Dottorato di Ricerca in *Visual and Media Studies* Ciclo XXXVII

Progetto PON *AI e sostenibilità*

INTELLIGENZA ARTIFICIALE E SOSTENIBILITÁ: UNA VIA NELL'INFERENZA ATTIVA

Maria Raffa

Matricola 1036908

Tutor:
Prof. Riccardo Manzotti

Cotutor:
Prof. Luciano Floridi

Coordinatore:
Prof. Vincenzo Trione

2023/2024

UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università e della Ricerca

PON RICERCA E INNOVAZIONE 2014 - 2020

REACT EU

ABSTRACT

Il presente lavoro di tesi dottorale indaga come l'Intelligenza Artificiale (IA) possa integrare i principi di sostenibilità sfruttando il modello cognitivo dell'inferenza attiva (AIF). Ciò avviene attraverso l'analisi delle pratiche di IA sostenibile e delle potenzialità delle architetture cognitive (CA) per la sostenibilità e la validazione di questi concetti tramite una simulazione implementata in un ambiente dinamico. Gli obiettivi principali del lavoro sono tre: offrire una riflessione critica e interdisciplinare sui limiti e sulle opportunità legati all'utilizzo dei modelli cognitivi per un'IA sostenibile dal punto di vista ambientale e sociale, promuovendo il dialogo tra filosofia della mente, scienze cognitive e IA; esplorare le potenzialità dell'AIF come strumento per progettare IA sostenibili; dimostrare, attraverso la simulazione proposta, la capacità di un sistema basato su AIF di bilanciare bisogni immediati e sostenibilità a lungo termine in contesti dinamici.

Per tali scopi, il lavoro è articolato in quattro capitoli, organizzati come segue: nel Capitolo I, si analizza la relazione tra IA e sostenibilità. Viene discusso il ruolo dell'IA nella promozione di pratiche sostenibili e di gestione efficiente delle risorse, in linea con gli Obiettivi di Sviluppo Sostenibili promossi dall'ONU. Vengono anche evidenziati i limiti e le criticità degli sviluppi contemporanei dell'IA, che talvolta rischiano di andare in direzione opposta rispetto a quella auspicata a causa dell'elevato consumo di energia e di risorse computazionali che comportano, nonché per il loro impatto sociale.

Dopo tale analisi, il Capitolo II esplora il ruolo dei modelli cognitivi nell'ambito dell'IA sostenibile, con un'attenzione particolare al concetto di IA spiegabile. Viene analizzato il principio di spiegabilità come strumento essenziale per garantire la sostenibilità sociale nei sistemi intelligenti. Il capitolo propone un approccio cognitivo per progettare IA che bilancino efficienza, robustezza e trasparenza. In questa prospettiva, l'adozione delle cosiddette CA non solo favorisce lo sviluppo di sistemi interpretabili, ma contribuisce alla creazione di IA più affidabili dal punto di vista sociale e ambientale. Infine, si introduce il principio di energia libera (FEP) come candidato esemplare di modello cognitivo sostenibile.

Nel Capitolo III si approfondiscono il FEP e l'AIF. Viene esaminato come il FEP, il *predictive processing* e l'AIF possano favorire modellazioni di IA che siano sostenibili. Tali modelli, ispirati ai processi biologici, permettono di predire e minimizzare la

discrepanza tra aspettative e realtà, con conseguente riduzione del consumo energetico e ottimizzazione delle risorse. Vengono poi illustrate le applicazioni dell'AIF per l'IA, con un'analisi delle analogie e differenze rispetto all'apprendimento per rinforzo, approccio per certi versi simile, e si analizzano le ragioni per cui l'AIF è appropriata all'implementazione di modelli di IA sostenibile.

Nel Capitolo IV, infine, è presentata una simulazione dove sono applicati i principi teorici precedentemente discussi. L'esperimento si concentra sulla gestione delle risorse in ambienti dinamici e mostra come un agente basato su AIF possa imparare a bilanciare le proprie necessità immediate con la sostenibilità a lungo termine. Vengono esaminati i risultati ottenuti e discussi i limiti dell'applicazione proposta, suggerendo sviluppi futuri per la ricerca, tra cui l'uso di modelli di apprendimento profondo avanzati e un confronto con dati reali.

Observe, infer, act, repeat.

INDICE

INTRODUZIONE

Il dibattito contemporaneo sembra essere dominato da due temi che, oltre a catturare l'immaginario collettivo, circoscrivono le priorità per il futuro della società: l'intelligenza artificiale (*Artificial Intelligence*, IA) e la sostenibilità. L'IA rappresenta una delle massime espressioni delle capacità tecnologiche umane, frutto di decenni di ricerca e innovazione. Essa ha il potenziale di trasformare profondamente settori chiave come la scienza, l'economia e la società, grazie alla capacità di elaborare grandi quantità di dati, apprendere autonomamente e automatizzare processi complessi. La sostenibilità, d'altro canto, sottolinea l'urgenza di preservare il pianeta e le sue risorse naturali, ponendo al centro il benessere delle generazioni presenti e future. Questo concetto si fonda sull'equilibrio tra esigenze ambientali, economiche e sociali e sulla promozione di soluzioni che garantiscano la continuità della vita sulla Terra e un uso responsabile delle risorse. Pur distinti, i due ambiti convergono in un'interazione complessa e complementare, testimoniata dall'ampia produzione scientifica e divulgativa che nell'ultimo decennio si è interessata a studiare, approfondire e problematizzare la loro relazione.

Molte delle sfide più pressanti dell'epoca contemporanea si collocano proprio all'intersezione tra IA e sostenibilità, due ambiti il cui sviluppo richiede un'analisi attenta e approfondita. Le IA odierne, pur straordinarie per capacità e applicazioni, necessitano di una comprensione critica per mitigarne i potenziali rischi. Tra questi spiccano problematiche etiche, come i *bias* degli algoritmi che possono condurre a discriminazioni; questioni di sicurezza, legate all'uso improprio o non autorizzato; e impatti sociali ed economici, tra cui la perdita di posti di lavoro e le crescenti disuguaglianze nell'accesso alle tecnologie. Affrontare e mettere a fuoco questi rischi è fondamentale per garantire che il potenziale trasformativo dell'IA venga utilizzato in modo responsabile ed equo. Parallelamente, la sostenibilità, intrecciandosi con il rapido progresso tecnologico, richiama un'urgenza globale che richiede uno sguardo critico e informato. È infatti necessario evitare fenomeni come il *greenwashing*, ovvero l'attribuzione ingannevole di benefici ambientali a prodotti o processi che in realtà nascondono impatti negativi, che compromette la fiducia pubblica e la reale efficacia delle soluzioni proposte (Floridi, 2022).

L'integrazione tra IA e sostenibilità non è soltanto auspicabile, ma indispensabile, considerando l'urgenza della crisi ambientale e sociale che il pianeta si trova ad affrontare. Questa situazione richiede soluzioni concrete, coordinate e ben calibrate, e l'IA offre l'opportunità di affrontare tali sfide con approcci innovativi ed efficaci. Tuttavia, l'IA rappresenta anche un'arma a doppio taglio: infatti, se da un lato può contribuire positivamente, dall'altro possiede caratteristiche che rischiano di aggravare le problematiche ambientali e sociali, a causa degli elevati costi energetici dei processi computazionali e delle potenziali discriminazioni introdotte dagli algoritmi. Per esplorare la connessione tra sostenibilità e IA è dunque necessario stabilire una base concettuale chiara che perimetri i due ambiti d'analisi e che consenta di analizzarne la mutua relazione.

L'Oxford Dictionary definisce l'IA come "the capacity of computers or other machines to exhibit or simulate intelligent behaviour"[1]. Questa definizione, volutamente ampia, si radica nel concetto altrettanto vasto di intelligenza. In questo lavoro si adotta una prospettiva operativa, definendo l'intelligenza come la capacità di una macchina di svolgere compiti comparabili a quelli umani. Tale scelta deriva dall'obiettivo scientifico di questa ricerca: contribuire allo sviluppo di modalità concrete e strategiche per affrontare le sfide sociali e culturali del nostro tempo.

La definizione operativa di intelligenza qui adottata si ispira al test di Alan Turing. Pioniere dell'informatica, già nel 1950 Turing sosteneva che chiedersi se una macchina potesse essere considerata intelligente fosse una domanda "absurd" (Turing, 1950, 433), suggerendo piuttosto di interrogarsi sulla possibilità che una macchina fosse in grado di svolgere specifici compiti in modo più efficace rispetto a un essere umano (Turing, 1950). Oggi molte IA superano ampiamente questa prova in ambiti come la generazione di contenuti testuali e visivi, ma continuano a mancare di alcune qualità intrinsecamente umane, come creatività, sarcasmo ed empatia[2].

La sostenibilità, a sua volta, è progressivamente emersa come valore centrale nel dibattito contemporaneo, proponendosi come una risposta alle sfide globali legate al degrado ambientale, alle disuguaglianze sociali e agli squilibri economici. Il suo obiettivo

[1] Oxford Dictionary online: https://www.oed.com/dictionary/artificial-intelligence_n?tl=true, consultato in data 22/10/2024.

[2] La discussione in merito è sterminata, e non è questo il luogo per trattarla. Per una panoramica sulla questione, si rimanda a M. Mitchell, 2022, *Intelligenza Artificiale: Una guida per esseri umani pensanti*, Einaudi, Torino.

principale è garantire uno sviluppo che soddisfi i bisogni delle generazioni presenti senza compromettere la capacità delle future generazioni di soddisfare i propri. Questo concetto si articola in tre dimensioni interdipendenti: ambientale, sociale ed economica. La sostenibilità ambientale mira a preservare le risorse naturali e a ridurre gli impatti negativi delle attività umane sull'ecosistema. La sostenibilità economica, invece, pone l'accento sulla creazione di modelli di crescita inclusivi e duraturi, che garantiscano prosperità condivisa. La sostenibilità sociale, infine, si concentra sul miglioramento della qualità della vita e sulla promozione di equità, inclusione e giustizia all'interno delle comunità.

Nell'indagine dell'intreccio tra i due ambiti, si considera da un lato la sostenibilità dell'IA per quanto riguarda i fini, che prevede l'implementazione di IA utili a compiere compiti che vanno in direzione del raggiungimento di obiettivi della sostenibilità, e dall'altro IA che siano sostenibili in quanto mezzi, dunque strutturate in modo da ridurre il loro impatto ecologico e sociale. L'aspetto della sostenibilità sociale assume in questo contesto un ruolo specifico e strategico, poiché si tratta di garantire che i sistemi siano progettati e utilizzati in modo etico, equo e accessibile, minimizzando le disuguaglianze e promuovendo il benessere collettivo. Un elemento chiave in questo ambito è l'IA spiegabile, ovvero il filone di ricerca che si occupa di rendere comprensibili e monitorabili le decisioni automatizzate. Questo aspetto è cruciale per favorire fiducia verso l'uso delle tecnologie, che devono essere trasparenti tanto agli sviluppatori quanto agli utenti, in modo da favorire il più possibile un utilizzo responsabile. Inoltre, la sostenibilità sociale nell'IA richiede che tali sistemi non solo rispettino la diversità e l'inclusione, ma anche che contribuiscano attivamente a ridurre le barriere di accesso e a evitare discriminazioni algoritmiche.

La relazione tra IA spiegabile, ovvero sostenibile dal punto di vista sociale, e sostenibilità ambientale – o ecologica – è poi particolarmente significativa. I modelli spiegabili, infatti, tendono ad avere una struttura meno intricata rispetto a quelli difficili da interpretare. Di conseguenza, richiedono un minor impiego di risorse computazionali sia durante l'addestramento sia nell'esecuzione. Questo si traduce in un impatto ambientale più contenuto, poiché una ridotta complessità contribuisce a diminuire il consumo energetico e le emissioni associate ai processi computazionali. In questo senso, l'IA ecologica e quella spiegabile si rafforzano reciprocamente, condividendo l'obiettivo di ridurre l'impatto ambientale e aumentare l'inclusività sociale.

In ambito scientifico, grande attenzione è stata dedicata allo studio di sistemi intelligenti per promuovere soluzioni sostenibili, mentre è meno esplorato il filone di ricerca incentrato sullo sviluppo di IA che siano sostenibili in sé. È proprio in questa seconda area che si colloca la presente ricerca. La progettazione di sistemi di IA che rispettino i principi di sostenibilità, infatti, richiede una riflessione integrata, in grado di considerare gli impatti ambientali, sociali ed economici di queste tecnologie.

Sebbene la dimensione economica sia imprescindibile per una valutazione completa della sostenibilità, poiché consente di bilanciare costi e benefici su scala globale, questo aspetto viene trattato qui solo marginalmente. L'attenzione è invece rivolta agli aspetti più rilevanti per l'implementazione di IA sostenibili sul piano ambientale e sociale. In particolare, lo studio si focalizza sul ruolo cruciale che i modelli cognitivi possono giocare in questo contesto, rappresentando il nucleo centrale della ricerca. La scelta di concentrare l'analisi sui modelli cognitivi è motivata dalla loro capacità di fornire una base teorica per lo sviluppo di sistemi di IA spiegabili, adattivi e affidabili. Questi modelli, infatti, permettono di integrare percezione, apprendimento e azione in maniera simile a quella dei sistemi biologici, promuovendo soluzioni che siano non solo tecnologicamente avanzate, ma anche sostenibili. Studiare il ruolo dei modelli cognitivi nell'ambito dell'IA sostenibile rappresenta una direzione di crescente importanza scientifica e interdisciplinare, che coinvolge IA, scienze cognitive e filosofia della mente. Questo approccio consente di colmare il divario tra la progettazione tecnica delle IA e la necessità di affrontare le sfide globali legate alla sostenibilità: l'esplorazione di modelli cognitivi, infatti, non solo mira a ottimizzare le prestazioni delle IA, ma anche a garantire che queste interagiscano con il contesto umano e ambientale in modo equilibrato e consapevole.

É in questo contesto interdisciplinare, all'intersezione tra studi sull'IA, sulla sostenibilità, filosofia della mente e scienze cognitive, che si colloca il presente lavoro, con lo scopo di contribuire alla comprensione e alla progettazione di sistemi intelligenti che, oltre a essere tecnologicamente avanzati, siano ecologicamente e socialmente sostenibili. Tali sistemi, capaci di operare efficacemente in ambienti complessi, rispondono alle sfide globali poste dall'innovazione tecnologica, contribuendo al contempo a mitigare le crisi ambientali e sociali del nostro tempo.

Tuttavia, gli ambiti citati spesso operano in isolamento, adottando linguaggi, strumenti e metodologie che raramente si incontrano. Questo isolamento rischia di

limitare la comprensione complessiva del problema, perpetuando lacune teoriche e pratiche. L'approccio adottato in questo lavoro mira quindi a creare ponti tra ambiti di ricerca scientifica che, pur condividendo oggetti di studio simili, spesso sembrano distanti. Per integrare prospettive teoriche e applicative, quindi, la tesi si fonda su una revisione critica e sistematica della letteratura scientifica negli ambiti della filosofia della mente, delle scienze cognitive e dell'IA per delineare lo stato dell'arte, evidenziandone potenzialità e criticità, e per analizzare modelli ispirati a processi biologici, ovvero il principio dell'energia libera e l'inferenza attiva, utili per l'implementazione di IA sostenibili. Inoltre, esplora soluzioni pratiche avvalendosi di un esperimento computazionale in Python per simulare un agente implementato con inferenza attiva che attua un comportamento sostenibile. Se, da un lato, la filosofia della mente e le scienze cognitive offrono una solida base teorica per l'analisi e la comprensione dei modelli, dall'altro la simulazione computazionale consente di testarne concretamente l'efficacia, valutandone le potenzialità e le implicazioni in contesti pratici.

Attraverso l'approccio integrato così delineato, il lavoro si propone non solo di contribuire a colmare un divario tra discipline, ma anche di provare a elaborare un linguaggio comune capace di favorire la collaborazione e di offrire soluzioni più complete e innovative alle sfide globali. L'interdisciplinarietà, in questa circostanza, non è solo un metodo, ma una strategia necessaria per affrontare con profondità e visione d'insieme i complessi problemi dell'IA sostenibile. Per mezzo di tale strategia, la tesi cerca di rispondere a tre principali domande di ricerca. La prima riguarda il modo in cui l'IA può integrarsi con i principi di sostenibilità. Questo implica un'esplorazione delle modalità con cui i sistemi di IA possono essere progettati per rispettare la sostenibilità ambientale e sociale, affrontando al contempo le implicazioni legate all'uso di queste tecnologie. La seconda domanda approfondisce il ruolo dei modelli cognitivi, con un focus sul principio dell'energia libera e sull'inferenza attiva, nella progettazione di IA sostenibili. Tali modelli, ispirati ai processi biologici, offrono una base teorica per sviluppare sistemi di IA trasparenti ed etici, capaci di migliorare la comprensibilità e l'affidabilità delle decisioni automatizzate. Infine, la ricerca si interroga su come validare questa integrazione in contesti concreti, traducendo le soluzioni teoriche in applicazioni pratiche. Ciò richiede l'identificazione di metodologie adeguate a testare l'efficacia e l'impatto delle IA sostenibili in scenari reali.

Alla luce di queste domande, gli obiettivi principali della ricerca sono tre. Il primo, più generale, è offrire una riflessione critica e interdisciplinare sui limiti e sulle opportunità legati all'utilizzo dei modelli cognitivi come paradigmi per un'IA sostenibile. Questo obiettivo mira a favorire il dialogo tra i diversi campi di ricerca che trattano questi concetti, chiarendo le loro implicazioni teoriche e pratiche. In particolare, si intende fare chiarezza su alcune ambiguità presenti nella letteratura filosofica, dove tali oggetti di ricerca sono spesso trattati in modo contraddittorio, e nelle scienze cognitive e nell'IA, dove talvolta vengono dati per acquisiti senza un'adeguata esplicitazione dei loro presupposti.

Il secondo obiettivo, più specifico, è esplorare le potenzialità del modello cognitivo dell'inferenza attiva come strumento per progettare IA sostenibili. Questa indagine si concentra sul modo in cui tali modelli possano guidare lo sviluppo di sistemi capaci di bilanciare esigenze contrastanti, come l'ottimizzazione delle risorse e la sostenibilità a lungo termine, mantenendo trasparenza e adattabilità.

Il terzo obiettivo, infine, consiste nel dimostrare, attraverso una simulazione, come un sistema basato sull'inferenza attiva possa bilanciare bisogni immediati e sostenibilità nel lungo periodo all'interno di un contesto dinamico. La dimostrazione intende non solo validare l'efficacia degli approcci teorici presentati precedentemente, ma anche offrire un contributo concreto alla progettazione di IA capaci di rispondere alle sfide globali in modo innovativo ed etico. Per perseguire tali scopi, la tesi è composta da quattro capitoli, organizzati come segue.

Nel Capitolo I, *IA e sostenibilità*, si analizza la relazione tra IA e sostenibilità. Si discute il ruolo dell'IA nella promozione di pratiche sostenibili e di gestione efficiente delle risorse, in linea con gli Obiettivi di Sviluppo Sostenibili promossi dall'ONU. Vengono anche evidenziati i limiti e le criticità degli sviluppi contemporanei dell'IA, che talvolta rischiano di andare in direzione opposta rispetto a quella auspicata a causa dell'elevato consumo di energia e di risorse computazionali che comportano, nonché per il loro impatto sociale.

Dopo tale analisi, il Capitolo II, *Modelli cognitivi per IA sostenibile*, esplora il ruolo dei modelli cognitivi nell'ambito dell'IA sostenibile, con un'attenzione particolare al concetto di IA spiegabile. É analizzato il principio di spiegabilità come strumento essenziale per garantire la sostenibilità sociale nei sistemi intelligenti. Il capitolo propone

un approccio cognitivo per progettare IA che bilancino efficienza, robustezza e trasparenza. In questa prospettiva, l'adozione delle cosiddette architetture cognitive non solo favorisce lo sviluppo di sistemi interpretabili, ma contribuisce alla creazione di IA più affidabili dal punto di vista sociale e ambientale. Infine, viene introdotto il principio dell'energia libera come candidato esemplare di modello cognitivo sostenibile.

Nel Capitolo III, *Il principio dell'energia libera e l'inferenza attiva*, sono approfonditi il principio dell'energia libera e l'inferenza attiva. Viene esaminato come essi, insieme al cosiddetto *predictive processing*, possano favorire modellazioni di IA sostenibili. Ispirati ai processi biologici, tali modelli permettono di predire e minimizzare la discrepanza tra aspettative e realtà, con conseguente riduzione del consumo energetico e ottimizzazione delle risorse. Questo consente di approfondire le applicazioni dell'inferenza attiva per l'IA, con un'analisi delle analogie e differenze rispetto all'apprendimento per rinforzo, approccio per certi versi simile, argomentando come l'inferenza attiva possa essere utilizzata per implementare modelli di IA sostenibile.

