

Modelli Cognitivi della Fiducia: risvolti teorici e applicazioni pratiche

Alessandro Sapienza, Rino Falcone and Cristiano Castelfranchi

Institute of Cognitive Sciences and Technologies, ISTC-CNR, Rome, Italy, 00185

alessandro.sapienza@istc.cnr.it, rino.falcone@istc.cnr.it, cristiano.castelfranchi@istc.cnr.it

Abstract

Il concetto di fiducia all'interno dell'AI ha un ruolo sempre più importante. La presenza di agenti artificiali in grado di compiere azioni sempre più complesse e in maniera sempre più autonoma ha comportato l'evoluzione dei modelli computazionali da forme più semplici a forme più elaborate di fiducia. Questo perché, al crescere della complessità dell'interazione, cresce conseguentemente anche l'importanza della fiducia. All'interno di questo contributo, partendo dal modello socio-cognitivo della fiducia, presentiamo 4 esemplificazioni di adattamento in differenti scenari.

1 Introduzione

La modellazione della fiducia [Castelfranchi e Falcone, 2010] è di fondamentale importanza nel campo dell'AI. Negli ultimi anni, l'applicazione di questo concetto è incrementata esponenzialmente grazie anche all'introduzione del trustworthy AI. Sebbene questo specifico settore stia coinvolgendo particolarmente l'attenzione della comunità scientifica, non è di certo l'unico a richiedere l'utilizzo della fiducia. La fiducia è infatti un concetto chiave per l'interazione sociale e non, umana o artificiale. Al crescere della complessità dell'interazione, cresce conseguentemente anche l'importanza della fiducia. Ce ne rendiamo conto osservandone l'evoluzione in molte aree tematiche. In alcuni casi (ad esempio i MAS) la fiducia è stata introdotta come un'esigenza teorica di base; in altri, come nel caso dell'IoT, questa è emersa come necessità strutturale: in scenari in cui bisogna gestire l'interazione di centinaia, migliaia o forse più dispositivi eterogenei per hardware, produttore e, soprattutto, scopi, l'introduzione del concetto di fiducia risulta una necessità. Non solo: più complessa è l'interazione che si intende simulare/studiare, più è necessario tener conto dei vari aspetti che un sofisticato modello teorico della fiducia mette a disposizione. Se, nelle forme più intuitive e semplici, la fiducia viene modellata come "misura della performance" o "probabilità di successo", quando si considerano interazioni più articolate, come quelle tra entità intelligenti in grado di compiere azioni autonome complesse, ci si accorge inevitabilmente che questo tipo di modellazione non è sufficiente. Risulta necessario fare riferimento a modelli di tipo cognitivo, in grado di spiegare ad esempio perché

un agente ha compiuto una data azione, cosa voleva ottenere, se il risultato ottenuto è dovuto a lui o a fattori esterni, se ha inteso bene i nostri scopi, in base a quali principi ha operato ecc.

All'interno di questo contributo, prendiamo in considerazione 4 esemplificazioni che abbiamo realizzato a partire dal modello socio-cognitivo della fiducia [Castelfranchi e Falcone, 2010]. Questi adattamenti ci hanno permesso di valutare tanto aspetti di ordine teorico che di natura applicativa.

I modelli sviluppati sono stati poi testati attraverso simulazioni ad agenti. Le simulazioni realizzate ci hanno permesso non solo di verificare e spiegare il comportamento teorizzato, ma anche di indagare possibili evoluzioni e ulteriori scenari ipotetici.

