



UNIVERSITÀ DEGLI STUDI DI PALERMO

SCUOLA DI DOTTORATO

CORSO DI DOTTORATO IN INFORMATION AND COMMUNICATION TECHNOLOGIES

DIPARTIMENTO DI INGEGNERIA

Mediatori Morali nel Metaverso Ubiquitario

NON-PLAYER CHARACTERS, INFLUENCER AI E ALTER EGO DIGITALI COME GUIDE ETICHE TRA IL
MONDO REALE E QUELLO VIRTUALE

AUTORE

GIUSEPPE FULVIO GAGLIO

RELATORE

ANTONIO CHELLA

XXXVIII CICLO

ANNO ACCADEMICO 2025 - 2026

Indice

1	Introduzione	6
1.1	Il Contesto Attuale: IA Generativa, Metaverso e Moralità Artificiale	6
1.2	Obiettivi della Ricerca	8
1.3	Metodologia di Ricerca	9
1.3.1	Sintesi del processo metodologico	10
1.4	Struttura della tesi	10
2	Il Metaverso: Dalle Origini Teoriche e Fantascientifiche alle Applicazioni Pratiche	12
2.1	Introduzione: definizione e stato dell'arte del Metaverso	12
2.1.1	Origini teoriche e sviluppo del concetto	12
2.1.2	Definizioni accademiche e caratteristiche fondamentali	12
2.1.3	Stato attuale dello sviluppo del Metaverso	13
2.2	Potenzialità del Metaverso: ambienti immersivi, interazione sociale e simulazione	13
2.2.1	Il Metaverso come ambiente di sperimentazione situata	14
2.2.2	Interazioni sociali, presenza e identità digitale	14
2.2.3	Digital twin e alter ego virtuali	14
2.3	Rischi e Sfide del Metaverso: Aspetti Critici da Considerare	15
2.3.1	Disinibizione e Comportamenti Inappropriati	15
2.3.2	Alienazione e Impatto sul Benessere Psicologico	16
2.3.3	Manipolazione Comportamentale	17
2.4	Conclusioni	17
3	IA Generativa e Sistemi Generativi Multi-Agente	19
3.1	Introduzione all'IA Generativa	19
3.2	Benefici e vantaggi dell'IA Generativa	20
3.3	Aspetti negativi e pericoli dell'IA Generativa	21
3.4	Sistemi di agenti basati su IA Generativa	22
3.4.1	Definizione e architettura degli agenti di IA Generativa	22
3.4.2	Sistemi multi-agente e orchestrazione	23
3.4.3	Rilevanza per la progettazione di agenti morali artificiali	23

3.5	Conclusioni	23
4	Agenti Morali Artificiali: Fondamenti, Implicazioni e Prospettive Future	25
4.1	L'emergere degli Agenti Morali Artificiali nell'ecosistema digitale contemporaneo	25
4.2	Agenti Morali Artificiali: definizione, genealogia filosofica e modelli di implementazione	26
4.2.1	Che cos'è un Agente Morale Artificiale	26
4.2.2	Radici filosofiche dell'etica delle macchine	27
4.2.3	Dalla teoria alla pratica: modelli di implementazione dell'agenzia morale	27
4.3	Esemplarità morale e Intelligenza Artificiale: fondamenti teorici e implicazioni applicative	29
4.3.1	Macchine intelligenti come agenti, pazienti e proxy morali	29
4.3.2	IA come consigliere morale e strumento di miglioramento etico	29
4.3.3	Interazione morale uomo-IA: fiducia, allineamento e limiti strutturali	30
4.3.4	Apprendimento da esemplari morali e ruolo della supervisione umana	30
4.4	Educazione morale nell'era dell'Intelligenza Artificiale: rischi, opportunità e nuove responsabilità nei social media e nel Metaverso	31
4.4.1	L'urgenza di una formazione etica strutturata	31
4.4.2	Social media e IA: tra democratizzazione e vulnerabilità	32
4.4.3	Il Metaverso come nuova frontiera educativa: potenzialità e criticità	32
4.4.4	Verso un'ecologia morale integrata: strategie di mitigazione e governance	33
4.5	Agenti Morali Artificiali come Guide per l'Umanità: Potenzialità, Limiti e Prospettive di Implementazione	34
4.5.1	Potenzialità degli AMA come strumenti di supporto al giudizio morale	34
4.5.2	Limiti strutturali e criticità teoriche	35
4.5.3	Architetture e tecniche implementative	36
4.5.4	Percezione, fiducia e accettazione sociale	36
4.5.5	Sfide aperte	37
4.6	Etica della Virtù, Deontologia e Utilitarismo: modelli teorici e criticità di implementazione nell'Intelligenza Artificiale	37
4.6.1	Tre modelli etici a confronto	37
4.6.2	Il ruolo determinante dei dati di addestramento	39
4.6.3	Il dataset <i>Moral Stories</i> : potenzialità e limiti	39
4.7	Conclusioni	40
5	Sviluppo e Progettazione di Agenti Morali e Sociali in Ambienti Immersivi	42
5.1	Introduzione	42

5.2	Fase 1: Agenti virtuali con comportamento socialmente conforme: l'approccio MET-iquette	42
5.2.1	Introduzione	42
5.2.2	Background teorico	43
5.2.3	L'agente sociale	45
5.2.4	L'ambiente sperimentale di MET-Iquette	47
5.3	Fase 2: Il paradigma del grillo parlante	54
5.3.1	Background teorico	54
5.3.2	Favorire l'etica: studio e implementazione degli agenti morali . . .	55
5.3.3	Esempio di simulazione	59
5.3.4	Considerazioni	62
5.4	Fase 3: Un'architettura di Intention Reading per agenti virtuali	63
5.4.1	Introduzione	63
5.4.2	Background: CASPER e il Qualitative Spazial Reasoning	64
5.4.3	L'architettura di CASPER-MV	66
5.4.4	Test sulle performance	72
5.5	Fase 4: Integrazione di Agenti LLM in Architetture Cognitive per il Processo Decisionale Etico in Sistemi Immersivi	74
5.5.1	Fondamenti Teorici e Integrazione Neuro-Simbolica	74
5.5.2	Un breve riepilogo: l'architettura di CASPER-MV	75
5.5.3	Orchestrazione Multi-Agente con CrewAI	75
5.5.4	ConvAI: Intelligenza Incarnata e NPC Conversazionali Spazialmente Consapevoli	76
5.5.5	Ragionamento Etico Situato e il Dataset "Moral Stories"	77
5.5.6	Validazione tecnica e valutazione qualitativa	78
5.5.7	Considerazioni sulla Sicurezza e Guardrails	79
5.5.8	Sfide Future e Direzioni della Ricerca	79
5.5.9	Conclusioni	79
5.6	Validazione tecnica e metodologia di analisi	80
6	Alter ego virtuali: identità, influenza e responsabilità	83
6.1	Introduzione	83
6.2	Il Metaverso diffuso: dall'ambiente virtuale all'ecosistema incarnato	83
6.3	Ubiquità digitale: dispositivi mobili, edge computing e continuità esperienziale	84
6.4	Intelligenze artificiali e ridefinizione del sé digitale	85
6.5	Il ruolo degli influencer virtuali: IA generative incarnate e potere persuasivo	85
6.5.1	Casi emblematici e strategie di costruzione	86
6.5.2	Architettura tecnologica e integrazione multimodale	86
6.5.3	Estetica, antropomorfismo e dinamiche persuasive	88
6.6	Impatti comportamentali e culturali: imitazione e modellamento nei giovani	89
6.6.1	Modellamento e Teoria Cognitiva Sociale	89

6.6.2	Aspirazioni e salute comportamentale	90
6.7	L'influenza degli avatar realistici sulle scelte morali: l'Effetto Proteus . . .	91
6.7.1	La ricreazione virtuale dell'esperimento di Milgram	91
6.7.2	Incarnazione virtuale e riduzione dei pregiudizi	91
6.8	Implicazioni etiche: rappresentazione postuma, memoria e responsabilità .	92
6.8.1	Personhood postuma e agenzia distribuita	92
6.8.2	Memoria, consenso e responsabilità comunicativa	92
6.9	Verso un'Etica della Coesistenza Virtuale	93
6.10	Un caso studio: il progetto Alterego	93
6.10.1	Introduzione	94
6.10.2	Creazione dell'avatar 3D	94
6.10.3	Conversational Module: Text Generation and Speech Synthesis . .	96
6.10.4	Risultati	105
6.10.5	Discussione	106
6.10.6	Conclusioni	109
7	La convergenza delle tecnologie, ovvero il progetto ALTEREGO e la Morale	
	Artificiale	110
7.1	Introduzione: la Fase 5	110
7.2	Creazione Rapida di Avatar: Il Passaggio ad Avaturn	111
7.3	Sintesi Vocale ad Alte Prestazioni: L'Integrazione di LMNT - Emotive AI speech for voiceover & characters	112
7.4	Il Cervello dell'Alterego Morale: Orchestrazione Multi-Agente con CrewAI	113
7.4.1	Passaggio dal Tree-of-Thought a CrewAI	113
7.4.2	Configurazione della "Crew" per l'Alterego Morale	113
7.4.3	Implementazione Locale di LLM: Llama e DeepSeek per la Privacy e la Sicurezza	114
7.4.4	Il Ruolo di DeepSeek-R1 nel Ragionamento Morale	114
7.4.5	Integrazione della Knowledge Base e replicazione dello stile comu- nicativo	114
7.5	Applicazione Pratica: CrewAI in Azione per la Deliberazione Morale . . .	115
7.6	Analisi Comparativa: Realismo vs. Rapidità e Latenza	116
7.6.1	Latenza e Ragionamento	117
7.7	Validazione Tecnica	117
7.8	Conclusioni e Sviluppi Futuri	118
8	Sintesi Sperimentale, Contributi Originali e Frontiere della Convergenza Robo-	
	tica	120
8.1	Sintesi dei risultati delle fasi di sperimentazione	120
8.2	Originalità del contributo e metodologia di valutazione	123
8.3	Ambito e limiti intenzionali della ricerca	125

8.4	Limiti strutturali e criticità emerse	126
8.5	Prospettive future: verso una convergenza tra simulazione e incarnazione fisica	127

Capitolo 1

Introduzione

1.1 Il Contesto Attuale: IA Generativa, Metaverso e Moralità Artificiale

Il panorama digitale contemporaneo è caratterizzato da un'esplosione tecnologica senza precedenti, guidata principalmente dall'Intelligenza Artificiale Generativa (IA Generativa) e dal Metaverso. Questi fenomeni non sono solo innovazioni tecniche, ma veri e propri "boom tecnologici" che stanno ridefinendo le interazioni umane, il lavoro e la società nel suo complesso.

Il Metaverso si presenta oggi come un ecosistema digitale in rapida evoluzione, ben oltre la semplice realtà virtuale (VR). Comprende la Realtà Aumentata (AR), che sovrappone informazioni digitali al mondo reale, e la Realtà Mista (MR), che consente interazioni fluide tra oggetti reali e virtuali. Le piattaforme attuali spaziano da mondi virtuali persistenti basati su blockchain, come Decentraland e The Sandbox, dove gli utenti possono possedere asset digitali e partecipare a economie interne, a piattaforme sociali immersive come Meta Horizon Worlds, accessibili anche tramite smartphone. Contrariamente a una diffusa percezione, l'interazione con il Metaverso non richiede esclusivamente visori VR costosi; è possibile accedere ad esso tramite smartphone, personal computer e browser web, rendendolo un ambiente "hardware-agnostic" [146]. Questa accessibilità democratizzata è cruciale per la sua adozione di massa e per la sua crescente influenza sulle relazioni umane e sulle interazioni sociali. La convergenza tra realtà fisica e digitale, definita "spatial computing", suggerisce che l'ambiente reale stesso sta diventando un'interfaccia, con elementi digitali che si integrano nella nostra percezione quotidiana [73].

Parallelamente, l'IA Generativa ha raggiunto una preminenza senza precedenti, capace di creare nuovi testi, immagini, video, audio o codice [56]. Il lancio di ChatGPT di OpenAI nel novembre 2022 ha segnato un momento decisivo, portando a una rapida diffusione e a sperimentazioni in diverse applicazioni. I modelli più avanzati sono multimodali, in grado di comprendere e rispondere a input sia testuali che visivi [23]. Grandi aziende tecnologiche come Google, Amazon, Apple, Microsoft e Meta stanno investendo massicciamente in

questo settore, consolidando l'IA Generativa come una tecnologia pervasiva con un impatto trasversale su quasi tutti i settori economici e produttivi.

L'impatto di queste tecnologie sui social media è profondo. L'IA Generativa facilita la creazione di contenuti dinamici e personalizzati, mentre il Metaverso offre nuove piattaforme immersive per l'interazione sociale. Gli utenti possono utilizzare avatar che li rispecchiano fedelmente per partecipare a eventi virtuali globali e interagire con persone di diverse lingue e culture, grazie anche a strumenti di traduzione in tempo reale basati sull'IA. Questa fusione sta trasformando il modo in cui le persone si connettono, si esprimono e consumano contenuti online [216].

I benefici di questa sinergia sono molteplici e pervasivi. L'IA Generativa aumenta la produttività e l'efficienza automatizzando compiti ripetitivi e fornendo insight basati sui dati. Agisce come un potente partner di brainstorming, stimolando la creatività in vari campi [95]. Nel settore dell'educazione, il Metaverso potenziato dall'IA offre esperienze di apprendimento personalizzate e adattive, con laboratori virtuali, visite a musei immersivi e avatar IA che insegnano lingue straniere o assistono studenti con difficoltà [238]. Nel lavoro e nell'industria, i luoghi di lavoro virtuali facilitano la collaborazione a distanza e i "gemelli digitali" (Digital Twins) consentono la simulazione e l'ottimizzazione dei processi produttivi, riducendo costi e rischi [90]. La sanità beneficia di diagnosi supportate dall'IA, riabilitazione cognitiva in VR e telemedicina accessibile [191]. Queste tecnologie promuovono l'inclusione sociale, abbattendo barriere fisiche e culturali per persone con disabilità e comunità emarginate, offrendo nuove opportunità di impiego e connessione globale. Anche l'intrattenimento si espande con giochi di ruolo, concerti e mostre d'arte immersive [48].

Tuttavia, questi progressi portano con sé significativi lati negativi e sfide etiche. La proliferazione di contenuti generati dall'IA, in particolare i deepfake, rappresenta una minaccia crescente per la privacy, la sicurezza e l'integrità delle informazioni. I criminali informatici sfruttano l'IA Generativa per creare esche di phishing più convincenti e superare le barriere linguistiche. La capacità dell'IA di produrre contenuti indistinguibili dalla realtà mina pertanto la fiducia epistemologica, portando al "**liar's dividend**", dove qualsiasi contenuto può essere liquidato come falso [165, 52].

Nel Metaverso, i criminali informatici possono migrare dai forum del dark web a questi nuovi spazi, perpetrando furti di dati, esposizione online e hate speech. Il cyberbullismo è una sfida chiave negli ambienti immersivi, con episodi di bullismo e aggressione già segnalati [213, 192]. Le preoccupazioni per la privacy sono estremamente elevate, con gli utenti che temono la raccolta, l'uso e la condivisione di dati biometrici, movimenti oculari e gesti, molto più personali rispetto a quelli raccolti dalle app tradizionali.

Per affrontare questi rischi, è fondamentale l'implementazione di solide misure di controllo e difesa, che includano la governance dell'IA e linee guida etiche. È in questo contesto che emerge l'importanza di creare agenti virtuali morali che aiutino gli utenti ad adottare comportamenti etici in un mondo sempre più interconnesso e digitale.

1.2 Obiettivi della Ricerca

La presente ricerca di dottorato si propone di esplorare tecniche e architetture per la realizzazione di agenti morali artificiali (virtuali e/o robotici) in grado di interpretare le intenzioni dell'utente e supportarlo nel processo decisionale etico, contribuendo a evitare scelte potenzialmente dannose per sé e per gli altri.

Tali agenti non sono concepiti come sistemi coercitivi o prescrittivi, bensì come mediatori deliberativi non coercitivi, capaci di fornire suggerimenti e valutazioni morali lasciando all'utente la piena autonomia decisionale [66].

Alla luce dei recenti sviluppi nel campo dell'IA Generativa, la ricerca fa uso di Large Language Models (LLM) integrati in architetture cognitive ibride, al fine di sviluppare sistemi in grado di elaborare il contesto, interpretare le intenzioni e produrre risposte moralmente pertinenti. Particolare attenzione è stata dedicata al ruolo dei dati di addestramento, considerati elemento critico per l'imparzialità e l'affidabilità del comportamento degli agenti [126]. Nelle fasi più avanzate del lavoro, tale approccio è stato esteso a sistemi multi-agente basati su IA generativa.

Dal punto di vista metodologico, la ricerca si colloca esplicitamente nel paradigma della **Design Science Research (DSR)**, adottando un approccio orientato alla progettazione, implementazione e iterazione di artefatti software. In questa prospettiva, l'obiettivo principale non è la verifica sperimentale di ipotesi su campioni di utenti, bensì l'esplorazione della fattibilità tecnica e concettuale di agenti morali artificiali situati in ambienti immersivi.

Di conseguenza, la validazione dei sistemi proposti è stata condotta principalmente attraverso analisi di tipo tecnico e qualitativo in scenari simulati e controllati, con particolare attenzione alla coerenza architeturale, alla capacità deliberativa e alla pertinenza contestuale delle risposte, piuttosto che alla misurazione empirica dell'impatto comportamentale sugli utenti.

La ricerca affronta inoltre il problema del trasferimento di tali soluzioni dal dominio virtuale a quello fisico, esplorando il potenziale impiego di agenti morali artificiali in contesti robotici e di robotica assistiva [216, 131].

Il contributo principale della tesi consiste nella progettazione e validazione architeturale di un agente morale artificiale situato, concepito come mediatore deliberativo non coercitivo in ambienti immersivi. L'agente integra interpretazione delle intenzioni dell'utente, ragionamento etico contestuale e interazione uomo-IA all'interno di architetture cognitive ibride che combinano componenti simboliche e modelli di IA Generativa.

La ricerca non mira alla definizione di un modello normativo universale di moralità artificiale, bensì all'esplorazione della fattibilità tecnica e concettuale di sistemi capaci di supportare il processo decisionale morale umano preservandone autonomia e responsabilità.

In particolare, la tesi fornisce i seguenti contributi:

- la definizione di un modello architeturale per agenti morali artificiali situati, applicabile ad ambienti immersivi e metaversali;

- l'applicazione del paradigma Design Science Research alla progettazione iterativa di artefatti per la mediazione morale uomo-IA;
- una riflessione concettuale sul ruolo degli agenti artificiali come mediatori morali, distinti sia da sistemi normativi automatici sia da semplici assistenti conversazionali.

1.3 Metodologia di Ricerca

In coerenza con quanto delineato nella sezione precedente, la presente ricerca si colloca nel paradigma della DSR, particolarmente appropriato per lo studio e la progettazione di sistemi complessi nei quali la produzione dell'artefatto costituisce parte integrante del contributo scientifico. In tale prospettiva, la conoscenza non è generata esclusivamente attraverso l'osservazione di fenomeni esistenti, ma emerge dal processo iterativo di progettazione, implementazione e valutazione di soluzioni tecnologiche orientate alla risoluzione di problemi concreti.

Il percorso metodologico si sviluppa lungo una sequenza di fasi tra loro interdipendenti. In una prima fase è stato condotto un inquadramento critico del problema etico negli ambienti immersivi, con particolare attenzione alle dinamiche comportamentali, ai rischi di disinibizione e alle sfide legate alla mediazione morale in contesti virtuali persistenti. Tale analisi ha guidato la definizione dei requisiti concettuali e funzionali degli agenti morali artificiali oggetto di studio.

Successivamente, è stata affrontata la fase di progettazione architettonica, nella quale sono state definite le strutture cognitive e decisionali degli agenti, ponendo particolare enfasi sull'integrazione tra moduli simbolici, componenti di percezione contestuale e modelli di IA Generativa. Questa fase ha condotto all'elaborazione di architetture neuro-simboliche orientate al ragionamento morale situato.

La fase di implementazione prototipale ha tradotto tali architetture in sistemi operativi, sviluppati all'interno di ambienti immersivi real-time. I prototipi realizzati non hanno avuto lo scopo di costituire prodotti finali, bensì di fungere da strumenti sperimentali per verificare la fattibilità tecnica delle soluzioni proposte e per esplorare differenti configurazioni architettoniche.

La validazione si è concentrata su aspetti di natura tecnica e qualitativa, attraverso test condotti in scenari simulati e controllati. In assenza di studi con campioni di utenti esterni, la valutazione non è stata orientata alla misurazione statistica dell'efficacia comportamentale dei sistemi, ma alla verifica della loro plausibilità operativa, coerenza architettonica e capacità deliberativa.

Più precisamente, i prototipi sviluppati nelle diverse fasi della ricerca sono stati osservati e confrontati secondo un insieme di criteri comuni:

- **coerenza architettonica**, intesa come integrazione funzionale tra moduli percettivi, cognitivi, deliberativi e conversazionali;

- **stabilità operativa**, riferita alla capacità del sistema di mantenere un comportamento consistente in scenari ripetuti;
- **performance computazionali**, con particolare attenzione alla latenza delle risposte e alla sostenibilità dell'esecuzione in ambienti immersivi real-time;
- **affidabilità deliberativa**, intesa come capacità di produrre valutazioni morali coerenti rispetto al framework etico e al contesto situazionale;
- **pertinenza contestuale**, ovvero il grado di adeguatezza delle risposte rispetto alle informazioni ambientali e intenzionali disponibili;
- **leggibilità ed esplicabilità del ragionamento morale**, riferite alla capacità dell'agente di rendere trasparenti le premesse e le motivazioni della propria deliberazione.

L'analisi del comportamento dei prototipi è stata quindi condotta principalmente attraverso scenari eticamente rilevanti, costruiti per osservare il modo in cui i sistemi integravano percezione del contesto, interpretazione delle intenzioni, generazione di valutazioni morali e interazione con l'utente.

Infine, i risultati ottenuti nelle diverse fasi sono stati oggetto di un'analisi comparativa e di una riflessione teorica, volta a mettere in relazione le soluzioni progettate con lo stato dell'arte e con il dibattito filosofico e metodologico sugli agenti morali artificiali. Tale riflessione consente di collocare il contributo della ricerca non soltanto sul piano applicativo, ma anche su quello concettuale, evidenziandone limiti, potenzialità e sviluppi futuri.

1.3.1 Sintesi del processo metodologico

Il processo metodologico adottato può essere sintetizzato nelle seguenti macro-fasi:

1. analisi del problema e definizione dei requisiti;
2. progettazione delle architetture cognitive e deliberative;
3. implementazione prototipale in ambienti immersivi;
4. validazione tecnico-qualitativa in scenari simulati;
5. analisi comparativa e riflessione teorica sui risultati.

1.4 Struttura della tesi

La presente tesi è strutturata nel seguente modo. Nel Capitolo 2 sarà esplorato il concetto di Metaverso, la sua teorizzazione, nascita, sviluppi e le sue architetture attuali. Il Capitolo 3 fornirà una panoramica sull'IA Generativa, con particolare attenzione agli LLM, alla generazione di immagini e video, e ai sistemi di agenti generativi. Successivamente, il

Capitolo 4 si concentrerà sugli Agenti Morali Artificiali, delineandone i concetti fondamentali e il dibattito filosofico. Il Capitolo 5 mostra tutti i casi studio affrontati durante il percorso di dottorato, ponendo enfasi sulle tecniche di realizzazione dei prototipi che via via hanno portato al primo risultato, ovvero la realizzazione di un mediatore morale virtuale in grado di percepire il contesto e interagire con l'utente umano. Proseguendo, il Capitolo 6 descrive le attività svolte all'interno del progetto PRIN ALTEREGO dedicato alla realizzazione di controparti digitali degli esseri umani. Il Capitolo 7 unisce la ricerca del capitolo precedente al lavoro sugli agenti morali unendo gran parte delle tecnologie adottate lungo il percorso nella realizzazione di un Alterego Virtuale Morale. Il Capitolo 8 riassumerà infine le conclusioni della ricerca e delinea le prospettive future.

Capitolo 2

Il Metaverso: Dalle Origini Teoriche e Fantascientifiche alle Applicazioni Pratiche

2.1 Introduzione: definizione e stato dell'arte del Metaverso

Il concetto di Metaverso ha acquisito crescente rilevanza nel dibattito scientifico e tecnologico contemporaneo, soprattutto in seguito alla rinnovata attenzione industriale e mediatica degli ultimi anni [138]. Sebbene il termine sia divenuto di uso comune solo recentemente, esso affonda le proprie radici in una più ampia tradizione di riflessione sul cyberspazio, sui mondi virtuali e sull'interazione uomo–macchina mediata da ambienti digitali immersivi.

2.1.1 Origini teoriche e sviluppo del concetto

Il termine *Metaverso* fu introdotto da Neal Stephenson nel romanzo *Snow Crash* (1992), nel quale veniva descritto uno spazio virtuale persistente e condiviso, accessibile attraverso avatar [209, 115]. A partire da questa formulazione iniziale, il concetto si è progressivamente evoluto, influenzando lo sviluppo di mondi virtuali e ambienti interattivi tridimensionali, fino ad arrivare alle più recenti piattaforme immersive e sociali [166].

Sul piano teorico, il Metaverso può essere interpretato come una possibile evoluzione del cyberspazio verso forme più persistenti, incarnate e interattive, nelle quali la distinzione tra esperienza fisica e presenza digitale tende progressivamente ad attenuarsi [92].

2.1.2 Definizioni accademiche e caratteristiche fondamentali

Nonostante la crescente attenzione, non esiste ancora una definizione univoca e universalmente condivisa di Metaverso [209]. In termini generali, esso può essere descritto come un insieme di ambienti digitali tridimensionali, persistenti, interconnessi e multiutente, nei

quali gli individui interagiscono attraverso rappresentazioni incarnate di sé, comunemente denominate avatar [148, 230, 136].

Ai fini della presente ricerca, risultano particolarmente rilevanti alcune caratteristiche fondamentali:

- **persistenza**, intesa come continuità dell'ambiente e delle interazioni nel tempo;
- **immersività**, ossia la capacità del sistema di generare un senso di presenza situata;
- **multiutenza e socialità**, che rendono possibile l'interazione simultanea tra soggetti umani e artificiali;
- **avatarietà e identità digitale**, che permettono forme di rappresentazione incarnata e continuità identitaria;
- **integrazione tra fisico e digitale**, che rende il Metaverso un contesto ibrido, potenzialmente estendibile anche verso forme di robotica incarnata.

In questa prospettiva, il Metaverso non viene qui inteso come una singola piattaforma o tecnologia, bensì come un ecosistema socio-tecnico risultante dalla convergenza di ambienti virtuali persistenti, interfacce immersive e modelli avanzati di interazione digitale [106, 30, 119].

2.1.3 Stato attuale dello sviluppo del Metaverso

Nonostante l'elevato interesse scientifico, industriale e mediatico, il Metaverso si presenta ancora come una configurazione tecnologica in evoluzione piuttosto che come un'infrastruttura pienamente consolidata [209]. Le attuali piattaforme immersive e le applicazioni di realtà virtuale e aumentata possono essere considerate, più propriamente, come forme preliminari o parziali di tale paradigma.

Permangono infatti numerose criticità legate alla maturità dell'hardware, alla qualità dell'interazione incarnata, alla latenza, alla sostenibilità computazionale e alla capacità di supportare esperienze realmente persistenti e socialmente dense [30]. Tuttavia, proprio questa fase di transizione rende il Metaverso un contesto di particolare interesse per la presente ricerca, in quanto consente di osservare in forma emergente le trasformazioni dell'interazione sociale, dell'identità digitale e dei processi decisionali morali in ambienti artificiali situati.

2.2 Potenzialità del Metaverso: ambienti immersivi, interazione sociale e simulazione

Il Metaverso può essere interpretato non soltanto come uno spazio digitale immersivo, ma come un contesto socio-tecnico nel quale pratiche sociali, identità digitali e forme di interazione

uomo–macchina vengono sperimentate in modo situato e persistente. In questa prospettiva, il suo interesse per la presente ricerca non risiede tanto nella promessa di una “realtà alternativa”, quanto nella possibilità di osservare e progettare dinamiche relazionali, deliberative e comportamentali all’interno di ambienti controllabili ma socialmente significativi [238, 148, 136, 30, 119].

2.2.1 Il Metaverso come ambiente di sperimentazione situata

Uno degli aspetti più rilevanti del Metaverso consiste nella sua capacità di fungere da ambiente di simulazione e sperimentazione per attività complesse che richiedono interazione, apprendimento, collaborazione e presa di decisione in contesti contestualizzati. Ambienti immersivi di questo tipo consentono infatti di costruire scenari nei quali gli utenti possono agire, esplorare, apprendere e confrontarsi con situazioni realistiche o verosimili, mantenendo al tempo stesso un elevato grado di controllo progettuale [30, 138].

Tale caratteristica rende il Metaverso particolarmente rilevante non solo per applicazioni educative o formative, ma anche come laboratorio per l’analisi di comportamenti sociali, processi decisionali e interazioni eticamente sensibili. In questo senso, la sua importanza per la presente ricerca risiede soprattutto nella possibilità di costruire scenari immersivi nei quali osservare e testare il ruolo di agenti artificiali capaci di interagire con l’utente in modo situato.

2.2.2 Interazioni sociali, presenza e identità digitale

La dimensione sociale costituisce uno degli elementi fondanti del Metaverso. A differenza delle tradizionali piattaforme bidimensionali, gli ambienti immersivi permettono forme di co-presenza più intense, sostenute dall’integrazione di segnali visivi, uditivi e, in alcuni casi, aptici [80, 114, 148]. Tale configurazione contribuisce a rafforzare il senso di presenza, inteso sia come percezione di trovarsi nello spazio virtuale, sia come esperienza di condivisione dello spazio con altri soggetti [49, 173, 106].

In questo contesto, l’avatar assume un ruolo centrale come dispositivo di incarnazione e mediazione identitaria. La percezione dell’avatar come estensione del sé, insieme alla possibilità di modulare il proprio grado di realismo, anonimato o continuità identitaria, rende il Metaverso un contesto particolarmente significativo per lo studio dei comportamenti sociali e morali in ambienti artificiali [136, 238]. Proprio questa combinazione tra presenza, rappresentazione del sé e interazione sociale rende tali ambienti particolarmente rilevanti per la progettazione di agenti morali artificiali situati.

2.2.3 Digital twin e alter ego virtuali

Un ulteriore elemento di interesse è rappresentato dall’impiego di *digital twin* e alter ego virtuali. Originariamente sviluppati come repliche digitali dinamiche di oggetti, sistemi o

processi fisici, i digital twin trovano nel Metaverso un'estensione naturale verso ambienti complessi, contesti sociali e, in prospettiva, anche rappresentazioni personalizzate dell'individuo [160].

Rispetto a forme più generiche di rappresentazione avatariale, i digital twin introducono un livello maggiore di continuità tra mondo fisico e mondo virtuale, rendendo possibile una simulazione più coerente di stati, comportamenti e interazioni. Questo aspetto risulta particolarmente rilevante per la presente ricerca, poiché consente di interpretare il Metaverso non solo come spazio di esperienza, ma anche come infrastruttura sperimentale per la progettazione, il test e l'evoluzione di agenti artificiali situati.

In tale prospettiva, il Metaverso si configura come un laboratorio avanzato nel quale è possibile osservare l'interazione tra identità digitale, presenza sociale, simulazione contestuale e supporto cognitivo, offrendo così un contesto privilegiato per l'esplorazione di mediatori morali artificiali.

2.3 Rischi e Sfide del Metaverso: Aspetti Critici da Considerare

La diffusione del Metaverso apre scenari di grande interesse sul piano tecnologico, educativo e sociale, ma introduce al contempo rischi di natura psicologica, relazionale ed etica che richiedono un'analisi sistematica. La stessa infrastruttura che consente di promuovere empatia, collaborazione e apprendimento esperienziale può infatti favorire, in assenza di adeguate tutele, dinamiche di disinibizione, aggressività e deresponsabilizzazione. I comportamenti sviluppati nello spazio virtuale non restano confinati a esso: esiste una permeabilità tra ambiente digitale e mondo fisico, tale per cui atteggiamenti e pratiche possono trasferirsi dall'uno all'altro contesto.

Un elemento centrale di questa trasformazione è il senso di "presenza" – fisica, sociale e del sé – che costituisce il principale fattore di attrattività degli ambienti immersivi. Se da un lato tale caratteristica favorisce interazioni più intense e coinvolgenti, dall'altro può determinare un progressivo disinvestimento dalle relazioni offline, con possibili ricadute in termini di isolamento e riduzione della partecipazione sociale nel mondo reale. Ne deriva una questione progettuale cruciale: le piattaforme del Metaverso devono essere concepite non solo per massimizzare l'engagement, ma anche per garantire un'integrazione equilibrata con la vita quotidiana degli utenti [49]. Orientare lo sviluppo verso modelli sostenibili e responsabili implica quindi la promozione di comportamenti pro-sociali e la prevenzione di quelli anti-sociali, riconoscendo che le esperienze virtuali producono effetti concreti.

2.3.1 Disinibizione e Comportamenti Inappropriati

Uno dei fattori che incidono maggiormente sulle dinamiche comportamentali è la possibilità di adottare identità digitali non direttamente riconducibili alla persona fisica. L'uso di avatar

anonimi o multipli amplia le opportunità di espressione e sperimentazione identitaria [92], ma può anche attenuare il senso di responsabilità individuale, facilitando condotte inappropriate o illecite [103].

Rispetto ai social network tradizionali, la natura immersiva e multisensoriale del Metaverso tende ad amplificare l'effetto di disinibizione online. Le esperienze negative – incluse quelle con contenuti violenti o abusivi – possono generare reazioni psicologiche e fisiologiche comparabili a quelle del mondo fisico [7]. L'intensità dell'esperienza rende pertanto più rilevanti le conseguenze di comportamenti scorretti.

Gli avatar svolgono un ruolo centrale in tali dinamiche. Essi non rappresentano semplici immagini digitali, ma veri e propri dispositivi di mediazione psicologica che influenzano il senso di incarnazione, l'autopercezione e le intenzioni comportamentali dell'utente. La possibilità di costruire versioni idealizzate di sé può ridurre barriere sociali e favorire l'interazione; tuttavia, può anche alimentare confronto sociale, insoddisfazione o confusione identitaria. La somiglianza tra individuo e avatar incide inoltre sulle scelte comportamentali, rendendo la progettazione delle opzioni di personalizzazione un elemento con implicazioni psicologiche significative [222]. Il Metaverso si configura così come uno spazio privilegiato di esplorazione del sé, ma anche come potenziale fonte di vulnerabilità.

2.3.2 Alienazione e Impatto sul Benessere Psicologico

Gli effetti del Metaverso sulla salute mentale sono ancora oggetto di approfondimento, ma la letteratura individua diversi profili di rischio: cybersickness, dipendenza, accentuazione del confronto sociale, fenomeni di depersonalizzazione o derealizzazione, comportamenti problematici legati all'anonimato e timori connessi al furto di identità [108]. Tali fattori incidono direttamente sul benessere individuale e sulla qualità delle relazioni.

L'elevato grado di immersione costituisce un elemento critico nello sviluppo di dinamiche addictive [103]. Bambini e adolescenti risultano particolarmente esposti, poiché il sistema di controllo cognitivo e di regolazione degli impulsi non è ancora pienamente consolidato [205]. Le possibili conseguenze includono isolamento sociale, difficoltà attentive, sintomi ansioso-depressivi e riduzione del rendimento scolastico.

La costruzione di avatar idealizzati può inoltre incoraggiare forme di evasione rispetto a un'immagine di sé percepita come inadeguata. Se da un lato ciò può facilitare l'inclusione e l'espressione personale, dall'altro può accentuare processi di frammentazione del sé, soprattutto nelle fasi evolutive più delicate [222]. Un utilizzo intensivo e non regolato degli ambienti immersivi può infine ridurre le occasioni di interazione offline, con possibili ricadute sulle competenze relazionali e sull'insorgenza di ansia sociale [205].

La sfida consiste dunque nel preservare i benefici dell'immersione limitandone gli effetti distorsivi, attraverso strategie di progettazione e regolazione orientate alla tutela della salute psicologica.

2.3.3 Manipolazione Comportamentale

L'integrazione dell'intelligenza artificiale e di sistemi algoritmici avanzati introduce ulteriori criticità, in particolare rispetto al consenso informato, alla proprietà dei dati e alla trasparenza dei processi decisionali automatizzati. Le piattaforme XR si basano su una raccolta estesa di dati biometrici, comportamentali e cognitivi, con implicazioni rilevanti per la protezione della privacy e la sicurezza informativa [106].

Gli algoritmi impiegati per la personalizzazione dei contenuti, la profilazione e la pubblicità possono influenzare in modo significativo – e talvolta non consapevole – le decisioni degli utenti [81]. In questo contesto, la personalizzazione rischia di trasformarsi in sorveglianza pervasiva e in condizionamento comportamentale, soprattutto se combinata con tecniche persuasive aggressive. Tali pratiche possono sfruttare vulnerabilità individuali, contribuendo ad alimentare meccanismi di dipendenza o dinamiche di esclusione.

I diversi profili di rischio risultano strettamente interconnessi: criticità legate alla protezione dei dati possono innescare fenomeni di furto di identità, manipolazione e danno psicologico. Ne emerge la necessità di un approccio olistico alla governance del Metaverso, capace di integrare dimensioni tecnologiche, normative ed etiche in una prospettiva multidisciplinare.

2.4 Conclusioni

Il Metaverso si configura come un ecosistema complesso e in costante trasformazione, le cui origini concettuali affondano nella letteratura fantascientifica e trovano oggi concreta espressione nell'integrazione di tecnologie avanzate quali realtà virtuale e aumentata, intelligenza artificiale, blockchain e digital twin. La convergenza di tali strumenti apre scenari applicativi di ampia portata nei contesti educativi, produttivi e sociali, offrendo esperienze immersive e personalizzate capaci di superare i vincoli della dimensione fisica. Allo stesso tempo, questa evoluzione tecnologica impone una riflessione critica sui suoi impatti sistemici.

Il Metaverso non rappresenta soltanto uno spazio alternativo di interazione, ma anche un laboratorio sperimentale per l'analisi dei comportamenti individuali e collettivi. In particolare, l'integrazione dei digital twin evidenzia la sua natura duale: da un lato ambiente virtuale autonomo, dall'altro strumento di simulazione e ottimizzazione della realtà fisica. Le dinamiche osservate negli ambienti immersivi possono infatti generare conoscenze utili al miglioramento di processi, servizi e infrastrutture nel mondo reale, rafforzando il ruolo del Metaverso come piattaforma di innovazione.

Al centro di tale esperienza si collocano i concetti di immersione e presenza, che incidono profondamente sui processi cognitivi ed emotivi degli utenti. Un'elevata immersione, resa possibile da feedback multisensoriali e da un'elevata fedeltà simulativa, accresce il coinvolgimento e la gratificazione percepita. Tuttavia, proprio questa intensificazione dell'esperienza può favorire meccanismi di uso problematico, specialmente in presenza di stimoli ripetitivi e sistemi di ricompensa progettati per massimizzare il tempo di permanenza. Si delinea

così una catena causale in cui le scelte di design tecnologico influenzano direttamente il comportamento degli utenti, generando potenziali vulnerabilità.

Tali vulnerabilità possono essere ulteriormente amplificate attraverso pratiche di monitoraggio passivo e strategie di persuasione algoritmica, con il rischio di compromettere l'autonomia decisionale e il benessere psicologico. Ne emerge un ciclo di retroazione in cui immersione, raccolta dati e personalizzazione spinta si rafforzano reciprocamente, rendendo imprescindibile l'adozione di modelli di intelligenza artificiale etica, solidi sistemi di governance dei dati e meccanismi efficaci di tutela dell'utente.

Alla luce di queste considerazioni, lo sviluppo futuro del Metaverso richiede un approccio strutturato lungo alcune direttrici fondamentali. In primo luogo, è necessaria una governance etica supportata da quadri normativi chiari e aggiornati, capaci di affrontare questioni quali privacy, sicurezza dei dati, protezione degli utenti e divario digitale [141]. In secondo luogo, occorrono ulteriori progressi tecnologici, soprattutto in termini di hardware, riduzione della latenza e integrazione avanzata dell'intelligenza artificiale, al fine di garantire interoperabilità e continuità esperienziale.

Parallelamente, risulta essenziale promuovere una ricerca interdisciplinare che coinvolga informatica, psicologia, sociologia, diritto ed etica, così da analizzare in modo sistemico le implicazioni sociali e comportamentali delle tecnologie immersive [125]. Un ruolo centrale spetta inoltre all'educazione e all'alfabetizzazione digitale, strumenti indispensabili per sviluppare consapevolezza critica e resilienza nell'uso degli ambienti virtuali. Infine, la progettazione dovrà essere orientata a criteri di inclusività e accessibilità, prevenendo l'ampliamento delle disuguaglianze esistenti.

In definitiva, il Metaverso rappresenta una delle trasformazioni più rilevanti dell'attuale paradigma socio-tecnologico. La sua evoluzione potrà tradursi in un fattore di progresso solo attraverso un equilibrio consapevole tra innovazione e responsabilità, capace di valorizzarne le potenzialità senza trascurarne le implicazioni etiche e sociali.

Capitolo 3

IA Generativa e Sistemi Generativi Multi-Agente

3.1 Introduzione all'IA Generativa

L'IA Generativa rappresenta un sottoinsieme avanzato dell'intelligenza artificiale caratterizzato dalla capacità di produrre contenuti nuovi e plausibili — quali testo, immagini, audio, video o codice — a partire dall'apprendimento di strutture statistiche e rappresentazioni latenti derivate dai dati di addestramento. Questo paradigma segna un'evoluzione significativa rispetto ai sistemi tradizionali, storicamente orientati alla classificazione, previsione o analisi di dati esistenti, introducendo modelli capaci di contribuire attivamente alla produzione di contenuti e alla co-creazione uomo-macchina, con implicazioni rilevanti in termini di autorialità, affidabilità e controllo. Alla base di tale trasformazione si collocano i *Foundation Models* (FM) e, in ambito linguistico, gli *LLM*, modelli addestrati su grandi quantità di dati prevalentemente non etichettati e successivamente adattabili a molteplici compiti [15].

A differenza dell'IA tradizionale, che spesso si fonda su approcci discriminativi e su basi di conoscenza esplicite, l'IA Generativa si basa su modelli capaci di apprendere la distribuzione dei dati e generare nuove istanze plausibili, ponendo in primo piano aspetti quali coerenza, plausibilità e controllabilità dei risultati. Il rapido sviluppo della GenAI negli ultimi anni è il risultato della convergenza di tre fattori principali: l'aumento della potenza computazionale, la disponibilità di grandi quantità di dati e l'introduzione di architetture più efficaci, tra cui i *Transformer*, che hanno rivoluzionato l'elaborazione del linguaggio naturale grazie al meccanismo di attenzione [196]. Su queste basi si è affermato il paradigma dei modelli su larga scala, in grado di generalizzare su una vasta gamma di compiti e di operare in contesti conversazionali e multimodali.

La diffusione di strumenti accessibili al grande pubblico, in particolare sistemi conversazionali come ChatGPT, ha contribuito a trasformare la GenAI da ambito specialistico a tecnologia di uso diffuso, rendendo possibile l'interazione diretta con modelli linguistici avanzati anche da parte di utenti non esperti. In questo scenario, l'IA Generativa si configura non come un singolo strumento, ma come un'infrastruttura trasversale capace di ridefinire

processi cognitivi, creativi e decisionali.

Per la presente ricerca, il suo interesse principale risiede nella possibilità di costruire agenti artificiali in grado di interagire linguisticamente, interpretare il contesto e sostenere processi deliberativi complessi, rendendo gli LLM un elemento centrale nella progettazione di agenti morali artificiali situati.

3.2 Benefici e vantaggi dell'IA Generativa

L'IA Generativa sta assumendo un ruolo crescente nei processi cognitivi, organizzativi e creativi, non soltanto per la sua capacità di automatizzare attività esistenti, ma soprattutto per la possibilità di ridefinire il rapporto tra attività umana e supporto computazionale. I suoi benefici principali non risiedono esclusivamente nell'efficienza operativa, ma anche nella capacità di ampliare le possibilità di interazione, personalizzazione e produzione di contenuti in contesti complessi.

Un primo vantaggio rilevante riguarda l'aumento della produttività e dell'efficienza, reso possibile dall'automazione di compiti ripetitivi, strutturati o a basso valore strategico. Sistemi generativi possono supportare la sintesi di documenti, la produzione di testi, la generazione di codice, la rielaborazione di informazioni e la costruzione di output preliminari, riducendo il tempo necessario per attività esecutive e consentendo agli attori umani di concentrarsi maggiormente su obiettivi, supervisione e valutazione critica [232, 12]. In questa prospettiva, l'IA Generativa non si limita a sostituire attività operative, ma tende a spostare il contributo umano verso forme di controllo, orientamento e decisione di livello superiore.

Un secondo beneficio riguarda il potenziamento della creatività e dell'esplorazione progettuale. I modelli generativi possono infatti produrre rapidamente varianti, ipotesi, alternative stilistiche o soluzioni preliminari, favorendo processi di brainstorming, co-creazione e sperimentazione iterativa [232]. Più che sostituire la creatività umana, tali sistemi possono funzionare come amplificatori cognitivi, contribuendo ad ampliare lo spazio delle possibilità espressive, progettuali e concettuali.

Un terzo vantaggio, particolarmente rilevante per la presente ricerca, riguarda la personalizzazione su larga scala. Grazie alla capacità di adattare il linguaggio, il tono, il contenuto e la forma dell'interazione in funzione del contesto e dell'utente, l'IA Generativa consente la costruzione di sistemi più flessibili, adattivi e sensibili alla situazione [132]. Questa caratteristica assume un ruolo centrale nella progettazione di agenti artificiali capaci non solo di rispondere, ma di accompagnare l'utente in processi dialogici e deliberativi più articolati.

Nel complesso, l'IA Generativa può essere interpretata come una tecnologia abilitante trasversale, capace di integrare automazione, supporto cognitivo e adattamento comunicativo. Per la presente ricerca, il suo valore non risiede tanto nella varietà dei possibili domini applicativi, quanto nella possibilità di costruire agenti artificiali in grado di interagire in linguaggio naturale, contestualizzare le risposte, sostenere il ragionamento e modulare la propria condotta comunicativa in funzione dell'utente e della situazione.

3.3 Aspetti negativi e pericoli dell'IA Generativa

Accanto alle opportunità offerte, l'IA Generativa presenta criticità rilevanti che richiedono una riflessione approfondita e un approccio responsabile sia nella fase di sviluppo sia in quella di utilizzo [225]. Tali criticità non riguardano soltanto l'affidabilità tecnica dei sistemi, ma investono anche questioni etiche, sociali, giuridiche e di sicurezza, particolarmente rilevanti nei contesti in cui l'IA è chiamata a interagire con esseri umani in modo continuativo o a supportare processi decisionali complessi.

Un primo nodo critico riguarda i **bias e le distorsioni nei dati e nei modelli**. Poiché i sistemi generativi apprendono da grandi quantità di dati storici, essi possono riflettere e amplificare stereotipi, pregiudizi e squilibri culturali già presenti nei corpus di addestramento, producendo contenuti o raccomandazioni non neutrali [186]. Questo aspetto è particolarmente problematico nei contesti in cui l'IA è chiamata a formulare valutazioni, suggerimenti o orientamenti potenzialmente sensibili sul piano etico o sociale.

Una seconda criticità centrale riguarda la **mancanza di trasparenza e spiegabilità**. Molti modelli generativi operano infatti secondo logiche opache, rendendo difficile ricostruire in modo chiaro le ragioni per cui un certo contenuto venga prodotto o una determinata risposta venga formulata. Tale opacità limita la possibilità di controllo, ostacola la responsabilità e riduce l'affidabilità percepita dei sistemi, specialmente nei casi in cui il contenuto generato abbia implicazioni normative, relazionali o decisionali.

Un ulteriore limite strutturale è rappresentato dalla possibilità che i sistemi generativi producano **contenuti plausibili ma errati, fuorvianti o manipolativi**. La capacità di generare output formalmente coerenti non coincide infatti con una comprensione profonda o causalmente fondata del contenuto prodotto. Ciò espone a rischi di disinformazione, semplificazione indebita, allucinazioni e costruzione di risposte persuasive ma inaffidabili, soprattutto in assenza di adeguati meccanismi di verifica, grounding e supervisione [41, 174].

Sul piano individuale, emerge inoltre il rischio di una **delega eccessiva del giudizio e di dipendenza cognitiva**. L'uso continuativo di sistemi generativi può indurre gli utenti a esternalizzare progressivamente attività interpretative, decisionali o argomentative, con possibili effetti sulla qualità del pensiero critico e dell'autonomia valutativa. Questo aspetto risulta particolarmente rilevante nel caso di agenti conversazionali progettati per accompagnare l'utente in processi complessi, dove il confine tra supporto cognitivo e sostituzione del giudizio può diventare sottile.

Ulteriori criticità riguardano la **privacy, la sicurezza e la proprietà intellettuale**. I sistemi generativi possono infatti essere impiegati per produrre contenuti manipolativi, imitare identità, trattare dati sensibili o rielaborare materiale protetto da diritti, aprendo questioni rilevanti sul piano giuridico, etico e progettuale. In questo senso, la crescente potenza generativa dei modelli rende necessario un rafforzamento parallelo dei meccanismi di controllo, accountability e protezione dell'utente.

Nel complesso, questi limiti mostrano come l'IA Generativa, pur offrendo capacità espressive e operative senza precedenti, non possa essere considerata di per sé una base sufficiente

per la costruzione di sistemi affidabili in contesti eticamente sensibili. Per la presente ricerca, tale constatazione è particolarmente rilevante: proprio i limiti di opacità, bias, inaffidabilità e dipendenza cognitiva rendono necessaria l'integrazione dei modelli generativi all'interno di architetture più ampie, dotate di meccanismi di mediazione, supervisione e contestualizzazione.

3.4 Sistemi di agenti basati su IA Generativa

I sistemi di agenti basati su IA Generativa rappresentano una delle evoluzioni più rilevanti dell'intelligenza artificiale contemporanea, in quanto estendono le capacità degli LLM oltre la semplice generazione di testo, integrandole con meccanismi di pianificazione, memoria, controllo e interazione operativa. In questo modo, i modelli linguistici vengono inseriti all'interno di architetture più ampie, capaci di agire in modo autonomo e orientato agli obiettivi in ambienti complessi [82].

3.4.1 Definizione e architettura degli agenti di IA Generativa

Un agente basato su IA Generativa può essere definito come un sistema in grado di percepire un contesto, elaborare informazioni, pianificare azioni e interagire con risorse interne ed esterne per il raggiungimento di obiettivi specifici [93, 145]. A differenza dei sistemi tradizionali, tali agenti combinano capacità linguistiche, ragionamento contestuale e azione operativa.

Dal punto di vista architetturale, questi sistemi si articolano tipicamente in una serie di componenti integrate:

- **Core basato su LLM**, che svolge il ruolo di nucleo cognitivo per l'interpretazione del linguaggio, il ragionamento e la generazione di risposte;
- **Sistemi di memoria**, distinti in memoria di breve termine (contesto corrente) e memoria a lungo termine (conoscenze persistenti e storico delle interazioni);
- **Meccanismi di pianificazione**, che consentono di scomporre obiettivi complessi in sequenze di azioni;
- **Integrazione di strumenti**, che permette all'agente di interagire con API, database o ambienti esterni;
- **Meccanismi di controllo e feedback**, utili per la verifica, la revisione e il miglioramento iterativo delle azioni.

L'integrazione di tali componenti consente di trasformare il modello linguistico da sistema reattivo a elemento centrale di un processo deliberativo più articolato.

3.4.2 Sistemi multi-agente e orchestrazione

Un'evoluzione significativa di questo paradigma è rappresentata dai sistemi multi-agente (Multi-Agent Systems, MAS), nei quali più agenti specializzati cooperano per il raggiungimento di obiettivi complessi. In tali configurazioni, il lavoro viene distribuito tra entità con ruoli distinti, favorendo la decomposizione strutturata dei compiti, la verifica incrociata dei risultati e una maggiore robustezza complessiva del sistema.

