FIDUCIA, AFFIDABILITÀ

UNA QUESTIONE FILOSOFICA E TECNOLOGICA

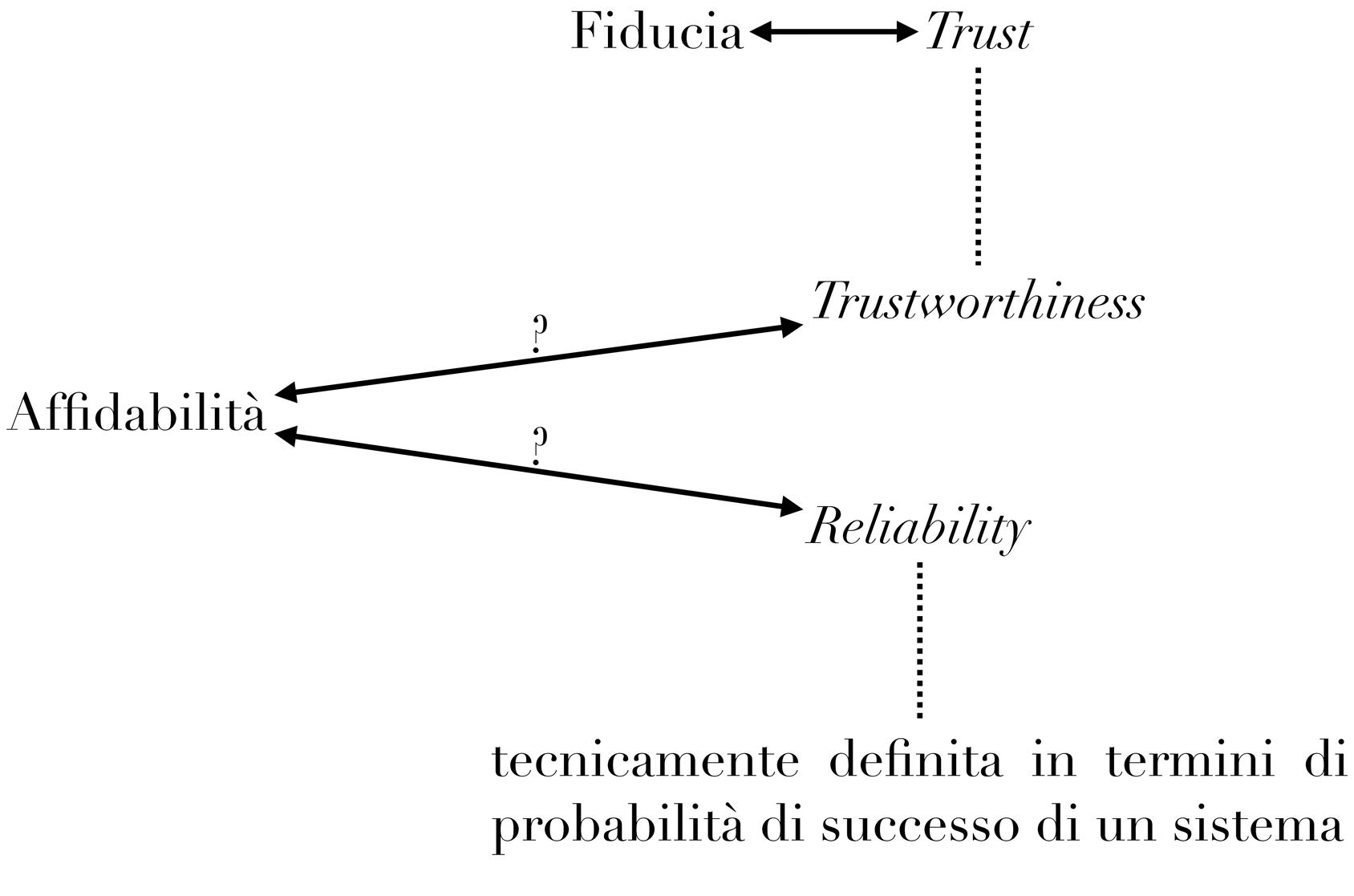
Francesco A. Genco
BRIO project (PRIN n. 2020SSKZ7R)
Laboratorio LUCI, Università di Milano

