BRiO - Bias, Risk and Opacity in AI: design, verification and development of Trustworthy AI.

Giuseppe Primiero[1], Daniele Chiffi[2], Fabio D'Asaro[1], Gianluca Fasano[3], Roberta Ferrario[3], Marcello Frixione[4], Daniele Porello[4], Roberto Prevete[5], Viola Schiaffonati[2], Guglielmo Tamburrini[5]

[1]Università degli Studi di Milano, [2]Politecnico di Milano, [3]ISTC-CNR, [4]Università di Genova, [5]Università degli Studi di Napoli Federico II

{giuseppe.primiero;fabio.dasaro}@unimi.it, {daniele.chiffi;viola.schoiaffonati}@polimi.it, {gianluca.fasano;roberta.ferrario}@cnr.it, {marcello.frixione;daniele.porello}@unige.it, {roberto.prevete;guglielmo.tamburrini}@unina.it

Abstract

BRiO (https://sites.unimi.it/brio) è un progetto recentemente finanziato dal MUR nel quadro PRIN2020, che si propone di sviluppare criteri di progettazione per l'IA affidabile basati sull'analisi filosofica e formale della trasparenza, del bias e del rischio.

1 Introduzione

L'intelligenza artificiale ha ormai assunto una posizione centrale nella vita di ogni giorno, immersa sempre più spesso in processi decisionali, anche in situazioni critiche, che possono mettere a rischio l'incolumità delle persone. Esempi ormai noti variano dal campo medico per la formulazione di diagnosi, trattamenti e ricerca, a quello assicurativo o bancario. Tuttavia i risultati non sono sempre rassicuranti. Watson for Oncology ha ottenuto risultati piuttosto deludenti (fino al 30% sotto lo standard) in uno studio pilota svolto a Copenhagen; le applicazioni di IA usate per l'affidamento di prestiti sono soggette a distorsioni (bias) e difficoltà legali. I problemi connessi con opacità, bias e rischio sono sempre più pressanti: i sistemi di riconoscimento facciale come ImageNet e di decisione come COMPAS hanno bias razziali e gli algoritmi di apprendimento per rinforzo spesso usano funzioni di utilità discutibili e generano le ben note polarizzazioni sui social media. Spiegare i problemi dei sistemi di decisione automatici non è semplice: gli algoritmi "a scatola nera" richiedono una particolare attenzione, specialmente se lavorano su dati personali, come richiesto dall'art. 12 del EU GDPR.

Diviene quindi fondamentale lo sviluppo di sistemi di intelligenza artificiale affidabili (Trustworthy AI - TAI). In tale contesto, le linee guida dell'UE chiedono alla TAI di essere rispettosa delle leggi, di principi etici, tecnicamente e socialmente robusta e di rispettare i principi dell'autonomia umana, della prevenzione del danno, dell'imparzialità e della spiegabilità.

L'idea di generare spiegazioni dei sistemi di machine learning (ML) che siano comprensibili dalle persone è al centro della eXplainable Artificial Intelligence (XAI) e dei suoi diversi metodi: da approcci di basso livello come il Layer-wise Relevance Propagation (LRP) [Bach *et al.*, 2015], che associa un valore di rilevanza a ogni input; a quelli di alto livello come la comprensione meccanicista dei modelli [Schölkopf, 2019] o i modelli di semantiche markoviane, per esempio con probabilità imprecise [Termine *et al.*, 2021a]; oppure, infine, strategie intermedie come quella tesa a sviluppare spiegazioni in termini di proprietà degli input di medio livello, che rappresentino le parti più salienti dal punto di vista percettivo e cognitivo [Apicella *et al.*, 2020].

L'imparzialità, intesa come assenza di bias nella progettazione dei dati e degli algoritmi, e la prevenzione del danno, intesa come identificazione e mitigazione dei rischi associati all'uso di IA, richiedono un nuovo approccio che integri la creazione di ontologie e modelli simbolici a partire dai dati con l'epistemologia del rischio.