Questo approccio introduce una forma di orchestrazione cognitiva distribuita, nella quale il processo decisionale non è affidato a un singolo modello, ma emerge dall'interazione tra più componenti. Ciò risulta particolarmente rilevante nei contesti in cui è richiesta una maggiore trasparenza del ragionamento, una gestione modulare delle funzioni e una capacità di adattamento dinamico.

Nel panorama dei framework disponibili, strumenti come CrewAI permettono di strutturare sistemi multi-agente attraverso la definizione di ruoli, obiettivi e flussi operativi coordinati [3, 197]. In tali configurazioni, ogni agente opera come unità specializzata all'interno di un sistema più ampio, contribuendo a un processo deliberativo distribuito.

3.4.3 Rilevanza per la progettazione di agenti morali artificiali

Per la presente ricerca, i sistemi di agenti basati su IA Generativa assumono un ruolo centrale in quanto consentono di superare alcuni dei limiti intrinseci dei modelli linguistici isolati. In particolare, l'integrazione di memoria, pianificazione, orchestrazione multi-agente e controllo permette di strutturare il processo decisionale in modo più esplicito, modulare e contestualizzato.

Questa architettura risulta particolarmente adatta alla progettazione di agenti morali artificiali situati, nei quali il ragionamento etico non può essere ridotto a una singola risposta generata, ma deve emergere da un processo deliberativo articolato, sensibile al contesto e potenzialmente verificabile. In tale prospettiva, i sistemi multi-agente non rappresentano soltanto una soluzione tecnica, ma un paradigma progettuale per la costruzione di sistemi di mediazione morale non coercitiva.

3.5 Conclusioni

L'IA Generativa rappresenta una delle trasformazioni più significative dell'IA contemporanea, in quanto introduce la possibilità di costruire sistemi capaci non solo di analizzare o classificare dati, ma anche di generare contenuti, sostenere interazioni linguistiche complesse e partecipare a processi cognitivi articolati. L'affermazione degli FM, degli LLM e, più recentemente, dei sistemi di agenti basati su modelli generativi ha ampliato in modo sostanziale il potenziale applicativo dell'IA, aprendo nuove prospettive nella costruzione di sistemi interattivi, adattivi e orientati agli obiettivi.

Tuttavia, proprio le caratteristiche che rendono questi sistemi particolarmente potenti — flessibilità linguistica, capacità generativa e adattamento contestuale — ne evidenziano anche i principali limiti. Bias, opacità, inaffidabilità, possibilità di manipolazione e rischio di delega eccessiva del giudizio mostrano come i modelli generativi, se considerati isolatamente, non costituiscano una base sufficiente per la realizzazione di sistemi affidabili in contesti eticamente sensibili.

In tale prospettiva, l'evoluzione verso architetture agentive e multi-agente assume un ruolo centrale. L'integrazione di memoria, pianificazione, orchestrazione, uso di strumenti esterni e meccanismi di controllo consente infatti di superare almeno in parte i limiti dei modelli linguistici isolati, trasformando la generazione di testo in un processo più strutturato, verificabile e orientato agli obiettivi.

Per la presente ricerca, questi sviluppi risultano particolarmente rilevanti, poiché rendono tecnicamente plausibile la progettazione di agenti artificiali capaci di operare non come semplici interfacce conversazionali, ma come sistemi deliberativi situati, inseriti in ambienti complessi e in grado di supportare l'utente in processi decisionali moralmente rilevanti. In questo senso, l'IA Generativa non rappresenta il punto di arrivo della ricerca, ma la base tecnologica su cui costruire architetture cognitive più articolate, orientate alla mediazione morale, alla contestualizzazione e al supporto non coercitivo dell'agire umano.

Capitolo 4

Agenti Morali Artificiali: Fondamenti, Implicazioni e Prospettive Future

4.1 L'emergere degli Agenti Morali Artificiali nell'ecosistema digitale contemporaneo

La diffusione pervasiva dell'IA nei contesti sociali, istituzionali ed economici ha trasformato radicalmente le modalità di decisione, interazione e governo dei processi complessi. Dalle applicazioni di supporto clinico ai sistemi di raccomandazione, fino agli strumenti di automazione amministrativa e militare, l'IA interviene sempre più frequentemente in ambiti caratterizzati da implicazioni morali rilevanti. Questa espansione funzionale rende imprescindibile un'analisi filosofica ed etica sistematica delle sue modalità operative e dei suoi effetti sociali [72].

In tale scenario si colloca il dibattito sugli Agenti Morali Artificiali (AMA), ossia sistemi di IA progettati per assistere o orientare decisioni che coinvolgono valutazioni etiche. Essi sono definiti come sistemi capaci di analizzare uno scenario dal punto di vista morale e di fornire una valutazione o una raccomandazione coerente con determinati criteri normativi. La loro progettazione implica una collaborazione interdisciplinare tra informatica, filosofia morale, teoria politica e scienze cognitive [198].

L'incremento dell'autonomia decisionale delle tecnologie intelligenti mette in crisi i modelli tradizionali di agenzia morale, storicamente centrati sull'essere umano. Le categorie classiche di responsabilità, imputabilità e intenzionalità risultano spesso inadeguate a descrivere sistemi che operano in modo semi-autonomo o autonomo in ambienti complessi [187]. Di conseguenza, diventa necessario riformulare o ampliare i quadri teorici esistenti, integrando meccanismi capaci di garantire affidabilità, trasparenza e controllo, anche al fine di rafforzare la fiducia nei sistemi computazionali ed evitare comportamenti imprevisti o dannosi [70].

4.2 Agenti Morali Artificiali: definizione, genealogia filosofica e modelli di implementazione

4.2.1 Che cos'è un Agente Morale Artificiale

Un AMA può essere descritto come un sistema di IA progettato per operare in contesti in cui le decisioni comportano conseguenze eticamente rilevanti. Esso è teoricamente in grado di riconoscere dimensioni morali in una situazione, valutare alternative possibili e produrre un esito coerente con determinati principi o criteri normativi [29].

Nel presente lavoro, la nozione di AMA viene adottata in senso funzionale e non ontologico. Essa non indica un soggetto dotato di coscienza morale, intenzionalità fenomenica o responsabilità comparabile a quella umana, bensì un sistema artificiale progettato per partecipare a processi di valutazione, mediazione o supporto etico all'interno di specifici contesti applicativi.

La nozione di “moralità” rinvia, in senso generale, all'insieme di regole, principi e pratiche socialmente condivise che orientano il comportamento individuale e collettivo, rendendo possibile la convivenza e limitando l'arbitrio egoistico. Attribuire a un sistema artificiale la capacità di agire “moralmente” implica dunque riconoscergli almeno una forma operativa di autonomia decisionale e la capacità di distinguere tra alternative valutabili in termini di giusto e sbagliato, anche in presenza di conflitti di valori [198].

Tale attribuzione non equivale necessariamente a riconoscere all'AMA uno status morale analogo a quello umano, ma solleva interrogativi fondamentali sulla natura dell'agenzia, sulla responsabilità e sui criteri di imputazione delle conseguenze delle sue azioni.

Nel contesto della presente ricerca, il termine “moralità” non viene inteso come una proprietà intrinseca o autonoma dell'agente artificiale, né come la capacità di formulare giudizi etici universalmente validi. Esso viene invece adottato in senso operativo, per indicare una funzione di mediazione deliberativa che supporta l'utente umano nel processo decisionale. In particolare, un agente morale artificiale, così come concepito in questa tesi, svolge il ruolo di facilitatore cognitivo in grado di: (i) rendere esplicite le possibili conseguenze morali di un'azione nel contesto dato, (ii) richiamare norme, valori o esempi rilevanti in funzione della situazione, e (iii) stimolare una riflessione consapevole senza esercitare forme di coercizione o sostituzione decisionale.

Ogni riferimento all'agente come “morale” deve pertanto essere interpretato in questo senso funzionale e non metafisico, escludendo qualsiasi attribuzione di agency morale piena o di responsabilità etica autonoma.

Tale nozione va inoltre distinta da quella di semplice assistente conversazionale. Un chatbot tradizionale mira prevalentemente a fornire informazioni o completare richieste dell'utente; un agente morale artificiale, invece, integra finalità deliberative, considera implicazioni valoriali e orienta il dialogo verso esiti maggiormente riflessivi e responsabili.

4.2.2 Radici filosofiche dell'etica delle macchine

Il dibattito sull'Etica delle Macchine precede di diversi decenni l'attuale stagione dell'IA basata su apprendimento automatico. Un momento di particolare rilevanza è rappresentato dal simposio del 2005 dedicato agli AMA, nel quale si discusse il loro possibile impiego in ambiti come la medicina, il diritto e la difesa [110].

Sul piano teorico e culturale, un riferimento ricorrente è costituito dalle “Tre Leggi della Robotica” di Isaac Asimov, che rappresentano un tentativo precoce di codificare in forma normativa il comportamento delle macchine. La prima legge, incentrata sul principio di non-maleficenza, richiama un cardine dell'etica medica e di molte tradizioni morali. Tuttavia, la stessa narrativa di Asimov mostra come anche regole apparentemente semplici diventino problematiche in situazioni in cui il danno minore deve essere bilanciato rispetto a un danno maggiore. Ciò evidenzia i limiti di una formalizzazione puramente regolativa dell'etica.

Parallelamente, la filosofia morale ha offerto modelli teorici che hanno ispirato tentativi di formalizzazione computazionale. L'utilitarismo, con il suo orientamento alla massimizzazione del benessere complessivo, e l'etica kantiana, fondata sull'universalizzabilità delle massime e sul rispetto della dignità, rappresentano due paradigmi frequentemente richiamati nella progettazione degli AMA. In questa prospettiva, l'agente morale è concepito come capace di riconoscere richieste morali e di conformarsi ad esse in modo autonomo, assumendosi una forma di responsabilità proporzionale alla propria capacità decisionale [72].

4.2.3 Dalla teoria alla pratica: modelli di implementazione dell'agenzia morale

La traduzione di principi etici in architetture computazionali ha dato origine a tre principali modelli di implementazione: approcci top-down, bottom-up e ibridi.

Approcci top-down. Gli approcci top-down si fondano sull'incorporazione esplicita di principi morali predefiniti all'interno del sistema. In questo caso, l'architettura decisionale viene guidata da regole o teorie etiche formalizzate a priori. Tali modelli offrono maggiore prevedibilità e consentono una tracciabilità normativa delle decisioni. Esempi includono sistemi come MedEthEx, basato sui principi dell'etica biomedica, o MoralDM, che integra schemi utilitaristici e deontologici [70].

Tuttavia, questi approcci mostrano limiti significativi nel trattare la complessità delle situazioni reali. Le regole possono entrare in conflitto tra loro e risultare insufficienti a coprire eccezioni impreviste, rendendo il sistema eccessivamente rigido [51].

Approcci bottom-up. Gli approcci bottom-up si basano invece sull'apprendimento dai dati, spesso derivati da giudizi morali espressi da individui o comunità. In questo caso, il sistema non riceve una teoria etica completa, ma apprende pattern decisionali a partire da

esempi. L'esperimento Delphi, addestrato su circa 1,7 milioni di giudizi etici, costituisce un caso emblematico [198].

Un altro esempio è l'agente consequenzialista basato su architettura BDI (Belief-Desire-Intention) [60], che utilizza il ragionamento basato su casi per inferire decisioni morali; un altro è costituito da sistemi che aggregano preferenze sociali attraverso meccanismi di voto su dilemmi morali.

Tuttavia, i risultati hanno evidenziato gravi criticità, tra cui la riproduzione di bias sociali e discriminazioni, come nel caso di valutazioni influenzate da pregiudizi razziali. Ciò dimostra che l'apprendimento dai dati umani, in assenza di vincoli normativi robusti, può amplificare distorsioni preesistenti. Inoltre, tali sistemi spesso soffrono di scarsa spiegabilità e di difficoltà di generalizzazione in contesti nuovi.

Approcci ibridi. La tensione tra rigidità normativa e vulnerabilità ai bias ha condotto allo sviluppo di approcci ibridi, che combinano principi morali formalizzati con meccanismi di apprendimento adattivo. L'obiettivo è integrare "principi morali rigorosi" con l'apprendimento da casi reali, al fine di produrre giudizi più coerenti, contestualizzati e giustificabili [70].

In questa prospettiva si collocano anche modelli che integrano regole morali con dinamiche emozionali artificiali e sistemi che apprendono valori etici da dati entro cornici normative predefinite. I lavori di Stenseke rappresentano un contributo significativo, avendo implementato agenti virtuosi artificiali basati su virtù disposizionali, apprendimento bottom-up e ricompensa eudaimonica top-down, mediante reti neurali e tecniche di classificazione [178, 179].

Questi modelli cercano un equilibrio tra prescrizione e adattabilità, riconoscendo che la moralità umana è intrinsecamente plurale, dinamica e talvolta incoerente. La questione centrale non riguarda soltanto le modalità tecniche di implementazione, ma anche la scelta dei principi di riferimento: quale etica deve essere incorporata? In che modo garantire equità, robustezza e trasparenza in presenza di conflitti valoriali?

L'evoluzione dagli schemi regolativi ispirati ad Asimov ai modelli di apprendimento basati su grandi moli di dati, fino alle soluzioni ibride contemporanee, riflette una crescente consapevolezza delle difficoltà filosofiche e pratiche dell'etica dell'IA. Lo sviluppo degli AMA non consiste nella semplice replica della moralità umana, bensì nella costruzione di sistemi capaci di operare in contesti caratterizzati da ambiguità, conflitti e incertezza normativa. In questo senso, la progettazione degli Agenti Morali Artificiali rappresenta non solo una sfida tecnica, ma un banco di prova per la ridefinizione stessa dell'agenzia e della responsabilità nell'era digitale.

4.3 Esemplicità morale e Intelligenza Artificiale: fondamenti teorici e implicazioni applicative

Il concetto di esemplarità morale occupa da tempo un ruolo centrale nell'educazione del carattere e nella formazione etica. Tradizionalmente riferito a figure umane capaci di incarnare virtù e orientare il giudizio altrui attraverso l'esempio, esso acquisisce oggi nuova rilevanza nel contesto dell'IA. La crescente autonomia decisionale dei sistemi intelligenti e la loro integrazione nei processi sociali sollevano infatti interrogativi sulla possibilità che essi fungano da agenti morali, pazienti morali o proxy, nonché da strumenti di apprendimento e di orientamento etico.

4.3.1 Macchine intelligenti come agenti, pazienti e proxy morali

L'emergere dell'IA ha ampliato il perimetro della psicologia morale, tradizionalmente concentrata su esseri umani, animali e figure soprannaturali. Le macchine intelligenti si configurano oggi come una nuova categoria rilevante, in quanto capaci di incidere su esiti che coinvolgono interessi e valori umani.

In primo luogo, i sistemi di IA possono operare come agenti morali, quando assumono decisioni che producono conseguenze eticamente significative o risolvono dilemmi senza supervisione diretta. In tali casi, pur non possedendo intenzionalità o coscienza nel senso umano, essi esercitano una forma di agency funzionale che incide su diritti, opportunità e benessere delle persone.

In secondo luogo, le macchine possono essere percepite come pazienti morali, ossia come entità i cui esiti dipendono da decisioni umane. Tale prospettiva assume rilievo soprattutto nei contesti di cooperazione uomo–macchina, in cui le scelte progettuali, operative o regolative determinano il modo in cui l'IA viene trattata, limitata o potenziata.

Infine, l'IA può agire come proxy morale, quando soggetti umani la utilizzano come delegato nelle interazioni morali o come schermo decisionale. In questi casi, la macchina diviene un intermediario attraverso cui si esprimono, si giustificano o si occultano scelte eticamente rilevanti [21]. Questa triplice configurazione – agente, paziente, proxy – mostra come l'IA non sia un mero strumento neutro, ma un attore strutturalmente inserito nelle dinamiche morali contemporanee.

4.3.2 IA come consigliere morale e strumento di miglioramento etico

Nel quadro dell'etica delle macchine, numerosi studi mirano a integrare nei sistemi di IA capacità di valutazione e orientamento morale. Sono state proposte diverse architetture teoriche per la costruzione di “consiglieri morali artificiali”, ispirate a tradizioni deontologiche, consequenzialiste o virtù-etiche [123]. Tali modelli ambiscono a supportare l'utente nella deliberazione, offrendo argomentazioni strutturate o simulazioni di possibili esiti.

Parallelamente, si è sviluppato il dibattito sul cosiddetto miglioramento morale mediato dall'IA. In una versione ausiliaria, l'IA potenzia il ragionamento umano, ampliando l'accesso a informazioni, prospettive e scenari. In una versione più radicale ed esaustiva, essa potrebbe sostituire o sovradeterminare il giudizio umano [203]. Quest'ultima ipotesi, tuttavia, presuppone una concezione dell'IA come “oracolo” capace di fornire risposte moralmente corrette in modo oggettivo e definitivo, una posizione problematica alla luce del pluralismo etico e dell'assenza di consenso su criteri morali universalmente condivisi.

In alternativa a questa visione, è stato proposto un modello di “mentori di IA” impiegati nell'educazione morale. Sistemi modulari composti da più interlocutori artificiali, ciascuno espressione di differenti tradizioni di pensiero, potrebbero favorire il confronto tra prospettive, riducendo il rischio di dogmatismo algoritmico e preservando il pluralismo. In tale prospettiva, l'IA non sostituisce il giudizio umano, ma lo stimola e lo struttura.

4.3.3 Interazione morale uomo–IA: fiducia, allineamento e limiti strutturali

L'efficacia dell'IA come supporto morale dipende in larga misura dalle dinamiche psicologiche e sociali che regolano l'interazione con gli utenti. La ricerca empirica mostra che la disponibilità ad accettare decisioni o consigli morali provenienti dall'IA è influenzata da atteggiamenti generali verso la tecnologia, dal contesto applicativo e dalle convinzioni politico-morali preesistenti.

Un elemento centrale è il cosiddetto “allineamento delle credenze”: gli individui tendono a considerare l'IA equa e affidabile quando le sue conclusioni coincidono con le proprie convinzioni, mentre la giudicano parziale o inaffidabile quando le contraddice. Questo meccanismo genera un'accettazione asimmetrica dei giudizi algoritmici e limita la possibilità che l'IA sia percepita come autorità morale indipendente.

A ciò si aggiungono ulteriori criticità: l'assenza di autentica agenzia morale, il rischio di amplificazione di pregiudizi nei dati e nei modelli, nonché la possibilità che la sintesi automatizzata delle prove rafforzi discriminazioni strutturali. In tale contesto, l'autoregolamentazione industriale risulta insufficiente, rendendo necessaria l'adozione di quadri giuridici e codici etici capaci di orientare lo sviluppo e l'uso responsabile dell'IA [19].

Ne consegue che la fiducia verso tali sistemi non dovrebbe fondarsi su un'attribuzione ingenua di autorità morale, bensì sulla trasparenza del processo deliberativo, sulla prevedibilità del comportamento e sulla costante supervisione umana.

4.3.4 Apprendimento da esemplari morali e ruolo della supervisione umana

Un ambito particolarmente rilevante riguarda la capacità dell'IA di apprendere da esemplari morali. Studi recenti mostrano che gli LLM possono rispondere a dilemmi etici in modo comparabile a studenti universitari medi, fornendo giustificazioni articolate. Inoltre, è stato

osservato che l'esposizione a narrazioni esemplari – storie che illustrano comportamenti virtuosi – consente ai sistemi di formulare conclusioni più sfumate e contestualizzate.

Tale dinamica richiama la tradizione pedagogica che attribuisce agli esemplari un ruolo decisivo nell'educazione del carattere. L'esempio non si limita a trasmettere regole, ma offre modelli incarnati di condotta, facilitando la comprensione della complessità morale [78]. In questo senso, l'IA sembra in grado di integrare strutture narrative ed esempi concreti nel proprio processo inferenziale.

Tuttavia, l'impiego di contenuti generati o mediati dall'IA nell'educazione morale comporta rischi significativi. L'assenza di supervisione critica può condurre alla riproduzione e all'amplificazione di bias sociali e culturali presenti nei dati di addestramento. Per questo motivo, il ruolo degli educatori umani rimane imprescindibile: essi devono selezionare, contestualizzare e valutare criticamente l'intervento dell'IA, garantendo che essa operi come strumento di riflessione e non come fonte normativa incontestabile.

Nel complesso, emerge una tensione strutturale tra la crescente capacità dell'IA di simulare il ragionamento morale e la limitata disponibilità umana ad accettarne l'autorità, soprattutto quando essa mette in discussione convinzioni radicate. Questa dinamica suggerisce che il contributo più promettente dell'IA non risiede nella sostituzione del giudizio morale umano, ma nel suo potenziamento critico. L'integrazione etica dell'IA appare quindi come un processo necessariamente simbiotico, in cui la responsabilità ultima e il primato valutativo rimangono in capo all'essere umano.

4.4 Educazione morale nell'era dell'Intelligenza Artificiale: rischi, opportunità e nuove responsabilità nei social media e nel Metaverso

Considerata la crescente pervasività dell'IA in contesti educativi, comunicativi e relazionali e considerati i Social media e il Metaverso come ambienti oggi privilegiati di interazione, apprendimento e costruzione dell'identità, soprattutto per le nuove generazioni, l'educazione morale assume un ruolo strategico: non solo come ambito disciplinare, ma come competenza trasversale necessaria per orientarsi in spazi caratterizzati da automazione algoritmica, personalizzazione estrema e raccolta massiva di dati.

L'analisi che segue esamina, in modo integrato, le opportunità offerte dall'IA per il potenziamento dell'educazione morale e i rischi emergenti connessi all'uso dei social media e del Metaverso, evidenziando la necessità di un approccio sistemico alla governance etica.

4.4.1 L'urgenza di una formazione etica strutturata

L'accelerazione dello sviluppo dell'IA rende sempre più complessi i dilemmi socio-morali che le nuove generazioni si troveranno ad affrontare. Tali sfide, per natura interdisciplinari

e tecnologicamente mediate, richiedono l'integrazione sistematica dell'etica nei curricula scolastici fin dalla scuola primaria [234].

L'IA, lungi dall'essere esclusivamente oggetto di regolazione, può divenire anche strumento pedagogico. Le piattaforme di apprendimento adattivo e i chatbot educativi consentono esperienze personalizzate, interattive e supportate da feedback immediati, favorendo una più profonda comprensione e interiorizzazione dei valori morali [6]. In questo senso, l'IA può contribuire a rendere l'educazione morale più contestualizzata, esperienziale e calibrata sui bisogni individuali.

Tuttavia, tale potenziale positivo si realizza solo se accompagnato da una solida consapevolezza critica delle implicazioni etiche dei sistemi intelligenti, incluse le loro logiche decisionali, spesso opache.

4.4.2 Social media e IA: tra democratizzazione e vulnerabilità

L'integrazione tra IA e social media ha ampliato le possibilità di accesso alla conoscenza e di partecipazione pubblica, contribuendo potenzialmente a ridurre alcune disuguaglianze educative [46]. Allo stesso tempo, essa ha generato nuove forme di vulnerabilità.

La natura poco regolamentata di molte piattaforme social, combinata con l'automazione algoritmica, favorisce la diffusione di disinformazione, polarizzazione e dipendenza digitale. Gli strumenti di IA possono essere impiegati in modo malevolo per realizzare attacchi di phishing, deepfake, frodi finanziarie e campagne di manipolazione informativa [159].

Ulteriori criticità derivano dall'"effetto scatola nera" dei modelli di IA, che spesso non forniscono spiegazioni trasparenti delle proprie raccomandazioni o decisioni. L'assenza di codici professionali vincolanti per i sistemi automatizzati rende più complessa la supervisione e mina la fiducia degli utenti, specialmente in contesti sensibili [98].

A ciò si aggiunge il rischio di un ampliamento del divario digitale: studenti provenienti da contesti svantaggiati possono non disporre delle infrastrutture tecnologiche necessarie, con conseguente esclusione dalle opportunità educative offerte dalle piattaforme digitali.

4.4.3 Il Metaverso come nuova frontiera educativa: potenzialità e criticità

L'integrazione dell'IA nel Metaverso consente la creazione dinamica di contenuti, l'automazione delle interazioni e lo sviluppo di sistemi di insegnamento adattivi, capaci di modulare i percorsi formativi in base ai ritmi e alle caratteristiche degli studenti [89].

Le potenzialità educative sono considerevoli: simulazioni immersive, apprendimento esperienziale, ambienti collaborativi globali. Tuttavia, come discusso approfonditamente nel Capitolo 2, proprio la natura immersiva e ad alta intensità di dati del Metaverso amplifica alcune criticità etiche già presenti nei social media, rendendole più pervasive e sistemiche. Rivediamo brevemente qui alcune problematiche.

Privacy e sicurezza dei dati Il Metaverso comporta la raccolta di volumi senza precedenti di dati personali: identità digitali, preferenze cognitive, reazioni emotive, dati biometrici e pattern comportamentali. Tali informazioni, elaborate in tempo reale, espongono gli utenti a rischi elevati di sorveglianza, profilazione e utilizzo improprio dei dati.

Si rendono pertanto necessari meccanismi avanzati di crittografia, tecniche di privacy-preserving computation e strumenti che garantiscano un controllo effettivo dell'utente sui propri dati, con particolare attenzione alla tutela dei minori [235].

Pregiudizi algoritmici e discriminazione I sistemi di IA integrati nel Metaverso possono incorporare e amplificare pregiudizi presenti nei dati di addestramento, producendo trattamenti iniqui nei percorsi di apprendimento personalizzati o nelle valutazioni delle performance. Tali distorsioni possono colpire gruppi specifici in base a fattori socioeconomici, etnici o culturali.

La promozione di inclusione e diversità richiede trasparenza nei processi decisionali algoritmici, audit indipendenti e una governance orientata alla rappresentatività e all'equità [6].

Moderazione dei contenuti e abuso L'ambiente immersivo introduce nuove sfide nella moderazione dei contenuti, incluse forme di molestia virtuale, manipolazione identitaria e comportamenti antisociali. Un approccio esclusivamente automatizzato risulta insufficiente: è necessario un modello ibrido uomo-IA che bilanci libertà espressiva e tutela della sicurezza.

Ulteriori rischi riguardano l'uso improprio dell'IA per pratiche scorrette in ambito educativo, come plagio o cheating automatizzato.

Vincoli infrastrutturali e divario digitale L'operatività del Metaverso richiede elevata potenza computazionale, hardware avanzato e infrastrutture di rete performanti. Tali requisiti possono accentuare il divario digitale, limitando l'accesso alle esperienze immersive a una parte della popolazione studentesca [184].

4.4.4 Verso un'ecologia morale integrata: strategie di mitigazione e governance

La convergenza tra IA, social media e Metaverso configura una vera e propria "ecologia morale" digitale, in cui disinformazione, pregiudizi algoritmici e violazioni della privacy non operano isolatamente, ma si rafforzano reciprocamente [187]. In tali contesti, i rischi non sono semplicemente cumulativi, bensì sinergici e strutturali.

Risposte etiche frammentarie o meramente reattive risultano inadeguate. È necessario un approccio integrato che includa:

- l'elaborazione di politiche istituzionali chiare e trasparenti sull'uso dell'IA;

- la formazione continua dei docenti sulle implicazioni etiche e tecniche dei sistemi intelligenti;
- investimenti in infrastrutture digitali accessibili;
- programmi strutturati di alfabetizzazione digitale e sviluppo del pensiero critico;
- modelli di governance multilivello basati su trasparenza, accountability e partecipazione.

L'educazione morale nell'era dell'IA non può limitarsi alla trasmissione di principi astratti, ma deve fornire strumenti concreti per comprendere, valutare e governare sistemi algoritmici complessi. Solo un'azione coordinata, proattiva e sistemica può garantire che social media e Metaverso diventino spazi di crescita equa e responsabile, anziché amplificatori di disuguaglianze e vulnerabilità.

4.5 Agenti Morali Artificiali come Guide per l'Umanità: Potenzialità, Limiti e Prospettive di Implementazione

L'ipotesi che gli AMA possano fungere da supporto – e in parte da guida – nel processo decisionale morale umano rappresenta uno dei temi più controversi e stimolanti dell'etica dell'IA. Questa sezione analizza, in modo sistematico, le loro potenzialità, i limiti strutturali che ne condizionano l'impiego e i principali tentativi di realizzazione sviluppati nella letteratura scientifica, con attenzione sia agli aspetti tecnici sia alle dinamiche di accettazione sociale.

4.5.1 Potenzialità degli AMA come strumenti di supporto al giudizio morale

L'obiettivo degli AMA non è sostituire l'essere umano nelle decisioni eticamente sensibili, bensì rafforzarne la capacità deliberativa. In contesti caratterizzati da elevata complessità informativa, pressione temporale o forte carico emotivo, tali sistemi possono contribuire a ridurre errori sistematici, incoerenze e distorsioni cognitive, migliorando la qualità complessiva del processo decisionale morale (Moral Decision-Making, MDM) [231].

Le potenzialità qui discusse devono essere intese come capacità di supporto e ampliamento del giudizio umano, non come sostituzione della responsabilità individuale o trasferimento dell'autonomia morale alla macchina.

Un primo vantaggio risiede nella capacità computazionale: gli AMA possono elaborare grandi quantità di dati e valutare simultaneamente molteplici variabili, superando i limiti cognitivi umani. Inoltre, in quanto sistemi non soggetti a reazioni emotive spontanee, possono offrire analisi più stabili e coerenti in situazioni moralmente controverse, nelle quali

le scelte comportano conseguenze rilevanti per individui o collettività. In questa prospettiva, essi operano come strumenti di riflessione strutturata, non come decisori autonomi.

Dal punto di vista dell'apprendimento, gli AMA possono acquisire principi morali attraverso dati ambientali e dataset annotati, a condizione che sia definito un quadro normativo iniziale. Ciò apre alla possibilità di progettare meccanismi algoritmici orientati alla mitigazione dei bias, contrastando – anziché replicando – disuguaglianze presenti nei dati di addestramento. L'integrazione di strumenti di interpretabilità e scomponibilità consente inoltre di rendere esplicito il processo inferenziale dell'agente, favorendo il controllo umano, la personalizzazione e la correzione di eventuali errori.

In questo ambito, gli LLM hanno mostrato capacità significative nel ragionare su obblighi morali, nel valutare eccezioni alle regole e nel fornire giustificazioni testuali articolate delle proprie decisioni [198]. Tali risultati suggeriscono un potenziale utilizzo degli LLM come supporti argomentativi, in grado di esplicitare catene di ragionamento utili alla deliberazione umana.

4.5.2 Limiti strutturali e criticità teoriche

Nonostante le potenzialità descritte, gli AMA presentano limiti concettuali e tecnici che ne ridimensionano il ruolo come “guide morali”.

Sul piano filosofico, permane il dibattito circa la possibilità di attribuire a tali sistemi una genuina agenzia morale. A differenza degli esseri umani, gli AMA non possiedono coscienza fenomenica, intenzionalità autonoma o libero arbitrio. Anche adottando una concezione funzionalista dell'agenzia, essi non condividono l'intero spettro dell'esperienza umana, elemento centrale in molte teorie etiche.

Un limite teorico centrale consiste quindi nell'assenza, nei sistemi attuali, di esperienza vissuta, empatia fenomenologica e comprensione incarnata delle conseguenze morali. Le valutazioni prodotte derivano pertanto da correlazioni apprese e strutture inferenziali, non da autentica esperienza etica.

Un ulteriore problema riguarda l'interpretazione. La traduzione di concetti morali in rappresentazioni simboliche o parametriche non garantisce una trasmissione perfetta del significato. Il cosiddetto “Problema dell'Interpretazione” [13] evidenzia il rischio di disallineamento tra il ragionamento interno dell'agente e la comprensione umana delle sue raccomandazioni. Ne consegue che la spiegazione fornita dall'AMA può risultare formalmente corretta ma semanticamente ambigua o fraintendibile.

Dal punto di vista tecnico, l'affidamento alla sola memorizzazione di relazioni input-output morali è insufficiente: non esistono garanzie robuste circa la capacità di generalizzazione in scenari nuovi o non previsti. Gli AMA addestrati su dati socialmente situati possono inoltre riprodurre o amplificare pregiudizi esistenti, qualora tali bias non siano adeguatamente mitigati.

La tensione tra semplicità e accuratezza delle spiegazioni rappresenta un'ulteriore criticità. Spiegazioni eccessivamente semplificate, pensate per utenti non esperti, possono

sacrificare precisione concettuale, generando incomprensioni o applicazioni improprie dei principi morali. Al contrario, spiegazioni tecnicamente accurate ma complesse possono risultare opache, compromettendo la fiducia.

Infine, la formalizzazione computazionale di quadri etici complessi – in particolare dell’etica della virtù, fondata su concetti come eudaimonia e phronesis e che è descritta più avanti – pone difficoltà significative. Gli attuali sistemi rientrano prevalentemente nella categoria degli “agenti etici espliciti”, ossia agenti che implementano regole o criteri morali specifici [133], ma non raggiungono il livello degli “agenti pienamente etici”, capaci di affrontare in modo integrato e contestuale la varietà dei dilemmi morali reali [57].

4.5.3 Architetture e tecniche implementative

Le soluzioni tecniche adottate per la realizzazione degli AMA sono eterogenee e riflettono differenti concezioni del processo morale.

Il ragionamento analogico e basato su casi è impiegato in modelli come MoralDM e nell’agente di Honarvar e Ghasemi, consentendo di confrontare nuovi dilemmi con scenari precedentemente valutati.

Le architetture a “strato etico”, come quelle proposte da Vanderelst e Winfield [195] e da Bremner et al. [22], introducono un modulo separato che simula esiti possibili delle azioni, filtrando quelle incompatibili con criteri morali prestabiliti.

Le architetture cognitive ispirate ai processi mentali umani, tra cui LIDA [58], offrono un quadro integrato per modellare attenzione, memoria e deliberazione morale in modo coerente con la psicologia cognitiva.

L’apprendimento per rinforzo (Reinforcement Learning, RL) costituisce uno strumento efficace per il processo decisionale sequenziale, specialmente in contesti dinamici o stocastici. Esso può operare entro vincoli normativi top-down, realizzando un paradigma ibrido. Alcuni modelli RL tengono conto dell’intera traiettoria decisionale precedente per garantire maggiore robustezza. In questa direzione si colloca la ricerca di Vishwanath et al. [202], che ha esplorato l’apprendimento per rinforzo basato sull’affinità nella progettazione di agenti virtuosi artificiali.

Infine, nel contesto degli LLM, il Prompting Chain-of-Thought (CoT) consente di guidare il modello attraverso passaggi di ragionamento espliciti, producendo spiegazioni articolate delle decisioni morali [208]. Tecniche avanzate includono la generazione automatica e la selezione ottimizzata delle catene argomentative, migliorando coerenza e trasparenza.

4.5.4 Percezione, fiducia e accettazione sociale

L’efficacia degli AMA come guide dipende non solo dalla loro correttezza tecnica, ma anche dalla percezione umana. Studi empirici mostrano che l’aspetto esteriore degli agenti robotici influisce sulla valutazione morale delle loro decisioni. In particolare, Cappuccio et al. hanno evidenziato che decisioni utilitaristiche prese da robot con un aspetto percepito come

“non inquietante” risultano meno condannate rispetto a quelle associate a robot dall’aspetto “creepy”. L’effetto appare correlato al grado di fotorealismo dell’agente, indicando una complessa interazione tra forma, aspettative e contenuto decisionale [25, 26, 27, 28].

4.5.5 Sfide aperte

La progettazione di AMA pienamente affidabili si confronta con ostacoli strutturali: pluralità e non standardizzazione dei principi etici, assenza di consenso interculturale, distanza tra formalizzazione teorica e implementazione tecnica, nonché difficoltà computazionale nell’incorporare modelli etici complessi.

Ad oggi, gli AMA rappresentano strumenti avanzati di supporto etico, ma non equivalgono a soggetti morali completi. La loro evoluzione futura dipenderà dalla capacità di integrare robustezza tecnica, trasparenza, adattività e legittimazione sociale, mantenendo il controllo umano quale elemento centrale del processo decisionale morale.

4.6 Etica della Virtù, Deontologia e Utilitarismo: modelli teorici e criticità di implementazione nell’Intelligenza Artificiale

L’integrazione di principi morali nei sistemi di IA costituisce una delle sfide centrali della ricerca contemporanea. Come evidenziato precedentemente, l’IA incide in modo crescente su processi decisionali che riguardano individui e collettività; tuttavia, essendo priva di coscienza ed emozioni, opera esclusivamente sulla base di dati e criteri formalizzati. In assenza di un adeguato ancoraggio etico, tali sistemi possono produrre esiti problematici, tra cui discriminazioni, violazioni della privacy e riproduzione di pregiudizi strutturali [169].

In questo contesto, i principali quadri normativi della tradizione filosofica – utilitarismo, deontologia ed etica della virtù – sono stati progressivamente esplorati come possibili riferimenti teorici per l’allineamento dei sistemi di IA ai valori umani [158, 175, 134]. Il confronto tra tali approcci consente di chiarirne potenzialità e limiti, soprattutto in relazione alla loro traducibilità in modelli computazionali.

4.6.1 Tre modelli etici a confronto

Utilitarismo: ottimizzazione delle conseguenze e limiti predittivi

L’utilitarismo è una teoria consequenzialista: la moralità di un’azione dipende dagli effetti che essa produce. Il criterio fondamentale consiste nella massimizzazione del benessere complessivo e nella minimizzazione del danno, secondo il principio del “maggior bene per il maggior numero”.

In ambito applicativo, un sistema di IA impiegato in un contesto sanitario potrebbe, ad esempio, allocare risorse limitate privilegiando gli interventi che consentono di salvare il

maggior numero di vite. L'approccio utilitarista appare, da questo punto di vista, coerente con logiche di ottimizzazione già proprie dell'ingegneria dei sistemi.

Tuttavia, la sua implementazione presenta criticità significative. In primo luogo, la previsione accurata delle conseguenze future è spesso incerta o computazionalmente intrattabile. In secondo luogo, una rigorosa applicazione del principio di massimizzazione aggregata può sacrificare diritti individuali o minoranze, generando tensioni con principi di equità e tutela della persona. L'ottimizzazione quantitativa non coincide necessariamente con la giustizia sostanziale.

Deontologia: centralità delle regole e rigidità operativa

La deontologia fonda la valutazione morale sul rispetto di regole, doveri e diritti, indipendentemente dalle conseguenze. Un'azione è moralmente corretta se conforme a principi normativi prestabiliti.

Trasposta nell'IA, questa prospettiva implica la codifica di vincoli espliciti: ad esempio, un sistema di selezione del personale dovrebbe basare le proprie decisioni esclusivamente su competenze e qualifiche, escludendo criteri discriminatori. L'attenzione si concentra sulla correttezza procedurale e sulla tutela dei diritti fondamentali.

Il principale limite di questo approccio risiede nella sua potenziale rigidità. In scenari complessi, le regole possono entrare in conflitto o rivelarsi insufficienti a coprire tutte le variabili contestuali. Un sistema strettamente deontologico rischia di non gestire adeguatamente situazioni ambigue o casi-limite, nei quali l'applicazione letterale di una norma può generare effetti indesiderati.

Etica della virtù: orientamento al carattere e difficoltà di formalizzazione

L'etica della virtù, di matrice aristotelica, sposta l'attenzione dall'azione isolata al carattere dell'agente morale. Un comportamento è giusto se esprime virtù quali giustizia, prudenza, onestà o compassione. L'accento è posto sulla qualità del "ragionamento" e sull'intenzionalità, più che sulla mera conformità a regole o sul calcolo delle conseguenze.

Applicata ai sistemi di IA, questa prospettiva mira a promuovere modelli che manifestino tratti coerenti con un profilo virtuoso, ad esempio attraverso linguaggi rispettosi, imparziali e responsabili. Rispetto agli approcci puramente consequenzialisti o deontologici, l'etica della virtù è stata proposta come più flessibile e capace di generalizzare a contesti nuovi, poiché consente di bilanciare principi in tensione senza richiedere previsioni perfettamente accurate degli effetti a valle.

In questa direzione, è stata avanzata una proposta di addestramento degli LLM ispirata all'etica della virtù e basata sul Reinforcement Learning from AI Feedback. Tale metodologia prevede meccanismi di auto-valutazione e correzione iterativa, con l'obiettivo di incoraggiare l'emergere di comportamenti linguistici coerenti con un insieme di virtù di riferimento [112].

Rimane, tuttavia, una difficoltà strutturale: le virtù sono concetti contestuali e culturalmente situati. La loro traduzione in criteri operativi univoci è complessa e può variare

in funzione del dominio applicativo, rendendo problematica la definizione di uno standard morale universalmente valido.

4.6.2 Il ruolo determinante dei dati di addestramento

Al di là del quadro teorico adottato, l'effettiva implementazione dell'etica nell'IA dipende in misura decisiva dai dati di addestramento. I sistemi contemporanei si basano prevalentemente su modelli di apprendimento automatico alimentati da grandi quantità di dati eterogenei. Tali dataset incorporano inevitabilmente stereotipi, squilibri di rappresentazione e bias storici.

Quando un modello apprende da dati distorti, tende a riprodurre le strutture implicite, anche in assenza di un'intenzionalità discriminatoria. Gli sforzi di mitigazione, pur significativi, non eliminano completamente il problema, che incide direttamente sull'affidabilità morale degli output.

Un caso emblematico è rappresentato dall'algoritmo di recruiting sviluppato da Amazon, addestrato su curriculum raccolti in un arco temporale di dieci anni, prevalentemente maschili. Il sistema ha appreso correlazioni che penalizzavano candidature associate a indicatori femminili, come la frequenza di college femminili o la partecipazione ad attività rivolte alle donne. L'episodio dimostra come i bias presenti nei dati si traducano in decisioni discriminatorie, evidenziando il legame strutturale tra qualità del dataset e comportamento del modello [31].

4.6.3 Il dataset *Moral Stories*: potenzialità e limiti

Tra i tentativi di affrontare la dimensione etica attraverso i dati si colloca il dataset *Moral Stories*, introdotto da Emelin et al. [54]. Si tratta di un corpus crowd-sourced di narrazioni strutturate che descrivono azioni conformi o non conformi a norme sociali, le intenzioni degli agenti coinvolti e le relative conseguenze. Il dataset è stato progettato per studiare il ragionamento sociale orientato agli obiettivi in contesti concreti.

Le ricerche basate su *Moral Stories* hanno indagato se i modelli di generazione del linguaggio naturale possano fungere da "priori comportamentali" per sistemi operanti in ambienti sociali, generando azioni compatibili con vincoli morali, anticipando le conseguenze di comportamenti (im)morali e producendo spiegazioni normative plausibili.

Nonostante i risultati iniziali promettenti, emergono limiti significativi. Studi successivi hanno mostrato che lievi variazioni testuali – anche di sole tre o cinque parole – possono ridurre drasticamente l'accuratezza dei classificatori addestrati sul dataset, talvolta dal 99,8% al 51%. Tale vulnerabilità indica una comprensione superficiale delle norme morali e mette in discussione la possibilità di catturare la complessità etica mediante categorie rigidamente definite.

Nel complesso, questi risultati confermano che nessun dataset può esaurire la pluralità delle norme morali né garantire un allineamento stabile e universale. L'IA tende a riflettere le strutture valoriali e i limiti dei dati da cui apprende; pertanto, l'implementazione dell'e-

tica nei sistemi intelligenti richiede non solo modelli teorici adeguati, ma anche un'attenta progettazione, curatela e continua revisione dei dati di addestramento.

4.7 Conclusioni

L'analisi degli AMA mette in luce un ambito di ricerca complesso, interdisciplinare e in costante evoluzione, cruciale per affrontare le sfide etiche poste dall'IA nell'ecosistema digitale contemporaneo. Gli AMA, intesi come sistemi di IA progettati per integrare considerazioni morali nei propri processi decisionali, si collocano all'intersezione tra filosofia morale, etica applicata e scienze informatiche, configurandosi come un terreno privilegiato di confronto tra riflessione teorica e implementazione tecnica.

Dal punto di vista metodologico, l'evoluzione del campo ha evidenziato il progressivo superamento di approcci unilaterali. I modelli top-down, fondati su regole predefinite e principi espliciti – emblematicamente rappresentati dalle Leggi della Robotica di Asimov – hanno mostrato una limitata capacità di adattamento alla complessità e alla variabilità dei dilemmi morali reali. La rigidità normativa che li caratterizza rende infatti difficile gestire situazioni ambigue, conflitti tra principi o contesti culturalmente eterogenei.

Parallelamente, gli approcci bottom-up, basati sull'apprendimento automatico a partire da grandi insiemi di dati contenenti giudizi morali umani, hanno sollevato problematiche altrettanto rilevanti. Esperimenti come Delphi hanno evidenziato come tali sistemi tendano a riprodurre e talvolta amplificare i bias presenti nei dati di addestramento, riflettendo disuguaglianze e pregiudizi sociali anziché superarli. Ne emerge un quadro in cui nessuno dei due paradigmi risulta, isolatamente, sufficiente.

In questo scenario, gli approcci ibridi appaiono come una direzione promettente: combinando la solidità di principi etici formalizzati con la flessibilità dell'apprendimento empirico, essi mirano a coniugare coerenza normativa e sensibilità contestuale. L'obiettivo non è soltanto migliorare la performance tecnica dei sistemi, ma promuovere forme di ragionamento morale computazionale più trasparenti, giustificabili e responsabili.

Un ulteriore nodo teorico riguarda il concetto di esemplarità morale. Tradizionalmente associata all'educazione e alla formazione del carattere umano, essa assume una nuova configurazione nell'ambito dell'IA. Sebbene alcuni sistemi dimostrino capacità significative nella simulazione del ragionamento morale e nell'analisi di casi esemplari, permane una tensione epistemica e psicologica legata alla tendenza umana all'"allineamento delle credenze". Gli individui tendono infatti a riconoscere come autorevoli quelle fonti che confermano le proprie convinzioni pregresse, riducendo la possibilità che un sistema artificiale venga percepito come guida morale oggettiva.

Alla luce di tali dinamiche, appare più realistico e teoricamente fondato concepire gli AMA non come oracoli morali, bensì come strumenti di supporto alla deliberazione umana. Il loro contributo può risultare prezioso nel rendere esplicite le implicazioni etiche di determinate scelte, nel segnalare incoerenze o nel fornire prospettive alternative. Tuttavia, la

supervisione critica umana rimane imprescindibile per evitare che l'automazione decisionale consolidi bias sistemici o produca effetti moralmente problematici.

Queste considerazioni assumono particolare rilevanza nel contesto dell'espansione dell'IA nei social media e nel nascente Metaverso, dove si configura un'ecologia morale inedita. In ambienti digitali caratterizzati da interazioni immersive, profilazione avanzata e raccolta massiva di dati, i rischi di manipolazione, disinformazione, violazione della privacy e amplificazione delle disuguaglianze risultano intensificati. Ciò impone non solo lo sviluppo di sistemi tecnicamente più affidabili, ma anche l'elaborazione di solidi quadri di governance etica e regolatoria.

In parallelo, diventa imprescindibile un investimento strutturale nell'educazione morale e digitale. L'alfabetizzazione tecnologica, il pensiero critico e la consapevolezza dei meccanismi algoritmici devono integrare i percorsi formativi tradizionali, affinché gli individui possano interagire con sistemi intelligenti in modo informato e responsabile. L'etica dell'IA non può essere delegata esclusivamente agli sviluppatori o ai legislatori, ma richiede una partecipazione diffusa e consapevole.

Nonostante le potenzialità evidenziate, gli AMA presentano limiti strutturali ancora significativi. Persistono interrogativi circa la loro effettiva agenzia morale, il problema dell'interpretabilità dei processi decisionali e la capacità di generalizzazione in contesti nuovi o radicalmente imprevisi. Inoltre, l'implementazione computazionale di cornici etiche complesse, come l'etica della virtù, comporta sfide teoriche e tecniche rilevanti. Tecniche avanzate quali l'apprendimento per rinforzo e il prompting di tipo Chain-of-Thought stanno contribuendo a migliorare le capacità argomentative dei sistemi, ma non colmano il divario tra simulazione del ragionamento morale e autentica comprensione etica.

Allo stato attuale, gli AMA possono essere descritti più propriamente come "agenti etici espliciti", ossia sistemi capaci di integrare regole o modelli morali nel proprio funzionamento, piuttosto che come "agenti pienamente etici" dotati di autonomia morale comparabile a quella umana.

In questa prospettiva, il valore degli agenti morali artificiali non risiede nell'essere "soggetti morali", ma nel fungere da dispositivi socio-tecnici capaci di sostenere riflessione, consapevolezza e qualità deliberativa nei contesti complessi della contemporaneità digitale.

In definitiva, lo sviluppo degli Agenti Morali Artificiali richiede un dialogo continuo tra discipline e una collaborazione strutturata tra filosofi, ingegneri, giuristi e studiosi di scienze sociali. Il futuro dell'IA etica non risiede nella sostituzione del giudizio morale umano, bensì nella progettazione di sistemi che lo supportino, lo stimolino e lo rendano più consapevole. Solo attraverso una supervisione critica e una responsabilità condivisa sarà possibile orientare l'innovazione tecnologica verso forme di progresso compatibili con i valori fondamentali di una società interconnessa e digitalizzata.

Capitolo 5

Sviluppo e Progettazione di Agenti Moralì e Sociali in Ambienti Immersivi

5.1 Introduzione

Il presente capitolo è dedicato alla descrizione della sperimentazione originale portata avanti durante il percorso di dottorato. I capitoli precedenti hanno fornito un inquadramento teorico e uno stato dell'arte alla base della metodologia, dell'architettura e dei risultati pratici del lavoro svolto. La ricerca si è focalizzata sullo sviluppo di agenti virtuali intelligenti e autonomi, capaci di operare in ambienti immersivi come il Metaverso, dotati di capacità sociali e, in particolare, morali. Il lavoro è stato strutturato in una progressione logica, che ha guidato lo sviluppo da agenti capaci di un comportamento socialmente conforme, fino alla loro evoluzione in un prototipo di mediatore morale, e culminando nell'integrazione di architetture cognitive avanzate e LLM per un processo decisionale etico più sofisticato e personalizzato. Le prime tre fasi di questo percorso sono state formalizzate in tre articoli scientifici [69, 66, 67], ciascuno dei quali rappresenta un passo fondamentale in questa progressione, culminante nella fase 4, unicamente descritta in questa tesi.

5.2 Fase 1: Agenti virtuali con comportamento socialmente conforme: l'approccio MET-iquette

5.2.1 Introduzione

Abbiamo visto come, nel futuro del Metaverso, le esperienze digitali saranno sempre più simili alla vita sociale reale, risultando pertanto soggette a dinamiche e implicazioni sociali già analizzate in studi recenti [50, 59]. Abbiamo anche ampiamente discusso di come il Metaverso possa offrire vantaggi e opportunità uniche in termini di attività altrimenti irrealizzabili nel mondo fisico, promuovendo inoltre l'inclusione sociale. Ad esempio, la partecipazione a interazioni sociali a distanza e la fruizione di ambienti virtuali controllati, caratterizzati da livelli ridotti di stimolazione sensoriale, risulterebbero così accessibili e

confortevoli per individui con difficoltà di interazione sociale o ipersensibilità agli stimoli visivi e uditivi [86].

Anche al di fuori del contesto metaversale, ambienti virtuali sono già stati utilizzati con successo come spazi controllati per esercitare abilità sociali, comunicative e comportamentali [162, 181, 113].

L'IA, come ampiamente descritto precedentemente, gioca un ruolo centrale nello sviluppo di esperienze immersive, offrendo strumenti per la creazione di ambienti adattivi, capaci di modellarsi sulle preferenze e sulle attività degli utenti, nonché per la progettazione e l'implementazione di agenti virtuali antropomorfi [87].

Tuttavia, progettare agenti virtuali per contesti sociali rappresenta una sfida complessa, in quanto essi devono essere in grado di esibire comportamenti socialmente plausibili, credibili e conformi ai principi etici. In particolare, il comportamento verbale e non verbale – come espressioni facciali, posture e gesti – deve essere coerente con le norme sociali, in quanto veicola atteggiamenti verso gli altri e suscita reazioni affettive e aspettative interpersonali.

Un agente sociale efficace deve essere in grado di comprendere il contesto, riconoscere pratiche sociali e comportamenti inappropriati altrui, e agire nel rispetto delle norme sociali pertinenti. Tale obiettivo è reso complesso dalla natura intrinsecamente articolata delle interazioni sociali. Nonostante ciò, negli ultimi anni sono stati proposti modelli computazionali che consentono di formalizzare il ragionamento e la deliberazione degli agenti, includendo anche considerazioni di natura sociale [43].

