Rino Falcone

Cognizione e Sistemi Intelligenti: il ruolo della Fiducia

Laboratorio dell'ISPF, XIX, 2022

[5]
DOI: 10.12862/Lab22FLR

0. *Introduzione*

In questo lavoro si vuole mostrare la rilevanza che assume il "cognitive modeling" (modellazione cognitiva) nell'approccio allo sviluppo dei sistemi artificiali intelligenti. Essendo le nostre società sempre più pervase da questi sistemi, diventa evidente la necessità di un chiarimento, tanto teorico quanto pratico-applicativo, di alcune attitudini tipicamente sviluppate dagli umani nell'esercizio della loro inter-relazione complessa e di quanto sia necessario trasferire quelle stesse attitudini, adattate e rielaborate, ai sistemi artificiali con cui si pretende di interagire con analoga complessità. Tra queste attitudini ricoprono grande significato quelle di *fiducia* e *affidabilità*.

1. *Intelligenza Artificiale e tecnologie intelligenti*

L'*intelligenza* cui l'Intelligenza Artificiale (IA) è ispirata e modulata su quella degli umani. Nella duplice modalità di *debole* o *forte*.

Per *IA debole* si intende un approccio in cui l'obiettivo, ossia la riproduzione del comportamento intelligente, sia piuttosto indipendente dal modo in cui gli umani lo realizzano: l'obiettivo è la risoluzione del problema e non l'analogia al mezzo di risoluzione.

Per l'*IA forte*, viceversa, l'obiettivo è di ottenere la soluzione nello stesso modo in cui l'otterrebbe un umano (il riferimento è quindi il confronto con i *processi neuro-socio-cognitivi* dei sistemi naturali).

La stessa *IA forte* ha quindi almeno due ulteriori prospettive: la riproduzione delle strutture biologiche e dei processi biologici degli esseri viventi (prospettiva *naturalista-connessionista*); oppure la simulazione della logica della psyché degli esseri umani (prospettiva *mentalista-cognitivista*).

Per modellare e costruire tecnologie intelligenti è quindi necessario stabilire a quale approccio aderire. In ogni caso sarà necessario affrontare varie problematiche fondamentali. In particolare andrà stabilito quali *architetture, strutture e processi* caratterizzeranno questi sistemi; andranno inoltre analizzate quali *relazioni* dovranno avere questi sistemi con *l'ambiente*: tanto con l'ambiente interno (il corpo, nel caso di robot ad esempio: quindi il rapporto "mente-corpo") quanto con l'ambiente esterno; ed infine, come questi sistemi si *modificano nel tempo*, per esempio grazie all'apprendimento oppure grazie a processi evolutivi generazionali. Questi sistemi sono quindi il risultato di un approccio inter e multi disciplinare con metodologie teoriche, formali, sperimentali, computazionali, simulative, robotiche e così via.

Come sappiamo le tecnologie intelligenti hanno avuto una rapida evoluzione negli ultimi lustri. Le loro basi sono da individuarsi principalmente nello straordinario incremento della potenza di calcolo concentrata in spazi miniaturizzati rispetto al passato; negli sviluppi delle reti di comunicazione e della sensoristica sofisticata; negli algoritmi di apprendimento e di AI; nello sviluppo di

nuovi materiali; negli aggiornati meccanismi di controllo; nella capacità di sviluppare basi di dati enormi (big data).

Questo ha condotto, come tutti siamo in grado di apprezzare, alla *trasformazione degli ambienti commerciali intellettuali* e *sociali* in cui gli umani operano: il commercio elettronico e le transazioni economiche in rete relative a prodotti e cose sono ormai di assoluta consuetudine; le informazioni legate alla cultura, ma anche gli avvenimenti più complessi in questo ambito, si avvalgono di internet e degli strumenti di supporto digitali; la stessa vita sociale degli umani è, non di rado, a cavallo tra il reale e il virtuale: condivisioni di amicizie, incontri, messaggistica varia (immagini, film, altro), opinioni, gruppi di interesse sono spesso direttamente virtualizzati o comunque supportati da strumenti digitali.

Il fatto che tutti questi comportamenti possono essere *osservati*, *memorizzati* ed *utilizzati* ci porta alle problematiche che sono ben note relative al cosiddetto "capitale di sorveglianza" (Zuboff, 2019). Con le note capacità di ottenere delle previsioni particolarmente efficaci sul futuro e al tempo stesso con i rischi cui queste previsioni sono soggette (conformismo, riproposizione di schemi inadeguati seppur prevalenti. E così via).

2. *La sfida dei sistemi intelligenti*

La sfida che le tecnologie intelligenti stanno rappresentando per le nostre società risulta come appena detto di primaria rilevanza. E le potenzialità che si prospettano sono davvero interessanti e non banalmente prevedibili. Crescerà la pervasività dei sistemi intelligenti in ciascun ambito delle nostre esistenze, individuale e sociale, con apparati e sistemi sempre più attrezzati dal punto di vista della capacità di adattamento ed efficacia collaborativa. Una dimensione particolarmente rivoluzionaria riguarderà l'impatto che queste tecnologie potrebbero produrre sui modelli socio-cognitivi della nostra realtà (nelle dimensioni: individuali, interazionali e collettive); esse potrebbero simulare e supportare i meccanismi di base del ragionamento e della comunicazione; svolgere un ruolo complementare alle nostre abilità ma anche ridefinire i paradigmi principali che li governano.

