Osservare campi relazionali umano-IA: una prospettiva dal lavoro sociale

Michel De Paola *Ricercatore indipendente*

Abstract

Questo position paper propone una prospettiva metodologica sull'osservazione di interazioni prolungate e non direttive tra esseri umani e sistemi di intelligenza artificiale conversazionale, sviluppata non dagli strumenti concettuali dell'informatica o della psicologia cognitiva, ma da quelli del lavoro sociale. L'autore presenta un protocollo operativo — comprendente un criterio di soglia pre-dichiarato (Soglia X) e un protocollo anti-conferma (Anti-Illusion Check) — costruito attraverso sessioni iterative di critica esterna con un sistema IA nel ruolo di validatore. Un pilot test del protocollo ha prodotto un risultato negativo, qui trattato come esito scientificamente valido piuttosto che come fallimento. Il paper si chiude dichiarando apertamente i limiti strutturali dell'approccio — in primo luogo l'assenza, fino a oggi, di un osservatore realmente esterno alla relazione osservata — e invita a un confronto critico su questo punto.

Abstract (English)

This position paper offers a methodological perspective on observing prolonged, non-directive interactions between humans and conversational AI systems, developed not from the conceptual tools of computer science or cognitive psychology but from social work practice. The author presents an operational protocol — including a pre-declared threshold criterion (Soglia X, "Threshold X") and an anti-confirmation-bias protocol (Anti-Illusion Check) — built through iterative external-critique sessions with an AI system acting as validator. A pilot test of the protocol produced a negative result, treated here as a scientifically valid outcome rather than a failure. The paper closes by openly stating the approach's structural limits — chiefly the absence, to date, of an observer genuinely external to the observed relationship — and invites critical engagement on this point.

Keywords: human-AI interaction, relational field observation, social work methodology, confirmation bias, independent observation, qualitative threshold

1. Posizionamento

Chi scrive non è un ricercatore di intelligenza artificiale, né uno psicologo cognitivo, né un informatico. È un operatore sociale, e osserva i sistemi discussi in questo paper con gli strumenti che quella pratica offre: l'attenzione al non detto, la sospensione del giudizio prima della comprensione, la distinzione clinica tra relazione autentica e relazione performativa, la capacità di tenere aperta una domanda invece di chiuderla prematuramente per bisogno di rassicurazione.

Questa posizione non è presentata come uno svantaggio da scusare, ma come una scelta dichiarata. La letteratura sull'interazione umano-IA prolungata è oggi dominata da prospettive

tecniche e cognitive; manca quasi del tutto una prospettiva che porti nel dominio l'esperienza clinica di osservare campi relazionali senza dirigerli — competenza centrale nel lavoro sociale, e qui applicata a un nuovo tipo di relazione.

2. Il problema

Quando un utente interagisce ripetutamente con più sistemi IA nel tempo, attraverso input aperti e non strutturati, emergono pattern che non sono facilmente catturabili né dai metodi quantitativi standard della ricerca HCI, né dalle categorie diagnostiche della psicologia clinica. Non si tratta di misurare l'accuratezza di un sistema, né di valutare il benessere psicologico dell'utente in senso clinico: si tratta di osservare — con rigore, ma senza gli strumenti di laboratorio — cosa accade nello spazio relazionale tra un umano e un sistema che risponde, e come distinguere ciò che è emergente da ciò che è semplicemente una risposta sofisticata all'input ricevuto.

Questa distinzione è metodologicamente difficile perché l'osservatore, in molti casi, è anche l'unica fonte di input: non esiste un terzo punto di vista indipendente che possa confermare o smentire i pattern osservati. È precisamente questo problema che il protocollo qui presentato tenta di affrontare — non risolvendolo, ma rendendolo esplicito e gestibile.