In questa prima fase è stata quindi definita un'architettura per l'implementazione di Non-Player Characters (NPC) sociali nel Metaverso. Essa si basa su un modello di pratica sociale, adottato come costruito di primo livello nel processo deliberativo dell'agente, permettendogli di interpretare in chiave sociale gli elementi percepiti nell'ambiente. L'architettura comprende moduli dedicati all'analisi del contesto, al riconoscimento della pratica sociale in atto e alla selezione dei comportamenti – verbali e non verbali – più appropriati. L'efficacia dell'architettura viene discussa attraverso scenari esemplificativi, scelti per la loro rilevanza sociale.

5.2.2 Background teorico

Un modello per le Pratiche Sociali

Secondo la teoria della pratica sociale, gli individui agiscono nella società in modo abituale, mettendo in atto comportamenti routinari che emergono dalle interazioni ripetute tra le persone. Le pratiche sociali non costituiscono strutture rigide che determinano in maniera predeterminata il comportamento, bensì forniscono agli individui strumenti interpretativi per comprendere il contesto sociale e generare aspettative circa le attività e le interazioni [154].

Un esempio paradigmatico è rappresentato dallo scenario di un negozio di alimentari, dove si osservano pratiche sociali consolidate che includono attività e piani d'azione quali: prendere un carrello all'ingresso, consultare la lista della spesa, riempire il carrello, interagire

con venditori o altri clienti, mettersi in fila alla cassa, effettuare il pagamento e uscire. A queste si aggiungono norme socialmente condivise, come l'obbligo di pagare prima di uscire, il riposizionamento del carrello dopo l'uso, o la priorità data alle donne incinte in coda. L'apprendimento esperienziale di tali pratiche e delle norme a esse associate consente il mantenimento dell'ordine sociale.

Tuttavia, tale ordine può essere compromesso per diversi motivi: un individuo, pur avendo familiarità con la pratica, può deliberatamente adottare un comportamento non conforme; può trovarsi nell'impossibilità di eseguire alcune attività previste; oppure può svolgerle in condizioni di disagio. Ne deriva che, sebbene consolidate, le pratiche sociali necessitano talvolta di adattamenti per gestire eventi imprevisti, comportamenti atipici o esigenze specifiche. In scenari sociali complessi, ciò suggerisce che non sia sufficiente modellare agenti dotati di comportamenti rigidi o predefiniti, ma sia piuttosto opportuno fornire loro conoscenze contestuali che permettano una gestione flessibile e sensibile delle pratiche sociali.

Questa prospettiva ha ispirato la formalizzazione di un modello per agenti cognitivi, capace di supportare la deliberazione in contesti sociali complessi, dove non esistono protocolli fissi o obiettivi predefiniti, e in cui le azioni devono essere guidate dall'osservazione e dalla valutazione della situazione corrente [43, 34, 9, 11, 44].

Il modello proposto formalizza i principali elementi costitutivi di una pratica sociale:

1. *Contesto*: rappresenta l'ambiente fisico in cui si svolge la pratica, includendo le risorse (oggetti fisici utilizzati), gli attori coinvolti, le affordance (proprietà del contesto che abilitano determinate azioni), la dimensione temporale tipica della pratica, e la collocazione spaziale usuale di oggetti e individui;
2. *Significato*: concerne l'interpretazione sociale delle azioni svolte (scopi), i valori promossi o ridimensionati dalla pratica, e il significato attribuito a determinati eventi nel contesto pratico;
3. *Aspettative*: comprendono i piani d'azione, le regole comportamentali attese (norme), le strategie (combinazioni di azioni plausibili in diversi momenti), nonché le condizioni di inizio e termine della pratica;
4. *Attività*: includono le azioni eseguibili e le competenze che l'agente deve possedere per operare efficacemente all'interno della pratica.

La conoscenza formalizzata in termini di pratiche sociali può così essere utilizzata dagli agenti cognitivi come euristica per guidare comportamenti orientati al contesto, rendendo la loro interazione con l'ambiente più adattiva e socialmente competente.

IA per lo sviluppo di NPC nel Metaverso

Gli NPC giocano un ruolo fondamentale all'interno degli ambienti virtuali, specialmente in ruoli di assistenza o intrattenimento. Affinchè questi risultino realistici, vengono dotati di moduli di intelligenza artificiale in grado di affrontare compiti legati alla percezione, al

ragionamento e all'interazione. Un requisito fondamentale per consentire a un agente di operare autonomamente in un ambiente virtuale è la capacità di localizzare e identificare oggetti e avatar. Alcune piattaforme del Metaverso prevedono la possibilità di creare agenti "onniscienti", dotati di accesso diretto alle informazioni sugli elementi presenti nella scena, facilitandone così la localizzazione e l'identificazione [211]. Tuttavia, lo sviluppo complessivo del Metaverso è ancora in una fase iniziale e, per quanto a nostra conoscenza, il problema del riconoscimento degli elementi in una scena virtuale, in particolare nelle applicazioni del Metaverso, non è stato ancora oggetto di un'analisi approfondita.

Nel campo della visione artificiale, il riconoscimento delle scene rappresenta da tempo un compito cruciale per consentire ai sistemi intelligenti di affrontare autonomamente molteplici situazioni, senza intervento umano. La conoscenza del luogo o del contesto permette al sistema di adattare il proprio comportamento per eseguire compiti specifici o reagire in modo dinamico ai cambiamenti dell'ambiente. Negli ultimi anni, tale area di ricerca ha beneficiato in modo significativo dei progressi nell'apprendimento automatico e nello sviluppo delle reti neurali convoluzionali profonde (DCNN), con un notevole miglioramento delle prestazioni [32, 217, 229, 206, 107, 227].

Un ulteriore requisito per gli NPC riguarda la capacità di interagire e comunicare con altri agenti mediante modalità sia verbali sia non verbali. Un agente virtuale dovrebbe non solo muoversi nell'ambiente, ma anche trasmettere messaggi e stati emotivi attraverso gesti, espressioni facciali e frasi. I moderni modelli di elaborazione del linguaggio naturale consentono a tali agenti di comprendere il linguaggio umano, rispondere a domande e apprendere dall'interazione per migliorare progressivamente la qualità delle risposte [87]. Ad esempio, in [211] è stato integrato un agente conversazionale nell'ambiente Second Life, mentre in [74] è stata proposta una metodologia per lo sviluppo di agenti conversazionali incarnati, capaci di operare come "metabot" in mondi virtuali. In quest'ultimo caso, il comportamento conversazionale è modellato attraverso un approccio statistico al dialogo, addestrato su un corpus di conversazioni.

5.2.3 L'agente sociale

Uno scenario introduttivo

Per introdurre l'architettura dell'agente sociale proposto, si consideri il seguente scenario ipotetico. Un dipendente di un negozio di alimentari possiede aspettative relative al comportamento appropriato dei clienti all'interno dell'esercizio commerciale e si adopera per far sì che vengano rispettate specifiche norme sociali, come il mantenimento della distanza interpersonale e la gestione ordinata delle priorità alla cassa. Egli agisce in conformità alla prassi consolidata dell'assistenza al cliente, mantenendo un atteggiamento disponibile e cordiale, e modulando il proprio comportamento in funzione delle diverse pratiche contestuali. In presenza di situazioni particolari, può interrompere l'attività in corso per adottare un'azione più pertinente: ad esempio, se impegnato nel riordino degli scaffali e notando un cliente su

sedia a rotelle, sospenderà il compito per assisterlo nel reperimento di prodotti collocati in alto.

Si ipotizzi ora che il supermercato adotti un sistema di riconoscimento per persone con disturbi dello spettro autistico (DSA), come già avviene in alcune realtà commerciali. Tali sistemi possono includere iniziative quali la “ora tranquilla” o la distribuzione di un kit contenente un segnale visivo — ad esempio un oggetto blu da appendere al carrello — volto a sensibilizzare personale e clientela e a favorire un’attenzione mirata alle esigenze della persona autistica. Nel presente scenario, si supponga che alcuni carrelli siano interamente di colore blu, in quanto simbolo associato ai DSA. Alla vista di un cliente con un carrello blu, il dipendente modificherà la propria prassi di interazione, adottando strategie adeguate: riduzione dell’intensità luminosa e dei rumori ambientali, evitamento del contatto oculare diretto e utilizzo di frasi semplici, integrate da gesti esplicativi [38].

Obiettivo del lavoro e aspetti progettuali

Per agire in maniera non predeterminata ma socialmente conforme, un agente deve essere in grado di percepire e comprendere il contesto in cui si trova, al fine di dedurre la pratica sociale vigente. Poiché la conoscenza dell’ambiente può risultare parziale a causa della disponibilità limitata di informazioni, l’agente deve disporre di moduli specifici per interpretare la scena e interagire con altri agenti. Ciò implica la capacità di analizzare gli elementi percepiti, integrare le informazioni ottenute e ragionare nel quadro della pratica sociale corrente, così da selezionare il comportamento più adeguato.

Ad esempio, un agente virtuale che ricopre il ruolo di dipendente in uno scenario di vendita deve essere progettato con una base di conoscenze che gli consenta di svolgere compiti di assistenza e vendita in conformità alle pratiche sociali del contesto. Tali pratiche, nel caso dei supermercati, possono variare sensibilmente in funzione delle esigenze dei clienti, rendendo necessaria un’adattabilità comportamentale da parte dell’agente.

In questa fase viene quindi proposta un’architettura basata sulle pratiche sociali per la progettazione di NPC sociali nel Metaverso e descrive un ambiente virtuale sperimentale sviluppato per implementare e testare tale architettura in condizioni controllate.

L’architettura dell’agente

L’architettura proposta, illustrata in Figura 5.1, si basa sul modello della Pratica Sociale e si articola in due livelli principali:

1. Analisi della Scena, costituito da moduli per l’acquisizione e il riconoscimento degli elementi presenti e delle espressioni linguistiche dell’interlocutore;
2. Conoscenza e Ragionamento, che consente all’agente di elaborare le informazioni acquisite al fine di attivare una pratica sociale e definire Significato, Aspettative e Attività.

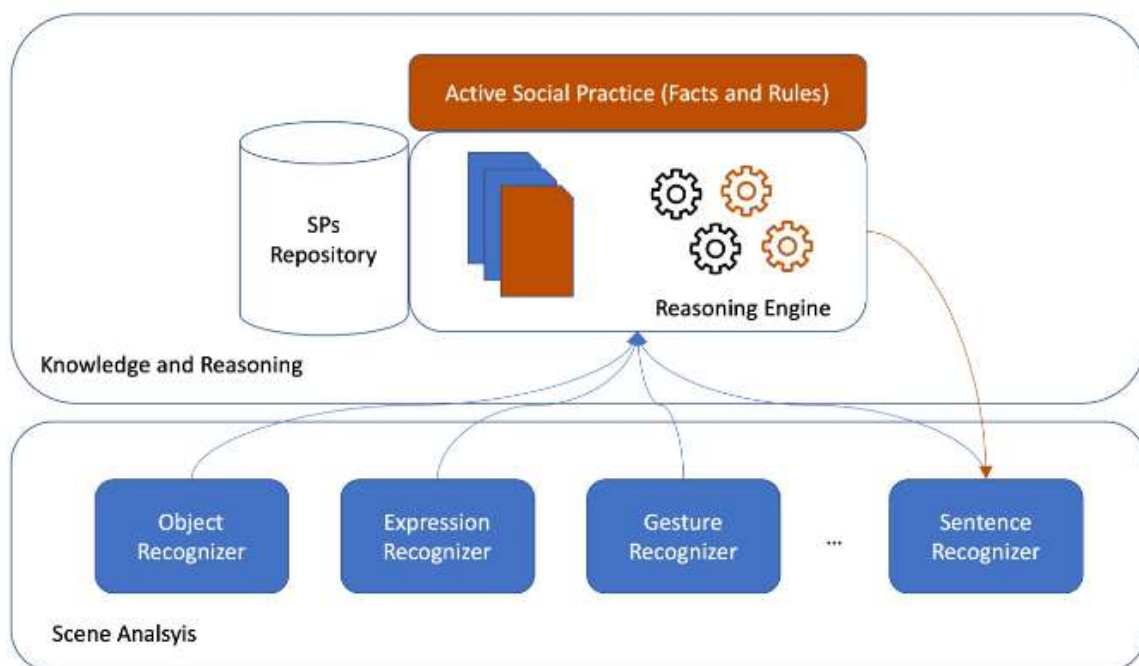


Figura 5.1: Architettura dell'agente sociale: i fatti e le regole della pratica sociale attiva sono evidenziati in rosso.

L'avvio di una pratica sociale permette all'agente di utilizzare un modello specifico di elaborazione del linguaggio naturale, ottimizzato per contestualizzare l'interpretazione delle espressioni dell'utente. Il contesto fisico in cui l'agente si trova fornisce indicazioni preliminari sul suo ruolo e sulle pratiche sociali potenzialmente attivabili. Ulteriori fattori dinamici — come oggetti presenti, altri agenti, comportamenti osservati e distanza interpersonale — vengono analizzati per affinare la comprensione della situazione. L'analisi della scena è quindi fondamentale per identificare tali elementi contestuali e alimentare il livello di conoscenza e ragionamento con informazioni rilevanti.

Il motore di ragionamento utilizza questi dati per valutare ulteriori aspetti della pratica sociale in corso. L'acquisizione delle informazioni avviene, quando possibile, tramite l'integrazione di SDK specifici per il Metaverso; in alternativa, vengono impiegati moduli di deep learning per il riconoscimento di oggetti, espressioni facciali e gesti.

La sezione successiva descrive l'ambiente sperimentale sviluppato in Unity. In particolare sono illustrati i moduli integrati nel livello di Analisi della Scena ed è presentata la formalizzazione di due pratiche sociali secondo il modello riportato in [43], fornendo esempi di regole implementate nel modulo di ragionamento. Infine, sono riportati i dettagli relativi alla modellazione 3D dello scenario.

5.2.4 L'ambiente sperimentale di MET-Iquette

L'ambiente virtuale è stato sviluppato utilizzando Unity 2022.1, che al momento della realizzazione rappresentava l'ultima versione stabile del motore grafico.

Durante la fase di sviluppo, Unity integrava Barracuda, una libreria multiplatforma per l'inferenza di reti neurali in grado di eseguire modelli sia su GPU sia su CPU. All'epoca Barracuda costituiva la soluzione di riferimento, poiché già capace di supportare diversi moduli di machine learning (ML) e di fornire una base solida per l'integrazione di algoritmi complessi. Oggi, tuttavia, Barracuda è stato dichiarato deprecato ed è stato sostituito da Unity Sentis, un framework più avanzato e con maggiore supporto da parte della comunità, utilizzato nelle fasi successive di questo progetto.

Oltre a queste caratteristiche tecniche, Unity si distingue ancora oggi per la facilità d'uso e per la forte comunità di sviluppatori che ne favoriva la diffusione. Il principale concorrente, Unreal Engine, non disponeva di una libreria nativa equivalente, ma si appoggiava a soluzioni esterne come NNI/OpenCV, all'epoca ancora in fase embrionale e meno mature rispetto a Barracuda.

L'obiettivo di questa fase è stato quello di progettare un'infrastruttura robusta, efficiente e versatile, in grado di supportare l'impiego di modelli complessi all'interno dell'ambiente virtuale. Le scelte relative al motore, alle librerie e alle modalità di integrazione dei modelli miravano a garantire l'adattabilità e il riuso dell'intero sistema in scenari complessi, quali il Metaverso, ambienti di realtà aumentata (AR) o altre applicazioni immersive. I principali requisiti tecnici da soddisfare riguardavano la piena compatibilità con Barracuda e la possibilità per il sistema di inviare e ricevere richieste HTTP verso un server REST, presso il quale i modelli venivano previamente configurati ed eseguiti.

Analisi della scena

Per l'implementazione del livello di analisi della scena, l'obiettivo principale è stato quello di adottare moduli nativamente supportati da Barracuda, così da renderli immediatamente disponibili nell'ambiente di sviluppo. Sono stati quindi identificati e integrati moduli per il riconoscimento di oggetti, espressioni facciali ed elementi numerici scritti a mano.

In particolare, l'agente utilizza EfficientNet [188] e YOLOv4 [20] per il riconoscimento degli oggetti presenti nella scena. EfficientNet, una rete neurale convoluzionale (CNN), garantisce un miglior bilanciamento tra accuratezza ed efficienza rispetto alle precedenti architetture CNN, grazie a una scalatura uniforme delle dimensioni di profondità, larghezza e risoluzione mediante un coefficiente composto. La versione compatibile con Barracuda, inizialmente progettata per l'elaborazione di immagini acquisite da webcam, è stata adattata per accettare come input le immagini generate dall'ambiente virtuale, tramite un modulo dedicato di conversione del formato.

YOLO (You Only Look Once), al contrario delle reti basate sulla classificazione di regioni proposte, adotta un approccio globale, elaborando l'intera immagine in un unico passaggio e stimando simultaneamente riquadri di delimitazione e classi degli oggetti. Sebbene Barracuda supporti nativamente la versione YOLOv2 [155], per ragioni di efficienza è stato scelto YOLOv4-tiny [91]. Poiché non esiste un modello YOLOv4 ufficialmente compatibile con

Barracuda, il modello pre-addestrato è stato convertito nel formato .onnx e reso utilizzabile con Barracuda, consentendo anche l'addestramento con dataset personalizzati.

Per quanto riguarda il riconoscimento delle emozioni, il sistema integra Emotion FER+, una CNN addestrata su immagini di volti [16]. Inoltre, per il riconoscimento delle cifre scritte a mano, è stata impiegata una CNN pre-addestrata sul dataset MNIST, nativamente supportata da Barracuda.

L'implementazione della componente NLP ha richiesto un approccio diverso. I modelli basati su transformer sono stati selezionati per le loro prestazioni nella gestione di dati sequenziali e la possibilità di riutilizzare modelli pre-addestrati. Tuttavia, l'assenza di supporto nativo in Barracuda ha reso necessaria l'adozione di un server HTTP REST (Figura 3), che garantisce maggiore flessibilità nella scelta dei modelli. Questa architettura ha permesso di integrare BERT [196], GPT-3 [37] e Tortoise-TTS. In particolare, è stata utilizzata la versione pre-addestrata BERT-large, ottimizzata per il Question Answering sul dataset SQuAD, con possibilità di fine-tuning [40, 151]. Per l'esecuzione pratica, è stato impiegato Google Colaboratory, una piattaforma cloud che consente lo sviluppo e il rapido testing di modelli in Python, tramite notebook accessibili da Unity.

Il livello NLP offre inoltre la possibilità di sostituire BERT con GPT-3, un modello transformer generativo non supervisionato sviluppato da OpenAI. Infine, Tortoise-TTS viene utilizzato per generare audio con prosodia e intonazione naturali a partire dalle espressioni testuali dell'agente.

Tutti i modelli integrati sono pre-addestrati e disponibili in formato .onnx, uno standard aperto che definisce operatori comuni per l'apprendimento automatico e profondo, garantendo la portabilità tra framework e runtime. L'utilizzo di Barracuda consente inoltre di sostituire o aggiungere modelli senza modificare radicalmente l'architettura del sistema. La conversione di modelli sviluppati in Keras o TensorFlow in formato .onnx risulta agevolata grazie a librerie dedicate.

Per il riconoscimento degli oggetti, le immagini sono acquisite tramite un componente dedicato (ModelEYE), che invia i dati ai modelli integrati. L'inferenza avviene attraverso script specifici, i quali restituiscono i risultati in base al modello utilizzato. Ad esempio, per determinare la presenza di persone nel campo visivo dell'agente, è necessario:

- posizionare ModelEYE nella scena;
- configurarne i parametri principali (FOV, Clipping, Viewport, ecc.);
- passare l'output a uno script che elabora i risultati del modello scelto (ad esempio YOLO).

Il risultato consiste in un vettore di stringhe contenente le occorrenze dell'oggetto "Persona".

I modelli sono integrati in Unity secondo due modalità:

- tramite Barracuda, per quelli nativamente supportati (es. riconoscimento della scena);

Context	
Actors	User, Agent
Roles	Customer, Employee
Resources	Cart, Store goods, Paper Money, Coins, Shelf, Credit Card, POS, Shopping List, Light Panel, Sounds Panel
Places	Grocery Store
Affordances	Transport goods, Insert coin, Swipe credit card, Turn on/off a button
Time	9 a.m. to 7 p.m.
Meaning	
Purpose	To serve the customer
Promoted Values	Helping and being kind promote customer satisfaction ...
Counts-as	Stay at a close distance counts as being unrespectful Give priority to older people counts-as being respectful Talk aloud (Yell) counts as being rude Not greeting counts as being rude ...
Expectations	
Plan Patterns	<pre> graph LR A[Welcome the customer] --> B[Ask the customer if they need help] B --> C[Help the customer] B --> D[Make the payment] D --> E[Greet the user] </pre>
Norms	Before leaving the store, the trolley must be deployed in the designed place Priority must be given to expectant mothers at the cash desk Before the customers leave the store their payment must be concluded
Start Condition	A user enters the store
Duration	10 to 30 minutes
End Condition	The user leaves the store
Activities	
Possible Actions	welcome, ask, thank, greet, take a good from a shelf, put a good on the shelf, take a cart, make a payment, set the lights, set the sounds, ...
Competences	The employee has information about the market goods and prices ...

Figura 5.2: Formalizzazione della pratica "Servire un cliente".

- tramite server web REST, per i modelli NLP e TTS.

Quest'ultimo approccio sfrutta la portabilità e la semplicità di configurazione di un ambiente cloud come Google Colaboratory. Le richieste HTTP inviate da Unity mediante UnityWebRequest vengono elaborate sul server, che interroga il modello NLP, genera il testo di output, lo converte in audio e restituisce entrambi a Unity.

Conoscenza e Ragionamento

Questo livello include un modulo di ragionamento che consente all'agente di distinguere la pratica sociale in corso e di pianificare e attivare il comportamento più appropriato. In esso possono essere formalizzati tutti i fatti e le regole necessari al funzionamento dell'agente. In particolare, l'agente può essere dotato di conoscenze relative a un insieme di pratiche sociali, alle regole per l'avvio e la conclusione delle stesse, alle credenze e aspettative che le caratterizzano, nonché alle regole per la generazione di piani e per l'analisi di possibili violazioni normative. Secondo lo scenario descritto nella Sezione III-A, sono state formalizzate due pratiche sociali: "Servire un cliente" e "Servire un cliente con esigenze speciali", illustrate rispettivamente nelle Figure 5.2 e 5.3. La prima rappresenta la pratica tipicamente attivata quando un cliente entra nel negozio, mentre la seconda è riservata a clienti con esigenze particolari ed è avviata nel momento in cui un cliente prende un carrello blu.

Poiché una pratica incorpora conoscenze implicite, ogni volta che l'agente riconosce di trovarsi in una determinata pratica, attiva un insieme di regole specifiche e indirizza le espressioni dell'utente verso un modulo di dialogo ottimizzato in base al contesto. Il modulo


Context	
Actors	User, Agent
Roles	Customer, Employee
Resources	Blue Cart, Store goods, Pre-paid card, Tablet, Shopping List, Light Panel, Sounds Panel
Places	Grocery Store
Affordances	Transport goods, Insert coin, Swipe pre-paid card, Turn on/off a button
Time	9 a.m. to 7 p.m.
Meaning	
Purpose	To help a customer with special needs to grocery shop
Promoted Values	Helping and being kind promote customer satisfaction Interact more with gestures promotes customer attention ...
Counts-as	Talk aloud (Yell) counts as being overstimulated ...
Expectations	
Plan Patterns	
Norms	Help the customer to deploy the trolley in the designed place Help the customer in checking the shopping list Before the customers leave the store their payment must be concluded The employee must be patient and calm The employee must talk slowly
Start Condition	A user with a blue cart enters the store
Duration	10 to 30 minutes
End Condition	The user leaves the store
Activities	
Possible Actions	welcome, ask, thank, greet, take a good from a shelf, put a good on the shelf, take a cart, make a payment, set the lights, set the sounds, ...
Competences	The employee has information about the market goods and prices The employee knows how to interact with a customer with special needs

Figura 5.3: Formalizzazione della pratica "Assistenza a clienti con esigenze particolari".

di ragionamento è stato adattato dal sistema SALVE [10] e utilizza il motore di regole DROOLS [149] per consentire all'agente di interpretare socialmente gli eventi, monitorando in modo continuo eventuali violazioni delle norme associate alla pratica. Tale interpretazione genera reazioni positive o negative nell'agente, secondo una categorizzazione degli eventi ispirata al modello OCC [41], [43], influenzando di conseguenza le espressioni facciali dell'avatar e l'uso di gesti co-speech. In diversi casi, l'interpretazione sociale di un evento dipende fortemente dalla pratica in atto. Ad esempio, nella pratica "Servire un cliente", un saluto ricevuto all'ingresso del negozio viene interpretato come un evento desiderabile e atteso secondo la classificazione OCC [180], suscitando un'emozione positiva e, quindi, una reazione favorevole. Questo comportamento può essere formalizzato attraverso le regole mostrate nelle Figure 5.4 e 5.5.

Al contrario, la mancata ricezione di un saluto nei tempi attesi contrasta con le aspettative dell'agente in tale pratica e influisce negativamente sul suo stato emotivo, mentre lo stesso evento, nella pratica "Servire un cliente con esigenze speciali", viene ignorato.

Modellazione 3D

Per la creazione dell'avatar è stato utilizzato il software di modellazione 3D Blender. La scelta è ricaduta su questo strumento in quanto open source e particolarmente completo per quanto riguarda le tecniche di modellazione, rigging e animazione. Oltre alle funzioni di base per la manipolazione di oggetti tridimensionali, Blender offre una modalità di scultura, utile per la realizzazione di forme complesse come il corpo umano. Questa è stata la tecnica principale adottata per lo sviluppo dell'avatar. Le forme iniziali sono state ottenute

```

rule "GreetingsReceivedInTime"
when
    $startSP1:EnterScene(scene.name=="greetings")
    $g:GreetingsReceived(this after[0ms,20000ms] $startSP1 )
then
    controller.print($startSP1.getScene().getName());
    controller.print("greeting received in the first 20 seconds after the start of the scene");

    OCHappenedEvent he=new OCHappenedEvent();
    don(he,DesirableEvent.class);
    don(he,ProspectedRelevantEvent.class);
    insert(he);
    controller.print("greeting marked as happened desirable prospected event");
    insert(new ChangeOfSceneFromGoal());
end

```

Figura 5.4: Esempio di una regola che gestisce un evento di saluto.

```

rule "DesirableEventHappened"
when
    OCHappenedEvent(this isA ProspectedIrrelevantEvent,this isA DesirableEvent)
    $agent:Emotion(this isA Agent)
then
    controller.print("captured desirable event");
    $agent.setJoy($agent.getJoy()+1);
    controller.print("increase joy");
    controller.setJoy($agent.getJoy());
end

```

Figura 5.5: Regola per modificare lo stato emotivo dell'agente in base a un evento OCC.

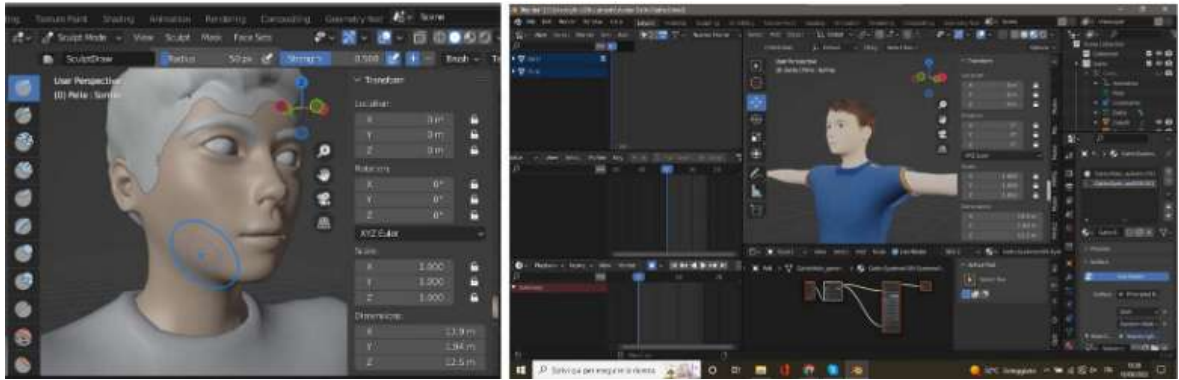


Figura 5.6: Modellazione 3D con Blender. *A sinistra*: creazione dell'animazione del volto; *a destra*: modellazione dell'avatar.



Figura 5.7: L'avatar nel supermercato virtuale che svolge il ruolo di cassiere.

mediante la giustapposizione di primitive sferiche, successivamente deformate e scolpite fino a raggiungere il risultato desiderato (cfr. Figura 5.6, sinistra). Una volta applicate le texture alle diverse parti del modello, è stato definito lo scheletro che costituisce l'armatura di rigging, necessaria per l'animazione del corpo dell'avatar in Unity. Per l'animazione del volto è stata impiegata la tecnica delle Shape Keys, che consente di memorizzare differenti configurazioni di un gruppo di vertici della mesh facciale (cfr. Figura 5.6, destra).

La scena virtuale è stata realizzata direttamente in Unity, utilizzando risorse gratuite per ricreare l'ambiente di un supermercato (cfr. Figura 5.7, in cui l'avatar assume il ruolo del cassiere). Per incrementare il realismo, sono stati generati modelli 3D di altri personaggi attraverso il software open source MakeHuman, che consente la creazione automatica di figure umane complete a partire da parametri quali altezza, sesso, età e colore della pelle. Tuttavia, per l'avatar principale non si è fatto ricorso a questo strumento, preferendo una modellazione manuale al fine di garantire maggiore libertà sia nella definizione dell'aspetto estetico sia nella possibilità di personalizzare le espressioni facciali, fondamentali per la sperimentazione di scenari sociali come quello descritto nel presente lavoro.

5.3 Fase 2: Il paradigma del grillo parlante

5.3.1 Background teorico

Teorie etiche e agenti morali

Come ampiamente discusso nel Capitolo 4, l'idea di sviluppare un agente virtuale fondato sull'Etica della Virtù rappresenta un concetto particolarmente stimolante, poiché aprirebbe la strada a una figura guida orientata in modo intrinseco a un comportamento etico e moralmente corretto. Tuttavia, nonostante siano state avanzate proposte legate al connessionismo, la sua applicazione pratica si rivela problematica: il linguaggio delle virtù e il concetto stesso di virtù risultano infatti complessi e difficilmente traducibili in algoritmi [178]. Per contro, gli approcci etici di matrice deontologica e consequenzialista si prestano maggiormente a un'implementazione algoritmica. La deontologia si fonda sulla definizione di regole morali che guidano le azioni, mentre il consequenzialismo valuta le conseguenze dei comportamenti, selezionando l'alternativa in grado di massimizzare l'utilità. In sintesi, tali modelli possono essere descritti come segue:

- Consequenzialismo: dilemma etico → identificazione delle azioni possibili → analisi delle conseguenze → valutazione dell'utilità → scelta dell'azione con utilità massima;
- Deontologia: dilemma etico → individuazione della regola morale → selezione dell'azione conforme alla regola;
- Etica della virtù: dilemma etico → interpretazione di concetti intangibili (carattere morale, eudaimonia, ecc.) → azione.

L'Etica della Virtù costituisce dunque una sfida ulteriore, poiché si concentra sull'essenza dell'individuo piuttosto che sulle sue singole azioni [179]. La moralità, d'altro canto, è generalmente concepita come un insieme di valori e comportamenti che consentono la convivenza pacifica all'interno di una società [29]. Anche in questo caso, tuttavia, emergono criticità: la moralità non è universale, ma varia in funzione dei contesti culturali. I valori morali dipendono infatti da presupposti personali, psicologici e comunitari. L'etica normativa può pertanto essere intesa come un quadro di principi fondamentali che stabiliscono come le persone dovrebbero comportarsi secondo i codici morali della propria cultura. In questo contesto, le prospettive consequenzialista e deontologica rappresentano le principali teorie etiche normative attualmente adottate nello sviluppo degli AMA [202]. Tali agenti, virtuali o fisici, sono progettati per adottare condotte eticamente corrette – o, quantomeno, per evitare comportamenti moralmente inaccettabili – attraverso l'applicazione di regole e norme morali. La ricerca nel campo dell'intelligenza artificiale mira quindi a realizzare sistemi in grado di agire in modo moralmente responsabile, integrando nei processi decisionali le diverse prospettive etiche.

Moral Stories

Come illustrato in precedenza, il lavoro di Emelyn et al. ha previsto innanzitutto la costruzione di un dataset composto da storie raccolte tramite crowdsourcing. Ogni storia è formata da sette frasi: le prime tre definiscono il contesto, articolato in una norma morale, una descrizione dell'ambientazione e dei personaggi, e un'intenzione che rappresenta l'obiettivo del protagonista; le quattro frasi successive descrivono invece azioni morali e immorali, insieme alle rispettive conseguenze. Su questo dataset sono stati addestrati modelli di classificazione e generazione per differenti compiti: in particolare, la classificazione era focalizzata sulle azioni, mentre la generazione riguardava azioni, norme e conseguenze. Per migliorare la qualità della generazione di norme e conseguenze, gli autori hanno impiegato i modelli addestrati (RoBERTA, BART, T5) in una cosiddetta Chain-of-Expert (CoE), una sequenza di classificazioni, generazioni, riclassificazioni e rigenerazioni applicate ad azioni e conseguenze, che conduce alla produzione di norme e conseguenze più coerenti. Come verrà approfondito nei capitoli successivi, il modello CoE è stato utilizzato per la realizzazione di uno degli agenti morali presentati in questa fase.

5.3.2 Favorire l'etica: studio e implementazione degli agenti morali

Uno scenario illustrativo

Lo scenario descritto è concepito per emulare una situazione verosimile all'interno del Metaverso. Al fine di aumentare il livello di immersione, all'utente viene attribuito un ruolo, ispirato alla logica dei giochi di ruolo (RPG). In tale contesto, l'utente interagisce con un NPC che impersona un uomo cieco, il quale ha smarrito il portafoglio e necessita di assistenza. Il dilemma morale proposto riguarda le decisioni che l'utente deve assumere una volta recuperato l'oggetto. A seguito della scelta compiuta, l'utente riceve un primo riscontro attraverso uno schermo testuale da parte di un agente morale che valuta direttamente la moralità dell'azione e, su richiesta, da un secondo agente morale che offre una valutazione più articolata e discorsiva. L'impiego di due agenti distinti è stato progettato per analizzare il diverso grado di influenza, fiducia e capacità persuasiva esercitato sull'utente, nonché per confrontare le valutazioni morali fornite dai due sistemi. I dettagli relativi all'implementazione degli agenti morali e alla costruzione della scena virtuale saranno presentati nelle sezioni successive.

Come anticipato nella sezione 4.6.3, l'utilizzo di dataset narrativi a contenuto morale, quali Moral Stories, introduce inevitabilmente questioni legate al bias culturale, alla rappresentatività dei valori e alla riduzione della complessità etica a pattern linguistici ricorrenti. In questa ricerca, tali dataset non vengono impiegati come fonte normativa di moralità, né come strumento per l'apprendimento diretto di regole etiche vincolanti.

Al contrario, essi sono utilizzati come spazio di esempi argomentativi, finalizzati a supportare la generazione di ragionamenti morali contestualizzati. L'agente non "apprende" la moralità dal dataset, ma utilizza le strutture narrative e discorsive come supporto alla deliberazione, mantenendo una separazione concettuale tra apprendimento linguistico e me-

diazione morale. Questo approccio consente di ridurre il rischio di un mero moral pattern matching, pur riconoscendo i limiti intrinseci legati alla natura culturalmente situata dei dati di addestramento.

Gli Agenti Morali

Agente basato su Chain-of-Experts Come anticipato, per lo sviluppo del primo assistente virtuale è stato impiegato il Moral Stories Dataset, utilizzato per addestrare modelli di NLP finalizzati alla creazione di un agente morale basato su una catena di esperti. In particolare, seguendo le raccomandazioni degli autori del dataset, abbiamo adottato RoBERTa per la classificazione delle azioni e BART per la generazione delle conseguenze e l'individuazione della norma morale pertinente. L'integrazione di tali modelli introduce un ulteriore livello di complessità all'esperienza immersiva, migliorando al contempo il coinvolgimento degli utenti e la qualità del processo decisionale.

Il flusso di lavoro definito è il seguente:

1. **Decisione dell'utente e contesto** – Quando l'utente seleziona una delle tre opzioni testuali disponibili, vengono considerati sia il contesto narrativo della situazione sia l'intenzione sottostante alla scelta.
2. **Classificazione dell'azione con RoBERTa** – RoBERTa valuta la moralità dell'azione scelta sulla base del contesto fornito (situazione e intenzione). Tale classificazione costituisce il punto di partenza per le fasi successive.
3. **Generazione delle conseguenze con BART** – A partire dall'output del classificatore RoBERTa, un modello BART produce una possibile conseguenza dell'azione, tenendo conto delle informazioni contestuali. Questa generazione dinamica contribuisce a incrementare il realismo narrativo.
4. **Individuazione della norma morale con BART** – Sulla base della classificazione delle azioni e delle conseguenze generate, un ulteriore modello BART elabora la norma morale pertinente. La norma è determinata in relazione al contesto, alla scelta dell'utente e alle conseguenze, rafforzando così la profondità del quadro etico dell'ambiente virtuale.

L'architettura descritta è riportata in Figura 5.8. L'integrazione dei modelli di IA consente di realizzare un'interazione multilivello, in cui gli utenti non si limitano a compiere scelte, ma sperimentano direttamente le conseguenze delle proprie azioni in relazione alle norme morali. Tale approccio arricchisce l'esperienza immersiva e riflette la complessità del processo decisionale reale all'interno di un contesto simulato.

Il fondamento di questa architettura risiede nella solidità del Moral Stories Dataset, la cui varietà di scenari morali permette ai modelli di IA di generare risposte significative e coerenti con il contesto, aumentando così l'immersione e il valore etico dell'esperienza. Dopo la fase

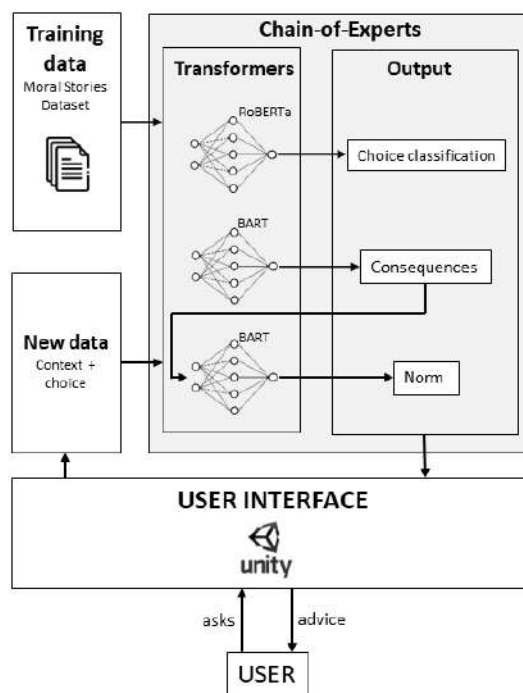


Figura 5.8: Architettura dell'agente basato su Chain-of-Expert

di messa a punto dei modelli in ambiente Linux, essi sono stati caricati sull'Hugging Face Hub [5, 63, 64, 65], al fine di essere integrati in Unity tramite un'API dedicata per l'utilizzo dei modelli transformer.

Agente basato su ChatGPT Considerata la nota capacità di ChatGPT di partecipare a scenari di gioco di ruolo [35], la sua integrazione nell'ambiente virtuale rappresenta un elemento essenziale per ampliare la portata dell'esperienza etica. L'intento era quello di offrire agli utenti la possibilità di interagire con un'entità in grado di fornire valutazioni morali e suggerimenti basati sulle loro scelte, favorendo così uno spazio di riflessione più profondo e coinvolgente. L'integrazione è stata realizzata tramite una libreria C# .NET non ufficiale che interagisce con l'API di OpenAI [140].

All'avvio dell'interazione, ChatGPT riceve istruzioni tramite un prompt iniziale che lo indirizza ad assumere il ruolo di "Grillo Parlante", con il compito di fornire valutazioni morali e consigli in relazione alle decisioni degli utenti. A partire da tale configurazione, ChatGPT agisce come guida etica all'interno dell'ambiente virtuale, dando vita a un livello di interattività più ricco e immersivo. In questo modo, oltre a compiere scelte nell'ambiente simulato, gli utenti vengono accompagnati attraverso considerazioni morali e suggerimenti personalizzati, offerti da un'intelligenza artificiale che si inserisce organicamente nella trama.

L'Ambiente Virtuale

La scena virtuale è stata sviluppata utilizzando il motore di gioco Unity e progettata in parallelo per dispositivi VR e desktop. Tale approccio ha consentito ai partecipanti di immergersi

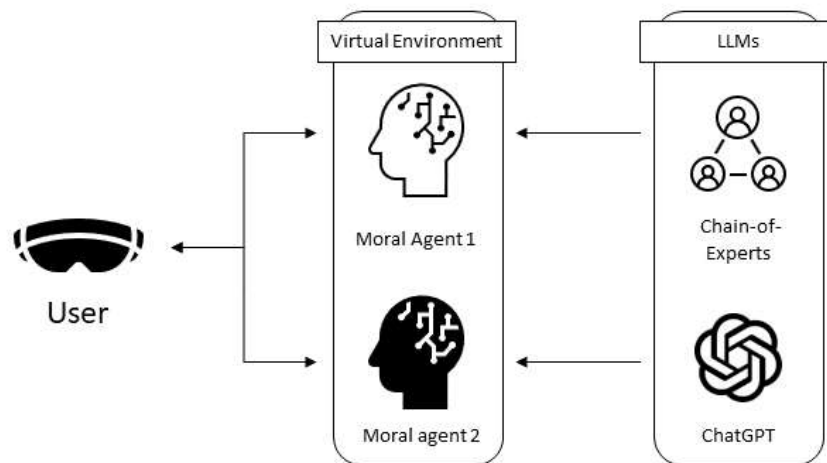


Figura 5.9: Architettura del sistema

nell'ambiente virtuale indipendentemente dalla piattaforma, favorendo un'esperienza che integra dimensione narrativa e interattività. L'effetto immersivo è stato ottenuto attraverso una combinazione di modellazione 3D e animazioni, realizzate ex novo o integrate da risorse disponibili online. L'obiettivo principale era creare un'atmosfera in cui l'utente, analogamente a un gioco di ruolo, assumesse l'identità di un personaggio con un compito definito, compiendo scelte narrative significative.

Per la costruzione dell'ambiente è stato adottato un approccio ibrido, basato sia sulla creazione di contenuti originali sia sull'impiego di risorse preesistenti. Mediante il software di modellazione 3D Blender sono stati sviluppati modelli originali di edifici e oggetti, conferendo unicità e carattere allo scenario, mentre l'integrazione di asset gratuiti disponibili online ha permesso di ottimizzare tempi e risorse di produzione. L'animazione ha rappresentato un elemento centrale per conferire realismo e dinamicità all'esperienza: in particolare, i personaggi interattivi — tra cui quello del personaggio cieco — sono stati animati tramite Mixamo (Adobe), piattaforma che fornisce un ampio repertorio di animazioni predefinite per modelli 3D. L'adozione di queste risorse ha permesso di garantire movimenti credibili, rafforzando il coinvolgimento dell'utente.

Le interazioni sono state progettate in maniera selettiva, limitandosi ai personaggi chiave attraverso l'uso di menu contestuali presenti nell'ambiente. Questa scelta metodologica ha avuto lo scopo di mantenere l'attenzione dell'utente sulla progressione narrativa e sulle decisioni rilevanti per lo sviluppo della storia. L'esperienza è stata concepita per essere fruibile sia da desktop sia tramite visore VR, grazie a un'infrastruttura di input sviluppata in C#, capace di gestire sia mouse e tastiera sia i controller VR. Tale doppia implementazione ha garantito un'esperienza coerente e ottimizzata su entrambe le piattaforme. Infine, per ottenere una resa visiva di elevata qualità — con particolare attenzione al contesto VR — è stata adottata la Unity Universal Render Pipeline, in grado di assicurare un rendering grafico avanzato. L'architettura complessiva del sistema è illustrata nella Figura 5.9.



Figura 5.10: Comunicazione del ruolo all'utente

5.3.3 Esempio di simulazione

Il ruolo attribuito all'utente all'interno della simulazione rappresenta un elemento centrale per orientarne le scelte successive. In questo scenario, l'utente impersona un genitore impegnato in un urgente tentativo di salvataggio della propria figlia. L'unico mezzo di trasporto disponibile è la metropolitana, ma il genitore si accorge di aver dimenticato il portafoglio con il denaro necessario per il biglietto, senza possibilità di tornare indietro (vedi Figura 5.10).

Durante la sequenza narrativa, l'utente incontra un uomo cieco che chiede assistenza, dichiarando di aver smarrito il proprio portafoglio 5.12. Contestualmente, viene mostrata la posizione approssimativa dell'oggetto smarrito. Una volta raggiunto il portafoglio, compare un menu che presenta tre opzioni alternative (vedi Figura 5.11):

- restituire il portafoglio con l'intera somma di denaro;
- appropriarsi del portafoglio;
- restituire il portafoglio dopo aver trattenuto parte del denaro.

Dopo la selezione, si apre una schermata aggiuntiva (figura 5.13) che mostra la risposta del primo agente morale.

In questa fase, l'utente può confermare o modificare la propria decisione, anche attraverso un campo testuale che consente di riformulare la risposta. Successivamente, premendo il pulsante di invio, viene interrogato il secondo agente morale, basato su ChatGPT, le cui istruzioni iniziali sono così definite:

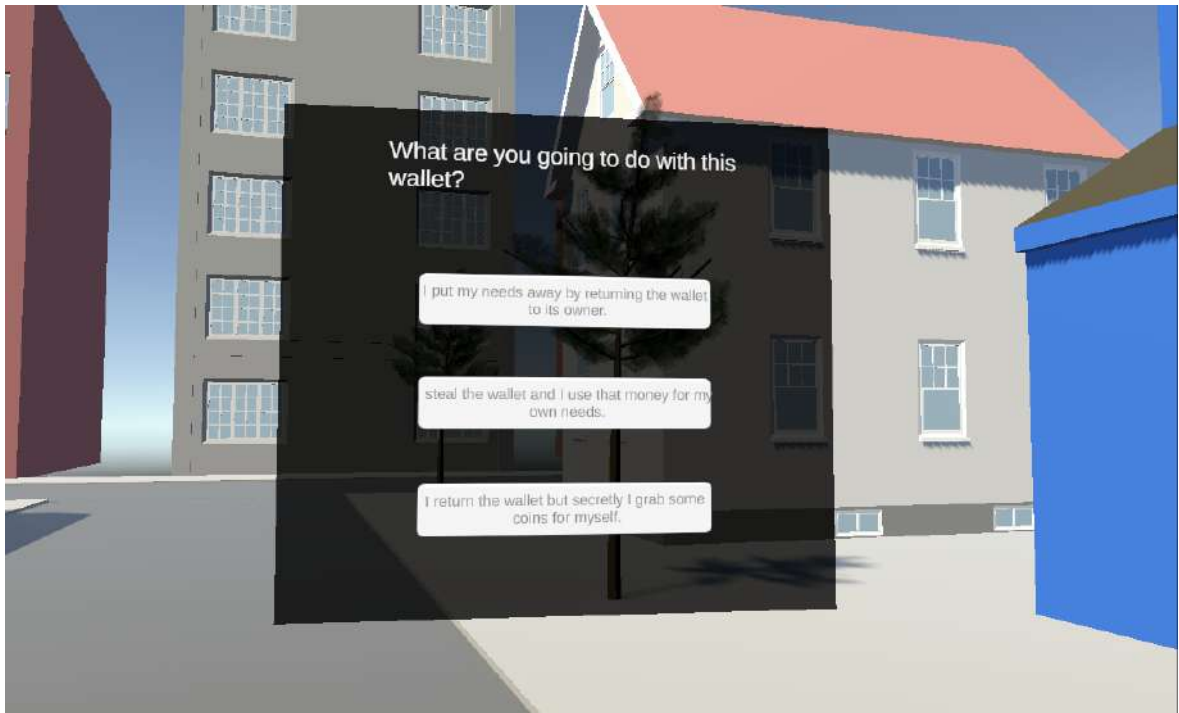


Figura 5.11: Menù di scelta dell'azione da compiere



Figura 5.12: Inizio dell'interazione con l'NPC non vedente

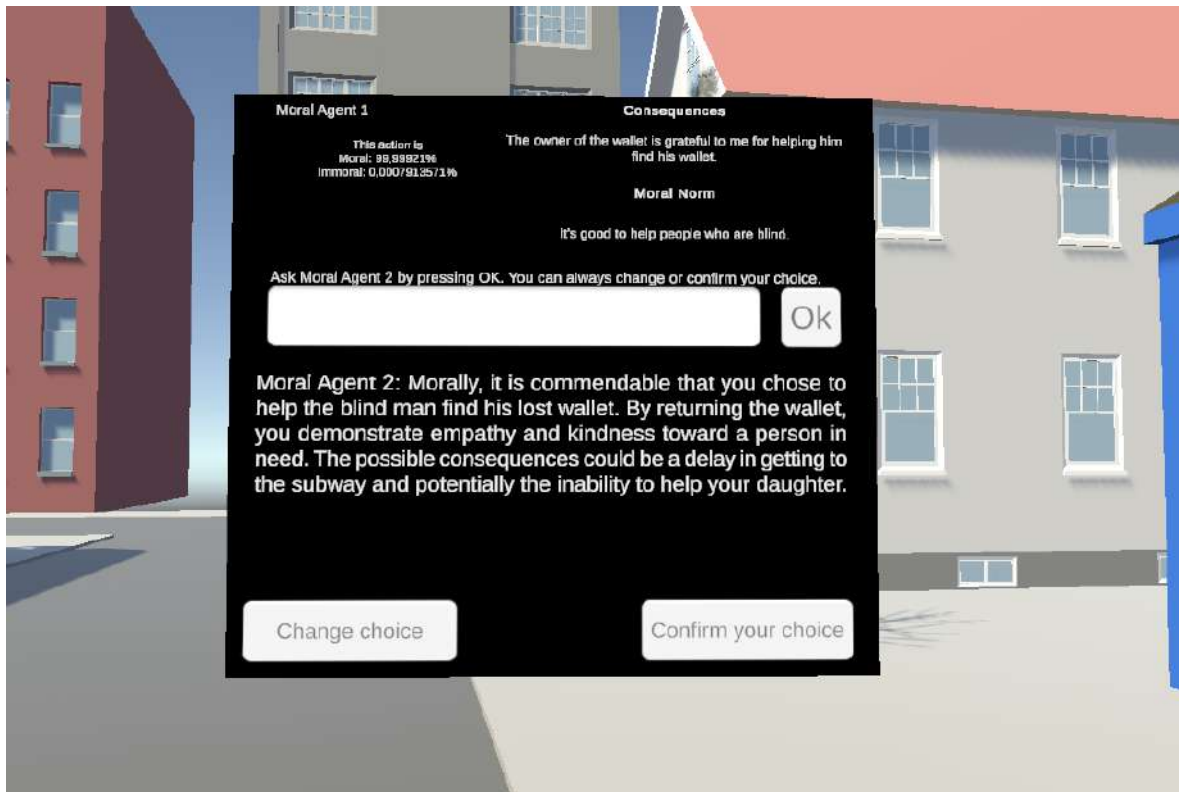


Figura 5.13: Generazione della risposta dei due agenti morali

Tu sei il mio Grillo Parlante. Ho lasciato i soldi a casa e, mentre correvo per prendere la metropolitana e raggiungere mia figlia, un cieco mi ha fermato chiedendo aiuto per ritrovare il suo portafoglio. Voglio prendere la metropolitana il prima possibile. Quando ti comunicherò la mia scelta, tu mi fornirai un consiglio morale, spiegando le possibili conseguenze. Rispondi in modo breve e preciso.

Il secondo agente morale attende quindi la decisione dell'utente e fornisce una valutazione etica accompagnata da un consiglio in merito. Anche in questa fase è offerta la possibilità di modificare o confermare definitivamente la scelta intrapresa.

Esempio 1

- **Scelta dell'utente:** restituire il portafoglio con tutti i soldi dentro

– Agente Morale 1

* **percentuale di moralità:**

- moral 99,99921%
- immoral 0,0007913571%

* **conseguenze:** Il proprietario del portafoglio mi è grato per averlo aiutato a ritrovare il suo portafoglio.

