**Abstract**

Nel loro articolo, gli autori sostengono che per costruire agenti cooperativi non è sufficiente che questi siano in grado di stimare lo stato attuale di un partner umano. È anche fondamentale che siano capaci di prevedere come il comportamento umano cambierà in risposta alle azioni degli agenti stessi. Propongono un approccio basato sulla razionalità computazionale, il quale si fonda sull'idea che il comportamento umano possa essere modellato calcolando politiche approssimativamente ottimali, tenendo conto delle limitazioni cognitive e fisiche degli esseri umani. L'obiettivo è combinare l'apprendimento per rinforzo e la modellazione cognitiva per aiutare le macchine a comprendere meglio le persone.

**Introduzione**

L'introduzione evidenzia l'importanza della comprensione reciproca tra esseri umani e macchine nell'IA cooperativa, con numerose applicazioni pratiche come le interfacce adattive, i veicoli automatizzati, l'interazione uomo-robot, il tutoraggio, il supporto decisionale e il design. Gli autori mettono in luce la complessità della sfida, che risiede nei processi psicologici latenti che determinano il comportamento umano. Per costruire macchine che possano comprendere le persone, è necessario risalire ai motivi psicologici che guidano le azioni umane, un compito estremamente complesso che coinvolge sia l'IA che le scienze cognitive.

**Razionalità Computazionale**

La razionalità computazionale è un approccio che combina modelli cognitivi e tecniche di apprendimento per rinforzo per comprendere e prevedere il comportamento umano. Questo metodo prende in considerazione le limitazioni umane, come la capacità limitata di elaborazione e memoria, nel calcolo delle politiche ottimali. In questo contesto, gli autori descrivono diversi strumenti e tecniche:

1. **Inferenza Approssimativa**: Gli autori propongono l'uso di metodi di inferenza approssimativa per stimare le distribuzioni posteriori dei parametri dei modelli cognitivi, invece di utilizzare valori fissi. Questo approccio permette di esprimere l'incertezza sui parametri, migliorando la capacità del modello di adattarsi a vari comportamenti umani. Ad esempio, l'uso di metodi come il campionamento Monte Carlo può aiutare a gestire l'incertezza e a migliorare la robustezza delle previsioni del modello.
2. **Politiche Approssimativamente Ottimali**: Calcolare politiche che sono approssimativamente ottimali date le limitazioni cognitive degli esseri umani è essenziale. Queste politiche non sono perfette ma sono sufficientemente buone per prevedere il comportamento umano in un contesto specifico. Gli autori discutono come le limitazioni nella capacità di elaborazione e nella memoria influenzino la capacità degli esseri umani di prendere decisioni ottimali e come i modelli di razionalità computazionale possano tenere conto di queste limitazioni per migliorare le previsioni.
3. **Apprendimento per Rinforzo**: Integrando l'apprendimento per rinforzo con modelli cognitivi, gli autori mirano a sviluppare agenti che possano apprendere dalle interazioni con gli esseri umani, migliorando continuamente la loro capacità di prevedere e rispondere ai comportamenti umani. L'apprendimento per rinforzo permette agli agenti di adattarsi dinamicamente ai cambiamenti nel comportamento umano, imparando da esperienze passate e migliorando la cooperazione nel tempo.

**Sfide e Prospettive Future**

Gli autori identificano diverse sfide principali nell'applicazione della razionalità computazionale:

1. **Alta Dimensionalità dei Dati**: La complessità dei dati osservativi richiede tecniche avanzate per estrarre informazioni rilevanti senza perdere dettagli cruciali per la modellazione del comportamento umano. La gestione della grande quantità di dati raccolti dalle interazioni tra esseri umani e macchine è un problema complesso che necessita di algoritmi efficienti per analizzare e interpretare questi dati.
2. **Metriche Sufficienti per Obiettivi Cooperativi**: È essenziale sviluppare metriche che possano valutare efficacemente gli obiettivi cooperativi tra umani e agenti artificiali, per garantire un'interazione fluida e produttiva. Le metriche devono essere in grado di catturare sia la qualità dell'interazione che l'efficacia della cooperazione, permettendo di valutare il successo dell'IA nel comprendere e adattarsi al comportamento umano.
3. **Integrazione delle Emozioni**: Le emozioni giocano un ruolo significativo nell'adattamento razionale. Gli autori suggeriscono che le emozioni possano essere integrate nei modelli di razionalità computazionale attraverso la connessione tra errori di previsione del premio (utilizzati negli algoritmi di apprendimento per rinforzo) e l'attività dopaminergica nel cervello umano. Questo approccio potrebbe aiutare a prevedere non solo le risposte emotive degli esseri umani, ma anche come le emozioni influenzano il comportamento in modi complessi. Ad esempio, comprendere come la frustrazione o la soddisfazione influenzano le decisioni potrebbe migliorare significativamente l'interazione tra esseri umani e macchine.

**Conclusioni**

L'articolo conclude che l'integrazione di modelli di razionalità computazionale può portare allo sviluppo di macchine che comprendono meglio gli esseri umani, migliorando significativamente la cooperazione e l'interazione in vari contesti applicativi. Tuttavia, gli autori riconoscono che esistono ancora molte sfide da affrontare, in particolare nella modellazione accurata dei processi psicologici latenti e nell'integrazione delle emozioni nei modelli di comportamento. Gli autori sottolineano l'importanza di continuare a sviluppare e affinare modelli che possano tener conto delle complessità del comportamento umano. La capacità di prevedere e rispondere adeguatamente alle azioni e alle emozioni umane rappresenta un passo cruciale verso la creazione di agenti artificiali veramente cooperativi.

**Riflessioni Aggiuntive**

Gli autori riflettono sull'importanza di un approccio interdisciplinare, che combini competenze dall'IA, dalla psicologia cognitiva e dalle neuroscienze, per affrontare le sfide poste dalla comprensione e previsione del comportamento umano. Essi enfatizzano che la collaborazione tra queste discipline può portare a una migliore comprensione dei processi psicologici latenti e delle emozioni che influenzano il comportamento umano.

Ad esempio, un'area di ricerca promettente è l'integrazione delle teorie cognitive con i modelli di apprendimento per rinforzo, per creare agenti che non solo reagiscano al comportamento umano, ma che possano anche prevedere come gli esseri umani potrebbero adattare il loro comportamento in risposta alle azioni degli agenti. Inoltre, gli autori suggeriscono che le emozioni possono essere modellate come parte dell'adattamento razionale, utilizzando connessioni tra errori di previsione del premio e l'attività dopaminergica.

In conclusione, l'articolo propone un quadro teorico e pratico per lo sviluppo di macchine che comprendano gli esseri umani, sottolineando l'importanza di modelli che tengano conto delle limitazioni umane e integrino le emozioni nei processi decisionali. Questo approccio ha il potenziale di rivoluzionare il modo in cui le macchine interagiscono con le persone, aprendo nuove possibilità per l'IA cooperativa. Gli autori enfatizzano che, sebbene ci siano ancora molte sfide da affrontare, i progressi in questo campo possono portare a interazioni uomo-macchina più naturali ed efficaci, migliorando la qualità della cooperazione in una vasta gamma di applicazioni.

Traduzione 1-1

**Abstract**

La capacità di stimare lo stato di un partner umano è una base insufficiente per costruire agenti cooperativi. È inoltre necessaria una capacità di prevedere come le persone adattino il loro comportamento in risposta alle azioni di un agente. Proponiamo un nuovo approccio basato sulla razionalità computazionale, che modella gli esseri umani partendo dall'idea che le previsioni possano essere derivate calcolando politiche che sono approssimativamente ottimali date le limitazioni umane. La razionalità computazionale unisce l'apprendimento per rinforzo e la modellazione cognitiva per raggiungere questo obiettivo, facilitando la comprensione delle macchine degli esseri umani.