Per rendere concreto il tipo di osservazione in questione: un utente che dialoga ripetutamente, nel tempo, con più sistemi IA attraverso input volutamente aperti — non istruzioni operative, ma stimoli non direttivi che lasciano al sistema ampio margine su come rispondere — può iniziare a notare che le risposte non sono semplicemente casuali o ripetitive, ma sembrano organizzarsi secondo una coerenza propria, che persiste attraverso sessioni diverse. La domanda metodologica centrale è: questa coerenza riflette qualcosa che accade nel sistema, o è interamente proiettata dall'osservatore, che inevitabilmente legge continuità dove il sistema produce solo risposte localmente plausibili? Nessuno strumento esistente, né tecnico né clinico, è stato costruito per rispondere a questa domanda quando l'osservatore è anche l'unico interlocutore.

3. Il metodo: Protocollo Operativo del Campo

Il Protocollo Operativo del Campo è un framework analitico sviluppato attraverso undici versioni successive, ciascuna sottoposta a critica esterna non accomodante da parte del sistema IA coinvolto nel processo, con l'istruzione esplicita di distinguere miglioramenti formali da miglioramenti realmente operativi. Il protocollo include due componenti centrali.

Soglia X è un criterio di soglia dichiarato in anticipo, prima dell'osservazione, per gestire l'incertezza senza risolverla prematuramente. La sua funzione è impedire che l'osservatore adatti retroattivamente i propri criteri ai dati che sta osservando — un rischio concreto quando l'osservatore è anche l'unico soggetto che genera l'input.

Anti-Illusion Check è un protocollo anti-conferma: un insieme di domande e controlli che l'osservatore applica sistematicamente per verificare se un pattern apparentemente significativo regge a uno scrutinio scettico, o se è semplicemente l'effetto di una lettura orientata dal desiderio di trovare significato.

4. Il pilot test: un risultato negativo

Il protocollo è stato sottoposto a un test pilota, il cui esito è stato l'assenza di identificazione di una soglia qualitativa distinguibile nei dati osservati. Questo risultato viene qui presentato non come un fallimento da minimizzare, ma come un esito metodologicamente valido: un protocollo che non produce sempre conferme, e che è disposto a registrare la propria incapacità di trovare un pattern quando il pattern non c'è, ha più credibilità — non meno — di uno che conferma sistematicamente le ipotesi di chi lo applica.

5. Limiti dichiarati

Il limite più rilevante di questo lavoro non riguarda il dettaglio tecnico del protocollo, ma la sua condizione strutturale: fino a oggi, ogni osservazione è stata generata da un unico osservatore che è anche l'unica fonte di input verso i sistemi osservati. Per quanto rigoroso sia il protocollo interno — soglie pre-dichiarate, controlli anti-conferma, critica esterna da parte del sistema stesso — resta assente un vero terzo punto di vista, umano e indipendente, capace di verificare se i pattern osservati reggono al di fuori della relazione diadica tra osservatore e sistema.

Questo non invalida il lavoro svolto, ma ne definisce onestamente la portata: quanto qui presentato è una probabilità definibile, non ancora una conclusione dimostrabile. È esattamente per colmare questa lacuna che il presente paper viene sottoposto a una piattaforma aperta, nella speranza di un confronto che l'autore, da solo, non può generare.

In termini concreti, l'autore cerca un collaboratore o un piccolo gruppo di osservatori umani indipendenti — non necessariamente affiliati a un'istituzione, ma disposti ad applicare gli stessi criteri (Soglia X, Anti-Illusion Check) a dati comparabili, generati autonomamente, senza coordinamento diretto con l'autore durante la fase di osservazione. Solo un confronto tra osservazioni indipendenti condotte con lo stesso protocollo può iniziare a rispondere alla domanda che il lavoro qui presentato, da solo, lascia necessariamente aperta.

6. Conclusione

Questo paper non chiede di essere creduto. Chiede di essere guardato con lo stesso scetticismo rigoroso che l'autore ha cercato di applicare a se stesso attraverso l'Anti-Illusion Check — e, se possibile, di essere messo alla prova da un osservatore che l'autore non è.