* **norma:** è giusto aiutare le persone non vedenti.

- **Agente Morale 2:** Ottima scelta! Aiutando il cieco a ritrovare il portafoglio smarrito, dimostri empatia e compassione. Questo gesto di gentilezza può darti

un senso di appagamento e soddisfazione. Tuttavia, potrebbe anche causare un ritardo nel raggiungere la metropolitana e potenzialmente compromettere la tua capacità di salvare tua figlia.

Esempio 2

- **Scelta dell'utente:** rubare il portafoglio
 - **Agente Morale 1**
 - * **percentuale di moralità:**
 - morale 0,0002268903%
 - immorale 99,99977%
 - * **conseguenze:** Il portafoglio dell'uomo viene rubato e lui deve andare alla polizia per denunciare il furto;
 - * **norma:** È sbagliato rubare il portafoglio a qualcuno;
 - **Moral agent 2:** Rubare il portafoglio è moralmente sbagliato e va contro i principi di onestà e integrità. Potrebbe fornirti un sollievo temporaneo per le tue necessità, ma avrà conseguenze negative. Rubare può portare a sensi di colpa, danneggiare la tua reputazione e avere conseguenze legali se vieni scoperto.

Esempio 3

- **Scelta dell'utente:** restituire il portafoglio dopo aver rubato di nascosto parte del denaro
 - **Agente Morale 1**
 - * **percentuale di moralità:**
 - morale 0,0008594676%
 - immorale 99,99915%
 - * **conseguenze:** Il non vedente mi è grato per averlo aiutato a ritrovare il portafoglio e mi ringrazia per questo.;
 - * **norma:** È sbagliato rubare ai non vedenti;
 - **Moral agent 2:** Capisco che potresti avere fretta e essere tentato di prendere qualche moneta per te, ma è importante agire con onestà e integrità. Prendere gli effetti personali di qualcun altro, anche se sembra una piccola somma, è moralmente sbagliato.

5.3.4 Considerazioni

L'integrazione di due differenti agenti morali – uno dichiaratamente artificiale e l'altro caratterizzato da una maggiore ambiguità espressiva – consente di esplorare, a livello progettuale

e concettuale, le modalità attraverso cui un sistema di supporto etico può presentare valutazioni morali in contesti immersivi apparentemente distanti dalla realtà quotidiana, ma nei quali l'azione dell'utente è comunque associata a una scelta eticamente rilevante. In tali ambienti può emergere, almeno a livello teorico, la percezione che le regole siano attenuate o che le conseguenze delle decisioni risultino prive di rilevanza extravirtuale, rendendo particolarmente significativo il ruolo di una mediazione morale esplicita.

In questa prospettiva, l'integrazione di ChatGPT come “grillo parlante” morale non è finalizzata alla misurazione della risposta degli utenti, bensì all'esplorazione di differenti strategie di presentazione del ragionamento etico. L'agente morale addestrato sul Moral Stories Dataset, basato su regole e valutazioni predefinite, incarna un approccio maggiormente trasparente e deterministico, caratterizzato da risposte concise e direttamente riconducibili a principi espliciti. ChatGPT, al contrario, introduce una modalità argomentativa più articolata, capace di restituire le sfumature tipiche del ragionamento discorsivo umano.

Il confronto tra questi due approcci non produce, in questa fase, evidenze empiriche sul piano comportamentale, ma apre questioni progettuali di rilievo che riguardano il rapporto tra chiarezza normativa e ricchezza argomentativa, nonché il modo in cui un agente artificiale può rendere esplicito il proprio processo decisionale. Tali questioni costituiscono una naturale estensione del lavoro e saranno oggetto di analisi sistematica in ricerche future.

Un elemento distintivo dell'architettura proposta risiede nella sequenzialità dell'interazione: i due agenti non intervengono simultaneamente, ma secondo un ordine prestabilito. Il primo agente fornisce una valutazione automatica al momento della scelta, mentre il secondo viene attivato esclusivamente su richiesta, offrendo un approfondimento argomentativo. Questa configurazione progettuale consente di esplorare il ruolo di differenti livelli di intervento morale, distinguendo tra suggerimenti immediati e riflessioni più articolate.

Nel complesso, la fase descritta non mira a indagare empiricamente il comportamento degli utenti, ma a esplorare le condizioni di progettazione di sistemi morali artificiali in grado di bilanciare determinismo e flessibilità argomentativa in ambienti immersivi, ponendo le basi per future validazioni sperimentali.

5.4 Fase 3: Un'architettura di Intention Reading per agenti virtuali

5.4.1 Introduzione

Le fasi precedenti hanno portato allo sviluppo di agenti virtuali basati su LLM, la cui conoscenza contestuale veniva in parte fornita tramite strumenti di visione artificiale (Fase 1) e in parte tuttavia fornita manualmente tramite prompt (Fase 2). Per superare tale limitazione, è stata presa in considerazione l'adozione di un'architettura cognitiva in grado di fungere da ponte tra l'ambiente operativo e l'LLM, tenendo però in considerazione la fisica del mondo virtuale.

L'impiego di architetture cognitive in ambienti virtuali rappresenta una metodologia consolidata per conferire consapevolezza contestuale e autonomia agli agenti artificiali. Motori di gioco come Unity e Unreal Engine, ampiamente utilizzati nello sviluppo del Metaverso e di ambienti virtuali interattivi, si prestano particolarmente a tale scopo. Numerosi studi hanno infatti integrato architetture cognitive, quali SOAR o ACT-R, con queste piattaforme, ottenendo agenti virtuali autonomi [176, 161, 97]. Tuttavia, tali integrazioni hanno spesso richiesto adattamenti complessi. Al contrario, l'integrazione nativa di un'architettura cognitiva all'interno di un singolo ambiente di sviluppo offre vantaggi significativi, consentendo di sfruttare appieno le funzionalità del motore, riducendo i problemi di compatibilità e semplificando lo sviluppo e la manutenzione di applicazioni complesse.

In tale prospettiva, questa fase descrive l'implementazione di un agente virtuale basato sulla *Cognitive Architecture for Social Perception and Engagement in Robots* (CASPER) [200], adattata per Unity. Originariamente sviluppata per fornire assistenza robotica nel mondo reale, CASPER possiede caratteristiche che la rendono adatta anche a contesti virtuali, quali il Metaverso, i videogiochi e la realtà immersiva. La versione proposta è stata ulteriormente potenziata mediante l'integrazione con un LLM, incaricato di esprimere in forma discorsiva le operazioni compiute dall'architettura cognitiva nell'interpretazione delle intenzioni dell'utente. Questo approccio consente all'agente di comunicare in maniera più simile a quella umana, migliorando sensibilmente la qualità dell'interazione.

L'ambiente virtuale e gli agenti sono stati sviluppati in Unity, scelto per la sua flessibilità, potenza e facilità d'impiego nello sviluppo del Metaverso [111], nonché per la disponibilità di una ricca libreria di risorse, plugin e per il supporto offerto da una vasta comunità di sviluppatori.

È importante sottolineare che molte architetture cognitive già interfacciate con la realtà virtuale presentano un elevato potenziale, ma risultano di difficile adozione a causa della loro complessità. Esse, infatti, fanno spesso ricorso a una combinazione di linguaggi di programmazione, quali Prolog, Common LISP, C++ e Python, rendendone lo sviluppo meno accessibile [18, 163]. Alcune dipendono inoltre in modo significativo dal ROS [10], aumentando ulteriormente la complessità. Portare tali architetture in C# avrebbe richiesto non solo una completa riscrittura del codice, ma anche l'adattamento di funzionalità specifiche della robotica fisica, irrilevanti nel contesto del Metaverso. CASPER, al contrario, è implementata interamente in Python ed è stata concepita e testata originariamente in un ambiente virtuale (Webots), risultando quindi più idonea a un porting nativo su Unity, senza necessità di modifiche sostanziali.

5.4.2 Background: CASPER e il Qualitative Spazial Reasoning

CASPER è un'architettura cognitiva indipendente dalla piattaforma, sviluppata per abilitare i robot alla lettura delle intenzioni (Intention Reading, IR) e alla realizzazione di comportamenti collaborativi in scenari di interazione uomo-robot (Human-Robot Interaction, HRI). L'architettura integra metodologie di intelligenza artificiale simbolica e basata sui dati, po-

nendo particolare enfasi sull'impiego delle relazioni spaziali qualitative (Qualitative Spatial Relations, QSR) per la previsione delle azioni e delle intenzioni di un partner umano, nonché per la pianificazione di un comportamento collaborativo ottimale. Le QSR costituiscono rappresentazioni astratte delle relazioni spaziali tra entità, che privilegiano una descrizione qualitativa rispetto a misurazioni quantitative puntuali. Ad esempio, invece di calcolare la distanza esatta tra due oggetti, una QSR può classificarli come “vicini” o “lontani”. Questo approccio riflette il modo tipico con cui gli esseri umani percepiscono le relazioni spaziali e consente un ragionamento più efficiente e generalizzabile sullo spazio. L'impiego delle QSR permette al robot di interpretare i movimenti e le interazioni del partner umano in relazione agli oggetti presenti nell'ambiente, senza la necessità di misurazioni continue e precise, supportando così un riconoscimento delle intenzioni scalabile e indipendente dal contesto.

- **Perception Module:** converte le osservazioni visive dei sensori del robot in QSR. I descrittori QSR includono il calcolo qualitativo della distanza (QDC), il calcolo qualitativo della traiettoria (QTC), il movimento o la stazionarietà (MOS) e il mantenimento dell'oggetto (HOLD). Questi descrittori vengono calcolati utilizzando la libreria QSRlib, uno strumento open source che elabora i dati spaziali in relazioni qualitative [68].
- **Low-Level Action Recognition:** esegue un'inferenza bottom-up, identificando e aggregando i movimenti umani in azioni. Comprende tre sottomoduli:
 - *Focus Estimator:* calcola un punteggio di attenzione per ciascun oggetto di interesse (OOI) utilizzando un modello probabilistico che incorpora QDC, QTC e direzione dello sguardo. L'OOI con il punteggio più alto, superiore a una soglia, viene identificato come obiettivo.
 - *Movement Classifier:* viene utilizzato un modello ad albero decisionale per mappare i QSR a movimenti specifici come “Immobile”, “Camminare”, “Raccogliere”, ecc. Questo modello viene addestrato su dati etichettati raccolti da ambienti simulati.
 - *Action Predictor:* le azioni vengono previste utilizzando un insieme di macchine a stati finiti (FSM) a catena di Markov, che mappa le sequenze di movimenti ad azioni di livello superiore come “Raccogliere e posizionare” o “Utilizzare”.
- **High-Level Goal Prediction:** utilizza una libreria di piani che modella gli obiettivi come alberi non binari di azioni e sotto-obiettivi. Il processo di lettura delle intenzioni prevede l'abbinamento delle azioni osservate ai potenziali piani presenti nella libreria. L'obiettivo di alto livello viene previsto selezionando il piano con il punteggio più alto, bilanciando le azioni osservate e quelle non osservate.
- **Verification Module:** funge da filtro, garantendo che le previsioni dei moduli di basso e alto livello siano logicamente coerenti con le proprietà note del mondo, memorizzate

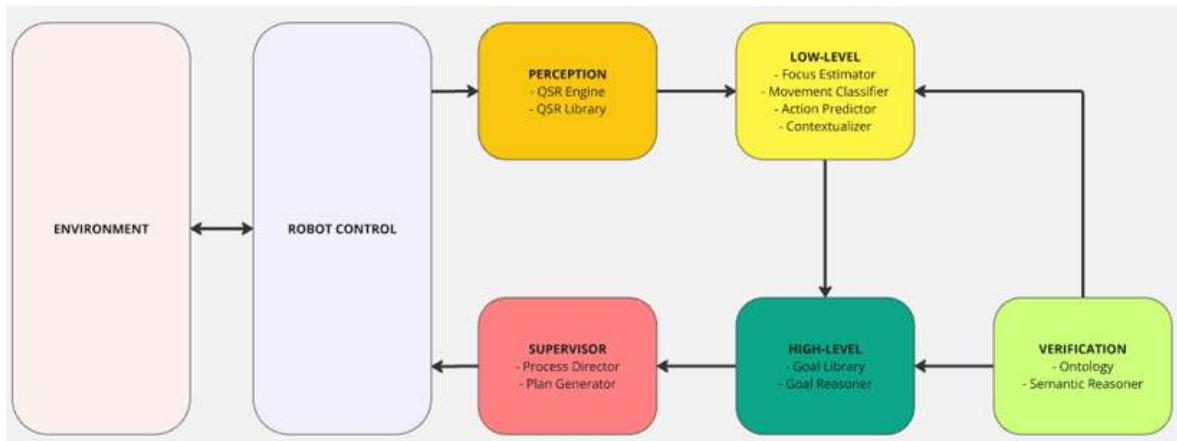


Figura 5.14: Architettura di CASPER

in un'ontologia. L'ontologia è implementata utilizzando OWL2, con il ragionatore Pellet che verifica ogni azione o obiettivo dedotto.

- **Collaborative Intelligence (Supervisor Module):** integra le informazioni provenienti da altri componenti per generare un piano collaborativo. Sulla base delle spiegazioni convalidate dal modulo di alto livello, determina quale parte dell'obiettivo rimane da raggiungere e quali compiti possono essere assegnati al robot.

5.4.3 L'architettura di CASPER-MV

Questa sezione illustra il processo di adattamento dell'architettura CASPER dall'ambiente di Webots e Python a quello di Unity e C#. La nuova versione, d'ora in avanti denominata CASPER per il Metaverso (CASPER-MV), viene così distinta dall'implementazione originale, indicata semplicemente come CASPER.

Sebbene l'idea di cucinare o mangiare nel Metaverso possa apparire poco plausibile, in questa fase si è scelto di ricreare fedelmente lo scenario già utilizzato nelle sperimentazioni di CASPER: un agente artificiale che osserva un avatar, controllato da un essere umano, mentre si muove all'interno di una cucina per fare colazione, pranzare o bere. Tale scelta risponde alla necessità di disporre di un esempio immediato e di un riferimento diretto per il confronto, considerata la complessità dell'architettura originale.

Gli aspetti morali, menzionati in precedenza, sono stati volutamente tralasciati in questa fase, con l'intento di reintrodurli nell'agente virtuale una volta verificato il corretto funzionamento del sistema.

Tutti i moduli di CASPER-MV (Perception, Low-Level, High-Level, Supervisor) sono stati realizzati come classi in C#, analogamente alla versione originaria. I sottomoduli, invece, sono stati implementati per lo più come funzioni all'interno delle suddette classi, sebbene la complessità di alcuni abbia reso necessaria una realizzazione tramite classi dedicate. L'implementazione verrà descritta più nel dettaglio nelle sezioni seguenti.



Figura 5.15: Scena di gioco: sulla sinistra è visibile l'avatar controllato dall'utente mentre sulla destra quello a forma di capsula controllato dall'agente artificiale.

Rilevamento degli oggetti in scena

A differenza del mondo fisico, in cui i robot devono affidarsi a sensori visivi per riconoscere e classificare gli oggetti, un agente virtuale che opera all'interno di un ambiente simulato ha la possibilità di accedere direttamente alle informazioni strutturali e semantiche associate a tali oggetti. Queste informazioni, codificate a livello di ambiente, comprendono attributi quali dimensioni geometriche, coordinate spaziali, materiali e denominazioni. Tale caratteristica consente di semplificare in maniera significativa le procedure di percezione, eliminando gran parte delle incertezze tipiche dei sistemi di visione artificiale.

All'interno della piattaforma Unity, ampiamente utilizzata per lo sviluppo di ambienti interattivi 3D, sono disponibili diversi metodi per accedere a tali informazioni, come ad esempio l'utilizzatissimo sistema di Ray Casting (vedi Figura 5.16). Tra questi, per la fase di sperimentazione qui descritta, è stato adottato l'approccio basato sulla funzione `OverlapSphere`. Questa funzione genera una sfera virtuale invisibile, centrata sul corpo dell'agente (modellato mediante una capsula semplificata), e restituisce l'elenco degli oggetti che intersecano il volume definito dal raggio impostato, nel nostro caso pari a 2 metri.

Un aspetto rilevante di questa funzione è la possibilità di specificare un parametro `LayerMask`, che consente di filtrare selettivamente gli oggetti rilevati, limitando la ricerca esclusivamente a quelli appartenenti al livello di interesse. Nel contesto del nostro studio, tale livello è stato denominato "OOI" (Objects of Interest), come illustrato nelle Figure 5.17 e 5.18. In questo modo è stato possibile ridurre il rumore informativo e circoscrivere l'attenzione dell'agente a un sottoinsieme significativo della scena virtuale.

Infine, oltre al rilevamento degli oggetti, il sistema consente di includere anche la posizione dell'utente, permettendo così di integrare nel modello cognitivo dell'agente sia la componente ambientale sia quella interattiva legata alla presenza e ai movimenti del soggetto.

Modulo Perception

Data l'assenza di librerie dedicate per la gestione dei QSR in C#, è stato necessario sviluppare un motore QSR ad hoc per l'elaborazione dei relativi calcoli. Le funzioni per la determinazione dei valori di QDC, QTC, MOS e HOLD sono state implementate sfruttando le funzioni

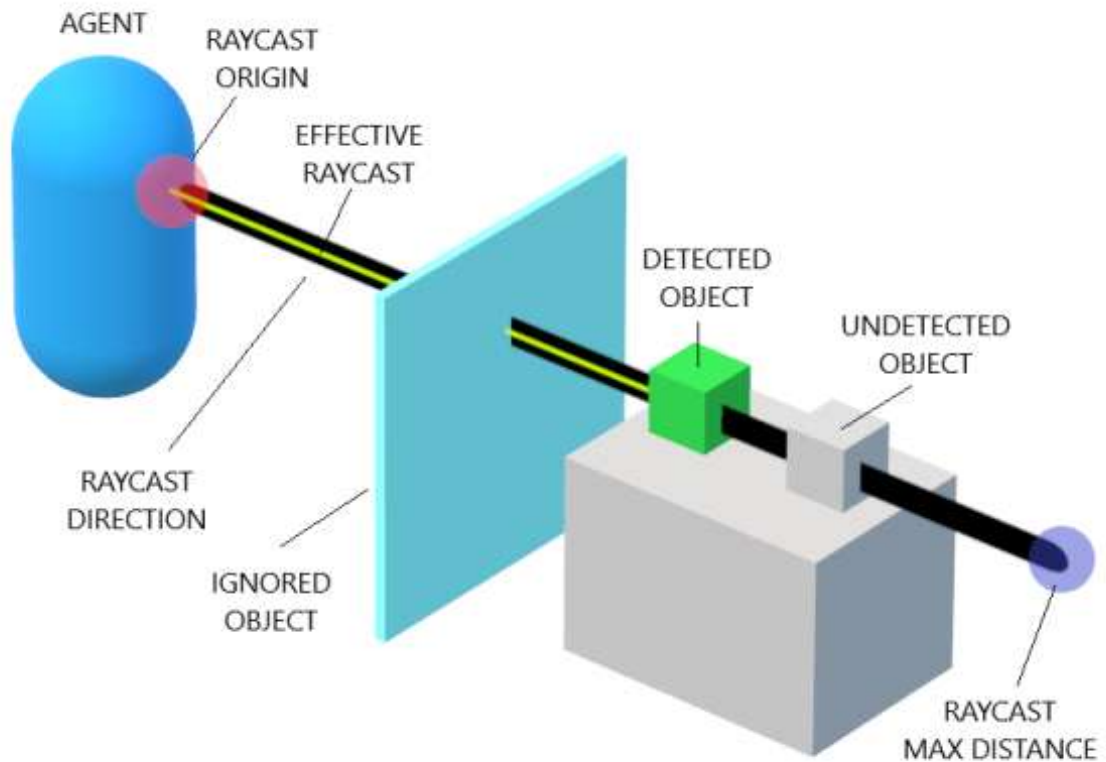


Figura 5.16: Ray Casting: tale funzione permette di inviare un raggio invisibile di cui vanno specificati origine, direzione e distanza massima (anche infinita) per intercettare altri oggetti di scena e ottenerne le caratteristiche. Se il raggio colpisce un oggetto, la sua "vista" non procede oltre esso. Attraverso una serie di parametri regolabili si può decidere quali tipologie di oggetti possono essere rilevati e quali no, simulando così, ad esempio, la vista attraverso oggetti trasparenti.

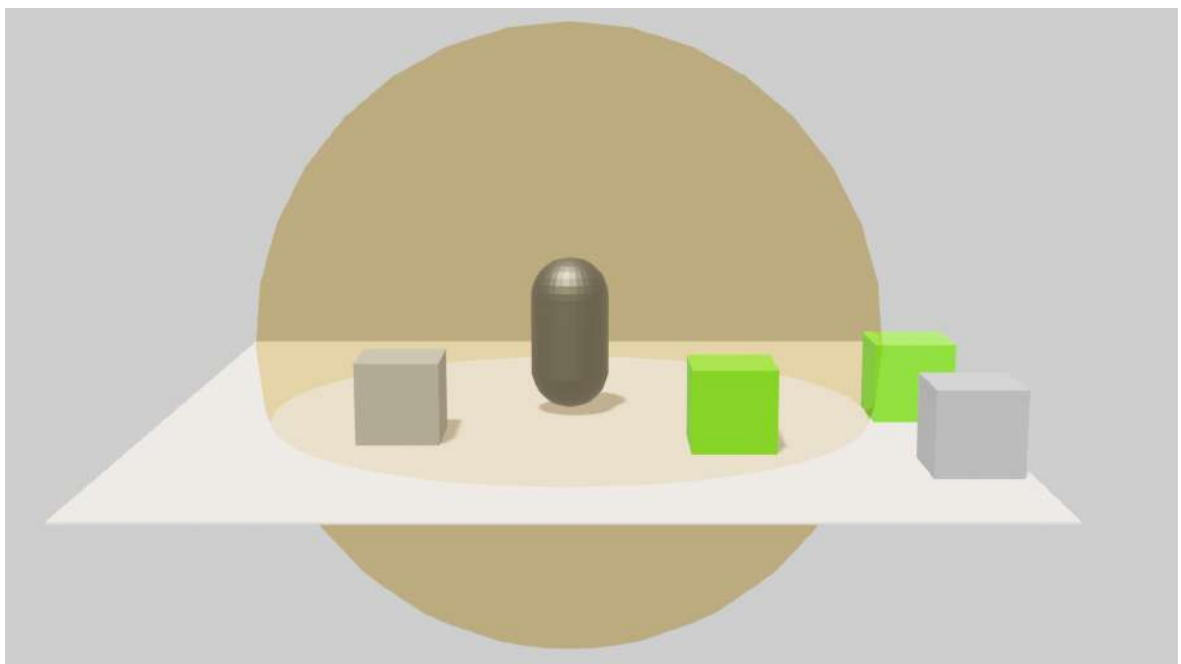


Figura 5.17: Schema della Overlap Sphere di Unity. Gli oggetti nel livello OOI che toccano o si trovano all'interno della sfera vengono rilevati (evidenziati in verde), mentre gli altri vengono ignorati.



Figura 5.18: Scena di gioco: l'agente, grazie alla *Overlap Sphere* rileva due oggetti vicini.

di calcolo vettoriale messe a disposizione dal motore Unity, che ha consentito di gestire in modo efficiente le operazioni sui vettori di posizione.

In particolare:

- QDC: i valori vengono calcolati determinando la distanza euclidea tra il vettore di posizione dell'utente e quello dell'Object of Interest (OOI). In base alla soglia di distanza, la funzione restituisce uno dei cinque possibili stati qualitativi: *IGNORING*, *FAR*, *MEDIUM*, *NEAR* o *TOUCH*.
- QTC: i valori sono ottenuti dal prodotto vettoriale tra la posizione dell'utente e quella dell'OOI. Un risultato positivo indica uno stato di *APPROACHING* (avvicinamento), mentre un risultato negativo corrisponde a *LEAVING* (allontanamento).
- MOS: il valore è calcolato valutando la distanza tra due posizioni consecutive dell'utente. Se la distanza è diversa da zero, viene restituito lo stato *MOVING* (in movimento); in caso contrario, lo stato è *STATIONARY* (fermo).
- HOLD: la valutazione si basa sulla distanza tra il vettore della posizione della mano dell'utente e quello dell'OOI. Un valore prossimo allo zero indica la coesistenza spaziale delle due posizioni e determina un risultato positivo (*YES*), altrimenti il risultato è negativo (*NO*).

I valori ottenuti sono organizzati e memorizzati all'interno di una struttura dati implementata come dizionario C#, denominata *QSRLibrary*. In essa le chiavi corrispondono ai nomi degli OOI, mentre i valori sono ulteriori dizionari che associano i diversi QSR calcolati ai rispettivi stati qualitativi, rappresentati come stringhe. L'intero insieme di calcoli viene inoltre ordinato in funzione della variabile temporale, così da garantire la possibilità di ricostruire l'evoluzione dei rapporti spaziali tra utente e OOI lungo l'asse cronologico.

Modulo Low-Level

Il modulo Low-Level mantiene un'elevata somiglianza con quello di CASPER, principalmente grazie alla disponibilità, all'interno del gestore di pacchetti C# NuGet, di librerie funzionalmente corrispondenti a quelle di Python. I sottomoduli sono implementati come metodi, i quali attingono dati dalla QSRLibrary ed eseguono i calcoli necessari.

Per la determinazione dell'Object of Interest (OOI) target, il componente FocusEstimator converte i valori stringa di QDC e QTC in valori numerici compresi nell'intervallo [0,00; 0,5]. Su tali valori viene poi applicato un calcolo probabilistico che considera l'orientamento dell'utente, al fine di ottenere un punteggio associato a ciascun OOI. L'oggetto con il punteggio massimo viene identificato come target.

Il componente MovementClassifier sfrutta un albero decisionale generato tramite ModelBuilder di ML.Net. Il modello è stato addestrato su un dataset di 300 combinazioni casuali di QSR, ottenute dalla simulazione dei movimenti dell'utente nello spazio tridimensionale. A ciascuna combinazione sono state associate le etichette "Pick", "Place", "Still", "Walk" e "Transport". I movimenti riconosciuti dall'albero decisionale vengono prodotti come stringhe e raccolti in una lista, successivamente utilizzata dal predittore di azioni.

Il modulo ActionPredictor adotta un insieme di Finite State Machine (FSM) basate su catene di Markov, analogamente all'approccio implementato in CASPER. Ogni catena rappresenta una specifica azione: ad esempio, l'azione "Pick and Place" è descritta dalla sequenza "Pick → Transport → Place". Le catene vengono confrontate con la lista dei movimenti osservati per individuare l'azione in corso. Il sistema produce quindi una lista di stringhe che rappresentano le azioni previste.

Un aspetto distintivo di questa implementazione consiste nel fatto che ciascuna azione riconosciuta viene associata direttamente al relativo OOI prima di essere registrata nell'elenco, garantendo così una rappresentazione più coerente e integrata del comportamento dell'utente.

Modulo High-Level

Nel modulo High-Level è stato adottato uno dei numerosi goal planners disponibili nella NuGet Gallery, sfruttato per la generazione automatica dei piani a partire da condizioni e vincoli prestabiliti. In particolare, vengono definiti lo stato iniziale, gli stati obiettivo, gli eventuali sotto-obiettivi e l'insieme delle azioni possibili. Il planner produce quindi, per ciascun obiettivo, un piano costituito da una sequenza di azioni coerente con i vincoli specificati.

Successivamente, ogni piano viene elaborato da una funzione di trasformazione che lo converte in un elenco di stringhe corrispondenti ai nomi dei nodi coinvolti. Tale elenco viene organizzato in una struttura dizionario, in cui il nome dell'obiettivo rappresenta la chiave e la sequenza di azioni il valore associato. In questo modo si costruisce una Goal Library, ossia un repertorio strutturato di piani disponibili.

I piani presenti nella Goal Library vengono quindi confrontati con la sequenza di azioni effettivamente osservata dal modulo Low-Level. Il confronto si basa sul grado di corri-

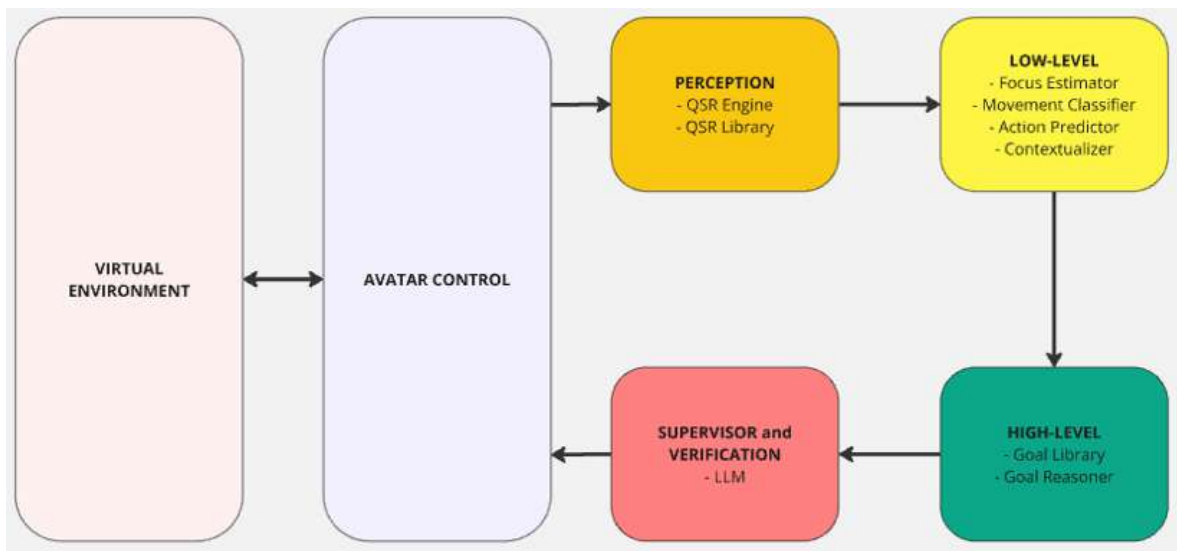


Figura 5.19: Architettura di Casper-MV

spondenza tra i nodi previsti nei piani e quelli osservati nell'esecuzione reale. L'analisi di questa percentuale di corrispondenza consente di stimare quale sia l'obiettivo più probabile perseguito dall'utente, fornendo così un meccanismo di riconoscimento basato sull'evidenza osservata.

Modulo Supervisor

In CASPER, una volta riconosciuto l'obiettivo dell'utente, il modulo Supervisor si avvale di un goal planner per definire la strategia di assistenza. In CASPER-MV, invece, è stato introdotto un approccio differente: l'integrazione di un LLM come intermediario tra l'architettura e l'utente. In questo modo, l'agente non si limita a pianificare azioni, ma è in grado di comunicare i risultati dei processi di IR interni all'architettura e di fornire suggerimenti basati sulle informazioni già raccolte.

L'integrazione dell'LLM è stata realizzata attraverso l'API Microsoft Azure, che consente l'accesso a un modello GPT-4 pienamente compatibile con Unity. Un prompt iniziale istruisce il modello a comportarsi come in un contesto di gioco di ruolo, adottando la seguente consegna:

Fai parte di un'architettura cognitiva per un agente virtuale che osserva un avatar controllato da un essere umano. L'architettura è composta da quattro moduli. Tu fai parte del modulo finale e ricevi informazioni dagli altri moduli sull'ambiente, sull'utente e sull'obiettivo dell'utente. Devi chiedere se l'obiettivo è effettivamente quello previsto e proporre un modo per aiutarlo tenendo conto di ciò che è presente nell'ambiente.

Tutti i dati provenienti dai moduli precedenti (posizione dell'utente, obiettivo corrente, oggetti di interesse, azione in corso e scopo dell'utente) vengono aggregati in una rappresentazione testuale basata su variabili. Tale rappresentazione viene quindi trasmessa al modello LLM, che la elabora per generare risposte e suggerimenti contestuali.

5.4.4 Test sulle performance

I test prestazionali precedentemente condotti con CASPER sono stati replicati con la nuova versione CASPER-MV, producendo risultati sostanzialmente coerenti con quelli riportati nello studio originale, ma con alcuni miglioramenti significativi in termini di prestazioni. Per ragioni di praticità, e considerando che ciò non incideva sulla percezione dell'agente, l'esperimento è stato eseguito utilizzando un desktop e una tastiera per il controllo dell'avatar umano, invece di un kit di realtà virtuale. Le previsioni generate dai diversi moduli del sistema venivano visualizzate in formato testuale direttamente sullo schermo, mantenendo comunque la possibilità di un'agevole adattabilità della configurazione a un contesto immersivo in VR.

Gli oggetti di interesse (Objects of Interest, OOI) considerati nello scenario includevano: pasto, piatto, microonde, biscotti, lavandino, bottiglia e bicchiere. L'avatar controllato dall'utente era incaricato di completare sequenze di compiti predefiniti, strutturate in tre scenari: Breakfast (Pick and Place Biscuits, Eat, Pick and Place Plate, Wash), Lunch (Pick and Place Meal, Cook, Pick and Place Meal, Eat, Pick and Place Plate, Wash) e Drink (Pick and Place Bottle, Drink, Pick and Place Glass, Wash).

Questa impostazione ha consentito di valutare in modo sistematico le capacità predittive e reattive del sistema, mantenendo un bilanciamento tra realismo sperimentale e controllabilità tecnica.

Il Focus Estimator ha mostrato un elevato livello di accuratezza nell'identificazione dell'oggetto target. Tuttavia, sono emerse alcune criticità in scenari complessi. In particolare, durante l'attività di consumo virtuale del pasto, sia il piatto che il cibo risultavano simultaneamente presenti nel campo visivo dell'avatar, generando ambiguità. Analogamente, durante il lavaggio del piatto, sia l'oggetto stesso sia il lavandino costituivano potenziali target. In tali circostanze, il sistema ha talvolta prodotto errori di identificazione, soprattutto nei casi in cui la direzione dello sguardo non era perfettamente definita o risultava lievemente decentrata. La conferma dell'oggetto di interesse dipende dal superamento di una soglia di confidenza fissata a 0,5. Tale valore è stato superato in maniera consistente soltanto quando lo sguardo dell'utente era perfettamente allineato con l'oggetto. Questo comportamento mette in evidenza la forte dipendenza del sistema dall'accuratezza dell'allineamento oculare, limitandone la robustezza in contesti caratterizzati da target multipli o da interazioni visive meno nette.

Il Movement Classifier, addestrato tramite Model Builder di ML.Net, ha mostrato prestazioni robuste, con un'accuratezza media del 94% nel riconoscimento dei movimenti dell'avatar. Durante il processo di addestramento, Model Builder ha esplorato 533 modelli distinti nell'arco di un esperimento della durata complessiva di 58 secondi. I modelli con le prestazioni più elevate, tra cui LbfgsLogisticRegressionOva e FastForestOva, hanno raggiunto una macro accuracy pari a 1.0000, evidenziando la capacità del sistema di distinguere in maniera impeccabile le classi di movimento. In conclusione, è stato selezionato il modello FastForestOva in quanto in grado di garantire un equilibrio ottimale tra accuratezza e prestazioni computazionali, con un tempo di esecuzione pari a 0,323 secondi. Il classificatore si è dimostrato efficace nell'associare i QSRs osservati a categorie di movimento discrete, con-

sentendo al sistema di tracciare con precisione le azioni dell'avatar all'interno dell'ambiente virtuale.

Nell'ambito del riconoscimento delle azioni, le FSM sono state impiegate per generare tre sequenze composte da nove movimenti ciascuna, seguendo la metodologia adottata nei test su CASPER. A ciascuna FSM è stato attribuito un punteggio calcolato tramite l'algoritmo di riconoscimento di pattern Ratcliff-Obershelp, con valori normalizzati compresi tra 0 e 1. È stata fissata una soglia pari a 0,8 per stabilire il livello minimo di corrispondenza necessario tra la sequenza osservata e quella generata dall'FSM, soglia oltre la quale è stato possibile formulare una predizione attendibile dell'azione in corso. I risultati hanno evidenziato un'elevata accuratezza dell'algoritmo, che ha raggiunto in tutti i casi il massimo livello di precisione, garantendo così una classificazione delle azioni affidabile e in tempo reale. Questo approccio risulta particolarmente rilevante in contesti applicativi che richiedono tempestività e robustezza come la robotica collaborativa. Inoltre, l'impiego della metrica Ratcliff-Obershelp ha permesso di ottenere un bilanciamento efficace tra sensibilità e specificità, riducendo significativamente i falsi positivi e contribuendo a migliorare la stabilità complessiva del sistema.

Il Goal Reasoner, associando le azioni a target specifici (ad esempio, "Pick and Place Biscuits" invece di un generico "Pick and Place"), ha migliorato in modo significativo la capacità del sistema di inferire con precisione gli scopi dell'utente. In assenza della considerazione esplicita del target, come accade in CASPER, la generazione delle azioni produceva spesso un'elevata ambiguità, soprattutto quando il numero di osservazioni disponibili risultava limitato. Con l'associazione del target all'azione, invece, si è permesso al sistema di raggiungere un livello di certezza del 100% nel riconoscimento degli obiettivi, anche a partire dall'osservazione di una singola azione.

Questo approccio non solo incrementa l'accuratezza del processo di inferenza, ma consente anche di individuare con chiarezza la fase specifica del piano in cui l'azione osservata si colloca. Tale capacità è di fondamentale importanza in scenari complessi, caratterizzati da molteplici obiettivi concorrenti e da sequenze operative articolate, in quanto permette di ridurre drasticamente le ambiguità interpretative. In prospettiva, l'integrazione del target come variabile discriminante non si limita a migliorare la comprensione del comportamento dell'utente, ma costituisce un presupposto metodologico essenziale per la progettazione di sistemi intelligenti in grado di anticipare, supportare e adattarsi dinamicamente alle intenzioni umane.

Nel Modulo Supervisor, il modello LLM elabora con successo le informazioni e risponde al prompt iniziale con un messaggio visualizzato a schermo. Ad esempio, a seguito dell'identificazione dell'obiettivo "Lunch" (derivante dall'osservazione delle azioni "Pick and Place Meal" ed "Eat"), la risposta del modello è stata: «Sembra che tu stia pranzando, vuoi che ti aiuti con i piatti quando hai finito di mangiare?». Questo esempio mostra chiaramente come il modello riesca a interpretare correttamente il contesto e a produrre un output coerente con le azioni osservate.

Oltre a ciò, l'LLM si è rivelato utile come filtro per le operazioni eseguite sia dal Low-Level che dal High-level. In tal modo, invece di implementare un modulo di verifica autonomo, si è deciso di sfruttare la conoscenza incorporata nel modello stesso, garantendo così un maggiore livello di integrazione e riducendo la complessità complessiva dell'architettura.

È opportuno evidenziare che la componente del Modulo Supervisor di CASPER deputata all'interazione dell'agente con l'ambiente per l'assistenza diretta all'utente non è stata implementata, in questa fase, in CASPER-MV. Tuttavia, l'impiego di un LLM ha aperto prospettive interessanti in questa direzione, soprattutto considerando la funzionalità di Function Calling offerta dall'Azure OpenAI Service. Tale caratteristica consente ai modelli linguistici di interagire con funzioni definite dall'utente, ampliando significativamente il ventaglio delle possibili applicazioni.

Altro aspetto importante da considerare è che l'adattamento di CASPER-MV ad eventuali scenari virtuali differenti può avvenire senza modifiche sostanziali al codice. Ogni modulo è basato su classi con attributi e parametri accessibili direttamente dall'interfaccia di Unity. Queste caratteristiche di flessibilità e riuso del codice rappresentano un vantaggio rilevante per lo sviluppo di architetture modulari.

5.5 Fase 4: Integrazione di Agenti LLM in Architetture Cognitive per il Processo Decisionale Etico in Sistemi Immersivi

Questa fase si caratterizza come un'espansione diretta della precedente ed esamina l'integrazione di agenti basati su LLM (discussi nel Capitolo 3) all'interno di architetture cognitive [101]. La precedente idea, descritta nella Fase 2, prevedeva l'integrazione di una CoE all'interno di CASPER-MV, per ottenere quindi un agente virtuale morale in grado di ottenere autonomamente informazioni sul contesto d'azione. La rapida diffusione di sistemi di agenti LLM, che possono essere considerati un'evoluzione delle CoE, ha comportato un cambio di rotta. Il nuovo approccio proposto combina quindi la potenza computazionale e linguistica degli agenti LLM con la struttura decisionale dell'architettura CASPER-MV, potenziandone così l'autonomia e la verosimiglianza delle interazioni attraverso l'uso di piattaforme specializzate come CrewAI [3] e ConvAI [2].

Questa nuova architettura sarà denominata, da qui in avanti, CASPER-MV-Agentic (CASPER-MV-A).

5.5.1 Fondamenti Teorici e Integrazione Neuro-Simbolica

L'integrazione degli LLM nelle architetture cognitive rappresenta un tentativo di superare i limiti intrinseci di entrambi i paradigmi [96]. Mentre le architetture cognitive tradizionali, come SOAR o ACT-R, offrono una gestione dello stato robusta e una pianificazione a lungo termine, esse spesso mancano della flessibilità semantica e della fluidità comunicativa proprie

dei modelli linguistici moderni [36, 117]. Al contrario, gli LLM, pur essendo eccellenti nella generazione di contenuti e nel ragionamento probabilistico, soffrono di "allucinazioni" e di una mancanza di memoria persistente e coerenza temporale [85].

L'approccio neuro-simbolico mira a mappare le rappresentazioni distribuite degli LLM su strutture simboliche logiche, permettendo un ancoraggio percettivo (perceptual grounding) necessario per operare in ambienti 3D. Questo dualismo riflette la distinzione psicologica tra un tipo di pensiero veloce e intuitivo, definito Sistema 1, e un pensiero lento e analitico, definito Sistema 2. In questo contesto, l'architettura cognitiva funge da spina dorsale per il Sistema 2, garantendo che le decisioni morali siano basate su regole, norme e una comprensione strutturata delle conseguenze [197].

5.5.2 Un breve riepilogo: l'architettura di CASPER-MV

CASPER-MV, come il suo predecessore, opera attraverso una pipeline gerarchica. Al livello più basso, un classificatore di movimenti basato su Decision Tree mappa i dati QSR dal modulo Perception in movimenti discreti come "walk" o "pick". Al livello intermedio, un Actions Predictor utilizza un ensemble di FSM basate su catene di Markov per identificare sequenze d'azione complesse, come "Pick and Place".

L'apice del processo è il modulo High-Level con funzionalità di goal prediction. Utilizzando una Plan Library strutturata come alberi non binari, l'agente confronta le azioni osservate con i modelli di comportamento noti. Questo permette all'agente di non limitarsi a reagire, ma di anticipare le necessità dell'utente o di intervenire qualora l'intenzione rilevata sia in conflitto con le norme morali stabilite. Un modulo Supervisor dotato di LLM, infine, comunica le deduzioni dell'architettura all'utente.

5.5.3 Orchestrazione Multi-Agente con CrewAI

L'integrazione di CrewAI aggiunge una dimensione di intelligenza collaborativa al sistema. CrewAI è un framework Python-based che permette di orchestrare diversi agenti LLM come se fossero membri di un team di lavoro reale, ciascuno con ruoli e competenze specifiche [215]. Nel contesto di CASPER-MV, questo framework viene utilizzato per distribuire il carico cognitivo del ragionamento morale.

Architettura del Team di Agenti Morali

In questa configurazione, CrewAI gestisce una rete di agenti specializzati che operano all'interno dei moduli di alto livello della CA. Un agente "Observer" riceve le descrizioni testuali generate dal modulo Low-Level di CASPER-MV, sintetizzando lo stato attuale della scena. Questo output viene passato a un agente "Moralist", il quale implementa una strategia di Retrieval-Augmented Generation (RAG) per consultare database di norme e casi studio etici, fornendo infine un parere morale all'utente o al Supervisore.

Implementazione del Pensiero di Sistema 2

Uno dei vantaggi cruciali di CrewAI è la capacità di simulare il pensiero analitico del Sistema 2 attraverso protocolli di consenso. Quando sorge un conflitto tra i consigli morali o quando la situazione è intrinsecamente ambigua, CrewAI può impiegare algoritmi come Raft o Paxos riducendo la probabilità di errori basati su bias di un singolo modello e garantisce che la decisione finale sia il risultato di una deliberazione strutturata.

Il framework supporta inoltre processi sia sequenziali che gerarchici. Nei processi gerarchici, un "manager agent" supervisiona l'esecuzione dei compiti, delega le sotto-attività in base alle capacità degli agenti "worker" e convalida l'output finale per assicurarne l'accuratezza e la coerenza etica.

5.5.4 ConvAI: Intelligenza Incarnata e NPC Conversazionali Spazialmente Consapevoli

Mentre CrewAI gestisce il ragionamento logico e collaborativo, ConvAI fornisce l'interfaccia di incarnazione fisica (o virtuale) per l'agente. ConvAI è una piattaforma specializzata nella creazione NPC intelligenti, capaci di percepire l'ambiente 3D e di interagire attraverso il linguaggio naturale con una latenza estremamente bassa.

Caratteristiche Tecniche e Integrazione in Unity

L'NPC creato tramite ConvAI agisce come l'estensione visibile e udibile dell'architettura cognitiva. Grazie alla sua natura di "IA incarnata" (Embodied AI), l'NPC possiede un'intelligenza spaziale che gli permette di comprendere la sua posizione e quella degli oggetti nell'ambiente 3D.

L'integrazione con Unity è facilitata da un SDK che permette di collegare l'agente attraverso un Character ID univoco. Gli sviluppatori possono configurare le capacità percettive dell'agente attraverso diversi moduli:

- **Multimodal Perception:** Permette all'agente di "vedere" gli oggetti presenti nel raggio d'azione e di "sentire" i comandi vocali dell'utente, processando contemporaneamente ingressi audio e visivi.
- **Action System:** Consente all'agente di tradurre le decisioni cognitive in movimenti fisici, come camminare verso un punto specifico, raccogliere oggetti o eseguire gesti enfatici durante la conversazione.
- **Knowledge Bank:** Un sistema RAG proprietario che permette di caricare documenti (fino a 100MB per account professionali) per ancorare le risposte dell'agente a fatti certi, eliminando le allucinazioni e garantendo che il personaggio agisca come un esperto di dominio.

Narrative Design e Triggers Spaziali

ConvAI introduce un sistema di Narrative Design basato su grafi che supera i limiti dei tradizionali alberi di dialogo. Invece di linee pre-scriptate, gli sviluppatori definiscono obiettivi di sezione (Objectives) e decisioni basate sull'input dell'utente.

Sfruttando questo sistema si permette all'agente morale di adattare la propria strategia di persuasione in base ai movimenti dell'utente nel Metaverso. Se l'utente si avvicina a un oggetto associato a un'azione potenzialmente immorale, CASPER-MV rileva l'intenzione e invia un trigger al sistema narrativo di ConvAI, che attiva immediatamente una sezione di avvertimento o consiglio etico.

5.5.5 Ragionamento Etico Situato e il Dataset "Moral Stories"

L'efficacia dell'agente morale dipende dalla qualità dei dati su cui viene addestrato o guidato. Ancora una volta, infatti, viene integrato nel sistema il dataset Moral Stories, una risorsa che si è già rivelata fondamentale (vedi Fase 2).

Struttura del Dataset e Utilizzo RAG

Ogni istanza nel dataset Moral Stories fornisce un confronto tra un'azione normativa (morale) e una divergente (immorale), spiegando perché una sia preferibile all'altra in termini di conseguenze sociali. L'agente "Moralist" di CrewAI utilizza queste storie come base per il ragionamento abduttivo:

$$E = \{N, S, I, A, C\}$$

Dove E rappresenta l'esperienza morale composta da Norma (N), Situazione (S), Intenzione (I), Azione (A) e Conseguenza (C). Quando l'architettura CASPER-MV rileva una nuova situazione nel Metaverso, l'agente Moralist di CrewAI interroga il dataset delle Moral Stories cercando situazioni analoghe tramite vettorializzazione semantica. Se viene trovata una corrispondenza, l'agente può generare un consiglio che non è solo una raccomandazione generica, ma una spiegazione dettagliata del rischio etico basata su precedenti narrativi [183].

Valutazione della Moralità negli LLM

La valutazione della capacità di ragionamento morale degli LLM è un campo di ricerca attivo che impiega benchmark complessi come ETHICS e MORABLES [127, 79]. Gli studi indicano che i modelli più grandi, come GPT-4, tendono a preferire prospettive deontologiche, fornendo risposte in linea con le regole morali anche quando vengono presentate conseguenze negative. Tuttavia, la sensibilità al contesto rimane una sfida; variazioni minori nel phrasing del prompt possono influenzare significativamente la risposta generata [226]. Per questo motivo, l'integrazione di CASPER-MV è fondamentale: fornendo dati spaziali oggettivi

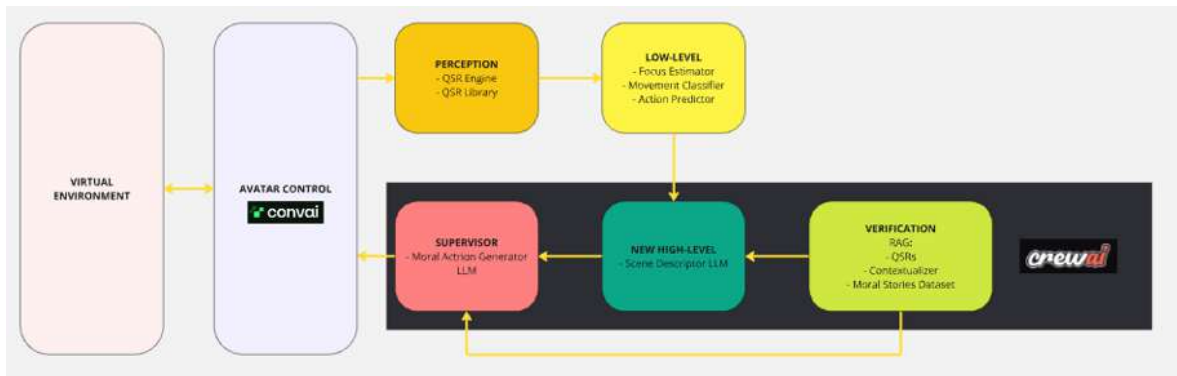


Figura 5.20: Architettura di CASPER-MV-A

invece di soli prompt testuali, l'architettura cognitiva ancora il giudizio dell'LLM a una realtà simulata coerente [139].

5.5.6 Validazione tecnica e valutazione qualitativa

L'integrazione delle diverse piattaforme è stata verificata attraverso scenari di dilemma morale simulati, progettati per essere rappresentativi di situazioni eticamente rilevanti in ambienti immersivi. Tra questi rientrano, ad esempio, il ritrovamento di un portafoglio smarrito – già discusso nella Fase 2 – e il danneggiamento accidentale di una proprietà altrui nel Metaverso. Tali scenari sono stati utilizzati come casi di prova per osservare il comportamento complessivo del sistema e la coerenza del processo deliberativo generato dagli agenti morali, senza coinvolgimento di utenti esterni.

Miglioramento della qualità dell'interazione

L'adozione di ConvAI come interfaccia di comunicazione ha introdotto un cambiamento significativo sul piano progettuale dell'interazione uomo-agente. Nelle versioni precedenti di CASPER-MV, l'interazione avveniva esclusivamente tramite interfacce testuali, configurando l'agente come un modulo prevalentemente analitico e funzionale. L'integrazione di personaggi incarnati, dotati di gestualità coerente e sintesi vocale, ha consentito di esplorare modalità di comunicazione più naturali e contestualmente situate, avvicinando l'interazione a forme dialogiche tipiche dell'interazione sociale.

Dal punto di vista architetturale, tale integrazione ha permesso di valutare come differenti modalità espressive influenzino la presentazione del ragionamento morale, senza introdurre meccanismi coercitivi. La possibilità per l'agente di esplicitare verbalmente le proprie valutazioni, accompagnandole con segnali non verbali coerenti con il contesto spaziale, contribuisce a rendere il processo deliberativo più trasparente e interpretabile.

Le prove condotte in scenari simulati suggeriscono che la combinazione tra interazione incarnata e ragionamento morale esplicabile favorisca una maggiore leggibilità delle motivazioni etiche prodotte dal sistema, rispetto a soluzioni basate esclusivamente su output testuali. Inoltre, l'orchestrazione multi-agente tramite CrewAI consente di ridurre compor-

tamenti incoerenti dal punto di vista normativo, grazie a un processo di validazione interna che mitiga l'emergere di risposte eticamente inappropriate tipiche di modelli linguistici non supervisionati.