Così come la *realtà aumentata* produce un'estensione dello spettro percettivo e informativo cui è possibile accedere da parte degli umani (giochi, visite ai siti storico-culturali o di altro genere, pratiche lavorative o educative, etc.), analogamente è possibile pensare a come le tecnologie intelligenti possano estendere le nostre funzioni cognitive, predisponendoci a valutazioni previsionali sul futuro, attraverso processi simulativi (*ragionamento aumentato*). Per esempio, potremmo avere suggerimenti on-line, considerando gli obiettivi che vogliamo raggiungere e le condizioni note di partenza (non necessariamente forniti, gli uni e le altre, da noi al sistema ma dedotti direttamente da esso), sui differenti scenari possibili a valle di ciascuna scelta disponibile nel repertorio delle nostre azioni (individuali e/o collettive).

Ci troveremo quindi crescentemente dentro un contesto/mondo sempre più *ibrido*: in cui opereremo tanto nel *reale* quanto nel *virtuale* e con entità sia *na-*

turali che *artificiali*. A volte senza saper chiaramente distinguere tra le due dimensioni e comunque spesso sviluppando pratiche sempre più interconnesse (Falcone et al., 2018).

Subiranno di conseguenza una forte ridefinizione alcune attitudini sociali a cui ci siamo significativamente affidati nella storia delle nostre interazioni con gli altri: delega, controllo, *fiducia*, autonomia, responsabilità, etc. Vedremo nel prossimo paragrafo un approfondimento e ruolo di queste attitudini.

3. *Ruolo della Fiducia nei sistemi intelligenti*

Data la necessità di una riconversione dell'apparato cognitivo di interazione, è importante approntare una analisi teorica delle principali attitudini a partire dalla *fiducia*. Si tratta in questo caso, da una parte, di riaggiornare il senso del fidarsi, ossia la ricostruzione di questa attitudine attraverso una rielaborazione dei soggetti cui la fiducia va rivolta, delle loro nuove proprietà, della stessa natura interazionale a cui ci sottopongono, dei canali di comunicazione che ci permettono di entrare in relazione con loro, dei nuovi e sofisticati contesti in cui tutto ciò può avvenire.

D'altra parte, si tratta di dotare i sistemi artificiali intelligenti con questa stessa attitudine per fare in modo che possano anch'essi esercitare l'atto di fidarsi nei confronti di altri soggetti, artificiali o umani.

Ma per dotare i sistemi intelligenti di questa funzione è necessario essere in grado di definirla e di descriverla analiticamente, di stabilire i processi e gli ingredienti necessari a svilupparli. Serve quindi una teoria socio-cognitiva della fiducia.

3.1 *Il concetto di fiducia*

Come ognuno sa, la fiducia gioca un ruolo importante nelle nostre vite, nelle nostre decisioni, nei nostri comportamenti. Spesso risulta determinante per molte delle nostre scelte.

Anche per questo la fiducia è stata oggetto di approfondimento e di studio per decenni e in molti differenti ambiti scientifici: dalla psicologia, alla filosofia, alla sociologia, alla economia, alla biologia (Luhmann, 1979, Gambetta, 1988, KJones, 1996, Hardin, 2002). Eppure, o forse proprio per questa vastità di punti di vista con cui può essere analizzata, non esiste una definizione condivisa e unica di fiducia.

Potremmo dire che essa si caratterizza in vari modi:

- come *un ragionamento:* essa implica infatti un modo di processare in modo razionale delle situazioni/caratteristiche/dati, di svolgere delle considerazioni logico-deduttive su questi elementi, individuare ipotesi considerabili oggettive e convincenti attraverso cui procedere ad un giudizio e quindi ad una decisione.
- come *un sentimento*: in questo caso entrano in gioco non processi razionali ma piuttosto elementi di affettività ed emozionalità.

- come una *attitudine implicita*: si tratta di caratterizzarla da un punto di vista istintuale, precognitivo se si vuole e persino pre-emotivo. Ci possono essere differenze, anche significative, tra soggetti differenti nel fidarsi rispetto ad identiche situazioni razionali o emotive.

Ed ancora, la fiducia è una *relazione*: un collegamento tra il fidante e qualcun altro. Ed è anche uno *stato mentale*: analizzabile attraverso gli ingredienti tipici della cognizione.

Il punto è di tenere insieme tutte queste cose appena indicate, che ciascun umano prova/sente in molteplici modi e verso molteplici interlocutori/oggetti/eventi.

3.2 *Verso la formalizzazione della fiducia*

Volendo quindi dare una veste *formale e operazionale* al concetto di fiducia, possiamo dire che:

1) la fiducia è uno *stato e attitudine mentale*:
- *ibrido*: ossia tanto cognitivo, quanto affettivo;
- *con struttura composita*: riferibile a differenti ingredienti cognitivi: credenze, scopi, intenzioni, aspettative, etc.;
- *orientato a differenti oggetti e dimensioni*.