2 Ricadute Teoriche

2.1 La generalizzazione della conoscenza all'interno di reti sociali virtuali

All'interno di un primo lavoro [Sapienza e Falcone, 2020], ci siamo occupati di modellare la fiducia all'interno di reti sociali virtuali formate da agenti cognitivi umani o artificiali. All'interno di questo scenario, abbiamo investigato il concetto di fiducia in mancanza di esperienza diretta. Infatti, in questa tipologia di reti, l'esperienza diretta potrebbe non essere disponibile nella maggior parte dei casi, data l'elevata dinamicità della rete e l'elevato numero di interconnessioni. E' quindi necessario fare riferimento a informazione non diretta: le raccomandazioni. Nello specifico contributo non ci limitiamo alle raccomandazioni di individui, ma prendiamo in considerazione il confronto con le raccomandazioni di categorie di individui. Infatti, tramite processi inferenziali, è possibile identificare caratteristiche comuni tra i singoli soggetti, stimare che queste caratteristiche comportano performance simili sullo stesso task o su task simili. Dunque, ci siamo focalizzati sul confrontare le raccomandazioni individuali con le raccomandazioni di intere categorie. Definiamo il *category value CV* come la percentuale di volte il cui le raccomandazioni di categoria forniscono un risultato più preciso rispetto alle raccomandazioni individuali. Le simulazioni agent-based ci hanno permesso di ottenere i seguenti risultati (figura 1):

- il *CV* aumenta all'aumentare del turnover nella rete, dunque in effetti la dinamicità della rete influenza positivamente il valore delle categorie;

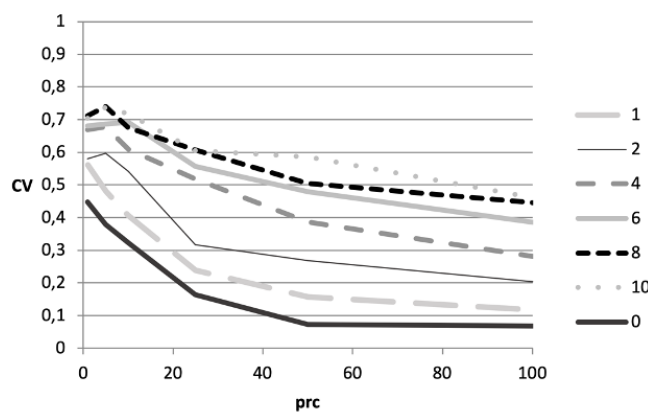


Figura 1: Effetto del turnover sul CV

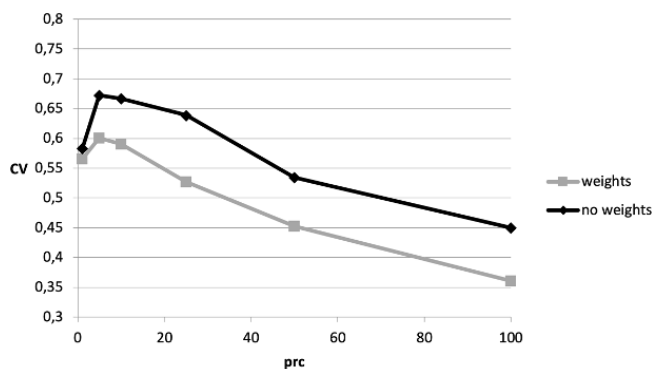


Figura 2: CV quando si conosce o meno l'affidabilità dei raccomandatori

- il *CV* aumenta in particolar modo quando non si conosce l'affidabilità dei raccomandatori.

Sebbene l'esperienza diretta sia di chiara importanza, non sempre questa è disponibile per la valutazione dei potenziali partner. In tal senso, possedere dei meccanismi di generalizzazione della conoscenza rappresenta un valore aggiunto. In questo lavoro, abbiamo effettivamente dimostrato che, in alcune specifiche condizioni le raccomandazioni di categoria hanno performance equiparabili o persino superiori a quelle sugli individui.