Nel Capitolo IV, *Un'applicazione: Modellare la sostenibilità tramite la gestione delle risorse*, infine, è presentata una simulazione dove si applicano i principi teorici discussi nei capitoli precedenti. L'esperimento si concentra sulla gestione delle risorse in ambienti dinamici e mostra come un agente basato su inferenza attiva possa imparare a bilanciare le proprie necessità immediate con la sostenibilità a lungo termine. Vengono esaminati i risultati ottenuti e discussi i limiti dell'applicazione proposta, suggerendo possibili direzioni per futuri sviluppi della ricerca, tra cui l'uso di modelli di apprendimento profondo avanzati e un confronto con dati reali.

Tale struttura permette di affrontare in modo sistematico i complessi temi dell'IA e della sostenibilità, offrendo una sintesi teorico-pratica e contribuendo al dibattito accademico e applicativo in questo campo cruciale.

CAPITOLO I

IA E SOSTENIBILITÁ

1.1 Quale sostenibilità

L'obiettivo del Capitolo I è delineare la relazione tra IA e sostenibilità, in modo da chiarire il contesto all'interno del quale si colloca il lavoro di ricerca qui esposto.

Se – con un'assunzione piuttosto forte – si considerano le IA odierne come simulatori degli esseri umani, si comprende bene come mai sia richiesto loro di confrontarsi con i medesimi problemi che devono essere fronteggiati dalla società. Uno dei quali – il più pressante probabilmente – è rappresentato dalla crisi climatica, che spinge a cercare soluzioni che vadano verso la direzione della sostenibilità. Occorre sottolineare che il concetto di crisi climatica è profondamente complesso e non può essere ridotto a una narrativa univoca. Esso viene associato al riscaldamento globale, che, se da un lato rappresenta una minaccia per molte regioni del pianeta, a causa dell'aumento delle temperature che aggrava eventi climatici estremi come siccità, inondazioni e uragani, dall'altro, in certe zone ha portato, almeno temporaneamente, a benefici locali. Ad esempio, in alcune aree dell'Artico e della Siberia, il riscaldamento globale ha reso accessibili risorse naturali precedentemente irraggiungibili e ha aperto nuove rotte commerciali marittime, riducendo i tempi di navigazione tra Asia ed Europa. Allo stesso modo, in regioni con climi tradizionalmente rigidi, come alcune parti del Canada settentrionale o della Scandinavia, l'innalzamento delle temperature ha esteso la stagione agricola e migliorato la produttività locale. Tuttavia, è importante considerare che questi vantaggi sono limitati nel tempo e nello spazio e non compensano gli impatti devastanti del riscaldamento a livello globale, come l'innalzamento dei mari, la perdita di biodiversità e l'aumento delle disuguaglianze socioeconomiche (Pachauri Meyer, 2014).

La duplice natura della crisi climatica evidenzia la necessità di affrontarla con un approccio integrato, che tenga conto delle dinamiche locali senza perdere di vista le conseguenze sistemiche e di lungo termine, che devono andare in direzione della

sostenibilità. É, pertanto, necessario chiarire cosa si intenda per sostenibilità, il secondo grande oggetto di ricerca sul quale indaga il presente lavoro.

Il concetto di sostenibilità affonda le radici in quella che può essere definita "la crisi di sviluppo", ovvero il fallimento, successivo alla Seconda Guerra Mondiale, dei piani di sviluppo internazionale che erano stati progettati e messi in campo per ovviare al generale impoverimento della popolazione mondiale (Caradonna, 2014; Purvis Mao Robinson, 2019). Nel 1990, circa il 38% della popolazione mondiale viveva con meno di 1,90 dollari al giorno, e la proporzione delle persone che vivevano in condizioni di estrema povertà riguardava circa 1 persona su 5 (Redclift, 1993). Tale soglia, nel 2022 è stata aggiornata a 2,15 dollari per tenere conto dell'inflazione. Nel 2019, la percentuale era scesa all'8,4%. Tuttavia, eventi recenti hanno interrotto questa tendenza positiva: la pandemia da COVID-19, in particolare, ha invertito i progressi, portando o riportando quasi 100 milioni di persone a vivere con meno di 2 dollari al giorno, e nel 2024 la Banca Mondiale ha stimato che l'8,5% della popolazione globale, pari a circa 692 milioni di persone, vive in condizioni di estrema povertà, ossia con meno di 2,15 dollari al giorno. Inoltre, quasi la metà della popolazione mondiale (il 43,6%) vive con meno di 6,85 dollari al giorno[3], evidenziando una diffusa vulnerabilità economica. Questi dati sottolineano la necessità di intensificare gli sforzi globali per affrontare le sfide legate alla povertà estrema e promuovere uno sviluppo sostenibile ed equo.

Chi si trova in povertà continua a vivere ai limiti della sopravvivenza, in situazioni deprecabili, caratterizzate da malnutrizione, malattie e scarse prospettive per la vita futura. Inoltre, in alcune zone del mondo come l'Asia meridionale, l'Africa subsahariana o il Sud America, questo spesso coincide con un disagio più esteso, radicato nel tessuto economico e sociale del luogo geografico in questione: infatti, si tratta di paesi schiacciati dal debito pubblico, con infrastrutture inadeguate, con un sistema scolastico spesso insufficiente rispetto alla popolazione e un sistema giudiziario carente, che determinano l'acuirsi di situazioni già di per sé critiche in termini di criminalità e violenza. A ciò si aggiunge la generale crisi ambientale ed energetica che comporta la riduzione delle risorse e ha drammatiche ripercussioni su un contesto già complesso. L'emergenza climatica, lungi dal colpire soltanto paesi in via di sviluppo, si ripercuote anche sulle

[3] Tutti i dati qui riportati sono consultabili sul sito della Banca Mondiale: https://data.worldbank.org/indicator/SI.POV.DDAY?locations=1W&start=1984&view=chart (consultato in data 30/12/2024).

nazioni più ricche, dal momento che il prezzo dell'energia aumenta e la biodiversità delle specie diminuisce a causa della distruzione, frammentazione e degradazione degli habitat dovuta alle attività umane e a calamità naturali.

Alla presa di consapevolezza di tale emergenza si deve la prima formulazione del concetto di "sviluppo sostenibile", così definito nel 1987 durante la *World Commission on Environment and Development*. Il documento che venne stilato in quell'occasione prese il nome di *Brundtland Report*, dal nome della presidente della commissione Gro Harlem Brundtland, o *Our Common Future*. Nello specifico, lo sviluppo sostenibile veniva lì definito come "development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (Brundtland, 1987, 41). Gli obiettivi della commissione erano riesaminare i punti critici che riguardavano lo sviluppo ambientale e formulare nuove proposte concrete per affrontarli. Inoltre, rafforzare la collaborazione internazionale per l'ambiente e lo sviluppo, mettere a punto nuove forme di cooperazione che potessero uscire fuori dai canoni preesistenti e influenzare politiche ed eventi nella direzione dei cambiamenti di cui necessita la società. Fondamentale anche l'obiettivo di aumentare la consapevolezza e l'impegno nell'azione da parte di individui, volontari e organizzazioni: "The Commission focused its attention in the areas of population, food security, the loss of species and genetic resources, energy, industry, and human settlements – realizing that all of these are connected and cannot be treated in isolation one from another" (Brundtland, 1987, 347).

Il *report* ha avuto il merito di riconoscere per la prima volta come lo sviluppo delle risorse umane debba essere inteso in termini di riduzione della povertà, uguaglianza tra i generi e ridistribuzione dei beni, e come il connubio tra questi aspetti sia cruciale per formulare strategie per la conservazione dell'ecosistema. Inoltre, esso rappresenta la prima presa di consapevolezza dell'effettiva esistenza di limiti nell'espansione umana e nella crescita economica, dovuta alla finitezza delle risorse ambientali all'interno delle società industrializzate. Il documento attesta come la povertà riduca e acceleri le pressioni sull'ambiente, creando la necessità di trovare un bilanciamento tra economia ed ecologia. Ancora, in esso viene enfatizzata l'importanza di integrare considerazioni che siano non solamente economiche ed ecologiche, ma anche sociali, per valutare in modo oculato le decisioni sullo sviluppo della società. É da queste consapevolezze che sono scaturiti i concetti etici di responsabilità e cura – preoccupazione – nei confronti delle generazioni

future, la cui vita e sopravvivenza è strettamente legata all'evoluzione del rapporto tra la natura e l'umanità. La rapida crescita della popolazione e il relativo consumo delle risorse disponibili, infatti, è risultato in uno squilibrio tra la capacità dei sistemi naturali e le attività umane che funzionano all'interno di tali sistemi. Il *Brundtland Report* suggerisce due imperativi principali per correggere questo squilibrio: in primo luogo, è necessario soddisfare i bisogni fondamentali di tutti gli esseri umani ed eliminare la povertà. In secondo luogo, occorre porre dei limiti allo sviluppo in generale, poiché la natura è finita.

La possibilità di venire incontro ai bisogni base di ciascuno è profondamente legata alla capacità della natura di soddisfare tali bisogni. La tecnologia deve essere sviluppata e applicata con criterio per fornire soluzioni che si accordino agli imperativi delineati sopra, senza, d'altra parte, influire negativamente sulla natura stessa, ovvero compromettendola a causa di un eccessivo consumo delle risorse, o danneggiandola in modo più diretto a causa delle conseguenze negative di alcune tecnologie.

La pubblicazione di *Our Common Future* e il lavoro della *World Commission* posero le basi per l'istituzione della Commissione per lo Sviluppo Sostenibile, che nacque nel 1992 sotto l'Assemblea Generale dell'ONU, e si concretizzò in una conferenza tenutasi a Rio de Janeiro. Durante il cosiddetto *Rio Earth Summit*, infatti, vennero stabilite le linee guida per un piano di sviluppo sostenibile su larga scala, che riguardasse tutti i paesi aderenti all'Unione Europea. Nel corso del *Rio Earth Summit*, inoltre, nove settori della società vennero identificati come i canali principali attraverso i quali facilitare un'ampia partecipazione alle attività delle Nazioni Unite relative allo sviluppo sostenibile. Essi sono chiamati ufficialmente "Gruppi Maggiori" e includono: donne, bambini e giovani, persone indigene, organizzazioni non governative, autorità locali, lavoratori e sindacati, affari e industria, comunità scientifica e tecnologica e contadini. Tra i documenti ratificati, vi furono l'*Agenda 2021* e la *Rio Declaration on Environment and Development*, che iniziarono a tratteggiare il concetto di sviluppo sostenibile, ma non ancora a specificare gli Obiettivi di Sviluppo Sostenibile (OSS), che vennero definiti solo successivamente. Alla Commissione per lo Sviluppo Sostenibile venne assegnato il compito di seguire e monitorare quanto definito dall'*Agenda 2021*, attribuendole di fatto un ruolo chiave nella promozione del dibattito sulla necessità di affinare e approfondire ulteriormente gli OSS e i traguardi da perseguire a livello mondiale.

Nel 2000, le Nazioni Unite adottarono gli Obiettivi di Sviluppo del Millennio (OSM), ovvero un insieme di otto obiettivi che si focalizzavano soprattutto sulla riduzione della povertà e il miglioramento di salute, educazione e ambiente. Nonostante gli OSM abbiano comportato progressi significativi in varie aree, emerse il bisogno di una visione che fosse più estesa e inclusiva. Spinte da questa motivazione, le Nazioni Unite iniziarono le negoziazioni per stabilire quelli che avrebbero sostituito gli OSM, dopo che il termine del 2015, che era stato fissato per la loro realizzazione, fosse trascorso. L'importanza dei nove Gruppi Maggiori fu riaffermata nel 2012 durante una nuova conferenza delle Nazioni Unite sullo Sviluppo Sostenibile, nota come Rio+20. Su questa scia, nel settembre 2015 i leader mondiali adottarono ufficialmente l'Agenda 2030 per lo Sviluppo Sostenibile durante una riunione generale dell'ONU. L'Agenda comprende 17 Obiettivi di Sviluppo Sostenibile (OSS), che coprono una vasta gamma di campi, tra cui: povertà, fame, educazione, genere, uguaglianza, giustizia sociale e azione climatica[4]. Ogni OSS comprende una lista di sotto-obiettivi, ovvero 169 *target*, ciascuno dei quali ha tra uno e tre indicatori di misurazione del progresso, per un totale di 232 indicatori. Uno dei punti di forza degli OSS è che i dati sui 17 obiettivi sono disponibili e comprensibili a quante più persone possibili: ciò avviene grazie al database globale degli indicatori OSS, ovvero una piattaforma che fornisce accesso a tutti i dati compilati e assemblati dal sistema delle Nazioni Unite. Essa raccoglie tutte le informazioni sullo sviluppo e l'implementazione di un quadro di riferimento globale di indicatori, relativo agli obiettivi e ai *target* dell'Agenda di Sviluppo Sostenibile del 2030, gestito dalla divisione di Statistica del Dipartimento di Economia e Affari Sociali delle Nazioni Unite[5].

La piattaforma è inclusa nel *Sustainable Development Report*, ovvero il primo studio globale volto a stabilire la posizione di ciascun paese nel raggiungimento degli OSS. Esso venne preparato da gruppi di esperti indipendenti dalla Rete di Soluzioni sullo Sviluppo Sostenibile delle Nazioni Unite (SDSN): differenti attori nello sviluppo possono

[4] In breve, gli OSS a livello globale sono: ridurre a zero la fame e la povertà; raggiungere benessere fisico e mentale per tutte le popolazioni; assicurare istruzione di qualità; raggiungere l'uguaglianza di genere; acqua potabile; energia pulita e accessibile; assicurare posti di lavoro e crescita economica; sostenere l'industria, l'innovazione e le infrastrutture; ridurre le disuguaglianze; progettare città e comunità sostenibili; consumi e produzione responsabili; azione climatica; sostenere gli ecosistemi di terra e mare; garantire pace, giustizia, istituzioni forti e collaborazione in vista del raggiungimento degli obiettivi stessi. Cfr. *Transforming our World: The Sustainable Development Agenda to 2030* e 2012, *The future we want. Resolution adopted by the general assembly on 27 July 2012*, A/RES/66/288 (New York, 2012), United Nations General Assembly, 2015.

[5] La piattaforma si trova al link: https://unstats.un.org/sdgs/dataportal, consultato in data 16/01/2025.

usare il *report* per identificare priorità nell'azione, comprendere le sfide, monitorare i progressi, assicurare la responsabilità e identificare i punti deboli o i passi mancanti per il raggiungimento degli OSS entro il 2030. Sulla piattaforma è consultabile anche la classifica dei paesi in base ai progressi rispetto agli OSS: la Danimarca si trova in cima alla classifica, con l'indice più alto: 85.2 su 100, mentre il paese con l'indice più basso è la Repubblica Centrale Africana, con un indice di 39.1.

In generale, la tendenza verso il raggiungimento degli OSS rimane preoccupante, così come rilevato dall'ultimo *report* del 2019: sono stati compiuti progressi in alcune aree critiche, inclusa quelle della povertà estrema, che è stata ridotta significativamente, e quella del tasso di mortalità infantile (che considera bambini sotto i 5 anni), che è arrivato al di sotto del 49% tra il 2000 e il 2017. Inoltre, i vaccini hanno salvato milioni di vite, e la maggioranza della popolazione adesso ha accesso all'elettricità (Pedemonte, 2020). Ciononostante, rimangono ancora sfide incalzanti da affrontare. Il tema più urgente di azione è il riscaldamento globale, poiché se non si ha una significativa riduzione delle emissioni dei gas serra il più presto possibile, il riscaldamento potrebbe aumentare di 1.5° nei prossimi decenni, rendendo alcune parti del globo inabitabili e provocando diversi disastri ambientali che colpirebbero soprattutto le fasce più indigenti della popolazione. Conseguentemente, ciò comporterebbe la crescita delle disuguaglianze, sia all'interno delle stesse nazioni sia tra nazioni diverse. Precarietà, fame e rischi per la salute continuano a essere concentrati tra i paesi e i popoli più poveri e vulnerabili.

Da quanto detto finora emerge chiaramente come il concetto di sostenibilità sia esteso e sfaccettato, con molteplici definizioni, e possa essere affrontato e compreso da svariati punti di vista (Gonzalez Coronado Martin Vaca-Tapia, 2021). Nello specifico, quando ci si riferisce alla sostenibilità, la letteratura di solito considera i cosiddetti "tre pilastri": ovvero l'ambiente, l'economia e la società (Purvis Mao Robinson, 2019), che sono compresi anche nella formulazione degli OSS. Nonostante ciò, la concettualizzazione della sostenibilità come oggetto tripartito sembra mancare di fondazioni teoretiche solide. Infatti, essi sono generalmente indicati come i tre aspetti che devono essere presi in considerazione per valutare se qualcosa – un dispositivo, un processo, una comunità – sia sostenibile. Ma, a ben vedere, non sembra esserci un testo originale da cui derivi questa formulazione: sembra che sia solo apparsa in letteratura e

comunemente adottata senza un esteso approfondimento. Ben Purvis e colleghi sottolineano come già nel 2001, questo approccio sia stato presentato come "commonplace throughout the literature" (Purvis Mao Robinson, 2019, 685) per trattare lo sviluppo sostenibile. Così diffuso da non richiedere nemmeno un riferimento. In ogni caso, i pilastri rappresentano delle linee guida teoriche che si dovrebbero seguire per evitare una categorizzazione parziale e vaga di "sostenibile". Si discute, inoltre, della possibilità di aggiungere anche un quarto pilastro, che il più delle volte è trascurato dalla letteratura e che dovrebbe acquisire spazio, ovvero la sostenibilità culturale (Trimarchi, 2004). L'utilità di questo quarto pilastro, infatti, si trova nel preservare e mantenere le credenze culturali delle comunità, nonché le loro pratiche e la conservazione del loro patrimonio attraverso il tempo, tenendo al sicuro la loro memoria (Loach Rowley Griffiths, 2016).

Oltre ai documenti chiave sulla sostenibilità menzionati precedentemente, vale la pena ricordare anche il regolamento *Tassonomia*, pubblicato nella Gazzetta Ufficiale UE il 22 giugno 2020[6], che costituisce uno strumento per valutare le attività economiche e gli investimenti dal punto di vista dell'ecosostenibilità. Questo regolamento rappresenta un pilastro fondamentale nell'ambito della finanza sostenibile, poichè fornisce un linguaggio comune e criteri chiari per determinare quali attività economiche possono essere considerate sostenibili dal punto di vista ambientale. Tale regolamento prevede inoltre l'obbligo per le grandi aziende di registrare e riportare l'impatto ambientale e sociale delle proprie attività, promuovendo una maggiore trasparenza e responsabilità aziendale.

Nello specifico, le aziende devono cercare di allinearsi a sei obiettivi ambientali chiave: la mitigazione dei cambiamenti climatici, l'uso sostenibile e la protezione delle acque e delle risorse marine, la prevenzione e la riduzione dell'inquinamento, l'adattamento ai cambiamenti climatici, la transizione verso un'economia circolare, e la protezione e il ripristino della biodiversità e degli ecosistemi. Questi obiettivi riflettono un approccio olistico alla sostenibilità, riconoscendo l'interconnessione tra diversi aspetti dell'ambiente e l'importanza di affrontare le sfide climatiche in modo integrato. Inoltre, per essere considerata ecosostenibile, un'attività deve contribuire in modo sostanziale a uno dei sei obiettivi delineati sopra, senza arrecare danno significativo a nessuno degli

[6] Consultabile al link: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32020R0852 (consultato in data 4/01/2025).

altri – principio noto come "Do No Significant Harm". Questo principio assicura che gli sforzi per migliorare un aspetto ambientale non compromettano altri obiettivi di sostenibilità, garantendo un approccio equilibrato e coerente. Altro requisito fondamentale per un'attività che sia sostenibile è rispettare le garanzie minime di salvaguardia sociali, come ad esempio quelle stabilite dall'Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE) e i Principi Guida delle Nazioni Unite su Imprese e Diritti Umani. Tali garanzie includono il rispetto dei diritti umani, le norme lavorative internazionali, la lotta alla corruzione e altre pratiche di buona governance (Regolamento L 198/13). Il doppio livello di requisiti, ambientali e sociali, sottolinea l'importanza di una sostenibilità integrata che tenga conto delle dimensioni ecologiche, sociali ed economiche, ovvero dei tre pilastri della sostenibilità che sono stati menzionati in precedenza.

