UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia Generale

Corso di Laurea Magistrale in Psicologia Clinica

Tesi di Laurea Magistrale

Confronto tra annotatori umani e LLMs nello scoring di ricordi autobiografici genuini e mentiti per la valutazione della credibilità

Comparison of human annotators and LLMs in scoring genuine and lied autobiographical memories for credibility assessment

Relatore
Prof. Giuseppe Sartori

Correlatore esterno
Dott. Riccardo Loconte

Laureanda: Chiara Pachera
Matricola: 2088974

Anno Accademico 2023/2024

INDICE

ABSTRACT

La ricerca forense ha come fulcro la comprensione della memoria umana e lo sviluppo di metodi per valutarne l'affidabilità. Tradizionalmente, gli esperti esaminano manualmente i resoconti di memoria, basandosi su fattori come la quantità e qualità dei dettagli riportati così come la loro verificabilità, un metodo però criticato a causa di problematiche come la soggettività dei valutatori, il costo delle risorse e il tempo richiesto. L'avvento dei *Large Language Models* (LLMs), ovvero reti neurali addestrate su vasti corpora in modo da predire la parola successiva in una sequenza, ha portato ad una serie di studi che hanno cercato di indagare la capacità del modello automatico di simulare il giudizio umano.

Il presente elaborato mira a confrontare le prestazioni degli annotatori umani e di un LLM, nello specifico GPT-4, nell'analisi di memorie autentiche e falsificate grazie all'uso di due *dataset*. Il primo (Monaro et al., 2020) comprende 62 narrazioni riguardanti vacanze passate trascritte da interviste videoregistrate in italiano. Ai valutatori umani è stato chiesto di estrapolare i dettagli dai testi forniti utilizzando il *Reality Monitoring* e il *Verifiability Approach*. Questa analisi è stata successivamente replicata su un sottoinsieme del *dataset Hippocorpus* (Sap et al., 2022) contenente 240 dichiarazioni scritte in inglese relative ad esperienze passate significative.

I risultati indicano che GPT-4 ha mostrato una concordanza eccellente con gli annotatori umani nel numero di dettagli identificati tramite il *Reality Monitoring*, mentre tale accordo è risultato moderato per quanto riguarda l'etichetta attribuita all'informazione individuata. Tuttavia, nell'applicazione del *Verifiability Approach*, il modello automatico ha fallito nella valutazione della verificabilità dei dettagli.

INTRODUZIONE

Lo svolgimento della presente ricerca si è reso necessario in quanto risulta fondamentale, all'interno della psicologia forense, la valutazione della credibilità della testimonianza di un individuo, cercando di automatizzare il più possibile tale processo. Tradizionalmente infatti, tale analisi è svolta manualmente, con il conseguente impiego di un gran numero di risorse nonché di tempo, e della possibile influenza della soggettività dei valutatori. Con lo sviluppo dei *Large Language Models* (LLMs) tuttavia, ovvero modelli di intelligenza artificiale basati su reti neurali e progettati per comprendere e generare testo naturale, si può auspicare ad un'automatizzazione della procedura.

Il presente elaborato risulta quindi organizzato in due sezioni principali. La prima è composta dai tre capitoli iniziali i quali, partendo da un'esaustiva definizione e descrizione della memoria autobiografica, proseguono con una revisione bibliografica circa gli strumenti utilizzati in psicologia forense per la valutazione della credibilità di una testimonianza, per poi presentare diverse tecniche di elaborazione automatica del linguaggio e il loro impiego in ambito psicologico.

La seconda sezione, di natura empirica, coinvolge invece gli altri tre capitoli. In essi, le parti relative al metodo, ai risultati e alla discussione descrivono dettagliatamente il processo di ricerca, portando infine alle conclusioni di rilevanza scientifica raggiunte.

CAPITOLO 1: LA MEMORIA AUTOBIOGRAFICA

La memoria può essere definita come l'abilità di mantenere una traccia più o meno duratura degli stimoli che sperimentiamo (Istituto dell'Enciclopedia italiana, 2017). Tuttavia, essa opera in modo differente a seconda del tempo a disposizione dell'individuo per immagazzinare le informazioni, permettendo così di descrivere una gerarchia della memoria dentro la quale si declinano le diverse tipologie di ricordi (*Figura 1.1*). Tale tassonomia prevede la distinzione di due tipologie di memoria.

La prima fa riferimento alla memoria a breve termine (MBT), di facile accesso, la quale riflette la capacità di mantenere l'informazione codificata per un periodo di circa 20 secondi e che si compone a sua volta della memoria sensoriale, ovvero l'acquisizione dell'informazione dall'ambiente esterno attraverso gli organi di senso, e della memoria di lavoro, ossia quel sistema cognitivo che permette di mantenere e manipolare temporaneamente l'informazione necessaria per lo svolgimento di compiti complessi come la comprensione del linguaggio, il ragionamento e l'apprendimento.

La seconda è costituita invece dalla memoria a lungo termine (MLT), la quale consente di immagazzinare, gestire e recuperare ricordi per periodi di tempo prolungati; essa è caratterizzata da una capienza illimitata e necessita del passaggio dell'informazione al magazzino a breve termine per lo svolgimento del compito. La memoria a lungo termine può essere ulteriormente distinta in memoria implicita o procedurale, in cui una persona ricorda inconsciamente vari tipi di informazione come quella necessaria per lo svolgimento di un compito fisico, e memoria esplicita o dichiarativa, grazie alla quale una persona rievoca consciamente un evento passato che si è verificato in un

determinato tempo e spazio. Quest'ultima rappresenta inoltre il connubio tra la memoria semantica, la quale riguarda la conoscenza generale del mondo e che comprende fatti, concetti e informazioni che non sono legati ad esperienze personali specifiche, e la memoria episodica, caratterizzata invece da ricordi di esperienze vissute in prima persona, inclusi i dettagli relativi al contesto temporale e spaziale in cui questi eventi si sono verificati (Nolen-Hoeksema et al., 2017).

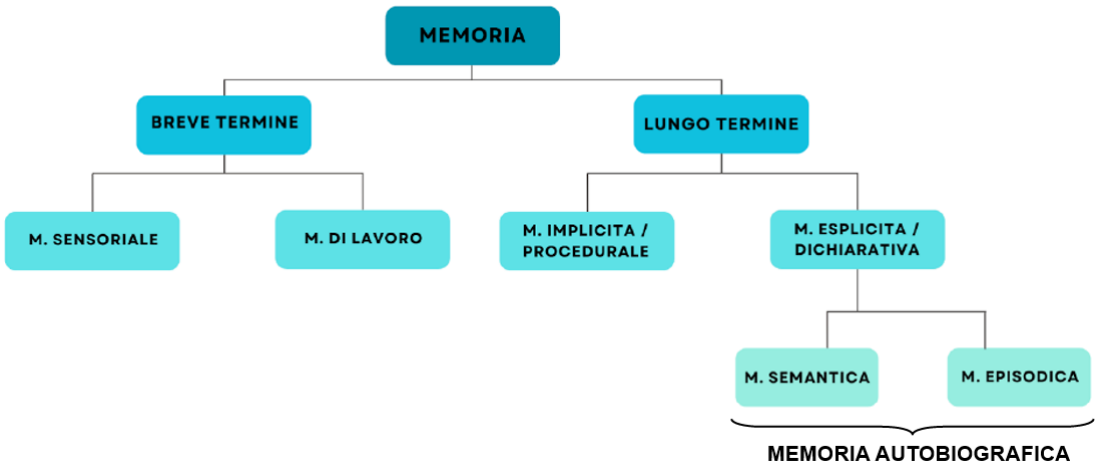

***Figura 1.1**: Tassonomia della memoria.*

All'interno del presente elaborato risulta tuttavia centrale la memoria autobiografica: essa viene definita come un insieme di conoscenze spazio-temporali e fattuali vissute in prima persona contestualizzate all'interno della storia personale del soggetto che le ricorda; la caratteristica principale di questo tipo di ricordi è la stretta relazione con il sé (Conway, 1996). Essa pone le sue radici all'interno di sistemi di funzionamento relativi alla memoria episodica e semantica, risultando tuttavia indipendente da esse (Magro et al., 2023).

È possibile differenziare due tipologie di memoria autobiografica, quella spontanea, o involontaria, quando un evento viene ricordato in maniera totalmente autonoma, e

quella incidentale, o volontaria, quando il ricordo viene sollecitato da fattori esterni, come ad esempio domande (Sartori, 2021).

Tra le funzioni attribuite alla memoria autobiografica, risultano essere centrali le seguenti: la "funzione sociale" (*Social function*), in quanto i ricordi autobiografici possono fungere da argomento di discussione facilitando in questo modo le interazioni interpersonali; la "funzione direttiva" (*Directive function*), che fa riferimento alla capacità degli individui di utilizzare le proprie esperienze passate per attribuire significati e interpretazioni alle vicende presenti e future; la "funzione del Sé" (*Self function*), che richiama l'importanza rivestita da questo tipo di memoria nella costruzione della propria identità (Magro et al., 2023).

Proprio a causa del forte legame con l'identità personale, la visione che la persona possiede circa sé stessa può andare ad influenzare il modo in cui ricorda determinati eventi del suo passato (Bartlett, 1995). Più nello specifico Ross (1989), nel tentativo di spiegare in che modo gli individui tendano a ricostruire la propria immagine passata, ha evidenziato l'esistenza di un processo a due fasi che prevede un'iniziale presa in considerazione dei sentimenti e delle attribuzioni nel qui ed ora, i quali risultano più facilmente accessibili a causa della vicinanza temporale, per poi procedere con la fabbricazione di un'immagine del sé passata coerente con quella attuale seguendo in questo modo teorie implicite che richiamano la stabilità dei propri attributi. In altre parole, gli esseri umani sono portati erroneamente a pensare che le attribuzioni, le caratteristiche e i comportamenti che conferiscono al sé passato siano in linea con quelli presenti, mettendo involontariamente in atto un bias mnestico e andando di conseguenza ad influenzare i propri ricordi (Wilson & Ross, 2003; Conway & Ross, 1984).

La memoria autobiografica risulta inoltre caratterizzata da tre stadi fondamentali: la **codifica**, ovvero la trasformazione delle informazioni in una tipologia di codice che la memoria è in grado di comprendere; l'**immagazzinamento**, ossia il mantenimento di tali contenuti; il **recupero**, il processo per mezzo del quale le informazioni vengono recuperate dalla memoria (Nolen-Hoeksema et al., 2017).

La microstruttura della memoria autobiografica può essere descritta come composta da tre livelli di specificità (*Figura 1.2*): il primo fa riferimento al periodo della vita in cui è avvenuto il ricordo autobiografico, con annessi fatti accaduti e persone coinvolte; il secondo allude all'evento generale che può incorporare sia accadimenti singoli che ricorrenti, o una serie di episodi legati da un tema comune; il terzo e ultimo livello riguarda l'evento specifico e tutta una serie di dettagli che caratterizzano il ricordo in sé (Conway & Pleydell-Pearce, 2000).

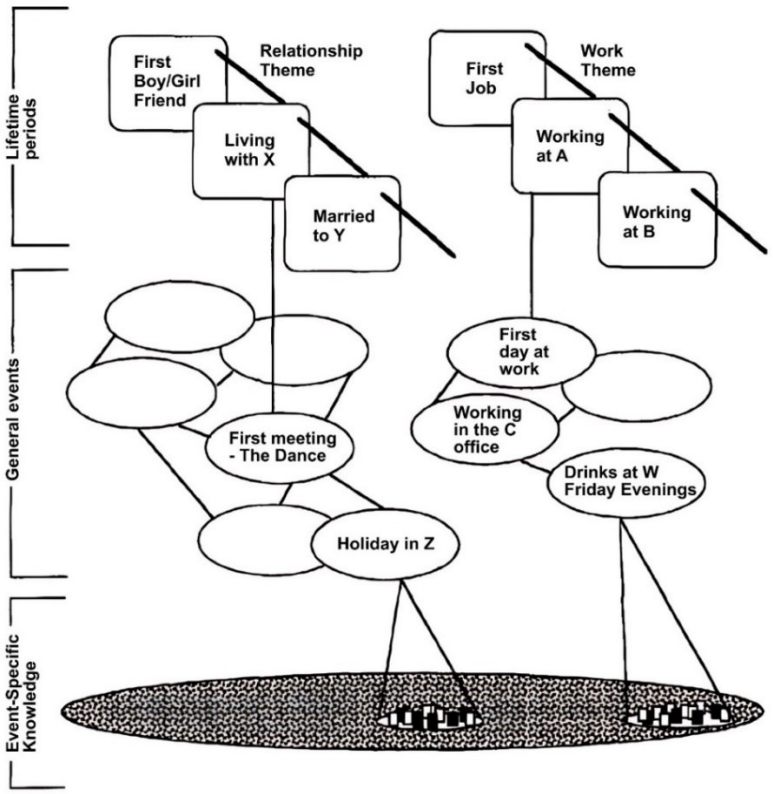

***Figura 1.2**: La microstruttura della memoria autobiografica (Conway & Pleydell-Pearce, 2000).*

Sulla base di tale strutturazione, all'interno della memoria autobiografica possono essere distinte due tipologie di dettagli relativi al fatto di interesse, ovvero **dettagli centrali**, che si riferiscono al *Chi*, al *Come*, al *Dove*, al *Quando* e al *Perché* dell'evento, e **dettagli periferici**, che nella maggior parte dei casi vengono identificati come specificazioni, emozioni, percezioni etc (Sartori, 2021).

Tuttavia, dato che la memoria è un processo di tipo costruttivo e non riproduttivo a causa del quale il ricordo di un evento può divergere sistematicamente dalla realtà obiettiva da cui si origina sia nel momento in cui si forma che in seguito, gli errori di memoria possono colpire qualsiasi dei tre stadi descritti in precedenza (Nolen-Hoeksema et al., 2017).

Un esempio è rappresentato dal semplice passaggio del tempo. Secondo la *legge di Ebbinghaus*, detta anche *legge dell'oblio*, contrariamente alla convinzione comune, la capacità di ricordare le informazioni, in assenza di revisione o pratica, tende a diminuire gradualmente e in modo significativo con l'aumentare del tempo secondo un processo fisiologico; tale decadimento sarebbe sostanziale nei primi anni successivi all'evento di interesse, per poi stabilizzarsi (Murre & Dros, 2015). Questo principio psicologico è inoltre valido sia per quei ricordi definiti confondibili, ovvero che possiedono caratteristiche simili ad altri eventi e che si presentano in maniera abbastanza frequente nella vita quotidiana, sia per le memorie atipiche, le quali si riferiscono ad avvenimenti poco comuni che si distinguono dalla normale gamma di esperienze di vita della persona, come ad esempio subire un'aggressione, e che possono avere un impatto significativo a livello emotivo ma non per questo rimanere maggiormente impresse nella memoria.

La capacità di ricordare un evento specifico dipende inoltre dalla sua tipologia. Diversi studi hanno infatti dimostrato non solo che i ricordi periferici, rispetto a quelli centrali, siano maggiormente soggetti al decadimento mnestico dovuto al passaggio del tempo, ma anche come, a parità di distanza temporale, l'accuratezza del ricordo di conversazioni sia peggiore di quella di eventi (Sartori, 2021).

Nel tentativo di fornire un ulteriore supporto empirico, Bruck e collaboratori (1999) hanno chiesto a delle madri di intervistare i loro bambini di quattro anni circa un'attività di gioco avvenuta pochi istanti prima durante la quale erano assenti; i risultati mostrano che, nonostante la metà delle partecipanti fosse stata avvertita che si trattasse di un esperimento che andava ad indagare le capacità mnestiche, la memoria delle madri circa le caratteristiche strutturali della conversazione, come ad esempio le domande poste al minore, era significativamente inferiore in termini di accuratezza rispetto al significato generale dell'interazione avvenuta con il proprio figlio. Tale problematica è presente anche in riferimento al ricordo delle ragioni alla base della messa in atto di un determinato comportamento, ovvero la memoria motivazionale (Sartori, 2021).

Un ulteriore fattore che può influenzare la capacità di recupero delle memorie autobiografiche è la presenza di una psicopatologia, di un disturbo di personalità o di una disabilità intellettiva.

Premettendo che la presenza di un disturbo psicopatologico non compromette automaticamente la capacità mnestica della persona che ne è affetta, risulta tuttavia fondamentale descrivere i casi in cui ciò può avvenire. Molti studi hanno dimostrato che la presenza di patologie come il Disturbo Borderline di Personalità (DBP; Jones et al., 1999), il Disturbo Acuto da Stress (DAS; Harvey et al., 1998), il Disturbo Post-Traumatico da Stress (DPTS; McNally et al., 1994) e il Disturbo Depressivo Maggiore

(DDM; Kuyken & Dalgleish, 1995) possono portare alla presenza di un fenomeno chiamato *Overgeneral Memory*, in italiano "memoria ipergenerale". Tale manifestazione fa riferimento al fatto che gli individui affetti da patologie come quelle appena elencate tendano a ricordare gli eventi relativi al passato in modo più generale e astratto rispetto ai soggetti sani, i quali riescono invece a riportare memorie più specifiche e dettagliate (Moore & Zoellner, 2007).

In aggiunta, la presenza di forme psicopatologiche relative all'area psicotica, come ad esempio la Psicosi paranoidea, caratterizzata da forte diffidenza e sospettosità nei confronti delle altre persone che tende a mantenersi anche in assenza di minacce reali (APA, 2013), può andare a modificare non tanto i ricordi autobiografici dal punto di vista quantitativo come visto finora, quanto a livello qualitativo, attribuendo intenzioni spesso ostili ai comportamenti degli altri e alterando in che modo la memoria viene immagazzinata (Sartori, 2021).

Il fatto che l'individuo sia affetto da una disabilità intellettiva come la Sindrome di Williams (SW), la Sindrome di Down (SD) o che semplicemente abbia un Quoziente Intellettivo (QI) inferiore a 70, può avere come possibile conseguenza una maggiore suggestionabilità, ovvero la tendenza di un individuo a modificare i propri ricordi in risposta a suggerimenti o informazioni esterne, e alle false memorie rispetto alla media della popolazione generale (Griego et al., 2019).

Un ulteriore elemento che potrebbe influenzare la capacità mnestica dell'individuo è quello che viene definito come *effetto arma*, ovvero un fenomeno per cui la presenza di un'arma durante un crimine può influenzare la memoria dei testimoni, portandoli a ricordare erroneamente dettagli dell'evento, spesso mostrando un'attenzione selettiva del testimone verso l'arma, sovrastimandone la presenza o il suo coinvolgimento.

Questo fenomeno può essere attribuito a diversi fattori, tra cui l'intensità emotiva dell'evento e l'effetto "sorpresa" derivato dalla presenza dell'arma (Loftus & Palmer, 1974; Steblay, 1992).

Infine, anche il periodo della vita dell'individuo in cui è avvenuto il fatto di interesse può influenzare la capacità del soggetto di ricordare. In primo luogo vi è l'*amnesia infantile*, ovvero un'incapacità quasi totale nel recuperare dalla memoria gli eventi che sono avvenuti dalla nascita a circa il quinto anno di età (Howe & Courage, 1993); tale fenomeno è dovuto non solo al fatto che in questo periodo della vita del minore la velocità dell'oblio è particolarmente elevata rispetto a quella dell'adulto, ma bisogna inoltre tenere in considerazione che la memoria autobiografica dipende dallo sviluppo del linguaggio, il quale non emerge prima del secondo anno di vita (Sartori, 2021), senza negare l'influenza di fattori come lo sviluppo del sé cognitivo o dei lobi frontali, e la nascita della comprensione della funzione sociale dei ricordi da parte del bambino (Mammarella & Di Domenico, 2011).

In secondo luogo è presente il cosiddetto *balzo del ricordo*, ovvero una maggior capacità dell'individuo nel recuperare e contestualizzare i ricordi che accadono nel periodo compreso tra i dieci e i trent'anni di età (Rubin et al., 1998); secondo alcuni autori (Mammarella & Di Domenico, 2011) gli eventi che avvengono all'interno di questo range temporale rappresenterebbero delle esperienze di definizione di sé e dunque andrebbero a comporre la parte più resistente della memoria autobiografica, diventando in questo modo più facilmente recuperabili.

CAPITOLO 2: LA VALUTAZIONE DELLA CREDIBILITA' DI UNA TESTIMONIANZA

2.1 La differenza fra attendibilità, accuratezza e credibilità

Nel mondo forense si può osservare come spesso vengano utilizzati i termini accuratezza, attendibilità e credibilità in modo interscambiabile, nonostante i loro diversi significati.

All'interno dell'ambito della testimonianza, l'**accuratezza** corrisponde alla capacità del testimone di ricordare e riportare correttamente i dettagli di un evento o di una situazione, ovvero quanto un resoconto risulta preciso, dettagliato e coerente sia sotto il profilo percettivo sia sulla base delle capacità mnestiche del soggetto (Scali et al., 2003). Tuttavia, tale valutazione può essere compiuta solo nel caso si abbia a disposizione ciò che a cui effettivamente il soggetto ha assistito: ciò che vi si avvicina maggiormente è l'attendibilità estrinseca, ovvero la verifica oggettiva del contenuto delle dichiarazioni dell'individuo sulla base di fattori esterni come ad esempio filmati di videocamere o documenti ufficiali.

In mancanza del riscontro esterno però, l'accuratezza viene desunta sulla base dei criteri dell'**attendibilità intrinseca**; si cerca quindi di stimare la valutazione del ricordo indirettamente sulla base dell'analisi della struttura della narrazione fornita dal testimone (Sartori, 2021).

Nel complesso, mentre i termini accuratezza e attendibilità fanno riferimento al contenuto delle dichiarazioni, quello di **credibilità** riguarda invece tutta una serie di aspetti motivazionali e personologici del testimone stesso, prendendo in considerazione

possibili influenze suggestive, come ad esempio il grado di parentela con l'accusato, che possono aver orientato le sue dichiarazioni portando il soggetto a mentire intenzionalmente (Tommasino et al., 2008).

