi noti) alla capacitá predittiva complessiva del prezzo da prevedere.

Figure: La gura mostra come cambiano le valutazioni sull'accoglienza di un reclamo bancario modi cando le parole del relativo testo, con intensitá crescente, per diversi modelli di AI. I modelli piú stabili, con indice di robustezza (RGR) vicino a 1, sono i migliori.

| Modello | Sust. | Acc. | Exp. | Prob | RMSE |
|---|---|---|---|---|---|
| MLP | 0.9661 | 0.4518 | 0.5114 | 0.7768 | 0.1046 |
| RBF | 0.9538 | 0.4519 | 0.5443 | 0.75454 | 0.0982 |
| NNAR | 0.7157 | 0.3718 | 0.2405 | 0.9361 | 0.1358 |
| LSTM | 0.9607 | 0.8186 | 0.1122 | 0.9118 | 0.0561 |
| GRU | 0.9244 | 0.8865 | 0.1778 | 0.8543 | 0.0439 |

Table: Confronto fra reti neurali MLP, RBF, NNAR, LSTM e GRU models, nella previsione di prezzi  nanziari, in termini di metriche SAFE vs il classico RMSE.

Lorenz MO. (1905). Methods of measuring the concentration of wealth. Publications of the American Statistical Association, 70.

Gini C. (1921). Measurement of Inequality of Incomes. Economic Journal, 31, 124-126.

Giudici P., Ra netti E. (2021). Shapley-Lorenz eXplainable Arti cial IntelligenceExpert Systems With Applications, 167, 114104.

Giudici P, Ra netti E. (2023).SAFE Arti cial Intelligence in Finance.Finance Research Letters, volume 56, 104088.

Giudici P, Ra netti E. (2024). RGA: a uni ed measure of predictive accuracy.Advances in Data analysis and classi cation.

Babaei, G., Giudici P, Ra netti E. (2025).A rank graduation Box for SAFE AI. Expert Systems With Applications