Il regolamento *Tassonomia,* dunque, non solo incentiva le aziende a contribuire attivamente alla sostenibilità ambientale, ma fa anche in modo che essi si impegnino a rendere conto del loro operato in maniera trasparente e a rispettare standard sociali elevati, promuovendo un modello di sviluppo sostenibile che possa fungere da riferimento a livello globale.

Il discorso sulla sostenibilità diventa ancora più complesso se, come accennato precedentemente, si rapporta con il secondo grande gigante dell'età contemporanea: l'IA.

1.2 IA e regolamentazione

Nel dibattito sulle sfide globali che il pianeta si trova ad affrontare, naturalmente le nuove tecnologie sono spesso chiamate in causa, e nello specifico lo è l'IA. Nel 2019, il gruppo di esperti che si occupa di IA all'interno della Commissione Europea ha dichiarato che l'IA è:

> a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation. In particular, AI systems can help to facilitate the achievement of the UN's Sustainable Development Goals, such as promoting gender balance and tackling climate change, rationalising our use of natural resources, enhancing our health, mobility and production processes (High-Level Expert Group on Artificial Intelligence, 2019, 4).

D’altro canto, utilizzare l’IA per sviluppare soluzioni sostenibili senza analizzarne in modo critico il significato e le implicazioni concrete può comportare rischi significativi. Tra questi, emerge il pericolo di adottare un atteggiamento tecno-entusiasta, ovvero, di abbracciare una visione eccessivamente ottimistica e acritica nei confronti dell’impiego della tecnologia (Coget, 2017). La manifestazione del tecno-entusiasmo può culminare in un fallimento nel riconoscere e valutare adeguatamente gli effetti negativi dell’IA. Tali effetti hanno ripercussioni su molteplici dimensioni: quella sociale, quella politica, quella etica e quella ambientale (Floridi, 2022). Le questioni etiche che riguardano l’IA ruotano soprattutto intorno al problema dell’IA responsabile e/o reprensibile che opera nella società in modo sempre più diffuso, con implicazioni molto profonde dal punto di vista morale. Le infrastrutture sociali, infatti, come l’energia e i trasporti pubblici, sono gestite sempre di più da macchine, il cui grado di sofisticatezza e intelligenza aumenta esponenzialmente. Rispetto a tale fenomeno, è d’obbligo domandarsi come vadano distribuite le responsabilità nel caso di incidenti dovuti a errori di sistema, o difetti di progettazione, oppure nel caso di operazioni svoltesi in modo effettivamente corretto, ma al di fuori dei vincoli che erano stati stabiliti e intesi originariamente.

Una delle questioni più dibattute in merito alla *governance* dell’IA è in quale sede dovrebbero essere prese le decisioni relative alla regolamentazione di questi strumenti, se nelle industrie o nei governi. Si tratta di un dilemma complesso, poiché coinvolge interessi diversi, spesso contrastanti. Da un lato, le aziende tecnologiche e le imprese che sviluppano soluzioni basate sull’IA sostengono che la regolamentazione dovrebbe essere determinata in primo luogo da coloro che meglio conoscono le tecnologie, ovvero i leader del settore. L’argomento principale a supporto di tale posizione è che una regolamentazione troppo rigida o imposta dall’esterno potrebbe soffocare l’innovazione e rallentare il progresso tecnologico, minando la competitività globale. D’altro canto, i governi e le istituzioni pubbliche sottolineano la necessità di regole più severe e centralizzate, in particolare per garantire che l’uso dell’IA sia in linea con i principi etici e sociali e per prevenire conseguenze potenzialmente dannose come discriminazioni, violazioni della privacy o automazione del lavoro senza una sufficiente protezione dei lavoratori. Molti sostengono che le decisioni sulla *governance* dell’IA non possono essere attribuite esclusivamente al settore privato, dato il potenziale impatto di queste tecnologie

su questioni sociali fondamentali come i diritti umani, la sicurezza nazionale e la stabilità economica. Una terza opzione, sempre più discussa, è l'idea di una *governance* collaborativa, in cui industrie e governi lavorino insieme per definire standard e regolamenti. Questo approccio mira a trovare un equilibrio tra innovazione tecnologica e responsabilità sociale, consentendo all'IA di svilupparsi in modo etico ed ecologicamente sostenibile (Floridi 2022).

In risposta alla crescente necessità di regolamentare le nuove tecnologie, il 21 maggio 2024 il Consiglio dell'Unione Europea ha approvato l'*AI Act*[7], ovvero il primo regolamento mondiale dedicato all'IA. Questo provvedimento ha l'obiettivo di garantire che i sistemi di IA introdotti sul mercato europeo e utilizzati nell'UE siano sicuri, rispettino i diritti fondamentali e i valori dell'Unione, e promuovano al contempo investimenti e innovazione nel settore in Europa. Uno degli obiettivi principali dell'*AI Act* è garantire la sicurezza e la tutela dei diritti fondamentali, promuovendo al contempo la trasparenza e la tracciabilità dei sistemi di IA. La proposta mira anche a gestire i rischi associati all'IA, stabilendo requisiti rigorosi per i sistemi considerati ad alto rischio, come quelli impiegati in ambiti critici quali la salute, la sicurezza e i diritti fondamentali. Per realizzare questi obiettivi, l'*AI Act* propone di classificare i sistemi di IA in base al loro livello di rischio: inaccettabile, alto, limitato, minimo. Sono considerati inaccettabili, e dunque vietati, i sistemi di IA che rappresentano una minaccia per la sicurezza, i mezzi di sussistenza e i diritti delle persone, come ad esempio i sistemi di *social scoring*[8] che in passato sono stati utilizzati da alcuni governi e i sistemi che utilizzano categorizzazione biometrica basati su caratteristiche sensibili. Un ulteriore esempio di sistema critico appartenente a questa categoria è rappresentato dai modelli che estrapolano indiscriminatamente immagini facciali da Internet o da registrazioni di telecamere a circuito chiuso, e le utilizzano per creare database dedicati al riconoscimento facciale. O ancora, i sistemi di riconoscimento delle emozioni sui luoghi di lavoro o nelle scuole, le pratiche di polizia predittiva − ovvero quelle metodologie che utilizzano analisi dati per

[7] Il testo dell'AI Act è consultabile integralmente al link: https://artificialintelligenceact.eu/the-act/ (ultimo accesso in data 8/01/2025).

[8] Un sistema di IA di *social scoring* è un tipo di sistema di valutazione che utilizza algoritmi di IA per analizzare, valutare e classificare il comportamento sociale, le attività online e offline, le interazioni e altre informazioni personali degli individui. Questi sistemi aggregano e interpretano una vasta gamma di dati per assegnare un punteggio o un *rating* alle persone, che può riflettere la loro affidabilità, credibilità o conformità a determinati standard sociali. Un esempio è il sistema di credito sociale della Cina, che valuta i cittadini sulla base del loro comportamento e conformità alle norme governative.

prevedere dove e quando potrebbero verificarsi crimini e quali individui potrebbero essere coinvolti in attività criminali – e i sistemi che manipolano il comportamento umano o sfruttano le vulnerabilità delle persone. I sistemi a rischio alto sono i sistemi di IA utilizzati in settori critici, come trasporti, sanità e applicazione della legge, che richiedono una valutazione rigorosa della conformità e misure di mitigazione dei rischi. La terza categoria comprende poi i sistemi a rischio limitato, ovvero quelli che richiedono obblighi di trasparenza specifici, come ad esempio le chatbot che devono dichiarare di non essere umane, e ricordarlo all'utente a più riprese durante l'interazione. Sono invece a rischio minimo quei sistemi che impattano poco o per nulla sui diritti e la sicurezza delle persone, che dunque non richiedono misure specifiche (UE 2024, *AI Act*).

In generale, le aziende e gli sviluppatori di IA dovranno conformarsi a vari requisiti imposti dall'*AI Act*, che includono valutazioni di conformità prima dell'immissione sul mercato, sorveglianza e monitoraggio post-commercializzazione, creazione di documentazione tecnica e reportistica dettagliata e adozione di misure di gestione del rischio e garanzia di trasparenza. Lo scopo ultimo del documento è creare un ambiente di fiducia per l'adozione dell'IA e al contempo proteggere i diritti dei cittadini. Tuttavia, la realizzazione delle norme non è banale, e presenta sfide significative: ad esempio, garantire che le normative non ostacolino l'innovazione, bilanciare la protezione dei diritti con la promozione della competitività industriale e affrontare le diverse interpretazioni e applicazioni nei vari Stati membri dell'UE. L'*AI Act* riprende alcuni principi del Regolamento Generale sulla Protezione dei Dati (GDPR), che è in vigore dal maggio 2018. Esso si applica a tutte le organizzazioni che trattano i dati personali di cittadini dell'UE, con l'obiettivo di garantire la protezione dei dati personali e la privacy degli individui. Tra i principi chiave del GDPR, si trova la garanzia di liceità, correttezza e trasparenza dei dati, che devono essere raccolti per finalità determinate e legittime, e circoscritti a quanto necessario. Inoltre, il GDPR sancisce che i dati possono essere conservati solo per un periodo di tempo limitato e trattati in modo da garantire sicurezza adeguata e assicurare i diritti degli individui.

L'*AI Act* e il GDPR si intersecano sui temi di trasparenza e consenso, in quanto entrambi i documenti insistono sulla necessità di informare gli utenti su come i dati vengono raccolti e utilizzati, e nel caso dell'*AI Act* viene sottolineato l'obbligo per i sistemi di IA di dichiarare la propria natura di tecnologie artificiali, in modo da assicurare

che l'utente sia costantemente conscio di stare interagendo con un'entità virtuale, e se ne assuma i rischi. Inoltre, come il GDPR prevede valutazioni di impatto sulla protezione dei dati per il trattamento di quei dati che possono comportare un alto rischio per i diritti e le libertà individuali, così l'*AI Act* introduce valutazioni di conformità specifiche per i sistemi di IA ad alto rischio, garantendo che essi non compromettano la sicurezza o i diritti fondamentali. Sia il GDPR sia l'*AI Act*, inoltre, prevedono sanzioni significative per le violazioni delle normative sulla protezione dei dati e per il non rispetto delle norme sull'IA per garantire la conformità e proteggere i diritti dei cittadini. Alla luce di ciò, è fondamentale garantire che l'*AI Act* e il GDPR siano coerenti e complementari, evitando conflitti normativi e assicurando un'applicazione armoniosa delle leggi.

L'integrazione dei due documenti è cruciale per creare un quadro normativo completo che protegga i diritti dei cittadini dell'UE e al contempo promuova l'uso responsabile e sicuro dell'IA. L'*AI Act*, inoltre richiama anche le *Ethics guidelines for trustworthy AI* sviluppate nel 2019 dal Gruppo di Esperti di Alto Livello sull'Intelligenza Artificiale (AI HLEG) nominati dalla Commissione Europea. Nelle suddette linee guida sono tratteggiati sette principi etici non vincolanti per l'IA, che hanno lo scopo di fornire un riferimento che assicuri che l'IA sia affidabile ed eticamente forte. Tali principi includono l'azione e la sorveglianza umane; robustezza e sicurezza tecniche; privacy e *data governance*; trasparenza; diversità, non discriminazione e giustizia (*fairness*); responsabilità e benessere sociali e ambientali. Nello specifico, con azione e sorveglianza umane si intende che i sistemi di IA devono essere sviluppati e usati come strumenti al servizio delle persone, in modo da rispettare la dignità umana e l'autonomia personale, da funzionare in modo che possa essere controllato e supervisionato dagli esseri umani. Robustezza e sicurezza significano che i sistemi di IA devono essere sviluppati e usati così da resistere nel caso di problemi tecnici e minimizzare il rischio che vengano utilizzati per scopi scorretti. Con tali termini si intende inoltre resilienza contro attacchi malevoli esterni, che mirino ad alterare le performance dei sistemi a beneficio di terze parti. I sistemi devono poi incontrare determinati criteri di privacy e *data governance*, nel senso che vanno sviluppati e utilizzati in accordo con specifiche regole di privacy e protezione dei dati e al contempo processare informazioni e dati che incontrino alti standard di qualità e integrità. Con trasparenza si indica che i sistemi di IA devono accordarsi anche a criteri di appropriata tracciabilità e spiegabilità (*explainability*),

ovvero, i processi degli algoritmi devono essere chiaramente comprensibili in tutte le fasi, da quella della raccolta dati a quelle dell'elaborazione, attraverso l'addestramento e la fase di test, per tutti i cicli previsti, fino all'output finale. Inoltre, come indicato esplicitamente dall'*AI Act*, è fondamentale che i sistemi dichiarino la propria natura di agenti artificiali, in modo che gli utenti siano costantemente informati e consapevoli del tipo di interazione in cui sono coinvolti. Ancora, gli utenti devono essere anche messi al corrente delle capacità e limitazioni di quello specifico sistema, e avere ben presenti i propri diritti. Con diversità, non discriminazione e giustizia si intende infine il fatto che tali sistemi vanno sviluppati e utilizzati così da includere attori differenti e promuovere un uso che sia il più possibile inclusivo, senza distinzioni di ceto sociale, razza, cultura, genere.

Una sfida cruciale per tali tecnologie è infatti rappresentata dal problema dei *bias*, ovvero la presenza di "systematic misrepresentations, attribution errors, or factual distortions that result in favoring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns" (Ferrara, 2023, 2). Nel contesto dei sistemi di IA, la propagazione dei *bias* dipende in larga parte dai dati che vengono forniti durante la fase iniziale per l'addestramento. Si distinguono diversi tipi di *bias*, che appartengono a varie categorie: le più comuni sono quelle dei *bias* demografici, culturali, linguistici, temporali, ideologici e politici. A tale concetto si lega la necessità di perseguire il benessere sociale e ambientale: i sistemi di IA, infatti, devono essere sviluppati e utilizzati in modo sostenibile rispetto all'ambiente, nonché a beneficio degli esseri umani, tenendo conto degli impatti sugli individui e la società sul lungo periodo. In generale, l'applicazione di tali principi dovrebbe guidare l'implementazione dei modelli di IA e "all stakeholders, including industry, academia, civil society and standardisation organisations, are encouraged to take into account as appropriate the ethical principles for the development of voluntary best practices and standards" (*AI Act*, 28).

In una tendenza per certi versi opposta all'UE, Negli USA, la nazione con la maggiore concentrazione al mondo di aziende e industrie che puntano sull'IA, nel settembre 2024 il governatore della California[9] Gavin Newsom ha scelto di non sottoscrivere il *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*

[9] La Silicon Valley, in California, è sede di 32 delle 50 compagnie attualmente leader nel settore dell'IA.

SB-1047[10], che puntava a regolare i modelli di IA più avanzati. Nello specifico, si riferiva a quelli addestrati

> using a quantity of computing power greater than 10^26 integer or floating-point operations, the cost of which exceeds one hundred million dollars ($100,000,000) when calculated using the average market prices of cloud compute at the start of training as reasonably assessed by the developer (*Senate Bill 1047*).

Tra le disposizioni in materia di sicurezza, vi era suggerita l'implementazione di misure di sicurezza prima di iniziare l'addestramento del modello e la possibilità per il procuratore generale della California di chiedere un provvedimento ingiuntivo, che facesse cessare a un'azienda le operazioni legate all'IA ritenute pericolose, nonché di citare in giudizio uno sviluppatore nel caso in cui il suo modello causasse un evento catastrofico. Tra gli oppositori al provvedimento figurano *OpenAI*, *Meta*, *Google* e *Microsoft* e l'*AI Alliance*[11], che ha dichiarato che il disegno di legge rallenterebbe l'innovazione, ostacolerebbe i progressi in materia di sicurezza e comprometterebbe la crescita economica della California, oltre a ridurre su scala globale l'accesso alle tecnologie all'avanguardia del settore. Il governatore Newsom ha deciso di non procedere con il disegno di legge sostenendo che, essendo limitato a pochi modelli di IA avanzati, esso avrebbe generato solamente un falso senso di sicurezza nella popolazione. Inoltre, nella dichiarazione di veto sul *Senate Bill 1047* da parte del governatore si legge che: "to keep the public safe, we must settle for a solution that is not informed by an empirical trajectory analysis of Al systems and capabilities. Ultimately, any framework for effectively regulating Al needs to keep pace with the technology itself"[12]. In un certo senso, dunque, sembra che in materia di regolamentazione dell'IA la California abbia assunto una posizione opposta a quella dell'UE (che ha approvato l'*AI Act*), lasciando da

[10] Il testo del *Senate Bill 1047* è consultabile in versione integrale sul sito delle informazioni legislative della California, al link https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047 (consultato in data 3/10/2024).
[11] L'*AI Alliance* "is a collaborative network of more than 100 companies, startups, universities, research institutions, government organizations, and non-profit foundations that are working at the forefront of AI technology, applications, and governance" (dal loro sito web: https://thealliance.ai/, consultato in data 3/10/2024). Tra i membri figurano *Meta*, *Sony*, *IBM*, il Cern, *Anaconda* e la Cornell University.
[12] La dichiarazione del governatore Newsom è consultabile in versione integrale sul sito delle informazioni legislative della California, al link: https://www.gov.ca.gov/wp-content/ uploads/2024/09/SB-1047-Veto-Message.pdf (consultato in data 3/10/2024).

parte i veti più rigidi e concentrandosi sulla necessità di implementare misure di trasparenza per i modelli.

1.3 IA per la sostenibilità

Alla luce delle crescenti preoccupazioni legate alle risorse energetiche e computazionali richieste da queste tecnologie, diventa sempre più urgente orientare l'implementazione dell'IA verso soluzioni intrinsecamente sostenibili. Si è visto che per affrontare in modo efficace il complesso intreccio tra IA e sostenibilità, è essenziale definire con precisione il concetto di IA sostenibile (Dauvergne, 2020). Ciò richiede la distinzione tra l'IA utilizzata come strumento per raggiungere obiettivi di sostenibilità e l'IA progettata per essere in sé sostenibile. Da un lato, infatti, l'IA può rappresentare un alleato strategico nel contrastare gli effetti devastanti della crisi ambientale, ad esempio supportando interventi per mitigare i cambiamenti climatici e i loro impatti sulla vita umana e sulla biodiversità. Dall'altro, la ricerca deve necessariamente affrontare il tema della sostenibilità delle tecnologie di IA, considerando aspetti quali il consumo energetico, i costi legati allo smaltimento dei materiali e l'impronta ecologica[13]. Oltre agli aspetti ambientali, sui quali si concentra maggiormente il dibattito, è importante non trascurare gli altri due pilastri fondamentali: l'aspetto sociale e quello economico. Per affrontare adeguatamente la relazione tra IA e sostenibilità, è essenziale dunque considerare questi tre elementi in modo integrato, definendo una scala di priorità che eviti approcci superficiali e valuti in maniera equilibrata i presunti benefici di tali tecnologie.

Come sottolinea Aimee van Wynsberghe, l'IA sostenibile dovrebbe costituire "a movement to foster change in the entire lifecycle of AI products towards greater ecological integrity and social justice" (van Wynsberghe, 2021, 217). In questa prospettiva, si possono individuare estesissime applicazioni dell'IA nel campo della sostenibilità, soprattutto a seguito dell'introduzione dei 17 OSS (Ferilli Girardi, 2021). Tra le iniziative più significative legate all'implementazione degli OSS spicca il progetto *AI for Good,* una piattaforma digitale gestita dall'Unione Internazionale delle

[13] L'impronta ecologica rappresenta la quantità di risorse naturali necessarie per sostenere il tenore di vita di un individuo, una comunità o una nazione, oltre che per assorbire i rifiuti prodotti, ed è generalmente espressa in ettari per quantificare la terra e l'acqua richieste per produrre e smaltire tali risorse.

Telecomunicazioni dell'ONU. Questo spazio virtuale è stato concepito come luogo di confronto per esperti e innovatori, mirato alla discussione e all'individuazione di soluzioni concrete per il raggiungimento degli OSS. Un contributo fondamentale in questo ambito è offerto dal lavoro di Luciano Floridi sull'etica dell'IA in relazione alla sostenibilità (Mazzi Floridi, 2023). Alla base dell'idea di IA per il bene comune risiede la speranza che lo sviluppo e l'uso dell'IA possano avere un impatto positivo su individui, società e ambiente (Floridi, 2019). Numerose sono, infatti, le applicazioni dell'IA orientate al bene sociale, che trovano spazio in molteplici settori (Floridi, 2020). Tuttavia, è altrettanto noto che l'IA può essere utilizzata in modo improprio o eccessivo, o persino per scopi non etici (King Aggarwal Taddeo, 2020).