In ambito scientifico e di ricerca tuttavia, quando si fa riferimento alla valutazione di una testimonianza, si utilizza il termine *credibility assessment*, il quale comprende qualsiasi tentativo di accertare la veridicità del racconto fornito (Yuille, 1989; Vrij, 2007; Raskin et al., 2013). All'interno del presente elaborato dunque, ogni riferimento al termine "credibilità" deve essere inteso esclusivamente nel contesto scientifico di determinazione della veridicità di una o più affermazioni indipendentemente dalle caratteristiche personologiche e motivazionali del testimone nel produrre il proprio resoconto.

Gli studi sulla valutazione della credibilità fanno parte della linea di ricerca definita *lie detection*, un insieme di strumenti utilizzati in ambito forense per la detezione della menzogna, ovvero per capire se una persona stia mentendo o meno (Sartori, 2021). Tali tecniche, nate originariamente per sopperire al fatto che le persone da sole non sono in grado di discriminare la verità dalla bugia con un'accuratezza superiore al caso (50%; Vrij, 2000), possono basarsi su rilevazioni di tipo fisiologico [1], espressivo [2], neuropsicologico[3], comportamentale[4], cognitivo o linguistico; in particolare, queste due ultime tipologie fanno riferimento alla *verbal lie detection*, ovvero l'analisi del linguaggio parlato per la detezione della menzogna.

[1] Poligrafo, *Guilty Knowledge Test* (GKT).
[2] *Facial Action Coding System* (FACS).
[3] Fa riferimento all'uso della risonanza magnetica funzionale (*functional Magnetic Resonance Imaging*, fMRI) o lo studio dei potenziali evento-relati (*Event-Related Potentials*, ERPs).
[4] *Keyboard dinamics*, *Mouse tracking,* aIAT.

Nei paragrafi successivi verranno esposte alcune delle principali tecniche di valutazione della credibilità di un ricordo autobiografico, con un focus iniziale sulle tecniche specifiche di intervista.

2.2 Le tecniche di intervista

2.2.1 *Cognitive Interview*

La *Cognitive Interview* (CI; Geiselman et al., 1984) rappresenta una tecnica di intervista utilizzata in ambito forense per ottenere testimonianze dettagliate e accurate da parte dei testimoni che hanno una elevata motivazione a collaborare. Essa ha alla base il seguente presupposto: chi dice la verità tende a riportare maggiori dettagli aggiuntivi in relazione alla propria testimonianza, sia a causa del fatto che chi mente cerca di fare il contrario, ovvero mantenere le proprie storie più semplici possibili, sia perché ciò rappresenta la conseguenza della loro inclinazione ad essere più disponibili e collaborativi (Bogaard et al., 2019).

Questa tecnica è composta da cinque fasi. Nella prima viene chiesto agli intervistati di riferire tutto ciò che ricordano dell'evento di interesse mediante un richiamo libero. Nella seconda fase viene chiesto di rievocare nuovamente il fatto di interesse, includendo però tutti i dettagli relativi ad esso come suoni, immagini, odori e pensieri. La terza fase consiste nel chiedere di ricordare da diverse prospettive, come quella della vittima, se presente, o semplicemente di un'altra persona: tale strategia ha alla base l'ipotesi teorica che un cambiamento di prospettiva costringa il soggetto ad aggiungere ulteriori informazioni alla propria storia. Nella quarta fase viene chiesto ai partecipanti di raccontare la vicenda in un ordine cronologico diverso da quello normale, partendo ad esempio dal centro, dalla fine o dall'evento più memorabile. Nella quinta ed ultima

fase infine, viene nuovamente richiesto di raccontare l'evento nel modo più dettagliato possibile (Vrij et al., 2022; Memon & Gawrylowicz, 2018).

Tale protocollo presenta inoltre una serie di nove domande chiuse con due alternative di risposta (ad esempio "Se fosse stato presente un agente di polizia, avrebbe notato qualcosa di sbagliato?") che vengono proposte tra le fasi descritte precedentemente; tali quesiti vengono proposti da una parte per aiutare il soggetto a riflettere maggiormente sull'accaduto, e dall'altra per riconoscere almeno in parte le strategie dei bugiardi, i quali rispondendo di "no" evitano la domanda di *follow-up* mantenendo così il racconto il più semplice possibile (Memon & Gawrylowicz, 2018).

In generale, la *Cognitive Interview* non solo aumenta del 25% la qualità dell'informazione circa l'evento di interesse rispetto all'uso dell'intervista tradizionale (Köhnken et al., 1999) ma è utile anche nel discriminare tra dichiarazioni vere e false (Hernández-Fernaud & Alonso-Quecuty, 1997). Inoltre, tale tecnica permette di aumentare il numero di dettagli riportati correttamente, con solo un leggero incremento dei dettagli non corretti (Memon et al., 2010).

2.2.2 *Reality Interview*

La *Reality Interview* (RI) è un protocollo di intervista standardizzato che rappresenta una variante della *Cognitive Interview* originale.

La differenza principale tra *Cognitive Interview* e *Reality Interview* è che mentre la prima si concentra sull'ottenimento di informazioni e la riduzione della contaminazione del ricordo, la seconda si occupa anche della rilevazione dell'inganno (Bogaard et al., 2019).

Più nel dettaglio, la RI si basa sul principio del *Differential Recall Enhancement* (DRE), una tecnica utilizzata sia per migliorare la capacità di una persona nel recuperare i dettagli specifici di un evento, sia per aumentare la difficoltà cognitiva dei soggetti che mentono. Esso si basa sul presupposto che la memoria non sia un processo passivo, ma un'attività dinamica influenzata da diversi fattori; di conseguenza, variando il modo in cui vengono poste le domande e il contesto dell'intervista, è possibile migliorare la quantità e la qualità delle informazioni recuperate dalla memoria del testimone (Colwell et al., 2013). Altre tecniche che rientrano del quadre DRE sono l'uso strategico delle prove (SUE) e l'imposizione di un carico cognitivo, le quali verranno descritte nel dettaglio all'interno dei paragrafi seguenti.

La *Reality Interview* risulta essere purtroppo uno dei protocolli meno esaminati all'interno dell'ambito forense e, quando presa in considerazione, si tende a valutare la capacità di discriminare tra verità e bugia nel suo complesso, senza esaminare la validità di ciascuna fase precedentemente descritta; alcuni studi hanno tuttavia dimostrato l'efficacia della richiesta di raccontare l'evento nell'ordine inverso nel rilevamento delle menzogne (Evans et al., 2013).

2.2.3 *Strategic Use of Evidence*

Il comportamento messo in atto da sospettati innocenti e colpevoli davanti ad un interrogatorio risulta essere differente: i primi infatti temono che l'investigatore non creda alla loro versione e cercano quindi di dare tutte le possibili informazioni, i secondi invece si preoccupano di non fornire quei dettagli che solo gli autori del reato posseggono (Vrij et al., 2011). Tali processi cognitivi si traducono nella messa in atto di strategie differenti: da una parte gli individui innocenti risultano essere molto

disponibili e aperti alle richieste di maggiori dettagli; dall'altra, i colpevoli utilizzano molto la negazione, ovvero si rifiutano di riconoscere una determinata informazione, o l'evitamento, evitando quindi di fornire determinati dettagli circa il fatto di interesse (Granhag & Hartwig, 2008).

Lo *Strategic Use of Evidence* (SUE) viene messo in atto quando gli intervistatori pongono al sospettato quesiti riguardanti delle prove che l'individuo non è consapevole essere in possesso dalle autorità (Vrij, 2018). La tecnica prevede dapprima l'uso di domande aperte per elicitare il racconto, seguite poi da domande chiuse specifiche senza rivelare la prova in possesso dagli investigatori; secondo la logica sopra menzionata, è probabile che i sospettati sinceri menzionino tale informazione in modo spontaneo, al contrario dei colpevoli (Vrij et al., 2011). Di fronte a questi quesiti, mentre coloro che dicono la verità tendono a fornire delle dichiarazioni che sono coerenti con le prove in possesso dagli investigatori, i mentitori tendono a modificare i propri racconti risultando in questo modo meno coerenti (Hartwig et al., 2014).

Per validare l'accuratezza della strategia appena descritta, Hartwig e collaboratori (2006) hanno istruito metà partecipanti ad intervistare i sospettati con la tecnica SUE e l'altra metà con una metodologia a loro scelta: i risultati mostrano che mentre il primo gruppo ha ottenuto un'accuratezza dell'85,4% nel discriminare sinceri e bugiardi, il secondo gruppo è arrivato solamente a 56,1%.

2.2.4 *Cognitive Credibility Assessment* (CCA)

Raccontare un'esperienza autobiografica richiede il coinvolgimento di meccanismi di recupero della memoria. La formulazione di una menzogna, invece, risulta più complessa, in quanto l'individuo deve dapprima sopprimere il recupero della traccia

mnestica autobiografica reale (*inhibition)* e produrre una versione alternativa del racconto (*shifting*), assicurandosi che la variante prodotta non contenga incongruenze logiche o contraddizioni. Ciò comporta che la produzione di una menzogna richiede uno sforzo cognitivo maggiore del raccontare la verità (Zuckerman et al., 1981). Inoltre, quando vi è la necessità che la sua risposta falsa venga creduta dall'interlocutore, il mentitore mette in gioco processi di monitoraggio per i) il controllo del proprio comportamento al fine di mostrarsi onesto e ii) il controllo delle reazioni dell'intervistatore per verificare che l'interlocutore stia percependo il contenuto delle affermazioni come veritiere (*monitoring*; Vrij et al., 2008). Di conseguenza, lo sforzo cognitivo derivato dall'ingaggio dei processi di *inhibition, shifting* e *monitoring* si riflette in tempi di risposta più elevati (Vrij et al., 2009).

Sulla base di tali evidenze sono state studiate e definite le seguenti strategie di intervista finalizzate ad aumentare il carico cognitivo del mentitore.

Imposizione di un carico cognitivo

Una prima strategia che potrebbe essere utilizzata per aumentare il carico cognitivo negli individui che mentono è quella di chiedere agli intervistati di raccontare la storia in un ordine diverso, ad esempio partendo dalla fine per arrivare all'inizio, oppure attraverso una prospettiva diversa dalla propria. Nel loro studio, Vrij e collaboratori (2007a) hanno chiesto a metà partecipanti sinceri e metà bugiardi di raccontare la loro versione di un evento in un ordine diverso da quello standard; i risultati mostrano che non solo gli intervistatori erano maggiormente in grado di discriminare i due gruppi di individui, ma anche gli osservatori esterni che guardavano le interviste registrate successivamente.

Un secondo metodo che può essere sfruttato per aumentare il carico cognitivo è quello di chiedere al sospettato di mantenere il contatto visivo con la persona che lo intervista. Come nel caso precedente, Vrij e colleghi (2007b) hanno riscontrato che nella condizione di contatto visivo, rispetto a quella di controllo, era osservabile un numero maggiore di segni di menzogna, riscontrabili anche da coloro che guardavano tali interazioni in differita.

Un'ulteriore strategia è quella detta del *turn-taking* forzato, in cui più soggetti vengono interrogati contemporaneamente: l'intervistatore, dopo aver posto la domanda, deciderà chi dovrà rispondere e dopo un breve tempo potrà decidere di interrompere tale persona per far continuare qualcun altro (Vernham et al., 2014).

Infine, si potrebbe chiedere alle persone sospettate di svolgere un compito secondario mentre rievocano la loro versione della storia. Dato che, come detto precedente, dire una bugia richiede più risorse cognitive che dire la verità, Vrij e collaboratori (2008) ipotizzano che le persone che mentono dovrebbero trovare più complesso questo doppio compito, rispetto ai sinceri. Il conseguente spostamento di risorse cognitive e attentive nello svolgimento del secondo compito invece che sul racconto, denoterebbe la necessità dell'individuo che il proprio racconto venga creduto, e di conseguenza l'identificazione del bugiardo (Vrij et al., 2008).

Richiesta di più informazioni

Chiedere agli intervistati di fornire maggiori informazioni sul fatto di interesse può aiutare a discriminare coloro che mentono da chi dice la verità. Tendenzialmente, il racconto di un individuo viene considerato tanto più credibile tanti più dettagli possiede (Bell & Loftus, 1989) anche se la qualità dei dettagli influenza tale credibilità. Ciò nonostante, la richiesta di fornire più informazioni risulta efficace perché da una parte

facilita coloro che dicono la verità a produrre più dettagli, e dall'altra mette in difficoltà i mentitori in quanto sono meno propensi ad aggiungere informazioni che potrebbero fornire indizi agli investigatori e svelare le proprie bugie (Vrij et al., 2017).

Domande inaspettate e domande sui processi mentali

La letteratura ha ormai dimostrato più volte il fatto che i bugiardi si preparino alle interviste cercando di prevedere alcuni possibili quesiti (Granhag et al., 2004); tuttavia, tale strategia risulta utile solo nella misura in cui tali soggetti riescano ad anticipare correttamente le domande che effettivamente verranno poste loro (Vrij et al., 2017). Di conseguenza un modo per prenderli alla sprovvista, aumentando quindi le risorse cognitive necessarie per rispondere, è quello di utilizzare domande inaspettate.

Una domanda viene considerata inaspettata quando prevede una risposta automatica e abbastanza rapida nella persona sincera e una risposta ragionata nel caso del bugiardo, avendo come ulteriore caratteristica il fatto che il soggetto mentitore non possa rifugiarsi nella risposta "non ricordo" oppure "non so" (Sartori, 2021). Tali domande devono inoltre riguardare parti centrali dell'evento di interesse e non secondarie (Vrij et al., 2017).

Di fronte ad un quesito del genere quindi, se da una parte l'individuo che dice la verità tenderà ad avere un carico cognitivo simile sia nei confronti delle domande attese che delle domande inaspettate, nei confronti di quest'ultime il bugiardo necessiterà di maggiori risorse cognitive che si manifesterà in tempi di risposta più elevati o in un maggior numero di risposte "non so" (Vrij et al., 2009; Vrij et al., 2017). Nei casi di falsa identità, un esempio di domanda inaspettata è quello di chiedere il proprio segno zodiacale (Sartori, 2021): l'onesto risponderà automaticamente a questa domanda, il

mentitore che invece finge di essere qualcun altro, per rispondere correttamente, dovrà ricavare il segno zodiacale a partire dalla data di nascita.

Un'ulteriore tipologia di domanda inaspettata riguarda l'esplorazione dei processi mentali, ossia delle fasi intermedie che conducono a una specifica conclusione, fatto o azione (Sartori, 2021). Chi mente, infatti, tende a fornire una grande quantità di dettagli riguardo all'evento in questione, mentre risulta notevolmente meno dettagliato quando viene chiamato a descrivere i processi mentali che hanno condotto a quel determinato evento.

2.2.5 *Verifiability Approach*

Il *Verifiability Approach* (VA) si basa sul presupposto che quando un individuo vuole mentire si trova automaticamente davanti ad un bivio, in quanto da una parte pensa che fornire un gran numero di dettagli circa la propria versione aumenterà la sua credibilità, ma dall'altra deve porre estrema attenzione al fatto che tali dettagli non possano essere verificati; di conseguenza, come spiegato da Nahari (2018), i mentitori tenderebbero a dare un numero di dettagli non verificabili simile a quello fornito dalle persone sincere, ma con meno informazioni controllabili. Al contrario, le persone che dicono la verità tendono ad essere più aperti e a fornire sia dettagli verificabili che non (Strömwall et al., 2006).

Per definizione, un dettaglio si dice verificabile quando la sua veridicità è potenzialmente controllabile; in particolare si fa riferimento a:

- attività svolte con persone nominate o identificabili in base alla descrizione fornita che possono successivamente essere chiamate come testimoni;

- attività a cui hanno assistito persone nominate o identificabili in base alla descrizione fornita e che possono successivamente essere chiamate come testimoni;
- attività che sono state documentate (moduli di registrazione, carte di credito o debito, telefoni cellulari), registrate o che l'intervistato ritiene possano essere state riprese da telecamere a circuito chiuso, tuttavia tale possibilità dovrebbe essere menzionata dal soggetto stesso.

Se la persona non menziona esplicitamente una documentazione ma si è sicuri che ci sia e lo sa anche lei, si può considerare come verificabile (ad esempio il fatto di essere andati in un negozio in cui si può pagare solo con la carta di credito); inoltre, se esiste una ragionevole possibilità di rintracciare il testimone o conoscente citato dall'individuo, è sufficiente considerarlo identificabile.

Data questa definizione, i dettagli verificabili sono generalmente di natura percettivo-sensoriale, spazio-temporale, o si riferiscono a persone o azioni compiute (Nahari et al., 2014); invece, i dettagli cognitivi ed emotivi sono considerati per loro natura non verificabili (Bogaard et al., 2019).

La validità del *Verifiability Approach* è stata dimostrata più volte (Vrij et al., 2016; Harvey et al., 2017) e anche in contesti diversi, come quello aeroportuale (Jupe et al., 2017). Inoltre, nella dettagliata meta-analisi effettuata da Palena e collaboratori (2021) è stato riscontrato come l'applicazione del protocollo informativo, ovvero rendere a conoscenza i partecipanti del fatto che le loro dichiarazioni sarebbero state valutate sulla base della verificabilità dei dettagli riportati, porti a rendere ancora più evidente il divario tra persone sincere e individui che mentono.

Un ulteriore studio in grado di dimostrare la rilevanza teorica ed empirica del *Verifiability Approach* è quello di Verschuere e collaboratori (2023) in cui è stata esaminata l'applicazione dell'euristica *use-the-best (and ignore-the-rest)*, un principio decisionale che suggerisce di fare affidamento sull'informazione più disponibile o rilevante, ignorando o dando poca importanza alle altre informazioni meno significative (Gigerenzer & Gaissmaier, 2011).

Gli autori hanno chiesto ai partecipanti di esprimersi circa la veridicità o meno di dichiarazioni scritte a mano oneste e ingannevoli, trascrizioni video, interviste videoregistrate o dal vivo. In particolare, nella condizione di controllo le persone hanno espresso il proprio giudizio in modo intuitivo basando la propria decisione su uno o più indizi scelti liberamente; nella condizione sperimentale invece, gli autori hanno invitato i soggetti ad utilizzare un singolo criterio, ovvero l'accuratezza e la verificabilità dei dettagli presenti nei testi. I risultati mostrano che in quest'ultimo caso i partecipanti riuscivano, con una buona accuratezza a discriminare tra bugia e verità (59-79%), mentre quando potevano utilizzare qualsiasi segnale o erano guidati a utilizzare segnali multipli le loro prestazioni erano vicine al livello del caso (50%). Tali risultati sono rimasti stabili anche quando confrontati con altre tecniche di *lie detection* come l'imposizione di un carico cognitivo o la richiesta di dare più informazioni.

Di seguito (*Tabella 2.1*) viene riportato un esempio di decodifica attuata tramite l'applicazione del *Verifiability Approach*.

After I left the Lab I went to the toilet that is situated in the groundflour of the library. That is on the same level as the cafeteria. Because I haven't eaten anything for breakfast yet I looked through the food that is offered at the cafeteria. I never went there so I was a bit overwhelmed at first. After checking whether or not they have vegan offers (which they do at some of the stands) I chose a tempeh sandwich at tashas (not sure about the name anymore since it was my first time eating there). The person at the counter said it is going to take 5 minutes to prepare the sandwich and in that time we chatted a little bit about tempeh and about veganism. After she gave me the sandwich I went to the cashier and tried to talk to them in English but they weren't good at it.	
After *I left* *the lab* *I went* *to the toilet* *that is situated in the groundflour of the library. That is on the same level as the cafeteria.* *Because I haven't eaten anything for breakfast yet* *I looked through the food that is offered* *at the cafeteria.* *I never went there so I was a bit overwhelmed at first.* *After checking whether or not they have vegan offers* *(which they do at some of the stands)* *I chose a tempeh sandwich* *at tashas* *(not sure about the name anymore since it was my first time eating there).* *The person at the counter said* *it is going to take 5 minutes to prepare the sandwich* *and in that time* *we chatted a little bit about tempeh and about veganism.* *After* *she gave me the sandwich* *I went to the cashier and* *tried to talk to them in English but they weren't good at it.*	 V V V V V V V V V

Tabella 2.1*: Esemplificazione di una decodifica effettuata tramite il Verifiability Approach; vengono indicati con "V" i dettagli considerati verificabili (adattata da Bogaard et al., 2019).*

In conclusione del paragrafo, viene presentata una tabella riassuntiva (*Tabella 2.2*) che mette a confronto le tecniche di intervista per la valutazione della credibilità di un racconto descritte finora.

SUE	**VA**	**CCA**	**RI**	**CI**
I metodi utilizzano cluster di variabili?				
No	Sì	Sì	Sì	Sì
Il metodo è standardizzato?				
Sì, in termini di procedura	Sì, in termini di procedura e domande poste	Sì, in termini di procedura e domande poste	Sì, in termini di procedura e domande poste	Sì, in termini di procedura e domande poste
Esiste un supporto empirico per il metodo?				
d = 1,06 (coerenza affermazione-prova)	g = 0,42 (dettagli verificabili) g = 0,80 (dettagli verificabili dopo l'implementa-zione del protocollo informativo)	*d* = 42; osservatori: accuratezza del 48% nelle interviste standard e del 60% nelle interviste CCA (76% se gli osservatori sono a conoscenza degli indizi)	Precisione del 75%	Aumenta del 25% la qualità dell'informa-zione rispetto all'uso di tecniche tradizionali di intervista
Il metodo è ampiamente studiato?				
Il campo è dominato da Granhag, Hartwig e colleghi	Il campo è dominato da da Nahari, Vrij e colleghi	Il campo è dominato da Vrij e colleghi	Il campo è dominato da Colwell e colleghi	Il campo è dominato da Geiselman, Fisher e colleghi
Sono necessarie prove indipendenti?				
Sì	Sì	No	No	No
Il metodo è stato testato su culture non WEIRD[5]?				
No	No	Sì, in parte	No	Dato non chiaro

[5] WEIRD: Western, Educated, Industrialized, Rich, Democratic.