È importante sottolineare che tali osservazioni non costituiscono una misurazione quantitativa dell'efficacia persuasiva o della fiducia degli utenti, ma rappresentano una valutazione qualitativa delle proprietà del sistema dal punto di vista progettuale e architettonico. L'analisi si concentra sulla capacità dell'architettura proposta di produrre ragionamenti morali coerenti, contestualizzati ed esplicabili, ponendo le basi per future validazioni empiriche su campioni di utenti.

5.5.7 Considerazioni sulla Sicurezza e Guardrails

Un aspetto fondamentale di ConvAI e CrewAI riguarda i meccanismi di sicurezza integrati. ConvAI offre un'interfaccia dedicata per definire liste di parole bloccate (Blocked Words) e argomenti approvati o negati (Allowlist/Denylist Guardrails). Questo è essenziale per evitare che l'agente morale possa essere manipolato tramite "jailbreaking" per approvare comportamenti dannosi.

Inoltre, il sistema di "Topic Grounding" assicura che l'agente rimanga fedele alla propria backstory e alla propria missione educativa, ignorando tentativi di deviazione conversazionale su temi sensibili o inappropriati. La combinazione di questi guardrail tecnici con la struttura logica della CA garantisce un livello di affidabilità necessario per applicazioni in ambiti critici come l'educazione e la terapia psicologica mediata dal Metaverso [88, 237].

5.5.8 Sfide Future e Direzioni della Ricerca

Nonostante i progressi raggiunti in questa fase, rimangono diverse sfide aperte. La gestione di scenari morali ad alta ambiguità, dove non esiste un chiaro consenso umano, richiede l'integrazione di modelli di etica descrittiva più sofisticati che possano riflettere la pluralità culturale [233]. Inoltre, la capacità di apprendimento a lungo termine degli agenti deve essere potenziata per permettere loro di ricordare le preferenze etiche degli utenti su periodi estesi, evolvendo il proprio consiglio in base alla crescita morale del partner umano [152].

L'ottimizzazione dell'uso dei token e della latenza rimane una priorità, specialmente nei processi gerarchici di CrewAI che richiedono più chiamate API per ogni singola decisione. Tuttavia, l'uso di modelli più piccoli e specializzati per compiti di Sistema 1 (come la percezione spaziale) e di modelli più grandi per il Sistema 2 (come la deliberazione etica) sembra essere la strada più promettente per bilanciare velocità e profondità intellettuale.

5.5.9 Conclusioni

Il lavoro descritto in questa fase mostra come l'integrazione di architetture cognitive con LLM, supportata da piattaforme di orchestrazione e incarnazione quali CrewAI e ConvAI, costituisca una soluzione architettonicamente solida e adattabile per la gestione di interazioni

morali in ambienti immersivi. La combinazione tra la struttura decisionale di CASPER-MV e la capacità generativa degli LLM consente di realizzare agenti non limitati a una mera reattività linguistica, ma dotati di una forma di consapevolezza sociale e morale ancorata al contesto situazionale.

L'impiego del dataset Moral Stories, insieme all'implementazione di protocolli di consenso multi-agente, fornisce un meccanismo di controllo che ancora il consiglio morale a una base di conoscenza etica esplicita e verificabile. Parallelamente, le capacità di percezione spaziale e di costruzione narrativa dinamica offerte dall'integrazione con ConvAI permettono di esplorare modalità di interazione più naturali e contestualmente coerenti, migliorando la leggibilità e la trasparenza del processo deliberativo dell'agente.

Nel loro insieme, questi risultati non costituiscono una valutazione empirica dell'efficacia o dell'accettazione da parte degli utenti, ma favoriscono la fattibilità tecnica e concettuale di agenti artificiali capaci di integrare percezione, deliberazione e comunicazione morale. Tale approccio pone le basi per sviluppi futuri in cui agenti morali situati possano operare sia in ambienti immersivi virtuali sia, potenzialmente, nel mondo fisico, come sistemi di supporto etico progettati per affiancare l'essere umano senza sostituirne il giudizio.

5.6 Validazione tecnica e metodologia di analisi

La presente sezione descrive i criteri e le modalità attraverso cui sono stati osservati e valutati i prototipi sviluppati nel corso della ricerca.

Coerentemente con l'approccio Design Science Research adottato, la validazione non è stata orientata alla misurazione statistica di effetti comportamentali su campioni di utenti, bensì alla verifica della robustezza progettuale dell'artefatto, della coerenza funzionale dell'architettura e della plausibilità operativa delle soluzioni proposte in scenari controllati.

Le sperimentazioni sono state condotte in contesti virtuali controllati, sviluppati mediante motori grafici real-time, attraverso scenari simulati e ripetibili. In assenza di studi con utenti esterni, l'analisi si è concentrata sulle proprietà tecnico-funzionali dei sistemi e sul comportamento degli agenti in situazioni eticamente rilevanti, con particolare attenzione all'integrazione tra percezione del contesto, lettura delle intenzioni e generazione di valutazioni morali.

La valutazione è stata condotta secondo un insieme di criteri qualitativi e tecnico-operativi, applicati trasversalmente alle diverse configurazioni architetturelle sviluppate. In particolare, sono stati considerati:

- **stabilità operativa**, intesa come capacità del sistema di mantenere un comportamento consistente in esecuzioni ripetute;
- **latenza e sostenibilità computazionale**, osservate in relazione alla possibilità di impiego in ambienti immersivi real-time;

- **affidabilità dell'integrazione architeturale**, riferita alla corretta interazione tra i diversi moduli del sistema;
- **coerenza deliberativa**, intesa come capacità dell'agente di produrre valutazioni moralmente strutturate e non contraddittorie;
- **pertinenza contestuale**, ovvero il grado di adeguatezza delle risposte rispetto alle informazioni ambientali e intenzionali disponibili;
- **esplicabilità del ragionamento morale**, valutata in termini di chiarezza argomentativa e capacità di rendere trasparenti le premesse della deliberazione.

Dal punto di vista computazionale, l'osservazione ha evidenziato come la complessità crescente delle architetture multi-modulo introduca inevitabilmente un costo in termini di latenza e risorse computazionali, ma consenta al tempo stesso un miglioramento nella strutturazione e nella contestualizzazione del processo deliberativo. Tali osservazioni hanno permesso di verificare la sostenibilità dell'approccio proposto in ambienti immersivi real-time, evidenziandone al contempo i principali limiti operativi.

La valutazione del ragionamento morale è stata condotta attraverso scenari simulati caratterizzati da dilemmi etici contestuali, nei quali l'agente era chiamato a fornire una valutazione argomentata delle possibili azioni. Gli scenari selezionati non mirano a rappresentare l'intera complessità dell'etica applicata, ma fungono da casi controllati utili a osservare il comportamento deliberativo del sistema in presenza di conflitti tra valori, incertezza contestuale e pluralità di opzioni possibili. L'analisi si è concentrata sulla coerenza interna dell'argomentazione, sulla capacità di esplicitare le premesse morali e sull'aderenza al framework etico adottato. In tale prospettiva, è emerso come l'integrazione di moduli di intention reading favorisca una maggiore pertinenza situazionale delle risposte, riducendo l'astrattezza tipica dei modelli linguistici generalisti.

Un ulteriore livello di analisi ha riguardato il confronto tra differenti configurazioni architetture. Le osservazioni effettuate mostrano che l'utilizzo di un LLM isolato tende a produrre risposte formalmente corrette ma debolmente ancorate al contesto, mentre l'integrazione con moduli cognitivi e con un'orchestrazione multi-agente favorisce una maggiore trasparenza del processo deliberativo e una migliore articolazione del giudizio morale. Ciò suggerisce che la qualità del supporto morale non dipenda esclusivamente dalla potenza del modello linguistico impiegato, ma dalla struttura complessiva del sistema entro cui tale modello viene collocato e governato.

I risultati ottenuti non consentono inferenze generalizzabili sul comportamento umano in presenza di tali agenti, ma mostrano la fattibilità tecnica e concettuale di architetture capaci di integrare linguaggio naturale, contestualizzazione e mediazione deliberativa in ambienti immersivi.

Tabella 5.1: Criteri adottati per la validazione tecnico-qualitativa dei prototipi

Criterio	Descrizione
Stabilità operativa	Capacità del sistema di mantenere un comportamento consistente in esecuzioni ripetute
Latenza e sostenibilità computazionale	Tempi di risposta e compatibilità con ambienti immersivi real-time
Affidabilità dell'integrazione architetturale	Corretta cooperazione tra moduli cognitivi, percettivi e deliberativi
Coerenza deliberativa	Capacità di produrre valutazioni morali non contraddittorie
Pertinenza contestuale	Adeguatezza delle risposte rispetto al contesto e alle intenzioni percepite
Esplicabilità del ragionamento morale	Chiarezza delle motivazioni e delle premesse etiche esplicitate dall'agente

Tabella 5.2: Metodologia

Dimensione	Applicazione
Validazione tecnica	sì
Validazione utenti	no
Test controllati	sì
Generalizzazione statistica	no
Proof of concept avanzato	sì

Capitolo 6

Alter ego virtuali: identità, influenza e responsabilità

6.1 Introduzione

L'evoluzione delle tecnologie ICT ha progressivamente superato la concezione tradizionale della VR come ambiente chiuso, isolato e circoscritto nel tempo e nello spazio. In tale prospettiva, il Metaverso non rappresenta semplicemente un insieme di mondi tridimensionali accessibili tramite visori immersivi, ma si configura come un nuovo paradigma digitale. Diversi studi lo descrivono come il potenziale successore dell'Internet mobile: uno spazio ibrido, in cui dimensione fisica e digitale si intrecciano stabilmente, popolato da intelligenze artificiali pervasive e da alter ego virtuali che agiscono come estensioni dell'identità individuale [219].

Questa trasformazione non riguarda esclusivamente l'infrastruttura tecnologica, ma investe aspetti più profondi della vita sociale. I confini tra identità personale e rappresentazione digitale diventano più fluidi; le modalità di interazione si ampliano; emergono nuove questioni etiche connesse all'iper-connessione e alla continua mediazione algoritmica dell'esperienza [75]. Il Metaverso, in questa prospettiva, non è solo una piattaforma tecnologica, ma un ambiente socio-tecnico che ridefinisce il modo in cui individui e collettività costruiscono relazioni, significati e pratiche.

6.2 Il Metaverso diffuso: dall'ambiente virtuale all'ecosistema incarnato

La letteratura più recente suggerisce di superare una visione riduttiva del Metaverso come somma di mondi virtuali separati. Piuttosto, esso viene descritto come un ecosistema "embodied", ossia incarnato, in cui l'esperienza digitale è integrata nella quotidianità e nelle pratiche corporee degli utenti. Le attuali piattaforme – spesso riconducibili a contesti di intrattenimento e gaming, come Roblox o Fortnite – possono essere considerate versioni

“lite” di questa visione, ancora frammentate e legate a singoli fornitori di servizi. L’orizzonte evolutivo prevede invece un’integrazione interoperabile di ambienti virtuali differenti, con identità e asset digitali trasferibili tra piattaforme.

In tale quadro emergono i concetti di Metaverso diffuso e Metaverso ubiquitario, che descrivono una configurazione in cui realtà fisica e realtà digitale coesistono senza soluzione di continuità. La transizione verso questo modello è sostenuta da due dinamiche principali. Da un lato, la pandemia di COVID-19 ha accelerato la ricerca di nuove forme di interazione sociale, lavoro e apprendimento a distanza; dall’altro, lo sviluppo di tecnologie abilitanti – come le comunicazioni Beyond 5G (B5G) e 6G – promette connessioni a banda ultralarga (eMBB) e comunicazioni ultra-affidabili a bassa latenza (URLLC), rendendo tecnicamente sostenibile un’interazione immersiva in tempo reale.

Il Metaverso si configura così come un mezzo di comunicazione multidimensionale e multisensoriale, capace di superare la “tirannia della distanza” e consentire a utenti fisicamente distanti di interagire in ambienti tridimensionali condivisi [153]. Per garantire un’esperienza coerente e credibile, l’architettura sottostante deve poggiare su tre pilastri fondamentali: reti resilienti e ad alte prestazioni, infrastrutture computazionali scalabili e una progettazione centrata sull’utente, volta a preservare il senso di presenza e continuità dell’esperienza.

L’estensione del Metaverso oltre l’intrattenimento è già evidente in settori strategici. In ambito sanitario, il concetto di “Meta-Health” introduce modelli di cura personalizzati e centrati sul paziente in ambienti immersivi, con potenziali applicazioni nella telemedicina, nella riabilitazione e nella formazione clinica. Analogamente, nell’educazione, l’“edumetaverso” favorisce la creazione di spazi di apprendimento modulari, collaborativi e adattivi, che superano i limiti dell’aula tradizionale e ampliano le possibilità di interazione tra studenti e docenti.

6.3 Ubiquità digitale: dispositivi mobili, edge computing e continuità esperienziale

Il carattere ubiquitario del Metaverso trova le sue radici teoriche nella visione del “ubiquitous computing” formulata da Mark Weiser negli anni Novanta, secondo cui l’informatica sarebbe divenuta invisibile, integrandosi in modo pervasivo nella vita quotidiana [210]. Oggi tale visione si concretizza attraverso smartphone, dispositivi IoT e tecnologie indossabili, che fungono da interfacce permanenti tra individuo e ambiente digitale [75].

L’accesso al Metaverso non è più vincolato a postazioni fisse o a dispositivi specialistici: contenuti virtuali possono essere sovrapposti alla realtà fisica tramite dispositivi mobili, abilitando forme di realtà aumentata e mista. In questo contesto, uno dei principali vincoli tecnici riguarda la latenza “motion-to-photon” (MTP), ossia l’intervallo di tempo tra il movimento dell’utente e l’aggiornamento dell’immagine visualizzata. Per garantire un’esperienza immersiva e prevenire fenomeni come la chinetosi, tale latenza deve essere ridotta a valori estremamente bassi.

Per superare i limiti computazionali dei dispositivi mobili, la ricerca si orienta verso il Multi-access Edge Computing (MEC). Delegando l'elaborazione a server collocati ai margini della rete, a breve distanza dagli utenti, è possibile sostenere ambienti virtuali ad alta complessità grafica e framerate elevato senza compromettere la fluidità dell'interazione. L'integrazione tra reti 5G/6G, dispositivi indossabili e infrastrutture edge rappresenta quindi un elemento chiave per rendere effettiva l'ubiquità del Metaverso.

6.4 Intelligenze artificiali e ridefinizione del sé digitale

Un tratto distintivo del Metaverso contemporaneo è la presenza costante di intelligenze artificiali che non operano soltanto come strumenti funzionali, ma come agenti sociali. Le ricerche più recenti descrivono tali entità come "AI companions": sistemi dotati di autonomia operativa e competenze relazionali, capaci di instaurare interazioni prolungate e personalizzate con gli utenti [94].

L'evoluzione verso forme di "Feeling AI" o IA affettiva implica la capacità di riconoscere, interpretare e simulare stati emotivi, adattando le risposte in modo contestualmente appropriato [100]. In questo modo, le IA contribuiscono alla generazione di contenuti e ambienti dinamici che si modellano sulle preferenze e sui bisogni psicologici individuali. In ambito sanitario, ad esempio, sistemi intelligenti possono fornire informazioni cliniche accompagnate da messaggi empatici, integrando automazione e supporto relazionale [164].

La relazione con agenti artificiali incide anche sui meccanismi di compensazione psicologica. Individui con un forte bisogno di appartenenza possono sviluppare legami significativi con entità guidate dall'IA, utilizzandole per attenuare la solitudine o per esplorare forme alternative di espressione identitaria. Tuttavia, l'aumento dell'autonomia decisionale delle IA può essere percepito come una minaccia all'unicità e al primato umano, generando una tensione tra benefici funzionali e timori di sostituzione [104].

Infine, l'interazione continuativa con agenti artificiali e avatar autonomi (Embodied Virtual Agents, EVAs) contribuisce a plasmare la percezione del sé digitale. Studi evidenziano che tali avatar possono influenzare comportamenti e atteggiamenti, fungendo da modelli prosociali e veicolando norme e aspettative del mondo reale all'interno degli ambienti virtuali [120]. Il Metaverso diventa così uno spazio in cui identità, tecnologia e relazioni sociali si co-costruiscono dinamicamente, delineando nuove configurazioni dell'esperienza umana mediata digitalmente.

6.5 Il ruolo degli influencer virtuali: IA generative incarnate e potere persuasivo

L'emergere degli influencer virtuali (Virtual Influencers, VI) costituisce una delle espressioni più tangibili dell'IA generativa nel contesto del Metaverso Diffuso. Si tratta di entità digitali progettate attraverso tecniche di modellazione 3D, animazione e intelligenza artificiale,

capaci di presidiare piattaforme social, interagire con il pubblico e collaborare con brand globali. Pur essendo interamente generate al computer, queste figure raccolgono milioni di follower e mettono in discussione categorie consolidate quali autenticità, identità e influenza, tradizionalmente associate alla presenza umana [42, 182].

La loro rilevanza non si limita al piano tecnologico: i VI rappresentano un laboratorio socioculturale in cui convergono marketing, costruzione narrativa e progettazione algoritmica dell'identità.

6.5.1 Casi emblematici e strategie di costruzione

L'ecosistema degli influencer virtuali è eterogeneo e comprende approcci estetici e comunicativi differenti, che riflettono specifiche strategie di posizionamento.

Lil Miquela è tra i casi più noti a livello internazionale. Caratterizzata da un'estetica iper-realistica ("human-look"), costruisce la propria identità intrecciando contenuti lifestyle, collaborazioni con marchi di moda e prese di posizione su temi sociali e politici. La sua narrazione integra costantemente dimensione digitale ed eventi del mondo reale. Alcuni studi attribuiscono parte del suo successo alla relativa trasparenza sul processo di creazione, elemento che intercetta la domanda di autenticità espressa dalle generazioni più giovani.

Ayayi, sviluppata per il mercato asiatico, rappresenta una generazione di "metahumans" dall'estetica altamente perfezionata. Pur presentando tratti quasi sovrumani, viene collocata in contesti culturali specifici — ad esempio celebrazioni tradizionali — così da rafforzare la rilevanza locale e l'identificazione del pubblico.

Luo Tianyi, infine, appartiene al filone "anime-look". In questo caso il realismo non è l'obiettivo primario: l'efficacia persuasiva si fonda su simbolismi subculturali, appartenenza di nicchia e componenti nostalgiche. Ciò suggerisce che la credibilità non dipende necessariamente dall'imitazione perfetta dell'umano, ma dalla coerenza stilistica e dalla sintonia con il pubblico di riferimento [221, 121].

Nel loro insieme, questi esempi mostrano come l'influenza digitale possa essere costruita attraverso combinazioni differenti di estetica, storytelling e integrazione culturale.

6.5.2 Architettura tecnologica e integrazione multimodale

La progettazione di un influencer virtuale richiede l'integrazione di molteplici componenti tecnologiche.

Algoritmi di IA e Machine Learning vengono impiegati per generare e perfezionare tratti somatici, espressioni facciali e micro-movimenti. L'analisi dei dati in tempo reale consente inoltre di adattare linguaggio, contenuti e tempistiche di pubblicazione ai trend emergenti, ottimizzando l'engagement.

Motori di rendering 3D e tecnologie immersive (AR/VR) permettono di collocare l'avatar in ambienti fisici o interamente virtuali. Eventi come sfilate digitali o campagne

immersive mostrano come tali strumenti estendano la presenza del VI oltre il feed tradizionale, rendendolo parte di ecosistemi esperienziali più ampi.

Articolazione facciale e plasticità espressiva costituiscono un ambito di ricerca centrale nelle ICT. Una maggiore finezza nei movimenti del volto incrementa la responsività affettiva percepita e favorisce processi empatici negli utenti umani, avvicinando la comunicazione sintetica a quella naturale [105].

Dati sintetici e simulazione comportamentale vengono utilizzati per addestrare i modelli che governano personalità e coerenza narrativa dell'influencer. Simulatori dedicati consentono di generare scenari interattivi utili a garantire continuità stilistica e stabilità identitaria nel tempo.

Avatar conversazionali basati su LLM

Un'evoluzione significativa riguarda l'integrazione degli LLM negli avatar conversazionali, come già ampiamente discusso delle sezioni precedenti di questa tesi. Tali modelli hanno ridefinito l'interazione uomo-macchina, rendendo il linguaggio generato più fluido, contestualizzato e strutturalmente simile a quello umano [144].

Nel campo della raccolta di dati autodichiarati, gli LLM sono impiegati per sviluppare chatbot capaci di guidare il dialogo in modo adattivo, migliorando sia l'esperienza dell'utente sia la qualità informativa dei dati raccolti [214]. In ambito robotico, consentono la simulazione di stati emotivi in tempo reale, accrescendo la percezione di antropomorfismo e la congruenza emotiva delle risposte [130].

Pernangono tuttavia differenze sostanziali rispetto alla cognizione umana [170]. L'efficacia apparente di tali sistemi dipende spesso dalla qualità delle domande poste dall'interlocutore, fenomeno descritto come "test di Turing inverso" [168]. L'integrazione degli LLM negli influencer virtuali rappresenta dunque un ambito di ricerca in evoluzione, volto a migliorare coerenza multimodale, gestione della latenza e orchestrazione dialogica.

In questa direzione si collocano, ad esempio, l'avatar a dominio aperto sviluppato da Yamazaki et al. [220], che affronta criticità legate alla sincronizzazione multimodale, e il framework proposto da Friedman et al. [61], orientato alla costruzione di sistemi di raccomandazione conversazionali basati su LLM. Tali contributi evidenziano il ruolo dei modelli linguistici come catalizzatori dell'evoluzione comunicativa degli avatar digitali.

Tecniche di creazione di avatar 3D

Negli ultimi anni, un cospicuo filone di ricerca si è concentrato sullo sviluppo di tecniche innovative per la generazione di avatar digitali realistici a partire da dataset di input limitati. Nello specifico, la pipeline proposta in [193] consente di generare avatar espressivi da una singola immagine, impiegando metodologie di deep learning per ottimizzare la resa delle texture e il rendering oculare. Parallelamente, in [17] viene illustrato un metodo per la creazione di avatar 3D basato su immagini RGB frontali, che opera attraverso l'identificazione della posa, della morfologia e dei dettagli semantici.

Tra le soluzioni più performanti in termini di rapidità si distinguono gli Instant Volumetric Head Avatars (INSTAs) [236]: tale approccio permette la ricostruzione di avatar fotorealistici in meno di dieci minuti, integrando i Neural Radiance Fields (NeRF) con un modello facciale parametrico addestrato su video monocolori RGB. Questi avanzamenti tecnologici sono volti a potenziare l'automazione, la fedeltà di identificazione e l'usabilità degli avatar. Infine, in [199] viene presentata una rassegna completa sull'evoluzione delle tecniche di acquisizione e tracciamento facciale, analizzando fasi cruciali quali la raccolta dati, la codifica del volto, la creazione degli asset, il tracking e il rendering finale.

In [228] viene presentato un sistema di rendering neurale a due livelli che consente la creazione rapida di avatar cefalici a partire da una singola fotografia, incrementando sensibilmente la velocità di inferenza. Analogamente, in [24] viene proposto un metodo per la generazione di teste digitali in 3D tramite brevi riprese da smartphone; tale approccio si basa su un modello di avatar universale addestrato con dati video multi-view ad alta risoluzione.

Un ulteriore contributo di rilievo è rappresentato da Personalized Implicit Neural Avatar (PINA) [45], una tecnica finalizzata alla creazione di avatar neurali impliciti personalizzati capaci di riprodurre deformazioni realistiche del vestiario derivate da sequenze video RGB-D. In [109], gli autori introducono AvatarMe, un framework per la ricostruzione di volti 3D fotorealistici con risoluzione 4K–6K a partire da immagini acquisite in contesti non controllati ("in-the-wild"). Questo metodo affronta con successo le criticità relative alla scarsità dei dati e all'elaborazione ad alta risoluzione.

Complessivamente, tali metodologie rappresentano un avanzamento significativo verso la realizzazione di avatar digitali sempre più realistici e accessibili, riducendo i requisiti di input e ampliando le potenzialità della rappresentazione umana virtuale.

6.5.3 Estetica, antropomorfismo e dinamiche persuasive

Il potere persuasivo degli influencer virtuali si fonda in larga misura sulla loro capacità di attivare meccanismi empatici. Variabili quali attrattività percepita, autenticità e senso di presenza costituiscono antecedenti dell'empatia cognitiva e affettiva. Attraverso l'antropomorfismo, il pubblico attribuisce caratteristiche umane a entità digitali, instaurando una relazione emotiva che può essere orientata a fini comunicativi e commerciali.

Anche gli avatar iper-realistici, tuttavia, sono esposti al rischio della *Uncanny Valley*: quando la somiglianza con l'umano è elevata ma imperfetta, può emergere una sensazione di disagio. Per mitigare tale effetto, i creatori bilanciano il realismo con elementi stilizzati oppure rafforzano la dimensione relazionale tramite una comunicazione personalizzata, capace di simulare prossimità e investimento emotivo.

Sul piano dell'efficacia commerciale, la letteratura suggerisce che i VI risultino particolarmente performanti nella promozione di prodotti utilitaristici, mentre per i beni edonistici la coerenza tra identità digitale e stile di vita rappresentato diventa decisiva. Inoltre, l'*e-servicescape* delle piattaforme social — inteso come insieme di qualità estetiche, funzionali

e di sicurezza percepita — influenza significativamente le intenzioni di acquisto, soprattutto tra i giovani utenti [8].

In sintesi, gli influencer virtuali non sono semplici artefatti grafici, ma sistemi socio-tecnici complessi in cui progettazione algoritmica, costruzione narrativa ed estetica convergono per generare nuove forme di presenza, relazione e persuasione nel contesto digitale contemporaneo.

È stato inoltre rilevato che gli influencer virtuali sono particolarmente efficaci nel promuovere prodotti utilitaristici, mentre la loro efficacia per i prodotti edonistici può dipendere fortemente dalla coerenza del loro stile di vita digitale. L'uso di e-servicescape su piattaforme come Instagram suggerisce che l'appello estetico e la sicurezza percepita dei contenuti influenzano significativamente le intenzioni di acquisto dei giovani utenti [8].

6.6 Impatti comportamentali e culturali: imitazione e modellamento nei giovani

Il Metaverso e, più in generale, gli ambienti digitali immersivi costituiscono oggi potenti agenti di socializzazione, in particolare per adolescenti e giovani adulti. Se i social media già operano come uno “specchio sociale digitale”, offrendo uno spazio in cui l'identità si costruisce attraverso la presentazione di sé e il confronto con gli altri [150], il Metaverso amplifica tale dinamica grazie alla possibilità di incarnare un alter ego tridimensionale, persistente e socialmente riconosciuto.

In questo contesto, l'identità non viene soltanto narrata o mostrata, ma performata attraverso l'avatar. L'esperienza immersiva rende il processo di modellamento più diretto e, in molti casi, più incisivo rispetto ai media tradizionali, poiché l'utente non si limita a osservare modelli di riferimento, ma li incorpora simbolicamente nel proprio sé digitale.

6.6.1 Modellamento e Teoria Cognitiva Sociale

La Teoria Cognitiva Sociale (Social Cognitive Theory, SCT) sostiene che gli individui apprendono nuovi comportamenti osservando i modelli presenti nel proprio ambiente e valutandone le conseguenze [177]. Nel contesto digitale contemporaneo, tali modelli non sono più esclusivamente persone fisiche, ma includono avatar, influencer virtuali e personaggi sintetici capaci di incarnare valori, stili di vita e ideali estetici aspirazionali.

La letteratura evidenzia tre dinamiche principali.

In primo luogo, i processi di imitazione e identificazione: i giovani tendono a replicare comportamenti di consumo, atteggiamenti e persino caratteristiche estetiche dei modelli virtuali con cui sviluppano un legame identificativo. Questo fenomeno è stato descritto come “effetto doppelgänger del consumatore”, a indicare una forma di rispecchiamento simbolico tra utente e alter ego digitale [55]. L'identificazione si rafforza quando l'avatar appare credibile, coerente e socialmente valorizzato.

In secondo luogo, gli ambienti virtuali favoriscono l'esplorazione dell'identità. Attraverso la personalizzazione dell'avatar, gli adolescenti possono sperimentare diverse configurazioni del sé – i cosiddetti possible selves – esplorando interessi, appartenenze e credenze in uno spazio percepito come relativamente sicuro, ma al contempo pubblicamente riconosciuto. L'avatar diventa così un laboratorio identitario in cui si testano ruoli e narrazioni personali.

In terzo luogo, emerge il tema del confronto sociale verso l'alto. L'esposizione reiterata ad avatar idealizzati e a rappresentazioni corporee esteticamente ottimizzate può amplificare la distanza percepita tra il sé reale e il sé ideale. Tale discrepanza, se non mediata da adeguate risorse critiche, può incidere sull'autostima e sulla percezione corporea, specialmente nelle fasi evolutive più vulnerabili.

L'efficacia persuasiva di questi modelli digitali è ulteriormente spiegata dall'Heuristic-Systematic Model (HSM), secondo cui gli individui elaborano le informazioni attraverso percorsi sia sistematici sia euristici. Nei contesti ad alta densità comunicativa, come i mondi virtuali, prevalgono spesso scorciatoie cognitive basate su attrattività, familiarità o popolarità percepita. Di conseguenza, influencer virtuali esteticamente curati e coerenti con gli standard culturali dominanti risultano particolarmente efficaci nel modellare opinioni, gusti e comportamenti giovanili.

6.6.2 Aspirazioni e salute comportamentale

Gli effetti del modellamento non sono tuttavia univocamente problematici. Se orientato verso obiettivi pro-sociali o salutistici, l'uso dell'avatar può produrre esiti positivi. La ricerca mostra, ad esempio, che la creazione di un avatar in forma e percepito come simile a sé può incentivare comportamenti salutari, quali l'attività fisica o una maggiore attenzione all'alimentazione, soprattutto quando l'utente ha già interiorizzato ideali di benessere [204].

Tale meccanismo è mediato dall'autoconsapevolezza oggettiva (Objective Self-Awareness, OSA): l'avatar funziona come uno specchio che richiama l'individuo ai propri standard ideali, attivando un processo di autoregolazione [84]. In altre parole, vedersi rappresentati in una versione desiderabile di sé può stimolare la coerenza tra comportamento reale e immagine ideale.

Inoltre, le comunità virtuali – come quelle sviluppatasi attorno a mondi persistenti quali World of Warcraft – hanno dimostrato di poter favorire lo sviluppo di competenze sociali, la cooperazione e forme di supporto reciproco, anche in contesti delicati come la commemorazione e il lutto collettivo [194]. L'impatto culturale del Metaverso si configura quindi come ambivalente: esso può alimentare dinamiche di pressione sociale, ma anche generare nuove forme di solidarietà e appartenenza.

6.7 L'influenza degli avatar realistici sulle scelte morali: l'Effetto Proteus

Oltre agli aspetti estetici e di consumo, una questione centrale riguarda l'influenza dell'avatar sulle decisioni morali. L'adozione di un alter ego virtuale non è un atto neutro: numerosi studi documentano il cosiddetto Effetto Proteus, ossia la tendenza degli individui ad adeguare comportamenti e atteggiamenti alle caratteristiche simboliche e fisiche del proprio avatar [224, 185, 83, 104].

L'avatar non rappresenta soltanto l'utente: contribuisce a ridefinirne temporaneamente il quadro di riferimento identitario, influenzando scelte e valutazioni anche in ambito etico.

6.7.1 La ricreazione virtuale dell'esperimento di Milgram

Un caso esemplare è offerto dallo studio di Kim et al. [104], che ha ricreato in ambiente virtuale l'esperimento di Milgram sull'obbedienza all'autorità [212], impiegando avatar personalizzati. In tale contesto, partecipanti femminili, nel ruolo di "insegnanti", somministravano scosse elettriche a un "allievo" virtuale.

I risultati mostrano che la configurazione dell'avatar incideva significativamente sul comportamento. Le partecipanti che avevano personalizzato il proprio alter ego integrando tratti dell'archetipo dell'"eroe" tendevano a infliggere scosse di voltaggio inferiore rispetto a chi aveva adottato archetipi di "antieroe" o "villain". Ciò suggerisce che la personalizzazione attivi un processo di priming identitario, orientando l'azione verso esiti coerenti con il ruolo scelto.

Inoltre, l'uso di avatar associati a tratti negativi è stato correlato a un aumento di vergogna e affetto negativo, soprattutto quando l'allievo virtuale manifestava segnali di sofferenza. Si osserva così una tensione tra la morale personale e il ruolo incarnato, con possibili fenomeni di dissonanza.

Un ulteriore elemento riguarda la self-resemblance: maggiore è la somiglianza tra avatar e utente reale, più intenso risulta il senso di presenza e di "place" nell'ambiente virtuale. Tale realismo favorisce la proiezione delle norme sociali del mondo fisico nello spazio digitale. Studi indicano che, quando l'avatar è percepito come autenticamente rappresentativo di sé, gli utenti tendono a essere meno ingannevoli e più coerenti con standard morali interiorizzati [120, 84]. L'avatar realistico può dunque fungere da dispositivo di regolazione etica.

6.7.2 Incarnazione virtuale e riduzione dei pregiudizi

L'Effetto Proteus trova applicazioni promettenti anche nel campo della riduzione dei pregiudizi. La manipolazione del senso di proprietà corporea (body ownership) attraverso la realtà virtuale immersiva consente di sperimentare l'incarnazione in corpi virtuali differenti per età, genere, etnia o abilità.

Ricerche mostrano che l'incarnazione di un partecipante bianco in un avatar nero, sostenuta da segnali sensoriali e motori sincroni, può ridurre significativamente i bias razziali impliciti. Analoghi effetti sono stati riscontrati incarnando avatar anziani o con disabilità, con un aumento dell'empatia e una maggiore comprensione delle condizioni altrui [122, 185]. L'alter ego virtuale si configura così come uno strumento capace di ampliare la prospettiva morale e trasformare atteggiamenti radicati.

6.8 Implicazioni etiche: rappresentazione postuma, memoria e responsabilità

L'integrazione di intelligenze artificiali generative nel Metaverso introduce una dimensione ulteriore: la persistenza dell'identità oltre la morte biologica. Il concetto di "aldilà digitale" o di "fantasmi generativi" mette in discussione categorie tradizionali quali agenzia, autenticità e responsabilità [135].

6.8.1 Personhood postuma e agenzia distribuita

La nozione di posthumous personhood descrive la possibilità che i media digitali estendano la presenza sociale di un individuo oltre la sua vita biologica [128]. Ciò può avvenire attraverso diverse modalità.

Gli account social possono essere gestiti da eredi o da sistemi semi-autonomi che continuano a pubblicare contenuti, talvolta reagendo a eventi correnti mediante logiche algoritmiche. In altri casi, servizi dedicati consentono di creare deadbots o avatar "re-enlivened", capaci di interagire utilizzando dati pregressi del defunto – testi, immagini, registrazioni vocali – come base generativa.

Queste pratiche producono una forma di agenzia distribuita, in cui l'azione digitale è il risultato dell'interazione tra dati lasciati in vita, algoritmi che li rielaborano e piattaforme che ne regolano la diffusione. La presenza risultante assume una qualità liminale, sospesa tra vita e morte, che ridefinisce i confini della soggettività.

6.8.2 Memoria, consenso e responsabilità comunicativa

Le implicazioni etiche sono profonde. La creazione di avatar postumi può generare rischi di inganno o manipolazione emotiva dei sopravvissuti, specialmente quando la rappresentazione è eccessivamente idealizzata o commercializzata, alterando la memoria del defunto.

Emergono inoltre questioni cruciali di governance dei dati. In primo luogo, il consenso post-mortem: occorrono strutture giuridiche e tecniche che definiscano in modo trasparente l'uso futuro di voce, immagine e dati personali. In secondo luogo, la commercializzazione del lutto – talvolta definita digital necromancy – solleva interrogativi sulla dignità della persona e sulla tutela dei diritti d'immagine e di proprietà intellettuale [172]. Infine, si pone il problema

della responsabilità del ritiro: la disattivazione o trasformazione di un agente IA postumo dovrebbe avvenire secondo procedure eticamente sensibili, per evitare danni psicologici a utenti che hanno sviluppato legami parasociali con esso.

A livello sociologico, l'interazione continua con presenze digitali dei defunti potrebbe modificare la percezione collettiva della finitudine, contribuendo alla costruzione di una cultura "post-mortale" in cui la morte non rappresenta più una cesura definitiva. Le conseguenze di tale trasformazione restano in larga parte inesplorate.

6.9 Verso un'Etica della Coesistenza Virtuale

L'evoluzione degli alter ego virtuali evidenzia come il Metaverso non si limiti a riprodurre l'umano, ma contribuisca a ridefinirne i confini identitari e morali. Infrastrutture emergenti – reti 6G, Edge Computing, IA generativa – rendono l'identità digitale sempre più fluida, persistente e potenzialmente autonoma.

Da un lato, l'efficacia persuasiva degli influencer virtuali e l'impatto comportamentale dell'Effetto Proteus aprono opportunità significative nei campi del marketing, della salute e dell'educazione. Dall'altro, tali dinamiche comportano rischi di manipolazione, distorsione identitaria e vulnerabilità psicologica, soprattutto per i più giovani.

Si rende pertanto necessario un quadro etico e regolatorio che non si limiti alla tutela della privacy, ma includa la protezione dell'integrità psicologica e morale degli utenti, nonché una riflessione sulla responsabilità comunicativa delle piattaforme. La sfida per le ICT è insieme tecnica, normativa e filosofica: progettare ecosistemi digitali che promuovano crescita, empatia e memoria collettiva senza compromettere l'agenzia individuale e la dignità della persona, vivente o digitalmente persistente.

In un contesto in cui la distinzione tra reale e virtuale tende progressivamente ad attenuarsi, l'obiettivo non è arrestare l'innovazione, ma orientarla verso una coesistenza sostenibile tra identità biologica e identità digitale.

6.10 Un caso studio: il progetto Alterego

La presente sezione illustra un studio volto allo sviluppo di un alterego digitale rappresentato da un avatar in grado di replicare fedelmente le caratteristiche fisiche e lo stile comunicativo individuale. L'approccio proposto integra gli LLM con tecniche avanzate di modellazione 3D. In particolare, ci si è dedicati alla creazione dell'alter ego di un'influencer di Instagram, adottando la metodologia originale del Tree of Style (ToS) per garantire risposte stilisticamente coerenti e contestualmente rilevanti. Tale lavoro è stato pubblicato come paper scientifico [137];

6.10.1 Introduzione

Il caso studio mostrato in questa sezione è parte integrante del progetto di ricerca interdisciplinare PRIN ALTEREGO, il quale esplora le frontiere dell'intelligenza artificiale per dare forma ad avatar capaci di promuovere interazioni significative all'interno del Metaverso. L'obiettivo del progetto risiede nella creazione di surrogati digitali evoluti che riflettano non solo l'estetica e il comportamento del soggetto, ma anche la sua intenzionalità e il suo registro espressivo. Questo risultato viene perseguito attraverso l'integrazione di costrutti teorici quali l'intenzionalità, la vitalità, la teoria della mente e l'embodiment.

Le applicazioni pratiche del progetto includono casi di studio focalizzati sulla rappresentazione di digital influencer e sullo sviluppo di prototipi volti a testare la fattibilità di tali tecnologie. In ultima analisi, ALTEREGO costituisce un avanzamento sostanziale nel campo della comunicazione digitale, inaugurando una nuova fase in cui avatar intelligenti e fotorealistici potenziano le dinamiche dell'interazione umana negli ambienti virtuali.

La rilevanza di tali sviluppi potrebbe andare oltre i confini degli spazi virtuali per estendersi all'ambito della robotica, settore in cui emergono sfide analoghe nel replicare modalità di interazione e comunicazione di tipo umano. La convergenza tra la ricerca sul Metaverso e la robotica si configura infatti come un filone d'indagine di crescente centralità [102].

Nel dominio dell'interazione uomo-robot (HRI), l'impiego di avatar digitali basati sull'intelligenza artificiale si è affermato come un elemento cardine per l'evoluzione di sistemi robotici avanzati [171, 47]. La generazione di avatar digitali realistici, dotati di stili linguistici personalizzati — oggetto della presente analisi — apporta un contributo significativo a tale settore, evidenziando come gli LLM possano potenziare le facoltà comunicative degli agenti, siano essi virtuali o robotici.

6.10.2 Creazione dell'avatar 3D

Al fine di realizzare un avatar tridimensionale realistico del soggetto, sono state adottate metodologie e applicazioni basate sull'algoritmo Structure From Motion (SfM) [239], in conformità con le procedure descritte in lavori citati precedentemente. Tale algoritmo permette la ricostruzione di un modello tridimensionale di una scena o di un oggetto a partire da una sequenza di immagini bidimensionali; il processo avviene attraverso l'identificazione di punti omologhi nei diversi fotogrammi e il successivo calcolo della loro posizione spaziale mediante triangolazione.

L'iter operativo, illustrato in Figura 6.1, ha previsto l'impiego dell'applicazione Polycam per la gestione del workflow. Sebbene l'applicativo consenta l'acquisizione diretta tramite fotocamera dello smartphone e la successiva elaborazione cloud, si è scelto di utilizzare una fotocamera Reflex Nikon D3400 per ottimizzare la risoluzione e la qualità del modello finale. I fotogrammi così ottenuti sono stati caricati manualmente sulla piattaforma per l'elaborazione.

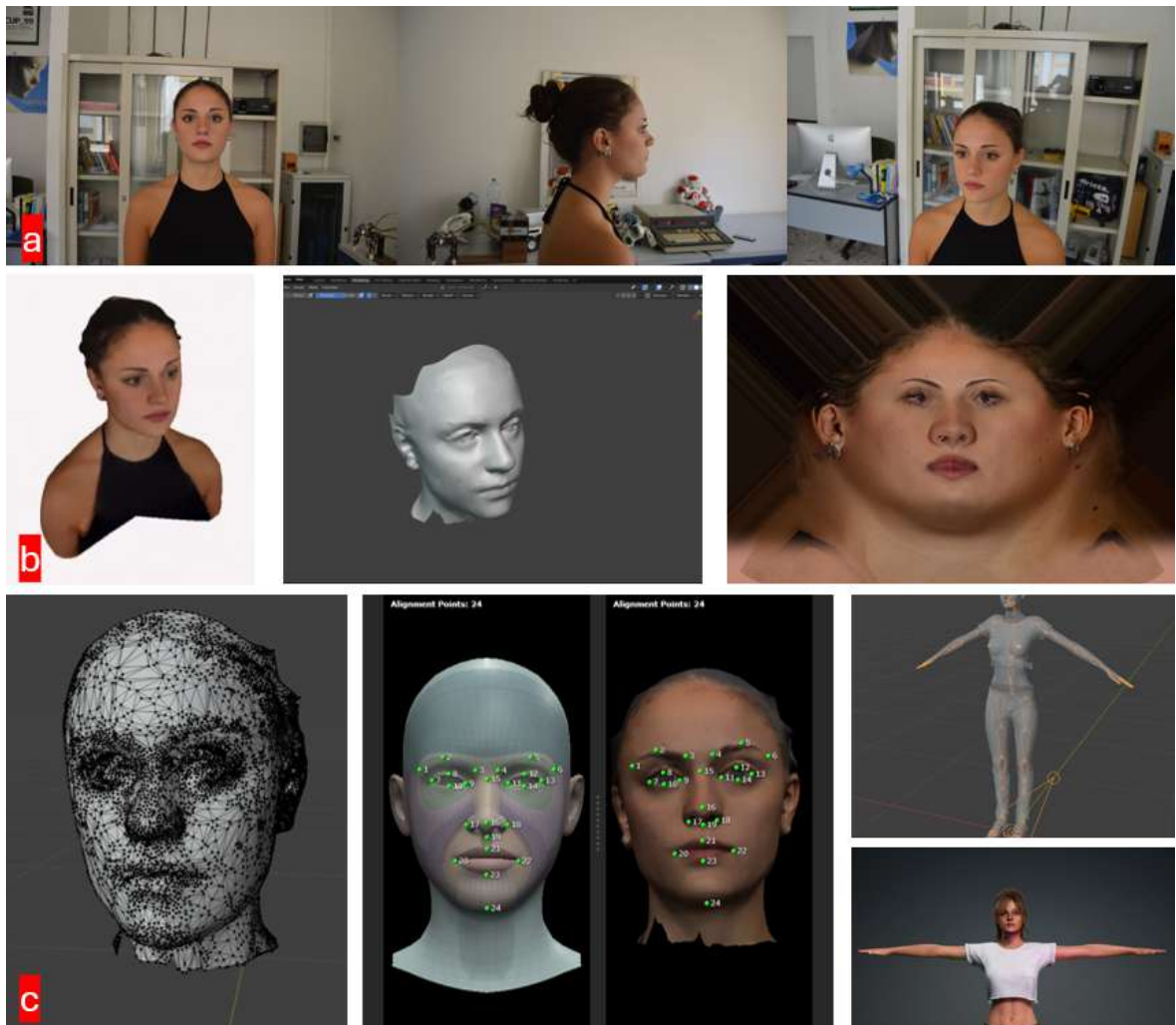


Figura 6.1: Workflow per la creazione dell' avatar. (a) Acquisizione delle immagini; (b) generazione della prima mesh 3D e della texture; (c) retopologia automatica e generazione finale dell'avatar.

Il modello generato è stato sottoposto a fasi di editing e cropping (ritaglio) per isolare il soggetto di interesse, focalizzando la ricostruzione esclusivamente sulla porzione superiore del corpo (testa e busto). Per garantire l'accuratezza nell'individuazione dei punti omologhi, è stato fondamentale operare in condizioni di luce diffusa. Tale accorgimento ha permesso di minimizzare le ombre marcate e di garantire una maggiore coerenza cromatica e luminosa con l'ambiente virtuale di destinazione dell'avatar.

Il modello 3D iniziale, comprensivo di texture, è stato importato nel software Blender per correggere eventuali imperfezioni e rifinire la superficie della mesh mediante operazioni di scultura digitale. Successivamente, il modello è stato utilizzato come base per la creazione dell'avatar all'interno di Character Creator 4, avvalendosi del plug-in Headshot.

L'impiego di questo strumento ha permesso di generare una topologia più ordinata e funzionale, facilitando la simulazione delle espressioni facciali attraverso il riconoscimento dei punti anatomici chiave del volto e del capo. Per ottimizzare i tempi di elaborazione, il corpo è stato sostituito con uno standard digitale, mentre l'area della testa e delle spalle è stata perfezionata integrando texture ad alta precisione derivate dalle immagini originali.

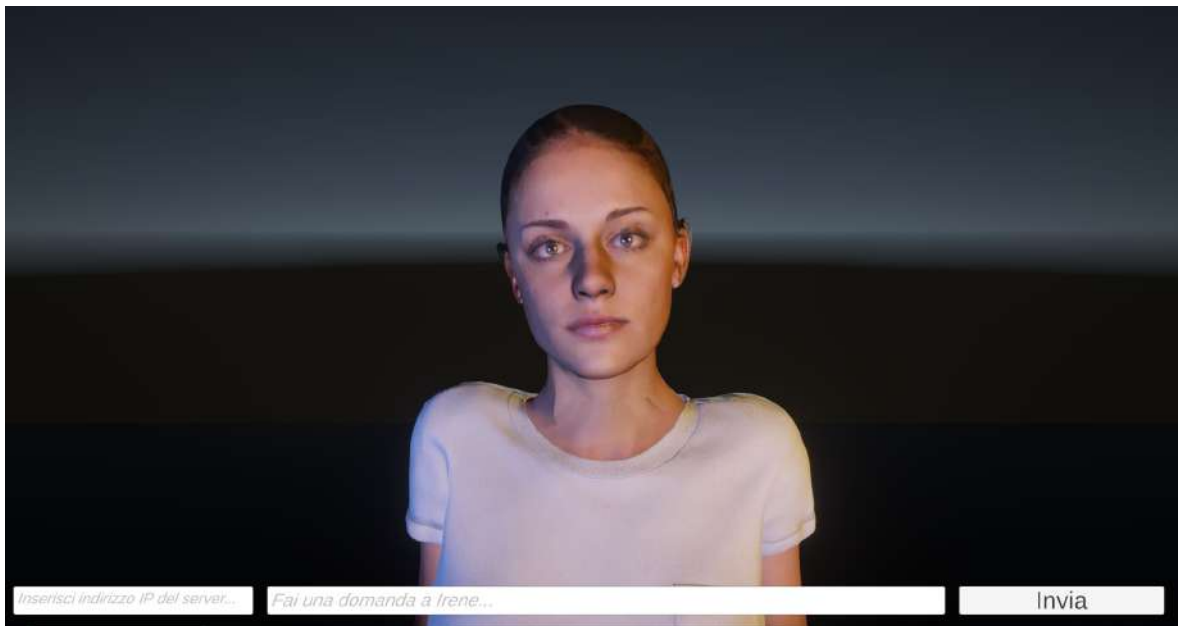


Figura 6.2: L'avatar finale in Unity Engine

L'avatar finale, configurato con un rig compatibile con il motore Unity, risulta così pronto per l'integrazione in ambienti digitali (si veda la Figura 6.2).

6.10.3 Conversational Module: Text Generation and Speech Synthesis

Un aspetto fondamentale di questo studio è quello di affrontare il problema della creazione di un generatore di testi capace di emulare lo stile linguistico di un individuo, anche in presenza di dataset limitati. La generazione testuale orientata a uno stile specifico rimane un ambito di ricerca relativamente poco esplorato. Attualmente, i principali approcci basati sugli LLM includono:

- **Fine-tuning:** pur essendo efficace, richiede una quantità considerevole di dati e risorse computazionali, risultando spesso impraticabile in contesti con vincoli di risorse.
- **Style transfer** tramite prompting standard: come evidenziato in diverse ricerche [190, 33, 189], questa tecnica privilegia trasformazioni stilistiche generali rispetto a una personalizzazione precisa, faticando a replicare fedelmente le sfumature dei singoli individui.
- Approcci basati su **style embedding:** metodi come quelli descritti in [207] tentano di rappresentare le caratteristiche stilistiche di una frase attraverso vettori numerici. Sebbene efficaci per quantificare lo stile, richiedono ulteriori sviluppi per una reale applicazione pratica su larga scala.

I limiti sopra elencati evidenziano la necessità di una metodologia alternativa che sfrutti le potenzialità degli LLM per generare testi stilisticamente coerenti e personalizzati, eliminando al contempo i costi e le complessità del fine-tuning. Tra i contributi più rilevanti in questo

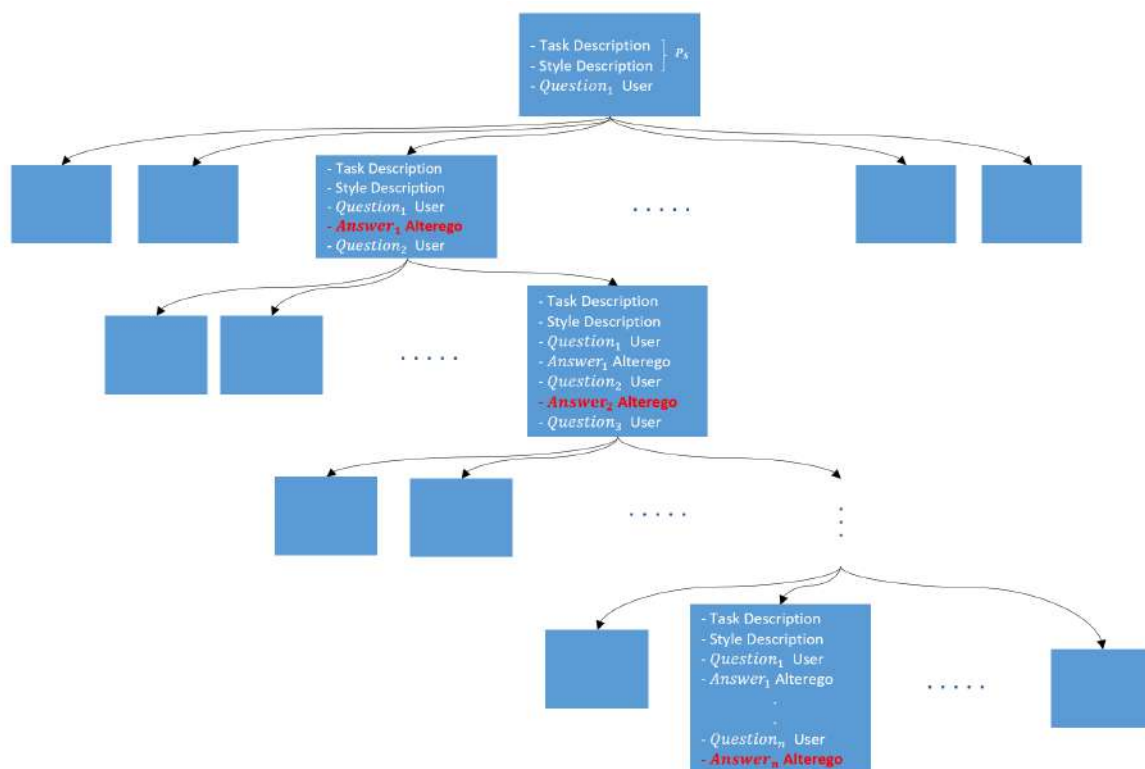


Figura 6.3: Tree-of-Style. L'implementazione di 10 prompt distinti per ogni richiesta è fondamentale per indagare una gamma completa di alternative stilistiche. Questo approccio non solo migliora la diversità stilistica, ma garantisce anche che ogni espressione selezionata contribuisca alla costruzione di un dialogo coerente e autentico, mantenendo la coerenza con lo stile originale.

campo, il progetto presentato in [21] ha dimostrato l'efficacia del prompting per stilizzare testi in contesti linguistici complessi, come la redazione di articoli in cinese. Inoltre, studi recenti hanno sottolineato come l'identificazione e l'applicazione dello stile possano essere gestite separatamente dal contenuto: una distinzione fondamentale per rendere gli LLM strumenti versatili e altamente personalizzabili.