Infine, il rispetto dell'autonomia umana è stato affrontato con lo sviluppo dei requisiti di Meaningful Human Control (MHC) per i sistemi che operano in situazioni critiche per la sicurezza o in altri domini eticamente sensibili e con l'elaborazione di criteri di equità in termini di accessibilità ai sistemi indipendentemente dalle caratteristiche individuali.

BRiO (https://sites.unimi.it/brio) è un progetto recentemente finanziato dal MUR nel quadro PRIN2020 che si propone di sviluppare tali obiettivi definendo criteri di progettazione per la TAI basati sull'analisi filosofica della trasparenza, del bias e del rischio, integrando il loro studio formale e la loro implementazione tecnica in una serie di piattaforme, basate su tecniche e sistemi di apprendimento sia supervisionati che non supervisionati.

2 Obiettivi

BRiO è articolato su quattro obiettivi principali, che saranno perseguiti congiuntamente da cinque unità di ricerca.

Obiettivo 1: offrire un'analisi normativa della IA in quanto minata da bias e rischi, non solo rispetto alla propria affidabilità, ma anche alla sua accettazione sociale. Il nostro obiettivo è di analizzare gli elementi epistemologici ed etici nelle componenti del trust e di fornire una caratterizzazione che tenga conto sia del livello epistemico che di quello etico, osservando queste tecnologie al lavoro nell'interazione con umani e con altri agenti artificiali.

Obitettivo 2: definire un'ontologia formale che includa una tassonomia dei bias e dei rischi e delle loro mutue relazioni per i sistemi decisionali autonomi. Lo scopo è quello di offrire una caratterizzazione sistematica dei tipi di bias e di renderla formalmente e automaticamente identificabile e al tempo stesso di aiutare a individuare possibili fonti di rischio già nella fase di design, così da poter evitare danni che possano essere cagionati alle persone dall'interazione con l'IA.

Obitettivo 3: articolare una rappresentazione dei bias di classificazione in ML con concetti con soglie definiti da una lista di features pesate che rappresenti la loro importanza per la classificazione e che rifletta la probabilità delle proprietà a cui si applica l'algoritmo di ML. Su queste basi, intendiamo modellare il comportamento incerto dei sistemi di IA complessi tramite processi inferenziali probabilistici che includano gli output attesi e le credenze degli agenti rispetto a essi. Svilupperemo un approccio sintattico e semantico per modellare la divergenza tra i risultati attesi e quelli ottenuti e i margini di rischio di ottenere effetti indesiderati. Questi modelli possono poi essere implementati per una verifica formale con assistenti automatici alla verifica di prove e modelli.

Obitettivo 4: sviluppare un innovativo framework computazionale per le capacità di spiegazione dei sistemi TAI, che rappresenti le risposte del modello di ML in termini di strutture gerarchiche e di proprietà composizionali delle features di medio livello coordinate con il modello formale.

Questi obiettivi permetteranno di sviluppare una roadmap verso un framework ibrido neuro-simbolico per la spiegazione nella TAI, che integri la spiegazione dei comportamenti del modello ML con approcci simbolici (ontologie formali, inferenza logica, specifica formale e metodi di verifica).

3 Risultati correnti e prospettive

I risultati attesi di questo progetto discendono da risultati correnti ottenuti dalle varie unità, dagli obiettivi prefissati e dalla loro interazione.