Il metodo viene utilizzato dai professionisti?				
Sì	Sì	Sì	Sì	Sì
Il metodo è facile da apprendere e utilizzare per i professionisti?				
No	No	No	No	No
Viene fornita una motivazione di fondo per cui il metodo dovrebbe funzionare?				
Sinceri: sono disponibili *Bugiardi*: evitano di riportare prove incriminanti	Sinceri: sono disponibili *Bugiardi*: evitano di riportare prove incriminanti	Sinceri: sono disponibili *Bugiardi*: mantengono le storie semplici	Sinceri: sono disponibili *Bugiardi*: mantengono le storie semplici	Sinceri: sono disponibili *Bugiardi*: mantengono le storie semplici
Può essere utilizzato sempre?				
Solo quando sono disponibili le prove	Solo quando le prove sono potenzialmente disponibili	Sì	Sì	Sì
Quando dovrebbe essere utilizzato?				
Quando le prove sono disponibili	Quando le prove sono potenzialmente disponibili	Quando non è possibile ottenere prove	Quando non è possibile ottenere prove	Quando non è possibile ottenere prove

Tabella 2.2*: Confronto tra SUE, VA, CCA, RI e CI (adattato da Vrij et al., 2022).*

2.3 Tecniche di valutazione della credibilità di un ricordo autobiografico

2.3.1 *Symptom Validity Assessment*

La *Symptom Validity Assessment* (SVA) è una metodologia di indagine per la determinazione della validità di una testimonianza (Mazzoni & Ambrosio, 2003). Essa si compone di tre fasi, un'intervista semi-strutturata utile sia per raccogliere informazioni circa il fatto di interesse che per valutare le capacità cognitive e linguistiche dell'individuo, soprattutto nel caso di un minore, *Criteria Based Content Analysis* (CBCA) e *Validity Checklist* (Gulotta, 2011).

Il CBCA rappresenta un insieme di 19 criteri che sono stati identificati dalla letteratura come in grado di discriminare qualitativamente tra ricordi genuini e fittizi (Sartori, 2021) in quanto più facilmente riscontrabili all'interno di dichiarazioni sincere (Vrij, 2018); in altre parole, i racconti veritieri rispettano un numero maggiore di criteri rispetto a quelli menzogneri (Vrij, 2016). Inizialmente sviluppato per indagare la credibilità dei minori nei casi di abusi sessuali, il CBCA può essere applicato indifferentemente dal tipo di procedimento e anche alle dichiarazioni di adulti (Amado et al., 2016). I criteri sono divisi in cinque aree (*Tabella 2.3*) e viene attribuito un punteggio di 0 nel caso in cui il criterio non sia presente, 1 se è parzialmente presente e 2 se è del tutto presente; i casi confermati, ovvero validi, si distribuiscono tra i 16 e i 34 punti, mentre quelli dubbiosi tra 0 e 10 (Raskin & Esplin, 1991).

Categorie	Criteri
Caratteristiche generali	Struttura logica
	Produzione non strutturata
	Quantità di dettagli
Contenuti specifici	Inserimento nel contesto
	Interazioni
	Riproduzione di conversazioni
	Complicazioni inattese
Particolarità di contenuto	Dettagli inusuali
	Dettagli superflui
	Dettagli riportati accuratamente ma fraintesi
	Associazioni esterne
	Descrizione dello stato mentale soggettivo
	Attribuzione all'accusato di uno stato mentale
Contenuti motivazionali	Correzioni o aggiunte spontanee
	Ammissioni di lacune di memoria
	Manifestazione di dubbi sulla propria testimonianza
	Autoaccuse
	Perdono dell'accusato
Elementi dell'offesa	Particolari caratteristici rispetto al crimine

***Tabella 2.3**: I 19 criteri CBCA (adattata da Gulotta, 2011).*

Alcune caratteristiche relative all'assegnazione dei punteggi sono le seguenti (De Leo et al., 2005): un criterio non è considerato soddisfatto se la risposta è data a una domanda diretta; ogni affermazione può rispondere a più parametri contemporaneamente; la ripetizione dello stesso elemento all'interno del racconto non incrementa la valutazione della presenza di un criterio.

La validità di questa metodologia è stata accertata, ad esempio, nella meta-analisi svolta da Oberlader (2019), il quale ha dimostrato che il CBCA è in grado di distinguere tra affermazioni basate sull'esperienza e quelle inventate, o nello studio di Roma e collaboratori (2011). In relazione alla *lie detection*, il CBCA, come descritto da Vrij (2008a), in ben 16 studi su 20 presenta punteggi più alti negli individui sinceri rispetto ai bugiardi; in particolare, il terzo criterio, ovvero la quantità di dettagli, è risultato centrale.

Tuttavia, dato che il punteggio attribuito al CBCA può essere condizionato da fattori esterni alla testimonianza, come ad esempio errori commessi dagli stessi valutatori, occorre determinare in che misura tali elementi abbiano influenzato il racconto attraverso l'uso della *Validity Checklist* (Vrij, 2018), la quale fa riferimento alle caratteristiche psicologiche dell'individuo, a quelle relative allo svolgimento dell'intervista, ai fattori motivazionali e agli aspetti investigativi (Gulotta, 2011); nel dettaglio (De Leo et al., 2005):

- caratteristiche psicologiche;
- linguaggio e conoscenze non appropriate;
- inadeguatezza delle emozioni;
- suggestionabilità;

- caratteristiche dell'intervista;
- domande suggestive, veicolanti o coercitive;
- inadeguatezza globale dell'intervista;
- motivazione (motivo per il quale è stata sporta la denuncia, motivazioni relative alle rivelazioni originali, pressione a rilasciare l'accusa);
- domande investigative (coerenza con l'ordine delle cose, con altri resoconti e con altre prove).

Tale metodologia possiede però delle criticità, soprattutto in termini di concordanza tra valutatori diversi: prima Gumbert e colleghi (1999) e poi Vrij (2008), hanno infatti dimostrato come il problema maggiore sia il mancato accordo sull'impatto che i fattori appartenenti alla *Validity Checklist* hanno sulla dichiarazione dell'individuo.

Inoltre, nella meta-analisi svolta da Hazlett (2006), appare chiaro come i criteri della SVA, da soli, non siano abbastanza affidabili e validi per essere utilizzati come tecnica di *lie detection*.

2.3.2 *Reality Monitoring*

Il punto centrale del funzionamento del *Reality Monitoring* (RM) è il fatto che i ricordi di eventi passati esperiti in prima persona, ovvero di origine esterna, differiscano qualitativamente da quelli immaginati, di origine interna. Nello specifico, dato che sotto un certo punto di vista anche i ricordi creati attraverso l'uso dell'immaginazione non possono essere considerati meno reali di quelli derivanti da esperienze vissute direttamente, con il termine *real* si fa riferimento a cose che esistono al di fuori del sé.

Johnson e Raye (1981) descrivono come il *Reality Monitoring* si basi su due momenti distinti all'interno del meccanismo di elaborazione dell'informazione: la codifica e il monitoraggio della realtà. La codifica della realtà di un evento fa riferimento alle informazioni caratterizzanti il ricordo, seguendo la logica secondo cui un avvenimento reale, differentemente da uno immaginato, presenterebbe un maggiore quantitativo di dettagli percettivo-contestuali e un numero ridotto di operazioni cognitive (Sporer, 2004). Il monitoraggio della realtà si riferisce invece alla riattivazione delle informazioni percettivo-contestuali presenti nella memoria a lungo termine e l'intervento di processi di natura decisionale che permettono di monitorare il reale avvenimento dell'evento (Mammarella & Di Domenico, 2011).

Secondo Masip e colleghi (2005), esisterebbero otto criteri alla base del funzionamento del *Reality Monitoring* (*Tabella 2.4*): mentre i primi sette prevarrebbero nelle dichiarazioni veritiere, le operazioni cognitive sarebbero maggiormente presenti all'interno delle bugie.

Criteri RM	Definizione	Truth/Lie
Chiarezza del ricordo	Vividezza e nitidezza del ricordo	*Truth*: i ricordi reali tendono ad essere più dettagliati e chiari rispetto a quelli immaginati, che possono risultare vaghi e poco definiti
Aspetti sensoriali	Dettagli derivanti dai cinque sensi: vista, udito, olfatto, tatto e gusto	*Truth*: i ricordi di eventi reali contengono più descrizioni sensoriali rispetto ai ricordi immaginati, che spesso mancano di tali dettagli o li possiedono in misura minore
Informazioni spaziali	Localizzazione spaziale degli eventi, disposizione degli oggetti, posizione di persone, configurazione ambientale	*Truth*: i ricordi reali forniscono una descrizione più coerente e precisa dello spazio rispetto a quelli immaginati
Informazioni temporali	Sequenza temporale degli eventi, inclusi l'ordine e la durata delle azioni	*Truth*: i ricordi reali tendono ad avere una sequenza temporale chiara e logica, mentre i ricordi immaginati possono presentare incongruenze o essere vaghi riguardo al tempo
Aspetti emotivi	Emozioni vissute durante l'evento di interesse	*Truth*: i ricordi reali sono spesso associati a emozioni genuine e specifiche, mentre i ricordi immaginati possono esserne deficitari, avere emozioni meno intense o più generiche
Ricostruibilità della memoria	Riguarda la coerenza e la struttura narrativa del ricordo	*Truth*: i ricordi reali presentano una sequenza logica e coerente degli eventi, i ricordi immaginati possono contenere elementi disordinati o incongruenti
Realismo	Plausibilità e verosimiglianza del ricordo	*Truth*: i ricordi reali appaiono più realistici e credibili, mentre i ricordi immaginati possono includere elementi improbabili o fantasiosi
Operazioni cognitive	Elaborazioni cognitive, inferenze e ragionamenti	*Lie*: i ricordi reali appaiono meno ricchi di operazioni cognitive, al contrario di quelli immaginati

***Tabella 2.4**: I criteri del Reality Monitoring (adattato da Masip et al., 2005; Bogaard et al., 2019).*

Con lo scopo di offrire una panoramica generale sull'argomento, vengono riportati di seguito i principali criteri utilizzati nella procedura di decodifica secondo il *Reality Monitoring* (Nahari et al., 2014; Elntib & Wagstaff, 2017; Bogaard et al., 2019):

- dettagli percettivo-sensoriali: possono fare riferimento a suoni, odori, gusti, sensazioni fisiche o informazioni visive (ad esempio "Mi ha detto che l'esame era difficile" o "Il caffè che ho bevuto era dolce");
- dettagli spaziali: vengono comprese le informazioni che collegano l'evento a luoghi o contesti particolari, azioni di entrata e uscita, la disposizione di persone e/o oggetti e direzioni (ad esempio "Mi trovavo nella biblioteca" o "Il libro era sullo scaffale di destra");
- dettagli temporali: quando è accaduto l'evento, la sua durata, avverbi temporali e la sequenza di avvenimenti (ad esempio "Era mattina presto" o "Siamo rimasti fino alle 17:00");
- dettagli emotivi: resoconti di stati mentali soggettivi (ad esempio "Mi ha fatto inorridire");
- dettagli cognitivi: ragionamenti, operazioni cognitive, inferenze e supposizioni (ad esempio "Sembrava abbastanza intelligente");
- azioni (ad esempio "Stavo camminando").

Tuttavia, dare per scontato che i ricordi generati sulla base di immaginazioni interne e non esperienze percettive esterne corrispondano ad una distorsione volontaria della realtà, con conseguente produzione di una bugia, può essere un'inferenza discutibile. Inoltre, le previsioni del *Reality Monitoring* nel discriminare tra dichiarazioni sincere e mentite possono essere influenzate da diverse variabili, la prima delle quali è rappresentata dalla modalità di presentazione: in generale, gli studi che hanno preso in

considerazione eventi accaduti in prima persona ai partecipanti mostrano maggiore sostegno per l'approccio RM, tuttavia, nessuno dei criteri descritti precedentemente è stato utile nel discriminare dichiarazioni veritiere e mentite dopo che gli individui avevano assistito ad un filmato (Masip et al., 2005).

La seconda variabile è la preparazione che hanno a diposizione i partecipanti nel ripresentare il ricordo dopo pochi minuti o alcuni giorni: da una parte Sporer (1997) ha dimostrato che nella condizione in cui i soggetti effettuavano un'intervista pochi minuti dopo il completamento della prima, ovvero la condizione immediata, le informazioni di tipo contestuale (temporali e spaziali) erano maggiormente presenti nei racconti sinceri, mentre le informazioni di tipo senso-percettivo non erano associate in modo specifico né alle dichiarazioni veritiere che a quelle mentite; al contrario, Alonso-Quecuty (1990) ha riscontrato che i dettagli sensoriali e contestuali erano maggiormente presenti nelle dichiarazioni sincere immediate, ma non in quelle ritardate di dieci minuti.

La terza variabile è costituita dalle dichiarazioni ripetute: mentre nello studio di Alonso-Quecuty e Hernández-Fernaud (1997) la ripetizione ha avuto conseguenze sulla presenza di informazioni contestuali e sensoriali aumentandone la frequenza, Granhag e colleghi (2001) non hanno riscontrato nessuna influenza sui criteri RM. L'età dei soggetti presi in considerazione rappresenta la quarta variabile: i ricercatori hanno osservato non solo che la capacità dei criteri RM nel discriminare tra verità e bugia può variare tra adulti e bambini (Sporer, 1997; Alonso-Quecuty, 1996), ma anche che si può osservare una correlazione positiva tra l'aumento della presenza di alcuni criteri e quello dell'età, indipendentemente dalla veridicità dell'affermazione (Santtila et al., 1998). Infine, nell'interpretazione dei risultati finali devono essere prese in considerazione anche le differenze individuali che caratterizzano ogni individuo.

I numerosi risultati contradditori appena presentati trovano in parte una spiegazione in due problematiche principali dello *scoring* effettuato tramite RM: il punteggio attribuito ai criteri, nonché la loro operazionalizzazione.

Il funzionamento di questa metodologia si basa infatti sulla suddivisone della trascrizione del ricordo autobiografico dell'individuo in singoli segmenti che vengono poi codificati attribuendo un'etichetta in base alla tipologia dell'informazione fornita in modo da distinguere un ricordo esperito in prima persona da uno immaginato. Tuttavia, nonostante la logica alla base del *Reality Monitoring* sia approvata dagli studiosi del settore, i parametri che lo definiscono posseggono ancora una condivisione instabile (Vrij, 2000).

A differenza del CBCA che è caratterizzato da un numero ben definito e specificato di criteri su cui basare la propria valutazione, il *Reality Monitoring* è stato utilizzato in modo diversificato da più autori, i quali hanno applicato definizioni e operazionalizzazioni diverse (Vrij, 2000). Una delle problematiche principali di questo approccio è quindi rappresentata dalla mancanza di definizioni condivise dei criteri da utilizzare, soprattutto per quanto riguarda le operazioni cognitive; tale miglioramento potrebbe risolvere, almeno in parte, i risultati contradditori visti in precedenza (Masip et al., 2005).

Per quanto riguarda il valore attribuito ai criteri, anche qui le differenze sono molte. Esistono infatti diverse metodologie per calcolare il punteggio RM, tra cui il conteggio delle volte in cui determinati criteri sono presenti in un racconto, ovvero la loro frequenza, contando ad esempio quante volte vengono menzionati i dettagli sensoriali, temporali o spaziali, la valutazione della presenza o assenza di un determinato criterio attraverso una metodologia binaria e l'utilizzo di una scala Likert con punteggi che

vanno da 1 a 5 con la quale determinare la rilevanza dei criteri RM in un racconto (Johnson & Raye, 1981; Sporer, 1997). Il punteggio finale viene poi calcolato sommando i valori rilevati, attribuendo se necessario pesi differenti ai criteri in base alla loro rilevanza; Vrij e colleghi (2004) sottolineano inoltre l'importanza, all'interno di comparazioni tra diversi testi, di rapportare tale punteggio con il numero totale delle parole, in modo da diminuire il più possibile le interferenze.

Tutte le problematiche appena descritte hanno come conseguenza un'ulteriore criticità, ovvero la mancanza di un *cut-off*. Tale questione presenta numerose ripercussioni, come la possibilità di interpretare le risposte secondo modalità differenti e di conseguenza attribuire un certo grado di attendibilità ad un racconto creando così dei possibili falsi positivi; ciò avrebbe come conseguenza la difficoltà nel confrontare più studi e valutazioni tra di loro, portando in parte alle discordanze presentate precedentemente. La mancanza di un *cut-off* nell'applicazione del *Reality Monitoring* può inoltre portare alla confusione tra memorie derivanti da eventi esperiti personalmente e ricordi che derivano da semplici ragionamenti e pensieri, identificando in questo modo una memoria inaffidabile (Johnson & Raye, 1981).

Per quanto riguarda l'accuratezza di questa metodologia, Masip e collaboratori (2005) hanno riscontrato che, nei dieci studi presi in considerazione in cui è stata utilizzata la tecnica RM per distinguere sinceri e bugiardi, la percentuale corrispondeva al 69%, con un ruolo fondamentale svolto dalle informazioni contestuali e percettive nel mettere in risalto il racconto veritiero. Sporer (1997) ha invece ottenuto una capacità discriminativa totale del 75% per i racconti veri e del 67,5% per quelli falsi. In sintesi si può quindi affermare che il *Reality Monitoring*, come strumento di *lie detection*, sia efficace (Vrij et al., 2015).

L'accuratezza di questa metodologia risulta ancora più incredibile se confrontata con altre, come ad esempio il CBCA. Una delle principali critiche che viene fatta al CBCA/SVA è la mancanza di un fondamento teorico: la tecnica è stata infatti creata ad hoc in maniera induttiva (*bottom-up*) partendo da esperienze concrete di bambini abusati; il *Reality Monitoring* invece si basa sulle solide fondamenta della teoria della memoria (*top-down*) argomentata da Johnson e Raye (1981) che permette di distinguere se un ricordo è il prodotto di una fonte interna o esterna.

Un'altra importante distinzione è che l'approccio RM presenta un numero minore di criteri rispetto al CBCA, rendendolo più semplice sia da utilizzare che da insegnare (Tommasino et al., 2008); inoltre, mentre il *Criteria Based Content Analysis* è formato da fattori che permettono solamente di identificare elementi di verità nel racconto, il *Reality Monitoring* possiede anche dei criteri, come le operazioni cognitive, che consentono di distinguere gli indizi relativi alla menzogna (Vrij et al., 2004).

Per quanto riguarda la capacità discriminativa dei due strumenti, dallo studio condotto da Vrij e collaboratori (2004) è emerso che, soltanto con l'uso del CBCA, è stato classificato correttamente il 60% dei racconti, mentre con l'utilizzo esclusivo dell'approccio RM si è arrivati ad un 74%. Calcolando invece i punteggi totali mettendo assieme le due tecniche, la capacità di distinguere correttamente un evento falso da uno vero arriva al 78,8% (Sporer, 1997).

Sporer (1997) ha inoltre effettuato uno studio in cui ha confrontato l'efficacia di CBCA e RM nel distinguere tra resoconti veritieri e inventati forniti da individui adulti dopo aver visionato un video, cercando di identificare possibili somiglianze tra i due strumenti. Tramite un'analisi fattoriale, l'autore ha identificato cinque fattori di base (*Tabella 2.5*).

Fattori di base	CBCA	RM
Consistenza logica e realismo	Dettagli inusuali (legato negativamente)	Ricostruibilità della storia
Quantità di dettagli e riferimenti contestuali	Dettagli superficiali e inusuali	Informazioni spaziali e temporali
Pensieri ed emozioni	Riferimenti ai propri stati mentali	Affetti e operazioni cognitive
Chiarezza	Produzione non strutturata (legato negativamente)	Chiarezza
Interazioni verbali e non verbali	Riproduzione di conversazioni e descrizione di interazioni, descrizione di stati mentali dell'accusato	Nessun criterio corrispondente

Tabella 2.5*: Cinque fattori di base individuati da Sporer (1997).*

In entrambi gli strumenti permane tuttavia il bisogno di definizioni più chiare e precise dei criteri da utilizzare; nel caso dell'approccio RM, ciò potrebbe tuttavia risultare più semplice grazie alla solida teoria che presenta alla base (Tommasino et al., 2008). Di seguito (*Tabella 2.6*) viene proposto un esempio di applicazione del *Reality Monitoring*.

First, I went back to my desk to get my UVA card. I saw that Lindy was also at the office so I said hi. Then I went to the ABC building to get myself a cup of tea and eat a mandarin. While I was drinking my tea I realised that I didn't have my UVA card anymore so I went back to the coffee place to look for it. I found it. Then I sent my friend Marie a text message.	
First,	T
I went	
back	
to my desk	S
to get my UvA card.	P
I saw that Lindy was also	P
at the office	
so I said hi.	P
Then	T
I went	
to the ABC building	S
to get myself a cup of tea and	P
eat a mandarin.	P
While	T
I was drinking my tea	P
I realised that I didn't have my UvA card anymore so	
I went	
back to the coffee place	S
to look for it.	P
I found it.	P
Then	T
I sent (a text message)	P
my friend Marie	P
a text message.	

***Tabella 2.6**: Esemplificazione di una decodifica effettuata tramite il Reality Monitoring; vengono indicati con "S" i dettagli spaziali, "T" i dettagli temporali e "P" i dettagli percettivi (adattata da Bogaard et al., 2019).*

2.3.3 *Scientific Content Analysis*

La *Scientific Content Analysis* (SCAN) è una tecnica di *lie detection* sviluppata da Avinoam Sapir, ex esaminatore poligrafico, che consiste in una stesura scritta da parte dell'esaminando delle attività svolte durante l'evento di interesse all'interno di un determinato range temporale; tale testo viene poi valutato da un esperto grazie all'utilizzo di una serie di criteri che ricordano in parte quelli del CBCA (Sapir, 1987). Secondo l'autore stesso, l'uso di riferimenti a sé stessi oppure la negazione delle accuse subite sono esempi di elementi caratterizzanti i racconti veritieri, mentre la mancanza di determinate informazioni è maggiormente attribuibile ad una dichiarazione falsa (Vrij et al., 2015).