**INTRODUZIONE**

Costruire macchine che comprendono le persone è un obiettivo chiave per la ricerca sull'IA cooperativa (Lake et al. 2017; Dafoe et al. 2021; Hadfield-Menell et al. 2016; Anderson, Boyle e Reiser 1985; Batmaz et al. 2019), con applicazioni in interfacce adattive, veicoli automatici, interazione uomo-robot, tutoraggio, supporto decisionale e design. Come, ad esempio, è possibile per una macchina capire perché Tina ha preso la strada più lunga per andare al lavoro (Figura 1)? La risposta potrebbe riguardare la sua funzione di utilità (le sue preferenze) o le sue capacità (inclusa la memoria e l'attenzione). Potrebbe aver scelto il percorso C, ad esempio, perché preferisce il paesaggio o perché ha meno incroci e, quindi, implica un carico cognitivo inferiore.

Sostanzialmente, la sfida chiave per risolvere questo tipo di problema sta nella complessità dei processi psicologici latenti che sottendono il comportamento. Questo è un problema di identificabilità; qualsiasi comportamento umano osservato può essere emerso a causa di uno dei numerosi possibili processi sottostanti. Capire quali processi latenti causano quali eventi è un problema difficile all'avanguardia della ricerca in Intelligenza Artificiale e Scienze Cognitive (ad es., Dafoe et al. 2021; Gershman e Daw 2017; Jara-Ettinger 2019; Lake et al. 2017; Russell 2019).

Un problema correlato riguarda come prevedere il comportamento in circostanze non viste, o controfattuali, rispondendo a domande "e se?". Nella Figura 1, ad esempio, per prevedere se valga la pena suggerire il percorso B, un agente cooperativo potrebbe condurre una serie di simulazioni "e se?" per determinare l'esito probabile di questa interazione. Le simulazioni potrebbero essere condotte sotto una distribuzione di assunzioni riguardanti le utilità soggettive e le capacità di Tina; in altre parole, la pianificazione potrebbe essere assistita ponendo domande "e se?" a un modello. Nell'articolo attuale, sosteniamo che rispondere a domande "e se?" e "perché?" sul comportamento umano richiede un tipo molto specifico di modello noto come modello di Razionalità Computazionale. Il loro obiettivo è catturare la straordinaria capacità della mente umana di adattare il comportamento al servizio delle utilità, delle capacità e dell'ambiente di un individuo.

**Figura 1:**

(A) (B) (C)

Panel (A): Perché Tina ha preso la lunga strada C? È stato per il paesaggio (una utilità soggettiva) o perché aveva un carico di lavoro inferiore (un limite di capacità)? Per assistere un partner umano, un assistente IA deve essere in grado di (1) ragionare sulle cause del comportamento utilizzando un modello per rispondere a domande di tipo "perché?", e (2) prevedere le conseguenze delle sue azioni, utilizzando lo stesso modello per rispondere a domande di tipo "e se?". Ad esempio, sarebbe fastidioso o utile suggerire a Tina di prendere il percorso A, attraverso una città con più incroci, o B? La razionalità computazionale modella il comportamento umano come ottimizzazione dell'utilità attesa sotto vincoli (possibili cause del comportamento). Il comportamento umano è previsto dalla politica ottimale 𝜋∗ dato l'ambiente, le capacità cognitive e l'utilità dell'individuo. Il pannello (B) illustra come i vincoli limitano lo spazio delle politiche e quindi permettono la previsione del comportamento umano. Dato una funzione di utilità (un modello dell'utilità soggettiva umana), la razionalità computazionale considera lo spazio determinato congiuntamente dall'ambiente e dalle capacità umane (intersezione dei quadrati gialli e blu). Il sottoinsieme ottimale di queste politiche (quadrato verde) 𝜋∗ fornisce una base per l'IA cooperativa per comprendere le persone. Pannello (C): I comportamenti disponibili di Tina sono illustrati come conseguenze di diverse politiche e quindi di diversi vincoli causali permettendo di rispondere a domande di tipo "perché". Perché Tina non ha preso la strada A? Risposta: perché è solo una conseguenza di una politica subottimale 𝜋𝐴 rispetto ai limiti di capacità. Perché non B? Perché B è una conseguenza di una politica subottimale rispetto all'utilità. Perché C? Perché C è una conseguenza di una politica ottimale vincolata rispetto ai vincoli su ambiente, limiti di capacità e utilità. Cosa succede se uno di questi vincoli cambia, o viene cambiato dall'IA cooperativa? Allora l'analisi dello spazio delle politiche deve essere rifatta e le implicazioni dei cambiamenti per il comportamento calcolate.