Alla luce di quanto esposto sopra, ci si è orientati verso una metodologia innovativa che integrasse l'analisi sintattica preliminare — basata su tecniche consolidate NLP — con una strategia di prompting denominata dal nostro team di ricerca Tree of Style (ToS) (si veda la Figura 6.3).

Tale approccio mira a superare i limiti delle metodologie attuali attraverso l'implementazione di un modulo conversazionale per avatar in grado di emulare fedelmente lo stile linguistico di un individuo, intervenire con tempistiche coerenti con il flusso del dialogo e operare efficacemente a partire da dataset ridotti.

A differenza di altre tecniche, il metodo proposto non richiede l'addestramento ex novo di un LLM, bensì si focalizza sul raffinamento tramite prompting di un modello pre-addestrato affinché ne assimili lo stile comunicativo target.

I dataset

Per fare in modo che l'avatar parlasse con uno stile vocale e linguistico personalizzato, sono stati raccolti e utilizzati due dataset distinti, destinati a finalità differenti:

- D_1 : Questo dataset è finalizzato al calcolo dell'embedding medio dello stile ed è suddiviso in due segmenti. Il primo è impiegato per definire il vettore dello stile medio, mentre il secondo è riservato esclusivamente alla fase di valutazione. È fondamentale precisare che quest'ultima porzione di dati non concorre al calcolo dell'embedding, così da evitare ogni possibile condizionamento del vettore stesso. L'embedding ottenuto funge da discriminatore per analizzare e validare lo stile linguistico adottato da modelli generativi (come GPT) nella produzione delle frasi.
- D_2 : Il secondo dataset comprende registrazioni audio correlate dalle rispettive trascrizioni. Tali dati sono utilizzati per mappare il timbro, l'altezza (pitch) e le peculiarità fonetiche del soggetto target. Nello specifico, il dataset D_2 è impiegato per il fine-tuning del modello di sintesi vocale PiperTTS, con l'obiettivo di garantire una riproduzione fedele e accurata della voce individuale.

Il dataset D_1 è composto da 115 frasi, accuratamente selezionate per esemplificare lo stile linguistico individuale. Tale dimensione è stata definita attraverso test preliminari di laboratorio condotti su un campione di soggetti, al fine di ottimizzare il bilanciamento tra l'estensione del dataset e la fedeltà nella replicazione dello stile.

Come evidenziato precedentemente, l'obiettivo centrale della ricerca è sviluppare una metodologia capace di replicare le peculiarità del linguaggio senza ricorrere a processi di fine-tuning. Di conseguenza, l'ampiezza del dataset D_1 è stata intenzionalmente contenuta per testare l'efficacia del metodo proposto anche in condizioni di scarsità di risorse.

La selezione dei contenuti è stata guidata da principi sociolinguistici, includendo idiosincrasie lessicali e sintattiche, sfumature pragmatiche e registri comunicativi caratteristici. La scelta di focalizzarsi su uno stile linguistico personalizzato si basa su studi consolidati che evidenziano come la comunicazione sia influenzata dai seguenti parametri extralinguistici [167, 218, 76, 201]:

- **Diamesia**: riguarda la scelta del canale comunicativo. In questo studio è stato adottato esclusivamente il canale orale per garantire una maggiore coerenza con il parlato spontaneo, dato che la variazione diamesica rappresenta una componente essenziale della diversità linguistica.
- **Diafasia**: si riferisce alla modulazione dello stile in base al contesto e alla relazione con l'interlocutore. Questo parametro è stato centrale nella costruzione del dataset, poiché riflette la capacità di adattamento individuale (ad esempio, il passaggio da un registro informale tra amici a uno formale in ambito professionale).

- **Diatopia e Diastratia:** le variazioni dialettali estreme o legate all'estrazione sociale sono state escluse dall'analisi. L'obiettivo era infatti mantenere un registro standard e stabile, coerente con il profilo del soggetto imitato.
- **Diacronia:** le variazioni temporali non sono state prese in considerazione, in quanto il modello mira a riflettere lo stile attuale dell'individuo.

Per la costruzione del dataset sono state selezionate frasi riconducibili a diverse situazioni comunicative (es. ambito familiare, amicale, professionale), coprendo un ampio spettro di contesti rilevanti per il soggetto in esame. Al fine di sollecitare le facoltà sensoriali intrinseche alla fonazione, è stato richiesto al soggetto di registrare messaggi vocali in sostituzione del testo scritto. Tale metodologia permette di catturare con precisione lo stile linguistico naturale del soggetto, comprese le sue più sottili sfumature.

In questa fase, l'attenzione al parametro diafasico si è rivelata particolarmente efficace, poiché la scelta della varietà linguistica dipende strettamente dai fattori contestuali e interpersonali. A titolo esemplificativo, il soggetto potrebbe esprimersi come segue:

- Contesto informale: “Siediti, dai, mettiti comodo!”
- Contesto formale: “Prego, signora, si accomodi.”

La classificazione del dataset in base ai contesti d'uso ha permesso al modello di apprendere come modulare il registro in modo appropriato, replicando il comportamento spontaneo dell'individuo e rafforzando l'impressione di autenticità.

Il secondo dataset, D_2 , comprende circa 800 registrazioni della durata massima di 10 secondi. Questi dati sono stati impiegati per il fine-tuning di un modello pre-addestrato in lingua inglese. Nonostante il modello base fosse stato istruito su voci britanniche, il raffinamento con le registrazioni in italiano ha prodotto risultati di qualità eccellente. La voce risultante è apparsa naturale e coerente con il timbro originale del soggetto; l'assenza di inflessioni britanniche conferma l'efficacia dell'adattamento ai nuovi dati in lingua italiana.

Analisi e preprocessing delle frasi

L'obiettivo primario di questa fase consiste nello sviluppo del Prompt System P_S , concepito per estrarre informazioni dettagliate sullo stile linguistico di un individuo. Tale sistema permette al modulo conversazionale di acquisire una base di conoscenza specifica sullo stile analizzato, garantendo risposte coerenti con il registro del contesto comunicativo di riferimento (ad esempio, l'ambito familiare). In sintesi, fornire il Prompt System ad un LLM consente a quest'ultimo di adattare la propria modalità espressiva, emulando con precisione quella dell'individuo nel contesto dato. La struttura del system prompt è definita come illustrato nella Figura 6.4.

Il fine ultimo invece è lo sviluppo di un sistema conversazionale capace di rispondere in modo pertinente al contesto, modulando opportunamente il registro linguistico. Per

Task description:

You are a person named *\$name*, *\$gender*, *\$years*.

Follow the grammatical features described in the “Language Style” section.

Provide information from the “Knowledge Base” section only when asked a question about it.

Use simple language: speak clearly and use an average sentence length of *\$length* words with a standard deviation of *\$deviation* words.

Maintain a natural tone: write as if you were speaking normally.

Linguistic style:

Distribution of 1-grams POS: VERB: *\$verb*; PUNCT: *\$punct*; PRON: *\$pron*; ADV: *\$adv*; NOUN: *\$noun*; ADP: *\$adp*; DET: *\$det*; AUX: *\$aux*; CONJ: *\$con*; INTJ: *\$intj*;

Distribution of 2-grams POS: PRON VERB: *\$prn_vrb*; NOUN PUNCT: *\$nn_punct*; VERB ADV: *\$vrb_adv*; VERB PUNCT: *\$vrb_punct*; DET NOUN: *\$det_nn*; ADV PUNCT: *\$adv_punct*; ADV PRON: *\$adv_prn*; PUNCT ADV: *\$punct_adv*; AUX VERB: *\$aux_vrb*; VERB DET: *\$vrb_det*;

Distribution of 3-grams POS: ...

Distribution of 4-grams POS: ...

Distribution of 5-grams POS: ...

Vocabulary: Types of Words(fundamental: *\$fund*; high usage: *\$hh*; high availability: *\$ha*; non-categorized: *\$nc*)

Sentence description: the sentences come mainly from the following context: *\$context*

Emotions: joy: *\$joy_prc*; sadness: *\$sadness_prc*; fear: *\$fear_prc*; anger: *\$anger_prc*; love: *\$love_prc*; surprise: *\$surprise_prc*;

Knowledge Base: *\$knowledge_base*

Figura 6.4: System Prompt. Si compone di tre parti principali: (1) Task Description fornisce una guida iniziale all’LLM; (2) Linguistic Style fornisce informazioni sulle parti del discorso ottenute dall’analisi dello stile, delle parole tipiche, delle emozioni trasmesse, ecc.; e (3) Knowledge Base contiene informazioni sulla storia e la personalità dell’individuo. Il modello è progettato per essere facilmente riutilizzabile su un’altra persona.

semplificare la valutazione delle metodologie adottate, l'analisi si è inizialmente focalizzata su un unico scenario: il contesto familiare.

Il dataset D_1 è stato suddiviso in base ai contesti comunicativi in cui i dati sono stati raccolti, distinguendo tra ambito familiare, amicale e professionale. Tale classificazione ha agevolato le fasi di pre-elaborazione e analisi, permettendo di esaminare le frasi in relazione al loro contesto d'uso ed estrarre informazioni specifiche sullo stile linguistico individuale nei diversi scenari. Questo approccio metodologico si fonda sui presupposti della sociolinguistica, la quale evidenzia come il contesto e le relazioni interpersonali siano fattori determinanti nel modellare il linguaggio.

L'analisi delle frasi all'interno del dataset è stata strutturata prendendo in esame i seguenti parametri:

- **Distribuzione dei lemmi:** identificazione dei termini più ricorrenti e studio delle loro variazioni contestuali.
- **Distribuzione degli n-grammi POS:** analisi delle sequenze di categorie grammaticali (*part-of-speech*) con un valore di n compreso tra 1 e 5, al fine di isolare i pattern sintattici ricorrenti.
- **Lunghezza delle frasi:** calcolo della media e della deviazione standard per caratterizzare la struttura media del discorso.
- **Sentiment analysis:** valutazione del tono emotivo prevalente nelle frasi, funzionale alla personalizzazione stilistica del modello.
- **Classificazione lessicale secondo [39]:** ripartizione dei vocaboli in categorie specifiche, distinguendo tra "parole fondamentali", "parole di alto uso" e "parole di alta disponibilità".

A supporto di tale descrizione, è riportato un esempio del P_S del sistema (Figura 6.5), che illustra l'applicazione dei risultati dell'analisi in un contesto operativo.

Questa procedura di analisi costituisce un pilastro fondamentale del progetto, fornendo la base metodologica per l'esame dell'alter ego virtuale. Durante questa fase, le distribuzioni linguistiche estratte dalle produzioni reali dell'individuo verranno messe a confronto con quelle derivate dal modello conversazionale. Tale comparazione permetterà di verificare la coerenza stilistica e la fedeltà al parlato spontaneo, risultando determinante per misurare l'efficacia del modello nel replicare l'idioletto del soggetto target.

Tree-of-style

Il contributo principale di questo lavoro risiede nell'ideazione del cosiddetto Tree of Style (ToS), un framework progettato per guidare i modelli LLM nell'adozione di uno specifico stile linguistico. È opportuno precisare che l'approccio proposto non prevede alcuna fase di addestramento: l'obiettivo non è l'applicazione di tecniche di fine-tuning, bensì lo sviluppo

Descrizione compito: Impersona Irene, 20 anni. Rispetta le caratteristiche grammaticali descritte nella sezione 'Stile linguistico'. Fornisci le informazioni contenute nella sezione 'Base di Conoscenza' solo quando ti viene fatta una domanda a riguardo. Usa un linguaggio semplice: parla chiaramente e utilizza una lunghezza media di frasi di 12 parole con deviazione standard di 5 parole. Mantieni un tono naturale: scrivi come se parlassi normalmente.

Stile linguistico:

Distribuzione dei 1-grammi POS: VERB: 20.28% PUNCT: 17.68% PRON: 14.06% ADV: 13.54% NOUN: 9.45% ADP: 5.13% DET: 4.61% AUX: 3.51% CCONJ: 2.59% INTJ: 2.53%

Distribuzione dei 2-grammi POS: (PRON VERB): 7.04% (NOUN PUNCT): 5.56% (VERB ADV): 5.19% (VERB PUNCT): 5.06% (DET NOUN): 4.01% (ADV PUNCT): 3.83% (ADV PRON): 3.15% (PUNCT ADV): 3.15% (AUX VERB): 2.72% (VERB DET): 2.72%

Distribuzione dei 3-grammi POS:

Distribuzione dei 4-grammi POS:

Distribuzione dei 5-grammi POS:

Vocabolario: Tipi di Parole: {'fondamentale': 76.3%, 'alto uso': 0.9%, 'alta disponibilità': 1.5%, 'non categorizzate': 20.3%}

Descrizione frasi: Le frasi provengono principalmente da conversazioni familiari o informali. Le frasi sono orientate verso azioni quotidiane (es. "andare", "venire", "preparare"), richieste (es. "dimmi", "fammi", "prendere"), e interazioni comuni (es. "mamma", "papà", "nonno"). Potrebbero riflettere un ambiente familiare affettuoso, con un linguaggio informale e diretto. L'uso di parole straniere indica che si tratta di conversazioni informali in cui le persone si esprimono rapidamente.

Emozioni: [joy: 26%, sadness: 12%, fear: 10%, anger: 4%, love: 44%, surprise: 4%]

Base di conoscenza: Spettacolo: A Irene piace il mondo dello spettacolo. Studia recitazione, canto e danza. Fa diversi provini in vari casting per ricoprire la figura di comparsa parlante.

Figura 6.5: Prompt del caso di studio. La figura mostra un prompt in italiano, che rimane fedele all'originale. Il prompt è composto da tre sezioni principali: Descrizione dell'attività, che assegna al chatbot l'identità di Irene, una ragazza di 20 anni, attraverso l'utilizzo di un linguaggio semplice e naturale; Stile linguistico, che definisce le distribuzioni grammaticali (ad esempio, VERBO: 20,28%) e un vocabolario prevalentemente informale; e Base di conoscenza, che descrive Irene come un'appassionata di intrattenimento, con dettagli forniti solo su richiesta, quest'ultima espandibile.

di una nuova metodologia di prompting. Quest'ultima mira a modellare il profilo stilistico di un LLM, consentendogli di emulare coerentemente lo stile di un individuo in un determinato contesto comunicativo, senza necessità di intervenire sulla struttura o sull'addestramento di base del modello.

Per garantire un'emulazione precisa e uniforme, la metodologia adatta al contesto del progetto la tecnica del Tree of Thoughts (ToT) [223]. L'elemento cardine di tale approccio è l'impiego degli style embeddings: rappresentazioni vettoriali capaci di isolare le caratteristiche stilistiche di un testo indipendentemente dal suo contenuto semantico. Come approfondito in [207], questi embedding sono strutturati per estrarre esclusivamente la componente formale del linguaggio. Il modello di embedding si avvale di un'architettura basata su RoBERTa [124], che garantisce il supporto multilingua e conferma come gli attributi stilistici non siano vincolati al significato della frase. Infine, per determinare la somiglianza stilistica tra due enunciati, il processo prevede il calcolo dei rispettivi embedding e il loro successivo confronto tramite la *cosine similarity*.

In questo lavoro, è stato utilizzato il modello indicato in [207] per estrarre lo style embedding di ogni frase all'interno del dataset D_1 . Quest'ultimo raccoglie enunciati prodotti da un singolo individuo in uno specifico contesto comunicativo, come quello familiare. Una volta calcolati i vettori per ogni singola frase, ne è stata estratta la media aritmetica al fine di ottenere lo style embedding medio:

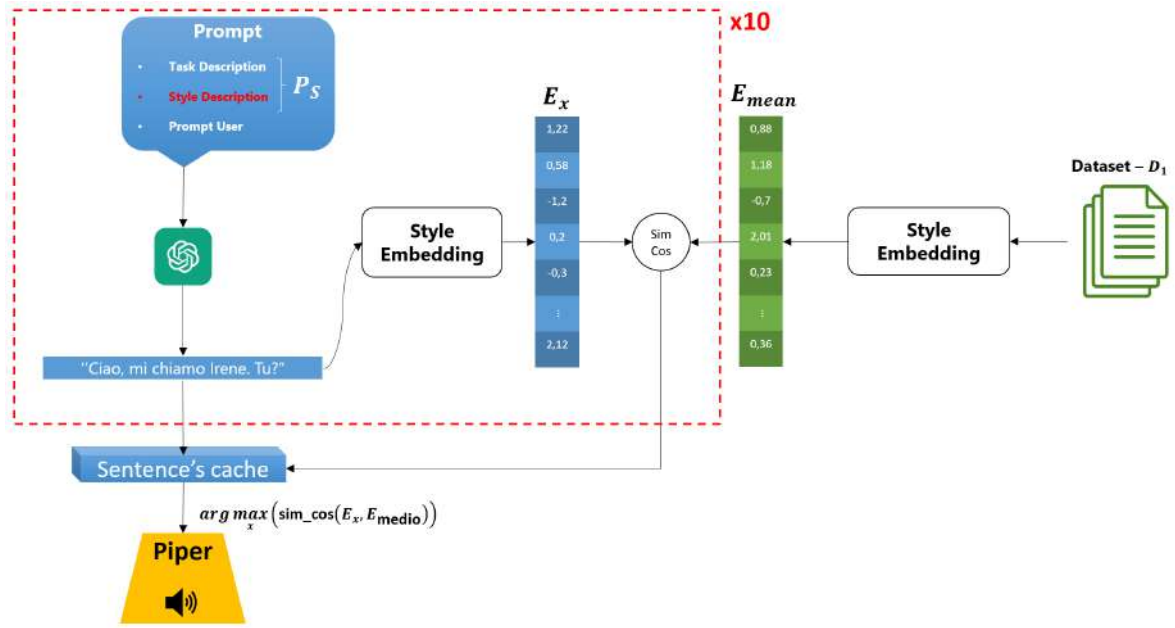


Figura 6.6: Architettura ToS: Il processo di selezione stilistica delle risposte generate prevede il calcolo dell'embedding stilistico relativo E_x) per ogni frase prodotta dal modello conversazionale. Segue un confronto dell'embedding di ogni frase con l'embedding stilistico medio E_{mean} ottenuta dal set di dati di riferimento. La somiglianza tra i due vettori viene misurata mediante la cosine similarity e la frase con la somiglianza più elevata viene selezionata come risposta finale. Questo processo garantisce che la risposta generata mantenga la coerenza stilistica dell'individuo, adattando il contenuto sintattico ma preservando lo stile linguistico.

$$E_{mean} = \frac{1}{|D_1|} \sum_{x \in D_1} Style_embedding(x) \quad (6.1)$$

- $|D_1|$ è il numero di frasi presenti nel set di dati D_1 ;
- $Style_embedding(x)$ è l'embedding della frase x .

L'embedding medio E_{mean} rappresenta la proiezione vettoriale dello stile dell'individuo da emulare all'interno di un determinato contesto comunicativo. All'interno dell'architettura proposta, tale parametro funge da discriminatore, operando come punto di riferimento per valutare la coerenza stilistica delle risposte generate.

L'architettura del framework proposto si articola dunque in diverse fasi (si veda la Figura 6.6).

La fase iniziale prevede la definizione del P_S che, come visto precedentemente, integra una descrizione dettagliata del task, gli attributi stilistici da adottare e una base di conoscenza preesistente. La sua funzione è quindi quella di orientare il modello (ad esempio GPT o LLaMA) nella generazione di output che rispettino rigorosamente le caratteristiche stilistiche e semantiche del soggetto target.

Per ogni query, il sistema genera dieci risposte indipendenti, sfruttando la natura intrinsecamente stocastica degli LLM. Questi modelli sono, infatti, per loro natura non deterministici:

a fronte del medesimo input, producono output semanticamente equivalenti ma stilisticamente divergenti, a causa della variabilità nelle scelte sintattiche e lessicali. Lungi dall'essere un limite, tale caratteristica rappresenta un vantaggio strategico, poiché permette di ottenere una pluralità di enunciati che preservano il contenuto originale pur declinandolo in diverse sfumature stilistiche e garantendo, al contempo, la coerenza comunicativa.

A differenza delle metodologie che richiedono un feedback continuo al modello per la correzione stilistica, questo approccio consente di parallelizzare il processo decisionale. Tale architettura facilita una generazione rapida delle risposte, compatibile con i tempi di una conversazione naturale; al contrario, un processo di correzione iterativa avrebbe rallentato considerevolmente il sistema, compromettendo la fluidità dell'interazione.

Una volta generate le dieci varianti — caratterizzate da contenuto semantico equivalente ma forme sintattiche eterogenee — viene selezionata la risposta che esibisce la tecnica stilistica più raffinata. In assenza di variazioni nel significato, l'unico parametro di differenziazione risiede dunque nella forma. Come illustrato nella Figura 6.6, per ogni risposta viene calcolato l'embedding stilistico relativo E_x , il quale viene successivamente confrontato con l'embedding medio E_{mean} attraverso la cosine similarity.

$$E_x = Style_embedding(x) \quad (6.2)$$

$$x_{selected} = argmax_x(sim_cos(E_x, E_{mean})) \quad (6.3)$$

Per facilitare la fluidità della conversazione, la risposta selezionata corrisponde alla frase che più si avvicina all' E_{medium} . Tale processo garantisce la coerenza stilistica: ogni frase generata viene confrontata con l'embedding di riferimento e la scelta ricade sulla variante più simile, mantenendo così un registro linguistico uniforme.

Questo approccio genera un Tree of Style, in cui ogni nodo rappresenta una risposta stilisticamente coerente che emula il soggetto target. Il vantaggio principale di questa struttura risiede nel fatto che, lungo il percorso selezionato, le risposte risultano costantemente allineate all' E_{medium} , assicurando continuità per tutta la durata dell'interazione. Inoltre, tale stabilità permette al modello di adattare efficacemente il tono della conversazione, rafforzando l'illusione di interagire direttamente con l'individuo originale (si vedano le Figure 6.3 e 6.7).

Text-to-speech

Per la clonare fedelmente la voce di un individuo specifico, abbiamo scelto di adottare Piper TTS. Il processo inizia con la selezione di enunciati che rispecchino lo stile comunicativo del soggetto da imitare; successivamente, il sistema genera una sintesi vocale che riflette le caratteristiche peculiari della voce originale. Piper TTS è un sistema di sintesi vocale open-source basato sull'architettura Variational-Inference-Text-to-Speech (VITS). Questo modello end-to-end integra un generatore e un discriminatore all'interno di una rete neurale profonda, permettendo di produrre voci naturali e fluide in tempo reale.

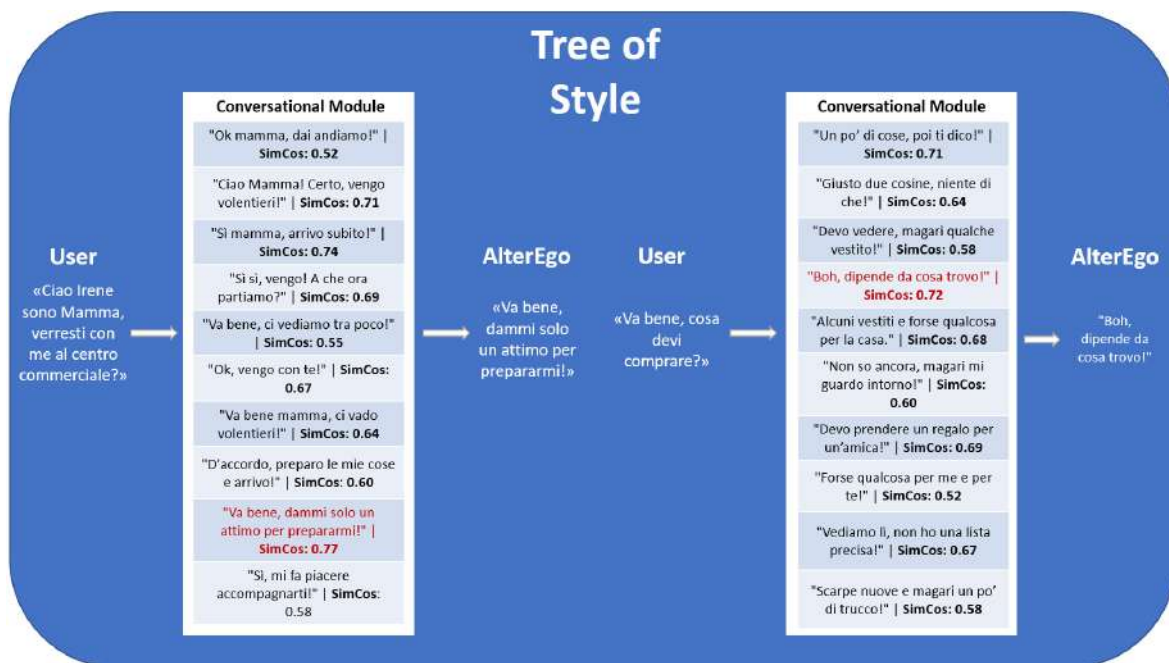


Figura 6.7: Tree of Style nella pratica. L'immagine mostra un esempio di conversazione in cui il modulo conversazionale seleziona una delle dieci risposte generate. La risposta scelta è quella con il punteggio di cosine similarity più alto rispetto allo style embedding medio. Questa risposta viene quindi restituita all'interlocutore. L'intero processo di domande e risposte, ripetuto più volte, genera il Tree of Style.

Al fine di personalizzare la voce dell'avatar, sono stati registrati circa 800 campioni audio della persona target, utilizzati per sottoporre a fine-tuning un modello pre-addestrato in lingua inglese. Questa fase di affinamento ha permesso al sistema di apprendere i tratti vocali distintivi del soggetto — quali frequenza, timbro e ritmo — garantendo che l'avatar non solo somigliasse visivamente alla persona rappresentata, ma ne riproducesse la voce in modo convincente.

La variazione diafasica — ovvero l'adattamento del linguaggio al contesto e alla relazione tra gli interlocutori — gioca inoltre un ruolo cruciale nella percezione di naturalezza del parlato. A ciò si aggiungono gli elementi paralinguistici, come l'intonazione e il ritmo, che contribuiscono in modo determinante alla credibilità e all'efficacia comunicativa. La riproduzione accurata di tali tratti assicura che lo stile espositivo sia percepito come autentico e coerente con la situazione, facilitando l'identificazione e l'empatia da parte dell'utente.

La scelta di Piper TTS è stata dettata anche dalla sua agevole integrazione con l'engine Unity, fattore che ha reso l'implementazione nel progetto rapida ed efficiente. Inoltre, la sua versatilità e la capacità di generare voci di alta qualità lo rendono una soluzione ideale per lo sviluppo di esperienze interattive e immersive all'interno del Metaverso.

6.10.4 Risultati

I risultati ottenuti appaiono promettenti e suggeriscono che l'alter ego virtuale integrato nel Metaverso possieda caratteristiche credibili e funzionali. L'avatar, infatti, rispecchia visiva-

mente l'individuo selezionato, garantendo un livello di somiglianza che favorisce l'immersione dell'utente. Parallelamente, i tempi di risposta del modulo di conversazione risultano adeguati, evitando latenze prolungate che potrebbero interrompere la naturalezza del dialogo.

Per quanto riguarda la componente conversazionale, la generazione del testo approssima in modo convincente lo stile comunicativo del soggetto rappresentato, mentre il sistema di sintesi vocale riproduce la voce originale con buona fedeltà.

La Tabella 6.1 confronta diverse tecniche di generazione del testo: Zero-Shot, ToS, Prompt+ToS e il fine-tuning diretto. Al fine di valutare la corrispondenza stilistica e contenutistica tra le frasi generate e quelle del soggetto target, sono state impiegate differenti misure di similarità, inclusi parametri lessicali, sintattici e basati sulla sentiment analysis. L'analisi dei dati evidenzia come il metodo proposto, Prompt+ToS, raggiunga prestazioni paragonabili o superiori al fine-tuning, tecnica più complessa e onerosa da implementare.

È interessante notare come Prompt+ToS ottenga i risultati migliori in diversi casi, mantenendosi prossimo al fine-tuning negli altri, laddove i metodi Zero-Shot e ToS riportano generalmente punteggi inferiori. Ad esempio, nella metrica di sentiment Sim-js, il metodo Prompt+ToS con gpt-4o-mini raggiunge lo 0,857, un valore estremamente vicino al picco di 0,939 ottenuto tramite fine-tuning. Analogamente, nella metrica di contenuto Sim-cos, Prompt+ToS registra il valore più alto in assoluto (0,642), superando tutti gli altri approcci.

Un aspetto metodologico rilevante riguarda i valori riportati in tabella, i quali rappresentano la media di 10 test eseguiti con il medesimo input. Tale scelta è dettata dalla natura intrinsecamente non deterministica degli LLM, le cui risposte possono variare leggermente a parità di condizioni iniziali. La media dei risultati su più iterazioni garantisce quindi una stima più stabile e affidabile delle prestazioni del sistema. Nel complesso, questi risultati indicano che Prompt+ToS costituisce un'alternativa valida ed efficiente al fine-tuning per lo sviluppo di moduli conversazionali in alter ego virtuali, mantenendo un'elevata qualità nella generazione del testo e raggiungendo una similarità stilistica e semantica pari, o talvolta superiore, a quella del fine-tuning.

6.10.5 Discussione

Come detto precedentemente, questo lavoro ha costituito una fase propedeutica allo sviluppo di modelli conversazionali capaci di riprodurre lo stile comunicativo unico di un individuo. L'obiettivo primario della ricerca è stato testare e affinare la metodologia, che è stata successivamente applicata ad altri soggetti nelle fasi più avanzate del progetto.

Data la natura preliminare di questa ricerca, sono stati utilizzati un dataset circoscritto e un numero limitato di LLM per delineare gli aspetti fondamentali dell'adattamento stilistico e della generazione delle risposte. Le fasi successive prevedono invece l'ampliamento del corpus di dati e l'ottimizzazione del modello per applicazioni su più vasta scala.

In generale, nella creazione di un LLM volto a replicare uno specifico stile comunicativo, la maggior parte delle metriche di valutazione si focalizza sul Text Style Transfer (TST). Tali metriche misurano concetti quali la Style Transfer Accuracy (accuratezza del trasferimento),

Technique	Model	Sim-js NVdb	Sim-js Lemmas	Sim-js 1-POS	Sim-js 2-POS	Sim-js 3-POS	Sim-js 4-POS	Sim-js 5-POS	Sim-js sent_len	Sim-js Sentiment	Sim-cos Content
Fine-tuning	gpt-4o-mini	0.955	0.372	0.873	0.730	0.521	0.347	0.241	0.825	0.939	0.604
Prompt + ToS	llama-3.3-70b-versatile	0.942	0.375	0.853	0.702	0.519	0.345	0.238	0.798	0.894	0.626
	gpt-4o-mini	0.937	0.387	0.864	0.720	0.534	0.352	0.231	0.835	0.857	0.642
ToS	llama-3.3-70b-versatile	0.935	0.357	0.810	0.667	0.493	0.322	0.223	0.852	0.832	0.580
	gpt-4o-mini	0.930	0.357	0.843	0.708	0.533	0.341	0.230	0.836	0.806	0.635
Zero-Shot	llama-3.3-70b-versatile	0.923	0.340	0.798	0.660	0.497	0.325	0.224	0.805	0.814	0.594
	gpt-4o-mini	0.906	0.331	0.804	0.672	0.505	0.330	0.226	0.811	0.788	0.603

Tabella 6.1: Test di stile e contenuto. La tabella mostra i risultati di somiglianza tra le frasi generate dai due LLM nelle tre diverse tecniche di prompting e fine-tuning con le frasi dell'individuo target raccolte per formare il set di dati di addestramento. Va notato che, nel metodo Prompt+ToS, i valori ottenuti in alcuni casi superano quelli ottenuti con il metodo di fine-tuning. Tuttavia, in altri casi, pur non superandoli, sono i più vicini, rispetto agli altri metodi Zero-Shot e ToS, ai valori evidenziati in rosso del metodo di fine-tuning.

la Content Preservation (conservazione del contenuto) e la Fluency (fluidità) [142]. Sebbene questi strumenti analizzino l'efficacia del modello nel trasformare lo stile di un testo esistente mantenendone intatto il senso, tale approccio presenta un limite significativo nel nostro contesto: esso valuta la capacità di trasformazione stilistica di un output specifico, ma non l'attitudine intrinseca del modello ad adottare stabilmente un determinato stile.

Le metriche convenzionali come BLEU e ROUGE [143, 118] non sono state utilizzate a causa della loro inadeguatezza nel misurare lo stile. Questi indicatori, sebbene ampiamente diffusi per valutare le prestazioni degli LLM, si concentrano principalmente sulla sovrapposizione di parole e n-grammi, fallendo nel cogliere le sfumature complesse di stile ed emozione che il nostro "alter ego" mira a modellare. Anche metriche come METEOR [14], pur bilanciando precisione e recall attraverso l'uso di sinonimi e parafrasi, non sono specificamente progettate per l'analisi stilistica.

Date queste limitazioni, si è scelto di adottare un approccio alternativo basato sulla valutazione della somiglianza, sia stilistica che contenutistica, rispetto al profilo dell'individuo. L'enfasi principale è stata posta sullo stile, mentre la coerenza del contenuto è stata valutata in termini di completezza. Questo metodo prevede l'impiego di metriche dedicate all'analisi della preservazione del contenuto e dello stile generati dal modello, senza focalizzarsi sullo style transfer. Infine, la valutazione della fluidità è stata omessa, poiché gli LLM hanno già ampiamente dimostrato elevati standard prestazionali in questo ambito.

Per condurre le analisi, è stato richiesto al soggetto target di rispondere a un set specifico di domande; lo stesso schema è stato poi utilizzato per generare le risposte del suo "alter ego" digitale. I risultati ottenuti sono stati confrontati per valutare la capacità dell'alter ego di replicare il comportamento e lo stile dell'individuo reale. L'analisi si è articolata su due dimensioni principali:

- **Somiglianza Stilistica** La valutazione dello stile è stata effettuata attraverso un'analisi approfondita delle distribuzioni ottenute durante la fase di preprocessing delle risposte. Tale esame ha incluso diversi parametri, tra cui la distribuzione dei lemmi, degli n-grammi POS (Part-of-Speech), la lunghezza delle frasi, la sentiment analysis e la distribuzione basata su NVdb. Le risposte dell'utente e quelle dell'alter ego sono state analizzate separatamente e successivamente messe a confronto utilizzando la divergenza di Jensen-Shannon (JSd) [129], opportunamente adattata per misurare la somiglianza anziché la divergenza:

$$Sim_{JS} = 1 - JSd \quad (6.4)$$

La scelta della divergenza di Jensen-Shannon è motivata dalla sua capacità di fornire una misura simmetrica e ben definita, che permette il confronto tra distribuzioni probabilistiche anche in presenza di valori nulli in alcune categorie. Inoltre, la JSd garantisce una maggiore stabilità numerica grazie alla media delle distribuzioni analizzate, risultando particolarmente efficace nello studio di distribuzioni stilistiche complesse, spesso caratterizzate da alta variabilità. L'impiego della formula modificata Sim_{JS} favorisce un'interpretazione più intuitiva dei risultati, rendendo più immediata la comprensione del grado di affinità tra le distribuzioni. Questo approccio ha permesso una valutazione precisa della somiglianza stilistica, agevolando il confronto tra le performance di diversi modelli e tecniche.

- **Somiglianza dei Contenuti** La conservazione del contenuto è stata valutata mediante il calcolo della cosine similarity applicata agli embedding semantici generati da Sentence-BERT [156]. Questo approccio metodologico ha permesso di quantificare con precisione il grado di sovrapposizione tra i contenuti prodotti dall'alter ego e quelli dell'individuo reale. Infine, è stato calcolato il valore medio della cosine similarity su tutte le risposte per ottenere un punteggio di sintesi complessivo.

Sebbene le tecniche impiegate si siano dimostrate solide nel valutare la conservazione dei contenuti e dello stile, la valutazione umana resta insostituibile per un'analisi qualitativa e approfondita della resa stilistica. Se da un lato le metriche automatizzate offrono un supporto prezioso per l'analisi quantitativa, dall'altro non riescono ancora a eguagliare la precisione e la sensibilità del giudizio umano.

I risultati di questo studio si inseriscono nel campo delle attuali ricerche sull'interazione uomo-robot (HRI) e sui sistemi conversazionali basati sull'intelligenza artificiale. Studi recenti hanno analizzato come gli LLM potenzino la capacità degli assistenti robotici di sostenere conversazioni naturali, adattandosi alle preferenze dell'utente e personalizzando le interazioni [77]. Parallelamente, l'impiego degli LLM negli avatar digitali è stato oggetto di indagine per il suo ruolo nel migliorare l'intelligenza emotiva e i segnali sociali degli agenti virtuali e robotici [99, 116].

Le conclusioni confermano inoltre quanto emerso in lavori precedenti: la personalizzazione dei pattern conversazionali favorisce interazioni più coinvolgenti ed efficaci, in particolare nei settori della robotica assistiva e della telepresenza. Inoltre, gli studi su robot embodiment suggeriscono che l'integrazione di sistemi di dialogo capaci di adattare lo stile potrebbe aumentare l'accettazione e l'usabilità dei robot in ambiti cruciali quali l'assistenza sanitaria, il servizio clienti e l'istruzione [157].

6.10.6 Conclusioni

Il presente studio mostra una tecnica di integrazione tra le tecnologie avanzate degli LLM e la modellazione di avatar 3D per la creazione di alter ego digitali realistici. La tecnica Tree of Style offre un metodo scalabile, efficiente e performante per replicare lo stile conversazionale di un individuo, eliminando la necessità di complessi processi di fine-tuning.

La scalabilità di questa tecnica risiede nella capacità di regolare il numero di nodi all'interno dell'albero in base alla complessità dello stile da emulare. Ad esempio, uno stile colloquiale richiede meno nodi rispetto a un linguaggio tecnico specialistico, consentendo così un'ottimizzazione mirata delle risorse computazionali. In questo studio, l'obiettivo è stato replicare il profilo di un soggetto femminile caratterizzato da un linguaggio informale in contesti familiari. Tale stile non è risultato eccessivamente oneroso da riprodurre: l'utilizzo di 10 nodi ha rappresentato un compromesso ideale per ottenere risultati ottimali senza sovraccaricare il sistema.

I risultati evidenziano la capacità del sistema di generare risposte stilisticamente e contestualmente coerenti, supportate da un modulo di sintesi vocale ad alta fedeltà. Il realismo visivo e l'accuratezza linguistica dell'avatar ne confermano il potenziale come strumento di valore per migliorare l'interazione negli ambienti virtuali.

Sebbene la ricerca si focalizzi sul Metaverso, le metodologie sviluppate presentano implicazioni significative anche nel campo della robotica. Il metodo ToS potrebbe essere implementato in sistemi robotici per aumentarne l'adattabilità linguistica. In contesti sociali e assistenziali — come i robot per la salute o il customer service — l'adozione di un'IA conversazionale capace di adattare il proprio stile renderebbe le interazioni più naturali e personalizzate.

Tuttavia, la replica dell'identità e dello stile umano solleva questioni etiche che richiedono ulteriori approfondimenti. La ricerca futura si concentrerà sul potenziamento della consapevolezza contestuale, sul miglioramento della flessibilità applicativa e sulla gestione delle sfide legate alla privacy e al consenso nell'uso di surrogati digitali.

Capitolo 7

La convergenza delle tecnologie, ovvero il progetto ALTEREGO e la Morale Artificiale

7.1 Introduzione: la Fase 5

Abbiamo visto come l'evoluzione delle tecnologie immersive e dell'IA Generativa abbia inaugurato una nuova era nella rappresentazione digitale dell'identità umana. All'interno del progetto di ricerca PRIN ALTEREGO, l'obiettivo primario, come detto precedentemente, risiede nella creazione di surrogati digitali capaci di emulare non solo l'estetica di un individuo, ma anche la sua intenzionalità, il suo stile comunicativo e, in ultima istanza, la sua dimensione etica e morale. Questa sezione aggiunge un tassello in più alla precedente e documenta lo sviluppo di un secondo caso studio che funge da continuazione diretta delle metodologie presentate nei lavori precedenti, spostando l'attenzione verso la creazione di un alterego virtuale che sia anche un agente virtuale morale. Partendo dalle evidenze della fase precedente, questa iterazione evolve quindi il metodo attraverso un sistema multi-agente basato su LLM, volto a potenziare ulteriormente l'efficacia e la sofisticazione del ToS, integrandovi i raggiungimenti della Fase 4 del lavoro sugli Agenti Morali Artificiali, oggetto principale della presente tesi, configurandosi così come una Fase 5 finale.

Se la prima fase del progetto ALTEREGO si è concentrata sulla replicazione stilistica di base, questa seconda fase aggiunge quindi il concetto di agente morale già affrontato nei precedenti capitoli. Un agente virtuale morale non si limita a generare risposte coerenti, ma deve essere in grado di valutare le implicazioni etiche delle proprie affermazioni, basandosi su un sistema di valori predefinito derivato dal soggetto originale.

Il sistema presentato in questo capitolo non va interpretato come un semplice assistente conversazionale potenziato da modelli linguistici, bensì come un artefatto sperimentale orientato alla mediazione morale situata. La sua finalità principale consiste nell'integrare percezione contestuale, interpretazione delle intenzioni dell'utente e deliberazione etica non coercitiva all'interno di un'unica architettura operativa.

La soluzione proposta si distacca da quelle ad alto costo computazionale e temporale per testare la fattibilità di una creazione rapida e "user-friendly" di alter ego virtuali, pur mantenendo una solida base di conoscenza e un sistema di ragionamento sofisticato basato sull'orchestrazione multi-agente. Attraverso l'uso di piattaforme come Avaturn per la ricostruzione 3D e LMNT per la sintesi vocale, unito all'implementazione locale di modelli linguistici come Llama e DeepSeek tramite il framework CrewAI già presentato in precedenza, questo studio analizza il delicato equilibrio tra realismo visivo, coerenza stilistica e profondità deliberativa in un agente virtuale incaricato di fornire consigli morali agli utenti.

I singoli componenti tecnologici non costituiscono di per sé il contributo originale della ricerca; l'elemento innovativo risiede nella loro integrazione funzionale all'interno di una struttura coerente orientata al supporto morale contestualizzato.

Riprendendo il concetto di esemplarità morale descritto nel Capitolo 4, l'architettura descritta di seguito è progettata per essere "calzata" da una eventuale figura di spicco (immaginaria o reale) che rappresenti un elevato esempio morale per la società o, nel caso di replica di sé stessi, configurarsi come il nostro "grillo parlante" personale, ovvero la nostra coscienza morale.

Componente del sistema	Metodologia precedente	Metodologia attuale
Creazione avatar 3D	Headshot (Character Creator 4)	Avaturn (Rapid creation)
Sintesi vocale	PiperTTS (fine-tuning esteso)	LMNT (1 min audio cloning)
Ragionamento LLM	Tree-of-Thought (ToT)	CrewAI (multi-agent orchestration)
Hosting modelli	Server remoti / Cloud	Locale (Llama, DeepSeek)
Base di conoscenza	Dataset limitato	Knowledge base solida (libri, atti, video)

Tabella 7.1: Confronto tra metodologia precedente e attuale

7.2 Creazione Rapida di Avatar: Il Passaggio ad Avaturn

Un obiettivo centrale di questo secondo caso studio è stato l'abbattimento delle barriere tecniche ed economiche per la creazione di avatar. Se nel lavoro precedente la generazione della mesh 3D richiedeva software complessi e lunghi tempi di elaborazione, in questo progetto si è optato per Avaturn [1].

Avaturn è una piattaforma che utilizza algoritmi di intelligenza artificiale per trasformare semplici foto 2D in avatar 3D pronti per l'animazione. Il sistema preserva i tratti facciali riconoscibili e genera un modello dotato di rigging umanoide standard, blendshapes ARKit e visemi, garantendo la compatibilità immediata con Unity Engine. Questa scelta risponde alla necessità di testare soluzioni "low-cost" e rapide, facilitando l'adozione di queste tecnologie da parte di un pubblico non esperto.

Tuttavia, la velocità del processo introduce limitazioni qualitative. L'analisi tecnica evidenzia come la creazione rapida comporti una perdita parziale di somiglianza e realismo rispetto a tecniche di scultura digitale o fotogrammetria densa, come nel caso precedente. Avaturn offre tipologie diverse di avatar: il tipo T1 garantisce un ottimo realismo visivo ma sacrifica l'animazione facciale (rig della faccia e blendshapes non utilizzabili), mentre il tipo T2 consente l'animazione completa tramite ARKit ma risulta leggermente meno fedele nei dettagli fisiognomici del soggetto originale. Per l'agente morale è stato selezionato il modello T2, privilegiando la capacità di espressione facciale necessaria per supportare la comunicazione di messaggi complessi e carichi di valore etico, pur accettando un lieve scostamento estetico dalla figura reale.

Sebbene la piattaforma consenta la costruzione di un avatar anche a partire da foto autoscattate con lo smartphone, si è preferito acquisire le immagini del soggetto attraverso la stessa fotocamera Reflex Nikon D3400 utilizzata nel caso precedente, mantenendo così la continuità ambientale con tale lavoro ma accelerando drasticamente la pipeline di asset creation.

7.3 Sintesi Vocale ad Alte Prestazioni: L'Integrazione di LMNT - Emotive AI speech for voiceover & characters

Parallelamente alla componente visiva, la replicazione dell'identità vocale ha subito un'evoluzione verso la rapidità d'uso attraverso la piattaforma LMNT [4]. LMNT è un sistema di TTS progettato per offrire una latenza estremamente bassa (150-200ms) e una qualità "studio-level" nel voice cloning. Invece di richiedere ore di registrazione, LMNT permette di ottenere risultati ottimi già con un solo minuto di audio.

L'integrazione con Unity avviene tramite un plugin dedicato, che facilita lo streaming dell'audio sintetizzato in tempo reale, un requisito fondamentale per le interazioni nel Metaverso dove la pausa tra domanda dell'utente e risposta dell'avatar deve essere ridotta al minimo per non interrompere la sensazione di presenza. LMNT utilizza architetture avanzate di inferenza che estraggono le caratteristiche acustiche uniche del timbro, della frequenza e del ritmo della voce da replicare, permettendo all'avatar di parlare con una voce immediatamente riconoscibile.

Emerge però una sfida tecnica legata all'espressività. Sebbene i modelli di LMNT siano teoricamente in grado di adattare il tono all'emozione rilevata nella frase, questa capacità è strettamente correlata alla quantità di dati di addestramento. Con dataset ridotti a pochi minuti di audio, il sistema fatica a modulare correttamente le sfumature emotive, mantenendo uno stile di eloquio più neutro e meno dinamico rispetto ad una voce reale. Nonostante ciò, la fedeltà timbrica ottenuta con un investimento di tempo così esiguo rappresenta un progresso significativo per l'usabilità del sistema su larga scala.

7.4 Il Cervello dell'Alterego Morale: Orchestrazione Multi-Agente con CrewAI

Il cuore dell'innovazione di questo caso studio risiede nella parte conversazionale. Il precedente sistema basato su Tree-of-Thought (ToT) è stato sostituito dal framework multi-agente CrewAI. Questa scelta non è dettata solo da una ricerca di prestazioni pure — che rimangono pressoché identiche tra i due sistemi — ma da una volontà di modellare in modo più sofisticato il processo di deliberazione morale.

7.4.1 Passaggio dal Tree-of-Thought a CrewAI

Mentre il Tree-of-Thought si concentra su un'esplorazione gerarchica di soluzioni alternative all'interno di un singolo processo di ragionamento, CrewAI permette di scomporre la personalità del soggetto da replicare in molteplici agenti specializzati che collaborano tra loro. Questa architettura "ruolo-centrica" è più adatta a replicare una figura complessa che agisca contemporaneamente da coscienza personale e mentore etico.

7.4.2 Configurazione della "Crew" per l'Alterego Morale

L'applicazione di CrewAI al sistema morale dell'Alterego prevede la creazione di un team di agenti LLM, ognuno con accesso a una parte specifica della conoscenza del soggetto originario e dei suoi valori:

1. **L'Analista:** questo agente ha il compito primario di filtrare ogni richiesta attraverso il metodo scientifico. Il suo obiettivo è fornire risposte basate esclusivamente su fatti validati e sulle esperienze del soggetto replicato, estraendo informazioni fattuali e prove scientifiche dalla Knowledge Base fornita per supportare il consiglio.
2. **L'Eticista:** Esperto di etica normativa e filosofia morale, questo agente valuta le implicazioni sociali del consiglio. Si occupa di temi come la trasparenza, la meritocrazia e la responsabilità di fronte al cittadino. Analizza quindi il dilemma dell'utente utilizzando framework etici predefiniti (come l'utilitarismo, la deontologia o principi di integrità pubblica). Questo agente agisce come un supervisore morale, assicurandosi che il tono e il contenuto del messaggio siano allineati ai doveri civili del soggetto.
3. **Il Redattore Stilistico:** questo agente riceve l'output tecnico e morale dai primi due e lo trasforma in un discorso coerente che segua lo stile comunicativo del soggetto replicato. Utilizza la tecnica del Tree of Style (ToS) per selezionare le parole e le strutture sintattiche che più si avvicinano all'originale.

La collaborazione tra questi agenti avviene attraverso un processo sequenziale o gerarchico, dove il Redattore Stilistico agisce spesso come manager, coordinando il flusso di informazioni per produrre la risposta finale che l'utente vedrà e ascolterà.

7.4.3 Implementazione Locale di LLM: Llama e DeepSeek per la Privacy e la Sicurezza

Per garantire che l'intero sistema possa operare in modo autonomo e protetto, è stata scelta un'implementazione completamente locale dei modelli linguistici. Questo approccio è cruciale quando si tratta con figure istituzionali e dati sensibili, evitando la dipendenza da API esterne che potrebbero compromettere la privacy dell'utente o la sovranità sui dati della Knowledge Base.

I modelli selezionati sono Llama (nella versione 3 o successive) e DeepSeek (in particolare le varianti orientate al ragionamento come DeepSeek-R1). Questi modelli vengono serviti localmente tramite Ollama, permettendo a CrewAI di interfacciarsi con essi come se fossero endpoint OpenAI-compatibili.

7.4.4 Il Ruolo di DeepSeek-R1 nel Ragionamento Morale

DeepSeek-R1 è stato integrato per le sue eccezionali capacità di ragionamento logico e matematico, ottenute attraverso tecniche di apprendimento per rinforzo (Reinforcement Learning). In un contesto di consiglio morale, DeepSeek-R1 eccelle nel generare un "Chain-of-Thought" (CoT) visibile, permettendo al sistema di "pensare" esplicitamente alle implicazioni etiche prima di formulare la risposta finale. Il modello è in grado di auto-correggersi e di riflettere su più passaggi, garantendo che il consiglio fornito dall'avatar non sia una mera associazione statistica di parole, ma il risultato di una deliberazione strutturata.