Tuttavia, in uno studio svolto da Nahari e collaboratori (2012) in cui hanno preso in considerazione 61 partecipanti, lo SCAN, a differenza del *Reality Monitoring* (71%), non è stato in grado di discriminare tra menzogneri e sinceri.

In conclusione, nonostante paradossalmente risulti uno degli strumenti maggiormente utilizzati nell'ambito della *lie detection* (Vrij, 2018), lo SCAN non ha un'accuratezza empirica dimostrata.

2.3.4 aIAT

Un'altra tecnica ampiamente validata per la valutazione della credibilità di un ricordo autobiografico è il *Test di Associazione Implicita Autobiografica* (*autobiographical Implicit Association Test*, aIAT; Sartori et al., 2008). Lo IAT rappresenta una variante del tradizionale *Implicit Association Test* (IAT; Greenwald et al., 1998), comunemente utilizzato per la misura degli atteggiamenti impliciti, e viene impiegato per verificare la

presenza della traccia mnestica di un presunto ricordo autobiografico dell'individuo (Agosta et al., 2011).

Lo aIAT si basa sull'assunto che l'individuo colpevole impiegherà un tempo maggiore per reagire ai target critici rispetto ai dettagli non critici, a cui risponde negativamente; ciò accade in quanto il soggetto è costretto a sopprimere la risposta autentica, più immediata e spontanea, ma non in linea con le istruzioni del compito (Suchotzki, 2018).

La struttura dello aIAT è la seguente: esso consiste in un compito computerizzato in cui vengono presentati degli stimoli al soggetto, il quale li deve categorizzare premendo due tasti del computer; tali stimoli appartengono a quattro possibili categorie (Agosta & Sartori, 2013):

a. frasi sempre vere nel momento in cui l'individuo svolge il test (ad esempio "sono davanti ad un computer");
b. frasi sempre false nel momento in cui l'individuo svolge il test (ad esempio "sto scalando una montagna");
c. prima versione dell'evento autobiografico di interesse (ad esempio "sono andato a Parigi per Natale"), che in ambito forense potrebbero riferirsi alla tesi dell'accusa;
d. versione alternativa dell'evento autobiografico di interesse (ad esempio "sono andato a New York per Natale"), che in ambito forense potrebbero riferirsi alla tesi della difesa.

Come descritto nel libro *Neuropsicologia forense* (Stracciari, Bianchi & Sartori, 2010), lo aIAT si suddivide in cinque blocchi, tre di categorizzazione semplice (il primo, il secondo e il quarto) e due di categorizzazione combinata (il terzo e il quinto): nel primo si richiede al soggetto di categorizzare semplicemente frasi sempre vere e frasi sempre

false con due tasti diversi, mentre nel secondo frasi relative alla tesi della difesa e frasi relative all'accusa; nel terzo blocco vengono invece categorizzate con lo stesso tasto frasi sempre vere assieme alle frasi della difesa e con un altro tasto frasi sempre false insieme alle frasi dell'accusa; nel quarto blocco si chiede nuovamente all'individuo di categorizzare le frasi dell'accusa e le frasi della difesa, invertendo però i tasti rispetto al blocco due; infine, si richiede la categorizzazione combinata tra frasi sempre vere e frasi dell'accusa con un tasto e frasi sempre false e frasi della difesa con l'altro, invertito quindi rispetto al terzo blocco (*Figura 2.1*).

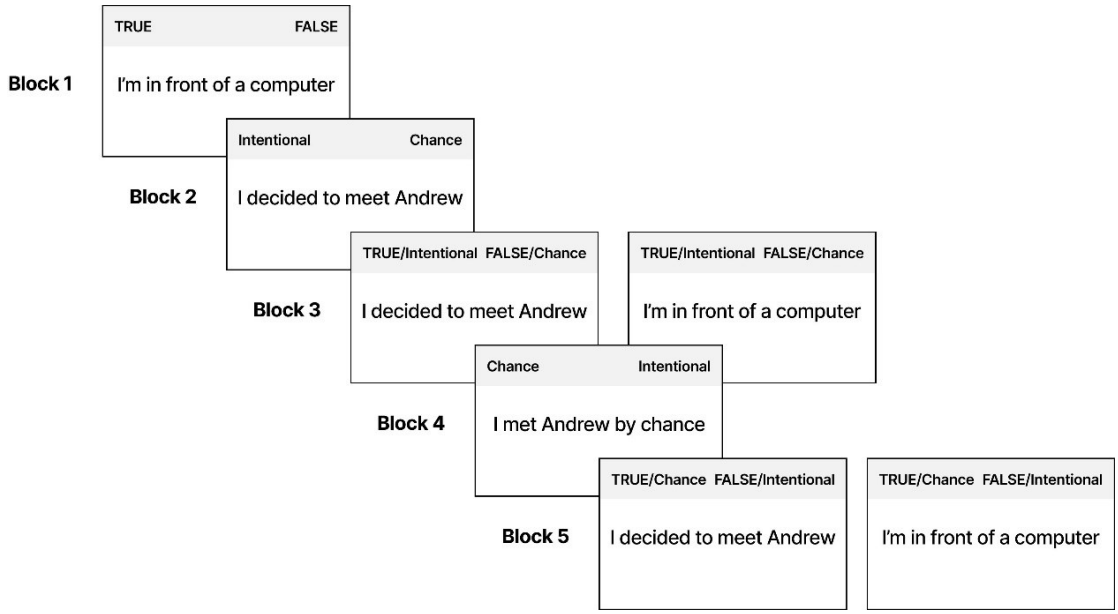

***Figura 2.1**: Esemplificazione della procedura sperimentale dello aIAT (adattata da Zangrossi et al., 2015)*

La logica da seguire è la seguente: le frasi che verranno categorizzate più velocemente, quindi con tempi di risposta più brevi, quando condividono la stessa risposta motoria con le frasi sempre vere, saranno quelle a cui corrisponde la traccia di memoria autobiografica (Sartori et al., 2018).

Come strumento di detezione della simulazione o della menzogna, lo aIAT presenta alcuni vantaggi, tra cui l'uso di tecnologie abbastanza semplici e reperibili, il fatto che possa essere somministrato rapidamente nel giro di 10-15 minuti, e anche a distanza, e il fatto che non preveda nessuna particolare formazione per l'utente (Agosta & Sartori, 2013). La validazione dello strumento è stata dimostrata da diversi studi di cui si può trovare un elenco dettagliato nell'articolo di Agosta e Sartori (2013), raggiungendo un'accuratezza del 92% (Stracciari et al., 2010). Un esempio presente in letteratura è quello proposto da Agosta e collaboratori (2013) in cui, attraverso l'uso dello aIAT, non solo è stato possibile discriminare tra soggetti sinceri e individui che producevano bugie bianche, ma anche distinguere la veridicità delle ragioni sottostanti.

Uno degli ulteriori vantaggi dello aIAT è la possibilità di generare stimoli specifici per il caso di interesse; proprio per questo motivo, nel corso del tempo sono nati IAT che riguardano diverse applicazioni (Sartori, 2021), come quello legato al doping (Petroczi et al., 2011), alla pedofilia (Babchishin et al., 2014), alla previsione di comportamenti suicidi (Nock & Banaji, 2007) o al pregiudizio raziale (Nosek et al., 2007).

CAPITOLO 3: VERSO UN'ANALISI AUTOMATICA DEI RICORDI AUTOBIOGRAFICI

3.1 Il problema dell'analisi testuale nello studio dei ricordi autobiografici

All'interno del mondo psicologico, in particolare quello cognitivo e forense, il processamento del linguaggio assume una particolare importanza. Negli ultimi anni gli studi che hanno indagato l'applicazione di modelli automatici per questo scopo si sono moltiplicati a dismisura e tale diffusione non è sorprendente se si pensa all'utilità che un esito positivo di tali sperimentazioni avrebbe nel risolvere alcune delle maggiori problematiche presenti all'interno dell'analisi linguistica, ovvero l'enorme richiesta di tempo, principalmente a causa dello svolgimento dello *scoring* manuale, il costo elevato e soprattutto l'accordo tra intervistatori diversi (*interrater agreement*). Uno dei tentativi dell'impiego di sistemi automatizzati nel processamento del linguaggio è dunque la sostituzione della codifica effettuata da esseri umani con quella algoritmica, più veloce e affidabile, in modo da ottenere un'estrazione delle caratteristiche linguistiche (Kleinberg et al., 2018).

In particolare, la concordanza tra valutatori risulta complicata. Essa rappresenta una misura di quanto siano d'accordo due o più osservatori nel fare le proprie valutazioni (Gisev et al., 2013); tale concetto assume un ruolo decisivo in quei contesti in cui più persone sono chiamate a fornire un giudizio indipendente e soggettivo, come appunto l'analisi linguistica. Un alto livello di *interrater agreement* è quindi indice non solo di una buona affidabilità dei valutatori, ma anche di un certo grado di validità in quanto se gli osservatori sono d'accordo, allora c'è una maggiore sicurezza nel fatto che la

variabile presa in considerazione vada effettivamente a valutare il costrutto che intende misurare.

Nei paragrafi successivi vengono descritte alcune delle componenti chiave utili per la determinazione dell'*interrater agreement*.

3.1.1 *Intraclass Correlation Coefficient*

L'*Intraclass Correlation Coefficient* (ICC) è una misura statistica utilizzata per valutare il grado di concordanza o similarità tra unità del gruppo rispetto a una o più variabili, risultando particolarmente utile per misurare l'affidabilità di valutazioni ripetute o eseguite da diversi valutatori. Il valore dell'ICC varia da 0 a 1, dove 0 indica nessuna concordanza e 1 una concordanza perfetta: di conseguenza, più l'indice si avvicina a 1 e maggiore sarà l'omogeneità delle misure all'interno dei gruppi rispetto a quelle tra i gruppi (Koo & Li, 2016).

Esistono due tipi di ICC: il *Single Rater/Measurement ICC*, utilizzato quando si valuta l'affidabilità di singole misurazioni, e l'*Average Rater/Measurement ICC*, per quando si vuole calcolare la media delle misurazioni o valutazioni di più osservatori diversi. L'ICC viene calcolato come la proporzione della varianza totale attribuibile alla varianza tra i gruppi (*Figura 3.1*).

$$ICC = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

***Figura 3.1**: Formula dell'ICC, dove $\sigma^2_{between}$ è la varianza tra i gruppi, mentre σ^2_{within} è la varianza all'interno dei gruppi.*

In termini di affidabilità, alcuni autori (Shrout & Fleiss, 1979; Cicchetti, 1994) hanno fornito delle linee guida per interpretare i valori di questo indice (*Tabella 3.1*).

ICC	Affidabilità
< 0.40	Scarsa
0.40 – 0.59	Discreta
0.60 – 0.74	Buona
≥ 0.75	Eccellente

***Tabella 3.1**: Interpretazione dell'Intraclass Correlation Coefficient.*

3.1.2 *Cohen's kappa*

Il *Cohen's kappa* (κ) è una misura statistica che viene utilizzata per valutare il grado di concordanza tra due osservatori indipendenti nello svolgimento di valutazioni qualitative; essa è particolarmente utilizzata quando si vuole determinare se la concordanza osservata assume un valore più alto di quello che ci si aspetterebbe basandosi sul caso. Il suo valore varia da -1 a 1, dove 1 indica una concordanza perfetta, 0 una concordanza equivalente a quella attesa per il caso, mentre i valori negativi sono indicatori di una discordanza peggiore di quella che ci si aspetterebbe per puro caso. La formula per calcolare questo indice è riportata nella *Figura 3.2*.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

***Figura 3.2**: Formula del Cohen's kappa, dove P_o è la proporzione di concordanza osservata e P_e la proporzione di concordanza attesa per caso.*

Secondo l'interpretazione adottata da Landis e Koch (1977), la concordanza calcolata secondo l'utilizzo del *Cohen's kappa* assume le seguenti caratteristiche (*Tabella 3.2*).

κ	Concordanza
< 0	Scarsa
0.01 – 0.20	Leggera
0.21 – 0.40	Discreta
0.41 – 0.60	Moderata
0.61 – 0.80	Buona
0.81 – 1.00	Eccellente

***Tabella 3.2**: Interpretazione del Cohen's kappa.*

3.1.3 *Fleiss's kappa*

Il *Fleiss's kappa* (κ) è una misura statistica utilizzata per valutare il grado di concordanza tra più esaminatori nel classificare le unità in categorie distinte; tuttavia, a differenza del *Cohen's kappa* che è limitato a solo due valutatori, il *Fleiss's kappa* è in grado di gestire situazioni in cui il numero di quest'ultimi è maggiore. Come nel caso precedente, anche questo indice assume valori che vanno da -1 a 1, dove quest'ultimo rappresenta la concordanza perfetta, lo 0 una concordanza equivalente a quella del caso, mentre i valori negativi una discordanza peggiore di quello dovuta al puro caso.

La formula e l'interpretazione del *Fleiss's kappa* corrispondono a quelle del *Cohen's kappa*.

3.1.4 *Krippendorff's alpha*

Il *Krippendorff's alpha* (α) è una misura statistica utilizzata per valutare l'affidabilità di più valutatori nel classificare unità diverse in categorie. Una caratteristiche particolare di questo indice è che non solo può essere applicato a tipologie di dati differenti, come nominali, ordinali, intervalli e rapporti, ma è anche in grado di gestire un qualsiasi numero di giudizi indipendenti, compreso il caso in cui non tutti i soggetti valutino ogni unità.

Come descritto dallo stesso autore (Krippendorff, 2018), l'interpretazione del valore del *Krippendorff's alpha* è equivalente a quella del *Cohen's kappa* e del *Fleiss's kappa*, mentre il suo calcolo viene effettuato seguendo la formula riportata in seguito (*Figura 3.3*).

$$\alpha = 1 - \frac{D_o}{D_e}$$

***Figura 3.3**: Formula del Krippendorff's alpha, dove D_o è la discordanza osservata e D_e la discordanza attesa dal caso.*

3.1.5 *F1 score*

L'indice *F1 score* (detto anche *F-score* o *F-measure*) è una misura utilizzata in statistica e nel campo del *machine learning* per valutare la precisione di un modello di classificazione binaria. Esso può essere utilizzato anche per valutare la concordanza tra due persone, in particolare quando queste devono valutare un insieme di dati, considerando la classificazione come "positiva" nel caso della presenza di una determinata condizione e "negativa" come la sua assenza.

L'indice F1 risulta particolarmente utile quando si vuole determinare l'equilibrio tra precisione (*precision*) e richiamo (*recall*): la prima si calcola come il numero di veri positivi, ovvero i casi correttamente identificati come positivi da entrambe le persone, diviso per il totale dei risultati positivi attribuiti da un solo individuo; il secondo invece, fa riferimento al numero di veri positivi diviso per il totale dei casi che avrebbero dovuto essere identificati come positivi, cioè la somma di veri positivi e falsi negativi (*Figura 3.4*).

$$\text{Precision} = \frac{\text{Vero Positivi (TP)}}{\text{Vero Positivi (TP)} + \text{Falsi Positivi (FP)}}$$

$$\text{Recall} = \frac{\text{Vero Positivi (TP)}}{\text{Vero Positivi (TP)} + \text{Falsi Negativi (FN)}}$$

***Figura 3.4**: Formule per calcolare precision e recall in riferimento all'indice F1.*

L'indice F1 risulta quindi dalla media armonica tra *precision* e *recall*, preferita a quella aritmetica in quanto richiede che entrambi i valori precedentemente descritti siano alti (*Figura 3.5*).

$$F1 = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

***Figura 3.5**: Formula per calcolare l'indice F1.*

Il valore dell'indice F1 varia da 0 a 1 in cui un valore alto indica una buona concordanza tra le due persone nella loro classificazione, mentre un punteggio basso suggerisce una scarsa concordanza (Derczynski, 2016).

3.2 Il *Natural Language Processing* per un'analisi automatizzata

La capacità di linguaggio è una caratteristica distintiva degli esseri umani che emerge nei primi anni di vita e continua ad evolversi nel corso dell'età adulta, consentendo loro di esprimersi e comunicare in modi sempre più complessi e sofisticati (Hauser et al., 2002). La creazione di sistemi automatizzati in grado elaborare enormi quantità di dati linguistici in tempi rapidi offre quindi l'opportunità di esplorare e comprendere meglio le complessità del linguaggio umano; tuttavia, risulta fondamentale affrontare anche le

sfide etiche e garantire lo sviluppo responsabile di tali tecnologie per massimizzarne i benefici senza compromettere al contempo la sicurezza e la privacy. La costruzione di un sistema di intelligenza artificiale che permetta di comprendere, interpretare e generare il linguaggio naturale, mimando in qualche modo l'interazione con un essere umano reale, risulta quindi tanto importante quanto difficoltosa.

Il primo passo in questa direzione è stato fatto grazie allo sviluppo del *Natural Language Processing* (NLP), in italiano *Elaborazione del Linguaggio Naturale*: esso viene definito come un sistema, nato dalla convergenza di diversi campi tra cui l'informatica, la linguistica, la statistica e la psicologia cognitiva, che impiega una serie di metodologie computazionali per trasformare vasti insiemi di dati linguistici in linguaggio informatico al fine di affrontare una varietà di compiti che richiedono comprensione e generazione di linguaggio naturale, simili a quelli eseguiti dagli esseri umani (Kang et al., 2020). Tuttavia, se tale modello è stato in grado da una parte di identificare le relazioni a breve termine tra diverse parole, dall'altra ha manifestato delle difficoltà nel lungo termine, specialmente nel caso di testi molto lunghi o in cui erano presenti due proposizioni all'interno della medesima frase (Sartori & Orrù, 2023).

Quindi, nonostante il *Natural Language Processing* sia nato in ambito informatico come strumento di analisi linguistica, il suo impiego nel campo della psicologia si sta diffondendo sempre di più, sia per quanto riguarda la sua abilità nella generazione del linguaggio sia, soprattutto, per la sua capacità di comprenderlo. Nei paragrafi successivi vengono presentate alcune tecniche di NLP e il loro impiego all'interno di studi di stampo psicologico e forense.

3.2.1 LIWC

Il LIWC (*Linguistic Inquiry and Word Count*) è un software di analisi testuale che consente di categorizzare il linguaggio scritto in diverse dimensioni psicologiche e linguistiche (Pennebaker et al., 2007; Tausczik & Pennebaker, 2010).

Tale processo è reso possibile grazie all'uso di un dizionario predefinito contenente dimensioni psicologiche (ad esempio emozioni come "tristezza" o "noia"), sociali (ad esempio "famiglia" o "amici") o cognitive (come ad esempio il pensiero analitico o l'insicurezza). Quando un testo viene inserito nel *software*, quest'ultimo confronta ogni parola del testo con il proprio dizionario classificandole in base alle categorie di appartenenza, permettendo quindi di quantificare la frequenza di utilizzo di termini appartenenti a ciascuna categoria. Al termine di tale procedimento, il *software* restituisce l'*output* sottoforma di percentuale di parole che rientrano in ciascuna categoria. Ciò risulta particolarmente utile per comprendere non solo il contenuto emotivo e cognitivo del testo proposto, ma anche eventuali pattern linguistici che potrebbero essere indicativi di determinate caratteristiche di personalità o stati d'animo del soggetto.

Poiché alcune tecniche di analisi del contenuto del testo hanno individuato marcatori linguistici che potrebbero essere utilizzati come indicatori di menzogna o di non attendibilità come ad esempio il *Reality Monitoring*, alcuni studi hanno cercato di applicare tali categorie al funzionamento del LIWC.

Bond e Lee (2005) ad esempio, hanno fatto uso di alcune categorie LIWC per codificare parole veritiere e ingannevoli nel linguaggio dei prigionieri utilizzando anche i modelli del *Reality Monitoring*; i risultati mostrano come i tassi di classificazione ottenuti dall'analisi di regressione logistica utilizzando alcune categorie LIWC fossero dei

predittori significativamente migliori rispetto sia all'accuratezza degli osservatori umani che del caso.

Tuttavia, altri risultati sembrano meno incoraggianti. Da una parte alcuni studi hanno riscontrato esiti opposti pur utilizzando le medesime categorie di parole; dall'altra alcune ricerche hanno ottenuto come risultato il fatto che determinate classi di parole fossero in grado di distinguere la verità dalla bugia, ma tale risultato è stato disconfermato da studi addizionali (Bond & Lee, 2005; Vrij et al., 2007). In aggiunta, altre ricerche hanno ottenuto esiti che contraddicono le teorie di base della *lie detection*, ovvero che le bugie conterrebbero un numero significativamente più elevato di parole rispetto alle dichiarazioni veritiere (Hancock, 2007).

In seguito a tali esiti, l'uso del LIWC e delle sue categorie come strumento di identificazione della veridicità e dell'attendibilità intrinseca di una dichiarazione dovrebbe essere maggiormente studiato e approfondito in quanto attualmente risulta essere fonte di diverse criticità (Masip et al., 2012).

3.2.2 *Sentiment analysis*

Il *Sentiment analysis*, o *analisi del sentiment*, è una tecnica utilizzata per determinare la polarità dell'atteggiamento dell'autore di un testo, il quale può essere classificato come positivo, negativo o neutro, attraverso l'elaborazione del linguaggio naturale (Onyenwe et al., 2020).

Tale strumento può essere applicato in molti ambiti diversi come il *marketing*, per monitorare le opinioni dei clienti su determinati prodotti attraverso l'analisi delle loro recensioni, la politica, per analizzare le opinioni pubbliche relative a candidati e partiti, la psicologia, ad esempio per dare supporto ai professionisti della salute mentale

nell'identificazione delle condizioni dei pazienti e delle loro emozioni o anche semplicemente per la detezione delle *fake news* (Alonso et al., 2021).