2.2 Fiducia, autonomia e controllo: un modello user-centered per l'interazione coi sistemi IOT

In questo secondo lavoro [Falcone e Sapienza, 2018], sosteniamo l'efficacia della fiducia come strumento per affrontare l'interazione degli utenti coi sistemi IoT. In un contesto in cui gli utenti si trovano a dover fare i conti con l'autonomia di sistemi/oggetti che solo parzialmente sono in grado di controllare, il concetto di fiducia entra in gioco spontaneamente. È necessario tenere presente quanto l'utente è disposto ad accettare un sistema IoT e in che misura vuole interfacciarsi con esso. Su quest'ultimo punto vorremmo soffermarci, il concetto di "accettazione da parte dell'utente"

Mentre la letteratura si è focalizzata molto sugli aspetti di sicurezza e privacy, minor attenzione è stata dedicata al punto

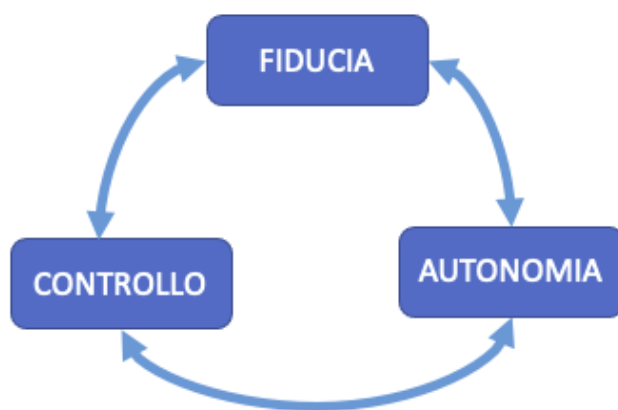


Figura 3: Rapporto tra fiducia, autonomia e controllo

di vista dell'utente e alla sua interazione con i sistemi IoT. Si tratta in realtà di un argomento chiave, poiché anche la tecnologia più sofisticata deve essere accettata dagli utenti, altrimenti semplicemente non verrà utilizzata.

In questo lavoro, proponiamo un modello cognitivo della fiducia basato sui concetti di fiducia e controllo (figura 3), con particolare riferimento al feedback. Da un lato, l'utente si interfaccia con i dispositivi concedendogli diversi livelli di autonomia nell'eseguire i task. Nell'interazione, risulta fondamentale l'esercizio di strumenti di controllo, esercitato tramite richieste di feedback sull'operato. Il controllo, in questo caso, non agisce diminuendo la fiducia. Al contrario, permette di superare situazioni in cui la fiducia non sarebbe sufficiente e diventa uno strumento di promozione dell'autonomia dei dispositivi. Dall'altro lato, i device modellano le credenze dell'utente sulla loro affidabilità, in termini di autonomia concessa ed effettive capacità nell'eseguire i compiti. Tale modellazione, permette ai dispositivi di identificare i loro limiti di azione ma gli dà anche la possibilità di capire se e quanto è possibile provare a ottenere maggiore autonomia. Proponiamo quindi un approccio ai sistemi IoT che permetta a tali sistemi di adattarsi allo specifico utente, in base alla disposizione di quest'ultimo verso tale tecnologia, con l'obiettivo di (1) identificare i limiti di accettazione dello specifico utente e (2) incoraggiare tale utente ad accettare sempre più questa tecnologia, stimolandone la fiducia. Dunque, tramite simulazione ad agenti, abbiamo analizzato i fattori che influenzano il sistema, cercando di capire quali effetti hanno e se rappresentano un vincolo per l'autonomia:

- *efficienza*: questa ha un effetto molto forte, tale che in presenza di un alto tasso di errore alcuni task potrebbero non essere più eseguiti.
- *profilo utente*: un profilo più aperto consente di ridurre l'impatto che la dimensione dell'efficienza ha sull'autonomia.
- *efficacia dei feedback*: tale fattore influenza la velocità con cui vengono raggiunti i corrispondenti valori finali di fiducia.

3 Ricadute Pratiche

Nei seguenti studi non ci occupiamo di modellare entità virtuali. Piuttosto, analizziamo il comportamento di individui umani e lo riproduciamo artificialmente, simulandone il cambiamento al variare degli input esterni.