Analisi etica ed epistemica. Poiché la fiducia è un concetto multidimensionale e i rischi e i bias possono riferirsi a elementi epistemici e non epistemici, abbiamo intenzione di integrare l'analisi normativa fornita dagli strumenti tipici dell'epistemologia e della filosofia della scienza con elementi offerti dall'etica applicata. Per valutare la dimensione di rischio epistemico della fiducia, ci baseremo sulla metodologia della valutazione probabilistica del rischio [Parkinson, 1993]. Le sue dimensioni non epistemiche saranno indagate principalmente attraverso la teoria del rischio [Roeser *et al.*, 2012], concentrandoci sulle implicazioni etiche e sociali. Indagheremo come affrontare gravi forme di incertezza legate alle tecnologie emergenti [Chiffi e Pietarinen, 2017] che si verificano in ambienti reali e complessi e che coinvolgono considerazioni di natura normativa. Per le dimensioni epistemiche del bias, indagheremo sugli errori sistematici che minano l'affidabilità e l'obiettività delle fonti alternative di dati e prove, riducendo la possibilità di amalgamare e replicare gli esperimenti. Per le componenti non epistemiche del bias nell'introduzione di una nuova tecnologia, proponiamo un'analisi normativa seguendo i 17 "sustainable goals" delle Nazioni Unite, in particolare, "salute e benessere", "uguaglianza di genere", "ridurre le disuguaglianze", "consumo e produzione responsabili". L'identificazione delle competenze morali da incorporare negli agenti di IA, la formulazione delle corrispondenti specifiche etiche e una tabella di marcia generale saranno realizzate alla luce degli attuali obiettivi e limiti dell'etica delle macchine [Dignum, 2018; Wallach e Allen, 2009]. L'analisi del requisito del Meaningful Human Control (MHC) sui sistemi di IA sarà condotta sulla base di modelli di limitazioni epistemiche degli agenti umani e artificiali, rispettivamente, sotto l'assunto ormai prevalente che una sola dimensione di MHC non si adatta a tutti i sistemi di IA e ai loro usi [Amoroso e Tamburrini, 2020].

Ontologia formale. Per procedere alla costruzione di un'ontologia completa di bias, rischio e fiducia, saranno analizzate e confrontate varie definizioni di tali nozioni tratte da diverse discipline (scienze cognitive, epistemologia, diritto), che saranno poi espresse attraverso assiomi formali. Per l'analisi e la formalizzazione ontologica faremo leva sulle tecniche sviluppate nella costruzione dell'ontologia DOLCE [Masolo *et al.*, 2003; Borgo e Masolo, 2009] e della metodologia OntoClean [Guarino e Welty, 2004], nonché sui recenti sviluppi nell'ontologia del rischio e della fiducia [Amaral *et al.*, 2019]. Una tale ontologia generale, fornita al momento della progettazione di sistema, può essere specializzata per modellare i sistemi decisionali che agiscono in condizioni di rischio.

Sistemi simbolici. Il nostro approccio alla progettazione dell'IA affidabile si basa sull'integrazione di approcci statistici puramente quantitativi con la modellazione simbolica delle proprietà dei sistemi e dei comportamenti degli utenti. Le "Description Logics" (DL) forniscono un'approssimazione computazionale implementabile dell'analisi ontologica riferita sopra. Le DL arricchite con concetti ponderati possono essere usate per fornire una rappresentazione simbolica, comprensibile all'uomo, delle classificazioni effettuate da un'interessante classe di algoritmi ML [Galliani *et al.*, 2020]. I processi di ragionamento incerto su queste classificazioni saranno modellati e la loro affidabilità computata attraverso calcoli di deduzione naturale [Primiero, 2016; Ceolin e Primiero, 2019; Primiero, 2020] e sistemi probabilistici tipati [D'Asaro e Primiero, 2021]. Indagheremo le proprietà meta-teoriche di tali sistemi per interpretare la sicurezza di un modello rispetto alla sua capacità di fornire un output previsto. Il ragionamento probabilistico può essere tradotto in un modello semantico dell'evoluzione temporale del sistema, attraverso la progettazione di logiche temporali computazionali per la valutazione di credenze ponderate sui comportamenti e la loro affidabilità [Chen *et al.*, 2016; Termine *et al.*, 2021b]. Questi approcci portano alla specifica delle proprietà formali e agli strumenti di verifica tramite assistenti alla prova, tecniche di model-checking e programmi di risposta. Basandosi su DL e i modelli inferenziali sviluppati, algoritmi di programmazione logica induttiva (ad esempio, programmi Prolog o Answer Set Programs con vincoli di peso) possono essere formulati per fornire accurate e succinte rappresentazioni umanamente comprensibili dei dati di allenamento, anche quando il dataset è piccolo, rumoroso e contiene relazioni di preferenza. Questi metodi applicati al ML possono migliorare la capacità di individuare automati-

camente i pregiudizi nascosti e indesiderati [D'Asaro *et al.*, 2020].