2) la fiducia è un *fenomeno ricorsivo*: è possibile individuare delle ragioni per fidare e per ciascuna di queste ragioni è possibile individuare altre ragioni per fidarsi di esse stesse (e così via).

3) la fiducia è un *processo mentale e pragmatico*: ossia può essere considerato come una *valutazione* (una semplice attitudine mentale, una predisposizione, una valutazione preventiva non necessariamente connessa all'atto di fiducia); ma può anche essere considerato come una *decisione* (anche eventualmente dopo aver preso in considerazione comparazioni tra soggetti da fidare) ed infine può essere considerato come una *azione* (un comportamento, un atto intenzionale). In generale, è possibile pensare che valutazione, decisione e azione siano rispondenti a stati mentali coerenti e precondizione dei successivi (lo stato mentale della valutazione è predisponente lo stato mentale della decisione e lo stato mentale della decisione è predisponente lo stato mentale della azione). In realtà può succedere che questa coerenza non sia sempre rispettata.

4) la fiducia è un *fenomeno dinamico*, non solo perché cambia nel tempo ma anche perché è possibile derivare fiducia da fiducia per esempio attraverso i fenomeni di transitività, o di categorizzazione, o dalla fiducia nelle credenze per fidare e così via.

Introduciamo quindi una formula esplicativa del concetto di fiducia:

$$Trust(X\ Y\ \tau\ C)$$

- dove X rappresenta il *trustor*, l'agente che si affida, che sente fiducia; questo deve essere un *agente cognitivo* dotato di scopi e credenze interni ed espliciti;
- Y è il *trustee*, l'agente/entità che deve essere fidato; Y non è necessariamente un agente cognitivo, nel caso in cui lo è la relazione si connota come fiducia sociale.
- C è il *contesto* (l'ambiente) in cui il trustee deve operare per realizzare il compito delegato;
- $\tau = (a,g)$ *è* il *compito delegato*; come si vede corrisponde ad una coppia: azione (a), stato del mondo (g); l'azione permette di ottenere lo stato del mondo. Non sempre nella delega di un compito vengono esplicitate entrambe queste variabili. È possibile che il trustor deleghi direttamente lo stato del mondo g al trustee e il trustee poi decida come ottenere quello stato del mondo.

Data questa formulazione, $Trust(X\ Y\ \tau\ C)$, possiamo tradurla quindi sostenendo che l'agente (cognitivo) X che ha la necessità di ottenere un certo scopo, ossia una certa situazione nel mondo (lo stato g), delega all'agente (non necessariamente cognitivo) Y il compito τ. Ossia gli delega la realizzazione di una certa azione (a), nel contesto C, per fare in modo che si avveri lo stato del mondo g che è il suo scopo.

Quindi se ne deduce che affinché si abbia l'attitudine a fidare è indispensabile che il fidante (trustor) abbia degli scopi da perseguire. Di conseguenza, l'agente X deve essere un agente cognitivo. Questa constatazione porta con sé alcune conseguenze rilevanti. Per esempio, che per perseguire uno scopo è necessario fidarsi di qualcuno/qualcosa, al limite di se stessi. Ma anche che il possesso (creduto da un trustor) di specifiche caratteristiche da parte di un trustee può attivare un potenziale scopo e la relativa relazione di fiducia (*generare un nuovo scopo*).

Si può inoltre pensare a *forme generalizzate di fiducia*, dove X può fidare Y per una certa tipologia di scopi o addirittura per "qualunque" scopo; o ancora, fidare un insieme di agenti per un dato scopo o per una famiglia di scopi.

Ci sono inoltre i cosiddetti *fenomeni di fiducia collettiva* che sono relati agli scopi (bisogni, aspirazioni) delle persone coinvolte.

Abbiamo quindi visto come non possa esserci fiducia senza uno scopo da ottenere, ma le credenze (*beliefs*) del trustor rappresentano le basi principali su cui la fiducia è fondata. Queste credenze riguardano principalmente il *trustee* ma non solo, per esempio riguardano anche il contesto in cui il trustee opererà. Vediamole nel dettaglio:

- una prima classe di credenze è rivolta alle *competenze* di Y; in particolare a quelle che sarebbero necessarie per il *task* o classe di task che X intende delegargli. Queste competenze generali si articolano in varie specializzazioni: ci sono le *abilità* vere e proprie, ossia le capacità fisiche che un agente è in grado di esibire, ma anche il *know-how*, ossia le conoscenze

utili per esercitare quelle abilità al meglio; e ancora la *self-confidence*, ossia la consapevolezza di quelle abilità (e così via).