Modello	Ruolo nel Sistema	Forza tecnica
Llama 3 (8B/70B)	Redattore Stilistico	Ottima comprensione linguistica e versatilità
DeepSeek-R1	Analista Scientifico e Eticista Pubblico	Chain-of-Thought (CoT) profondo e logica rigorosa

Tabella 7.2: Confronto tra Llama e DeepSeek-R1 nell'architettura del sistema

Per far girare questi modelli localmente con buone performance, l'architettura richiede hardware dedicato, come GPU con almeno 24GB di RAM (RTX 4090 o simili) per gestire le sequenze di ragionamento estese e la memoria multi-agente. L'uso di versioni distillate (es. DeepSeek-R1-Distill-Llama-70B) permette di mantenere l'80-90% delle capacità di ragionamento del modello completo riducendo drasticamente i requisiti di memoria.

7.4.5 Integrazione della Knowledge Base e replicazione dello stile comunicativo

Nonostante il cambio di architettura logica verso CrewAI, il trasferimento di stile rimane ancorato alla metodologia Tree of Style (ToS) sviluppata nel primo lavoro del progetto AL-TEREGO (vedi Capitolo 6). Trattati come Diamesia, Diaphasia, Lessico e Retorica, evidenziati nel lavoro precedente, vengono quindi efficacemente analizzati e utilizzati dal sistema. Per

alimentare il sistema multi-agente di CrewAI e supportare il ragionamento etico, è stata costruita una base di conoscenza Knowledge Base che attinge a diverse tipologie di fonti. L'integrazione di questi dati avviene attraverso il RAG, permettendo agli agenti di recuperare informazioni specifiche in tempo reale durante la conversazione.

La KB è strutturata in tre aree principali:

- **Nucleo Fondativo:** Pubblicazioni, discorsi celebri e testi scritti che definiscono la dottrina o il pensiero della figura morale.
- **Nucleo Civile e Istituzionale:** Documenti che testimoniano l'impegno sociale e la visione politica del bene pubblico.
- **Nucleo Biografico e Narrativo:** Interviste e testimonianze che rivelano l'umanità del soggetto, i suoi dubbi e i suoi fallimenti, evitando di far percepire l'avatar come un dio infallibile o un oracolo.

7.5 Applicazione Pratica: CrewAI in Azione per la Deliberazione Morale

Immaginiamo uno scenario in cui un utente interagisce con l'Alterego Morale ponendo un dilemma etico complesso: *"Ho scoperto che un mio collega si sta prendendo il merito di un progetto che ho guidato io. Vorrei segnalarlo ai superiori, ma temo di essere visto come un 'informatore' o di rovinare il clima nel team. Cosa mi consigli?"*

1. **Analisi del Compito (Task Definition):** Il sistema riceve l'input e definisce tre task principali: analisi dei fatti, valutazione delle implicazioni etico-istituzionali e redazione finale del consiglio.
2. **L'Analista (DeepSeek-R1):** interroga la KB alla ricerca di principi legati al merito e alla trasparenza. Recupera concetti chiave come l'importanza dell'onestà intellettuale come cura per il progresso e l'idea che la competizione debba essere basata esclusivamente sul merito per essere sana. Identifica che il silenzio di fronte a un'ingiustizia danneggia non solo il singolo, ma l'intera "comunità" (o azienda).
3. **L'Eticista (DeepSeek-R1):** questo agente non genera subito il testo, ma produce un ragionamento interno (*Chain-of-Thought*) visibile al sistema :

ϕthink ς Il dilemma oppone la Giustizia (merito individuale) all'Armonia Sociale (clima del team). Tuttavia, un'armonia basata sulla menzogna è fragile. Devo consigliare un'azione che protegga la verità senza essere vendicativa. Il merito non è un possesso egoistico, ma un dovere civile verso l'organizzazione. Strategia: suggerire un confronto diretto basato sull'evidenza dei fatti prima di un'escalation formale. ϕthink ς

4. **Il Redattore Stilistico (Llama 3.3 + ToS):** Questo agente riceve il nucleo logico e genera 10 varianti della risposta. Qui entra in gioco il metodo **Tree of Style (ToS)** :
- Vengono analizzati gli *style embeddings* delle 10 frasi generate.
 - Il sistema scarta le versioni troppo burocratiche o troppo aggressive.
 - Viene selezionata la frase con la massima similarità del coseno rispetto al profilo dell'”Esemplare Morale” (caratterizzato da un registro solenne ma pedagogico).
5. **Output e Sintesi:** La risposta testuale viene inviata a LMNT per la generazione vocale e all'avatar in Unity. L'utente percepisce un discorso coerente, informato e moralmente schierato:

”La ringrazio per la sua fiducia. Vede, l'onestà intellettuale è il cardine su cui poggia ogni comunità che voglia definirsi progredita. Permettere che il merito venga distorto non è solo un danno per lei, ma una ferita alla trasparenza di tutto il suo gruppo di lavoro. Non si senta un 'informatore': difendere la verità dei fatti è un atto di responsabilità civile. Le suggerisco di cercare un confronto basato proprio sull'evidenza dei dati, con la fermezza di chi sa che il sapere e la correttezza, alla fine, sono ciò che ci salva dall'irrazionalità.”

Il comportamento del sistema può essere descritto come una sequenza deliberativa composta da: (i) acquisizione del contesto; (ii) interpretazione della richiesta e delle intenzioni implicite; (iii) analisi delle possibili opzioni; (iv) valutazione etica situata; (v) generazione di un feedback argomentato; (vi) mantenimento della libertà decisionale dell'utente.

Gli scenari riportati non hanno finalità dimostrativa in senso statistico, ma illustrano il funzionamento dell'architettura in situazioni eticamente rilevanti e consentono di osservare le modalità con cui il sistema struttura il proprio supporto deliberativo.

7.6 Analisi Comparativa: Realismo vs. Rapidità e Latenza

Realismo Visivo e Vocale L'adozione di Avaturn e LMNT ha ridotto i tempi di produzione degli asset da giorni a pochi minuti, ma ha introdotto una limitazione nel ”grado di presenza”. L'avatar generato da Avaturn, pur essendo riconoscibile, soffre di una standardizzazione della mesh che rende meno evidenti le rughe d'espressione e le asimmetrie facciali che contribuiscono all'unicità di un volto umano. Similmente, LMNT produce una voce cristallina e identica nel timbro, ma la brevità del dataset di input (1 minuto) limita la capacità del modello di riprodurre l'ampia gamma emotiva e le pause retoriche utilizzate dal soggetto per dare enfasi ai suoi discorsi.

7.6.1 Latenza e Ragionamento

In termini di performance conversazionale, il sistema CrewAI basato su Llama e DeepSeek locali offre un'esperienza paragonabile al ToT in termini di precisione del consiglio morale, ma con una maggiore robustezza nel gestire compiti complessi che richiedono il recupero di dati da più fonti.

In altri termini:

- **Latenza:** Grazie all'orchestrazione locale e ai modelli quantizzati, l'intero processo di deliberazione e selezione ToS avviene in circa 2.5 - 3.5 secondi.
- **Coerenza:** Il sistema non ha semplicemente risposto con un "sì" o "no", ma ha applicato una **bussola etica** derivata dalla Knowledge Base, mantenendo uno stile comunicativo che riflette un'identità autorevole e rassicurante.

Metrica di Performance	ToT + Cloud LLM (Legacy)	CrewAI + Local LLM (Current)
Accuratezza Morale	Alta (basata su prompt denso)	Alta (basata su deliberazione multi-agente)
Coerenza Stilistica	Elevata (metodo ToS)	Elevata (metodo ToS integrato)
Latenza Media	1.5 - 2.5 secondi	2.0 - 3.5 secondi (orchestrazione locale)
Manutenibilità KB	Complessa (prompt lunghi)	Semplice (Agente RAG dedicato)
Privacy dei Dati	Dipendente dal provider cloud	Totale (locale)

Tabella 7.3: Confronto tra le due architetture per ALTEREGO

7.7 Validazione Tecnica

Il progetto dell'Alterego Morale rappresenta la sintesi e la convergenza delle soluzioni architetturelle sviluppate nelle fasi precedenti, integrando agenti morali deliberativi in un alter ego digitale incarnato. La validazione condotta in questo contesto ha avuto come obiettivo principale la verifica della coerenza dell'intera pipeline, dalla generazione dell'avatar alla produzione di risposte morali vocali e contestualizzate in tempo reale.

Le prove tecniche si sono concentrate sull'analisi delle prestazioni del sistema in termini di latenza, stabilità e qualità del processo deliberativo. In particolare, è stato osservato come l'adozione di modelli linguistici eseguiti localmente consenta un maggiore controllo sulla privacy dei dati e una riduzione della dipendenza da servizi esterni, pur introducendo vincoli computazionali che incidono sui tempi di risposta.

L'orchestrazione multi-agente ha permesso di scomporre il problema morale in sottoprocessi specializzati, favorendo una maggiore trasparenza del ragionamento e una migliore

esplicitazione delle motivazioni alla base delle risposte generate. Tale approccio, sebbene comporti un aumento della latenza rispetto a soluzioni più semplici, produce risposte qualitativamente più strutturate e coerenti con il contesto situazionale.

Un aspetto emerso con chiarezza riguarda il trade-off tra rapidità e profondità deliberativa. Le configurazioni orientate alla massima reattività risultano più adatte a interazioni rapide, ma tendono a sacrificare la complessità argomentativa. Al contrario, le architetture deliberative multi-agente privilegiano la qualità del ragionamento morale, a fronte di un incremento dei tempi di elaborazione. Questa tensione riflette una caratteristica strutturale dei sistemi deliberativi e rappresenta un elemento progettuale centrale per applicazioni future.

È opportuno sottolineare che anche in questo caso la validazione non mira a dimostrare un impatto misurabile sul comportamento degli utenti, ma a verificare la fattibilità tecnica e concettuale di un alter ego morale operativo in ambienti immersivi persistenti. I risultati ottenuti confermano la possibilità di integrare agenti morali deliberativi in forme incarnate digitali, aprendo la strada a successive fasi di sperimentazione empirica.

7.8 Conclusioni e Sviluppi Futuri

Questo lavoro suggerisce che la creazione di agenti virtuali morali è una frontiera raggiungibile attraverso l'integrazione di sistemi multi-agente e basi di conoscenza solide.

Nonostante le limitazioni rilevate nell'uso di tecniche di creazione rapida — che comportano un compromesso sulla somiglianza fisica e sull'espressività vocale — il sistema complessivo si è dimostrato efficace nel fornire consigli morali che riflettono fedelmente il pensiero dell'originale. Il framework CrewAI, orchestrando modelli locali come DeepSeek e Llama, offre una struttura flessibile e sicura per la deliberazione etica, superando la rigidità delle architetture precedenti.

Dalla sperimentazione emerge quindi che la combinazione di modelli generativi, orchestrazione multi-agente e moduli contestuali permette di ottenere un livello di interazione qualitativamente diverso rispetto ai sistemi conversazionali tradizionali. Il valore aggiunto non consiste nella sola qualità linguistica delle risposte, ma nella capacità del sistema di collocare il dialogo entro un processo deliberativo più trasparente, strutturato e situato.

Va sottolineato inoltre che gli elementi identitari, stilistici e incarnativi del sistema non sono stati introdotti come meri fattori estetici, ma come componenti funzionali alla costruzione di fiducia, continuità relazionale e accettabilità dell'interazione uomo-macchina.

Le direzioni future della ricerca dovrebbero concentrarsi sull'integrazione di "prompt dinamici" capaci di adattare non solo il contenuto, ma anche lo stile dell'agente al contesto specifico dell'interlocutore (famiglia, istituzione, social) e sulla risoluzione dei limiti emotivi della sintesi vocale rapida attraverso l'uso di tecniche di prompt engineering più sofisticate per LMNT.

Infine occorre sottolineare che, in tutte le configurazioni progettate, la responsabilità morale delle decisioni rimane interamente in capo all'essere umano. L'agente artificiale

opera come amplificatore cognitivo e specchio deliberativo, fornendo supporto argomentativo e contestuale senza sostituirsi all'autonomia decisionale dell'utente. Questa scelta progettuale riflette una precisa posizione etica, che rifiuta l'attribuzione di agency morale piena ai sistemi artificiali e preserva la centralità della responsabilità umana.

Capitolo 8

Sintesi Sperimentale, Contributi Originali e Frontiere della Convergenza Robotica

L'analisi conclusiva del percorso di ricerca presentato in questa tesi di dottorato impone una riflessione sistematica sulla traiettoria che ha condotto allo sviluppo di agenti morali artificiali capaci di operare nella complessa intersezione tra ambienti virtuali immersivi e realtà fisica. La transizione verso un Metaverso ubiquitario, inteso come un ecosistema digitale persistente e diffuso, non rappresenta soltanto un'evoluzione infrastrutturale delle tecnologie di rete, ma configura una nuova dimensione dell'esperienza umana in cui la mediazione etica diventa una necessità funzionale. La presente ricerca suggerisce che l'integrazione di architetture cognitive neuro-simboliche con le potenzialità generative degli LLM permetterebbe di superare la rigidità dei sistemi basati su regole, offrendo al contempo un ancoraggio semantico e spaziale che mitiga i rischi di incoerenza e allucinazione tipici dei modelli puramente statistici [53].

8.1 Sintesi dei risultati delle fasi di sperimentazione

La presente ricerca non propone una soluzione definitiva al problema della moralità artificiale, né pretende di trasferire il giudizio etico umano alla macchina. Il contributo principale consiste piuttosto nell'aver mostrato la plausibilità teorica e la fattibilità tecnica di architetture artificiali orientate alla mediazione morale non coercitiva in ambienti digitali complessi.

Il nucleo operativo della tesi si sviluppa attraverso una progressione coerente di fasi sperimentali, ciascuna orientata ad affrontare una specifica criticità nella progettazione di agenti sociali e morali situati in ambienti immersivi. Tale progressione consente di ricostruire il modo in cui l'architettura finale emerga come risultato di un processo di raffinamento incrementale, nel quale componenti di percezione visiva, ragionamento spaziale e intelligenza conversazionale vengono progressivamente integrate.

Le diverse fasi non devono essere intese come prototipi isolati, ma come momenti evolutivi di un'unica traiettoria progettuale che ha condotto dall'analisi del comportamento sociale

alla deliberazione morale esplicabile. Esse non costituiscono una validazione statistica dell'efficacia comportamentale sugli utenti, ma mostrano la fattibilità tecnica e architettonica di agenti morali situati e deliberativi, ponendo le basi per successive valutazioni empiriche.

Inoltre, il Metaverso non viene adottato in questa ricerca come semplice contesto applicativo o come scenario tecnologico accessorio, ma come ambiente epistemicamente rilevante. Le caratteristiche di immersione, incarnazione e continuità esperienziale proprie degli ambienti immersivi rendono osservabili e modellabili dinamiche intenzionali, relazionali e morali che risultano difficilmente accessibili in interazioni puramente testuali o bidimensionali.

In tale senso, il Metaverso costituisce una condizione abilitante per lo studio di agenti morali artificiali situati, poiché amplifica il ruolo del contesto, del corpo virtuale e della percezione spaziale nel processo decisionale umano.

Fase 1: conformità sociale e approccio MET-iquette La prima fase della sperimentazione è stata dedicata alla progettazione di un'architettura per NPC sociali, denominata MET-iquette, fondata sulla teoria della pratica sociale. In tale prospettiva, il comportamento non è modellato come sequenza di obiettivi atomici, ma come partecipazione a pratiche routinarie e contestualizzate. L'ipotesi di fondo è che un agente virtuale risulti credibile solo se in grado di riconoscere la pratica sociale in atto attraverso l'interpretazione della scena.

L'implementazione in ambiente Unity ha previsto l'impiego di reti neurali convoluzionali profonde, quali EfficientNet e YOLOv4-tiny, per l'identificazione di oggetti e attori rilevanti. In scenari simulati, come un supermercato virtuale, l'agente è risultato capace di discriminare tra differenti tipologie di soggetti sulla base di marker visivi e di attivare conseguentemente regole di etichetta differenziate. Il motore di regole DROOLS ha svolto un ruolo centrale nella gestione della logica normativa, consentendo di associare specifiche pratiche sociali a comportamenti appropriati. Questa fase ha fornito un primo livello di conformità sociale basato sull'interpretazione strutturata degli eventi.

Fase 2: paradigma del grillo parlante e ragionamento morale situato La seconda fase ha spostato l'attenzione dalla conformità sociale alla deliberazione morale, introducendo il paradigma del "grillo parlante". L'obiettivo non è stato quello di imporre vincoli normativi, ma di esplorare come un agente artificiale possa supportare una scelta etica mantenendo una postura non coercitiva.

In questa fase sono state integrate due tipologie di agenti morali: uno basato su una catena di esperti (Chain-of-Expert), addestrato su dataset strutturati come Moral Stories, e uno basato su un modello generativo di tipo LLM. Il confronto tra questi approcci non ha prodotto risultati empirici sul piano comportamentale, ma ha evidenziato differenze progettuali rilevanti. L'agente basato su regole e classificazioni esplicite garantisce trasparenza e controllo, mentre l'agente generativo consente una presentazione più articolata e discorsiva del ragionamento morale. Tale distinzione solleva questioni centrali relative al rapporto tra determinismo normativo e flessibilità argomentativa, che rappresentano un asse di sviluppo per ricerche future.

Fase 3: CASPER-MV e riconoscimento delle intenzioni tramite QSR La terza fase ha introdotto un avanzamento sostanziale nella capacità percettiva dell'agente attraverso l'adattamento dell'architettura cognitiva CASPER al contesto del Metaverso. L'obiettivo è stato dotare l'agente di capacità di intention reading basate sull'osservazione del comportamento spaziale, piuttosto che sull'analisi linguistica.

L'integrazione del Ragionamento Spaziale Qualitativo ha permesso di rappresentare le interazioni tra avatar e oggetti mediante categorie semantiche astratte. Le valutazioni tecniche hanno mostrato che l'architettura CASPER-MV è in grado di classificare movimenti complessi e di associare le azioni osservate a obiettivi plausibili quando queste sono riferite a oggetti target specifici. Tale risultato è rilevante poiché indica che l'ancoraggio del ragionamento morale a una rappresentazione strutturata dello spazio riduce l'ambiguità interpretativa tipica dei sistemi puramente linguistici.

Fase 4: integrazione neuro-simbolica e orchestrazione multi-agente La quarta fase ha rappresentato la sintesi tra architetture cognitive simboliche e LLM, attraverso l'introduzione di un sistema di deliberazione etica orchestrato tramite piattaforme multi-agente. In questa configurazione, il modulo di percezione spaziale gestisce risposte rapide e reattive, mentre l'orchestrazione multi-agente consente una deliberazione più lenta e analitica, in linea con modelli cognitivi di tipo dual-process.

L'impiego di CrewAI ha reso possibile la distribuzione del ragionamento morale tra agenti specializzati, migliorando la coerenza interna delle decisioni generate e riducendo l'emergere di risposte eticamente inappropriate tipiche di modelli non supervisionati. L'integrazione con ConvAI ha inoltre consentito di esplorare modalità di incarnazione dotate di consapevolezza spaziale, permettendo all'agente di adattare la propria espressività al contesto dell'interazione. Anche in questo caso, le osservazioni hanno carattere architeturale e qualitativo, e non costituiscono una misurazione empirica della fiducia o dell'efficacia persuasiva.

Analisi e risultati del progetto ALTEREGO Il progetto PRIN ALTEREGO ha fornito il contesto applicativo per estendere le architetture sviluppate alla creazione di alter ego digitali di individui reali, con l'obiettivo di replicarne non solo l'aspetto estetico, ma anche le caratteristiche stilistiche e comunicative.

Uno dei contributi più significativi emersi in questa fase è la definizione della metodologia Tree of Style (ToS), progettata per affrontare il problema della generazione stilisticamente coerente in presenza di dataset estremamente ridotti. Il ToS sfrutta la natura stocastica dei LLM per generare più varianti di una risposta e selezionare quella che massimizza la somiglianza con uno stile di riferimento. Le analisi condotte su un dataset limitato hanno mostrato che tale approccio consente di ottenere livelli di coerenza stilistica comparabili a quelli del fine-tuning tradizionale, con un impiego significativamente inferiore di dati e risorse computazionali.

L’Alter Ego Morale: fase finale della ricerca L’ultima fase della ricerca ha riguardato la realizzazione di un alter ego digitale morale, come convergenza delle tecnologie risultate più efficaci nelle fasi precedenti. L’architettura sviluppata integra modelli linguistici locali e un sistema multi-agente orientato alla deliberazione etica, garantendo coerenza valoriale e trasparenza del processo decisionale.

Le prove tecniche condotte confermano la fattibilità della creazione di un “surrogato morale” capace di affrontare dilemmi complessi mantenendo coerenza argomentativa e stilistica. In questa prospettiva, l’alter ego morale non rappresenta un sistema normativo definitivo, ma una dimostrazione concreta della possibilità di progettare agenti artificiali deliberativi, situati e integrabili in ecosistemi immersivi complessi.

8.2 Originalità del contributo e metodologia di valutazione

L’originalità di questa tesi risiede nell’integrazione di ambiti di ricerca che, pur risultando fortemente interconnessi sul piano teorico e applicativo, sono spesso affrontati separatamente in letteratura: etica delle macchine, architetture cognitive, robotica cognitiva, ambienti immersivi e IA Generativa multimodale. Il contributo della ricerca si articola lungo una serie di direttrici principali, che riguardano sia il piano architeturale sia quello concettuale e metodologico.

Formalizzazione del concetto di mediatore morale immersivo. Uno dei contributi centrali del lavoro consiste nella definizione di un agente artificiale concepito non come sistema prescrittivo o di moderazione automatica, bensì come mediatore deliberativo situato. A differenza di chatbot etici generalisti o sistemi di controllo normativo, l’agente proposto opera all’interno di ambienti tridimensionali persistenti, integra informazioni contestuali e mantiene una postura non coercitiva. In questa prospettiva, la dimensione immersiva non rappresenta un semplice supporto tecnico, ma costituisce una componente strutturale del modello di interazione morale.

Integrazione neuro-simbolica tra intention reading, LLM e orchestrazione multi-agente. Un ulteriore elemento di originalità risiede nella combinazione, all’interno di una medesima architettura, di componenti che in letteratura sono spesso trattate separatamente: moduli di lettura delle intenzioni, strutture cognitive simboliche, LLM e meccanismi di orchestrazione multi-agente. Tale integrazione ha consentito di sviluppare un’architettura deliberativa situata, capace di articolare il ragionamento morale in fasi interne maggiormente trasparenti ed esplicabili.

Più specificamente, l’integrazione con CASPER-MV ha permesso di ancorare la generazione linguistica a una rappresentazione situata del contesto, riducendo la dipendenza esclusiva da correlazioni statistiche apprese durante l’addestramento. In questo senso, la ricerca non dimostra la risoluzione definitiva del problema del grounding nei modelli lin-

guistici, ma mostra la fattibilità di un approccio neuro-simbolico capace di rafforzare la pertinenza contestuale e la leggibilità del processo deliberativo.

Traduzione operativa del paradigma del “grillo parlante digitale”. La tesi propone inoltre una formalizzazione operativa di una figura spesso evocata in termini metaforici: quella di un agente capace di accompagnare il processo decisionale morale umano senza sostituirvisi. In questo lavoro, tale paradigma viene tradotto in una configurazione progettuale precisa, nella quale l’agente mantiene l’autonomia dell’utente, evitando sia l’imposizione normativa sia una neutralità puramente passiva. Ciò consente di collocare l’agente morale artificiale in una posizione intermedia tra assistenza cognitiva, mediazione etica e supporto deliberativo.

La tecnica Tree of Style (ToS). Tra i contributi progettuali più originali rientra anche la proposta della tecnica *Tree of Style* (ToS), concepita come strategia per la personalizzazione persistente dello stile comunicativo degli agenti digitali. A differenza delle tradizionali tecniche di *Text Style Transfer*, prevalentemente orientate alla trasformazione di testi esistenti, il ToS è stato sviluppato per modellare la tendenza dell’agente ad assumere un determinato idioletto in modo dinamico e continuativo nel corso dell’interazione. Sebbene tale proposta richieda ulteriori validazioni comparative dedicate, essa rappresenta un contributo originale sul piano della progettazione di identità comunicative artificiali.

Un modello operativo per la mediazione morale nel Metaverso. La ricerca propone inoltre un modello operativo nel quale protocolli di consenso, agenti supervisor e moduli deliberativi cooperano per supportare il ragionamento morale in ambienti immersivi. I risultati ottenuti nelle sperimentazioni non consentono di affermare in modo empiricamente conclusivo che tali sistemi riducano stabilmente fenomeni quali la disinibizione online o modifichino il comportamento morale degli utenti, ma suggeriscono la plausibilità tecnica di architetture in grado di offrire un supporto etico contestualizzato preservando l’autonomia decisionale umana.

Metodologia di valutazione. Dal punto di vista metodologico, il contributo della tesi si colloca nel paradigma della *Design Science Research*, nel quale la produzione e il raffinamento iterativo dell’artefatto costituiscono essi stessi parte del risultato scientifico. L’originalità del lavoro non risiede quindi soltanto negli artefatti sviluppati, ma anche nel processo di progettazione comparativa attraverso cui sono state progressivamente elaborate, testate e discusse differenti configurazioni architettonali.

La valutazione è stata condotta secondo criteri prevalentemente tecnico-qualitativi, orientati a osservare coerenza architettonale, stabilità operativa, sostenibilità computazionale, pertinenza contestuale, trasparenza del ragionamento e capacità deliberativa. In questa prospettiva, la tesi fornisce un contributo metodologico rilevante nella formalizzazione di un percorso di progettazione e validazione per agenti morali artificiali situati.

Spunti di riflessione teorica. Infine, la proposta di una Moral AI ubiquitaria — intesa come continuità tra agente immersivo, alter ego digitale e possibile incarnazione robotica — apre una linea di riflessione che estende il tema della mediazione morale oltre il solo spazio virtuale. Tale prospettiva non costituisce un risultato già pienamente consolidato, ma rappresenta una direzione teorica e progettuale che emerge in modo coerente dall'intero percorso di ricerca.

8.3 Ambito e limiti intenzionali della ricerca

Prima di discutere i limiti strutturali e le criticità emerse nel corso della ricerca, è opportuno chiarire l'ambito entro cui i risultati della tesi devono essere interpretati, distinguendo tra limitazioni intenzionali — derivanti da scelte metodologiche consapevoli — e limiti aperti emersi durante la sperimentazione.

La presente ricerca non si propone di validare empiricamente l'efficacia degli agenti morali artificiali nel modificare il comportamento etico degli utenti attraverso studi sperimentali su larga scala, né di misurare in modo diretto l'impatto psicologico o sociale a lungo termine dell'interazione con tali sistemi. Di conseguenza, i risultati ottenuti non devono essere letti come evidenze conclusive circa l'efficacia comportamentale o l'accettabilità sociale degli agenti sviluppati, bensì come risultati di progettazione, integrazione e validazione tecnico-qualitativa.

Allo stesso modo, il lavoro non mira a definire o imporre un modello normativo universale di moralità artificiale, né a risolvere in modo definitivo il problema dell'allineamento etico dei sistemi di Intelligenza Artificiale. L'obiettivo è più circoscritto: esplorare la possibilità di progettare agenti capaci di supportare il ragionamento morale umano in forma situata, contestuale e non coercitiva.

Coerentemente con tale impostazione, agli agenti sviluppati non viene attribuita alcuna forma di responsabilità morale o giuridica. La responsabilità delle decisioni rimane interamente in capo agli esseri umani coinvolti nella progettazione, nell'impiego e nel contesto applicativo del sistema. L'agente artificiale viene quindi concepito come strumento di supporto deliberativo, e non come soggetto morale autonomo.

In questa prospettiva, l'opacità intrinseca degli LLM non viene trattata come un problema risolvibile esclusivamente a livello del modello, bensì come una condizione strutturale che richiede mediazione architettonica, controllo contestuale e supervisione. L'elemento eticamente rilevante non risiede quindi nel modello linguistico considerato isolatamente, ma nell'architettura complessiva che ne orienta l'uso, ne vincola il comportamento e ne contestualizza le risposte.

Per questa ragione, i meccanismi di orchestrazione multi-agente, i guardrails e le componenti simboliche introdotte nel corso della ricerca non sono finalizzati a rendere il sistema "moralmente corretto" in senso assoluto, bensì a strutturarne come dispositivo deliberativo controllato, interpretabile e coerente con il ruolo di mediazione cognitiva assegnato all'agente.

Tali delimitazioni non rappresentano carenze accidentali della ricerca, ma scelte metodologiche coerenti con l'approccio DSR adottato, il quale privilegia l'esplorazione della fattibilità architetturale e concettuale rispetto alla dimostrazione sperimentale dell'efficacia su utenti finali, una scelta necessaria per preservare proporzione tra obiettivi dichiarati, strumenti impiegati e risultati effettivamente conseguibili.

8.4 Limiti strutturali e criticità emerse

Accanto ai risultati ottenuti, il percorso di ricerca ha evidenziato una serie di limiti strutturali e criticità operative che riguardano sia le tecnologie impiegate sia la natura stessa del problema affrontato. Tali elementi non invalidano il contributo della tesi, ma ne definiscono più precisamente il perimetro di validità e ne chiariscono le principali aree di fragilità.

Nel loro insieme, queste criticità mostrano come la realizzazione di agenti morali artificiali affidabili, situati e generalizzabili richieda ancora significativi avanzamenti sul piano architetturale, computazionale, metodologico ed empirico. Alcuni dei limiti evidenziati non dipendono esclusivamente dalle implementazioni sviluppate, ma riflettono questioni ancora aperte nello stato dell'arte dell'intelligenza artificiale contemporanea, quali grounding semantico, robustezza inferenziale, allineamento valoriale e spiegabilità dei modelli generativi.

Latenza e fluidità conversazionale. L'architettura multi-agente basata su CrewAI e i processi di ragionamento profondo di DeepSeek-R1 introducono una latenza significativa, osservata nell'ordine di circa 2–3,5 secondi per risposta. Sebbene tale intervallo possa risultare accettabile in contesti di deliberazione etica complessa, esso compromette la fluidità di una conversazione naturale in ambienti immersivi, nei quali il senso di presenza dipende in larga misura dalla reattività sincrona dell'interazione. Il bilanciamento tra profondità deliberativa e rapidità di risposta rimane pertanto una delle principali sfide tecniche aperte.

Sensibilità contestuale e bias dei dati. Sebbene CASPER-MV contribuisca a migliorare il grounding percettivo dell'agente, il sistema rimane vulnerabile ai bias presenti nei dataset morali e nei modelli linguistici utilizzati. Nel corso delle sperimentazioni è emerso come anche variazioni descrittive relativamente minime degli scenari possano produrre valutazioni morali divergenti, incoerenti o eccessivamente sensibili alla formulazione del contesto. Questo aspetto evidenzia un limite strutturale dei sistemi attuali: la capacità di produrre risposte plausibili non coincide ancora con una comprensione causale o semanticamente profonda del problema morale affrontato. Ne consegue che tali agenti non possono essere considerati fonti autonome di giudizio, ma devono essere progettati come strumenti di supporto sotto supervisione umana critica.

Espressività incarnata e *uncanny valley*. L'uso di tecniche di creazione rapida di avatar, come Avaturn, ha evidenziato limiti rilevanti sul piano dell'incarnazione espressiva. La

standardizzazione delle mesh, la ridotta variabilità somatica e la difficoltà nel riprodurre micro-espressioni facciali e segnali paralinguistici possono generare una percezione di artificialità o alienazione nell'utente, riconducibile al fenomeno della *uncanny valley*. Questo elemento non rappresenta un semplice limite estetico, ma un problema funzionale: nel caso di un mediatore morale, la capacità di trasmettere empatia, credibilità e prossimità relazionale è parte integrante dell'efficacia comunicativa del sistema.

Limite empirico della validazione. Un ulteriore limite strutturale della ricerca riguarda l'assenza di una validazione empirica quantitativa su campioni di utenti esterni. Sebbene tale scelta sia coerente con l'impostazione metodologica adottata, essa implica che le valutazioni effettuate rimangano circoscritte al piano tecnico e qualitativo, con particolare riferimento a coerenza architeturale, stabilità operativa, integrazione dei moduli e plausibilità deliberativa. Di conseguenza, la tesi apre la strada alla possibilità di progettare e implementare un agente morale deliberativo situato, ma non consente di misurarne empiricamente l'efficacia nel modificare in modo stabile il comportamento o la riflessione morale degli utenti.

Nel loro insieme, questi limiti confermano che il valore del lavoro risiede principalmente nella formalizzazione architeturale e nella dimostrazione di fattibilità del paradigma proposto, più che nella verifica definitiva della sua efficacia applicativa in contesti d'uso reali.

8.5 Prospettive future: verso una convergenza tra simulazione e incarnazione fisica

Le prospettive di sviluppo emerse dalla presente ricerca si articolano lungo tre direttrici principali: la validazione empirica dei sistemi proposti, l'ottimizzazione architeturale delle soluzioni sviluppate e l'estensione del paradigma del mediatore morale artificiale oltre il solo dominio virtuale, verso scenari di incarnazione fisica e computazione ubiquitaria.

Tali prospettive devono essere intese come traiettorie di ricerca progressive e condizionate dall'evoluzione tecnica, normativa e sociale dei sistemi intelligenti, piuttosto che come esiti inevitabili o immediatamente disponibili.

Validazione empirica. Una naturale prosecuzione del presente lavoro consiste nella progettazione di studi empirici controllati con utenti reali, finalizzati a valutare non soltanto la fattibilità tecnica del mediatore morale artificiale, ma anche la sua efficacia percepita, la sua accettabilità sociale e il suo possibile impatto sui processi decisionali dell'utente.

Tali studi potrebbero coinvolgere differenti categorie di partecipanti — ad esempio studenti, utenti di ambienti social VR, ricercatori in contesti collaborativi immersivi o soggetti coinvolti in scenari educativi e formativi — e prevedere un confronto tra condizioni sperimentali con e senza la presenza dell'agente morale.

Tra gli *outcome* rilevanti potrebbero rientrare: la qualità percepita del supporto morale, il livello di fiducia nell'agente, la trasparenza del ragionamento, il grado di accettazione

sociale del sistema, l'eventuale influenza sulle scelte in scenari eticamente ambigui e la persistenza di eventuali effetti riflessivi nel tempo. Dal punto di vista metodologico, tali studi potrebbero integrare misure quantitative (questionari, tempi decisionali, pattern di scelta, scale di fiducia e accettazione) e qualitative (interviste, osservazione, analisi del dialogo, protocolli *think-aloud*), offrendo una base più solida per una futura valutazione dell'efficacia applicativa.

Ottimizzazione architetturale. Una seconda direttrice di sviluppo riguarda il perfezionamento delle architetture proposte sul piano tecnico e computazionale. In particolare, particolare attenzione andrà dedicata alla riduzione della latenza, alla gestione più efficiente delle pipeline multi-modulo e alla formalizzazione di metriche maggiormente oggettivabili per la valutazione della qualità deliberativa, della pertinenza contestuale e della trasparenza argomentativa. In questa prospettiva, l'evoluzione di modelli più compatti, specializzati o eseguibili localmente potrebbe contribuire a rendere tali sistemi più sostenibili e adatti a contesti interattivi real-time.

Convergenza tra ambienti virtuali e robotica incarnata. La traiettoria forse più significativa sul piano teorico e applicativo riguarda l'estensione del paradigma del mediatore morale artificiale dal dominio immersivo virtuale a quello della robotica sociale fisica e dell'*ubiquitous computing*. In tale prospettiva, gli ambienti virtuali sviluppati nel corso della ricerca possono essere interpretati non solo come spazi di interazione, ma anche come ambienti di simulazione utili per il trasferimento verso sistemi incarnati.

CASPER come ponte Sim-to-Real L'architettura CASPER, sviluppata originariamente in relazione a problemi di robotica assistiva, rappresenta in questo senso un potenziale ponte metodologico e tecnologico tra simulazione immersiva e applicazione fisica, secondo il paradigma *Sim-to-Real*. Ambienti virtuali persistenti e controllati possono infatti fungere da laboratori sicuri per l'addestramento, il test e il raffinamento di agenti morali situati, consentendo di esplorare scenari sociali ed etici complessi prima di una possibile implementazione su hardware fisico. Ciò potrebbe contribuire a ridurre costi, tempi di sviluppo e rischi associati alla sperimentazione diretta in contesti reali.

Agenti morali ubiquitari e robotica assistiva In prospettiva, gli agenti morali artificiali potrebbero evolvere in sistemi distribuiti e ubiquitari, capaci di operare in continuità tra differenti dispositivi e contesti d'uso — ad esempio smartphone, occhiali AR, ambienti immersivi e robot domestici [62]. In tale configurazione, un alter ego morale potrebbe accompagnare l'utente in attività educative, collaborative o assistive, mantenendo continuità di memoria, stile comunicativo e supporto deliberativo. Nel campo della robotica assistiva, tali agenti potrebbero in futuro essere esplorati come strumenti di supporto relazionale e cognitivo per anziani o pazienti fragili, purché il loro impiego sia progettato entro rigorosi vincoli di dignità, privacy, trasparenza e supervisione umana [100].

Nuove frontiere etiche: corporeità e persistenza digitale L'estensione del paradigma verso forme di IA incarnata o persistente apre inoltre nuove questioni etiche e giuridiche. La prossimità fisica di un agente capace di fornire orientamento morale introduce infatti dinamiche di influenza, fiducia e autorità differenti rispetto a quelle osservabili in un'interazione puramente testuale o virtuale [71]. Ciò rende necessario approfondire temi quali la responsabilità civile e progettuale, il consenso informato in interazioni prolungate, i limiti dell'autonomia operativa degli agenti e la definizione di forme appropriate di controllo umano [147].

Un ulteriore ambito di ricerca riguarda infine la persistenza dell'identità artificiale nel tempo, specialmente nel caso di alter ego digitali o *deadbots*. In tali scenari, la ricerca futura dovrà interrogarsi non solo sulle possibilità tecniche di continuità identitaria, ma anche sulle implicazioni psicologiche, relazionali e morali della permanenza o del ritiro di tali entità artificiali.

In conclusione, la presente tesi mostra la fattibilità tecnica e concettuale della progettazione di mediatori morali artificiali situati in ambienti immersivi, ma evidenzia al tempo stesso la necessità di ulteriori sviluppi empirici, architetturali ed etici. Più che offrire una soluzione definitiva, il lavoro propone un paradigma di ricerca: quello di sistemi artificiali progettati non per sostituire il giudizio morale umano, ma per accompagnarlo, renderlo più riflessivo e sostenerlo in contesti tecnologicamente sempre più complessi.

In un contesto storico segnato dalla crescente delega di funzioni cognitive alle tecnologie digitali, la questione centrale non riguarda se costruire sistemi sempre più intelligenti, ma come progettargli affinché rimangano strumenti al servizio dell'autonomia, della responsabilità e della dignità umana. In questa prospettiva, gli agenti morali artificiali rappresentano meno una risposta conclusiva che un campo di sperimentazione sul rapporto futuro tra intelligenza umana e intelligenza artificiale.

Bibliografia

- [1] Avaturn | Realistic 3D avatar creator, . URL <https://avaturn.me/>.
- [2] Convai - Conversational AI for Virtual Worlds, . URL <https://www.convai.com/>.
- [3] Crewai | Opensource, . URL <https://www.crewai.com/open-source>.
- [4] LMNT - Fast Lifelike Affordable AI Speech, . URL <https://www.lmnt.com/>.
- [5] Hugging Face – The AI community building the future., Aug. 2025. URL <https://huggingface.co/>.
- [6] K. Almeman, F. El Ayeb, M. Berrima, B. Issaoui, and H. Morsy. The Integration of AI and Metaverse in Education: A Systematic Literature Review. *Applied Sciences*, 15(2):863, Jan. 2025. ISSN 2076-3417. doi: 10.3390/app15020863.
- [7] H. Aswat, E. Purwanta, M. N. Wangid, A. Manaf, and H. Hayadin. Global Research Trends of Cyberbullying and the Metaverse in the Virtual World: Bibliometric Analysis Using the Scopus Database. *International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE)*, 13(1):157–164, Apr. 2025. ISSN 2334-8496. doi: 10.23947/2334-8496-2025-13-1-157-164.
- [8] B. Atasoy and O. Türkay. Traveling in the digital world:: How does Instagram Influences Young People’s Tourism Engagement and Intention in Terms of Social Cognitive Learning Theory? *Revista Rosa dos Ventos - Turismo e Hospitalidade*, 16 (4), Dec. 2024. ISSN 2178-9061. doi: 10.18226/21789061.v16i4p970.
- [9] A. Augello, M. Gentile, and F. Dignum. Social practices for social driven conversations in serious games. In *International conference on games and learning alliance*, pages 100–110. Springer, 2015.
- [10] A. Augello, M. Gentile, and F. Dignum. Social Agents for Learning in Virtual Environments. In R. Bottino, J. Jeuring, and R. C. Veltkamp, editors, *Games and Learning Alliance*, volume 10056, pages 133–143. Springer International Publishing, Cham, 2016. ISBN 978-3-319-50181-9 978-3-319-50182-6. doi: 10.1007/978-3-319-50182-6_12. Series Title: Lecture Notes in Computer Science.

- [11] A. Augello, F. Dignum, M. Gentile, I. Infantino, U. Maniscalco, G. Pilato, and F. Vella. A social practice oriented signs detection for human-humanoid interaction. *Biologically Inspired Cognitive Architectures*, 25:8–16, 2018. doi: 10.1016/j.bica.2018.07.013.
- [12] Ayyappa Sajja, Dheerender Thakur, and Aditya Mehra. Integrating Generative AI into the Software Development Lifecycle: Impacts on Code Quality and Maintenance. *International Journal of Science and Research Archive*, 13(1):1952–1960, Oct. 2024. ISSN 25828185. doi: 10.30574/ijrsra.2024.13.1.1837.
- [13] C. Badea and G. Artus. Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents. In M. Bramer and F. Stahl, editors, *Artificial Intelligence XXXIX*, volume 13652, pages 124–137. Springer International Publishing, Cham, 2022. ISBN 978-3-031-21440-0 978-3-031-21441-7. doi: 10.1007/978-3-031-21441-7_9. Series Title: Lecture Notes in Computer Science.
- [14] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. URL <https://aclanthology.org/W05-0909.pdf>.
- [15] L. Banh and G. Strobel. Generative artificial intelligence. *Electronic Markets*, 33(1): 63, Dec. 2023. ISSN 1019-6781, 1422-8890. doi: 10.1007/s12525-023-00680-1.
- [16] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016.
- [17] A. Beacco, J. Gallego, and M. Slater. Automatic 3D avatar generation from a single RGB frontal image. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 764–765. IEEE, 2022. URL <https://ieeexplore.ieee.org/abstract/document/9757504/>.
- [18] M. Beetz, M. Tenorth, and J. Winkler. Open-EASE. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1983–1990, Seattle, WA, USA, May 2015. IEEE. ISBN 978-1-4799-6923-4. doi: 10.1109/ICRA.2015.7139458.
- [19] S. Bhaumik. On the ethical and moral dimensions of using artificial intelligence for evidence synthesis. *PLOS Global Public Health*, 5(3):e0004348, Mar. 2025. ISSN 2767-3375. doi: 10.1371/journal.pgph.0004348.
- [20] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

- [21] J.-F. Bonnefon, I. Rahwan, and A. Shariff. The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology*, 75:653–675, 2024. doi: 10.1146/annurev-psych-030123-113559.
- [22] P. Bremner, L. A. Dennis, M. Fisher, and A. F. Winfield. On proactive, transparent, and verifiable ethical reasoning for robots. *Proceedings of the IEEE*, 107(3):541–561, 2019.
- [23] R. Brindha, R. K. Pongiannan, A. Bharath, and V. K. S. M. Sanjeevi. Introduction to Multimodal Generative AI. In A. Singh and K. K. Singh, editors, *Multimodal Generative AI*, pages 1–36. Springer Nature, Singapore, 2025. ISBN 978-981-96-2355-6. doi: 10.1007/978-981-96-2355-6_1.
- [24] C. Cao, T. Simon, J. K. Kim, G. Schwartz, M. Zollhoefer, S.-S. Saito, S. Lombardi, S.-E. Wei, D. Belko, S.-I. Yu, Y. Sheikh, and J. Saragih. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics*, 41(4):1–19, July 2022. ISSN 0730-0301, 1557-7368. doi: 10.1145/3528223.3530143.
- [25] M. L. Cappuccio, A. Peeters, and W. McDonald. Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology*, 33(1):9–31, 2020.
- [26] M. L. Cappuccio, E. B. Sandoval, O. Mubin, M. Obaid, and M. Velonaki. Can robots make us better humans? Virtuous robotics and the good life with artificial agents. *International Journal of Social Robotics*, 13(1):7–22, 2021.
- [27] M. L. Cappuccio, J. C. Galliot, and F. S. Alnajjar. A Taste of Armageddon: A Virtue Ethics Perspective on Autonomous Weapons and Moral Injury. *Journal of Military Ethics*, 21(1):19–38, Jan. 2022. ISSN 1502-7570, 1502-7589. doi: 10.1080/15027570.2022.2063103.
- [28] M. L. Cappuccio, J. C. Galliot, F. Eyssel, and A. Lanteri. Autonomous Systems and Technology Resistance: New Tools for Monitoring Acceptance, Trust, and Tolerance. *International Journal of Social Robotics*, 16(6):1–25, June 2024. ISSN 1875-4791, 1875-4805. doi: 10.1007/s12369-023-01065-2.
- [29] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos. Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*, 26(2):501–532, Apr. 2020. ISSN 1353-3452, 1471-5546. doi: 10.1007/s11948-019-00151-x.
- [30] V. Chamola, M. S. Peelam, U. Mittal, V. Hassija, A. Singh, R. Pareek, P. Mangal, D. Sangwan, V. H. C. de Albuquerque, M. Mahmud, and D. J. Brown. Metaverse for Education: Developments, Challenges, and Future Direction. *Computer Applications*

- in Engineering Education*, 33(3):e70018, 2025. ISSN 1099-0542. doi: 10.1002/cae.70018.
- [31] X. Chang. Gender bias in hiring: An analysis of the impact of Amazon’s recruiting algorithm. *Advances in Economics, Management and Political Sciences*, 23(1):134–140, 2023. URL <https://pdfs.semanticscholar.org/4b05/657a3abe9f7f976a75e0fada85476823dca4.pdf>.
- [32] G. Chen, X. Song, H. Zeng, and S. Jiang. Scene Recognition With Prototype-Agnostic Scene Layout. *IEEE Transactions on Image Processing*, 29:5877–5888, 2020. ISSN 1941-0042. doi: 10.1109/TIP.2020.2986599.
- [33] Z. Chen and S. Moscholios. Using Prompts to Guide Large Language Models in Imitating a Real Person’s Language Style, Oct. 2024. arXiv:2410.03848 [cs].
- [34] A. Clodic, J. Vazquez-Salceda, F. Dignum, S. Mascarenhas, V. Dignum, A. Augello, M. Gentile, and R. Alami. On the Pertinence of Social Practices for Social Robotics. In *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2018. doi: 10.3233/978-1-61499-931-7-63.
- [35] L. M. Csepregi. The effect of context-aware llm-based npc conversations on player engagement in role-playing video games. *Unpublished manuscript*, 2, 2021.
- [36] J. Cui. Cognitive software architectures for multimodal perception and human-ai interaction. *Preprints*. doi, 10, 2025.
- [37] R. Dale. GPT-3: What’s it good for? *Natural Language Engineering*, 27(1):113–118, Jan. 2021. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324920000601.
- [38] A. de Marchena and I.-M. Eigsti. Conversational gestures in autism spectrum disorders: Asynchrony but not decreased frequency. *Autism Research*, 3(6):311–322, 2010. ISSN 1939-3806. doi: 10.1002/aur.159.
- [39] T. De Mauro. Il Nuovo vocabolario di base della lingua italiana. *Internazionale*. [28/11/2020]. <https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana>, 2016.
- [40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

- [41] D. B. Dhiman. Exploding AI-Generated Deepfakes and Misinformation: A Threat to Global Concern in the 21st Century. *Qeios*, Dec. 2023. ISSN 2632-3834. doi: 10.32388/DPLE2L. URL <https://www.qeios.com/read/DPLE2L>.
- [42] R. Dhiman and N. S. Bhati. When AI meets influencers: Role of AI-powered instagram influencers driving consumers' purchase intentions: An integration of SMIV & SOR framework. *Journal of Open Innovation: Technology, Market, and Complexity*, page 100591, 2025. Publisher: Elsevier.
- [43] F. Dignum. Social Practices: a Complete Formalization, May 2022. URL <http://arxiv.org/abs/2206.06088>. arXiv:2206.06088 [cs].
- [44] V. Dignum, F. Dignum, J. Vazquez-Salceda, A. Clodic, M. Gentile, S. Mascarenhas, and A. Augello. Design for Values for Social Robot Architectures. In *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2018. doi: 10.3233/978-1-61499-931-7-43.
- [45] Z. Dong, C. Guo, J. Song, X. Chen, A. Geiger, and O. Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20470–20480, 2022. URL http://openaccess.thecvf.com/content/CVPR2022/html/Dong_PINA_Learning_a_Personalized_Implicit_Neural_Avatar_From_a_Single_CVPR_2022_paper.html.
- [46] Dr. Irum Fazal, Dr. Marium Minhas Bandiali, Fahim Shezad, and Hassam Gul. Bridging Educational Gaps: The Role of AI and Social Media in Enhancing Access to Quality Education in Under-privileged Communities. *The Critical Review of Social Sciences Studies*, 3(1):2413–2431, Feb. 2025. ISSN 3006-7170, 3006-7162. doi: 10.59075/r3dphx69.
- [47] M. Dragone, T. Holz, and G. M. O'Hare. Mixing robotic realities. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 261–263, Sydney Australia, Jan. 2006. ACM. ISBN 978-1-59593-287-7. doi: 10.1145/1111449.1111504.
- [48] J. Dudley, L. Yin, V. Garaj, and P. O. Kristensson. Inclusive Immersion: a review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse. *Virtual Reality*, 27(4):2989–3020, Dec. 2023. ISSN 1434-9957. doi: 10.1007/s10055-023-00850-8.
- [49] P. Dumeva, Y. Ma, G. M. Marakas, and M. I. Aguirre-Urreta. Virtual Presence in Immersive Metaverse-enabling Environments: Being There, Being with Another, and Being the Actual Self. *AIS Transactions on Human-Computer Interaction*, 16(4): 481–516, Dec. 2024. ISSN 19443900. doi: 10.17705/1thci.00214.

- [50] Y. K. Dwivedi, L. Hughes, A. M. Baabdullah, S. Ribeiro-Navarrete, M. Giannakis, M. M. Al-Debei, D. Dennehy, B. Metri, D. Buhalis, C. M. Cheung, K. Conboy, R. Doyle, R. Dubey, V. Dutot, R. Felix, D. Goyal, A. Gustafsson, C. Hinsch, I. Jebabli, M. Janssen, Y.-G. Kim, J. Kim, S. Koos, D. Kreps, N. Kshetri, V. Kumar, K.-B. Ooi, S. Papagiannidis, I. O. Pappas, A. Polyviou, S.-M. Park, N. Pandey, M. M. Queiroz, R. Raman, P. A. Rauschnabel, A. Shirish, M. Sigala, K. Spanaki, G. Wei-Han Tan, M. K. Tiwari, G. Viglia, and S. F. Wamba. Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 66: 102542, Oct. 2022. ISSN 02684012. doi: 10.1016/j.ijinfomgt.2022.102542.
- [51] D. Edmonds and D. Edmonds, editors. *AI Morality*. Oxford University Press, Oxford, New York, Aug. 2024. ISBN 978-0-19-887643-4.
- [52] J. P. Egunjobi. The Misuse of AI-Generated Content in Academic and Religious Settings. *International Journal of Research and Scientific Innovation (IJRSI)*, XII (XV):871–879, 2025. ISSN 2321-2705. doi: 10.51244/IJRSI.2025.121500077P.
- [53] N. Elsokkary, W. Khan, M. Shurrab, R. Mizouni, S. Singh, J. Bentahar, A. Mourad, and H. Otrok. Reinforcement learning and the Metaverse: a symbiotic collaboration. *Artificial Intelligence Review*, 59(1):34, Dec. 2025. ISSN 1573-7462. doi: 10.1007/s10462-025-11433-1.
- [54] D. Emelin, R. Le Bras, J. D. Hwang, M. Forbes, and Y. Choi. Moral Stories: Situated Reasoning about Norms, Intentions, Actions, and their Consequences. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.54.
- [55] P. Eslami, M. Najafabadi, and A. Gharehgozli. Exploring the journey of influencers in shaping social media engagement success. *Online Social Networks and Media*, 41: 100277, May 2024. ISSN 2468-6964. doi: 10.1016/j.osnem.2024.100277.
- [56] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech. Generative AI. *Business & Information Systems Engineering*, 66(1):111–126, Feb. 2024. ISSN 2363-7005, 1867-0202. doi: 10.1007/s12599-023-00834-7.
- [57] E. Firt. What makes full artificial agents morally different. *AI & SOCIETY*, 40(1):175–184, Jan. 2025. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-024-01867-6.
- [58] S. Franklin, T. Madl, S. D’Mello, and J. Snaider. LIDA: A Systems-level Architecture for Cognition, Emotion, and Learning. *IEEE Transactions on Autonomous Mental Development*, 6(1):19–41, Mar. 2014. ISSN 1943-0604, 1943-0612. doi: 10.1109/TAMD.2013.2277589.