Sviluppo. Infine, ci proponiamo di sviluppare un framework XAI generale basato sulle indagini ottenuti nei precedenti obiettivi, da applicare a diverse definizioni computazionali di caratteristiche di medio livello ogni volta che

a) l'input di un sistema ML può essere codificato e decodificato sulla base di caratteristiche di medio livello organizzate gerarchicamente, e

b) un metodo LRP-like può essere applicato sia sul modello ML che sul decoder.

A questo scopo, progetteremo, svilupperemo e testeremo:

i) diversi algoritmi per un sistema encoder-decoder gerarchico in termini di caratteristiche di medio livello a diversi livelli di astrazione, partendo dalle principali soluzioni proposte in letteratura;

ii) algoritmi che integrano opportunamente il decoder gerarchico con il modello ML per ottenere un nuovo modello utilizzando le stesse codifiche di input a diversi livelli di astrazione;

iii) algoritmi di tipo LRP per retro-propagare i valori di rilevanza dall'output del nuovo modello al suo input, ottenendo valori di rilevanza per le caratteristiche di medio livello al livello gerarchico selezionato.

Le caratteristiche di medio livello identificate saranno valutate in termini di interpretabilità umana sia con metodi qualitativi (utenti umani) che quantitativi. Infine, per includere spiegazioni controfattuali, gli approcci computazionali saranno utilizzati anche per identificare le caratteristiche di medio livello relative alle risposte scartate.

In conclusione, BRiO si prefigge la creazione di una suite di tecniche di ML che producano modelli più spiegabili basati sull'identificazione del bias e del rischio, ma continuando a garantire un alto livello di performance di apprendimento.

Riferimenti bibliografici

[Amaral *et al.*, 2019] Glenda Amaral, Tiago Prince Sales, Giancarlo Guizzardi, e Daniele Porello. Towards a reference ontology of trust. In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, CTC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings*, page 3–21, Berlin, Heidelberg, 2019. Springer-Verlag.

[Amoroso e Tamburrini, 2020] Daniele Amoroso e Guglielmo Tamburrini. Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues. *Current Robotics Reports*, 1(4):187–194, December 2020.

[Apicella *et al.*, 2020] Andrea Apicella, Francesco Isgrò, Roberto Prevete, e Guglielmo Tamburrini. Middle-level features for the explanation of classification systems by sparse dictionary methods. *Int. J. Neural Syst.*, 30(8):2050040:1–2050040:17, 2020.

[Bach *et al.*, 2015] S. Bach, A. Binder, G. Montavon, F. Klauschen, K-R. Müller, e W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015.

[Borgo e Masolo, 2009] Stefano Borgo e Claudio Masolo. Foundational choices in DOLCE. In Steffen Staab e Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 361–381. Springer, 2009.

[Ceolin e Primiero, 2019] Davide Ceolin e Giuseppe Primiero. A granular approach to source trustworthiness for negative trust assessment. In Weizhi Meng, Piotr Cofta, Christian Damsgaard Jensen, e Tyrone Grandison, editors, *Trust Management XIII - 13th IFIP WG 11.11 International Conference, IFIPTM 2019, Copenhagen, Denmark, July 17-19, 2019, Proceedings*, volume 563 of *IFIP Advances in Information and Communication Technology*, pages 108–121. Springer, 2019.

[Chen *et al.*, 2016] Taolue Chen, Giuseppe Primiero, Franco Raimondi, e Neha Rungta. A computationally grounded, weighted doxastic logic. *Stud Logica*, 104(4):679–703, 2016.

[Chiffi e Pietarinen, 2017] Daniele Chiffi e Ahti-Veikko Pietarinen. Fundamental uncertainty and values. *Philosophia*, 45(3):1027–1037, 2017.