L'analisi etica rappresenta dunque un elemento imprescindibile per orientare lo sviluppo dell'IA verso il bene sociale. Questo tipo di riflessione deve essere sostenuto sia dalle aziende private sia dai governi che integrano l'IA nelle loro strategie nazionali. Il principio cardine è quello della beneficenza, che impone che l'IA debba generare benefici concreti tanto per il pianeta, quanto per le persone, promuovendo soluzioni sostenibili che rispettino sia l'ambiente sia il tessuto sociale (Mazzi Floridi, 2023).

Le potenziali applicazioni dell'IA in relazione agli OSS evidenziano il suo ruolo cruciale nella comprensione dei problemi, nella ricerca di soluzioni innovative e nel supporto ai processi decisionali. Uno studio del McKinsey Global Institute, pubblicato nel novembre 2018, ha individuato 135 casi d'uso dell'IA a livello globale per sostenere gli OSS (Chui Harrysson Manyika, 2018). Tra questi, l'obiettivo con il maggior numero di applicazioni è l'OSS n. 3, "Salute e benessere", che conta 29 casi identificati. La maggior parte di queste applicazioni si concentra sui temi di "Salute e fame", con specificità come "Predizione e Prevenzione" e "Cura e Trattamento", per un totale di 28 esempi. Al contrario, l'area meno rappresentata è quella della "Verifica e Validazione dell'Informazione", con soli quattro esempi di utilizzo dell'IA, come strumenti contro la disinformazione e le *fake news*.

Inoltre, una revisione più recente della letteratura, condotta da Agung Bella Putra Utama e colleghi (2024), conferma che gli OSS affrontati mediante tecniche di apprendimento profondo (*deep learning*, DL) includono l'OSS n. 3, "Salute e benessere", con applicazioni per la previsione di epidemie e la diagnosi precoce, l'OSS n. 7, "Energia pulita e accessibile", con previsioni della domanda energetica e ottimizzazione delle

energie rinnovabili, l'OSS n. 11, "Città e comunità sostenibili", con attività di monitoraggio della qualità dell'aria, gestione del traffico e uso del suolo, e l'OSS n. 13, "Lotta contro il cambiamento climatico", con previsioni di eventi naturali come alluvioni e velocità del vento (Utama Wibawa Handayani, 2024).

Questi risultati sottolineano il potenziale trasformativo dell'IA nel contribuire a molteplici obiettivi globali, con particolare attenzione ai settori della salute, dell'energia sostenibile, dello sviluppo urbano e del cambiamento climatico. Tuttavia, è importante non confondere tali dati con una reale valutazione delle potenziali applicazioni dell'IA in un'area o in un obiettivo specifico. Piuttosto, rappresentano un valido indicatore dell'evoluzione e della scoperta delle possibilità offerte dall'IA. In generale, il primo e più diffuso impiego di IA per gli OSS è stato nel campo della diagnostica medica per immagini, principalmente indirizzato all'OSS "Salute e benessere". A partire da gennaio 2015, infatti, non meno di ⅓ delle *startup* di IA nel settore sanitario ha stanziato fondi per investire nella diagnostica per immagini (Varadharajan Lee, 2018.

Anche Ricardo Vinuesa e colleghi (2020) hanno discusso ampiamente le come l'IA può incoraggiare o, al contrario, inibire il raggiungimento dei 17 OSS stabiliti dall'Agenda per il 2030. È interessante notare come all'inizio del loro lavoro si sincerino di partire dalla definizione di IA, dal momento che in letteratura mancava ancora una definizione che non fosse ambigua:

> We consider as AI any software technology with at least one of the following capabilities: perception – including audio, visual, textual and tactile (e.g., face recognition), decision-making (e.g., medical diagnosis systems), prediction (e.g., weather forecast), automatic knowledge extraction and pattern recognition from data (e.g., discovery of fake news circles in social media), interactive communication (e.g., social robots or chat bots), and logical reasoning (e.g., theory development from premises). This view encompasses a large variety of subfields, including machine learning (Vinuesa Hossein Leite, 2020, 1).[14]

Vinuesa e colleghi procedono con una panoramica dei modelli di apprendimento automatico (*machine learning*, ML) attualmente usati nell'ambito del raggiungimento degli OSS. Ad esempio, alcuni tipi di modelli ibridi basati sulla cosiddetta regressione di

[14] La definizione di IA di Vinuesa e colleghi riprende alcuni dei criteri delineati da Allen Newell per la definizione di un'architettura cognitiva, che verranno discussi nel §2.1.

vettori a supporto (SVR)[15], oppure su ottimizzazione per sciami di particelle (PSO)[16] possono essere usati per predire l'utilizzo di energia atteso da dati forniti. Per predizioni sul cambiamento climatico vengono utilizzati anche alcuni tipi di algoritmi predittivi basati su reti neurali ricorrenti, particolarmente appropriati nella computazione di serie temporali. Si tratta di reti che hanno capacità di memoria per periodi di tempo sia brevi sia lunghi e che presentano una certa robustezza (Sirmacek Gupta Mallor, 2023). In modo simile, nella progettazione delle cosiddette *smart cities*, in particolare per compiti che richiedono capacità di generalizzazione e gestione e modellazione dei trasporti automatizzati, trovano largo impiego le reti generative avversarie (GAN)[17] che apprendono rappresentazioni profonde senza dover fare uso di una grande quantità di dati già etichettati. Ad esempio, la rilevazione del traffico è un compito molto complesso nella modellazione e gestione dei trasporti, e i ricercatori si servono soprattutto di GAN per farlo.

Per quanto riguarda le problematiche sociali, sono state sviluppate applicazioni di IA per fronteggiare l'OSS n. 1 ("Riduzione della povertà"), come un modello di ML di classificazione per aiutare a tenere traccia e monitorare la povertà in alcuni paesi specifici (Alsharkawi Al-Fetyani Dawas, 2021). Sulla stessa linea sono i progetti che puntano a occuparsi dell'OSS n. 3 ("Salute e benessere"): come visto, l'utilizzo di IA in campo medico si sta diffondendo sempre nella diagnostica per immagini: la diagnostica dei tumori è molto dibattuta (Alshuhri Al-Musawi Al-Alwany, 2023), mentre sembra che sia più affidabile quella della retinopatia diabetica, che è una delle cause principali di cecità.

[15] La regressione a vettori di supporto (*Support Vector Regression*, SVR) è un algoritmo di apprendimento automatico utilizzato per la predizione di valori numerici. Si basa sui concetti della regressione lineare e utilizza vettori di supporto per trovare la relazione tra le variabili di input e i relativi valori di output. L'obiettivo è ottenere una funzione che approssimi i dati di addestramento con la massima precisione possibile, mantenendo al contempo una certa tolleranza rispetto agli errori. In breve, SVR è una tecnica di regressione che sfrutta i principi dei vettori di supporto per modellare e predire valori numerici.

[16] L'ottimizzazione per sciami di particelle (*Particle Swarm Optimization*, PSO) è un algoritmo di ottimizzazione metaeuristica ispirato al comportamento sociale degli uccelli in stormo o dei pesci in banco. In PSO, una popolazione di soluzioni potenziali, chiamate particelle, si muove attraverso lo spazio di ricerca per trovare la soluzione ottimale. Ogni particella regola la propria posizione in base alla propria esperienza (migliore personale) e all'esperienza dell'intero sciame (migliore globale). Attraverso l'aggiornamento iterativo delle loro posizioni, le particelle convergono verso la soluzione ottimale nel tempo.

[17] Le reti generative avversarie (*Generative Adversarial Networks*, GAN) sono un tipo di architettura di reti neurali artificiali composte da due reti neurali – un generatore e un discriminatore – che lavorano in opposizione. Il generatore cerca di creare campioni che somiglino ai campioni provenienti dall'insieme di dati di addestramento, mentre il discriminatore cerca di distinguere tra campioni reali e falsi. Le GAN sono utilizzate per generare dati sintetici, come immagini, suoni o testo, che possono essere indistinguibili dai dati reali.

I ricercatori di *Google* hanno sviluppato un algoritmo di IA capace di rilevare i primi sintomi da immagini della retina, con un'accuratezza che supera quella degli esperti nel campo (Upreti Singh Nagpal, 2023). D'altra parte, la precisione della diagnosi dipende dalla qualità delle immagini, e questo può rappresentare un grosso ostacolo soprattutto nelle regioni del mondo dove le risorse sono limitate. Inoltre – e questo vale in generale per tutte le IA usate in campo medico – è molto problematico individuare il modo giusto per inserire l'utilizzo di tali strumenti all'interno del normale flusso di lavoro nel campo sanitario. Lo stesso vale per la radiologia, dove l'IA può essere usata per l'interpretazione di immagine mediche ottenute tramite raggi X, TAC e risonanza magnetica. In particolare, i ricercatori dell'università di Stanford sono riusciti a sviluppare un algoritmo capace di diagnosticare la polmonite da radiografie del petto. Tale algoritmo ha raggiunto performance paragonabili a quelle di radiologi esperti e ha dimostrato le potenzialità del supportare i professionisti della sanità nel fornire diagnosi più rapide e accurate. Anche in questo caso, un grande limite è rappresentato dalla necessità di assicurare la trasversalità del funzionamento dei diversi dispositivi, in quanto la varietà degli strumenti disponibili e i diversi formati delle immagini possono entrare in conflitto tra loro. Infine, un'altra applicazione dell'IA per la salute riguarda l'evitare ricoveri reiterati. In questo caso, le analitiche predittive basate su IA giocano un ruolo cruciale nell'identificare i pazienti con un alto rischio di ricovero. In uno studio condotto dai ricercatori della University of Chicago Medicine, è stato sviluppato un modello che prevede quali pazienti hanno probabilità di essere riammessi entro 30 giorni dopo essere stati dimessi, analizzando le cartelle cliniche elettroniche. Il modello di IA ha ottenuto risultati migliori rispetto ai metodi tradizionali di valutazione del rischio, offrendo approfondimenti per interventi mirati e migliori strategie di gestione delle cure. Questo caso dimostra come l'analisi predittiva basata sull'IA possa aiutare le organizzazioni sanitarie a ridurre i tassi di riammissione e a migliorare l'uso efficiente delle risorse sanitarie (Upreti Singh Nagpal, 2023).

Inoltre, vale la pena menzionare gli sforzi compiuti per ovviare all'emergenza della mancanza di acqua (OSS n. 6), attraverso l'implementazione di modelli di ML per predire la pressione idrica in alcune aree del pianeta, in modo da ottimizzare il lavoro di pompaggio dell'acqua per evitare sprechi o altri tipi di danni che nuocerebbero alle comunità locali. Tale modello in particolare venne implementato nei laboratori cinesi

dell'azienda *SAP*[18], e include anche la possibilità di seguire in tempo reale lo stato generale del sistema e simulare i risultati delle diverse azioni. Un altro modello correlato era già in uso per generare avvisi di allerta basati sulle tendenze dei consumatori (Pedemonte, 2020). *SAP* ha sviluppato anche un modello di regressione (ovvero un modello predittivo di variabili quantitative basato su ML) per analizzare i fattori chiave dell'equità retributiva usando i dati tratti dalla gestione del capitale umano svolto dai reparti di risorse umane. La realizzazione di tale progetto ha permesso la creazione di una relazione sull'equità salariale, che riferisce quali sono i fattori principali per raggiungerla e in che modo si possono mitigare gli eventuali ostacoli o *bias* che ne limitano la realizzazione. Lo scopo del progetto va incontro all'OSS n. 5 che concerne l'uguaglianza di genere.

Ancora, modelli di IA sono stati implementati per compiti che includono processi decisionali più generali per la gestione delle risorse idriche (Bromley, 2005; Phan Smart Capon, 2016). Per i processi decisionali, è molto diffuso l'uso di algoritmi basati su reti bayesiane[19] nell'analisi statistica dei dati, come nel caso dell'approccio usato da Leonardo Sierra e colleghi (2018) per supportare processi decisionali nella progettazione di infrastrutture in modo da ottimizzare la sostenibilità sociale (Mazzi Floridi, 2023). David Requejo-Castro (2021) ha proposto un approccio simile in riferimento all'OSS n. 6: il suo lavoro combina opinioni di esperti e dati quantitativi a supporto di processi decisionali informati. Nello specifico, si serve di algoritmi di apprendimento basati su reti bayesiane per replicare quadri concettuali composti da vari indicatori, che rappresentano le conoscenze degli esperti e identificano le interconnessioni associate a un contesto complesso, insieme alla tecnica di *bootstrapping*[20], per ridurre l'incertezza dei risultati, e

[18] *SAP* (*Systems, Applications and Products*) è una società multinazionale tedesca con sedi in tutto il mondo, attualmente uno dei principali produttori mondiali di *software* applicativi per la centralizzazione della gestione dei dati e il miglioramento dei processi aziendali.

[19] Una rete bayesiana, come verrà approfondito nel Capitolo III, è un modello grafico probabilistico che rappresenta le relazioni causali tra diverse variabili utilizzando un grafo diretto aciclico. In una rete bayesiana, i nodi del grafo rappresentano le variabili, mentre gli archi rappresentano le dipendenze probabilistiche tra di esse. Questo modello consente di rappresentare e inferire le relazioni di dipendenza probabilistica tra le variabili, facilitando il processo di ragionamento e decisione in contesti complessi.

[20] Il *bootstrapping* è una tecnica statistica di campionamento che consiste nell'estrazione ripetuta di campioni con sostituzione dai dati disponibili. Tale metodo viene utilizzato quando non si dispone di un grande numero di dati campionati o quando è necessario stimare la distribuzione di un parametro statisticamente, senza fare ipotesi sulla sua distribuzione teorica. Con il *bootstrapping* è possibile ottenere stime robuste di varianza, errori standard e intervalli di confidenza per i parametri di interesse senza dover fare alcuna assunzione sulla forma della distribuzione dei dati.

a un'analisi completa della robustezza dei risultati. I risultati, validati sull'OSS n. 6, dimostrano che questo approccio combinato migliora la capacità di inferenza del modello, identifica le interconnessioni tra le variabili considerate e può essere utile per l'analisi delle complessità in diversi contesti. Per quanto riguarda i modelli bayesiani, interessante è anche il lavoro di Juhwan Kim e colleghi (2018), che hanno sviluppato un metodo statistico che combina l'analisi delle reti sociali e la modellazione bayesiana. Lo scopo è fornire un modello gerarchico che può essere applicato per comprendere e misurare la sostenibilità di una determinata tecnologia di IA. In questo caso, Kim e colleghi hanno utilizzato come parametro di riferimento i codici della Classificazione Internazionale dei Brevetti dei documenti brevettuali relativi alle tecnologie IA considerate sostenibili, e la valutazione finale dei risultati è stata affidata agli esperti del dominio (Kim Jun Jang, 2018).

Interessanti sono anche le applicazioni di IA che riguardano l'OSS n. 16, ovvero "Pace, giustizia e istituzioni forti": nello specifico, l'OSS n. 16 è dedicato alla promozione di società pacifiche ed inclusive ai fini dello sviluppo sostenibile, e si propone inoltre di fornire l'accesso universale alla giustizia e a costruire istituzioni responsabili ed efficaci a tutti i livelli. A questo proposito, l'Istituto per la ricerca sul disarmo delle Nazioni Unite (UNIDIR) sta sviluppando strumenti di IA per monitorare il traffico di armi e i flussi finanziari illeciti. Queste tecnologie consentono un'analisi rapida e approfondita dei dati, migliorando le capacità di rilevamento, prevenzione e contrasto di attività che minano la stabilità internazionale e promuovendo un ambiente più sicuro e giusto per tutti. L'adozione di tali soluzioni innovative potrebbe avere un impatto significativo sul miglioramento della *governance* globale e sulla lotta contro la corruzione, rafforzando la cooperazione internazionale nella promozione della pace e della sicurezza. Anche l'organizzazione indipendente *International Crisis Group*[21] si muove verso la realizzazione dell'OSS n. 16 esplorando gli usi dell'IA per la mediazione e la risoluzione

[21] L'*International Crisis Group* (ICR) è un'organizzazione internazionale indipendente e non-profit dedicata alla prevenzione e risoluzione dei conflitti, fondata nel 1995. L'obiettivo dell'ICR è fornire analisi approfondite e raccomandazioni politiche per prevenire guerre e gestire situazioni di crisi nei punti più critici del mondo. Il *Crisis Group* svolge un'importante funzione di monitoraggio e analisi delle situazioni di conflitto, con un approccio multidisciplinare che coinvolge esperti locali, diplomatici, politici e accademici. L'organizzazione pubblica regolarmente rapporti e *briefing* sui conflitti in corso, cercando di influenzare le politiche di governi, organizzazioni internazionali come le Nazioni Unite e l'Unione Europea, oltre a proporre soluzioni diplomatiche e negoziali.

di conflitti. Un'applicazione in questo senso è *PaxAI*, un software che analizza dati dai social network per identificare segnali di potenziali tensioni o violenze in Nigeria (Marwala, 2023).

Esistono poi tecnologie che possono contribuire simultaneamente a più OSS. Un esempio significativo è rappresentato da *M-Pesa*, una piattaforma per telefoni cellulari lanciata da Vodafone in Kenya e Tanzania. Questo sistema consente di caricare e trasferire denaro, oltre a facilitare operazioni bancarie di vario genere, risultando particolarmente utile in aree prive di accesso a servizi bancari tradizionali (Mbiti Weil, 2011). Tra le varie funzioni di *M-pesa* c'è la possibilità di pagare in modo agile le assicurazioni sanitarie, e in questo senso tale strumento si rivela utile sia per la realizzazione dell'OSS n. 3 su salute e benessere, sia per la realizzazione dell'OSS n. 6 ("Imprese, innovazione e infrastrutture"), che riguarda "costruire una infrastruttura resiliente e promuovere l'industrializzazione e una innovazione equa, responsabile e sostenibile". Nella stessa direzione si collocano altri sistemi basati su IA impiegati in Ruanda e Uganda, come *Zipline,* un drone utilizzato per la consegna di forniture mediche essenziali, tra cui sangue e altri presidi sanitari, in aree difficili da raggiungere (Ackerman Koziol, 2019). Questo sistema ha rivoluzionato il trasporto di materiali sanitari, garantendo tempestività e precisione nelle consegne, soprattutto in situazioni di emergenza.

Svariati altri lavori di recente hanno esplorato le applicazioni di IA per gli OSS. Tra quelli rilevanti, vale la pena citare Kumar Kar e colleghi (2022), che hanno revisionato 287 articoli selezionati da un corpus cospicuo e analizzato il loro potenziale nel contribuire allo sviluppo sostenibile della società, citando vari casi. Osama Nasir e colleghi (2021) hanno analizzato in quale misura gli OSS vengono affrontati dalla letteratura sull'IA. Hanno fatto ciò estraendo dati da Scopus attraverso l'uso di parole chiave come "AI", "machine learning", "SDG", e usando gli *abstract* e le affiliazioni degli autori per mapparne geograficamente la distribuzione nel mondo. Ivan Palomares e colleghi (2021) hanno tracciato una panoramica dei processi della relazione tra IA e OSS considerando diverse aree, ovvero "knowledge", "representation", "natural language processing", "computer vision", "machine learning", "automated reasoning" (De Magistris Del Bimbo 2023). Josh Cowls e colleghi (2019) hanno presentato un *dataset* di riferimento che raccoglie tutti i progetti di IA che si riferiscono agli OSS. Un lavoro simile

è stato portato poi avanti dall'*International Communication Union*, che ha lanciato una *repository* globale per identificare i progetti relativi a IA che possono accelerare il processo verso il raggiungimento degli OSS.

1.4 Robot per la sostenibilità

Un ambito strettamente connesso all'IA per la sostenibilità è quello della robotica, che sta assumendo un ruolo sempre più centrale nel supportare e promuovere lo sviluppo sostenibile. L'impiego dei robot in questo contesto si sta rapidamente diffondendo grazie alla loro capacità di affrontare una vasta gamma di sfide, contribuendo in modo significativo a diversi settori. Tra le applicazioni più rilevanti vi è l'ottimizzazione dei processi produttivi, che consente di ridurre gli sprechi e migliorare l'efficienza energetica (Galati Mantriota Reina, 2022), nonché la gestione delle risorse naturali, come la monitorizzazione di ecosistemi, la conservazione della biodiversità e il controllo degli effetti del cambiamento climatico (Gadd De Martini Pitt, 2024). Oltre agli aspetti tecnici, la robotica si distingue anche per il suo potenziale nel sensibilizzare e educare il pubblico sui temi ambientali. Attraverso l'uso di robot interattivi e strumenti educativi innovativi, infatti, è possibile coinvolgere persone di tutte le età in iniziative volte a promuovere comportamenti più sostenibili (Scheutz Law Scheutz, 2021).