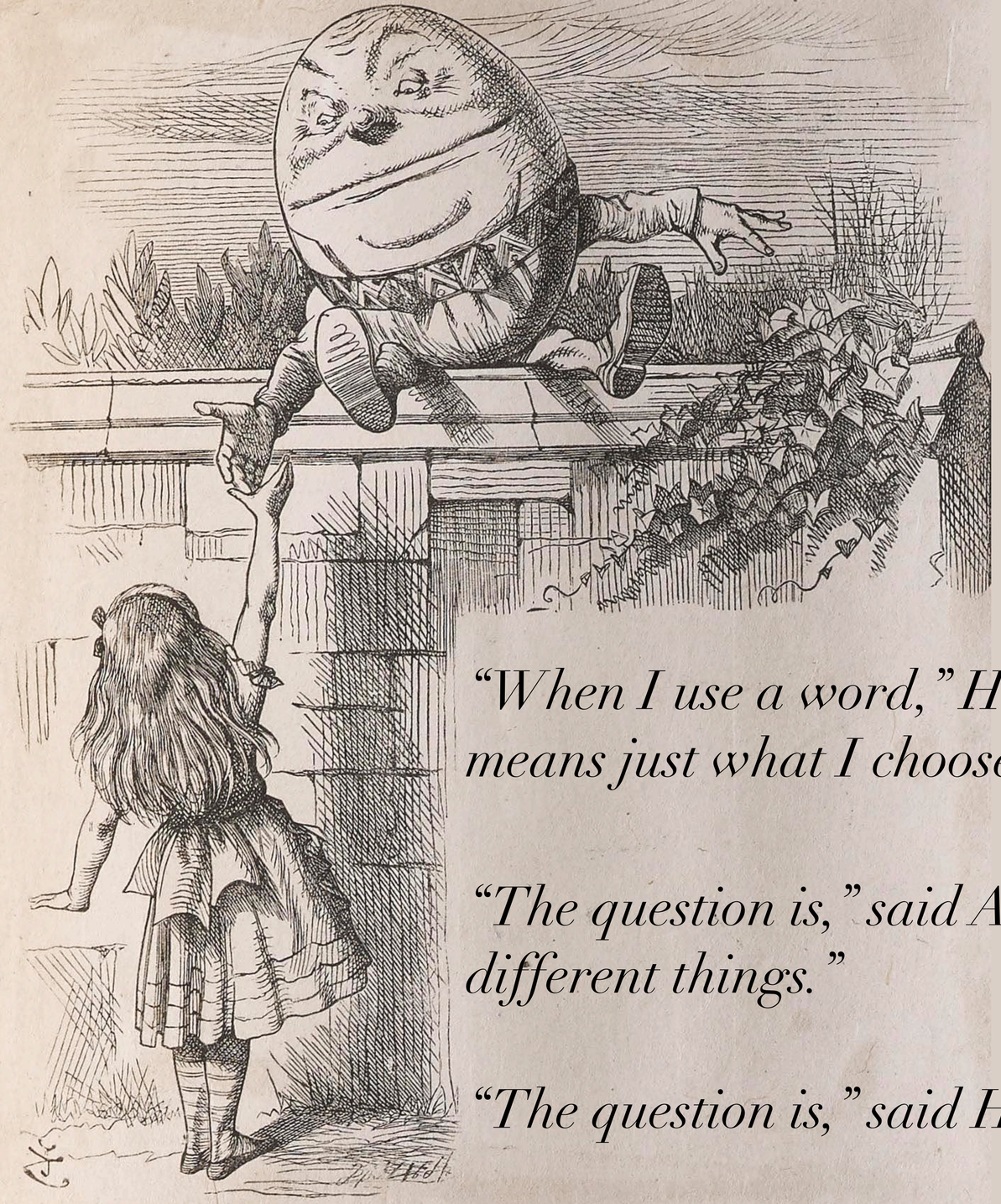
XV Convegno di Informatica Giuridica
Collegio Ghislieri, 23 novembre 2023

In Italia si parla di «AI affidabile», nella letteratura in lingua inglese, invece, di «*Trustworthy AI*».

Il problema è lo stesso—ovvero, sviluppare metodi per controllare che i sistemi di AI facciano quello che ci si aspetta facciano—ma…

UNA DISTINZIONE





L. Carroll. *Alice Through the Looking-Glass*, 1871.

"When I use a word," Humpty Dumpty said in rather a scornful tone, "it means just what I choose it to mean—neither more nor less."

"The question is," said Alice, "whether you CAN make words mean so many different things."

"The question is," said Humpty Dumpty, "which is to be master—that's all."