Tuttavia, nonostante i termini *analisi del sentiment* e *analisi delle emozioni* siano spesso usati in modo interscambiabile, il loro significato differisce: mentre il primo ha come obiettivo un'analisi puramente soggettiva della polarità del trascritto, il secondo rappresenta una misura più oggettiva e precisa dello stato psicologico ed emotivo del soggetto.

Quindi, nonostante l'utilità che questo strumento possiede, sono ancora molte le sfide che deve affrontare, tra cui: l'ambiguità del linguaggio, ovvero il problema che parole e frasi possano assumere significati diversi a seconda del contesto; il fatto che la maggior parte delle risorse è disponibile solamente il lingua inglese; la difficoltà nel comprendere l'ironia e il sarcasmo, i quali possono essere in grado di invertire la polarità di un'affermazione; le differenze culturali e linguistiche che possono tradursi in discrepanze nell'uso delle espressioni emotive; l'uso del nuovo *slang* e di una grammatica scorretta (come ad esempio "*u*" invece di "*you*") (Nandwani & Verma, 2021); la rilevazione della soggettività, ovvero la distinzione tra affermazioni fattuali e opinabili (per un approfondimento di questo specifico punto, fare riferimento a Chaturvedi e collaboratori (2018), in cui gli autori hanno cercato di identificare contenuti ingannevoli concentrandosi su affermazioni soggettive e cariche di sentimento piuttosto che su quelle neutrali).

3.2.3 *POS tagging*

Il *POS tagging* (*Part-Of-Speech tagging*), o assegnazione delle categorie grammaticali, rappresenta il processo di attribuzione di etichette grammaticali a ogni parola all'interno di un testo, come verbi, aggettivi o sostantivi; tale passaggio risulta fondamentale all'interno dell'elaborazione del linguaggio naturale in quanto facilita l'analisi sintattica e semantica del testo, migliorando compiti come la traduzione automatica (Martinez, 2012).

Le component principali del *POS tagging* sono il *corpus* annotato, cioè un *dataset* di test precedentemente etichettato con le categorie grammaticali, il quale viene poi utilizzato per l'addestramento e la valutazione dei modelli automatici, e il *tagset*, ovvero un insieme di etichette grammaticali standardizzate utilizzate per annotare le parole (Chiche & Yitagesu, 2022). Tra quelli disponibili, uno dei più utilizzati è il *Penn Treebank tagset* (*Figura 3.6*) che contiene 45 etichette (Marcus et al., 1993).

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+, %, €*
CD	Cardinal number	*one, two*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, uh, oops*
EX	Existential 'there"	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past particip.	*eaten*
JJR	Adj. comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj. superlative	*biggest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1,2,3*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, singular/mass	*dog, snow*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*dogs*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singul.	*Marco*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Alps*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	"
POS	Possessive ending	*'s*	"	Right quote	"
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	*([{ <*
PP$	Possessive pronoun	*my, your*	)	Right parenthesis	*)] } >*
RB	Adverb	*never, often*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final pun	*. ! ?*
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punt.	*: ; … -*
RP	Particle	*up, on ,off*			

***Figura 3.6**: Penn Treebank tagset.*

Nel processo di categorizzazione, i *tag* vengono in genere aggiunti alla fine della parola dopo "/" (*Figura 3.7*).

The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

***Figura 3.7**: Esemplificazione di una decodifica con POS tagging.*

Nello studio di Soldner e colleghi (2019, June) gli autori hanno sviluppato un *dataset* multimodale contenente conversazioni tra partecipanti che hanno giocato al gioco *Box of Lies*, in cui illustravano un oggetto nascosto e gli altri individui dovevano capire se la descrizione fosse vera o falsa; tale giudizio veniva poi confrontato con quello di modelli automatici di classificazione della menzogna. Tra le metodologie utilizzate, sono stati estratti anche i *tag* POS per ogni parola nelle trascrizioni (come ad esempio nomi, verbi, aggettivi) e tali informazioni sono state poi utilizzate come *features* aggiuntive per migliorare la capacità dei modelli di apprendimento automatico nel distinguere tra comportamenti veritieri e menzogneri. I risultati mostrano che l'inclusione delle caratteristiche dialogiche ha migliorato le prestazioni di classificazione, superando sia il giudizio umano che la classificazione casuale.

3.2.4 *Named Entity Recognition*

La *Named Entity Recognition* (NER), in italiano *riconoscimento di entità denominate*, è una parte integrante del *Natural Language Processing* e rappresenta un metodo volto all'individuazione e alla categorizzazione di entità specifiche all'interno di testi definiti che variano a seconda dell'algoritmo utilizzato, come nomi di persone, luoghi, enti

organizzativi, valute, unità di misura e altro ancora, che può essere sfruttata per risolvere domande, sintetizzare contenuti testuali o condurre operazioni di traduzione (Li et al., 2020). Di seguito (*Figura 3.8*) viene riportato un esempio di codifica effettuata tramite l'uso della NER.

We met yesterday [*DATA*] at 11:30 am [*ORA*] on Coronado [*ENTITA' GEOPOLITICA*] beach, then went to Starbucks [*ORGANIZZAZIONE*] and paid $3.50 [*QUANTITATIVO MONETARIO*] for a coffee.

***Figura 3.8**: Esemplificazione dell'applicazione della NER (adattato da Kleinberg et al., 2018)*

L'uso della NER prevede innanzitutto un processo di *file annotation* manuale delle entità rilevanti all'interno del testi eseguito da esseri umani necessario per creare *dataset* etichettati; tale procedimento, nonostante richieda molto tempo (Strobl et al., 2022, June), risulta particolarmente utile quando i dati includono annotazioni incomplete o non abbastanza specifiche per lo scopo che si sta perseguendo (Stollenwerk et al., 2023). Per incrementare l'efficienza del processo manuale si possono adottare strategie come l'impiego di modelli già esistenti per effettuare un'annotazione preliminare dei dati, successivamente perfezionata da esperti del settore di riferimento (Sageder & Karampatakis, 2021, September).

L'intervento umano risulta quindi fondamentale per rendere più efficaci le prestazioni di riconoscimento della NER: Komiya e collaboratori (2018) hanno ulteriormente dimostrato come sistemi addestrati su *dataset* etichettati manualmente avevano prestazioni migliori rispetto a quelli che avevano utilizzato banche di dati annotati in modo semiautomatico.

In ambito forense e nello specifico della *verbal lie detection*, la NER ha avuto diverse applicazioni e lo studio di Kleinberg e colleghi (2018) ne rappresenta un ottimo esempio. Gli autori hanno utilizzato il riconoscimento di entità denominate per consolidare ulteriormente tre principi teorici già diffusi, tra cui il fatto che chi dice la verità tende a fornire più dettagli all'interno del proprio racconto, e le logiche alla base del *Reality Monitoring* e del *Verifiability Approach*, ovvero che tali informazioni dovrebbero contenere maggiori riferimenti contestuali e in generale un quantitativo più elevato di dettagli verificabili. Essi hanno quindi proceduto a confrontare la prestazione di due strumenti NER, lo *spaCy* e lo *Standford's N.E.R.*, con una misura della specificità della frase applicata principalmente nell'analisi di titoli di giornali chiamata *speciteller*, e il LIWC (Tausczik & Pennebaker, 2010), per identificare recensioni veritiere e ingannevoli di hotel. I risultati dello studio mostrano come la NER sia stata più efficace nel discriminare dichiarazioni mentite e sincere rispetto agli altri due sistemi, andando ulteriormente a validare la maggior presenza di dettagli specifici come date o menzioni di altre persone in queste ultime, e confermando l'ipotesi iniziale degli autori secondo cui vi sarebbe un numero maggiore di entità denominate nelle affermazioni veritiere rispetto a quelle menzognere.

L'addestramento del modello automatico è proprio ciò che differenzia i due approcci proposti, sia dal punto di vista qualitativo che quantitativo: se da una parte la NER è in grado di apprendere direttamente dai dati già pre-annotati, dall'altra i modelli automatici si possono basare su istruzioni dettagliate fornite dall'utente che definiscono le entità di interesse, ed eventualmente degli esempi di come esse vengano menzionate nei testi di riferimento, in modo da poterle facilmente individuare e comprendendo al

contempo anche il contesto (y Arcas, 2022). Entrambe le tecniche, hanno dunque come *output* lo *scoring* automatico del materiale fornito.

3.2.5 Tecniche di *embedding*

Le tecniche di *embedding* sono dei metodi utilizzati per trasformare immagini o parole in rappresentazioni vettoriali; tali vettori vengono poi utilizzati in diversi modelli di apprendimento automatico permettendo loro di lavorare con dati complessi in modo più efficiente, mantenendo tuttavia intatte le relazioni semantiche e strutturali tra gli oggetti rappresentati (Mikolov et al., 2013).

Uno dei modi in cui tale strumento è stato testato in campo psicologico è attraverso il paradigma *Deese-Roediger-McDermott* (DRM). Esso rappresenta un compito di falsa memoria in cui viene chiesto ai partecipanti di codificare diverse parole (ad esempio "tavolo", "seduto", "divano") che sono inserite in una lista di 15 vocaboli, le quali sono correlate semanticamente ad una parola esca (ad esempio "sedia"); successivamente, in seguito ad un breve compito di distrazione, ai soggetti viene chiesto di indicare se una data parola fa parte o meno degli elenchi presentati in precedenza (Gatti et al., 2024). In generale, durante il compito di riconoscimento, i partecipanti riportano come precedentemente memorizzate un numero abbastanza elevato di parole esche, nonostante esse non siano state presentate, creando così dei falsi ricordi (Gallo, 2010).

Normalmente, le liste di parole utilizzate per lo svolgimento di questo paradigma sono selezionate in modo intuitivo dagli sperimentatori in base alla somiglianza semantica con la parola esca. Gatti e collaboratori (2022) hanno tuttavia sviluppato un metodo più avanzato ed automatico per la creazione di queste liste basandosi su tecniche di *embedding* in grado di rappresentare, attraverso vettori numerici, parole appositamente

selezionate sulla base dell'uso di modelli semantici distribuzionali addestrati su linguaggio naturale e in grado di analizzare la relazione semantica tra di esse: la distanza e la direzione tra i vettori rappresentano le relazioni semantiche tra i vocaboli, di conseguenza parole che hanno significati simili o che appaiono in contesti analoghi avranno vettori vicini tra di loro (*Figura 3.9*).

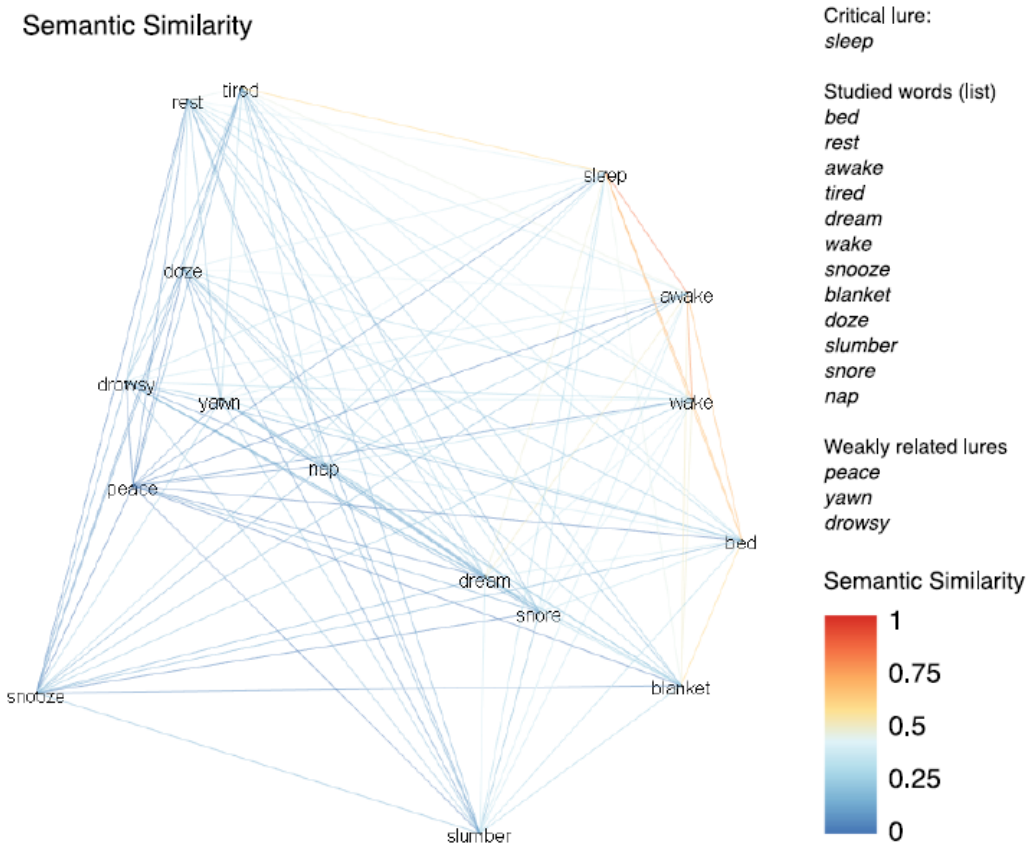

***Figura 3.9**: Proiezione bidimensionale della struttura di somiglianza semantica tra le parole che compongono la lista "sleep"; i colori più caldi rappresentano una maggiore somiglianza semantica (Gatti et al., 2022).*

La novità introdotta da Gatti e colleghi (2022) risulta fondamentale su tre punti differenti. In primo luogo permette di avere una maggiore precisione nella selezione delle parole da inserire nelle liste: l'uso di modelli semantici distribuzionali consente di identificare vocaboli che sono maggiormente in grado di generare falsi ricordi poiché si

basano su analisi quantitative piuttosto che sull'intuizione soggettiva. In secondo luogo il metodo basato sui dati risulta più sistematico e replicabile, migliorando in questo modo la consistenza degli esperimenti e la capacità di confrontare i risultati tra diversi studi. In terzo luogo questo nuovo approccio consente una miglior comprensione dei processi cognitivi coinvolti nella formazione delle false memorie evidenziando come la somiglianza semantica influenzi direttamente questi processi.

3.3 *Large Language Models* e il loro impiego in ambito psicologico

Per capire come si è riusciti ad arrivate agli attuali modelli di intelligenza artificiale, risulta tuttavia essenziale delineare un passaggio fondamentale oltre all'associazione di ogni parola ad un vettore distinto, ovvero l'integrazione dell'autoattenzione (*self-attention*). Questo termine si riferisce alla capacità del modello di rappresentare le parole come vettori di contesto, i quali vengono poi confrontati in modo da attribuire un'importanza relativa a ciascuno di loro; ciò permette di assegnare pesi diversi alle varie parti dell'*input* di testo in base alla rilevanza posseduta per il compito richiesto.

Tale componente si è rivelata essere fondamentale per lo sviluppo di famosi *Transformer* (Vaswani et al., 2017), i quali combinano il meccanismo di *self-attention* con classici *layer* lineari e connessioni residuali tipici del *deep learning*, una branca dell'intelligenza artificiale che consente alle reti neurali di apprendere complesse rappresentazioni di dati attraverso una gerarchia di livelli di astrazione.

Nel tempo, si è scoperto inoltre che l'ampliazione della dimensione dei modelli avrebbe potuto portare ad un miglioramento nelle loro prestazioni (Zhao et al., 2023); per questo motivo si è deciso di sviluppare dei modelli linguistici pre-addestrati su un ampio *set* di dati, chiamati *Large Language Models* (LLMs).

Gli attuali LLMs vengono definiti come un tipo di modello di intelligenza artificiale addestrato sulla base di enormi raccolte di dati di linguaggio naturale e di testo al fine di sviluppare la capacità di comprendere e generare linguaggio umano in modo naturale e coerente (Sartori & Orrù, 2023). Tali modelli sono addestrati su compiti generici, noti come *general task*, che coinvolgono una vasta gamma di attività linguistiche, come la comprensione del linguaggio naturale, la traduzione o la generazione di un testo, il riassunto automatico e molto altro. In particolare, dato un nuovo compito e dei nuovi dati mai visti in precedenza, viene utilizzato il termine *fine tuning* per indicare il processo di addestramento di un modello pre-addestrato su un *dataset* diverso in modo da rendere più specializzato il suo *output* (Zhao et al., 2023).

Per quanto riguarda nello specifico l'ambito psicologico, i LLMs possono essere utilizzati come strumenti di ricerca per diversi scopi: riassumere e revisionare in modo sistematico la letteratura presente su un determinato argomento (Van Dis et al., 2023); *peer review* e supporto nella scrittura di documenti accademici (Dergaa et al., 2023); generazione di ipotesi sulla base degli studi già effettuati e assistenza nella progettazione sperimentale (Park et al., 2024); analisi dei dati raccolti (Patel & Fan, 2023).

3.3.1 GPT-4

Nel 2018 *OpenAI*, una società di ricerca sull'intelligenza artificiale con sede negli USA, ha rilasciato sul mercato il *Generative Pre-trained Transformer* (GPT), un modello di *machine learning* progettato per generare testo naturale in modo comprensibile e coerente, addestrato su un ampio *set* di dati (Floridi & Chiriatti, 2020). Nel 2022 l'azienda ha sviluppato *ChatGPT*, un LLM in grado di fornire delle risposte a domande

e richieste fornite dall'utente. Grazie al veloce sviluppo tecnologico, negli anni successivi sono state rilasciate versioni aggiornate del modello fino ad arrivare alla più recente, ovvero GPT-4, proposta agli utenti il 14 marzo 2023.

GPT-4 accetta *prompt*, ovvero *input* forniti dall'utente, sia in formato testo che immagine, il che permette all'individuo di rendere esplicito qualsiasi compito sia di tipo visivo che linguistico; inoltre, nel fornire il proprio *output*, GPT-4 mostra capacità simili a quelle che ha con *input* puramente testuali anche in risposta a immagini e fotografie. GPT-4, rispetto alle sue versioni precedenti, riduce significativamente gli errori di ragionamento e le allucinazioni, ovvero risposte date dall'IA con sufficiente fiducia ma non derivanti dai dati con i quali è stata addestrata (Ji et al., 2023) (*Figura 3.10*).

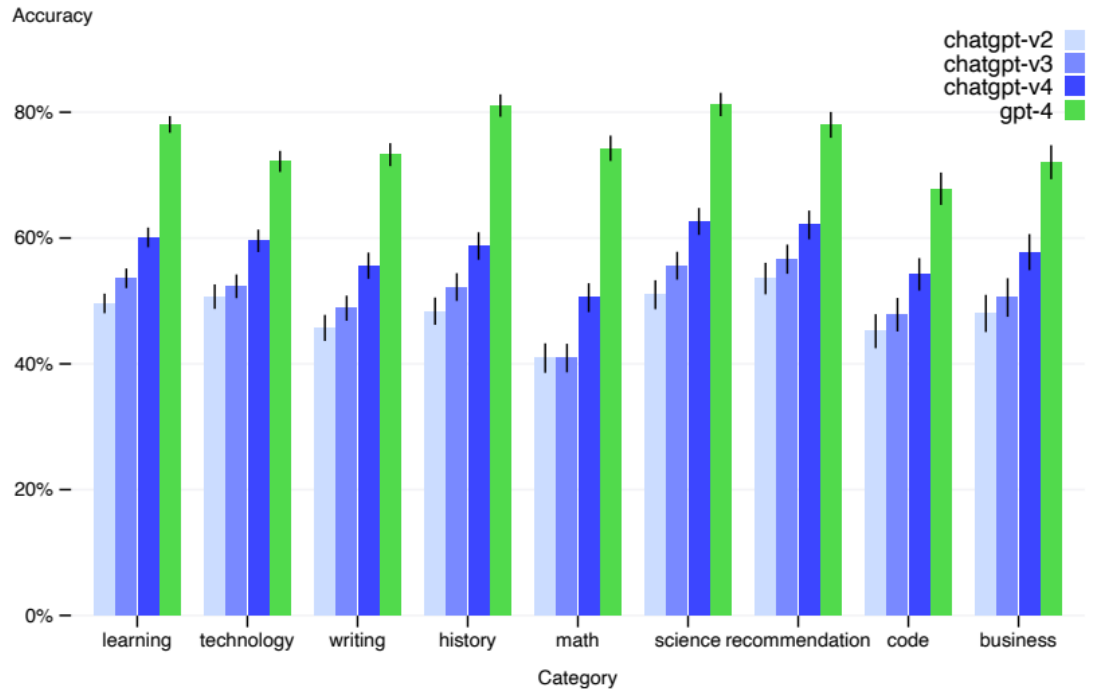

***Figura 3.10**: Prestazioni di GPT-4 su nove valutazioni di factualità interne; l'accuratezza, rappresentata sull'asse delle ordinate, è tanto migliore quanto più elevato è il suo valore, per cui un'accuratezza di 1.0 significa che le risposte del modello sono giudicate in accordo con le risposte ideali umane per tutte le domande della valutazione (Achiam et al., 2023).*

Nel complesso, nonostante le problematiche ancora presenti come la comprensione limitata del contesto o i limiti nella generazione creativa, GPT-4 rappresenta un significativo passo in avanti verso sistemi di intelligenza artificiale utili alla ricerca (Achiam et al., 2023), superando anche i più comuni test cognitivi portando ad un risultato mai visto prima (Dhingra et al., 2023), e pertanto è stato selezionato quale sistema automatico da adoperare nell'ambito del progetto sperimentale oggetto del presente elaborato.