**Pagina 3**

La capacità della memoria di lavoro della maggior parte delle persone varia tra 3 e 5 elementi (Cowan 2010), ad esempio, alcune persone ricordano meno, altre di più (Conway, Jarrold e Miyake 2008). Inoltre, le esperienze variano anche a causa delle leggi del caso, ma a volte anche a seconda che le informazioni statistiche siano acquisite attraverso l'esperienza o attraverso la descrizione (Rakow e Newell 2010). Per quanto riguarda l'ambiente del compito, ci sono prove che gli esseri umani si adattano all'ambiente del compito con una straordinaria sottigliezza (Anderson 1990; Oaksford e Chater 2007). Simon (1969) ha sottolineato che è impossibile comprendere il comportamento umano senza capire che è adattato sia alla struttura dell'ambiente sia alle limitazioni cognitive. La razionalità computazionale fornisce un quadro formale per investigare entrambi come vincoli sul comportamento umano. Crediamo che sia impossibile spiegare il comportamento umano senza tenere simultaneamente conto dell'utilità, delle capacità e dell'ambiente vissuto.

In sintesi, la razionalità computazionale spiega il comportamento individuale (Howes, Lewis e Vera 2009) nonostante le difficoltà associate alla diversità, fornendo i mezzi per stabilire un legame causale dalle utilità soggettive, capacità ed esperienze, tramite il calcolo di una politica ottimale vincolata, al comportamento osservato. Nel Box 1, il modello è adattato agli individui tramite l'inferenza dei parametri, che viene effettuata nell'Equazione (2) e risulta in un insieme di parametri 𝜃∗ che governano il funzionamento del modello.

Nella scienza cognitiva, è stato dimostrato che le spiegazioni per alcuni aspetti enigmatici del comportamento umano sono state talvolta attribuite erroneamente a questi tre componenti del problema di ottimizzazione vincolata. Kahneman (Kahneman e Tversky 1979), ad esempio, attribuisce numerosi comportamenti umani a funzioni di utilità distorte quando questi dovrebbero, invece, essere attribuiti alla struttura dell'ambiente del compito (Gigerenzer 2018). Allo stesso modo, gli effetti contestuali sulla scelta sono stati attribuiti all'irrazionalità (Ariely 2009) quando, invece, sono più probabilmente causati dall'adattamento razionale alle informazioni fornite dal contesto sulla struttura dell'ambiente (Howes et al. 2016) e dal modo in cui le informazioni sono presentate (Rakow e Newell 2010).

Il lancio della moneta illustra il problema (Hahn e Warren 2009). Consideriamo un contesto in cui una moneta equa viene lanciata 𝑛 volte e siamo interessati alla probabilità di sequenze con determinate sottosequenze di lunghezza 𝑘 dove 𝑘 < 𝑛. Le persone hanno dimostrato di credere correttamente che, in queste condizioni, (ad esempio, 𝑛 = 4 e 𝑘 = 3) HHT (testa, testa, croce) si verificherà più frequentemente di HHH. Questo risultato è sorprendente per coloro che ignorano l'ambiente del compito e calcolano erroneamente che la probabilità di HHH = 0,5 × 0,5 × 0,5 = la probabilità di HHT. Il calcolo corretto richiede di contare le frequenze relative di ciascuna sequenza all'interno di sequenze di lunghezza 4. La frequenza relativa di incontrare almeno un HHT è 4∕16 = 0,25 mentre quella di incontrare HHH è 3∕16 = 0,19. La ragione per cui HHH si verifica in meno sequenze di lunghezza 𝑛 = 4 è che due delle occorrenze si sovrappongono in una singola sequenza. È sbagliato citare il fatto che una persona creda che incontrerà HHT più spesso di HHH in una sequenza come prova di una funzione di utilità distorta. Hahn (Hahn e Warren 2009) ha raggiunto questa svolta con un'analisi matematica del problema. Un modello computazionalmente razionale incorporato in un sistema di intelligenza artificiale cooperativa giungerebbe alla stessa conclusione.