FIDUCIA

Rischio

«Trust is a solution to specific problems of risk»

N. Luhmann. *Familiarity, confidence, trust: Problems and alternatives*, 2000.

FIDUCIA

Assenza di garanzie

«Trust […] refer[s] to expectations which may lapse into disappointments»

N. Luhmann. *Familiarity, confidence, trust: Problems and alternatives*, 2000.

FIDUCIA

Dipendenza dal comportamento altrui per trarne vantaggio

«You can avoid taking the risk, but only if you are willing to waive the associated advantages.»

N. Luhmann. *Familiarity, confidence, trust: Problems and alternatives*, 2000.

FIDUCIA

Una decisione consapevole

«If you choose one action in preference to others in spite of the possibility of being disappointed by the actions of others, you define the situation as one of trust. […] In the case of [disappointed] trust you will have to consider an internal attribution and eventually regret your trusting choice.»

N. Luhmann. *Familiarity, confidence, trust: Problems and alternatives*, 2000.

FIDUCIA

Non un semplice calcolo razionale

«[T]rust is only possible in a situation where the possible damage may be greater than the advantage you seek»

N. Luhmann. *Familiarity, confidence, trust: Problems and alternatives*, 2000.

Queste componenti dei rapporti di fiducia tra umani:

- Rischio
- Assenza di garanzie complete
- Decisione di dipendere da altri per il proprio vantaggio
- Decisione che trascende un semplice calcolo razionale

possono essere presenti anche nel rapporto tra esseri umani e sistemi di intelligenza artificiale.

LA COMPONENTE MORALE DELLA FIDUCIA

Il fatto che si parli di *tradire* la fiducia di un individuo, però, suggerisce che possa esserci anche una componente morale nelle relazioni di fiducia.

> «*Plausible conditions for proper trust will be that it survives consciousness, by both parties, and that the trusted has had some opportunity to signify acceptance or rejection, to warn the trusting if their trust is unacceptable.*»

Il mio fidarmi di un individuo sembrerebbe implicare il pensiero, da parte mia, che questo individuo sia consapevole di una qualche forma di accordo che lo lega a me.

A. Baier. *Trust and Antitrust*, 1986.

FIDUCIA E MACCHINE

È possibile parlare di fiducia in un sistema di intelligenza artificiale senza ignorare le possibili implicazioni morali di una tale definizione?

FIDUCIA E MACCHINE

C’è chi sostiene che, con lo sviluppo della tecnologia, si raggiungerà un momento in cui non sarà più possibile considerare certi agenti artificiali come privi di *responsabilità morale*.

Se così fosse, nulla ci impedirà allora di definire un agente artificiale come *degno di fiducia*. Nulla tranne il comportamento dell’agente stesso e il suo atteggiamento nei nostri confronti, ovviamente.

FIDUCIA E MACCHINE

Altri, invece, ritengono insensato attribuire *responsabilità morale* a un agente artificiale, indipendentemente da qualsiasi possibile sviluppo tecnologico.

Questi ritengono infatti che

- la capacità di comprendere la rilevanza morale di certe azioni,
- la capacità di provare senso di colpa o empatia,
- la capacità di provare dolore (a seguito, ad esempio, di una punizione)

non siano artificialmente riproducibili ma esclusivamente simulabili.

FIDUCIA, MACCHINE E AFFIDABILITÀ

È possibile, in ogni caso, ridurre il problema della *fiducia* nei sistemi di intelligenza artificiale a quello della loro *affidabilità* (nel senso di *reliability*).

Si mette da parte, in questo modo, la possibile componente morale della questione per potersi concentrare sul problema tecnico relativo alla prevedibilità del comportamento del sistema e alla sua valutazione secondo criteri di sicurezza ed efficienza.

Un problema tradizionale, per risolvere il quale esistono strumenti efficaci.

UNA DIFFICOLTÀ ULTERIORE

L'imprevedibilità del comportamento dei sistemi di intelligenza artificiale, però, è intimamente legata alla loro autonomia e alle scarsa comprensibilità dei meccanismi interni che li governano, caratteristiche che li distinguono dai sistemi artificiali tradizionali.

IL PROBLEMA DELLA SPIEGAZIONE
SISTEMI TRASPARENTI

In un sistema di calcolo tradizionale (*simbolico*), ogni simbolo sintattico ha un significato. Un simbolo può, ad esempio, rappresentare un dato, un concetto, un'azione…

Formare astrazioni comprensibili del comportamento del sistema è dunque possibile, e questo permette di spiegare il comportamento di sistemi anche molto complessi.