Con l'avvento di GPT-4 e di altri modelli linguistici avanzati quindi, le tecniche di NLP viste in precedenza possono essere applicate semplicemente tramite le tecniche di *prompting*, ovvero istruzioni fornite al modello utilizzando il linguaggio naturale, tra cui *zero-shot* (chiedere al modello di svolgere un compito senza fornire esempi, come "Traduci questa frase dall'italiano all'inglese"), *few-shot* (fornire al modello alcuni esempi di un compito prima di chiedere di eseguire un'azione semplice, come "Ecco due esempi di frasi tradotte, ora traduci questa frase") e *chain-of-thoughts* (richiedere al modello di spiegare il ragionamento dietro una risposta, passo dopo passo, come "Descrivi come risolveresti questo quesito matematico") (Liu et al., 2023).

La vera novità sta quindi nel fatto di poter utilizzare il linguaggio naturale, senza bisogno di codice in *Phyton* o altri linguaggi di programmazione, per ottenere risultati complessi di NLP. Ciò non rende solo la tecnologia maggiormente accessibile ad un pubblico più ampio, ma semplifica anche notevolmente il processo di interazione con i modelli linguistici avanzati.

Tutto ciò appare fondamentale all'interno del presente elaborato. L'applicazione del *Reality Monitoring* e del *Verifiability Approach* risulta infatti un processo dispendioso a livello temporale a causa della necessità dello *scoring* manuale: riuscire ad ottenere gli

stessi risultati da un LLM fornendo solamente istruzioni in linguaggio naturale senza necessità di programmazione, rappresenterebbe un passo avanti enorme.

CAPITOLO 4: ESPERIMENTO 1

L'obiettivo dei due esperimenti che verranno descritti di seguito è stato quello di testare la capacità di un LLM nell'applicazione degli approcci *Reality Monitoring* e *Verifiability Approach* per valutare la credibilità di una serie di ricordi autentici e mentiti, confrontando le sue *performance* con quelle umane.

L'ipotesi dei due studi è che GPT-4, un modello di linguaggio avanzato, possa valutare i resoconti di memoria autentici e ingannevoli con una precisione paragonabile a quella degli umani.

4.1 Materiali e metodi

4.1.1 *Dataset*

Il *dataset* utilizzato nel primo esperimento è stato ricavato da Monaro e collaboratori (2022). Gli autori, con l'obiettivo di testare la capacità di modelli di apprendimento automatico nel distinguere tra persone che mentono e coloro che dicono la verità basandosi sulle microespressioni facciali, hanno analizzato il *set* di dati precedentemente raccolto da Monaro e colleghi (2020).

Tale *dataset* consiste in 62 videoregistrazioni di partecipanti italiani, di cui 19 maschi e 43 femmine con un'età compresa tra i 20 e 29 anni, intervistati riguardo una vacanza passata. A 32 soggetti è stato chiesto di dire la verità, ovvero di raccontare una vacanza reale avvenuta negli ultimi 12-18 mesi, mentre agli altri 30 è stata assegnata la condizione di "bugiardo", cioè dovevano riferire una vacanza finta o che non avevano mai vissuto.

Per evitare che i bugiardi includessero dettagli veritieri nelle loro storie, è stato fornito loro un modulo precompilato contenente informazioni sulla vacanza fittizia che dovevano descrivere. Allo stesso modo, i partecipanti che raccontavano la verità sono stati invitati a tralasciare dettagli su cui erano incerti e ad utilizzare fotografie e video in modo da ridurre la distorsione dei ricordi causata dal passaggio del tempo.

Ogni video era composto da tre fasi: la *baseline*, in cui venivano chiesti i dati anagrafici; il racconto libero, in cui il partecipante ricordava la vacanza per circa 2 minuti; le domande inaspettate, poste in modo da aumentare il carico cognitivo richiesto per narrare la storia, uguali per entrambe le condizioni. La durata media dei video è stata di 9,56 minuti.

L'esperimento 1 del presente elaborato prende quindi in considerazione le trascrizioni di tali videoregistrazioni in lingua italiana.

4.1.2 *Codebook* per il *Reality Monitoring*

È stato inizialmente creato un *codebook*, disponibile in appendice e reso possibile grazie all'unione delle informazioni ricavate da Nahari e collaboratori (2014), Elntib e Wagstaff (2017) e Bogaard e colleghi (2019), il cui primo passaggio prevede l'applicazione del *Reality Monitoring* con la distinzione di:

- dettagli percettivi: informazioni sulle esperienze sensoriali (suoni, odori, gusti, sensazioni fisiche e dettagli visivi) codificate con "PERC";
- spaziali: informazioni sui luoghi o sulla disposizione di persone e/o oggetti, o che collegano l'evento a luoghi o contesti spaziali, così come azioni di entrata ed uscita, direzioni o verbi di movimento, codificate con "SPACE";

- temporali: informazioni circa quando è accaduto il fatto di interesse e la sua durata, sequenze di eventi e avverbi temporali, codificate con “TIME”;
- affettivi: resoconti di sensazioni emotive, codificati con “AFFECT”;
- cognitivi: evidenze nelle narrazioni di varie attività cognitive, come pensieri o ragionamenti e supposizioni cognitive di esperienze sensoriali, includendo anche le descrizioni delle inferenze fatte dal partecipante al momento dell'evento, codificate con “COG”.

Nel documento è specificato che un dettaglio viene considerato tale quando aggiunge nuove informazioni che non possono essere estratte da altre, di conseguenza mentre “l’albero verde” verrà considerato come un dettaglio unico, “la casa blu” come due distinti; inoltre, l’elenco di persone presenti all’evento (ad esempio “mio padre, mia madre..”) viene contato come un dettaglio solo. Un’ulteriore considerazione da fare, è che l’annotazione dei dettagli è stata fatta prendendo in considerazione solamente l’evento autobiografico centrale del racconto, il quale non sempre corrispondeva con l’intero testo; di conseguenza alcune frasi, essendo puramente contestuali, non sono state codificate.

4.1.3 Istruzioni per il *Verifiability Approach*

Il secondo passaggio del *codebook* prevede l’applicazione del *Verifiability Approach* e quindi la determinazione della verificabilità (indicata con “VE”) o della non verificabilità (indicata con “UNVE”) del dettaglio precedentemente individuato.

Un’informazione viene considerata verificabile quando la sua veridicità può essere potenzialmente controllabile: di conseguenze, solo i dettagli percettivi, temporali e

spaziali possono essere verificati mentre quelli cognitivi e affettivi, per definizione, vengono considerati sempre non verificabili.

In particolare, un dettaglio può essere verificato quando si tratta di:

- attività svolte con persone nominate o identificabili in base alla descrizione fornita;
- attività a cui hanno assistito persone nominate o identificabili in base alla descrizione fornita;
- attività che sono state documentate (moduli di registrazione, utilizzo di carte di debito, telefoni cellulari..) registrate o che l'intervistato ritiene possano essere state riprese da telecamere a circuito chiuso (la possibilità di telecamere a circuito chiuso dovrebbe essere esplicitamente menzionata dall'intervistato stesso).

Se la persona non menziona esplicitamente una documentazione ma si è sicuri che ci sia e lo sa anche lei, si può considerare come verificabile (ad esempio un negozio in cui si può pagare solo con la carta di credito); in aggiunta, se esiste anche una ragionevole possibilità di rintracciare la persona nominata, ciò è sufficiente per considerarla identificabile.

Per la creazione delle istruzioni per il *Verifiability Approach* sono stati utilizzati i lavori di Bogaard e colleghi (2019) e Loconte e collaboratori (2023).

4.1.4 Procedura di annotazione umana

È stato inizialmente eseguito un *training* per entrambi gli annotatori umani circa l'utilizzo dei due approcci, consentendogli così di familiarizzare con le tecniche di

annotazione e gli strumenti necessari; in seguito, sono stati istruiti a codificare il 20% dei testi del *dataset* attraverso *Doccano*, uno strumento *open source* di annotazione testuale che permette l'etichettatura di dati linguistici (*Figura 4.1*), applicando le regole del *Reality Monitoring*.

***Figura 4.1**: Schermata esemplificatrice dell'annotazione umana con l'uso di Doccano.*

Dopo aver controllato la loro concordanza circa il numero e la tipologia dei dettagli individuati (ICC_{freq}=0.98; $F1_{RM}$=0.75), le etichette identificate dal primo annotatore sono state utilizzate da entrambi per l'applicazione del *Verifiability Approach*. In seguito alla valutazione dell'accordo tra i due ($F1_{VA}$=1), uno solo dei due annotatori ha proseguito con la codifica del restante 80% del *dataset* tramite l'applicazione di entrambi gli approcci.

4.1.5 Annotazione GPT-4

Per dare le istruzioni al *Large Language Model* è stata utilizzata l'interfaccia API (*Application Programming Interface*; model: gpt-4-turbo; temperature: 0.7; max tokens: 1000) fornendo il *codebook* precedentemente descritto e la trascrizione del testo di riferimento contenente il ricordo autobiografico; l'*output* richiesto era l'identificazione delle etichette del *Reality Monitoring* (*Figura 4.2*).

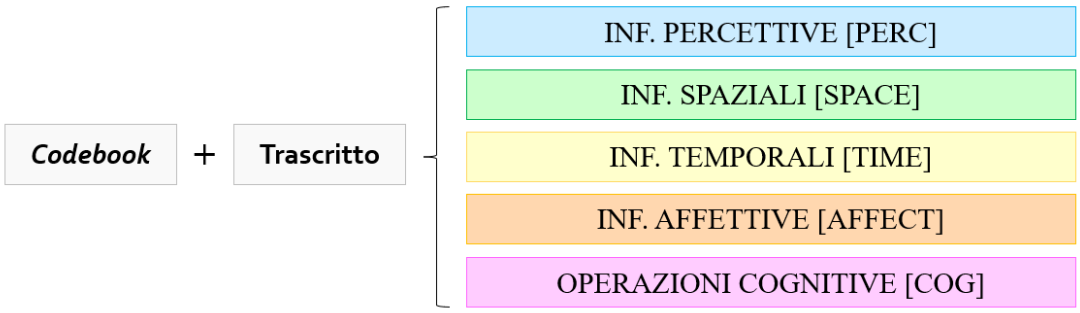

***Figura 4.2**: Schema rappresentativo del materiale fornito a GPT-4 e dell'output richiesto.*

Successivamente, così come per l'annotazione eseguita da umani, è stato chiesto al modello automatico di valutare la verificabilità o meno dei dettagli percettivi, spaziali e temporali precedentemente individuati tramite l'applicazione del *Reality Monitoring*, avendo a disposizione le istruzioni presenti nel *codebook* e il trascritto del testo (*Figura 4.3*).

***Figura 4.3**: Schema rappresentativo del materiale fornito a GPT-4 e dell'output richiesto.*

4.1.6 Piano d'analisi

Le analisi dettagliate che hanno consentito di confrontare la *performance* di GPT-4 con quella degli annotatori umani si sono basate su tre aspetti principali.

Innanzitutto, per quanto riguarda l'applicazione del *Reality Monitoring*, è stato preso in considerazione non solo il numero complessivo di dettagli individuati, ma anche la loro tipologia (*sequence-classification*); ciò ha permesso di valutare la capacità di GPT-4 nel riconoscere e distinguere tra dettagli percettivi, spaziali, temporali, cognitivi e affettivi, fornendo una misura precisa della sua efficacia nel monitoraggio della realtà. Per il *Verifiability Approach* invece, sono state confrontate le etichette poste dal modello automatico e dagli annotatori umani.

4.2 Risultati

4.2.1 *Reality Monitoring*: numero di dettagli

Relativamente al numero di dettagli di tipo percettivo, spazio-temporale, affettivo e cognitivo, rilevati nei testi tramite l'applicazione del *Reality Monitoring*, la concordanza tra *performance* umana e del modello automatico è eccellente (*Tabella 4.1*), con GPT-4 che individua un numero minore di dettagli percettivi, cognitivi e affettivi rispetto al primo annotatore, a fronte di un numero maggiori di dettagli spaziali (*Grafico 4.1*).

Type	Point Estimate	Lower 95% CI	Upper 95% CI
ICC 3,1	**0.952**	0.940	0.962

***Tabella 4.1**: Concordanza tra la performance umana e quella di GPT-4 nel numero di dettagli individuati tramite l'applicazione del Reality Monitoring.*

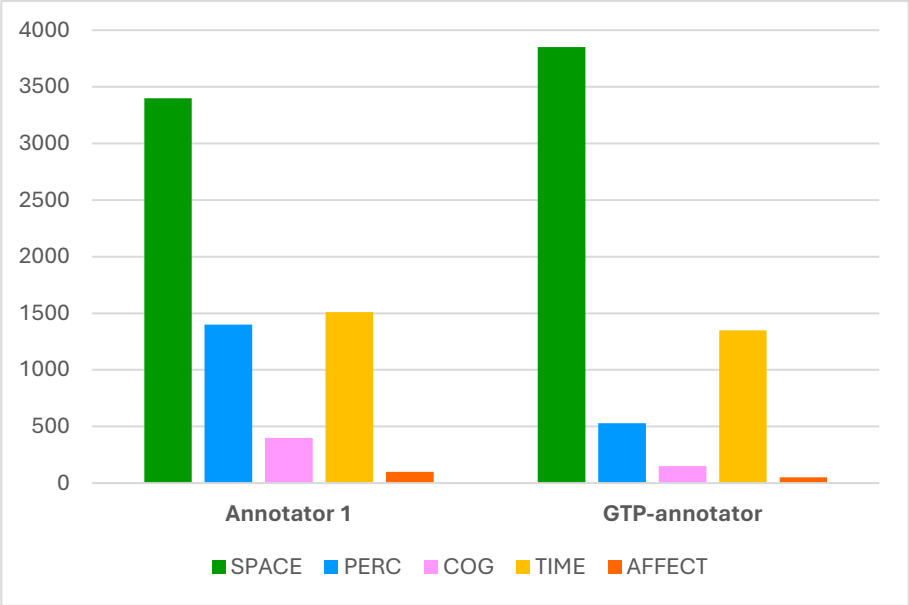

***Grafico 4.1**: Confronto tra il numero di dettagli individuati da GPT-4 e dall'annotatore umano tramite l'applicazione del Reality Monitoring.*

4.2.2 *Reality Monitoring*: *sequence-classification*

Per quanto riguarda la tipologia di dettagli individuati grazie all'applicazione del *Reality Monitoring*, emerge che la concordanza tra la *performance* umana e quelle del modello automatico risulta moderata (*Tabella 4.2*), indicando una discreta capacità del modello GPT-4 di allinearsi con gli esseri umani nell'identificazione e classificazione dei dettagli.

Tag	F1 score	Occurrences
AFFECT	0.04	66
PERC	0.22	1349
TIME	0.44	1546
COG	0.07	369
O	0.74	11252
SPACE	0.51	3349
Overall	**0.61**	**17931**

***Tabella 4.2**: Concordanza tra la performance umana e quella di GPT-4 nella tipologia di dettagli individuati tramite l'applicazione del Reality Monitoring.*

4.2.3 *Verifiability Approach*

In merito alla concordanza nella frequenza di etichettamento ottenuta attraverso l'applicazione del *Verifiability Approach*, emerge un dato contrastante rispetto a quelli precedenti: la concordanza tra la *performance* umana e quella del modello automatico GPT-4 risulta essere molto bassa (*Tabella 4.3*).

Tag	F1 score	Occurrences
VE	0.21	1337
UNVE	0.66	1490
Overall	**0.45**	**2827**

***Tabella 4.3**: Concordanza tra la performance umana e quella di GPT-4 nell'etichettamento dei dettagli tramite l'applicazione del Verifiability Approach.*

Questo risultato evidenzia una discrepanza sostanziale tra la capacità del modello di intelligenza artificiale di etichettare correttamente le informazioni rispetto alla capacità degli esseri umani, sia nell'uso dell'etichetta riferibile alla verificabilità che alla non verificabilità (*Grafico 4.2*).

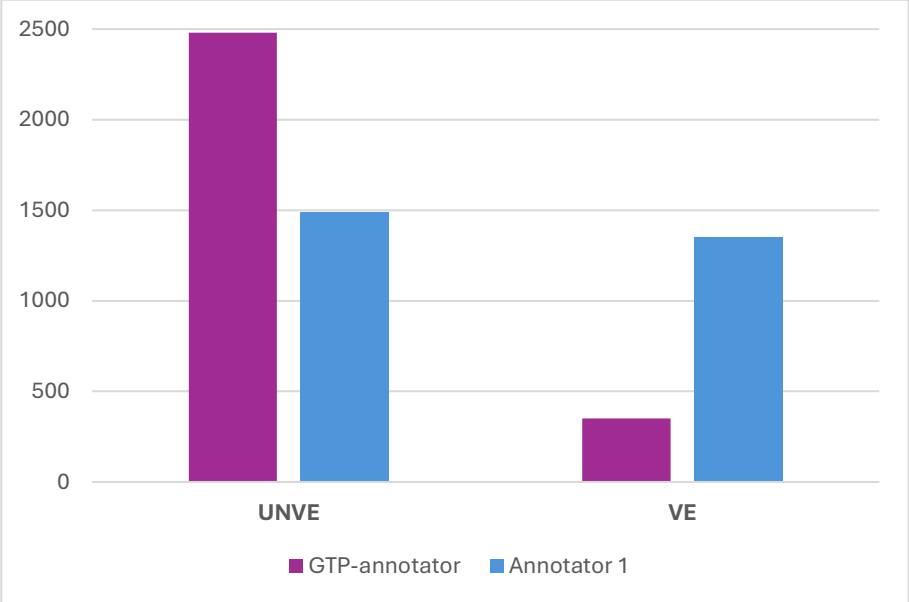

***Grafico 4.2**: Confronto tra il numero di dettagli identificati come verificabili o meno da GPT-4 e dall'annotatore umano tramite l'applicazione del Verifiability Approach.*

CAPITOLO 5: ESPERIMENTO 2

5.1 Materiali e metodi

5.1.1 *Dataset*

Il *dataset* dell'esperimento 2, denominato *Hippocorpus*, fa riferimento ai dati raccolti da Sap e colleghi (2022): nel loro studio, gli autori hanno messo insieme 6.854 racconti in lingua inglese simili a diari che contenevano esperienze di vita importanti, per poi selezionare un sottoinsieme di 240 storie distinte equamente in tre fasi.

La prima (*recall*) prevedeva che i partecipanti scrivessero di un'esperienza personale autobiografica ricordata poco dopo l'accaduto e ne fornissero una rispettiva sintesi. In questa fase, l'accento era posto sull'immediatezza del ricordo e sulla freschezza delle emozioni ad esso associate, al fine di catturare il massimo livello di dettaglio e di autenticità nelle narrazioni.

La seconda fase (*imagined*) richiedeva agli individui di scrivere una storia simile a un diario partendo da un breve riassunto fornito dagli sperimentatori. Questa sintesi era creata per stimolare l'immaginazione dei partecipanti e indurli ad elaborare una narrazione coerente e dettagliata, sebbene basata su esperienze non vissute direttamente da loro. In questo modo, i ricercatori potevano esplorare le differenze tra narrazioni autobiografiche reali e storie inventate, osservando come la creatività e la costruzione narrativa variassero in base alla fonte dell'ispirazione.

La terza fase (*retold*) coinvolgeva i soggetti della prima, i quali ricevevano il loro riassunto originale e venivano invitati a raccontare nuovamente la storia dopo un periodo compreso tra i 3 e i 6 mesi. Questo passaggio permetteva di osservare le

variazioni nei ricordi nel tempo, analizzando come gli eventi venissero reinterpretati o modificati in base alla distanza temporale dall'accaduto. Inoltre, consentiva di studiare il processo di rielaborazione dei ricordi e l'influenza della memoria a lungo termine sulla narrazione autobiografica.

Sia nel caso dei testi *recall* che *retold*, gli sperimentatori avevano chiesto ai partecipanti di indicare il tempo trascorso da quando avevano vissuto l'evento (*timesinceevent*), in settimane o mesi, e anche la frequenza con cui avevano pensato o parlato di esso (*freqofrecall*) su una scala Likert a 5 punti da "mai" a "costantemente".

5.1.2 Procedura di annotazione umana

La procedura di annotazione umana, così come nell'esperimento precedente, ha previsto dapprima un breve *training* per entrambi gli annotatori, sia nell'uso del *Reality Monitoring* che del *Verifiability Approach*, utilizzando gli 80 testi della categoria *retold* del *dataset* precedentemente descritto, sempre attraverso l'uso dello strumento *open source Doccano* (*Figura 4.4*).

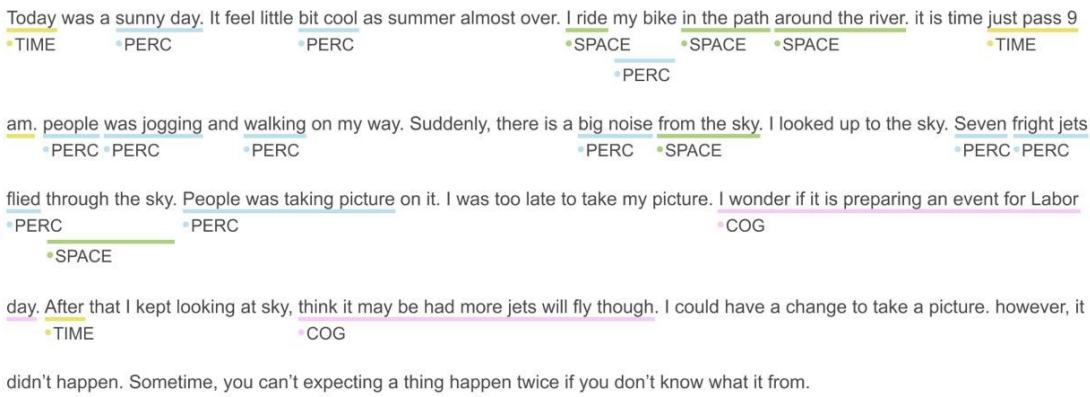

***Figura 4.4**: Schermata esemplificatrice dell'annotazione umana con l'uso di Doccano.*

In seguito, entrambi i valutatori hanno eseguito l'annotazione del 20% dei restanti 160 racconti (*recall* e *imagined*) raggiungendo un'ottima concordanza (ICC_{freq}=0.99; $F1_{RM}$=0.82; $F1_{VA}$=0.99). Il primo annotatore ha così potuto in seguito completare il restante 80% da solo.