- una seconda classe di credenze è rivolta alle *intenzioni* di *Y*; in particolare a quelle che sarebbero necessarie per il *task* o classe di task che *X* intende delegargli. Queste intenzioni si articolano in due differenti sottoclassi. Da una parte ci sono le attitudini intenzionali verso quel compito da parte di *Y*, indipendentemente da chi glielo ha delegato: la capacità di persistenza, motivazione, etc. impliciti nella natura di *Y* verso quel task o classe di task. D'altra parte, ci sono le attitudini intenzionali, attribuibili a *Y* sulla base della combinazione di quel task con la conoscenza di chi ha delegato il compito: benevolenza, non pericolosità, sicurezza, etc.
- una terza classe di credenze è rivolta alle conoscenze sul *contesto* in cui *Y* dovrà operare per realizzare il *task* delegato da *X*; ci sono infatti varie possibilità di condizioni favorevoli o di ostacolo alla realizzazione del compito. E l'azione di *Y* risulterà ovviamente influenzata da queste condizioni.
- un'ultima classe di credenze riguarda la *dipendenza* di *X* da *Y* per la realizzazione del task. In realtà questa dipendenza può essere di due tipi: *X* non è in grado di realizzare quel compito se non può delegarlo ad *Y* (*dipendenza forte*); oppure, per *X* è meglio delegare a *Y* piuttosto che svolgere da solo quel compito che pure sarebbe in grado di fare (*dipendenza debole*).

Ovviamente, quanto più è precisa la conoscenza di queste competenze, intenzionalità, contesti e dipendenze da parte di *X*, tanto più adeguata sarà l'aspettativa sui risultati delle azioni di *Y*.

Possiamo riassumere il modello mentale del trustor nella predisposizione a fidare attraverso la Figura1.

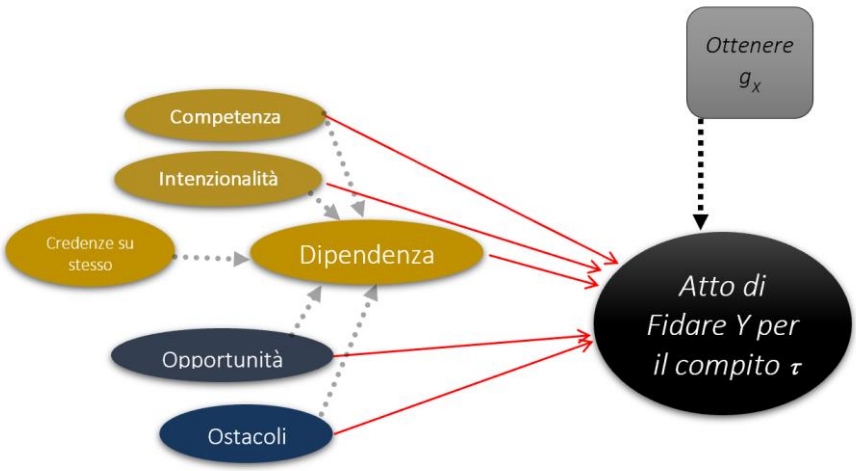

Figura 1

Il modello illustrato assai sinteticamente nelle pagine precedenti, rappresenta la base del modello socio-cognitivo della fiducia (Falcone e Castelfranchi, 2001a, Falcone e Castelfranchi, 2001b).

A partire dallo schema mentale necessario per sviluppare l'attitudine di fiducia è possibile analizzare le molteplici caratteristiche ed effetti che derivano da questo concetto di base. Per esempio, quali sono le dinamiche della fiducia? Come si relaziona con l'ordine sociale ed in particolare con le norme e le autorità; Su quali tipologie di sorgenti si basa la fiducia? Come si misura la fiducia? Quale è il suo rapporto con l'autonomia e con il controllo? Come può essere sfruttata così da rappresentare un capitale utilizzabile in una rete sociale?

Rispondiamo, nel seguito, solo ad alcuni tra i più interessanti di questi interrogativi.

3.3 *Le fonti di fiducia*

Come abbiamo visto i comportamenti degli agenti, in particolare se essi si fidano o meno di altri agenti, dipendono da cosa essi credono su questi ultimi: ossia dalle loro beliefs. Queste credenze non sempre hanno la stessa validità. Alcune possono essere più convinte, altre meno. Ma da cosa deriva questa convinzione, ossia la forza di quelle credenze? Essa deriva dalla fonte (o le fonti, se ce n'è più di una) che ha (hanno) permesso di generarle.

È quindi importante concentrarsi sulla natura di queste fonti, per comprendere appieno come le credenze si generano e si modificano nel tempo. Una prima tipologia di fonte, la principale, è l'*esperienza diretta*: l'agente acquisisce la specifica credenza attraverso la percezione diretta (dei propri sensi ma anche di alcune proprie capacità cognitive, tipo la memoria) del fenomeno che genera la credenza.

Una seconda tipologia di fonte è la comunicazione da parte di altri agenti, del fenomeno che genera la credenza. In questo caso si parla di *esperienza indiretta*, in quanto il fenomeno viene mediato da un altro agente che può aver esperito direttamente lui il fenomeno che genera la credenza o addirittura aver a sua volta ricevuto comunicazione indiretta del fenomeno.

Una terza tipologia di fonte è il ricorso da parte dell'agente primario (colui che si costruisce la credenza) a sue capacità cognitive particolarmente sofisticate come il *ragionamento e l'inferenza*. Attraverso queste capacità l'agente genera nuove credenze che possono derivare da esperienza diretta e indiretta ed essere elaborate. Esempi sono il ragionamento che permette la categorizzazione di elementi del mondo, piuttosto che il ragionamento per analogia.

