FATTORI ALLA BASE DELLA ATTRIBUZIONE DI FIDUCIA IN SISTEMI DI INTELLIGENZA ARTIFICIALE

1. Progetto di Ricerca

Il progetto di ricerca si propone di esplorare, in un contesto di presa di decisione, i fattori che influenzano l'attribuzione di fiducia nei confronti dei sistemi basati su Intelligenza Artificiale (IA), in particolare focalizzandosi sull'interazione tra utenti umani e sistemi di supporto alla decisione (Decision Support Systems, DSS). La fiducia riveste un ruolo fondamentale nell'accettazione di sistemi DSS da parte degli utenti. Alcuni dei fattori che modulano la fiducia nei DSS sono la capacità del DSS di adattarsi alle esigenze degli utenti, la chiarezza delle spiegazioni fornite e la consistenza nei risultati nel corso del tempo (Morandini et al., 2023). Inoltre, è fondamentale la possibilità di spiegare in base a cosa il DSS abbia preso una certa decisione (exaplainability; Miller, 2019), in termini comprensibili per l'utente (Floridi, 2022). Inoltre, dato che utenti umani e DSS collaborano nell'esecuzione di un compito (agency condivisa), è importante considerare quanto le decisioni prese dal DSS ricalchino quelle umane o se ne discostino (Engen et al., 2016).

Il progetto di ricerca è articolato in due studi che si occupano di indagare, in un contesto sperimentale (studio 1) e applicativo (studio 2), quali fattori modulino l'attribuzione di fiducia a un sistema artificiale. In particolare, sarà indagato il ruolo dell'accordo tra decisioni umane e di un DSS, dell'accuratezza del risultato ottenuto, e della capacità di dare una spiegazione della decisione presa dal DSS.

Obiettivo dello studio sperimentale è indagare, in scenari di presa di decisione di vita quotidiana, l'attribuzione di fiducia in funzione dell'accordo tra decisioni umane e suggerimenti di un sistema DSS, dell'accuratezza dei risultati, e della capacità di fornire spiegazioni. Ad esempio, è possibile che, in caso di esiti positivi di una decisione, sia percepito come relativamente poco importante il grado di accordo tra decisione umana e artificiale; tuttavia, se un sistema IA propone una decisione che si discosta molto da quelle umane e questa porta a esiti negativi, la fiducia nel sistema IA potrebbe calare nettamente.

Nello studio applicativo, si andrà a studiare un contesto in cui decisori umani (planner) devono risolvere un complesso problema di pianificazione. Per risolvere questo problema i planner hanno a disposizione la loro esperienza passata e la attuale soluzione adottata per la pianificazione (human knowledge e status quo), e valutano soluzioni ottimizzate proposte da sistemi artificiali con l'obiettivo di ridurre costi, impatto ambientale, o aumentare l'efficienza del sistema. Obiettivo dello studio sarà quindi indagare quali sono i fattori che

modulano la fiducia dei planner umani nei confronti di diverse soluzioni proposte da sistemi artificiali, e quali sono gli aspetti delle soluzioni proposte che sono tenuti maggiormente in considerazione della scelta e nella attribuzione di fiducia da parte dei planner.

Nel complesso, il progetto permetterà di acquisire informazioni rispetto a quali fattori influenzino l'attribuzione di fiducia. In un contesto in cui strumenti di Intelligenza Artificiale sono costantemente sviluppati per essere integrati con le attività umane, progettare sistemi in grado di adattarsi alle richieste cognitive umane, ad esempio scegliendo tra due alternative simili in termini di costi/benefici quella più vicina alle aspettative umane potrebbe aiutare ad aumentare la fiducia in un sistema IA in caso di esiti positivi, e a ridurre la sfiducia in caso di esiti negativi.

Engen, Pickering, & Walland (2016). Machine Agency in Human-Machine Networks; Impacts and Trust Implications. Proceedings from the 18th International Conference on Human-Computer Interaction International, Toronto, Canada

Floridi, L. (2022). Etica dell'Intelligenza Artificiale. Raffaello Cortina Editore

Morandini et al., (2023). Examining the Nexus between Explainability of AI Systems and User's Trust: A Preliminary Scoping Review. Proceedings from the 1st World Conference on eXplainable Artificial Intelligence, Lisbon, Portugal

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

2. Piano di Attività

Il piano di attività consisterà in :

- Preparazione e raccolta dati dei due studi indicati (studio sperimentale e studio applicato). Per entrambi gli studi, saranno preparati i materiali sperimentali richiesti (scenari per la presa di decisione, interviste da somministrare ai pianificatori, etc.) e saranno strutturate le sessioni di raccolta dati.
- Revisione della letteratura di riferimento sulla attribuzione di fiducia a sistemi di Intelligenza Artificiale.
- I risultati degli studi e della ricerca in letteratura saranno periodicamente riportati ai labmeeting.
- I risultati di maggiore interesse potranno essere sottomessi per la pubblicazione a riviste scientifiche o per la presentazione a conferenze di riferimento.