5.1.3 Annotazione GPT-4

Come nel primo esperimento, l'interazione con il *Large Language Model* è avvenuta tramite l'interfaccia API. Il modello ha ricevuto il *codebook* utilizzato anche dagli annotatori umani e il testo scritto da codificare; nuovamente, l'*output* richiesto era l'identificazione delle etichette del *Reality Monitoring*. Successivamente, è stato chiesto a GPT-4 di valutare la verificabilità dei dettagli percettivi, temporali e spaziali utilizzando il *Verifiability Approach*.

Questa uniformità nella procedura ha garantito che le condizioni di analisi fossero identiche tra le due fasi della ricerca, assicurando un confronto accurato e valido dei risultati ottenuti.

5.1.4 Piano d'analisi

La seconda ricerca ha seguito esattamente la stessa procedura adottata precedentemente. In particolare è stato esaminato non solo il numero complessivo di dettagli identificati da GPT-4 e dagli annotatori umani tramite il *Reality Monitoring*, ma anche la loro tipologia; si è proceduto in seguito al confronto della *performance* umana con quella del modello automatico in merito alle etichette assegnate grazie all'applicazione del *Verifiability Approach*.

5.2 Risultati

5.2.1 *Reality Monitoring*: numero di dettagli

Per quanto riguarda il numero di dettagli individuati grazie all'applicazione del *Reality Monitoring*, la concordanza tra la *performance* umana e quella del modello automatico GPT-4 risulta buona (*Tabella 4.4*).

Type	Point Estimate	Lower 95% CI	Upper 95% CI
ICC 3,1	**0.896**	0.881	0.910

***Tabella 4.4**: Concordanza tra la performance umana e quella di GPT-4 nel numero di dettagli individuati tramite l'applicazione del Reality Monitoring.*

Le differenze maggiori possono essere notate a livello del numero di dettagli percettivi, spaziali e temporali (*Grafico 4.3*).

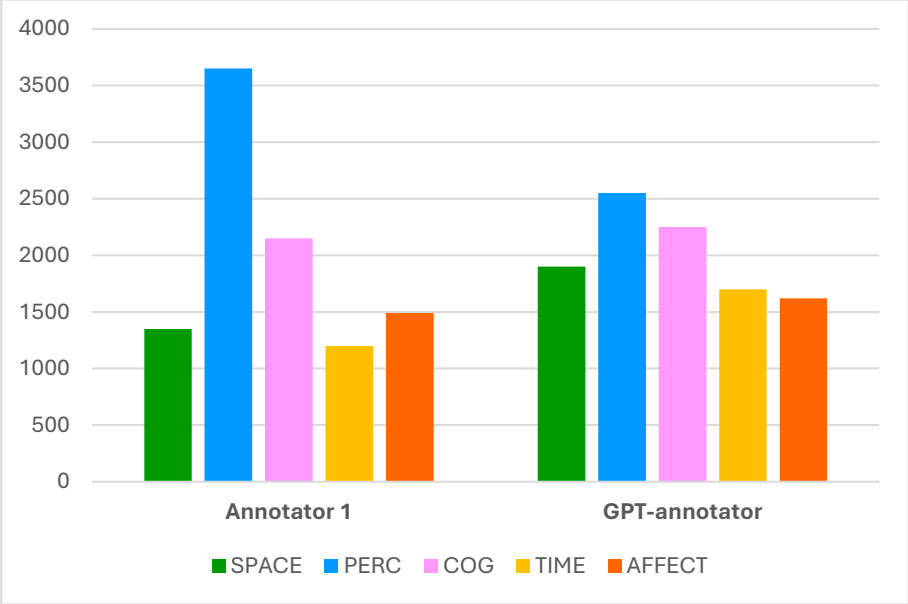

***Grafico 4.3**: Confronto tra il numero di dettagli individuati da GPT-4 e dall'annotatore umano tramite l'applicazione del Reality Monitoring.*

5.2.2 *Reality Monitoring*: *sequence-classification*

In merito alla tipologia di dettagli che sono stati individuati grazie all'uso dell'approccio *Reality Monitoring*, risulta che la concordanza tra la *performance* umana e quella fornita da GPT-4 è bassa (*Tabella 4.5*).

Tag	F1 score	Occurrences
O	0.70	17778
COG	0.29	2164
PERC	0.30	3696
AFFECT	0.38	1547
SPACE	0.41	1375
TIME	0.38	1237
Overall	**0.57**	**27797**

***Tabella 4.5**: Concordanza tra la performance umana e quella di GPT-4 nella tipologia di dettagli individuati tramite l'applicazione del Reality Monitoring.*

5.2.3 *Verifiability Approach*

In relazione alla concordanza sulla frequenza di etichettamento ottenuta tramite l'uso del *Verifiability Approach*, si osserva un valore moderato (*Tabella 4.6*).

Tag	F1 score	Occurrences
UNVE	0.71	1238
VE	0.45	1032
Overall	**0.60**	**2270**

***Tabella 4.6**: Concordanza tra la performance umana e quella di GPT-4 nell'etichettamento dei dettagli tramite l'applicazione del Verifiability Approach.*

In particolare, si può notare la grande differenza nell'attribuzione delle etichette "VE" e "UNVE" da parte di GPT-4 rispetto al primo annotatore (*Grafico 4.4*).

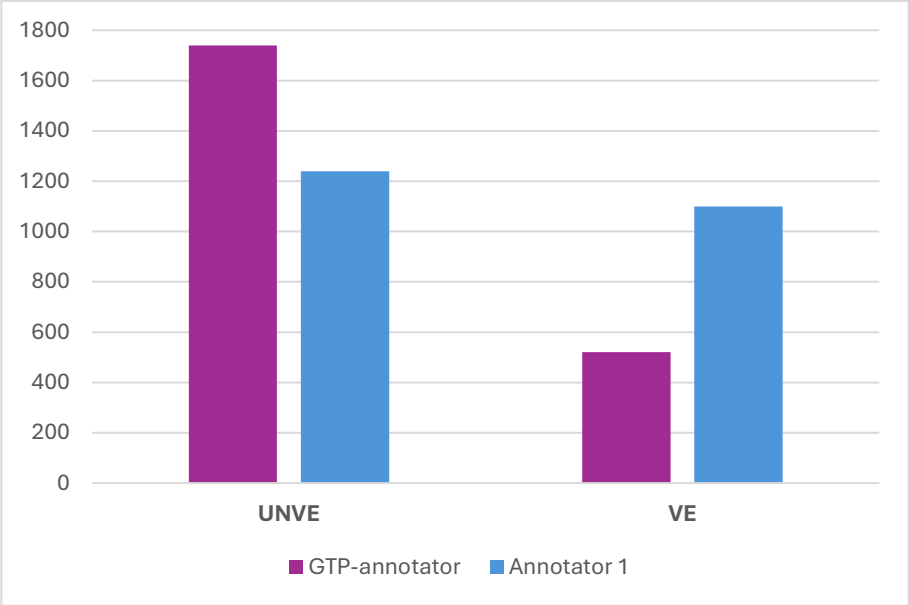

***Grafico 4.4**: Confronto tra il numero di dettagli identificati come verificabili o meno da GPT-4 e dall'annotatore umano tramite l'applicazione del Verifiability Approach.*

CAPITOLO 6: DISCUSSIONE GENERALE

6.1 Discussione

L'obiettivo del presente elaborato era quello di confrontare le prestazioni di annotatori umani e di GPT-4 nell'analisi di memorie autentiche e falsificate tramite l'utilizzo del *Reality Monitoring* e del *Verifiability Approach*. Per fare ciò sono stati presi in considerazione due *dataset*: il primo (Monaro et al., 2020) conteneva 62 narrazioni di vacanze passate, di cui 32 autentiche e 30 ingannevoli, trascritte da interviste videoregistrate di partecipanti italiani; il secondo era rappresentato da un sottoinsieme del *dataset Hippocorpus* (Sap et al., 2022) contenente 240 dichiarazioni di memorie veritiere e false. Sia gli annotatori umani che GPT-4 avevano a disposizione lo stesso *codebook*, ovvero un insieme delle definizioni e delle regole da seguire nella codifica dei testi.

Lo scopo di tale procedura era l'individuazione del numero dei dettagli presenti nei testi, la loro tipologia e in seguito la loro verificabilità; secondo le logiche alla base dei due approcci utilizzati infatti, le dichiarazioni veritiere non solo sarebbero maggiormente ricche di dettagli di tipo percettivo-contestuali mostrando invece una carenza negli aspetti cognitivi (Loconte et al., 2023), ma conterrebbero anche un numero più elevato di informazioni verificabili, ovvero la cui veridicità può essere confermata (Strömwall et al., 2006).

Nella fase di analisi, gli annotatori umani e GPT-4 hanno esaminato ciascuna narrazione identificando e categorizzando i dettagli secondo il *codebook* fornito. Successivamente, è stata valutata la verificabilità di tali informazioni per determinare se potessero essere confermate da fonti esterne.

Dall'analisi dei risultati emerge che, in merito al numero di dettagli individuati grazie all'applicazione del *Reality Monitoring*, la concordanza tra umani e GPT-4 appare eccellente (ICC=0.952) nel primo studio e buona (ICC=0.896) nel secondo. Questo significa che sia gli annotatori umani che il modello automatico sono stati in grado di identificare un numero simile di informazioni rilevanti nelle narrazioni analizzate, dimostrando un alto livello di affidabilità del modello di intelligenza artificiale nell'imitare la capacità umana. Tuttavia, una certa precauzione deve essere comunque posta in quanto nel momento in cui si va a verificare la tipologia di frasi selezionate, le differenze tra umani e modello automatico si fanno più forti.

In relazione all'etichetta assegnata al dettaglio rilevato infatti, la concordanza risulta meno positiva in entrambi gli esperimenti: nel primo (F1=0.61) assume tuttavia un valore per il quale può essere considerata come moderata; nel secondo invece (F1=0.57), il decremento ulteriore è sufficiente per farla rientrare nella categoria bassa. Rispetto ai risultati precedenti, ciò suggerisce una minor abilità di GPT-4 nell'analisi qualitativa dei dettagli individuati rispetto alla *performance* umana, risultando invece più competente nell'analisi quantitativa.

I risultati ottenuti dal confronto tra l'elaborazione testuale umana e quella automatica eseguita tramite l'applicazione del *Verifiability Approach* appaiono tuttavia più scoraggianti. L'*interrater agreement* tra le due annotazioni assume infatti un valore molto basso (F1=0.45) nel primo studio, aumentando invece di poco nel secondo (F1=0.60). Tali evidenze indicano un'estrema difficoltà da parte di GPT-4 nel comprendere la possibile verificabilità di un'informazione pur avendo a disposizione l'intera narrazione e quindi l'ambiente in cui gli avvenimenti analizzati si sono svolti. Gli esseri umani infatti, possiedono una capacità innata di interpretare il contesto e le

sfumature delle narrazioni, che consente loro di assegnare etichette con maggiore precisione. Al contrario, GPT-4, pur essendo un modello avanzato di intelligenza artificiale, si basa su pattern di linguaggio preesistenti e può avere difficoltà a cogliere le sottili differenze tra diverse categorie di dettagli.

Essendo le istruzioni fornite agli annotatori umani e al modello automatico identiche per entrambi gli esperimenti in quanto contenute nel *codebook*, una possibile spiegazione all'incremento della *performance* di GPT-4 nel secondo studio può essere rappresentata dalla differenza della lingua in cui erano scritti i testi da elaborare. Il fatto che il secondo *dataset* fosse in inglese infatti, potrebbe aver in qualche modo agevolato il modello di intelligenza artificiale nella comprensione del contesto della frase e quindi nell'attribuzione delle etichette. Tale inferenza potrebbe trovare sostegno nel fatto che non solo la gran parte dei dati su cui GPT-4 è stato addestrato appartiene alla lingua inglese, ma anche che le istruzioni presenti nel *codebook* stesso.

Alla luce dei risultati ottenuti dai due esperimenti presentati in questo elaborato, l'ipotesi inizialmente formulata, secondo cui GPT-4 avrebbe la capacità di valutare la credibilità di resoconti di memoria autentici e ingannevoli con una precisione comparabile a quella degli annotatori umani, risulta parzialmente confermata per quanto riguarda l'approccio del *Reality Monitoring*, mentre non trova riscontro nell'uso del *Verifiability Approach*. Di conseguenza, una possibile implicazione potrebbe essere l'adozione di un modello automatico per la fase di pre-annotazione nell'ambito dell'approccio *Reality Monitoring*, la quale dovrebbe tuttavia essere seguita da una valutazione umana, portando comunque ad un risparmio di tempo nello svolgimento dell'intera procedura e fornendo un *output* solo grazie a istruzioni in linguaggio naturale, non necessitando quindi di programmazione del modello.

6.2 Limiti e prospettive future

La presente indagine ha fornito un approfondito confronto tra annotatori umani e modelli di linguaggio di grandi dimensioni (LLMs) nell'ambito dell'annotazione di un testo, utilizzando due approcci metodologici: il *Reality Monitoring* e il *Verifiability Approach*. Tuttavia, come accade per ogni ricerca, è fondamentale riconoscere i limiti e le criticità incontrate, poiché farne emergere la consapevolezza non solo rende lo studio più trasparente, ma fornisce anche spunti preziosi per migliorare le future indagini in questo campo di ricerca.

Una prima criticità emersa riguarda i fattori che potrebbero aver influenzato il calcolo della concordanza tra annotatori umani e GPT-4. La discordanza nei risultati può essere stata causata dalla diversa modalità di annotazione adottata dalle due parti. Ad esempio, uno dei problemi più rilevanti ha riguardato l'errata segmentazione dei dettagli. In diversi casi, GPT-4 ha interpretato come un unico dettaglio due informazioni che invece gli annotatori umani consideravano distinte. Un caso esemplare è l'annotazione di frasi come "*we went to / the hotel*": mentre un annotatore umano avrebbe separato il movimento ("*we went to*") e il luogo ("*the hotel*") come due dettagli spaziali distinti, GPT-4 ha erroneamente codificato l'intera frase come un solo dettaglio, portando a un calo della concordanza. Inoltre, è stato osservato che il numero di parole selezionate variava significativamente tra il modello e gli annotatori umani. Un esempio concreto è l'identificazione di dettagli temporali: se un annotatore umano codifica "*for three days*" come un unico dettaglio, GPT-4 potrebbe considerare soltanto "*three days*", causando ulteriori discrepanze e abbassando la concordanza complessiva. Per mitigare questo problema, nelle future ricerche sarebbe opportuno definire con maggiore precisione i

limiti e i criteri per la selezione delle parole da considerare come "dettaglio" all'interno del processo di annotazione.

Un ulteriore elemento di criticità è rappresentato dalla difficoltà di GPT-4 nell'identificare correttamente le informazioni rilevanti all'interno del testo. Da un lato, il modello tende a codificare ripetutamente la stessa informazione, violando quanto stabilito nel *codebook* che definisce tali ripetizioni come irrilevanti per l'analisi. Dall'altro, GPT-4 sembra includere nel processo di codifica parti del testo che non corrispondono al ricordo autobiografico e che pertanto non dovrebbero essere considerate per la valutazione della credibilità del racconto. Mentre la prima problematica può essere interpretata come un errore del modello, poiché l'istruzione relativa alla gestione delle ripetizioni era chiaramente indicata, la seconda criticità offre un'opportunità di miglioramento per future applicazioni sperimentali. Una possibile soluzione sarebbe quella di esplicitare in modo più dettagliato all'interno del *codebook* che tutte le informazioni che non corrispondono al ricordo autobiografico devono essere escluse dal processo di codifica. L'inclusione di istruzioni più precise potrebbe migliorare significativamente la *performance* del modello nell'applicazione del *Verifiability Approach*, garantendo una codifica più accurata e coerente.

Inoltre, per ottimizzare ulteriormente la *performance* di GPT-4 e migliorare la concordanza con gli annotatori umani, un'importante prospettiva futura potrebbe consistere nel modificare i parametri utilizzati per il modello. Cambiare indici come la temperatura, che influenza il grado di variabilità e creatività delle risposte, o altri settaggi che controllano l'accuratezza e la coerenza della generazione di testo, potrebbe portare a risultati diversi e potenzialmente più in linea con le annotazioni umane.

L’esplorazione di questi parametri permetterebbe di adattare meglio il modello alle esigenze specifiche dell’annotazione, riducendo le discrepanze osservate.

Un’altra possibile area di miglioramento riguarda la revisione e l’ottimizzazione delle istruzioni fornite sia a GPT-4 che agli annotatori umani. La variazione delle linee guida contenute nel *codebook* potrebbe aiutare a ridurre le ambiguità che si sono verificate nel processo di annotazione. Ad esempio, fornire istruzioni più chiare e dettagliate su cosa considerare come dettaglio, potrebbe garantire una maggiore uniformità nei criteri di annotazione, riducendo ulteriormente le differenze tra il modello automatico e gli esseri umani.

Un ulteriore sviluppo interessante sarebbe quello di creare una versione specializzata di GPT focalizzata esclusivamente su un tipo specifico di dettaglio, come ad esempio i dettagli temporali, spaziali o percettivi. Addestrare il modello a riconoscere con precisione solo una determinata categoria di informazione potrebbe portare a un incremento notevole nella qualità delle annotazioni, riducendo il rischio di includere dettagli non rilevanti o ridondanti. Questo approccio settorializzato permetterebbe di sfruttare al meglio la potenza di GPT concentrandola su aree specifiche di interesse, anziché cercare di farlo eccellere in una vasta gamma di compiti annotativi.

Infine, guardando al futuro, gli studi potrebbero estendere l'analisi ad altri modelli di linguaggio di grandi dimensioni oltre GPT-4, come *Claude 3* o *Llama 3*. Questi modelli potrebbero offrire diverse prospettive e capacità di annotazione, aprendo la strada a un confronto più ampio non solo tra le *performance* umane e quelle di un singolo LLM, ma anche tra diversi LLMs tra loro. Tale approccio permetterebbe di esplorare le specifiche competenze e i limiti di ciascun modello, contribuendo così a un avanzamento complessivo delle metodologie di annotazione automatica dei testi. Questo tipo di

confronto potrebbe rivelarsi cruciale per identificare quale modello riesca a offrire le prestazioni più affidabili, sia in termini di accuratezza nell'annotazione che di aderenza alle istruzioni fornite, permettendo così di ottimizzare l'uso degli LLMs in studi futuri.

CONCLUSIONI

La ricerca forense si occupa da sempre di comprendere il funzionamento della memoria umana e le metodologie per valutarne l'affidabilità. Tradizionalmente, gli esperti analizzano manualmente i resoconti di memoria basandosi su criteri come la quantità e la qualità dei dettagli riportati, nonché la loro verificabilità, richiedendo tuttavia costi elevati e dispendio di tempo. L'emergere dei *Large Language Models* (LLMs) ha tuttavia stimolato una serie di studi volti a esaminare la capacità di questi modelli automatici di emulare il giudizio umano.

Il presente elaborato mette a confronto le prestazioni degli annotatori umani con quelle di GPT-4 nell'analisi di memorie autentiche e falsificate, utilizzando due *dataset* distinti, di cui il primo comprendente 62 narrazioni riguardanti esperienze di vacanze passate in lingua italiana. I valutatori umani sono stati incaricati di estrapolare i dettagli dai testi forniti utilizzando i metodi del *Reality Monitoring* e del *Verifiability Approach*. Questa analisi è stata successivamente replicata su un secondo *dataset* contenente 240 dichiarazioni scritte in inglese relative a esperienze significative del passato.

I risultati dello studio indicano che GPT-4 ha mostrato un'ottima concordanza con gli annotatori umani nel numero di dettagli identificati attraverso il *Reality Monitoring*, mentre la concordanza è risultata moderata per quanto riguarda l'identificazione del tipo di dettaglio. Tuttavia, nell'applicazione del *Verifiability Approach*, il modello automatico ha incontrato difficoltà, non riuscendo a etichettare correttamente le informazioni rilevate.

Questo confronto offre spunti interessanti per future ricerche e applicazioni nel campo della valutazione della memoria, come ad esempio l’uso di un modello automatico come

pre-annotatore di un testo o la sperimentazione con l'uso di altri LLMs. Questi risultati, seppur con qualche criticità, sono tuttavia incoraggianti.

RIFERIMENTI BIBLIOGRAFICI

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology*, *25*(2), 299-306.

Agosta, S., Pezzoli, P., & Sartori, G. (2013). How to detect deception in everyday life and the reasons underlying it. *Applied Cognitive Psychology*, *27*(2), 256-262.

Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. *Frontiers in Psychology*, *4*, 519.

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. *Electronics*, *10*(11), 1348.

Alonso-Quecuty, M. L. (1990). Recuerdo de la realidad percibida vs. imaginada. Buscando la mentira. *Boletín de Psicología*, *29*, 73-86.

Alonso-Quecuty, M. L. (1996). Detecting fact from fallacy in child and adult witness accounts. *Psychology, law, and criminal justice: International developments in research and practice*, 74-80.

Alonso-Quecuty, M., & Hernández-Fernaud, E. (1997). Tócala otra vez Sam: repitiendo las mentiras. *Estudios de psicología*, *18*(57), 29-37.

Amado, B. G., Arce, R., Farina, F., & Vilarino, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology*, *16*(2), 201-210.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: American Psychiatric Publishing.

Babchishin, K. M., Nunes, K. L., & Kessous, N. (2014). A multimodal examination of sexual interest in children: A comparison of sex offenders and nonsex offenders. *Sexual Abuse*, *26*(4), 343-374.

Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*. Cambridge university press.

Bell, B. E., & Loftus, E. F. (1989). Trivial persuasion in the courtroom: The power of (a few) minor details. *Journal of personality and social psychology*, *56*(5), 669.