Per quanto riguarda l'agricoltura, i robot sono utili nel monitoraggio delle condizioni del suolo e della salute delle piante e degli animali e per adattare le azioni da compiere a ciascun caso diverso – anche pianta per pianta ove necessario – oltre ad avere ruoli nella coltivazione e raccolta delle colture (Oliveira Moreira Silva 2021). Analogamente, possono monitorare i parametri di inquinamento dell'aria e dell'acqua, nonché aumentare la resa nella produzione alimentare. Ad esempio, un robot usato per la mungitura può aumentare il numero di litri prodotti al giorno da una mucca, poiché la mucca può servirsi del robot in qualsiasi momento (Bugmann Siegal Burcin, 2011).

Una direzione di ricerca emergente è poi la robotica sociale per la sostenibilità, ovvero l'applicazione di tecnologie di robotica sociale che sfruttano modalità di comunicazione multimodali basate su indicazioni sociali (come emozioni e linguaggio del corpo) per promuovere comportamenti sostenibili tra gli utenti e contribuire al raggiungimento degli OSS (Alfieri Fleres Damiano, 2022). Questo tipo di robot sociali

sono progettati per interagire in modo naturale con gli esseri umani, influenzando le loro abitudini e comportamenti in un'ottica di sostenibilità.

La maggior parte della ricerca sulla sostenibilità si concentra sul cambiamento delle abitudini e dei comportamenti umani e pertanto la robotica si sta orientando principalmente verso l'utilizzo di robot sociali persuasivi. Un robot sociale persuasivo è un agente fisico capace di interagire socialmente con gli esseri umani e di influenzare o modificare in modo significativo il loro comportamento, i loro atteggiamenti o i loro processi cognitivi (Siegel Breazeal Norton, 2009). Studi recenti hanno evidenziato come i robot possano agire come agenti persuasivi, incoraggiando in modo efficace le persone ad adottare comportamenti sostenibili (Beheshtian Moradi Ahtinen, 2020). In particolare, i robot sociali possono incrementare la consapevolezza degli utenti riguardo alle risorse ambientali. Ad esempio, possono contribuire alla riduzione del consumo energetico domestico, fornendo *feedback* diretti come congratulazioni o rimproveri agli utenti durante l'impostazione di una lavatrice (Ham Midden, 2014). Questo tipo di interazione dimostra come i robot possano facilitare cambiamenti significativi nei comportamenti quotidiani, favorendo una gestione più sostenibile delle risorse.

In ambito educativo, i robot trovano ulteriori applicazioni interessanti. Progetti innovativi come i sistemi di giardinaggio robotico intelligente combinano tecnologia robotica e orticoltura per insegnare pratiche sostenibili e promuovere la riduzione dei rifiuti elettronici (Araiza Morris Integlia, 2019). Inoltre, i robot sociali possono contribuire a migliorare la differenziazione dei rifiuti da parte dei bambini. Un esempio notevole è il robot Pepper, che, giocando in una competizione con i bambini, li incoraggia a essere più consapevoli riguardo al riciclo dei materiali. Pepper, dotato di un'interfaccia amichevole e interattiva, riesce a rendere il processo di apprendimento divertente ed efficace. Inoltre, Pepper viene utilizzato anche per incoraggiare comportamenti sostenibili all'interno di spazi condivisi, promuovendo pratiche come la gestione efficiente delle risorse e la riduzione degli sprechi (Castellano De Carolis D'Errico, 2021; Behestian Moradi Ahtinen, 2020).

Nel medesimo contesto, vale la pena menzionare anche una ricerca condotta da Shih-Yu Lo e colleghi (2022), i quali hanno esplorato l'efficacia dei robot nell'incoraggiare il riciclo. L'esperimento prevede il confronto tra l'uso di un robot e di un tablet per fornire istruzioni sul corretto smaltimento dei rifiuti. I risultati hanno

mostrato che i partecipanti che si sono confrontati con il robot hanno smistato i rifiuti in modo più accurato rispetto a quelli che hanno fatto uso del tablet. Questo effetto è attribuito alle caratteristiche antropomorfe del robot, che evocano maggiore empatia rispetto a un tablet, rendendo i robot strumenti più efficaci per promuovere comportamenti sostenibili (Lo Lai Liu, 2022; Indurkhya Sienkiewicz, 2024).

Tali approcci evidenziano il crescente potenziale della robotica nel favorire comportamenti sostenibili e nel promuovere una cultura della sostenibilità attraverso educazione e interazione sociale. Grazie ai continui progressi tecnologici e alla maggiore consapevolezza dell'importanza della sostenibilità, le applicazioni della robotica sociale si stanno espandendo rapidamente. Tali robot, infatti, non solo educano e sensibilizzano il pubblico, ma offrono soluzioni pratiche per affrontare le sfide ambientali, facilitando l'adozione di pratiche sostenibili. La loro capacità di influenzare comportamenti attraverso l'interazione sociale li rende strumenti efficaci per guidare la transizione verso un futuro più sostenibile.

D'altra parte, non può sfuggire il fatto che i robot in sé non siano dispositivi *eco-friendly*, in quanto alcuni materiali con cui vengono costruiti possono non essere riciclabili o avere impronte ecologiche pesanti (Joshi, 2018). Inoltre, la crescente complessità dei robot comporta il problema dello smaltimento dei rifiuti elettronici al termine del loro ciclo di vita, rendendo necessari metodi di smaltimento e riciclo adeguati per mitigare il degrado ambientale. Lo sviluppo robotico richiede un'alta intensità di risorse, tra cui materie prime ed energia, e comporta la necessità di un approvvigionamento sostenibile e di pratiche di produzione efficienti dal punto di vista energetico. Inoltre, la manutenzione e la riparabilità dei robot pongono sfide che, se non affrontate, possono portare a un aumento dei rifiuti e a uno smaltimento prematuro. È quindi essenziale sviluppare robot che siano facilmente riparabili e che possano essere mantenuti per prolungare il loro ciclo di vita e ridurre al minimo l'impatto ambientale.

Si muove in questa direzione la cosiddetta robotica *soft*, che si ispira alla natura e punta a creare robot leggeri, flessibili e biodegradabili, in grado di collaborare in modo sicuro con gli esseri umani e di interagire con l'ambiente in maniera ecologica. I robot *soft*, infatti, sono costruiti con polimeri biodegradabili, come l'Ecoflex, e materiali derivati da fonti rinnovabili, come l'alginate e il glicerolo. La produzione richiede inoltre

tecniche di fabbricazione innovative come la stampa a iniezione[22] e la stampa 3D. È necessaria anche una certa attenzione al design del robot, che deve prevedere facilità nello smontaggio e nel riciclo alla fine del ciclo di vita del prodotto. Essi possono utilizzare fonti di energia rinnovabile, come quella solare, per funzionare in modo autonomo e sostenibile. I robot *soft* trovano impiego soprattutto nel campo biomedico, ad esempio con attuatori pneumatici biodegradabili utilizzati per la somministrazione di farmaci e supporto nella guarigione delle ferite. Si sta infine esplorando anche l'uso di materiali "viventi", come muscoli ingegnerizzati e il micelio (ovvero l'apparato vegetativo) dei funghi, per costruire robot che possono crescere e autoripararsi. Tali materiali offrono nuove capacità per la robotica, inclusa la capacità di adattarsi ed evolversi in base all'ambiente circostante (Hartmann Baumgartner Kaltenbrunner, 2021).

1.5 L'insostenibilità dell'IA

Negli ultimi due paragrafi sono state analizzate le applicazioni dell'IA e della robotica per la sostenibilità. Tuttavia, riguardo ai robot è stato evidenziato come il loro utilizzo possa presentare anche aspetti non sostenibili, una criticità che coinvolge anche l'IA e che riconduce alla necessità di IA che siano in sé sostenibili.

Infatti, come sottolinea Kate Crawford (2021), il termine "intelligenza artificiale" evoca l'idea di algoritmi – parola la cui definizione esatta sembra non essere poi così chiara ai più – misteriosi processi informatici e vaghi *software* sparsi nel *cloud*. Tuttavia, è fondamentale riconoscere che l'efficacia di molte soluzioni tecnologiche per affrontare le emergenze globali dipende da un elemento fisico: il litio, risorsa cruciale per la produzione di batterie e presente in grandi quantità in paesi come Cile, Cina e Australia. I processi di estrazione del litio sollevano numerose problematiche ambientali e sociali, diventando spesso oggetto di accese proteste, in particolare negli Stati Uniti, dove gli ambientalisti denunciano i gravi impatti ecologici associati a tali attività: "the mining that makes AI is both literal and metaphorical. The new extractivism of data mining also encompasses and propels the old extractivism of data mining" (Crawford, 2021, 31).

[22] La stampa a iniezione è un processo di produzione utilizzato per produrre parti in plastica o altri materiali attraverso l'iniezione del materiale fuso in uno stampo. Questo processo è ampiamente utilizzato nell'industria manifatturiera per la produzione di componenti di varie forme e dimensioni con alta precisione e ripetibilità.

L'attività mineraria porta con sé una storia di sfruttamento di terre e popoli, e rappresenta solo la prima delle caratteristiche non sostenibili dell'IA. Infatti, dopo che il litio è stato estratto e raffinato, viene utilizzato per costituire i componenti di tutti i processi computazionali dei modelli di IA. Come è noto, essi richiedono grandi quantità di elettricità, che è in gran parte fornita da fonti energetiche basate su combustibili fossili. Di conseguenza, ciò comporta emissioni di gas serra che aggravano il cambiamento climatico. L'IA mette anche sotto pressione le reti energetiche, il che può portare a carenze di energia, con particolare impatto sulle regioni che hanno scarse risorse. Inoltre, i requisiti energetici dei sistemi di IA richiedono la costruzione di nuove centrali elettriche, che a loro volta contribuiscono al degrado ambientale (Crawford, 2021).

A tale proposito, Luciano Floridi e colleghi (2022) riportano l'esempio di GPT-3 – ormai già datato rispetto alle versioni attuali di GPT-4 e GPT-4o – notissimo modello di generazione automatica del testo, targato *OpenAI*. Secondo la documentazione rilasciata da *OpenAI* nel maggio 2020, GPT-3 richiedeva una potenza di calcolo di diversi ordini di grandezza superiore rispetto al predecessore GPT-2, rilasciato solo un anno prima. Per provare a stimare il costo ecologico di una singola sessione di addestramento, devono essere monitorati diversi fattori: il sistema di *hardware* utilizzato, la durata di una sessione, il numero di reti addestrati, l'ora del giorno, l'impiego di memoria e le risorse adoperate dalla rete energetica che fornisce l'elettricità (Henderson Hu Romoff, 2020). Tali dati sono solo alcuni di quelli necessari, e l'assenza di parte di essi può naturalmente distorcere le valutazioni dell'impronta ecologica. Nonostante varie limitazioni, come la mancanza di informazioni sul numero di modelli addestrati per ottenere risultati pubblicabili, Floridi e colleghi (2022) sono riusciti a stimare che una singola sessione di addestramento di GPT-3 avrebbe prodotto 223.920 chilogrammi di anidride carbonica. Tale stima è stata realizzata sulla base delle informazioni relative alla quantità di potenza di calcolo e al tipo di *hardware* utilizzato dai ricercatori di *OpenAI* per addestrare il modello (Brown Mann Ryder, 2020), facendo ipotesi sul resto delle condizioni di addestramento del modello (Cowls Tsamados Taddeo, 2021) e usando un calcolatore dell'impatto di carbonio (Lacoste Luccioni Schmidt, 2019), senza considerare le tecniche di contabilizzazione e compensazione dei fornitori per ottenere emissioni "pari a zero". Per fare un confronto, un'automobile negli Stati Uniti emette circa 4600 chilogrammi di anidride carbonica l'anno: dunque, una singola sessione di GPT-3 emetterebbe circa

quanto 49 automobili in un anno. Inoltre, bisogna precisare che anche la geografia ha un peso: infatti, costa dieci volte di più in termini di anidride carbonica equivalente[23] addestrare un modello utilizzando le reti energetiche in Sudafrica piuttosto che in Francia (Floridi, 2022).

Come anticipato precedentemente, le criticità dell'IA dal punto di vista ambientale sono quelle più evidenti, ma non dovrebbero essere sottovalutati nemmeno i problemi legati all'economia e alla società. Vinuesa e colleghi (2020), nella loro analisi dell'impatto dell'IA sugli OSS, concludono che le tecnologie attuali, se utilizzate su larga scala, possono contribuire al raggiungimento del 90% degli obiettivi ambientali, del 70% degli obiettivi economici e dell'82% degli obiettivi sociali (Vinuesa Hossein Leite, 2020). Tuttavia, queste affermazioni devono essere esaminate attentamente. Infatti, ad esempio, bisogna considerare il rischio di sviluppare un tipo di IA che miri a un OSS a discapito di un altro: "if an AI application, e.g., for analysing strategies for protecting biodiversity, recommends the abandonment of long-inhabited, culturally relevant settlements (and thus calls for displacement of some peoples), the net effect on sustainability on earth may be zero or negative" (Heilinger Kempt Nagel, 2023, 3). Ovviamente, non ogni tentativo di raggiungere un OSS implica la violazione di un altro, ma quando si implementano tecnologie di IA mirate a singoli OSS è necessario prendere in considerazione i potenziali conflitti di interessi tra i diversi obiettivi (Sætra, 2021).

Un altro aspetto da non trascurare, come sottolineato da Crawford (2021) e Mark Coeckelbergh (2021), è che l'utilizzo dell'IA per la sostenibilità potrebbe anche condurre a una concentrazione di potere, e al cosiddetto "Leviatano verde": infatti, imporre a tutte le attività umane l'obbligo dell'adattamento o della mitigazione climatica potrebbe portare alla violazione di libertà legittime. Il rischio di utilizzare l'IA per la sostenibilità come strumento per accrescere la concentrazione di potere in nome dell'azione climatica diventa ancora più preoccupante se si considera che l'IA stessa potrebbe non essere così efficace come spesso si sostiene, soprattutto se genera in governi e comunità la falsa percezione che l'impegno intrapreso sia già sufficiente (Heilinger Kempt Nagel, 2023), come addotto dal governatore Newsom.

[23] La CO_2 eq è una misura che rappresenta tutti i gas a effetto serra convertendoli nella quantità equivalente di CO_2.

È poi particolarmente allarmante l'idea che sia possibile sviluppare applicazioni tecnologiche capaci di ottimizzare o compensare completamente i processi e le abitudini non sostenibili dell'umanità. Questa visione, nota come tecno-entusiasmo, può degenerare in un approccio di tecno-soluzionismo, particolarmente pericoloso quando si tratta di affrontare problemi di natura sociale (Morozov, 2013). Tentare di rendere la società più sostenibile attraverso la tecnologia implica spesso la presunzione che le ingiustizie sociali possano essere risolte semplicemente grazie all'automazione. Tuttavia, questa prospettiva riduce drasticamente la complessità delle dinamiche che alimentano tali ingiustizie, le quali richiedono invece interventi integrati di natura politica e sociale, spesso indipendenti dall'apporto tecnologico (Heilinger Kempt Nagel, 2023). Un esempio concreto di sostenibilità sociale compromessa riguarda i modelli di business adottati da alcune aziende di IA, caratterizzati dallo sfruttamento del lavoro e da pratiche di raccolta dati discutibili (Chan Okolo Terner, 2021). Come spesso si afferma, i dati sono il nuovo petrolio: ciò implica che il controllo esercitato da queste aziende sui dati personali delle persone contribuisce a creare dinamiche sociali sconvenienti. Questo crescente potere basato sull'accumulazione di informazioni personali alimenta ulteriori disuguaglianze, mettendo in discussione la sostenibilità etica e sociale della tecnologia stessa.

Considerando tutto ciò, il punto centrale della discussione su IA e sostenibilità è che dovrebbero essere considerate tutte le diverse sfaccettature dell'argomento, che spesso sono trascurate dalla letteratura: la sostenibilità non riguarda solo l'ambiente, e la tecnologia non è sempre la panacea per tutti i mali. Su questa base, Jan-Christoph Heilinger e colleghi (2023) hanno proposto una valutazione di qualsiasi tecnologia basata sull'IA che tenga conto degli aspetti legati alle dimensioni dell'ambiente e della società. La valutazione di Heilinger e colleghi riguarda due dimensioni diverse e complementari: il primo aspetto è costituito da una categorizzazione in base ai fini e agli esiti dell'IA, che cioè considera gli effetti reali dell'utilizzo dell'IA per promuovere la sostenibilità ambientale o sociale. Il secondo aspetto è una valutazione in base ai mezzi, ovvero considerando la sostenibilità ambientale e sociale dell'IA come strumento in sé. Le conclusioni di Heilinger e colleghi sono rilevanti nel dibattito poiché propongono una distinzione terminologica tra "thin and thick accounts of sustainability" (Heilinger Kempt Nagel, 2023, 9): solo se sono rispettate tutte le dimensioni menzionate – che coinvolgono

tanto l'aspetto sociale quanto quello ambientale della sostenibilità, considerandoli sia dal punto di vista dei mezzi sia da quello del fine – un'IA può essere definita sostenibile "in modo spesso" ("thick sustainable"). In tutti gli altri casi, può essere al massimo giustificata come sostenibile "in modo sottile" ("thin sustainable").

Definire sostenibile qualcosa che in realtà non lo è, è ovviamente una distorsione della realtà, che porta a un'operazione di *ethics-washing*. Infatti, catalogare una tecnologia come "sostenibile" è attraente non solamente per il suo effettivo contributo alla sostenibilità sociale e ambientale, ma anche per ragioni esterne, legate ad aspetti economici e politici. Aziende, investitori e sviluppatori hanno un vasto interesse a essere percepiti come sostenibili sia rispetto ai prodotti sia rispetto alle strategie adottate. I problemi più cruciali riguardano il fatto che la crescita economica è, a spese della crescita in termini di efficienza, direttamente legata a un aumento dell'uso delle risorse, e dunque raramente si può classificare come sostenibile; inoltre, d'altra parte, la crescita dell'implementazione di IA è legata a doppio filo ad alti costi sociali – come già accennato in precedenza – come, ad esempio, lo sfruttamento dei popoli e dei territori. Tali criticità, tuttavia, vengono sistematicamente sottostimate – quando non ignorate del tutto – nella foga di assicurare ad acquirenti e consumatori che il prodotto scelto sia effettivamente *green*. L'*ethics-washing* non riguarda solamente la presunta riduzione dell'impatto ambientale, ma anche l'impatto sociale delle tecnologie, come quando ad esempio le tecnologie basate su IA vengono spacciate come utili per aumentare l'inclusione o l'uguaglianza, laddove in fin dei conti perpetuano – se non addirittura cementano ulteriormente – preesistenti ingiustizie sociali o divari digitali (Eubanks, 2018).

La distinzione tra "thin sustainable" e "thick sustainable" è utile nell'ottica di trovare una valutazione dell'IA che coinvolga quattro aspetti della sostenibilità: i primi due riguardano la targettizzazione rispetto ai fini e ai prodotti, ovvero l'uso effettivo dell'IA per avere avanzamenti ambientali e sociali, e gli altri due la sostenibilità sociale e ambientale dell'IA come strumento in sé, ovvero della tecnologia come mezzo. Preso atto del fatto che molte tecnologie basate su IA sono sostenibili rispetto a uno o più dei quattro aspetti della sostenibilità che abbiamo citato, bisogna comunque tenere a mente che nessuno di essi da solo è condizione sufficiente per un giudizio comprensivo di un'IA come sostenibile. Al contrario, tutti e quattro – nella misura in cui sono applicabili – sono condizioni necessarie per giustificare l'etichettatura complessiva di sostenibile.

Secondo queste premesse, nessuna tecnologia basata sull'IA dovrebbe, al momento, essere qualificata come sostenibile. Tuttavia, una distinzione così forte è necessaria poiché "sustainability is a powerful notion and should be kept as an ambitious goal, not be cheapened to accommodate, e.g., the economic or political interests […] at the expense of environmental and social concerns" (Heilinger Kempt Nagel, 2023, 9). Dal punto di vista sociale, è fondamentale democratizzare le tecnologie digitali, il che comporta la regolamentazione delle infrastrutture di telecomunicazione – come i centri dati, le reti in fibra ottica e le infrastrutture informatiche in generale – con l'obiettivo di evitare il monopolio da parte di entità private e garantire invece che siano trattate come beni pubblici. Ciò è legato alla necessità di connessioni a Internet affidabili e accessibili, che possano essere utilizzate in modo efficace in contesti educativi, sanitari ed economici. Inoltre, la spinta alla democratizzazione digitale deve affrontare anche il requisito dell'efficienza computazionale, un aspetto fondamentale che, al momento, sembra essere stato ampiamente trascurato da molte aziende (Marwala Mbuvha Mungwe, 2023). Garantire che questi sistemi funzionino in modo efficiente non è vitale solamente per il progresso tecnologico, ma anche per la sostenibilità. Inoltre, è imperativo considerare l'uso responsabile delle risorse naturali, come l'acqua e l'energia, che vengono consumate in grandi quantità dalle infrastrutture digitali.