3.1 Impatto del Covid-19 sul rapporto dei cittadini con l'informazione

In questo studio [Falcone e Sapienza, 2020], ci siamo occupati di investigare l'information seeking behavior (ISB) dei cittadini italiani durante le prime fasi della pandemia da Covid-19, al fine di identificare sia il rapporto tra cittadini e informazione, ma anche se e come questo abbia influenzato l'opinione pubblica.

Nello specifico, a partire da un dataset di soggetti ($N=4260$), abbiamo individuato l'ISB dei cittadini in termini di frequenza d'utilizzo delle fonti d'informazione e affidabilità attribuitagli. L'analisi dell'ISB è stata condotta per età e genere, data la forte correlazione tra questi attributi e la mortalità da Covid-19. Successivamente, abbiamo simulato l'evoluzione della loro opinione sull'importanza delle misure restrittive tramite un modello di opinion dynamics basato su [Hegselmann *et al.*, 2002]. Tramite simulazione ad agenti, siamo in grado di identificare l'opinione media complessiva della popolazione e di singole classi sociali, oltre a quanto queste deviano da quelle ottimali. In aggiunta, abbiamo confrontato i nostri dati sull'ISB con quelli estratti da [Zucco *et al.*, 2018], relativamente all'ISB in ambito sanitario della popolazione italiana. Tale confronto ci ha permesso di riscontrare effettivamente un utilizzo virtuoso delle fonti d'informazione. Ciò ha avuto un'effetto positivo sull'opinione della popolazione. A sua volta, l'utilizzo virtuoso delle fonti d'informazione potrebbe aver favorito l'accettazione delle misure di contrasto necessarie per fronteggiare il Covid-19.

3.2 Accettazione delle restrizioni dovute al Covid-19

Anche quest'ultimo lavoro [Falcone e Sapienza, 2021] è relativo al contesto Covid-19. In questo caso, si è deciso di investigare l'effetto della fiducia nelle Istituzioni sull'accettazione delle restrizioni imposte per fronteggiare la pandemia. Applicando la regressione lineare sul campione raccolto ($N=4260$), si è modellata la fiducia dei cittadini Italiani a partire dalle principali componenti cognitive, facendo particolare riferimento agli aspetti di competenza e motivazione, ma anche a dimensioni relative agli scopi e alla loro importanza soggettiva, o alle aspettative sul comportamento altrui. Dunque, sempre tramite regressione lineare, si è modellato il legame tra la fiducia nelle istituzioni e l'effettiva accettazione delle restrizioni imposte. Quindi, tramite modellazione ad agenti, abbiamo investigato le modalità con cui l'evoluzione della fiducia istituzionale influenza l'accettazione delle misure restrittive.

I risultati (figure 4 e 5) offrono notevoli spunti di riflessione. In primo luogo, confermano il ruolo chiave delle dimensioni cognitive di competenza e motivazione in quanto componenti della fiducia. In aggiunta, l'analisi suggerisce possibili spunti di ottimizzazione. Emerge infatti che, ai fini

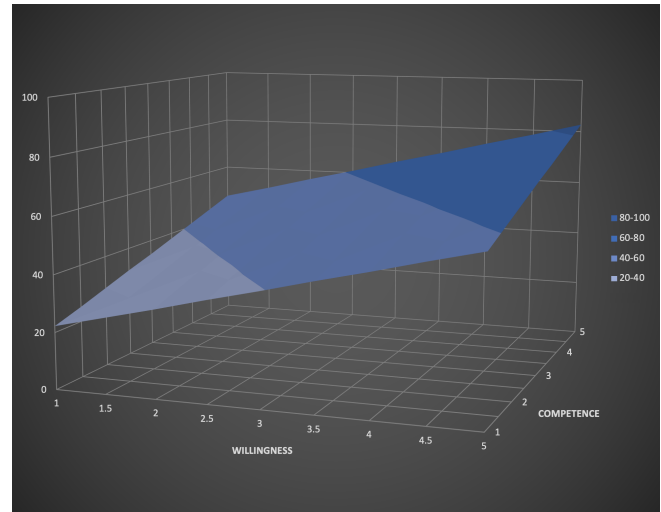


Figura 4: Evoluzione della fiducia media nelle Istituzioni, al variare di competenza e motivazione