Bogaard, G., Colwell, K., & Crans, S. (2019). Using the reality interview improves the accuracy of the criteria-based content analysis and reality monitoring. *Applied Cognitive Psychology, 33(6)*, 1018-1031.

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, *19*(3), 313-329.

Bruck, M., Ceci, S. J., & Francoeur, E. (1999). The accuracy of mothers' memories of conversations with their preschool children. *Journal of Experimental Psychology: Applied*, *5*(1), 89.

Chaturvedi, I., Cambria, E., Welsch, R. E., & Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, *44*, 65-77.

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, *9*(1), 10.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284.

Colwell, K., Hiscock-Anisman, C., & Fede, J. (2013). Assessment criteria indicative of deception: An example of the new paradigm of differential recall enhancement. In *Applied issues in investigative interviewing, eyewitness memory, and credibility assessment* (pp. 259–291). New York: Springer.

Conway, M. A. (1996). Autobiographical memory. In *Memory* (pp. 165-194). Academic Press.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological review*, *107*(2), 261.

Conway, M., & Ross, M. (1984). Getting what you want by revising what you had. *Journal of personality and social psychology*, *47*(4), 738.

De Leo, G., Scali, M., & Caso, L. (2005). *La testimonianza. Problemi, metodi e strumenti nella valutazione dei testimoni*. Il Mulino.

Derczynski, L. (2016, May). Complementarity, F-score, and NLP Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 261-266).

Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biology of sport*, *40*(2), 615-622.

Dhingra, S., Singh, M., Vaisakh, S. B., Malviya, N., & Gill, S. S. (2023). Mind meets machine: Unravelling gpt-4's cognitive psychology. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *3*(3), 100139.

Elntib, S., & Wagstaff, G. (2017). Are reality monitoring differences between truthful and deceptive autobiographical accounts affected by standardisation for word-count and the presence of others?. *Psychology, Crime & Law, 23*(7), 699-716.

Evans, J. R., Michael, S. W., Meissner, C. A., & Brandon, S. E. (2013). Validating a new assessment method for deception detection: Introducing a Psychologically Based Credibility Assessment Tool. *Journal of Applied Research in Memory and Cognition*, *2*(1), 33-41.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & cognition*, *38*, 833-848.

Gatti, D., Rinaldi, L., Marelli, M., Mazzoni, G., & Vecchi, T. (2022). Decomposing the semantic processes underpinning veridical and false memories. *Journal of Experimental Psychology: General*, *151*(2), 363.

Gatti, D., Rinaldi, L., Mazzoni, G., & Vecchi, T. (2024). Semantic and episodic processes differently predict false memories in the DRM task. *Scientific Reports*, *14*(1), 256.

Geiselman, R. E., Fisher, R. P., Firstenberg, I., Hutton, L. A., Sullivan, S. J., Avetissian, I., and Prosk, A. (1984). Enhancement of eyewitness memory: an empirical evaluation of the cognitive interview. *Journal of Police Science and Administration, 12*, 74±80.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, *62*(1), 451-482.

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338.

Granhag, P. A., Andersson, L. O., Strömwall, L. A., & Hartwig, M. (2004). Imprisoned knowledge: Criminals' beliefs about deception. *Legal and Criminological Psychology*, **9**, 103–119.

Granhag, P. A., & Hartwig, M. (2008). A new theoretical perspective on deception detection: On the psychology of instrumental mind-reading. *Psychology, Crime & Law*, *14*(3), 189-200.

Granhag, P. A., Strömwall, L., & Olsson, C. (2001, June). Fact or fiction? Adults' ability to assess children's veracity. In *11th European Conference on Psychology and Law, Lisbon, Portugal*.

Greenwald, G. A., McGhee, E. D., and Schwartz, K. L. J. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol*. 74, 1464–1480.

Griego, A. W., Datzman, J. N., Estrada, S. M., & Middlebrook, S. S. (2019). Suggestibility and false memories in relation to intellectual disability and autism spectrum disorder: a meta-analytic review. *Journal of Intellectual Disability Research*, *63*(12), 1464-1474.

Gulotta, G. (2011). *Compendio di psicologia giuridico-forense, criminale e investigativa* (Vol. 53). Giuffrè Editore.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, *45*(1), 1-23.

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic use of evidence during investigative interviews: The state of the science. *Credibility assessment*, 1-36.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Kronkvist, O. (2006). Strategic use of evidence during police interviews: When training to detect deception works. *Law and human behavior*, *30*(5), 603.

Harvey, A. C., Vrij, A., Leal, S., Lafferty, M., & Nahari, G. (2017). Insurance based lie detection: Enhancing the verifiability approach with a model statement component. *Acta psychologica*, *174*, 1-8.

Harvey, A. G., Bryant, R. A., & Dang, S. T. (1998). Autobiographical memory in acute stress disorder. *Journal of consulting and clinical psychology*, *66*(3), 500.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve?. *science, 298(5598), 1569-1579.*

Hazlett, G. (2006). Research on detection of deception: What we know vs. what we think we know. *NDIC, Educing information interrogation: Science and art foundations for the future*, 45-62.

Hernández-Fernaud, E., & Alonso-Quecuty, M. (1997). The cognitive interview and lie detection: A new magnifying glass for Sherlock Holmes?. *Applied Cognitive*

Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, *11*(1), 55-68.

Howe, M. L., & Courage, M. L. (1993). On resolving the enigma of infantile amnesia. *Psychological bulletin*, *113*(2), 305.

Istituto dell'Enciclopedia italiana (2017). *Vocabolario Treccani 2017*: *Il Treccani*. Roma: Giunti T.V.P.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023, December). Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1827-1843).

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological review*, *88*(1), 67.

Jones, B., Heard, H., Startup, M., Swales, M., Williams, J. M. G., & Jones, R. S. P. (1999). Autobiographical memory and dissociation in borderline personality disorder. *Psychological medicine*, *29*(6), 1397-1404.

Jupe, L. M., Leal, S., Vrij, A., & Nahari, G. (2017). Applying the verifiability approach in an international airport setting. *Psychology, Crime & Law*, *23*(8), 812-825.

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139-172.

Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2018). Using named entities for computer-automated verbal deception detection. *Journal of forensic sciences*, *63*(3), 714-723.

Köhnken, G., Milne, R., Memon, A., & Bull, R. (1999). The cognitive interview: A meta-analysis. *Psychology, crime and law*, *5*(1-2), 3-27.

Komiya, K., Suzuki, M., Iwakura, T., Sasaki, M., & Shinnou, H. (2018). Comparison of methods to annotate named entity corpora. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *17*(4), 1-16.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kumawat, D., & Jain, V. (2015). POS tagging approaches: A comparison. *International Journal of Computer Applications*, *118*(6).

Kuyken, W., & Dalgleish, T. (1995). Autobiographical memory and depression. *British journal of clinical psychology*, *34*(1), 89-92.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, *34*(1), 50-70.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1-35.

Loconte, R., Russo, R., Capuozzo, P., Pietrini, P., & Sartori, G. (2023). Verbal lie detection using Large Language Models. Scientific Reports, 13(1), 22849.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior*, *13*(5), 585-589.

Magro, T., Sartori, G., & Benatti, F. (2023). *La memoria autobiografica*. Libreriauniversitaria.it.

Mammarella, N., & Di Domenico, A. (2011). *La memoria autobiografica*. Carocci.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313-330.

Martinez, A. R. (2012). Part-of-speech tagging. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(1), 107-113.

Masip, J., Bethencourt, M., Lucas, G., SEGUNDO, M. S. S., & Herrero, C. (2012). Deception detection from written accounts. *Scandinavian Journal of Psychology*, *53*(2), 103-111.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law*, *11*(1), 99-122.

Mazzoni, G., & Ambrosio, K. (2003). L'analisi del resoconto testimoniale in bambini: impiego del metodo di analisi del contenuto CBCA in bambini di 7 anni. *Disponibile online: http://www. psicologiagiuridica.com/numero*, *20006*.

McNally, R. J., Litz, B. T., Prassas, A., Shin, L. M., & Weathers, F. W. (1994). Emotional priming of autobiographical memory in post-traumatic stress disorder. *Cognition & Emotion*, *8*(4), 351-367.

Memon, A., & Gawrylowicz, J. (2018). The cognitive interview. *The handbook of communication skills*, 511-530.

Memon, A., Meissner, C. A., & Fraser, J. (2010). The Cognitive Interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, public policy, and law*, *16*(4), 340.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Monaro, M., Capuozzo, P., Ragucci, F., Maffei, A., Curci, A., Scarpazza, C., ... & Sartori, G. (2020). Using blink rate to detect deception: A study to validate an automatic blink detector and a new dataset of videos from liars and truth-tellers. In *Human-Computer Interaction. Human Values and Quality of Life: Thematic Area, HCI 2020, Held as Part of the 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part III 22* (pp. 494-509). Springer International Publishing.

Monaro, M., Maldera, S., Scarpazza, C., Sartori, G., & Navarin, N. (2022). Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models. *Computers in Human Behavior*, *127*, 107063.

Moore, S. A., & Zoellner, L. A. (2007). Overgeneral autobiographical memory and traumatic events: an evaluative review. *Psychological bulletin*, *133*(3), 419.

Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PloS one*, *10*(7), e0120644.

Nahari, G. (2018). The applicability of the verifiability approach to the real world. In *Detecting concealed information and deception* (pp. 329-349). Academic press.

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Does the truth come out in the writing? Scan as a lie detection tool. *Law and Human Behavior*, *36*(1), 68.

Nahari, G., Vrij, A., & Fisher, R. P. (2014). The verifiability approach: Countermeasures facilitate its ability to discriminate between truths and lies. *Applied Cognitive Psychology*, *28*(1), 122-128.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, *11*(1), 81.

Nock, M. K., & Banaji, M. R. (2007). Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. *Journal of consulting and clinical psychology*, *75*(5), 707.

Nolen-Hoeksema, S., Fredrickson, B. L., Loftus, G.R., Lutz, C. (2017). *Atkinson & Hilgard's Introduzione alla psicologia* (16ª ed.). Piccin.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European review of social psychology*, *18*(1), 36-88.

Oberlader, V. (2019). *Meta-Analyses on the Validity of Verbal Tools for Credibility Assessment* (Doctoral dissertation, Universitäts-und Landesbibliothek Bonn).

Onyenwe, I., Nwagbo, S., Mbeledogu, N., & Onyedinma, E. (2020). The impact of political party/candidate on the election results from a sentiment analysis perspective using# AnambraDecides2017 tweets. *Social Network Analysis and Mining*, *10*, 1-17.

Palena, N., Caso, L., Vrij, A., & Nahari, G. (2021). The verifiability approach: A meta-analysis. *Journal of Applied Research in Memory and Cognition*, *10*(1), 155-166.

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 1-17.

Patel, S. C., & Fan, J. (2023). Identification and description of emotions by current large language models. *bioRxiv*, 2023-07.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [Computer software]. *Austin, TX: liwc. net*, *135*.

Petroczi, A., Uvacsek, M., Nepusz, T., Deshmukh, N., Shah, I., Aidman, E. V., ... & Naughton, D. P. (2011). Incongruence in doping related attitudes, beliefs and opinions in the context of discordant behavioural data: in which measure do we trust?. *PLoS One*, *6*(4), e18804.

Raskin, D. C., & Esplin, P. W. (1991). Statement validity assessment: Interview procedures and content analysis of children's statements of sexual abuse. *Behavioral Assessment*.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (Eds.). (2013). Credibility assessment: Scientific research and applications.

Roma, P., San Martini, P., Sabatello, U., Tatarelli, R., & Ferracuti, S. (2011). Validity of Criteria-Based Content Analysis (CBCA) at trial in free-narrative interviews. *Child abuse & neglect*, *35*(8), 613-620.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological review*, *96*(2), 341.

Rubin, D. C., Rahhal, T. A., & Poon, L. W. (1998). Things learned in early adulthood are remembered best. *Memory & cognition*, *26*, 3-19.

Sageder, C., & Karampatakis, S. (2021, September). Annotating Entities with Fine-Grained Types in Austrian Court Decisions. In *SEMANTiCS* (pp. 139-153).

Santtila, P., Roppola, H., & Niemi, P. (1998). Assessing the Truthfulness of Witness Statements Made by Children (Aged 7--8, 10--11, and 13--14) Employing Scales Derived from Johnson and Raye's Model of Reality Monitoring. *Expert evidence*, *6*(4), 273-289.

Sapir, A. (1987). The LSI course on scientific content analysis (SCAN). *Phoenix, ZA: Laboratory for Scientific Interrogation*.

Sap, M., Jafarpour, A., Choi, Y., Smith, N. A., Pennebaker, J. W., & Horvitz, E. (2022). Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, *119*(45), e2211715119.

Sartori, G. (2021). *La memoria del testimone: Dati scientifici utili a magistrati, avvocati e consulenti*. Giuffrè Francis Lefebvre.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., and Castiello, U. (2008). How to accurately detect autobiographical events. *Psychol. Sci*. 19, 772–780.

Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology*, *14*, 1279317.

Sartori, G., Zangrossi, A., & Monaro, M. (2018). Deception detection with behavioral methods: the autobiographical implicit association test, concealed information test–reaction time, mouse dynamics, and keystroke dynamics. In *Detecting Concealed Information and Deception* (pp. 215-241). Academic Press.

Scali, M., Calabrese, C., & Biscione, M. C. (2003). *La tutela del minore: le tecniche di ascolto*. Carocci.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Soldner, F., Pérez-Rosas, V., & Mihalcea, R. (2019, June). Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1768-1777).

Sporer, S. L. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *11*(5), 373-397.

Sporer, S. L. (2004). *Reality monitoring and detection of deception*.

Stollenwerk, F., Fastlund, N., Nyqvist, A., & Öhman, J. (2023). Annotated Job Ads with Named Entity Recognition. *arXiv preprint arXiv:2310.11769*.

Stracciari, A., Bianchi, A., Sartori, G. (2010). Neuropsicologia forense. *Il Mulino*.

Strobl, M., Trabelsi, A., & Zaïane, O. (2022, June). Named entity recognition for partially annotated datasets. In *International Conference on Applications of Natural Language to Information Systems* (pp. 299-306). Cham: Springer International Publishing.

Strömwall, L. A., Hartwig, M., & Granhag, P. A. (2006). To act truthfully: Nonverbal behaviour and strategies during a police interrogation. *Psychology, Crime & Law*, *12*(2), 207-219.

Suchotzki, K. (2018). Challenges for the Application of Reaction Time–Based Deception Detection Methods. In *Detecting Concealed Information and Deception* (pp. 243-268). Academic Press.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Tommasino, M. G., Carillo, B. F., & Grattagliano, I. (2008). Statement validity analysis e reality monitoring: analisi critica di due strumenti per valutare le affermazioni dei testimoni. *Rassegna Italiana di Criminologia*, (2), 409-431.

Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224-226.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vernham, Z., Vrij, A., Mann, S., Leal, S., & Hillman, J. (2014). Collective interviewing: Eliciting cues to deceit using a turn-taking approach. *Psychology, Public Policy, and Law*, 20, 309–324.

Verschuere, B., Lin, C. C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., ... & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature human behaviour*, *7*(5), 718-728.

Vicianova, M. (2015). Historical techniques of lie detection. *Europe's journal of psychology*, *11*(3), 522.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. Chichester, UK: Wiley.

Vrij, A. (2007). Credibility assessments in a legal context. *Applying psychology to criminal justice*, 81-96.

Vrij, A. (2016). Baselining as a lie detection method. *Applied Cognitive Psychology*, *30*(6), 1112-1119.

Vrij, A. (2018). Verbal lie detection tools from an applied perspective. In *Detecting concealed information and deception* (pp. 297-327). Academic Press.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *36*(2), 113.

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1-21.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1-2), 39-43.

Vrij, A., Granhag, P. A., Ashkenazi, T., Ganis, G., Leal, S., & Fisher, R. P. (2022). Verbal lie detection: Its past, present and future. *Brain Sciences*, *12*(12), 1644.

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science*, *20*(1), 28-32.

Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R. P., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and human behavior*, *33*, 159-166.

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior*, *31*, 499-518.

Vrij, A., Mann, S., Leal, S., & Fisher, R. (2007b). *'Look into my eyes': Can an instruction to maintain eye contact facilitate lie detection?* Submitted.

Vrij, A., Mann, S., Fisher, R., Leal, S., Milne, B., & Bull, R. (2007a). Increasing cognitive load to facilitate lie detection: The benefit or recalling an event in reverse order. *Law and Human Behavior* (In press).

Vrij, A., Nahari, G., Isitt, R., & Leal, S. (2016). Using the verifiability lie detection approach in an insurance claim setting. *Journal of Investigative Psychology and Offender Profiling*, *13*(3), 183-197.

Vrij, A., Taylor, P., & Picornell, I. (2015). Verbal lie detection. *Communication in investigative and legal contexts: Integrated approaches from forensic psychology, linguistics and law enforcement*, 259-286.

Wilson, A., & Ross, M. (2003). The identity function of autobiographical memory: Time is on our side. *Memory*, *11*(2), 137-149.

y Arcas, B. A. (2022). Do large language models understand us?. *Daedalus*, *151*(2), 183-197.

Yuille, J. C. (Ed.). (1989). *Credibility assessment* (No. 47). Springer Science & Business Media.

Zangrossi, A., Agosta, S., Cervesato, G., Tessarotto, F., & Sartori, G. (2015). “I didn't want to do it!” The detection of past intentions. *Frontiers in Human Neuroscience*, *9*, 608.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223.*

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In *Advances in experimental social psychology* (Vol. 14, pp. 1-59). Academic Press.

SITOGRAFIA

OpenAI. GPT-4. Disponibile a: https://openai.com/research/gpt-4

APPENDICE

CODING SCHEME FOR REALITY MONITORING AND THE VERIFIABILITY OF DETAILS

STEP 1: Code details in memory reports according to the following Reality Monitoring (RM) categories:

- **Perceptual/sensory detail**: information about sensory experiences obtained through the senses:
    - Sounds (shouting, noises from the street, reproduction of conversations): code two separate details only if they add information to the detail (e.g., "sarcastic whisper" not "*loud scream*"). Examples:
        - "*He told me that the exam was difficult*"
        - "*He really shouted at him*".
    - Smells (e.g. "*I could smell her perfume*").
    - Tastes. Examples:
        - "*I had a coffee*"
        - "*The tea was sweet*".
    - Physical sensations. Examples:
        - "*It hurt*"
        - "*I was pushed*"
        - "*My heart was beating fast*".
    - Visual details: do not code repetitions when something has been seen before and is referred to again; code when a person/object/place becomes the subject of a sentence. Examples:
        - "*I saw the cashier sitting at the desk*"

- "*One of my classmates got hurt*"
- "*I saw the nurse enter the ward*"
- "*I saw him entering the room*".

Encodes perceptual details with [PERC].

- **Spatial detail**: information about locations or spatial arrangement of people and/or objects, information linking the event to particular places or spatial contexts such as street names, locales, city names.
  - Place information. Examples:
    - "*I was in the library*"
    - "*We were on vacation in a city*"
    - "*It was in a park*".
  - Arrangement of people and/or objects. Examples:
    - "*The book was on the shelves on the right*"
    - "*The lamp was partially hidden behind the curtains*"
    - "*The man was sitting left from his wife*".
  - Information linking the event to spatial locations or contexts, details such as street names, locales, city names (e.g. "*We were close to Barcelona*").
  - Verbs expressing movements (e.g. "*I left the room*").
  - When movement verbs are accompanied by places they count as two different details (e.g., I arrived [SPACE] at home [SPACE]).
  - Directions. Examples:
    - "*Into*"
    - "*Out of*"

    - "*Facing away*".
- Do not count vague descriptions (e.g. "*It was somewhere over there*").

Encodes spatial details as [SPACE].

- **Temporal detail**: information about when the event happened, the duration of an activity, or the sequence of events or linking the event to a particular time by giving date or time information. This type of detail falls under the category "verifiable details (VE)". Encodes these details with [TIME]:
    - When the event happened by also giving an indication of the date or time. Examples:
        - "*early in the morning*".
    - Duration of an event. Examples:
        - "*20 minutes*"
        - "*until 5 p.m*".
    - Sequence of events. Examples:
        - "*First*"
        - "*Then*"
        - "*As soon as the guy entered the pub the girl started smiling*".
    - Temporal adverbs. Examples:
        - "*While*"
        - "*When*"
        - "*Immediately*"
        - "*Again*".

- **Affective detail**: when the person remembers feelings of the event, accounts of subjective mental states. Examples:
    - "*Joseph was very scared*"
    - "*I was frightened*"
    - "*I was scared* [AFFECT], *it horrified me* [AFFECT]"
    - "*It really hurt my feelings*".

Encodes affective details with [AFFECT].

- **Cognitive operations**: evidence in narratives of various cognitive activities, such as thoughts or reasoning and cognitive suppositions of sensory experiences. This criterion also includes descriptions of inferences made by the participant at the time of the event:
    - Cognitive reasoning and operations. Examples:
        - "*It made me think how nice it might be if I had never been there*"
        - "*I must have had my coat on, as it was very cold that night*".
    - Thoughts. Examples:
        - "*I thought it would have been nice to join that party*".
    - Inferences and assumptions. Examples:
        - "*He seemed pretty smart*"
        - "*She seemed quite clever*".

Encodes cognitive operations as [COG].

WARNING:

- When describing who was present, this counts as one detail only (e.g. "*My brother, my parents...*").
- It is considered a detail when it adds new information that cannot be abstracted from the other information (e.g. "*The blue house*", blue cannot be extracted, so two details are considered; "*The green tree*", it is presumed that the tree is green, so it does not add new information, so it is considered as one detail) .

Example of how an annotated text should appear: "*I went* [SPACE] *to the kitchen* [SPACE], *got a gun* [PERC] *and shot him* [PERC]. *I thought it was the right thing to do* [COG], *although then I began to tremble* [AFFECT]. Then, [TIME] *he was scared* [AFFECT] *and he left* [PERC]".