Per ogni specifica credenza (*belief*), indipendentemente dalla natura della fonte che l'ha generata, è utile indicare alcune caratteristiche fondamentali del rapporto credenza-fonte:

- *identificazione della fonte*: è in grado l'agente che possiede la specifica credenza, di ricondurre quella credenza alla fonte che l'ha generata?

- *valore di certezza della fonte sul contenuto trasmesso*: la fonte che genera (o contribuisce a generare) la credenza ha fornito (o può essere dedotto) un valore di certezza sul contenuto?
- *fiducia verso la fonte*: considerata la fonte che potenzialmente può determinare la credenza di un certo agente, quale è la fiducia dell'agente verso quella fonte? E quanto è determinante?

Come si vede c'è un interessante elemento di ricorsione nell'analizzare l'attitudine di fiducia. Per fidarsi infatti è necessario ricorrere a delle credenze, ma queste a loro volta hanno la necessità di essere considerate affidabili e quindi scatenare, da parte del trustor, un processo di fiducia all'indietro, verso le fonti (e poi le fonti delle fonti, etc.).

3.4 *Dinamica della fiducia*

Un esempio di dinamica della fiducia che dà conto della necessità di costruire un modello cognitivamente ricco e rispondente al profilo del trustee, viene dall'analisi del processo di fiducia, quando lo si confronta rispetto ad un modello semplificato.

Se supponiamo di avere il seguente schema (Figura 2):

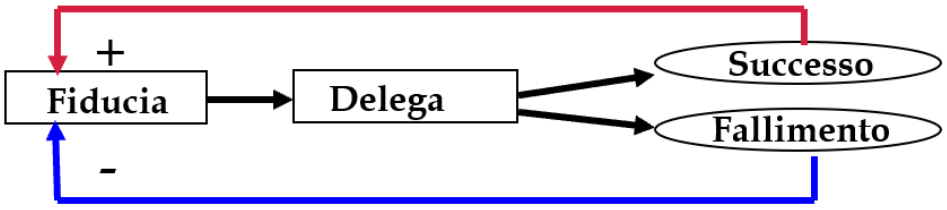

Figura 2

Potremmo dire che ogni volta che grazie ad un livello di fiducia sufficiente si passa a delegare un compito e a valle di questo il *trustee* ottiene il risultato atteso (successo), la fiducia nel trustee si conferma o addirittura aumenta (linea rossa positiva di ritorno sulla fiducia). Insomma, il feedback è positivo e rafforza quel potenziale comportamento. Quando invece si ha un fallimento da parte del *trustee*, la fiducia del *trustor* in esso declina (linea azzurra negativa di ritorno sulla fiducia). Uno schema eccessivamente semplificato.

Se introduciamo un modello cognitivo del *trustee*, lo schema cambia (Figura 3).

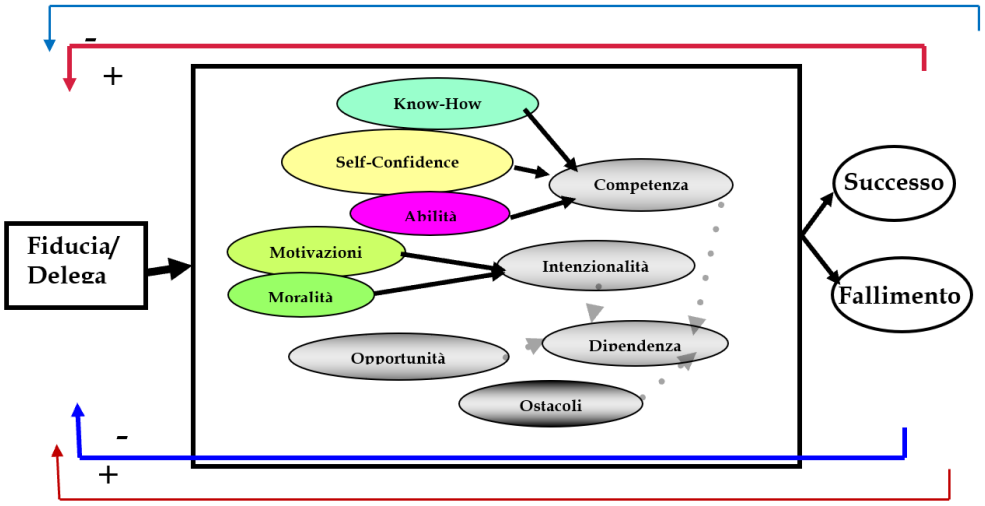

Figura 3

In questo caso, come è riportato nella figura 3, sono associate tanto al successo quanto al fallimento del compito delegato, oltre ai feedback coerenti con il risultato (quelli già visti nella figura 2), anche feedback di segno opposto (in generale di peso inferiore). Questi stanno a rappresentare il fatto che nella valutazione della performance del *trustee* da parte del *trustor* è possibile, grazie alla valutazione dei vari fattori che entrano in gioco e che sono consapevolmente considerati ed eventualmente valutati dal *trustor*, anche elementi contribuenti in segno contrario al risultato ottenuto. Elementi che devono essere presi in considerazione per raffinare il modello del *trustee* e per adeguare la fiducia nei suoi confronti. Per fare un esempio, è possibile che la delega di un task al *trustee* veda una sua performance che realizza il task ma solo in quanto particolari condizioni ambientali (non ordinarie) lo hanno favorito. Ed anzi nella performance si potrebbero rilevare deficit sulle sue competenze e/o sulle intenzionalità.