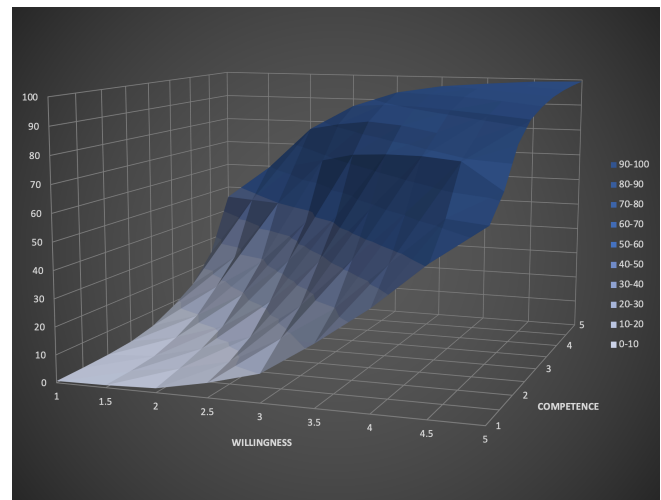


Figura 5: Variazione dell'acceptance rate della popolazione, al variare di competenza e motivazione

dell'ottimizzazione della risposta comunitaria (acceptance rate), un generico tentativo di migliorare la fiducia media, oltre che essere molto dispendioso, può non rappresentare la strategia migliore. Di contro, è possibile individuare delle specifiche categorie sociali di cittadini e agire specificatamente su di esse, andando a sollecitare le particolari componenti che ottimizzano maggiormente la loro fiducia istituzionale.

4 Conclusioni

All'interno di questo contributo, sono stati introdotti 4 lavori rappresentanti lo stato dell'arte della modellazione cognitiva della fiducia. I modelli si ripropongono non solo di stimare la fiducia, ma anche di indagarne e modellarne le ragioni. Dunque, attraverso un'implementazione simulativa, si è mostrato la potenzialità di questi modelli e i risultati che si possono ottenere.

Riferimenti bibliografici

- [Castelfranchi e Falcone, 2010] Christiano Castelfranchi e Rino Falcone. *Trust theory: A socio-cognitive and computational model*, volume 18. John Wiley & Sons, 2010.
- [Falcone e Sapienza, 2018] Rino Falcone e Alessandro Sapienza. On the users' acceptance of iot systems: A theoretical approach. *Information*, 9(3):53, 2018.
- [Falcone e Sapienza, 2020] Rino Falcone e Alessandro Sapienza. How covid-19 changed the information needs of italian citizens. *International Journal of Environmental Research and Public Health*, 17(19):6988, 2020.
- [Falcone e Sapienza, 2021] Rino Falcone e Alessandro Sapienza. An agent-based model to assess citizens' acceptance of covid-19 restrictions. *Journal of Simulation*, pages 1–15, 2021.
- [Hegselmann *et al.*, 2002] Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.
- [Sapienza e Falcone, 2020] Alessandro Sapienza e Rino Falcone. Evaluating agents' trustworthiness within virtual societies in case of no direct experience. *Cognitive Systems Research*, 64:164–173, 2020.
- [Zucco *et al.*, 2018] Rossella Zucco, Francesco Lavano, Rosa Anfosso, Aida Bianco, Claudia Pileggi, e Maria Pavia. Internet and social media use for antibiotic-related information seeking: Findings from a survey among adult population in italy. *International journal of medical informatics*, 111:131–139, 2018.