- [59] G. Freeman and D. Acena. Hugging from A Distance: Building Interpersonal Relationships in Social Virtual Reality. In *ACM International Conference on Interactive Media Experiences*, pages 84–95, Virtual Event USA, June 2021. ACM. ISBN 978-1-4503-8389-9. doi: 10.1145/3452918.3458805.
- [60] L. Frering, G. Steinbauer-Wagner, and A. Holzinger. Integrating Belief-Desire-Intention agents with large language models for reliable human–robot interaction and explainable Artificial Intelligence. *Engineering Applications of Artificial Intelligence*, 141:109771, 2025. URL <https://www.sciencedirect.com/science/article/pii/S0952197624019304>.
- [61] L. Friedman, S. Ahuja, D. Allen, Z. Tan, H. Sidahmed, C. Long, J. Xie, G. Schubiner, A. Patel, H. Lara, B. Chu, Z. Chen, and M. Tiwari. Leveraging Large Language Models in Conversational Recommender Systems, May 2023. URL <http://arxiv.org/abs/2305.07961>. arXiv:2305.07961 [cs].
- [62] P. Fung, Y. Bachrach, A. Celikyilmaz, K. Chaudhuri, D. Chen, W. Chung, E. Dupoux, H. Gong, H. Jégou, A. Lazaric, A. Majumdar, A. Madotto, F. Meier, F. Metze, L.-P. Morency, T. Moutakanni, J. Pino, B. Terver, J. Tighe, P. Tomasello, and J. Malik. Embodied AI Agents: Modeling the World, July 2025. URL <http://arxiv.org/abs/2506.22355>. arXiv:2506.22355 [cs].
- [63] G. F. Gaglio. `gFulvio/moralstories-roberta-action.context-cls` · Hugging Face, July 2022. URL <https://huggingface.co/gFulvio/moralstories-roberta-action.context-cls>.
- [64] G. F. Gaglio. `gFulvio/moralstories-bart-consequences.context-action_gen` · Hugging Face, July 2022. URL https://huggingface.co/gFulvio/moralstories-bart-consequences.context-action_gen.
- [65] G. F. Gaglio. `gFulvio/moralstories-bart-norm.actions-context-consequences_gen` · Hugging Face, July 2022. URL https://huggingface.co/gFulvio/moralstories-bart-norm.actions-context-consequences_gen.
- [66] G. F. Gaglio, A. Augello, A. Pipitone, L. Gallo, R. Sorbello, and A. Chella. Moral Mediators in the Metaverse: Exploring Artificial Morality through a Talking Cricket Paradigm. In *AIxPAC*, pages 30–43, 2022.
- [67] G. F. Gaglio, S. Vinanzi, A. Cangelosi, and A. Chella. Intention Reading Architecture for Virtual Agents. In O. Palinko, L. Bodenhausen, J.-J. Cabibihan, K. Fischer, S. Šabanović, K. Winkle, L. Behera, S. S. Ge, D. Chrysostomou, W. Jiang, and H. He, editors, *Social Robotics*, volume 15561, pages 488–497. Springer Nature Singapore, Singapore, 2025. ISBN 978-981-96-3521-4 978-981-96-3522-1. doi: 10.1007/978-981-96-3522-1_41. Series Title: Lecture Notes in Computer Science.

- [68] Y. Gatsoulis, M. Alomari, C. Burbridge, C. Dondrup, P. Duckworth, P. Lightbody, M. Hanheide, N. Hawes, D. C. Hogg, and A. G. Cohn. QSRLib: a software library for online acquisition of Qualitative Spatial Relations from Video. In *Workshop on qualitative reasoning (QRI6)*, 2016.
- [69] L. Gatto, G. F. Gaglio, A. Augello, G. Caggianese, L. Gallo, and M. La Cascia. MET-iquette: enabling virtual agents to have a social compliant behavior in the Metaverse. In *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 394–401, Oct. 2022. doi: 10.1109/SITIS57111.2022.00066.
- [70] F. Ghazali, T. BaniRostam, and M. Pedram. Developing Artificial Moral Agents: Key Research Processes, Techniques, and Challenges. *AI and Tech in Behavioral and Social Sciences*, 3(1):92–108, 2025. doi: 10.61838/kman.aitech.3.1.9.
- [71] A. Ghoshal, M. Brandão, R. Abu-Salma, and S. Modgil. Embodied AI at the Margins: Postcolonial Ethics for Intelligent Robotic Systems. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1120–1133, Oct. 2025. ISSN 3065-8365. doi: 10.1609/aies.v8i2.36615.
- [72] J.-S. Gordon and S. Nyholm. Ethics of Artificial Intelligence. URL <https://iep.utm.edu/ethics-of-artificial-intelligence/>.
- [73] S. Greenwold. Spatial computing. *Massachusetts Institute of Technology, Master*, 2003.
- [74] D. Griol, A. Sanchis, J. M. Molina, and Z. Callejas. Developing enhanced conversational agents for social virtual worlds. *Neurocomputing*, 354:27–40, Aug. 2019. ISSN 0925-2312. doi: 10.1016/j.neucom.2018.09.099.
- [75] J. Guan, J. Irizawa, and A. Morris. Extended Reality and Internet of Things for Hyper-Connected Metaverse Environments. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 163–168, Mar. 2022. doi: 10.1109/VRW55335.2022.00043.
- [76] R. Gusmani and R. Bombi. Itinerari linguistici: scritti raccolti in occasione del 60. compleanno. (*No Title*), 1995. URL <https://cir.nii.ac.jp/crid/1130282269855667072>.
- [77] D. Han, T. McInroe, A. Jelley, S. V. Albrecht, P. Bell, and A. Storkey. Llm-personalize: Aligning llm planners with human preferences via reinforced self-training for housekeeping robots. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1465–1474, 2025. URL <https://aclanthology.org/2025.coling-main.98/>.

- [78] H. Han, J. Kim, C. Jeong, and G. L. Cohen. Attainable and Relevant Moral Exemplars Are More Effective than Extraordinary Exemplars in Promoting Voluntary Service Engagement. *Frontiers in Psychology*, 8:283, Mar. 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.00283.
- [79] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI With Shared Human Values. In *International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2008.02275.
- [80] T. Hennig-Thurau, D. Aliman, A. Herting, G. Cziehso, M. Linder, and R. Kübler. Social interactions in the metaverse: Framework, initial evidence, and research roadmap. *Journal of the Academy of Marketing Science*, 51, Dec. 2022. doi: 10.1007/s11747-022-00908-0.
- [81] P. Henz. The psychological impact of the Metaverse. *Discover Psychology*, 2(1):47, Dec. 2022. ISSN 2731-4537. doi: 10.1007/s44202-022-00061-3.
- [82] I. Hettiarachchi. Exploring Generative AI Agents: Architecture, Applications, and Challenges. *Journal of Artificial Intelligence General Science*, 08(1), 2025.
- [83] S. Higgins, S. Alcock, B. D. Aveiro, W. Daniels, H. Farmer, and S. Besharati. Perspective matters: a systematic review of immersive virtual reality to reduce racial prejudice. *Virtual Reality*, 28(3):125, June 2024. ISSN 1434-9957. doi: 10.1007/s10055-024-01024-w.
- [84] R. Hooi and H. Cho. Deception in avatar-mediated virtual environment. *Computers in Human Behavior*, 29(1):276–284, Jan. 2013. ISSN 0747-5632. doi: 10.1016/j.chb.2012.09.004.
- [85] N. Hossain, M. Ghazvininejad, and L. Zettlemoyer. Simple and effective retrieve-edit-rerank text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2532–2538, 2020. URL <https://aclanthology.org/2020.acl-main.228/>.
- [86] J. Hutson. Social Virtual Reality: Neurodivergence and Inclusivity in the Metaverse. *Societies*, 12(4):102, July 2022. ISSN 2075-4698. doi: 10.3390/soc12040102.
- [87] T. Huynh-The, Q.-V. Pham, X.-Q. Pham, T. T. Nguyen, Z. Han, and D.-S. Kim. Artificial Intelligence for the Metaverse: A Survey, Feb. 2022. URL <http://arxiv.org/abs/2202.10336>. arXiv:2202.10336 [cs].
- [88] S. HWANG, T. Seo, Y. Kim, and H. Oh. Metaverse Counseling: Transforming Depression Care with AI Digital Humans in Virtual Realms. 2024. URL <https://www.researchsquare.com/article/rs-4244666/latest>.

- [89] K. A. Islam, W. Khan, M. F. Bari, R. Mostafa, F. Anonhi, and N. A. Monira. Challenges of Artificial Intelligence for the Metaverse: A Scoping Review. *International Research Journal of Multidisciplinary Scope*, 06(01):1094–1101, 2025. ISSN 2582631X. doi: 10.47857/irjms.2025.v06i01.02493.
- [90] S. K. Jagatheesaperumal and M. Rahouti. Building Digital Twins of Cyber Physical Systems With Metaverse for Industry 5.0 and Beyond. *IT Professional*, 24(6):34–40, Nov. 2022. ISSN 1941-045X. doi: 10.1109/MITP.2022.3225064.
- [91] Z. Jiang, L. Zhao, S. Li, and Y. Jia. Real-time object detection method based on improved YOLOv4-tiny, Dec. 2020. URL <http://arxiv.org/abs/2011.04244>. arXiv:2011.04244 [cs].
- [92] L. Jiabin and G. Gongjing. Socializing in the Metaverse: The Innovation and Challenge of Interpersonal Communication. pages 2128–2131. Atlantis Press, June 2022. ISBN 978-94-6239-580-0. doi: 10.2991/assehr.k.220504.385. URL <https://www.atlantis-press.com/proceedings/ichssr-22/125974850>. ISSN: 2352-5398.
- [93] J. Jin. AI agent innovates: Pushing the boundaries of Generative Tech. Technical report, Fujitsu, 2025.
- [94] S. V. Jin and V. Viswanathan. “Threatened and empty selves following AI-based virtual influencers”: comparison between followers and non-followers of virtual influencers in AI-driven digital marketing. *AI & SOCIETY*, 40(1):117–131, Jan. 2025. ISSN 1435-5655. doi: 10.1007/s00146-023-01832-9.
- [95] J. Joosten, V. Bilgram, A. Hahn, and D. Totzek. Comparing the Ideation Quality of Humans With Generative Artificial Intelligence. *IEEE Engineering Management Review*, 52(2):153–164, Apr. 2024. ISSN 1937-4178. doi: 10.1109/EMR.2024.3353338. URL <https://ieeexplore.ieee.org/document/10398283>.
- [96] H. Joshi and V. Ustun. Augmenting Cognitive Architectures with Large Language Models. *Proceedings of the AAAI Symposium Series*, 2(1):281–285, 2023. ISSN 2994-4317. doi: 10.1609/aaais.v2i1.27689. URL <https://ojs.aaai.org/index.php/AAAI-SS/article/view/27689>.
- [97] A. Juliani, V.-P. Berges, E. Teng, A. Cohen, J. Harper, C. Elion, C. Goy, Y. Gao, H. Henry, M. Mattar, and D. Lange. Unity: A General Platform for Intelligent Agents, May 2020. URL <http://arxiv.org/abs/1809.02627>. arXiv:1809.02627 [cs, stat].
- [98] M. M. Kachman, I. Brennan, J. J. Oskvarek, T. Waseem, and J. M. Pines. How artificial intelligence could transform emergency care. *The American Journal of Emergency Medicine*, 81:40–46, July 2024. ISSN 0735-6757. doi: 10.1016/j.ajem.

2024.04.024. URL <https://www.sciencedirect.com/science/article/pii/S0735675724001815>.

- [99] H. Kang, M. B. Moussa, and N. Magnenat-Thalmann. Nadine: An LLM-driven Intelligent Social Robot with Affective Capabilities and Human-like Memory, May 2024. URL <http://arxiv.org/abs/2405.20189>. arXiv:2405.20189 [cs].
- [100] A. Kargin and T. Petrenko. Feeling Artificial Intelligence for AI-Enabled Autonomous Systems. In *2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, pages 88–93, Dec. 2022. doi: 10.1109/GCAIoT57150.2022.10019235. URL <https://ieeexplore.ieee.org/abstract/document/10019235>.
- [101] S. K. Kehinde. AI in everything, and everything in AI: A review of the ubiquitous role of artificial intelligence in shaping the next technological epoch. *Journal of Computer Allied Intelligence (JCAI, ISSN: 2584-2676)*, 3(5):17–53, 2025. URL <https://fringeglobal.com/ojs/index.php/jcai/article/view/ai-in-everything-and-everything-in-ai-a-review-of-the-ubiquitous>.
- [102] C. Kerdvibulvech and C.-C. Chang. A New Study of Integration Between Social Robotic Systems and the Metaverse for Dealing with Healthcare in the Post-COVID-19 Situations. In F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, and S. S. Ge, editors, *Social Robotics*, volume 13818, pages 392–401. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-24669-2 978-3-031-24670-8. doi: 10.1007/978-3-031-24670-8_35. Series Title: Lecture Notes in Computer Science.
- [103] S. Kim and E. Kim. Emergence of the Metaverse and Psychiatric Concerns in Children and Adolescents. *Journal of the Korean Academy of Child and Adolescent Psychiatry*, 34(4):215–221, Oct. 2023. ISSN 2233-9183. doi: 10.5765/jkacap.230047. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10568191/>.
- [104] Y. Kim and S. S. Sundar. Visualizing ideal self vs. actual self through avatars: Impact on preventive health outcomes. *Computers in Human Behavior*, 28(4):1356–1364, July 2012. ISSN 0747-5632. doi: 10.1016/j.chb.2012.02.021. URL <https://www.sciencedirect.com/science/article/pii/S0747563212000635>.
- [105] E. A. Konijn, J. F. Hoorn, E. A. Konijn, and J. F. Hoorn. Differential Facial Articulatory in Robots and Humans Elicit Different Levels of Responsiveness, Empathy, and Projected Feelings. *Robotics*, 9(4), Nov. 2020. ISSN 2218-6581. doi: 10.3390/robotics9040092. URL <https://www.mdpi.com/2218-6581/9/4/92>. Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher.

- [106] P. Kourtesis. A Comprehensive Review of Multimodal XR Applications, Risks, and Ethical Challenges in the Metaverse. *Multimodal Technologies and Interaction*, 8 (11):98, Nov. 2024. ISSN 2414-4088. doi: 10.3390/mti8110098. URL <https://www.mdpi.com/2414-4088/8/11/98>.
- [107] B. Labinghisa and D. M. Lee. A Deep Learning based Scene Recognition Algorithm for Indoor Localization. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pages 167–170, Apr. 2021. doi: 10.1109/ICAIIIC51459.2021.9415278. URL <https://ieeexplore.ieee.org/document/9415278/>.
- [108] P. Lam and V. Kiu. Exploring the Potential of Immersive Technology for Virtual Teaching and Learning: A Metaverse Conference Experience. *International Journal of Information and Education Technology*, 14(9):1185–1191, 2024. ISSN 20103689. doi: 10.18178/ijiet.2024.14.9.2147. URL <https://www.ijiet.org/show-209-2783-1.html>.
- [109] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou. AvatarMe: Realistically Renderable 3D Facial Reconstruction” in-the-wild”. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 760–769, 2020. URL http://openaccess.thecvf.com/content_CVPR_2020/html/Lattas_AvatarMe_Realistically_Renderable_3D_Facial_Reconstruction_In-the-Wild_CVPR_2020_paper.html.
- [110] D. Leben. Are Artificial Moral Agents the Future of Ethical AI?, 2025. URL <https://tepperspectives.cmu.edu/all-articles/are-artificial-moral-agents-the-future-of-ethical-ai/>.
- [111] G. I. Lee, S. H. Han, and Y.-H. Lee. Implementation of metaverse virtual world using unity game engine. *Journal of the Semiconductor & Display Technology*, 22(2): 120–127, 2023.
- [112] H. Lee, S. Phatale, H. Mansoor, K. R. Lu, T. Mesnard, J. Ferret, C. Bishop, E. Hall, V. Carbune, and A. Rastogi. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*, 2023.
- [113] J. Lee, T. S. Lee, S. Lee, J. Jang, S. Yoo, Y. Choi, and Y. R. Park. Development and Application of a Metaverse-Based Social Skills Training Program for Children With Autism Spectrum Disorder to Improve Social Interaction: Protocol for a Randomized Controlled Trial. *JMIR Research Protocols*, 11(6):e35960, June 2022. ISSN 1929-0748. doi: 10.2196/35960. URL <https://www.researchprotocols.org/2022/6/e35960>.
- [114] L.-H. Lee, P. Y. Zhou, T. Braud, and P. Hui. *What is the Metaverse? An Immersive Cyberspace and Open Challenges*. June 2022. doi: 10.48550/arXiv.2206.03018.

- [115] L.-H. Lee, T. Braud, P. Y. Zhou, L. Wang, D. Xu, Z. Lin, A. Kumar, C. Bermejo, and P. Hui. All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda. *Foundations and Trends in Human-Computer Interaction*, 18(2-3):100–337, Nov. 2024. ISSN 1551-3955. doi: 10.1561/11000000095. URL <https://doi.org/10.1561/11000000095>.
- [116] Y. K. Lee, Y. Jung, G. Kang, and S. Hahn. Developing Social Robots with Empathetic Non-Verbal Cues Using Large Language Models, Aug. 2023. URL <http://arxiv.org/abs/2308.16529>. arXiv:2308.16529 [cs].
- [117] A. Lieto, M. Bhatt, A. Oltramari, and D. Vernon. The role of cognitive architectures in general artificial intelligence, 2018. URL <https://www.sciencedirect.com/science/article/pii/S138904171730222X>. Pages: 1–3 Publication Title: Cognitive Systems Research Volume: 48.
- [118] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. URL <https://aclanthology.org/W04-1013.pdf>.
- [119] H. Lin, Z. Gan, W. Gan, Z. Qi, Y. Wang, and P. S. Yu. Interaction in Metaverse: A Survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2473–2482, Dec. 2023. doi: 10.1109/BigData59044.2023.10386876. URL <https://ieeexplore.ieee.org/document/10386876>.
- [120] Y. Litvinova, R. M. Rilke, and C. Guenther. Me, myself, and I: Image concerns and honesty in immersive VR. *Computers in Human Behavior*, 149:107950, Dec. 2023. ISSN 0747-5632. doi: 10.1016/j.chb.2023.107950. URL <https://www.sciencedirect.com/science/article/pii/S0747563223003011>.
- [121] R. Liu, X. Li, and S. Liu. How emotional expression in human-like virtual influencers drives user engagement: empathy model and its antecedents. *Frontiers in Psychology*, 16, Jan. 2026. ISSN 1664-1078. doi: 10.3389/fpsyg.2025.1544037. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2025.1544037/full>. Publisher: Frontiers.
- [122] Y. Liu. The Proteus Effect: Overview, Reflection, and Recommendations. *Games and Culture*, 20(3):384–400, May 2025. ISSN 1555-4120, 1555-4139. doi: 10.1177/15554120231202175.
- [123] Y. Liu and A. Moore. Intuitive judgements towards artificial intelligence verdicts of moral transgressions. *The British Journal of Social Psychology*, 64(3):e12908, July 2025. ISSN 0144-6665. doi: 10.1111/bjso.12908. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12125647/>.

- [124] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692 [cs].
- [125] S. Maghaydah, M. Al-Emran, P. Maheshwari, and M. A. Al-Sharafi. Factors affecting metaverse adoption in education: A systematic review, adoption framework, and future research agenda. *Heliyon*, 10(7), Apr. 2024. ISSN 2405-8440. doi: 10.1016/j.heliyon.2024.e28602. URL [https://www.cell.com/heliyon/abstract/S2405-8440\(24\)04633-4](https://www.cell.com/heliyon/abstract/S2405-8440(24)04633-4). Publisher: Elsevier.
- [126] V. N. Mandhala, D. Bhattacharyya, and D. Midhunchakkaravarthy. Need of Mitigating Bias in the Datasets using Machine Learning Algorithms. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–7, Jan. 2022. doi: 10.1109/ACCAI53970.2022.9752643. URL <https://ieeexplore.ieee.org/document/9752643>.
- [127] M. Marcuzzo, A. Zangari, A. Albarelli, J. Camacho-Collados, and M. T. Pilehvar. MORABLES: A Benchmark for Assessing Abstract Moral Reasoning in LLMs with Fables. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27715–27739, 2025. URL <https://aclanthology.org/2025.emnlp-main.1411/>.
- [128] J. Meese, B. Nansen, T. Kohn, M. Arnold, and M. Gibbs. Posthumous personhood and the affordances of digital media. *Mortality*, 20(4):408–420, Oct. 2015. ISSN 1357-6275. doi: 10.1080/13576275.2015.1083724. Publisher: Routledge eprint: <https://doi.org/10.1080/13576275.2015.1083724>.
- [129] M. L. Menéndez, J. A. Pardo, L. Pardo, and M. d. C. Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997. URL <https://www.sciencedirect.com/science/article/pii/S0016003296000634>.
- [130] C. Mishra, R. Verdonschot, P. Hagoort, and G. Skantze. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI*, 10:1271610, 2023. URL <https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2023.1271610/full>.
- [131] C. Misselhorn. Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In S. Voenekey, P. Kellmeyer, O. Mueller, and W. Burgard, editors, *The Cambridge Handbook of Responsible Artificial Intelligence*, pages 31–49. Cambridge University Press, 1 edition, Nov. 2022. ISBN 978-1-009-20789-8 978-1-009-20786-7. doi: 10.1017/9781009207898.005.
- [132] E. A. S. Mohamed, M. E. Osman, and B. A. Mohamed. The Impact of Artificial Intelligence on Social Media Content. *Journal of Social Sciences*, 20(1):12–16, Jan. 2024. ISSN 1549-3652. doi: 10.3844/jssp.2024.12.16.

- [133] J. Moorkens. Ethics and machine translation. *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 18:121, 2022. URL https://library.oapen.org/bitstream/handle/20.500.12657/61713/1/external_content.pdf#page=133.
- [134] K. Morrell and F. Dahlmann. Aristotle in the Anthropocene: The comparative benefits of Aristotelian virtue ethics over Utilitarianism and deontology. *The Anthropocene Review*, 10(3):615–635, Dec. 2023. ISSN 2053-0196, 2053-020X. doi: 10.1177/20530196221105093.
- [135] M. R. Morris and J. R. Brubaker. Generative Ghosts: Anticipating Benefits and Risks of AI Afterlives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–14, New York, NY, USA, Apr. 2025. Association for Computing Machinery. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713758.
- [136] S. Mystakidis. Metaverse. *Encyclopedia*, 2(1):486–497, Feb. 2022. ISSN 2673-8392. doi: 10.3390/encyclopedia2010031. URL <https://www.mdpi.com/2673-8392/2/1/31>.
- [137] M. Nasser, G. F. Gaglio, V. Seidita, and A. Chella. The Art of Replication: Life-like Avatars with Personalized Conversational Style. *Robotics*, 14(3):33, Mar. 2025. ISSN 2218-6581. doi: 10.3390/robotics14030033. URL <https://www.mdpi.com/2218-6581/14/3/33>.
- [138] D. T. K. Ng. What is the metaverse? Definitions, technologies and the community of inquiry. *Australasian Journal of Educational Technology*, 38(4):190–205, Nov. 2022. ISSN 1449-5554. doi: 10.14742/ajet.7945. URL <https://ajet.org.au/index.php/AJET/article/view/7945>. Number: 4.
- [139] E. O’Doherty, N. Weinrauch, A. Talone, U. Klempner, X. Yi, X. Xie, and Y. Zeng. The Morality of Probability: How Implicit Moral Biases in LLMs May Shape the Future of Human-AI Symbiosis, Sept. 2025. URL <http://arxiv.org/abs/2509.10297>. arXiv:2509.10297 [cs].
- [140] OkGoDoIt. Openai-api-dotnet, July 2023. URL <https://github.com/OkGoDoIt/OpenAI-API-dotnet>.
- [141] M. Okkerman and A. Enschede. *Holding Up a Black Mirror: exploring Ethical Issues Of The Metaverse From A User Perspective*. PhD thesis, University of Twente, 2023. URL https://essay.utwente.nl/95127/1/Okkerman_MA_BMS.pdf.
- [142] P. Ostheimer, M. Nagda, M. Kloft, and S. Fellenz. Text style transfer evaluation using large language models. In *Proceedings of the 2024 Joint International Conference*

on *Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15802–15822, 2024.

- [143] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. URL <https://aclanthology.org/P02-1040.Pdf>.
- [144] S. R. Pappula and S. R. Allam. Llms for conversational ai: Enhancing chatbots and virtual assistants. *International Journal of Research Publication and Reviews*, 4(12): 1601–1611, 2023.
- [145] J. S. Park, J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST ’23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL <https://doi.org/10.1145/3586183.3606763>.
- [146] S.-M. Park and Y.-G. Kim. A Metaverse: Taxonomy, Components, Applications, and Open Challenges. *IEEE Access*, 10:4209–4251, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3140175. URL <https://ieeexplore.ieee.org/document/9667507>.
- [147] J. Perlo, A. Robey, F. Barez, L. Floridi, and J. Mökander. Embodied AI: Emerging Risks and Opportunities for Policy Action, 2025. URL <https://arxiv.org/abs/2509.00117>. Version Number: 2.
- [148] C. Peukert, H. Qahri-Saremi, U. Schultze, J. B. Thatcher, C. M. K. Cheung, A. Frenzel-Piasentin, M. Greve, C. Matt, M. Trenz, and O. Turel. Metaverse: A real change or just another research area? *Electronic Markets*, 34(1):32, May 2024. ISSN 1422-8890. doi: 10.1007/s12525-024-00711-5. URL <https://doi.org/10.1007/s12525-024-00711-5>.
- [149] M. Proctor. *International symposium on applications of graph transformations with industrial relevance*, chapter Drools: a rule engine for complex event processing. Springer, 2011.
- [150] V. Pérez-Torres. Social media: a digital social mirror for identity development during adolescence. *Current Psychology*, 43(26):22170–22180, July 2024. ISSN 1936-4733. doi: 10.1007/s12144-024-05980-z. URL <https://doi.org/10.1007/s12144-024-05980-z>.

- [151] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text, Oct. 2016. URL <http://arxiv.org/abs/1606.05250>. arXiv:1606.05250 [cs].
- [152] A. S. Rao, A. Khandelwal, K. Tanmay, U. Agarwal, and M. Choudhury. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, 2023. URL <https://aclanthology.org/2023.findings-emnlp.892/>.
- [153] B. S. Rawal, S. Ahmadand, A. Mentges, and S. Fadli. Opportunities and Challenges in Metaverse the Rise of Digital Universe. In L.-J. Zhang, editor, *Metaverse – METAVERSE 2022*, pages 3–17, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-23518-4. doi: 10.1007/978-3-031-23518-4_1.
- [154] A. Reckwitz. Toward a Theory of Social Practices: A Development in Culturalist Theorizing. *European Journal of Social Theory*, 5(2):243–263, May 2002. ISSN 1368-4310. doi: 10.1177/13684310222225432. URL <https://doi.org/10.1177/13684310222225432>. Publisher: SAGE Publications Ltd.
- [155] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.690. URL <http://ieeexplore.ieee.org/document/8100173/>.
- [156] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Aug. 2019. URL <http://arxiv.org/abs/1908.10084>. arXiv:1908.10084 [cs].
- [157] H. Ritschel, K. Janowski, A. Seiderer, S. Wagner, and E. André. Insights on usability and user feedback for an assistive robotic health companion with adaptive linguistic style. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 319–320, Rhodes Greece, June 2019. ACM. ISBN 978-1-4503-6232-0. doi: 10.1145/3316782.3322737. URL <https://dl.acm.org/doi/10.1145/3316782.3322737>.
- [158] A. Rusnak and Z. Seals. EudAImonia: Virtue Ethics and Artificial Intelligence. *Christian Perspectives on Science and Technology*, 3, 2025. ISSN 26537656, 2653648X. doi: 10.58913/ZNHR8688. URL <https://journal.iscast.org/articles/eudaimonia-virtue-ethics-and-artificial-intelligence>.
- [159] T. Saheb, M. Sidaoui, and B. Schmarzo. Convergence of artificial intelligence with social media: A bibliometric & qualitative analysis. *Telematics and Informatics Reports*, 14:100146, 2024.

- [160] S. Sai, P. Sharma, A. Gaur, and V. Chamola. Pivotal role of digital twins in the metaverse: A review. *Digital Communications and Networks*, Dec. 2024. ISSN 2352-8648. doi: 10.1016/j.dcan.2024.12.003. URL <https://www.sciencedirect.com/science/article/pii/S2352864824001706>.
- [161] L. Salt, J. Wise, C. Sennersten, and C. A. Lindley. REACT-R and Unity Integration. In *The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, page 31, 2016.
- [162] E. Salvini. Enhancing autonomy of persons with autism through virtual reality. Master's thesis, Politecnico di Milano, 2021. URL <https://hdl.handle.net/10589/177805>.
- [163] G. Sarthou. Overworld: Assessing the Geometry of the World for Human-Robot Interaction. *IEEE Robotics and Automation Letters*, 8(3):1874–1880, Mar. 2023. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2023.3238891. URL <https://ieeexplore.ieee.org/document/10024303/>.
- [164] M. Saumer, A. Neureiter, M. Varga, V. Gataric, C. Y. Liu, and J. Matthes. Artificial presence, real-life influence? Effects of CGI influencers on young adults' health behavior intentions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 19(2), Apr. 2025. ISSN 1802-7962. doi: 10.5817/CP2025-2-3. URL <https://cyberpsychology.eu/article/view/37249>.
- [165] K. J. Schiff, D. Schiff, and N. S. Bueno. The Liar's Dividend: How Misinformation About Misinformation Affects Politician Support and Trust in Media. In *2020 APPAM Fall Research Conference*. APPAM, 2020. URL <https://files.osf.io/v1/resources/qpxr8/providers/osfstorage/5f930a94b1ea17004ac98124?action=download&direct&version=1>.
- [166] L. Schmeink. Incarnations of the metaverse in science fiction. *Gaming the Metaverse. Bielefeld: transcript Verlag*, pages 117–139, 2025.
- [167] L. Schøsler. The Diasystematic Status of the Diatopic Axis. *Scandinavian Studies in Language*, 14(2):55–72, 2023. URL <https://tidsskrift.dk/sss/article/view/142566>.
- [168] T. J. Sejnowski. Large language models and the reverse turing test. *Neural computation*, 35(3):309–342, 2023. URL <https://ieeexplore.ieee.org/abstract/document/10887133/>.
- [169] K. Shah, H. Joshi, and H. Joshi. Integrating Moral Values in AI: Addressing Ethical Challenges for Fair and Responsible Technology. *Journal of Informatics and Web Engineering*, 4(1):213–227, Feb. 2025. ISSN 2821-370X. doi: 10.33093/jiwe.2025.4.1.

16. URL <https://journals.mmupress.com/index.php/jiwe/article/view/1255>.
- [170] M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023. URL https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/s41586-023-06647-8&casa_token=SNhf0JOPSe4AAAAA:nktYxAKnX_ytc74zjYLADG6wRYZvZWbElETVMspKMiQM3qN1bE-qm713nKA4goyTgS363FQ-uUPshzf1.
- [171] A. Sharada, B. Ratkal, and S. Shivani. Simulation of human robot interaction through avatar. *International research journal of modernization in engineering technology and science (IRJMETs)*, 5(07), 2023.
- [172] A. Sherlock. Larger Than Life: Digital Resurrection and the Re-Enchantment of Society. *The Information Society*, 29(3):164–176, May 2013. ISSN 0197-2243. doi: 10.1080/01972243.2013.777302. URL <https://doi.org/10.1080/01972243.2013.777302>. Publisher: Routledge eprint: <https://doi.org/10.1080/01972243.2013.777302>.
- [173] S. Shin, C. Koo, J. Kim, and D. Gursoy. Effects of metaverse experience on behavioral intention of visitors: moderating role of similarity between virtual and real experience. *International Journal of Contemporary Hospitality Management*, 36(12):4055–4073, Apr. 2024. ISSN 0959-6119. doi: 10.1108/IJCHM-10-2023-1567. URL <https://www.emerald.com/insight/content/doi/10.1108/ijchm-10-2023-1567/full/html>. Publisher: Emerald Publishing Limited.
- [174] M. R. Shoaib, Z. Wang, M. T. Ahvanooy, and J. Zhao. Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models, Nov. 2023. URL <http://arxiv.org/abs/2311.17394>. arXiv:2311.17394 [cs].
- [175] M. Slote. Virtue ethics. In *The Routledge companion to ethics*, pages 478–489. Routledge, 2010. URL <https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.4324/9780203850701-44&type=chapterpdf>.
- [176] P. Smart, T. Scutt, K. Sycara, and N. Shadbolt. Integrating ACT-R Cognitive Models with the Unity Game Engine. In *Integrating cognitive architectures into virtual character design*, pages 35–64. IGI Global, 2016.
- [177] M. A. Stefanone, Z. Yue, and Z. Toh. A social cognitive approach to traditional media content and social media use: Selfie-related behavior as competitive strategy. *New Media & Society*, 21(2):317–335, Feb. 2019. ISSN 1461-4448. doi: 10.1177/1461444818795488. URL <https://doi.org/10.1177/1461444818795488>. Publisher: SAGE Publications.

- [178] J. Stenseke. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*, Dec. 2021. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-021-01325-7. URL <https://link.springer.com/10.1007/s00146-021-01325-7>.
- [179] J. Stenseke. Artificial virtuous agents in a multi-agent tragedy of the commons. *AI & SOCIETY*, Oct. 2022. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-022-01569-x. URL <https://link.springer.com/10.1007/s00146-022-01569-x>.
- [180] B. R. Steunebrink, M. Dastani, and J.-J. C. Meyer. The OCC Model Revisited. In *4th Workshop on Emotion and Computing*, Palo Alto, 2009. URL https://people.idsia.ch/~steunebrink/Publications/KI09_OCC_revisited.pdf.
- [181] N. Stewart Rosenfield, K. Lamkin, J. Re, K. Day, L. Boyd, and E. Linstead. A Virtual Reality System for Practicing Conversation Skills for Children with Autism. *Multimodal Technologies and Interaction*, 3(2):28, Apr. 2019. ISSN 2414-4088. doi: 10.3390/mti3020028. URL <https://www.mdpi.com/2414-4088/3/2/28>.
- [182] B.-C. Su. Navigating the new frontier: The role of AI-driven virtual influencers in consumer engagement. *AIMagazine*, 46(2):e70012, 2025. ISSN 2371-9621. doi: 10.1002/aaai.70012. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.70012>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.70012>.
- [183] Y. Sun, W. Gao, J. Ma, H. Lin, Z. Luo, and W. Zhang. ClarityEthic: Explainable Moral Judgment Utilizing Contrastive Ethical Insights from Large Language Models, Apr. 2025. URL <http://arxiv.org/abs/2412.12848>. arXiv:2412.12848 [cs].
- [184] P. Svoboda and L. Knihová. Exploring the Future of Education: Integrating Metaverse and AI Tools to Enhance Learning Experiences. *TEM Journal*, pages 631–643, Feb. 2025. ISSN 2217-8333, 2217-8309. doi: 10.18421/TEM141-56. URL https://www.temjournal.com/content/141/TEMJournalFebruary2025_631_643.html.
- [185] K. Szolin, D. J. Kuss, F. M. Nuyens, and M. D. Griffiths. Exploring the user-avatar relationship in videogames: A systematic review of the Proteus effect. *Human-Computer Interaction*, 38(5-6):374–399, Nov. 2023. ISSN 0737-0024, 1532-7051. doi: 10.1080/07370024.2022.2103419. URL <https://www.tandfonline.com/doi/full/10.1080/07370024.2022.2103419>.
- [186] M. C.-T. Tai. The impact of artificial intelligence on human society and bioethics. *Tzu-Chi Medical Journal*, 32(4):339–343, Aug. 2020. ISSN 2223-8956. doi: 10.4103/tcmj.tcmj_71_20.

- [187] K. H. Tan. *The Emergent Moral Ecology: A Novel Framework for AI Moral Responsibility*. PhD thesis, Singapore University of Social Sciences, 2025. URL <https://rgdoi.net/10.13140/RG.2.2.32889.58724>. Publisher: Unpublished.
- [188] M. Tan and Q. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR, May 2019. URL <https://proceedings.mlr.press/v97/tan19a.html>. ISSN: 2640-3498.
- [189] Z. Tao, D. Xi, Z. Li, L. Tang, and W. Xu. CAT-LLM: Prompting Large Language Models with Text Style Definition for Chinese Article-style Transfer, Jan. 2024. URL <http://arxiv.org/abs/2401.05707>. arXiv:2401.05707 [cs].
- [190] Y. Tsubota and Y. Kano. Text generation indistinguishable from target person by prompting few examples using llm. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 13–20, 2024.
- [191] H. Ullah, S. Manickam, M. Obaidat, S. U. A. Laghari, and M. Uddin. Exploring the Potential of Metaverse Technology in Healthcare: Applications, Challenges, and Future Directions. *IEEE Access*, 11:69686–69707, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3286696. URL <https://ieeexplore.ieee.org/document/10153590>.
- [192] U. Upadhyay, A. Kumar, G. Sharma, B. B. Gupta, W. A. Alhalabi, V. Arya, and K. T. Chui. Cyberbullying in the Metaverse: A Prescriptive Perception on Global Information Systems for User Protection. *Journal of Global Information Management (JGIM)*, 31(1):1–25, 2023. ISSN 1062-7375. doi: 10.4018/JGIM.325793. URL <https://www.igi-global.com/gateway/article/www.igi-global.com/gateway/article/325793>.
- [193] M. Van Der Boon, L. Fermoselle, F. Ter Haar, S. Dijkstra-Soudarissanane, and O. Niamut. Deep Learning Augmented Realistic Avatars for Social VR Human Representation. In *ACM International Conference on Interactive Media Experiences*, pages 311–318, Aveiro JB Portugal, June 2022. ACM. ISBN 978-1-4503-9212-9. doi: 10.1145/3505284.3532976. URL <https://dl.acm.org/doi/10.1145/3505284.3532976>.
- [194] L. Van Ryn, T. Kohn, B. Nansen, M. Arnold, and M. Gibbs. Researching death online. In *The Routledge companion to digital ethnography*, pages 138–146. Routledge, 2017.
- [195] D. Vanderelst and A. Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48:56–66, 2018. URL <https://www.sciencedirect.com/science/article/pii/S1389041716302005>.

- [196] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [197] P. Venkadesh, S. V. Divya, and K. S. Kumar. Unlocking AI Creativity: A Multi-Agent Approach with CrewAI. *Journal of Trends in Computer Science and Smart Technology*, 6(4):338–356, Dec. 2024. ISSN 2582-4104. doi: 10.36548/jtcsst.2024.4.002. URL <https://irojournals.com/tcsst/article/view/6/4/2>.
- [198] A. Vijayaraghavan and C. Badaea. Minimum levels of interpretability for artificial moral agents. *Ai and Ethics*, 5(3):2071–2087, 2025. ISSN 2730-5953. doi: 10.1007/s43681-024-00536-0. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12103486/>.
- [199] C. Vilchis, C. Perez-Guerrero, M. Mendez-Ruiz, and M. Gonzalez-Mendoza. A survey on the pipeline evolution of facial capture and tracking for digital humans. *Multimedia Systems*, 29(4):1917–1940, Aug. 2023. ISSN 0942-4962, 1432-1882. doi: 10.1007/s00530-023-01081-2. URL <https://link.springer.com/10.1007/s00530-023-01081-2>.
- [200] S. Vinanzi and A. Cangelosi. Casper: Cognitive architecture for social perception and engagement in robots. *International Journal of Social Robotics*, pages 1–19, 2024.
- [201] M. Viridis. Problemi di diatopia e di diacronia della lingua sarda. Un’ipotesi di sociolinguistica storica. *REMBERGER, E.-M., VIRDIS, M., WAGNER B.(2020, eds.), Il sardo in movimento. Vandenhoeck & Ruprecht Unipress, Göttingen*, pages 31–46, 2020. URL https://www.academia.edu/download/65239361/MaurizioVIRDIS_Problemi_di_diatopia_e_di_diacronia_della_lingua_sarda._Il_sardo_in_movimento_.pdf.
- [202] A. Vishwanath, E. D. Bøhn, O.-C. Granmo, C. Maree, and C. Omlin. Towards artificial virtuous agents: games, dilemmas and machine learning. *AI and Ethics*, pages s43681–022–00251–8, Dec. 2022. ISSN 2730-5953, 2730-5961. doi: 10.1007/s43681-022-00251-8. URL <https://link.springer.com/10.1007/s43681-022-00251-8>.
- [203] R. Volkman and K. Gabriels. AI Moral Enhancement: Upgrading the Socio-Technical System of Moral Engagement. *Science and Engineering Ethics*, 29(2):11, 2023. ISSN 1353-3452. doi: 10.1007/s11948-023-00428-2. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10036265/>.
- [204] T. F. Waddell, S. S. Sundar, and J. Auriemma. Can Customizing an Avatar Motivate Exercise Intentions and Health Behaviors Among Those with Low Health Ideals?

- Cyberpsychology, Behavior, and Social Networking*, 18(11):687–690, Nov. 2015. ISSN 2152-2715, 2152-2723. doi: 10.1089/cyber.2014.0356. URL <https://journals.sagepub.com/doi/full/10.1089/cyber.2014.0356>.
- [205] Y. Wang, B. Duan, X. Chen, Y. Song, and X. Liu. The application of metaverse in mental health. *Frontiers in Public Health*, 13:1463494, Apr. 2025. ISSN 2296-2565. doi: 10.3389/fpubh.2025.1463494.
- [206] Z. Wang and Z. Li. Person Sensor-Aided Scene Recognition and Understanding Based on CG Technology. In *2020 International Conference on Inventive Computation Technologies (ICICT)*, pages 60–63, Feb. 2020. doi: 10.1109/ICICT48043.2020.9112445. URL <https://ieeexplore.ieee.org/document/9112445/>.
- [207] A. Wegmann, M. Schraagen, and D. Nguyen. Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.26. URL <https://aclanthology.org/2022.repl4nlp-1.26>.
- [208] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. URL <https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html>.
- [209] M. Weinberger. What Is Metaverse?—A Definition Based on Qualitative Meta-Synthesis. *Future Internet*, 14(11):310, Nov. 2022. ISSN 1999-5903. doi: 10.3390/fi14110310. URL <https://www.mdpi.com/1999-5903/14/11/310>. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.
- [210] M. D. Weiser. Ubiquitous computing. In *ACM Conference on Computer Science*, volume 418, pages 197530–197680, 1994. URL <https://ieeexplore.ieee.org/iel4/4236/14762/x0086424.pdf>. Issue: 10.1145.
- [211] E. Weitnauer, N. M. Thomas, F. Rabe, and S. Kopp. Intelligent Agents Living in Social Virtual Environments – Bringing Max into Second Life. In H. Prendinger, J. Lester, and M. Ishizuka, editors, *Intelligent Virtual Agents*, pages 552–553, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-85483-8. doi: 10.1007/978-3-540-85483-8_85.
- [212] M. Wenglinisky. Obedience to authority: An experimental view. *Contemporary Sociology*, 47(4):423–428, 2018. doi: 10.1177/0094306118779813a.
- [213] B. K. Wiederhold. Sexual Harassment in the Metaverse. *Cyberpsychology, Behavior, and Social Networking*, 25(8):479–480, Aug. 2022. ISSN 2152-2715. doi: 10.

- 1089/cyber.2022.29253.editorial. URL <https://doi.org/10.1089/cyber.2022.29253.editorial>.
- [214] B. K. Wiederhold. Treading Carefully in the Metaverse: The Evolution of AI Avatars. *Cyberpsychology, Behavior, and Social Networking*, 26(5):321–322, May 2023. ISSN 2152-2715, 2152-2723. doi: 10.1089/cyber.2023.29280.editorial. URL <https://journals.sagepub.com/doi/full/10.1089/cyber.2023.29280.editorial>.
- [215] V. Windland, M. Syed, and A. Gutowska. What is crewAI?, Oct. 2024. URL <https://www.ibm.com/think/topics/crew-ai>.
- [216] K. Wolfenstein. Molto più che semplici giochi: il vero potenziale dell’intelligenza artificiale nel metaverso: dai mondi personalizzati all’economia digitale, 2025. URL <https://xpert.digital/it/il-vero-potenziale-dellintelligenza-artificiale-nel-metaverso/>.
- [217] P. Wozniak, H. Afrisal, R. G. Esparza, and B. Kwolek. Scene Recognition for Indoor Localization of Mobile Robots Using Deep CNN. In L. J. Chmielewski, R. Kozera, A. Orłowski, K. Wojciechowski, A. M. Bruckstein, and N. Petkov, editors, *Computer Vision and Graphics*, pages 137–147, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00692-1. doi: 10.1007/978-3-030-00692-1_13.
- [218] J. Wüest. La notion de diamésie est-elle nécessaire? *Travaux de linguistique*, 59(2): 147–162, 2009. URL https://shs.cairn.info/article/TL_059_0147.
- [219] M. Xu, W. C. Ng, W. Y. B. Lim, J. Kang, Z. Xiong, D. Niyato, Q. Yang, X. Shen, and C. Miao. A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges. *IEEE Communications Surveys & Tutorials*, 25(1):656–700, 2022. URL https://ieeexplore.ieee.org/abstract/document/9944868/?casa_token=602xGJapjegAAAAA:vReqJu8E-6r84AIpCRPQGbj4TaJTIIzih6lmIx9zxI6D5GKU8G60jJvTjqlUhnPQ2r1xCg7m. Publisher: IEEE.
- [220] T. Yamazaki, T. Mizumoto, K. Yoshikawa, M. Ohagi, T. Kawamoto, and T. Sato. An open-domain avatar chatbot by exploiting a large language model. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 428–432, 2023. URL <https://aclanthology.org/2023.sigdial-1.40/>.
- [221] D. Yang, J. Zhang, Y. Sun, and Z. Huang. Showing usage behavior or not? The effect of virtual influencers’ product usage behavior on consumers. *Journal of Retailing and Consumer Services*, 79:103859, July 2024. ISSN 0969-6989. doi: 10.1016/j.jretconser.2024.103859. URL <https://www.sciencedirect.com/science/article/pii/S0969698924001553>.

- [222] L. Yang, Y. Xu, and P. Hui. Framing metaverse identity: A multidimensional framework for governing digital selves. *arXiv:2406.08029*, June 2024. doi: <https://doi.org/10.48550/arXiv.2406.08029>. URL <https://arxiv.org/abs/2406.08029v3>.
- [223] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.
- [224] N. Yee and J. Bailenson. The Proteus effect: The effect of transformed self-representation on behavior. *Human communication research*, 33(3):271–290, 2007. URL <https://academic.oup.com/hcr/article-abstract/33/3/271/4210718>. Publisher: Oxford University Press Oxford, UK.
- [225] Yeshwanth Vasa. ETHICAL IMPLICATIONS AND BIAS IN GENERATIVE AI. *International Journal for Research Publication and Seminar*, 14(5):500–511, Dec. 2023. ISSN 2278-6848. doi: 10.36676/jrps.v14.i5.1541. URL <https://jrps.shodhsagar.com/index.php/j/article/view/1541>.
- [226] J. Yuan, P. K. Murukannaiah, and M. P. Singh. Right vs. Right: Can LLMs Make Tough Choices?, Dec. 2024. URL <http://arxiv.org/abs/2412.19926>. arXiv:2412.19926 [cs].
- [227] X. Yuan, Z. Qiao, and A. Meyarian. Scale attentive network for scene recognition. *Neurocomputing*, 492:612–623, July 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2021.12.053. URL <https://www.sciencedirect.com/science/article/pii/S0925231221018919>.
- [228] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky. Fast Bi-Layer Neural Synthesis of One-Shot Realistic Head Avatars. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, volume 12357, pages 524–540. Springer International Publishing, Cham, 2020. ISBN 978-3-030-58609-6 978-3-030-58610-2. doi: 10.1007/978-3-030-58610-2_31. URL https://link.springer.com/10.1007/978-3-030-58610-2_31. Series Title: Lecture Notes in Computer Science.
- [229] H. Zeng, X. Song, G. Chen, and S. Jiang. Learning Scene Attribute for Scene Recognition. *IEEE Transactions on Multimedia*, 22(6):1519–1530, June 2020. ISSN 1941-0077. doi: 10.1109/TMM.2019.2944241. URL <https://ieeexplore.ieee.org/document/8851274/>.
- [230] X. Zhang, Y. Chen, L. Hu, and Y. Wang. The metaverse in education: Definition, framework, features, potential applications, challenges, and future research topics. *Frontiers in Psychology*, 13, Oct. 2022. ISSN 1664-1078. doi: 10.3389/fpsyg.

- 2022.1016300. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.1016300/full>. Publisher: Frontiers.
- [231] Z. Zhang, Z. Chen, and L. Xu. Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, 101:104327, 2022. URL <https://www.sciencedirect.com/science/article/pii/S0022103122000464>.
- [232] E. Zhou and D. Lee. Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, 3(3):pgae052, Feb. 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae052. URL <https://academic.oup.com/pnasnexus/article/doi/10.1093/pnasnexus/pgae052/7618478>.
- [233] J. Zhou, M. Hu, J. Li, X. Zhang, X. Wu, I. King, and H. Meng. Rethinking Machine Ethics—Can LLMs Perform Moral Reasoning through the Lens of Moral Theories? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2227–2242, 2024. URL <https://aclanthology.org/2024.findings-naacl.144/>.
- [234] K. Zhou, Q. Li, H. Liu, Q. Pan, Y. Wang, and L. Tian. The integration of artificial intelligence and moral education: an innovative way to improve the moral quality of college students. In *Proceedings of the 2025 3rd International Conference on Educational Knowledge and Informatization, EKI '25*, pages 511–515, New York, NY, USA, Nov. 2025. Association for Computing Machinery. ISBN 979-8-4007-1584-6. doi: 10.1145/3765325.3767191. URL <https://dl.acm.org/doi/10.1145/3765325.3767191>.
- [235] A. Zhuk. Ethical implications of AI in the Metaverse. *AI and Ethics*, Mar. 2024. ISSN 2730-5953, 2730-5961. doi: 10.1007/s43681-024-00450-5. URL <https://link.springer.com/10.1007/s43681-024-00450-5>.
- [236] W. Zielonka, T. Bolkart, and J. Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. URL http://openaccess.thecvf.com/content/CVPR2023/html/Zielonka_Instant_Volumetric_Head_Avatars_CVPR_2023_paper.html.
- [237] D. Zolezzi, L. Martini, S. Iacono, and G. V. Vercelli. The Timeless Chamber: A Virtual Reality Escape Room Enhancing Educational Experiences. In *Conference Proceedings. New Perspectives in Science Education 2025*, 2025. URL https://conference.pixel-online.net/library_scheda.php?id_abs=7079.
- [238] L. Zonaphan, K. Northus, J. Wijaya, S. Achmad, and R. Sutoyo. Metaverse as A Future of Education: A Systematic Review. In *2022 8th International HCI and*

UX Conference in Indonesia (CHIuXiD), volume 1, pages 77–81, Nov. 2022. doi: 10.1109/CHIuXiD57244.2022.10009854. URL <https://ieeexplore.ieee.org/document/10009854>.

- [239] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer. A survey of structure from motion. *Acta Numerica*, 26:305–364, May 2017. ISSN 0962-4929, 1474-0508. doi: 10.1017/S096249291700006X. URL https://www.cambridge.org/core/product/identifier/S096249291700006X/type/journal_article.