Heilinger e colleghi (2023) a questo proposito riportano un esempio efficace, che si riferisce a un articolo di Emily Bender e colleghi (2021) sui modelli linguistici di grandi dimensioni (*Large Language Models*, LLM). Nell'articolo, viene problematizzato l'alto costo di tali modelli rispetto alla sostenibilità ambientale, date le considerevoli quantità di energia impiegata nell'addestramento e le emissioni di gas serra conseguenti. Dunque, la valutazione degli LLM è negativa rispetto alla dimensione dell'IA ecologica. Inoltre, non è chiaro, dal punto di vista della valutazione dei fini, in che modo gli LLM dovrebbero contribuire direttamente alla sostenibilità, e quindi al più si può dire che essi sono neutrali rispetto alla sostenibilità ambientale. Dal punto di vista sociale, inoltre, l'addestramento degli LLM è problematico a causa alla selezione preventiva che occorre fare per i dati di addestramento, in modo da evitare che il modello adotti un linguaggio sessista, razzista o in generale discriminatorio, e dunque che possa essere usato in modo sicuro in tutti i contesti conversazionali. La questione è piuttosto complessa, in quanto se da un lato gli sviluppatori degli LLM sostengono che addestrarli in tale maniera li renda socialmente

sostenibili, in quanto *politically correct*, dall'altro è stato fatto notare che la selezione preventiva introduce dei *bias* nel linguaggio, che comportano la cancellazione di determinati modi di dire e culture, comportando l'offuscamento di certe identità (Heilinger Kempt Nagel, 2023).

Heilinger e colleghi (2023), in conclusione, osservano come una sostenibilità "thin", in linea di principio, sia naturalmente preferibile a una sostenibilità del tutto assente. La soddisfazione di parametri di sostenibilità che considerino tutte e quattro le dimensioni sopracitate, infatti, è molto complessa da raggiungere. Dopo tutto, non sarebbe comunque un passo avanti accogliere qualsiasi miglioramento, anche in ambiti specifici come l'efficienza energetica, pur se le altre dimensioni della sostenibilità restano inesplorate? Tuttavia, la sostenibilità è una nozione complessa e potente, che non dovrebbe essere ridotta a un concetto parziale o strumentalizzato per favorire interessi economici o politici, spesso a scapito delle problematiche ambientali e sociali. La sfida è mantenere una visione olistica: concentrarsi su un miglioramento in una sola area, trascurandone o addirittura compromettendone altre, significa tradire il senso autentico della sostenibilità, che mira a preservare le risorse per un uso continuo e a garantirne la disponibilità per le generazioni future.

Giungendo a tali considerazioni, in questo capitolo si è cercato di delineare, in termini generali, il rapporto tra l'IA e la sostenibilità, offrendo un quadro di riferimento per i due macrotemi centrali di questa tesi. Come anticipato nell'introduzione, il lavoro adotta un approccio basato sulle scienze cognitive, proponendolo come strumento utile per sviluppare modelli di IA che rispettino i criteri di sostenibilità. I capitoli successivi si svilupperanno dunque in questa direzione, approfondendo il ruolo dei modelli cognitivi per la progettazione di sistemi di IA che siano in sé sostenibili.

CAPITOLO II

MODELLI COGNITIVI PER IA SOSTENIBILE

2.1 IA spiegabile: un approccio cognitivo

Nel capitolo precedente è stata delineata la relazione tra l'IA e la sostenibilità e sono state evidenziate le principali problematiche relative agli aspetti non sostenibili dell'IA. Su tali basi, il Capitolo II punta a esplorare in che modo un'IA sostenibile dal punto di vista sociale sia anche spiegabile, e come i modelli cognitivi abbiano delle potenzialità significative in tale ambito.

Come è stato visto, la prima sfida cruciale posta alla sostenibilità dall'IA riguarda l'impatto ambientale. I sistemi di IA richiedono notevoli risorse energetiche per l'estrazione dei materiali necessari alla produzione dei componenti *hardware*, che spesso comporta un uso intensivo del suolo e lo sfruttamento delle popolazioni locali. A ciò si aggiungono i costi ambientali associati alla manutenzione delle infrastrutture digitali, al consumo di energia necessario per alimentare i centri dati e i dispositivi e, infine, allo smaltimento dei dispositivi stessi, che contribuisce in modo significativo ai rifiuti elettronici. Il secondo aspetto critico riguarda poi l'impatto sociale di alcuni tipi di IA. A livello sistemico, l'addestramento di algoritmi su grandi insiemi di dati che incorporano pregiudizi culturali o sociali rischia di perpetuare e amplificare tali pregiudizi all'interno della società, potenzialmente influenzando le decisioni automatizzate e discriminando i gruppi vulnerabili. Inoltre, a livello tecnico, l'implementazione di modelli di IA non spiegabili che funzionano come scatole nere (*black box*) impedisce sia ai tecnici sia agli utenti di comprendere appieno i processi decisionali degli algoritmi. La mancanza di trasparenza solleva notevoli problemi etici e di sicurezza, soprattutto nella gestione di sistemi che influenzano decisioni critiche.

Dall'urgenza di affrontare tali problematiche scaturisce il filone di ricerca dell'IA spiegabile (*explainable*), il cui obiettivo è sviluppare modelli di apprendimento automatico che siano comprensibili, affidabili e facilmente interpretabili (Guidotti Monreale Ruggieri, 2018). L'IA spiegabile punta a rendere trasparente il processo decisionale delle macchine, consentendo agli utenti di comprendere il ragionamento alla base delle scelte effettuate e di valutare l'accuratezza e l'affidabilità dei risultati (Longo Brcic Cabitza, 2024). I recenti sforzi per regolamentare l'IA, a cui si è fatto riferimento nel Capitolo I, potrebbero rendere la spiegabilità un requisito fondamentale per l'implementazione su larga scala di qualsiasi sistema di IA. Negli Stati Uniti, ad esempio, il *National Institute of Standards and Technology* (NIST) ha pubblicato nel 2023 l'*Artificial Intelligence Risk Management Framework* (RMF), che definisce la spiegabilità e l'interpretabilità come elementi essenziali per un sistema di IA affidabile. L'RMF è stato progettato per fornire alle aziende del settore tecnologico linee guida per la gestione dei rischi legati all'IA e potrebbe diventare uno standard del settore. Analogamente, il senatore Chuck Schumer ha promosso un'iniziativa legislativa per introdurre norme nazionali sull'IA, con un punto chiave: la necessità di fornire spiegazioni su come l'IA giunge alle sue conclusioni (Drake Ong Hansen, 2023). Per quanto riguarda le norme seguite dall'Unione Europea, come trattato precedentemente, nel giugno del 2024 è entrato in vigore l'*AI Act*.

In letteratura, i termini "spiegabile", "interpretabile" e "comprensibile" vengono spesso utilizzati in modo intercambiabile. Tuttavia, questa equivalenza non è del tutto accurata. In particolare, mentre "spiegabile" e "interpretabile" possono essere considerati sinonimi in molti contesti, il termine "comprensibile" presenta una sfumatura distinta che lo differenzia dagli altri due (Charmet Tanuwidjaja Ayoubi, 2022). Con "spiegabilità", si intende "the ability to provide the meaning of the relationships a model's inputs and its outcomes have, in a human-readable form" (Molnar, 2018). Dunque, la spiegabilità è il grado di comprensione da parte degli esseri umani delle decisioni prese da un modello di IA: maggiore è la spiegabilità, più è facile per gli utenti comprendere come mai un modello ha preso una determinata decisione. La *comprensione*, d'altra parte, si riferisce alla capacità di un modello di IA di far capire a un essere umano la propria funzione, senza necessariamente rivelare i dettagli del suo funzionamento interno (Barredo Arrieta Díaz-Rodríguez Del Ser, 2020).

Tuttavia, per essere veramente comprensibile, un modello dovrebbe essere almeno parzialmente trasparente, in modo da fornire informazioni sufficienti per garantire che le sue decisioni siano interpretabili e affidabili. La mancanza di trasparenza può derivare da diversi fattori, tra i quali si annoverano: l'impossibilità cognitiva per gli esseri umani di interpretare modelli algoritmici e insiemi di dati di dimensioni estremamente elevate; l'assenza di strumenti adeguati per la visualizzazione e il monitoraggio di grandi volumi di codice e dati; la presenza di codice e dati strutturati in modo tale da risultare illeggibili; e, infine, i continui aggiornamenti e interventi umani sul modello, che ne complicano ulteriormente la tracciabilità e la comprensione, ovvero la cosiddetta "malleabilità algorimica" (Floridi, 2022, 154). Quest'ultima caratteristica consente agli sviluppatori di monitorare e ottimizzare un algoritmo già implementato. Tuttavia, può anche essere utilizzata per oscurare la storia del processo di sviluppo, lasciando gli utenti finali nell'incertezza riguardo alle modalità con cui si è giunti a specifici output (Ananny Crawford, 2018). La mancanza di trasparenza è una caratteristica tipica degli algoritmi di autoapprendimento, in quanto essi alterano la loro logica decisionale producendo nuovi insiemi di regole durante il processo di apprendimento, e ciò rende difficile per gli sviluppatori mantenere una comprensione dettagliata del motivo per cui sono state apportate alcune modifiche specifiche (Burrell, 2016). Tuttavia, ciò non sempre si traduce in risultati opachi. Infatti, anche senza comprendere ogni passaggio logico, gli sviluppatori possono modificare i parametri che regolano il processo di addestramento per verificare vari output (Floridi, 2022).

Il fatto che l'opacità sia una caratteristica intrinseca di molti algoritmi di ML non significa però che non possano essere attuati dei miglioramenti. Aziende come *Google* e *IBM*, infatti, hanno compiuto sforzi notevoli per rendere gli algoritmi di ML più interpretabili e inclusivi, rendendo disponibili strumenti come *Explainable AI*, *AI Explainability 360* e *What-If Tool*. In cantiere è anche lo sviluppo di

> altre strategie e strumenti, che offrano agli sviluppatori e al pubblico in generale interfacce visive interattive che migliorino la leggibilità umana, esplorino vari risultati del modello, forniscano ragionamenti basati su casi, regole direttamente interpretabili e identifichino e mitighino anche i pregiudizi non voluti negli insiemi di dati e nei modelli algoritmici (Floridi, 2022, 158).

In generale, la comprensione delle macchine – dei computer – è stata oggetto di studio fin dai primordi della loro diffusione. Il tema ha acquisito una maggiore sistematizzazione come campo di ricerca scientifica negli anni ’80, periodo in cui si è cominciato a definire in modo più chiaro la questione dell’interazione uomo-macchina (Héder, 2023). Terry Winograd e Fernando Flores (1986) furono tra i primi a occuparsi in modo sistematico dei problemi legati alla spiegabilità e alla trasparenza delle macchine, mettendo ordine nelle definizioni di termini vaghi come “user-friendly”, “easy-to-learn” e “self-explaining” e inserendoli in una riflessione scientifica approfondita che si basasse sulla fenomenologia e sulle scienze cognitive. La loro idea principale era che un sistema deve riflettere il modo in cui è strutturata la rappresentazione mentale[24] dell’utente riguardo al dominio di utilizzo di quel sistema.

Dunque, l’interfaccia e le funzionalità di un sistema devono essere progettate in modo da corrispondere al modo in cui l’utente percepisce e comprende il contesto o le informazioni con cui interagisce. Sebbene oggi questa possa apparire un’idea scontata, all’epoca rappresentava un approccio innovativo. Winograd e Flores (1986) hanno contribuito significativamente all’evoluzione della progettazione delle interfacce, orientando la ricerca verso lo sviluppo di sistemi che rispecchiassero le rappresentazioni mentali degli utenti, promuovendo così un’interazione più intuitiva ed efficace. Contemporaneamente, emergeva l’approccio alla programmazione orientata agli oggetti, che facilitò la comprensione e la gestione del codice da parte degli sviluppatori. L’implementazione di processi automatici che fossero comprensibili, infatti, fu affidata a “oggetti”, ovvero strutture dati organizzate secondo regole logiche ben definite. L’IA dell’epoca, dunque, era progettata per compiere inferenze logiche e ragionamenti basati su dati organizzati e rappresentazioni simboliche – come liste, grafici, matrici e tabelle. Tali vincoli portavano a sistemi che gli sviluppatori e gli utenti potevano facilmente

[24] Come sottolinea Dan Ryder (2009) “there are many different types of things discussed in the psychological and philosophical literature that are candidates for representation-hood”. In questa sede, ci limitiamo a riportare che il termine “rappresentazione mentale” è innanzitutto un costrutto teoretico delle scienze cognitive, e, in termini minimi è “something that possesses semantic properties: a truth value, a satisfaction value (i.e. satisfied or not), truth conditions, satisfaction conditions, reference, or content” (Ryder, 2009, 234). Si tratta di uno dei concetti base della teoria della mente computazionale secondo cui gli stati e i processi cognitivi sono costituiti dalla comparsa, dalla trasformazione e dall’immagazzinamento (nella mente/cervello) di strutture portatrici di informazioni (rappresentazioni) di un tipo o di un altro (per un approfondimento si rimanda a https://plato.stanford.edu/entries/mental-representation, consultato in data 23/10/2024).

supervisionare, dal momento che ogni passaggio logico dei processi di inferenza e computazione era auto esplicativo (Héder, 2023).

Con l'avvento del ML, la progettazione dei sistemi di IA è profondamente mutata. I metodi di apprendimento automatico più moderni, infatti, si basano sull'utilizzo di grandi quantità di dati e modelli statistici complessi che apprendono automaticamente comportamenti e conoscenze, senza una struttura predefinita, come ad esempio le reti neurali. Precedentemente è stato menzionato il termine *black box*, che venne usato per la prima volta da Frank Rosenblatt nel 1957 in riferimento a un singolo neurone di una rete neurale artificiale, ovvero il percettrone, indicando l'inaccessibilità di alcuni processi interni (Rosenblatt, 1957). Tuttavia, a causa della natura dei computer, è piuttosto evidente come ogni dettaglio e ogni componente di tali sistemi dovrebbe poter essere esaminato facilmente. Sarebbe dunque più corretto parlare di mancanza di comprensione e spiegabilità per i comportamenti che siano curiosi o imprevisti. La comprensione è un valore epistemico, e dunque è utile adottare il termine "opacità epistemica" (Héder, 2020), il cui opposto è la trasparenza epistemica, che è una caratteristica dei sistemi che permette la comprensione umana e la supervisione intellettuale.

I ricercatori stanno esplorando vari approcci per l'implementazione di IA spiegabile, ma non si è ancora giunti a una metodologia univoca e rigorosa per definire esattamente la spiegabilità e la correttezza (*fairness*) di un sistema (Tsamados Aggarwal Cowls, 2022; Barredo Arrieta Díaz-Rodríguez Del Ser, 2020; Doshi-Velez Kim, 2017). Principalmente, sono state individuate due strategie per rendere i sistemi di ML trasparenti, ovvero comprensibili: la prima è l'interpretazione del modello intero, e in questo caso la spiegazione che ne risulta viene chiamata "globale". La spiegazione globale può essere raggiunta attraverso un modello surrogato, che è di aiuto nel rappresentare fedelmente il modello originale e al contempo ne permette la semplificazione attraverso l'uso di elementi che gli esseri umani possono comprendere facilmente. Se i modelli surrogati sono implementati con accuratezza, l'intero modello è reso trasparente ed è possibile predire il comportamento causato da determinati input prima che il processo di computazione sia completato, permettendo il controllo intellettuale dei sistemi. Altre spiegazioni globali implicano la visualizzazione del modello o la mappatura dei concetti principali utilizzati dallo stesso. All'approccio globale si oppone il cosiddetto approccio "locale". Quest'ultimo si riferisce al concetto di

"local fidelity" (Héder, 2023, 10), ovvero, una spiegazione per un input particolare del sistema potrebbe equivalere a una spiegazione per altri input simili, dove la similarità è misurata come distanza in uno spazio matematico (Holzinger Saranti Molnar, 2022).

In generale, l'approccio epistemico alla trasparenza è naturalmente legato alle conoscenze delle singole persone che provano a raggiungere la comprensione intellettuale dei sistemi con cui si rapportano. Esso è un paradigma chiave del sopracitato *AI Act* dell'UE, che utilizza un approccio basato sugli *stakeholder,* ovvero su tutte le persone coinvolte nell'implementazione e nell'utilizzo di un determinato sistema. Queste vengono suddivise in: pubblico generale, utenti (coloro che utilizzano direttamente il sistema), spettatori (coloro che non usano il sistema direttamente ma possono essere influenzati da esso) ed esperti (coloro che possiedono competenze specifiche o svolgono un ruolo tecnico nel sistema). A loro volta, gli esperti si caratterizzano in agenti di certificazioni e revisori, investigatori di incidenti e consulenti esperti in controversie legali. L'obiettivo di un approccio di questo tipo è garantire che ogni gruppo venga preso in considerazione e rispettato nelle decisioni prese rispetto alle implementazioni che si devono portare a termine.

Il bisogno di interpretabilità e verificabilità umana risiede dunque nella necessità di sciogliere i processi automatici in ragionamento esplicito: la comprensione delle performance e dei potenziali *bias* dei sistemi di IA è cruciale per una loro allocazione che sia etica e responsabile. Tale comprensione deve essere estesa oltre la valutazione dei sistemi automatici sulla base dei parametri di riferimento tracciati dall'accademia e per gli specifici compiti che sono interessanti solo in circoscritti campi della ricerca, andando verso una comprensione globale di cosa i modelli rappresentano e imparano, così come gli algoritmi su cui sono basati (Guest Martin, 2023). Le considerazioni legate alla trasparenza devono essere presenti durante l'intero ciclo di vita di un sistema di IA, cominciando con l'identificazione dei problemi sociali che potenzialmente potrebbero emergere e per i quali diventa necessario sviluppare una soluzione, a partire dalla fase di raccolta dei dati fino al punto in cui i sistemi di IA vengono diffusi e migliorati progressivamente. La trasparenza funge da apripista per lo sviluppo di altre dimensioni dell'IA etica, come l'interpretabilità, la responsabilità e la sicurezza (Hipólito, 2023).

Spiegabile, come visto, è ciò che è interpretabile dall'essere umano. La trasparenza e la comprensibilità sono considerati anche principi biologici e il tentativo di avere sistemi

artificiali interpretabili si può tradurre nella tendenza a produrre sistemi che imitino la biologia umana (Lawrence El Shazly Seal, 2024). Più precisamente, l'uso delle macchine per predire e spiegare il comportamento dei sistemi biologici risale alla metà del XIX secolo, con la teorizzazione della cibernetica di Norbert Wiener (1948).

La cibernetica nacque come tentativo di promuovere una visione meccanicistica degli esseri umani, in contrasto con la visione vitalistica di Henri Bergson (1911) e l'uso del principio della forza vitale per spiegare l'evoluzione e l'adattamento. Wiener applicò alla biologia il concetto tecnico di *feedback*, in un'operazione fondamentale per la modellazione formale del metabolismo cellulare e della stabilità biologica. Inoltre, introdusse per la prima volta l'idea che la cognizione[25] può essere concettualizzata come un'attività biologica di auto-regolazione (Damiano Cañamero, 2012). Tale ritorna nell'"embodied cognitive science" (Clark, 1999), che a partire dagli anni '80 si pose in contrapposizione alla visione tradizionale delle scienze cognitive ovvero, quella computazionalista, per cui i processi mentali sono processi computazionali e il cervello è equiparato a un computer (Nunez Freeman, 1999). L'approccio *embodied* alle scienze cognitive, infatti, enfatizza la rilevanza del corpo fisico di un agente per gli atti cognitivi. L'interazione del corpo con l'ambiente, in particolare, gioca un ruolo chiave nella cognizione[26]. Parallelamente al lavoro di Wiener, John von Neumann (1958) trasferiva lo schema del computer digitale dalla produzione ingegneristica alla descrizione della cognizione biologica. Il lavoro interconnesso di Wiener e von Neumann diede origine alla tesi per cui la cognizione biologica può essere caratterizzata in termini di processo di informazioni e guidò l'orientamento delle scienze cognitive moderne (Damiano Cañamero, 2012).

Negli stessi anni, lo sviluppo teorico fu accompagnato dalla costruzione di macchine cibernetiche, come le tartarughe di William Grey Walter[27], l'omeostato di

[25] In questa sede, per semplicità, adottiamo l'ampia definizione di cognizione tratta dalla voce "cognition" dell'Oxford Dictionary, ovvero "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses" (https://web.archive.org/web/20160129060026/http://www.oxforddictionaries.com/definition/english/cognition, consultato in data 26/10/2024).