Un sistema di questo genere viene detto *trasparente*.

IL PROBLEMA DELLA SPIEGAZIONE
SISTEMI OPACHI

Nell'ambito dell'intelligenza artificiale contemporanea, invece, le tecniche maggiormente usate sono quelle di *Machine Learning*.

Queste tecniche consistono, prevalentemente, nell'impiego di sistemi di elaborazione dell'informazione le cui componenti formali non hanno necessariamente un significato definito.

Questi sistemi di elaborazione dell'informazione vengono addestrati in modo che sviluppino (in maniera più o meno autonoma) la capacità di risolvere un dato problema.

IL PROBLEMA DELLA SPIEGAZIONE
SISTEMI OPACHI

Anche a seguito dell'addestramento non è sempre possibile attribuire un significato alle componenti formali di un sistema di Machine Learning.

Quindi, spiegare perché il sistema ha un certo comportamento può essere estremamente problematico.

Un sistema di questo genere viene detto *opaco*.

1. I sistemi di Machine Learning si addestrano (più o meno autonomamente) su collezioni di dati e possono sviluppare comportamenti indesiderati dovuti a difetti di queste.

2. L'*opacità* di questi sistemi rende difficile individuare eventuali comportamenti indesiderati e riconoscerli come tali.

3. I compiti che svolgono non sono sempre associabili a criteri ovvi di successo o correttezza.

In conclusione, i motivi per i quali i sistemi di Machine Learning sono versatili, autonomi ed efficaci sono gli stessi per i quali questi sistemi possono risultare imprevedibili e problematici in relazione alla verifica del comportamento.

Sia nel bene che nel male, la situazione è molto diversa rispetto a quella che presentano i sistemi tradizionali.

Anche se ci limitiamo all'uso del termine *affidabilità* (*reliability*), lo usiamo in un un'accezione che sembra, di fatto, avvicinarlo considerevolmente al campo semantico della *fiducia* (*trust*).

ALCUNI SISTEMI DI CONTROLLO

λ-Trust e TPTND (Trustworthy Probabilistic Typed Natural Deduction)

Definizione e verifica di criteri di affidabilità

Applicazione BRIO x Alkemy

Bias detection

Analisi delle caratteristiche significative dell'input

Ontologie formali per la costruzione di alberi di decisione

https://sites.unimi.it/brio/brio-x-alkemy-it/

Genco e Primiero. *A Typed λ-Calculus for Establishing Trust in Probabilistic Programs*.

D'Asaro, Genco e Primiero. *Checking Trustworthiness of Probabilistic Computations in a Typed Natural Deduction System*.

Confalonieri, Galliani, Kutz, Porello, Righetti e Troquard. *Towards Knowledge-driven Distillation and Explanation of Black-box Models*.

Apicella, Isgrò, Prevete e Tamburrini. *Middle-Level Features for the Explanation of Classification Systems by Sparse Dictionary Methods*.

https://sites.unimi.it/brio

SISTEMI DI CONTROLLO

λ-Trust & TPTND (Trustworthy Probabilistic Typed Natural Deduction)

Sistemi formali per la definizione e la verifica di criteri di affidabilità rispetto a sistemi di intelligenza artificiale.

Applicazione BRIO x Alkemy

Sistema software per la *detection* di *bias* in sistemi predittivi.

https://sites.unimi.it/brio/brio-x-alkemy-it/

Genco e Primiero. *A Typed λ-Calculus for Establishing Trust in Probabilistic Programs*.
D'Asaro, Genco e Primiero. *Checking Trustworthiness of Probabilistic Computations in a Typed Natural Deduction System*.

SISTEMI DI SPIEGAZIONE

Analisi delle caratteristiche significative dell'input

Strumenti di analisi di quelle componenti dell'input che sono rilevanti rispetto al risultato ottenuto e che corrispondono a parti percettivamente salienti dell'input stesso.

Ontologie formali per la costruzione di alberi di decisione

Costruzione di *alberi di decisione* guidata tramite *ontologie formali* per rendere più comprensibili le spiegazioni del comportamento di sistemi di intelligenza artificiale.

Confalonieri, Galliani, Kutz, Porello, Righetti e Troquard. *Towards Knowledge-driven Distillation and Explanation of Black-box Models*.
Apicella, Isgrò, Prevete e Tamburrini. *Middle-Level Features for the Explanation of Classification Systems by Sparse Dictionary Methods*.