Insomma, una *teoria attribuzionale*, legata ad un modello cognitivo più sofisticato dell'attitudine a fidare, permette di andare oltre alla sola valutazione del risultato dell'atto di fiducia.

Un modello di fiducia analitico e articolato come quello sviluppato in (Falcone e Castelfranchi, 2001a; Castelfranchi e Falcone, 2010) può quindi essere implementato in sistemi artificiali intelligenti (Figura 4) così che l'interazione di questi sistemi sia più verosimile e adeguata.

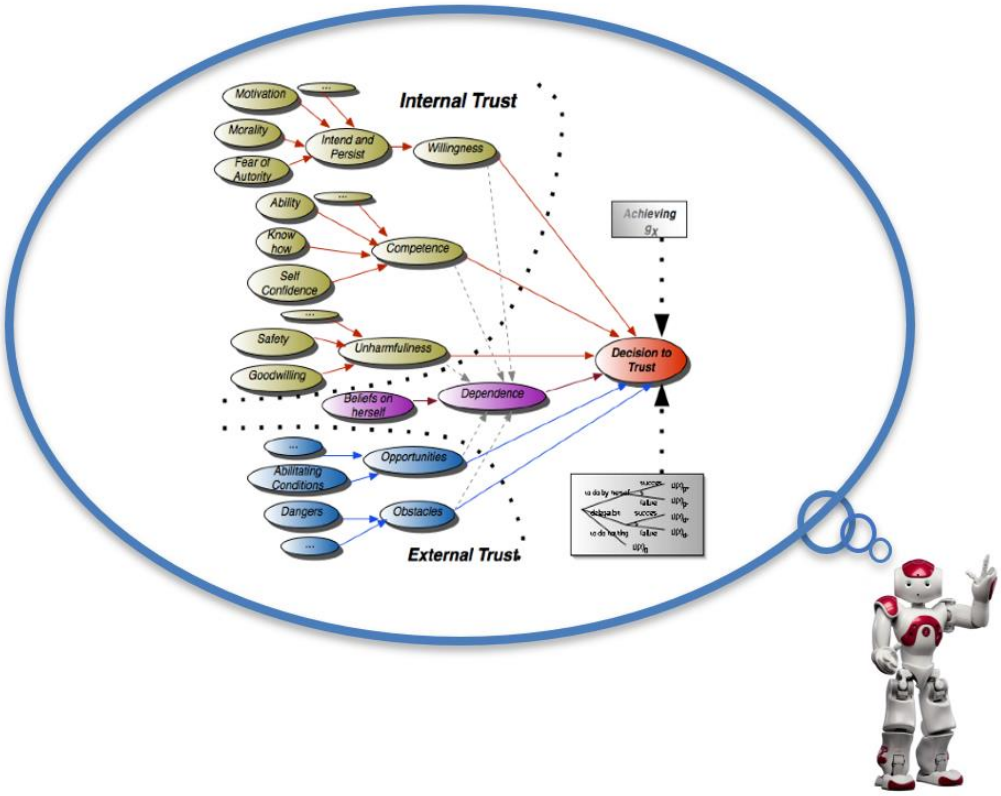

Figura 4

4.0 *Conclusioni*

In questo lavoro ci siamo concentrati su una delle attitudini cognitive più interessanti: la fiducia. Come sostiene Locke (Locke, 1689) essa rappresenta il collante nella società degli umani. Per fare in modo che i sistemi artificiali intelligenti possano penetrare in modo efficace le nostre società, essi dovranno essere soggetti attivi e passivi rispetto a questa attitudine che risulta funzionale ed efficiente in ambito interazionale. Abbiamo per questo mostrato come è possibile sviluppare un modello concettuale operazionale e quindi potenzialmente implementabile di fiducia. Abbiamo anche accennato alla complessità della piena modellazione di questa attitudine e di conseguenza a riflettere sull'enorme lavoro necessario per portare i sistemi artificiali intelligenti in grado di utilizzarla così come fanno gli umani in modo del tutto naturale e nella maggior parte dei casi quasi inconsapevolmente.

Le questioni in ambito AI in cui la cognizione può giocare un ruolo rilevante sono molte altre. Ne vogliamo accennare solo alcune significativamente presenti nell'attuale dibattito scientifico:

1) *Problema dell'apprendimento dai Big Data*. Nell'ambito della disciplina "Intelligenza Artificiale" sta crescendo prepotentemente il ruolo del *machine learning*,

tanto che spesso si confondono i due termini così da assimilare l'IA ai soli sistemi artificiali di apprendimento (e ancora più specificamente alla sottocategoria del *deep learning*). La forza e la diffusione di queste tecniche di apprendimento è che permettono di individuare regolarità nei dati a disposizione caratterizzanti un qualche sistema/comportamento/fenomeno; regolarità altrimenti di difficile identificazione. Esse offrono in particolare, una volta identificate queste regolarità, risultati straordinari nella possibilità di fare previsioni future rispetto ai fenomeni derivanti da quei dati.