[D'Asaro *et al.*, 2020] Fabio Aurelio D'Asaro, Matteo Spezialetti, Luca Raggioli, e Silvia Rossi. Towards an inductive logic programming approach for explaining black-box preference learning systems. In Diego Calvanese, Esra Erdem, e Michael Thielscher, editors, *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, September 12-18, 2020*, pages 855–859, 2020.

[D'Asaro e Primiero, 2021] Fabio Aurelio D'Asaro e Giuseppe Primiero. Probabilistic typed natural deduction for trustworthy computations. In Dongxia Wang, Rino Falcone, e Jie Zhang, editors, *Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021) Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021), London, UK, May 3-7, 2021*, volume 3022 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[Dignum, 2018] Virginia Dignum. Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Inf. Technol.*, 20(1):1–3, mar 2018.

[Galliani *et al.*, 2020] Pietro Galliani, Guendalina Righetti, Oliver Kutz, Daniele Porello, e Nicolas Troquard. Perceptron connectives in knowledge representation. In C. Maria Keet e Michel Dumontier, editors, *Knowledge Engineering and Knowledge Management - 22nd International Conference, EKAW 2020, Bolzano, Italy, September 16-20, 2020, Proceedings*, volume 12387 of *Lecture Notes in Computer Science*, pages 183–193. Springer, 2020.

[Guarino e Welty, 2004] Nicola Guarino e Christopher Welty. An overview of ontoclean. In Steffen, editor, *Handbook on Ontologies*, pages 151–159, Germany, 2004. Springer.

[Masolo *et al.*, 2003] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, e Alessandro Oltramari. WonderWeb deliverable D18 ontology library (final). Technical report, IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web, 2003.

[Parkinson, 1993] David Parkinson. Risk: Analysis, perception and management. report of a royal society study group: Pp 201. the royal society. 1992. paperback£ 15.50 isbn 0 85403 467 6, 1993.

[Primiero, 2016] Giuseppe Primiero. A calculus for distrust and mistrust. In Sheikh Mahbub Habib, Julita Vassileva, Sjouke Mauw, e Max Mühlhäuser, editors, *Trust Management X - 10th IFIP WG 11.11 International Conference, IFIPTM 2016, Darmstadt, Germany, July 18-22, 2016, Proceedings*, volume 473 of *IFIP Advances in Information and Communication Technology*, pages 183–190. Springer, 2016.

[Primiero, 2020] Giuseppe Primiero. A logic of negative trust. *J. Appl. Non Class. Logics*, 30(3):193–222, 2020.

[Roeser *et al.*, 2012] Sabine Roeser, Rafaela Hillerbrand, Per Sandin, e Martin Peterson. *Essentials of risk theory*. Springer Science & Business Media, 2012.

[Schölkopf, 2019] Bernhard Schölkopf. Causality for machine learning. *CoRR*, abs/1911.10500, 2019.

[Termine *et al.*, 2021a] Alberto Termine, Alessandro Antonucci, Giuseppe Primiero, e Alessandro Facchini. Logic and model checking by imprecise probabilistic interpreted systems. In Ariel Rosenfeld e Nimrod Talmon, editors, *Multi-Agent Systems - 18th European Conference, EUMAS 2021, Virtual Event, June 28-29, 2021, Revised Selected Papers*, volume 12802 of *Lecture Notes in Computer Science*, pages 211–227. Springer, 2021.

[Termine *et al.*, 2021b] Alberto Termine, Giuseppe Primiero, e Fabio Aurelio D'Asaro. Modelling accuracy and trustworthiness of explaining agents. In Sujata Ghosh e Thomas Icard, editors, *Logic, Rationality, and Interaction - 8th International Workshop, LORI 2021, Xi'ian, China, October 16-18, 2021, Proceedings*, volume 13039 of *Lecture Notes in Computer Science*, pages 232–245. Springer, 2021.

[Wallach e Allen, 2009] W. Wallach e C. Allen. *Moral machines: teaching robots right from wrong*. OUP, 2009.