[26] Per un approfondimento esteso e completo dell'approccio *embodied* nell'ambito delle scienze cognitive, si rimanda alla voce "embodied cognition" della Stanford Encyclopedia of Philosophy: https://plato.stanford. edu/entries/embodied-cognition/, consultato in data 26/10/2024.

[27] W. G. Walter (1910-1977) fu uno dei pionieri dell'applicazione dell'elettroencefalografia alla clinica neurologica e psichiatrica. Costruì le cosiddette "tartarughe" – per via della loro forma e della lentezza nei movimenti – per illustrare il funzionamento di alcuni meccanismi cerebrali. Si trattava di macchine autonome semoventi a tre ruote, capaci di muoversi verso una fonte di luce, in grado di trovare la via per una stazione di ricarica delle proprie batterie (Jones, 2016).

William R. Ashby[28] e la macchina *Musicolour* di Gordon Pask[29] (Pickering, 2010). Questi oggetti vennero concepiti dai loro creatori come modelli scientifici, che fungessero da strumento di esplorazione dei processi biologici. L'idea era creare "modelli sintetici" dei processi viventi e cognitivi, ovvero sistemi artificiali in grado di rappresentare, a livello operativo, ipotesi scientifiche sui meccanismi alla base dei comportamenti degli esseri viventi.

È in questi stessi anni che i filosofi cominciarono ad avvertire l'esigenza di "naturalizzare" la fenomenologia, ovvero di renderla uno strumento utile per lo studio delle scienze naturali, come la biologia e le scienze cognitive (Gallagher Zahavi, 2008). A partire dalla lezione di Maurice Merleau-Ponty (1962), scienziati e filosofi, come visto, si contrapposero al concetto di cognizione disincarnata. Ciò andava di passo con i progressi in ambito neuroscientifico: gli avanzamenti nel campo delle tecnologie di *neuroimaging* permisero infatti di cominciare a osservare direttamente i processi neurali in corso, e resero possibili esperimenti basati sui resoconti delle esperienze dei soggetti.

Poiché si trattava di osservazione dell'esperienza percettiva soggettiva, necessariamente entrava in gioco la fenomenologia (Gallagher Zahavi, 2008). In particolare, la necessità di rivalutare la fenomenologia come strumento per studiare le scienze cognitive venne avvertita principalmente da Francisco Varela e Humberto Maturana (1974), i quali furono i padri della cosiddetta neurofenomenologia, ovvero, un quadro teorico per spiegare il comportamento degli organismi biologici, all'interno del quale l'atto cognitivo non è più un processo formato da rappresentazioni e limitato a una funzione epistemica, ma diventa attività e movimento che influenzano sia l'ambiente interno di un organismo sia quello esterno. Secondo Varela e Maturana, la caratteristica principale di un organismo vivente è la cosiddetta autopoiesi, ovvero, la tendenza a stabilire un equilibrio omeostatico tra il tutto e le parti dell'organismo stesso. Così, lo

[28] W. R. Ashby (1903-1972), neurologo e pioniere nello studio dell'organizzazione e del controllo dei sistemi complessi. Ideò l'omeostato per imitare le proprietà di autoregolazione degli organismi viventi: esso consiste in un sistema di quattro dispositivi elettromeccanici (il cui stato può essere variato tramite opportuni commutatori, che simulano l'ambiente esterno) collegati elettricamente tra loro in modo che lo stato di ciascuno di essi dipenda dagli stati dagli altri tre (Ashby, 1960).

[29] G. Pask (1928-1966), ingegnere e inventore, diede molteplici contributi alla cibernetica, alla psicologia, alla tecnologia educativa, all'epistemologia applicata, al calcolo chimico, all'architettura e all'arte dei sistemi. Tra le sue invenzioni ci furono le cosiddette macchine reattive, tra cui la macchina *Musicolour*, che era uno strumento in grado di rispondere in tempo reale a un input sonoro con uno spettacolo di luci diverso a ogni esecuzione (Werner, 2019; Tanni, 2023).

scopo principale di un organismo vivente è mantenersi in vita grazie alle sole componenti che produce internamente.

In questo contesto, l'atto cognitivo è il "corporeal know-how with which any with which any organism is endowed" (Cappuccio, 2009, 22), ovvero, la competenza che si acquisisce agendo e interagendo in e con l'ambiente, in modo da cambiare continuamente la relazione con esso. Come accennato sopra, dunque, l'atto cognitivo non è una rappresentazione, laddove con "rappresentazione" si intende una sorta di riflesso interno del mondo basato sull'adattamento delle informazioni provenienti dal mondo esterno nel momento in cui entrano nell'organismo. Piuttosto, l'atto cognitivo è un processo di costruzione attiva, che coinvolge da un lato l'organismo e dall'altro l'ambiente, allo stesso momento, in quanto uno presente all'altro. Dunque, nel modello autopoietico la cognizione non è più un processo epistemico basato su rappresentazioni interne, ma diventa azione e movimento che riorganizza la struttura interna di un essere vivente, modificando al contempo l'ambiente esterno in cui esso si trova a vivere e ad agire. L'organismo e l'ambiente sono interconnessi così profondamente, e ciascuno di essi dà vicendevolmente forma all'altro, che è impossibile scinderli, o considerarli realtà distinte. In questo senso, Varela e Maturana sfidano la tradizionale separazione tra interno ed esterno nello studio della mente, e cercano di superare la visione dualistica dei processi cognitivi, che fino ad allora erano considerati solamente come atto del soggetto percipiente su un oggetto percepito esterno e passivo (Cappuccio, 2009).

L'aspirazione cibernetica a modellare la vita e i processi cognitivi in modo sintetico e le basi epistemiche della biologica autopoietica di Varela e Maturana vennero ereditate dal filone di ricerca scientifica cosiddetto *Artificial Life* (AL), che emerse alla fine degli anni '80 con lo scopo di produrre modelli artificiali (*software*, *hardware* e *wetware*) di processi vitali e cognitivi al fine di esplorare sperimentalmente aspetti della vita e della cognizione non facilmente accessibili dai sistemi e scenari di ricerca biologici (Cordeschi, 2008). In questo filone si inserì il lavoro pionieristico di Allen Newell (1980), a cui si deve il concetto di "architettura cognitiva" (*cognitive architecture*, CA), emerso proprio dall'intento di creare una base per studiare i processi cognitivi della mente umana con l'aiuto dell'IA. Con tale termine ci si riferisce sia ai modelli cognitivi astratti in agenti naturali e artificiali, sia alle instanziazioni di tali modelli all'interno di *software* per applicazioni in IA (Lieto Bhatt Oltramari, 2018). La categorizzazione di CA venne

introdotta per tre ragioni principali: dal punto di vista della prospettiva cognitivista (Vernon, 2014), per catturare a livello computazionale i meccanismi della cognizione umana, compresi quelli sottostanti alle funzioni di ragionamento, controllo, apprendimento, memoria, adattività, percezione e azione (Oltramari Lebiere, 2012). Dal punto di vista cosiddetto emergentista: per formare la base dello sviluppo delle capacità cognitive attraverso ontogenesi in un periodo esteso di tempo. E infine per raggiungere il livello di intelligenza umana, ovvero per catturare l'intero insieme delle abilità mentali umane unificandole in un'unica architettura. A tale obiettivo ci si riferisce spesso nei termini di ricerca dell'Intelligenza Artificiale Generale (AGI) (Langley, 2006; Goertzel, 2014). Tuttavia, storicamente, le ricerche in IA si sono concentrate maggiormente sull'implementazione di meccanismi più semplici, che mirano a replicare aspetti discreti della cognizione umana, focalizzandosi su domini particolari come la visione o l'elaborazione del linguaggio (Acciai Angius Perconti, in pubblicazione).

Insieme a Herbert Simon (1996), Newell pose le basi per il metodo simulativo, o sintetico, in IA, che consiste nell'usare sistemi computazionali per testare ipotesi cognitive su come funzionino gli organismi viventi (Winsberg, 2010; Durán, 2018). A partire dalla fine degli anni '90, tale metodologia è stata adottata e utilizzata in modo sempre più diffuso per analizzare fenomeni complessi – dal metabolismo biologico alla riproduzione, dalla mappatura concettuale dell'ambiente al ragionamento logico, dal linguaggio allo studio delle emozioni (Damiano Cañamero, 2012). Lo slogan associato al metodo sintetico è "understanding by building" (Pfeifer Scheier, 2000), che ben ne racchiude il senso, e in quegli anni venne applicato alle scienze cognitive, comportamentali e biologiche.

Nell'ambito delle scienze cognitive, si tratta di rappresentare i sistemi naturali attraverso un modello matematico, che di solito è un insieme di equazioni differenziali, implementando un modello computazionale, che può essere una simulazione virtuale da eseguire in modo da ottenere predizioni su comportamenti del sistema target che si è scelto di esaminare. La caratteristica principale di tali simulazioni è che le predizioni per essere affidabili devono essere ottenute attraverso l'imitazione dell'evoluzione del sistema *target* (Angius Perconti Plebe, 2024), e in generale i modelli devono allinearsi con le condizioni di incertezza tipiche del mondo reale (Gigerenzer, 2020).

Su tale scia, la biologia sintetica iniziò a diffondersi all'inizio degli anni 2000 nell'intersezione tra biologia e ingegneria (Chiarabelli Stano Luisi, 2009), principalmente per progettare e costruire parti o sistemi biologici non esistenti in natura, con lo scopo di raggiungere obiettivi pratici, come la biosintesi di prodotti chimici fini, biocarburanti o prodotti farmaceutici. Tuttavia, a poco a poco, la biologia sintetica superò gli scopi meramente applicativi e contribuì in modo nuovo ad affrontare l'esplorazione scientifica della vita (Damiano Stano, 2023). Il metodo sintetico assunse il ruolo di indagine epistemologica, sancendo l'inversione della tendenza fino allora più diffusa nella ricerca scientifica, ovvero l'osservazione di un comportamento e poi la costruzione di un modello corrispondente: il metodo sintetico, infatti, prevede al contrario che i ricercatori in un primo momento formulino ipotesi sui meccanismi vitali e cognitivi all'interno di sistemi artificiali, e successivamente esaminino i comportamenti che producono. Tale approccio è basato sulla distinzione teoretica tra organizzazione e realizzazione fisico-chimica dei sistemi viventi e cognitivi e sulla relativa ipotesi per la quale tali sistemi e la loro fenomenologia possono essere riprodotti realizzando la loro medesima organizzazione all'interno di media fisici: per questo, se applicato alle scienze cognitive tale metodo viene detto anche "di simulazione" – "simulative" (Angius Perconti Plebe, 2024).

Il metodo sintetico, dunque, punta a produrre un tipo di conoscenza scientifica che fornisca spiegazioni operative dei fenomeni esplorati, ovvero, spiegazioni orientate non a indagare fenomeni, bensì a definire i meccanismi che li generano. Si tratta di spiegazioni generali, che si focalizzano non solamente su sistemi e comportamenti reali ma anche possibili.

Nel metodo sintetico si ritrovano le medesime basi epistemologiche della neurofenomenologia di Varela e Maturana, dal momento che i comportamenti cognitivi non sono considerati dipendenti esclusivamente dagli organismi che li adottano, ma da una serie di relazioni tra l'organismo, le sue componenti e il sistema in cui essi operano e con il quale interagiscono (Damiano Cañamero, 2012). I modelli dei sistemi viventi e cognitivi che vengono prodotti sono semplici e generativi, ovvero, sono modelli in grado di generare comportamenti complessi e inaspettati, e non solamente riprodurre meccanismi lineari. Si tratta, infatti, di modelli progettati per essere parte di dinamiche interattive tra sistemi, componenti e ambiente. Così, infatti, essi possono essere veri generatori di conoscenza, dal momento che possono esibire comportamenti imprevisti

rispetto a quelli che gli sviluppatori si aspettano e produrre nuove intuizioni sulle ipotesi per cui sono stati concepiti (Damiano Hiolle Cañamero, 2011).

La forza del metodo sintetico risiede nel fatto che l'analisi non è più subordinata alla sintesi, come era tradizionalmente. Esso, infatti, apre la via verso l'indagine della complessità della natura, non più in chiave riduzionista ma costruttivista: la ricerca scientifica è un atto di costruzione, da esercitare su un sistema che non esisterebbe senza tale atto, e richiede ai ricercatori di pensare e implementare la scienza come una forma di conoscenza che crea in modo attivo – non riflette passivamente – gli oggetti esplorati (Damiano Cañamero, 2012).

Il metodo sintetico nelle scienze cognitive ha delle analogie con il lavoro di Newell e Simon, soprattutto nella cosiddetta *Information Processing Psychology* (Newell Simon, 1972). Nell'approccio di Newell e Simon, a un agente umano viene dato un compito di risoluzione di un problema – come un esercizio di logica o la scelta di una mossa a scacchi – e gli viene chiesto di pensare a voce alta, in modo da avere un resoconto dei processi mentali in corso man mano il compito richiesto viene portato a termine. I resoconti verbali vengono poi analizzati e usati per sviluppare un programma che simuli il comportamento dell'agente umano (Angius Perconti Plebe, 2024). Più di recente, il metodo sintetico – simulativo – è stato applicato anche nel campo della biorobotica, ad esempio nella simulazione della chemiotassi[30] nelle aragoste (Grasso Consi Mountain, 2000), della fonotassi[31] nei grilli (Webb, 2002), della costruzione di formicai (Lambrinos Möller Labhart, 2000) e dei movimenti dei topi (Burgess Donnett Jeffert, 1997).

A Newell (1980, 1990) si attribuisce anche la formulazione di una serie di criteri utili a valutare in che misura un sistema artificiale può fornire una base computazionale adeguata a supportare una teoria della cognizione o, meglio, a offrirne una rappresentazione simulativa. John Anderson e Christian Lebiere (2003) hanno condensato la lista in dodici criteri chiamati "Newell Test for a Theory of Cognition". Secondo gli

[30] La chemiotassi è un fenomeno biologico in cui una cellula o un organismo unicellulare si muove in risposta a un gradiente chimico nell'ambiente circostante. Molti microrganismi utilizzano tale processo per trovare cibo (spostandosi verso concentrazioni maggiori di nutrienti) o per evitare sostanze tossiche (allontanandosi da concentrazioni elevate di tossine). Esso si applica anche nei processi cellulari degli organismi multicellulari, come il movimento delle cellule immunitarie verso il sito di un'infiammazione.
[31] La fonotassi è il fenomeno biologico per cui un organismo è in grado di orientarsi verso la fonte di un suono. Questo fenomeno è particolarmente comune in molte specie animali, come insetti, anfibi e uccelli, dove i suoni prodotti da individui della stessa specie servono a segnalare la presenza di un potenziale partner, predatore o altra risorsa importante.

autori, tali criteri sono vincoli funzionali per le CA. I primi nove delineano gli obiettivi fondamentali che una CA deve perseguire per riprodurre le capacità intellettive umane, mentre gli ultimi tre specificano i vincoli relativi alle modalità con cui tali funzioni devono essere implementate. Di conseguenza, i criteri non includono tutte le aspettative attese di una teoria cognitiva. Essi sono: comportamento flessibile – ovvero, un sistema che rifletta una teoria cognitiva dovrebbe essere abbastanza flessibile da imparare e attuare compiti cognitivi in modo sufficientemente corretto; performance in tempo reale − ovvero, i compiti devono essere portati a termine nella stessa quantità di tempo che impiegherebbe un essere umano; comportamento adattivo: le CA dovrebbero avere meccanismi che agevolano la loro adattività; vasta conoscenza di base e comportamento dinamico: un sistema con una teoria cognitiva deve essere in grado di avere a che fare con l'incertezza e con un ambiente in cambiamento. A questi si aggiunge la necessità di saper integrare le conoscenze: diversi tipi di conoscenza dovrebbero essere integrati in una CA per fornire un ampio spettro di capacità inferenziali, simile a quello degli esseri umani, come induzione, deduzione, abduzione, analogia, capacità decisionale. Tipica di una CA dovrebbe essere anche la capacità di manipolare il linguaggio naturale, nonché di apprendere, acquisire competenze, svilupparsi ed evolversi: ciò significa che le abilità complessive finora menzionate dovrebbero crescere nel tempo, riflettendo i processi evolutivi che hanno condotto alla selezione specifici meccanismi e strategie. Newell suggerisce anche che le componenti delle CA dovrebbero essere mappate sulle strutture cerebrali e che questi collegamenti dovrebbero portare a un'implementazione neurale, in modo tale che il calcolo delle strutture neurali corrisponda a quello delle componenti assegnate – in altre parole, una CA dovrebbe possedere qualcosa che ricordi un cervello. Questo criterio mette in luce come Newell, inizialmente poco interessato alle neuroscienze, abbia progressivamente riconosciuto l'importanza della biologia come elemento chiave per arricchire i vincoli posti ai sistemi cognitivi, specialmente ai livelli più elevati di astrazione (Lieto, 2021). Infine, una CA dovrebbe possedere una coscienza, ovvero includere una teoria della coscienza e saperla modellare.

Quest'ultimo aspetto risulta particolarmente controverso, in quanto non esiste un consenso univoco su alcuna teoria della coscienza[32]. Tuttavia, numerosi modelli

[32] Il dibattito sulla coscienza, sia umana sia artificiale, è estremamente vasto e complesso, e non può essere trattato in modo esaustivo in questa sede. Tuttavia, affrontare il tema delle CA senza fare riferimento alla

computazionali tentano di affrontare questa sfida, esplorando diverse prospettive per comprendere e rappresentare la coscienza. La natura della coscienza, infatti, si configura come una questione straordinariamente complessa. Sin dalla sua definizione come *hard problem* da parte di David Chalmers (1995), non è stata raggiunta una spiegazione universalmente accettata o soddisfacente. Il dibattito sulla possibilità di sviluppare una coscienza artificiale è emerso come uno dei temi centrali negli studi contemporanei sulla mente. L'assenza di una definizione condivisa sottolinea quanto sia arduo replicare in un sistema artificiale un fenomeno così sfuggente. Gli approcci più comuni in questo senso, come ricostruiti da Riccardo Manzotti e Antonio Chella (2020), sono il computazionalismo, il cosiddetto "sciovinismo" biologico e neuronale, il cognitivismo e il cosiddetto "corporeismo" (Manzotti, 2019). Il computazionalismo nell'ambito dello studio della coscienza è l'idea per cui l'informazione o la computazione siano una sostanza aggiuntiva che viene in qualche modo generata da una macchina computazionale (Marconi, 2001). La nozione di informazione, in particolare, ha assunto negli anni una sempre maggiore autonomia, fino a essere quasi considerata qualcosa di concreto che viene acquisita, immagazzinata, processata e trasmessa (Alexander Gamez, 2020; Tononi, 2004). Sebbene la metafora dell'intelligenza umana come meccanismo di elaborazione delle informazioni domini l'immaginario comune e appaia ampiamente supportata dalla comunità scientifica, essa richiede una valutazione critica. Questo approccio, infatti, è stato oggetto di numerose critiche da parte di neuroscienziati ed esperti di IA, che ne mettono in discussione la validità e i limiti interpretativi (Brette, 2019; Epstein, 2016; Manzotti, 2012).

Proseguendo con la categorizzazione di Manzotti e Chella, lo "sciovinismo" biologico e neuronale è quell'approccio che associa la coscienza esclusivamente ai processi biologici e neuronali. I sostenitori di questa prospettiva, come gli enattivisti, considerano l'autopoiesi – la capacità di un organismo di auto-organizzarsi secondo Varela e Maturana – un elemento fondamentale della mente (Di Paolo, 2002; Froese Taguchi, 2019; Thompson, 2007). Tuttavia, non vi sono prove solide che colleghino necessariamente vita e coscienza (Manzotti Chella, 2020). Una variante di questo

nozione di coscienza artificiale significherebbe ignorare un aspetto cruciale della relazione tra cognizione e IA.

approccio è lo sciovinismo neuronale, che attribuisce al cervello proprietà emergenti speciali, mai confermate empiricamente (Manzotti Rossi, 2023).

Anche il cognitivismo ripropone la convinzione che la coscienza emerga dalla struttura cerebrale, senza però fornire una spiegazione chiara di come ciò avvenga (Manzotti Chella, 2020). Infine, il "corporeismo", strettamente legato alla versione dell'*embodied cognition*, presenta ulteriori problematiche. Nonostante si riconosca il ruolo causale del corpo e dell'ambiente nel plasmare coscienza e cognizione, manca una dimostrazione convincente di come esse si costituiscano attraverso l'interazione tra corpo e ambiente (Block, 2005; Gallagher Nelson, 2003).