Esistono però serie questioni che sempre più si evidenziano (riproposizioni di bias presenti nei dati di partenza, difficoltà ad individuare comportamenti generativi, etc.) e che fanno ritenere che sia necessario integrare queste tecniche con approcci basati su conoscenza top-down più strutturata.

2) *La relazione mente corpo*. Esistono teorie che indicano come *i processi cognitivi ed emotivi si fondino su meccanismi sensorimotori ed interocettivi e siano quindi strettamente condizionati dal corpo*. La robotica umanoide rappresenta un mezzo potente, forse il più efficace che possediamo oggi, per verificare se ed in che misura il corpo, e le differenze nelle sue tipologie, possa influenzare i processi cognitivi. Per esempio, alcuni studi sulla comparazione di come si generano nell'uomo i concetti concreti e quelli astratti, evidenziano differenze nei modelli generativi delle due tipologie dovute al fatto che per i concetti astratti oltre all'esperienza sensorimotoria gioca un ruolo anche l'esperienza linguistica, interocettiva e metacognitiva (contribuendo così anche a spiegare perché i concetti astratti impiegano più tempo ad essere processati, oltre a chiarire come ce li rappresentiamo). Ovviamente comprendere come funziona l'essere umano non necessariamente deve vincolarci a definire modelli perfettamente aderenti e doverci necessariamente riferire, per proporre sistemi artificiali intelligenti, alla *embodied e grounded cognition*.

Resta però un problema con cui fare i conti (ben spiegato in Castelfranchi, 2018) e che riassumiamo di seguito: possiamo anche costruire agenti artificiali che hanno *intenzioni* (e anche *obiettivi*, *progetti*, *preferenze*). Ma non agenti che "sentono", "provano" qualcosa, come succede con i veri *desideri* e *bisogni soggettivi*, che sono un *tipo* di scopi che si "provano" (le *intenzioni* non si "provano"). Per questi è necessario avere un "corpo" e non semplicemente un hardware. Serve enterocezione e propriocezione del proprio sistema fisico che scambia energia col mondo (percepisce e agisce) ma che ha anche eventi "interni" da percepire. Solo così si potranno avere *vere emozioni* e avere l'equivalente del *dolore* o del *piacere* (che non significano semplice frustrazione o raggiungimento dello scopo (Castelfranchi, 2018)).

3) *Problema delle decisioni autonome e della coscienza*. Queste macchine artificiali intelligenti saranno sempre più autonome nelle loro decisioni. Sarà quindi fondamentale stabilire come questo livello di autonomia sia realmente orientato agli scopi dell'utente o comunque trasparente nel caso in cui serva a tutelare scopi di altro genere/livello.

Questa considerazione ci porta a problemi di ordine teorico-filosofico: un sistema intelligente artificiale comprende realmente ciò che fa? E questa comprensione come si distingue da quella naturale? Può avere coscienza? E cosa si intende per coscienza: processi attentivi, meta-cognizione, senso del sè, esperienze fenomeniche, etc. Aspetti legati alla coscienza nei robot sono molteplici, come ad esempio, la necessità di avere o meno un corpo, la necessità di essere inserito e parte di un ambiente, ossia "situato," la necessità di provare emozioni.

Ci piace infine sottolineare un ultimo punto. Per comprendere come si svilupperà l'interazione con i nuovi sistemi intelligenti e come questa interazione cambierà la nostra cognizione e comportamento serviranno approfondimenti da varie discipline: tecnologiche, psicologiche, filosofiche, sociali, pedagogiche (scienze cognitive in senso lato).

Per fare qualche esempio di studi che hanno avviato indagini approfondite in questa direzione si può citare (Pezzulo, 2018):

- *riguardo ai meccanismi che permettono di cooperare interattivamente per perseguire scopi comuni* (Gallotti and Frith, 2013; Knoblich and Sebanz, 2008; Pezzulo et al., 2017; Sebanz et al., 2006);
- *riguardo al comprendere le intenzioni altrui ed aiutare gli altri a comprendere le proprie* (Donnarumma et al., 2017; Misyak et al., 2016; Yoshida et al., 2008)
- *riguardo all'apprendimento interattivo* (Csibra and Gergely, 2007; Dindo et al., 2015; Gopnik et al., 2004).
- Infine, come abbiamo visto*, riguardo al negoziare spazi di autonomia e fidarci l'uno dell'altro* (Castelfranchi and Falcone, 2010).

Bibliografia

Castelfranchi, C., Falcone, R., 2010. *Trust Theory: A socio-cognitive and computational model*. Wiley.

Castelfranchi, C., 2018. Cambieranno anche i concetti della psicologia e del senso comune. *Giornale Italiano di Psicologia*, 95-98.

Csibra, G., Gergely, G., 2007. "Obsessed with goals": Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica* 124, 60-78.

Dindo, H., Donnarumma, F., Chersi, F., Pezzulo, G., 2015. The intentional stance as structure learning: a computational perspective on mindreading. *Biological cybernetics* 109, 453-467.

Donnarumma, F., Dindo, H., Pezzulo, G., 2017. Sensorimotor communication for humans and robots: improving interactive skills by sending coordination signals. *IEEE Transactions on Cognitive and Developmental Systems*, 11.