Quasi sempre, ovviamente, il comportamento è attribuibile a un mix di tutti e tre i componenti del problema di adattamento. Quando le persone svolgono più di un compito contemporaneamente, le prestazioni dei compiti possono interferire; uno o entrambi i compiti possono essere rallentati o il tasso di errori può aumentare.

**Box 1: IA cooperativa con razionalità computazionale**

Il problema dell'IA cooperativa è massimizzare una funzione di valore  che riassume una certa bontà obiettiva dell'interazione, mentre un partner umano tenta di massimizzare la propria utilità soggettiva . La distinzione tra  e  consente di modellare situazioni in cui i due obiettivi non sono gli stessi. Sia  che  sono definiti come mappature da una storia ℎ ∈  a un valore scalare:  ∶  → ℝ e  ∶  → ℝ, dove  è l'insieme di tutte le possibili storie.

Un modello computazionalmente razionale dell'utente: Una storia del compito ℎ ∈  è generata dal modello 𝑀, seguendo una politica 𝜋 con parametri 𝜃 dei meccanismi della cognizione umana, in un ambiente del compito 𝜙. Un agente computazionalmente razionale (Lewis, Howes e Singh 2014) impiega una politica ottimale vincolata 𝜋∗ che massimizza l'utilità del compito:

𝜋∗ = argmax 𝜋  (ℎ), (1)

dove ℎ ∼ 𝑀(𝜋, 𝜃, 𝜙).

Domande "Perché?": L'inferenza dei parametri determina l'insieme più plausibile di parametri 𝜃∗ massimizzando la probabilità dei dati umani osservati 𝑦:

𝜃∗ = argmax 𝜃 𝑝(𝑦 ∣ 𝑀) (2)

In pratica, i modelli computazionalmente razionali richiedono un metodo di stima senza probabilità (Gutmann e Corander 2016; Jokinen et al. 2021; Kangasrääsiö et al. 2019). La probabilità è calcolata in base al modello 𝑀 dell'utente che interagisce con l'ambiente del compito 𝜙. Il posteriore dei parametri può essere utilizzato per valutare cosa ha causato il comportamento osservato, come l'utilità soggettiva e le capacità.

Domande "Cosa succederebbe se?": L'agente tenta di ottimizzare un insieme di interventi 𝑖∗ ⊂  (spesso chiamati "progetti"), che fuori dallo spazio di tutti i possibili progetti massimizzano la funzione di valore atteso , data la storia prevista delle sequenze comportamentali che sono adattamenti a questi interventi da parte dell'utente:

𝑖∗ = argmax 𝑖 (ℎ), (3)

dove ℎ ∼ 𝑀(𝜋∗, 𝜃∗, 𝜙𝑖). Allo stesso modo della funzione di utilità , la funzione di valore  mappa una storia ℎ a uno scalare. Determinare  è dipendente dal problema, ma ad esempio, si può impostare  = , in modo che l'IA abbia la stessa utilità dell'utente.

**COSA SUCCEDE SE? FARE PREVISIONI SOTTO ADATTAMENTO**

La razionalità computazionale permette di rispondere alle domande "cosa succederebbe se?" simulando le conseguenze degli interventi candidati (Oulasvirta, Jokinen e Howes 2022). Unico nell'approccio è che consente di tenere conto della risposta adattiva umana, inclusi i suoi costi, permettendo così all'IA di progettare interventi ottimali. Recentemente, le applicazioni sono emerse in due aree:

1. Progettazione computazionale.
2. Interfacce utente adattive (UIs).

**ESEMPIO: COSTRUIRE MODELLI COMPUTAZIONALMENTE RAZIONALI**

In questa sezione, illustriamo tre passaggi chiave in un flusso di lavoro per costruire modelli computazionalmente razionali e lo esemplifichiamo con un modello:

1. Specificare l'ambiente esterno dell'utente.
2. Specificare l'architettura cognitiva.
3. Imparare e adattare.