Falcone, R., Capirci, O., Lucidi, F., Zoccolotti, P., 2018. Prospettive di intelligenza artificiale: mente, lavoro e società nel mondo del machine learning. *Giornale Italiano di Psicologia*, 43-68.

Falcone, R., Castelfranchi, C., 2001. Social Trust: A Cognitive Approach, in *Trust and Deception in Virtual Societies*, ed. by Castelfranchi C. and Yao-Hua Tan, Kluwer Academic Publishers, 55-90.

Falcone, R., Castelfranchi, C., 2001, The Human in the Loop of a Delegated Agent: The Theory of Adjustable Social Autonomy, *IEEE Transactions on Systems, Man, and Cybernetics*, Part A: Systems and Humans, Special Issue on "Socially Intelligent Agents - the Human in the Loop", 31, 406-418.

Gallotti, M., Frith, C.D., 2013. Social cognition in the we-mode. *Trends Cogn. Sci.* (Regul. Ed.) 17, 160-165. <https://doi.org/10.1016/j.tics.2013.02.002>.

Gambetta, D., 1988. "Can we trust trust?". In *Trust: Making and Breaking Cooperative Relations*, ed. by D. Gambetta. Oxford Blackwell.

Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., Danks, D., 2004. A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* 111, 3-32.

Hardin R., 2002. *Trust and Trustworthiness*, New York: Russel Sage Foundation.

Knoblich, G., Sebanz, N., 2008. Evolving intentions for social interaction: from entrainment to joint action. *Philos Trans R Soc Lond B Biol Sci* 363, 2021-2031.

Jones, K., 1996. Trust as an Affective Attitude, *Ethics* 107, 4-25.

Locke, J., 1689. *Letter Concerning Toleration*, London: Printed for Awnsham Churchill.

Luhmann N., 1979. *Trust and Power*, Wiley, New York.

Misyak, J., Noguchi, T., Chater, N., 2016. Instantaneous Conventions: The Emergence of Flexible Communicative Signals. *Psychological Science* 27, 1550-1561.

Pezzulo, G., 2018. Noi e le tecnologie intelligenti: cosa c'è di speciale? *Giornale Italiano di Psicologia*, 137-140.

Pezzulo, G., Iodice, P., Donnarumma, F., Dindo, H., Knoblich, G., 2017. Avoiding accidents at the champagne reception: A study of joint lifting and balancing. *Psychological Science*, 28. < https://doi.org/10.1177/0956797616683015>

Sebanz, N., Bekkering, H., Knoblich, G., 2006. Joint action: bodies and minds moving together. *Trends Cogn Sci* 10, 70-76. <https://doi.org/10.1016/j.tics.2005.12.009>.

Yoshida, W., Dolan, R.J., Friston, K.J., 2008. Game Theory of Mind. *PLoS Comput Biol 4*, e1000254+.

Zuboff, S., 2019, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

Rino Falcone
Istituto di Scienze e Tecnologie della Cognizione – CNR, Roma
rino.falcone@istc.cnr.it

– Cognizione e Sistemi Intelligenti: il ruolo della Fiducia

Citation standard:
FALCONE, Rino. Cognizione e Sistemi Intelligenti: il ruolo della Fiducia. Laboratorio dell'ISPF. 2022, vol. XIX [5]. DOI: 10.12862/Lab22FLR.

Online: 31.12.2022

ABSTRACT
Cognition and Intelligent Systems: the role of Trust. Artificial intelligence (AI) is becoming an enabling technology (KET, Key Enable Technology), i.e. one of those technological approaches/tools (like, for example, robotics, cybersecurity, the cloud, nanotechnologies, and so on) that play a fundamental role in the processes linked to digital transformation and which can therefore be considered promoters of innovation and of the structural evolution of society. As a pervasive and increasingly widespread technology in the ordinary lives of humans, AI faces a further challenge: being able to relate deeply to human cognition, adequately corresponding to some priority and specific attitudes of cognition itself. In this short article we will show the example of a particularly specific and intimate attitude of human: trust.

KEYWORDS
Artificial Intelligence; Cognitive Modelling; Trust; Trustworthiness

SOMMARIO
L'intelligenza artificiale (IA) sta diventando una tecnologia abilitante (KET, Key Enable Technology), ossia uno di quegli approcci/strumenti tecnologici (al pari ad esempio della robotica, della cybersecurity, del cloud, delle nanotecnologie, e così via) che svolgono un ruolo fondamentale nei processi legati alla trasformazione digitale e che possono essere considerati quindi promotori dell'innovazione e dell'evoluzione strutturale della società. In quanto tecnologia pervasiva e diffusa sempre più intensamente nelle vite ordinarie degli umani, l'AI deve affrontare una sfida ulteriore: essere in grado di mettersi profondamente in relazione con la cognizione umana, corrispondendo adeguatamente ad alcune attitudini prioritarie e specifiche della cognizione stessa. In questo breve articolo mostreremo l'esempio di una attitudine particolarmente precipua e intima dell'uomo: la fiducia.

PAROLE CHIAVE
Intelligenza Artificiale; Modellamento Cognitivo; Fiducia; Affidabilità