In generale, le attuali teorie della coscienza soffrono spesso di ragionamenti circolari, assumendo ciò che dovrebbero dimostrare (Manzotti Chella, 2020). Ciò rende complicato sviluppare una teoria esaustiva e coerente sulla natura della coscienza, sia biologica sia artificiale, che sarebbe auspicabile in quanto contribuirebbe all'implementazione di IA etiche (Chella, 2023).

Per via di difficoltà teoriche e applicative, dunque, lo sviluppo di CA avanzate finora si è concentrato prevalentemente sulla riproduzione di specifiche capacità cognitive, come apprendimento, percezione e riconoscimento, senza necessariamente implicare la presenza di una consapevolezza simile a quella umana. Negli ultimi decenni sono state realizzate svariate CA dalle diverse caratteristiche e sono stati testati agenti che basandosi su tali strutture svolgono compiti cognitivi che coinvolgono apprendimento, percezione, esecuzione di azioni, attenzione selettiva e riconoscimento (Thórisson Helgasson, 2012). Tra queste, *SOAR* (Laird, 2012), *ACT-R* (Anderson Bothell Byrne, 2004), *CLARION* (Sun, 2005) e *iCub* (Vernon Metta Sandini, 2007).

Nello specifico, *SOAR* (*State, Operate And Result*) venne progettata da Newell e John Laird negli anni '80 con l'obiettivo che fosse una teoria generale della cognizione, in grado di simulare molteplici attività cognitive. Utilizzava una struttura basata su regole e sfruttava un meccanismo di apprendimento detto *chunking*, che permette al sistema di creare nuove regole in base all'esperienza modellando la cognizione come una sequenza di decisioni che trasformano lo stato corrente in un nuovo stato. *SOAR* è stata utilizzata in vari ambiti, come la risoluzione di problemi e la pianificazione (Laird Newell Rosenbloom, 1987).

ACT-R (*Adaptive Control of Thought-Rational*), sviluppata da John Anderson agli inizi del Duemila, è una CA basata sulla teoria dell'elaborazione simbolica, con obiettivo la spiegazione di come la mente umana organizza la conoscenza per svolgere compiti cognitivi complessi. L'architettura è composta da moduli specializzati (come, ad esempio, un modulo per la memoria dichiarativa e uno per la memoria procedurale), ciascuno con una funzione specifica. Essa è usata per simulare processi cognitivi come linguaggio, memoria e attenzione (Anderson Lebiere, 1998).

CLARION (*Connectionist Learning with Adaptive Rule Induction On-line*), sviluppata da Ron Sun negli stessi anni di *ACT-R*, integra invece modelli simbolici per rappresentare la cognizione umana. É composta da vari livelli, tra cui uno implicito che utilizza reti neurali per l'apprendimento, e uno esplicito, che rappresenta la conoscenza attraverso regole simboliche, ed è stata utilizzata per studiare i processi decisionali e di apprendimento (Sun, 2005).

Icub (*indexed Cognitive Universal Biped*)[33], infine, non è una CA in senso tradizionale, bensì un robot umanoide sviluppato come progetto europeo dall'Istituto Italiano di Tecnologia di Genova, progettato per la ricerca sulla cognizione e lo sviluppo di abilità motorie e cognitive, in particolare per studiare come i bambini imparano attraverso l'interazione con l'ambiente. *Icub* è dotato infatti di capacità motorie e percettive avanzate, che permettono l'esplorazione della cognizione incarnata. La CA di *iCub* si basa su moduli *software* che imitano funzioni cognitive umane come il riconoscimento di oggetti, la percezione spaziale e l'apprendimento motorio, dunque, è utile per testare teorie cognitive in ambienti dinamici e complessi (Metta Natale Nori, 2010).

Il design delle CA ha seguito approcci differenti, basati sulle diverse specificità degli obiettivi che si volevano raggiungere attraverso tali strumenti. In particolare, le architetture che cercano di imitare un modello unico standard della mente, sono progettate secondo l'approccio cosiddetto "cognition in the loop", ispirato alla tradizione cibernetica e sintetica menzionata precedentemente (Cordeschi, 2002), in cui si ritiene che la simulazione computazionale dei processi biologici e cognitivi giochi un ruolo epistemologicamente centrale nello sviluppo delle teorie sugli elementi che caratterizzano il comportamento intelligente. All'interno di tale quadro teorico, il dibattito tra i modelli

[33] Per un approfondimento su iCub: https://icub.iit.it/ (consultato in data 8/11/2024).

funzionalisti, che considerano solo una vaga equivalenza in termini di organizzazione funzionale tra i processi cognitivi e l'IA, e i modelli strutturalisti, basati su una sovrapposizione maggiore tra attività cognitiva e IA, ha visto prevalere i modelli funzionalisti, principalmente per ragioni di natura pratica (Lieto Bhatt Oltramari, 2018). Tuttavia, è stato spesso sottolineato come i modelli progettati in accordo alla prospettiva funzionalista non siano dei buoni candidati per i progressi nell'ambito dell'IA cognitiva, in quanto i meccanismi e le scelte di design complessive adottate per costruire tali strumenti impediscono loro di avere qualsivoglia ruolo esplicativo rispetto ai loro presunti analoghi in natura (Lieto, 2021). L'approccio strutturalista, d'altra parte, riconosce la necessità di una connessione maggiore e più forte tra la progettazione dei sistemi artificiali e delle loro architetture e processi interni, e le architetture corrispondenti disponibili in natura. I modelli e i sistemi artificiali che rispondono a dei criteri strutturali possono essere utili sia per portare avanti la ricerca in IA in termini di avanzamento tecnologico, sia a giocare il ruolo di "esperimenti computazionali" in grado di fornire risultati utili a rifinire o ripensare aspetti teoretici che riguardano i sistemi naturali target usati come fonte di ispirazione (Cordeschi, 2002). Una criticità dello strutturalismo è che non è possibile riprodurre una replica realistica artificiale di un sistema naturale, o che, anche se è possibile farlo, sorgono problemi relativamente alla sua interpretabilità (Kitano Hamahashi Luke, 1998). La ricerca di metodi strutturali, infatti, rischia di produrre una regressione sempre maggiore al mondo microscopico, fino al notissimo paradosso di Wiener, per il quale "the best material model of a cat is another, or preferably the same, cat" (Rosenblueth Wiener, 1945). Ciononostante, rimane centrale l'esigenza di individuare il livello corretto di rappresentazione e applicare i vincoli necessari per ottenere una computazione che imiti quella umana. Da questa prospettiva, il progresso può avvenire solo attraverso lo sviluppo di modelli strutturali credibili della cognizione umana, basati su una corrispondenza più rigorosa tra i processi di IA e i processi cognitivi umani.

Pertanto, una soluzione da perseguire si trova nel considerare la dicotomia funzionalismo/strutturalismo come i poli estremi di uno spettro: tra la mancanza di utilità esplicativa tipica dei modelli artificiali puramente funzionali e l'infattibile realizzabilità di modelli puramente strutturali, è possibile individuare una via di mezzo, ovvero una

quantità di modelli *proxy*, con diversi gradi di spiegabilità rispetto al sistema naturale considerato come fonte di ispirazione (Lieto, 2021).

Ad oggi, alcuni modelli funzionali hanno raggiunto risultati impressionanti. Nel 2012, il gruppo dell'università di Toronto guidato da Geoffrey Hinton, l'inventore del DL, vinse ImageNet, la competizione di classificazione di immagini più complicata al mondo, con il modello di rete neurale convoluzionale profonda AlexNet. Nel 2016, la compagnia *DeepMind* fondata da Demis Hassabis e acquisita da *Google*, sconfisse il campione mondiale di *Go* (Silver Huang Maddison, 2016). Tuttavia, nonostante la propaganda dei media, il cosiddetto *Alpha Go* non poteva essere considerato un "cognitive computing system"[34] (Lieto, 2021, 22), in quanto non forniva nessuna spiegazione su come gli esseri umani prendono decisioni quando pianificano le loro mosse giocando a un gioco come *Go*.

Tali risultati inaugurarono la cosiddetta era del Rinascimento in IA (Perconti Plebe, 2020; Tan Lim, 2018), con un inaspettato successo del DL (Plebe Grasso, 2019): esso, infatti, contiene solo piccoli miglioramenti rispetto alle reti neurali artificiali (RNA), campo che era stato stagnante per molto tempo e al quale lo stesso Hinton aveva contribuito sostanzialmente (Hinton McClelland Rumelhart, 1986). Le tecniche convenzionali di ML, infatti, erano limitate nella capacità di elaborare dati naturali nella loro forma grezza: per decenni, costruire un sistema di apprendimento automatico aveva richiesto abilità ingegneristiche elevate e profonda conoscenza del dominio specifico di lavoro per progettare un cosiddetto estrattore di caratteristiche che trasformasse i dati grezzi (come i valori in pixel di un'immagine) in una rappresentazione interna adeguata ("feature vector"), tramite cui il sottosistema di apprendimento, spesso un classificatore, rintracciasse o classificasse i modelli presenti nell'input (LeCun Bengio Hinton, 2015). Analogamente, i metodi di DL sono metodi di apprendimento per rappresentazione strutturati però su livelli multipli, ottenuti componendo moduli semplici ma non lineari[35],

[34] I media presentarono nello stesso modo anche il sistema *IBM Watson*, che era in grado di rispondere a quesiti di cultura generale e nel 2010 sconfisse il campione del gioco a quiz *Jeopardy!*. *IBM Watson*, tuttavia, non forniva alcuna conoscenza rispetto a come gli esseri umani immagazzinano e usano le informazioni per rispondere domande in tali situazioni (Lieto, 2021).

[35] Una struttura non lineare è un'organizzazione di elementi in cui questi non seguono un ordine sequenziale e possono essere collegati a più elementi contemporaneamente. Le strutture non lineari permettono quindi connessioni complesse tra gli elementi, consentendo l'accesso e la navigazione tramite percorsi multipli. Esempi di strutture non lineari includono alberi e grafi. Nelle strutture non lineari, è possibile attraversare i dati in più modi, il che le rende adatte per rappresentare gerarchie, reti o relazioni complesse.

ciascuno dei quali trasforma una rappresentazione che si trova a uno specifico livello in una rappresentazione a un livello più alto, cioè di astrazione maggiore. Con la composizione di un certo numero di trasformazioni di questo tipo, possono essere apprese funzioni molto complesse: un caso indicativo è quello dei compiti di classificazione, in cui i livelli più alti di rappresentazione amplificano gli aspetti dei dati in input che sono importanti per la differenziazione tra *pattern* ed eliminano le variazioni irrilevanti. Per esempio, un'immagine è rappresentata come una matrice di valori di pixel, e nelle reti di DL i livelli iniziali identificano caratteristiche di base, come la presenza o assenza di bordi con determinate orientazioni e posizioni. I livelli successivi rilevano motivi più complessi, combinando i bordi per formare parti di oggetti familiari, fino a identificare l'oggetto completo. La caratteristica fondamentale del DL è che queste caratteristiche non sono progettate dagli ingegneri, ma apprese dai dati attraverso un procedimento di apprendimento generico. È evidente quindi che il DL ha compiuto grandi progressi nel risolvere problemi che hanno a lungo afflitto la comunità di IA, dimostrandosi molto efficace nel rilevare strutture complesse in dati ad alta dimensionalità e applicabile a molti settori, dalla scienza, all'economia e al governo (LeCun Bengio Hinton, 2015).

Una delle differenze principali tra la prima generazione di RNA e gli attuali sistemi di DL si trova nella modifica fondamentale delle intenzioni con cui avviene l'implementazione: infatti, le prime RNA nacquero con lo scopo di studiare l'apparato cognitivo umano, mentre con il DL si ha un drastico cambiamento di direzione verso obiettivi prettamente ingegneristici. Gli sviluppatori, infatti, rinunciarono a perseguire indagini cognitive, e pertanto la modellazione delle reti neurali guadagnó moltissimo in termini di libertà, in quanto nell'implementazione si cominciarono a usare soluzioni matematiche del tutto estranee ai processi mentali. Tuttavia, è innegabile che il DL sia strettamente legato alle scienze cognitive, poiché ha permesso di sviluppare modelli artificiali in grado, per la prima volta, di svolgere compiti cognitivi complessi al pari degli esseri umani, e talvolta persino di superarli (Plebe Perconti, 2020). Tale cambiamento di paradigma nella ricerca in IA, se da un lato ha permesso un indiscusso avanzamento dal punto di vista tecnologico, dall'altro ha condotto a una progressiva perdita di interpretabilità dei modelli di DL, che, come menzionato precedentemente, sono finiti con l'essere identificati con *black boxes* (Castelvecchi, 2016; Gunning, 2017). Una via per

ritrovare la spiegabilità perduta potrebbe essere quella di rivolgersi nuovamente ai modelli cognitivi biologici.

Come messo in luce da Davide Vernon (2017), i due approcci alla relazione tra CA e IA delineati finora – funzionalista e strutturalista – pur non essendo necessariamente complementari, sono entrambi importanti e dovrebbero essere mantenuti per osservare se e in che misura elementi di successo di un approccio possano essere adattati all'altro. Un possibile terreno comune per valutare i progressi in questo senso è analizzare i problemi che sono facili da risolvere per gli esseri umani, ma molto complessi per le macchine, come il ragionamento attraverso senso comune su spazio, azione, cambiamento e categorizzazione del linguaggio, l'attenzione selettiva, l'integrazione della percezione multi-modale, l'apprendimento da pochi esempi e l'integrazione di meccanismi di pianificazione, azione e monitoraggio degli obiettivi (Lieto Bhatt Oltremari, 2018). Tali processi sono particolarmente complessi da implementare, in quanto derivano dall'integrazione di funzioni cognitive diverse, e pertanto richiedono un certo livello di astrazione architetturale, andando oltre lo studio di ciascun componente singolo. Si tratta di sistemi devono essere robusti, resilienti e capaci di soddisfare specifici criteri qualitativi (Lieto Bhatt Oltremari, 2018). Ci sono inoltre tratti cognitivi che andrebbero considerati nei sistemi di IA spiegabile (soprattutto in quelli rivolti direttamente agli utenti finali). Ad esempio, gli esseri umani tendono a cercare spiegazioni selezionate: data la limitata capacità di elaborazione umana, l'intera catena di connessioni causali alla base di una decisione algoritmica in molti casi è troppo complessa e difficile da comprendere. Al contrario, una spiegazione sintetica che evidenzi i punti essenziali della catena causale è maggiormente gestibile (Lieto, 2021).

Dall'excursus condotto in questo paragrafo, dunque, emerge come l'indagine sul mondo e sulla natura possa essere guidata dalla costruzione di artefatti tecnologici. Se costruire è la via più diretta per comprendere, avere strumenti spiegabili – interpretabili – è a sua volta una strategia per studiare i fenomeni che non si comprendono. Tale tendenza, come affrontato nel prossimo paragrafo, si sposa con l'esigenza di un'IA che sia sostenibile.

2.2 Spiegabile come sostenibile

Come illustrato nel Capitolo I, il concetto di sostenibilità è un tema complesso e articolato, che abbraccia molteplici dimensioni: ecologica, economica e sociale. All'interno della sostenibilità sociale, un elemento centrale è rappresentato dall'etica dell'IA, che include anche l'esigenza di garantire la spiegabilità e la correttezza degli algoritmi (Floridi, 2022). Infatti, la ricerca nel campo dell'IA socialmente sostenibile coinvolge il modo in cui le IA interpretano e rispondono al mondo e, viceversa, il modo in cui gli sviluppatori e gli utenti si rapportano ad essa.

Floridi (2022) identifica cinque principi etici chiave per l'IA, ovvero *beneficenza*, *non maleficenza*, *autonomia*, *giustizia* e *esplicabilità*, "che include sia il senso epistemologico di 'intelligibilità' sia il senso etico di 'responsabilità'" (Floridi, 2022, 101). I primi quattro principi sono quelli tradizionali della bioetica: la beneficenza implica che la tecnologia promuova il benessere e preservi la dignità di tutte le creature, sostenendo il pianeta. La non maleficenza riguarda la cautela rispetto alle violazioni della privacy personale e il principio di autonomia significa "trovare un equilibrio tra il potere decisionale che ci riserviamo e quello che deleghiamo agli agenti artificiali" (Floridi, 2022, 98). La decisione di prendere o delegare decisioni non è distribuita in modo equo nella società, e pertanto "lo sviluppo dell'IA dovrebbe promuovere la giustizia e cercare di eliminare tutti i tipi di discriminazione" (*Dichiarazione di Montréal*, 2017[36]). A corollario di tali principi, si aggiunge in modo conseguente e naturale l'esigenza di spiegabilità, ovvero esplicabilità:

> Questo principio completa gli altri quattro: affinchè l'IA sia benefica e non malefica, dobbiamo essere in grado di comprendere il bene o il danno che sta effettivamente facendo alla società e in quali modi; affinchè l'IA promuova e non limiti l'autonomia umana la nostra "decisione su chi dovrebbe decidere" deve essere informata dalla conoscenza di come l'IA agirebbe al nostro posto e, in tal caso, di come migliorare le sue prestazioni; e, affinchè l'IA sia giusta, dobbiamo capire chi ritenere eticamente o legalmente responsabile in caso di un esito grave e negativo, il che richiederebbe a sua volta un'adeguata comprensione del perché tale esito si sia prodotto (Floridi, 2022, 101).

[36] La *Dichiarazione di Montréal per l'IA responsabile* (Università di Montréal, 2017) è stata uno dei primi tentativi concreti di stabilire linee guida etiche per l'IA a livello internazionale. È stata formulata sotto gli auspici dell'università di Montréal attraverso un processo partecipativo, coinvolgendo esperti di diversi settori e cittadini, a seguito del *Forum sullo sviluppo socialmente responsabile dell'IA* tenutosi nel novembre 2017 (Floridi, 2022, 94).

Preso atto della necessità dell'IA per il bene sociale, tale spinta va di pari passo con il principio di beneficenza: l'uso dell'IA, infatti, deve fornire benefici alle persone (preferibilità sociale) e al mondo naturale (sostenibilità) (Floridi, 2022, 229). É in questo senso che IA socialmente sostenibile e IA spiegabile si incontrano e fondono, dal momento che la necessità di strumenti interpretabili garantisce non solo trasparenza e fiducia nelle decisioni prese dai sistemi intelligenti, ma anche la capacità di identificare e correggere eventuali errori o pregiudizi. Ciò ha ripercussioni anche sulla dimensione della sostenibilità ambientale: un'IA spiegabile permette infatti la valutazione dell'impatto delle sue azioni e decisioni sull'ambiente e sulla società, rendendo possibile l'adozione di un approccio etico e responsabile che consideri tanto la giustizia sociale quanto l'equilibrio ecologico. Gli obiettivi dell'IA spiegabili, dunque, sono i medesimi dell'IA sostenibile, ovvero la produzione di sistemi che siano non solo tecnicamente avanzati, ma anche eticamente consapevoli e orientati al benessere globale.

Come discusso nel §1.2, il modello dominante di regolamentazione dell'IA si basa sull'autoregolamentazione delle aziende, un approccio che solo di recente si sta cercando di integrare in un quadro normativo sistematico e completo. Finora, l'autoregolamentazione ha comportato, "an appearance of responsible design and deployment while at the same time avoiding genuine external accountability and explainability to vulnerable recipient populations and economies" (Mazzi Floridi, 2023, 52). Inoltre, l'autoregolazione permette alle grandi aziende di eludere parte delle critiche mosse contro di loro. Ciò avviene non solo attraverso la denigrazione di ricerche considerate una minaccia, ma anche tramite il finanziamento strategico di critici meno incisivi, favorendo la diffusione di una nozione di IA etica vaga e imprecisa (Whittaker, 2021).

Vinuesa e colleghi (2020) sottolineano come sia possibile usare metodi di IA spiegabile per rendere i modelli di IA e i loro risultati interpretabili dal pubblico generale, nonché per promuovere strategie efficaci dal punto di vista delle politiche aziendali per gli OSS. Per fare un esempio, Vinuesa e Beril Sirmacek (2021) hanno analizzato uno studio di Neal Jean e colleghi (2016), che da immagini satellitari di territori abitati individuano determinate caratteristiche (come intensità delle luci notturne, materiale di rivestimento dei tetti, distanza dalle aree urbane) di cui si servono per fare predizioni sui consumi economici medi giornalieri pro capite. Partendo da qui, Vinuesa e Sirmacek

(2021) mostrano che aggiungere interpretabilità a questo modello sarebbe utile per comprendere l'influenza di ciascuno dei parametri sul risultato finale, in modo da ottenere uno strumento efficace per il tracciamento della povertà (OSS n. 1). Attraverso una rappresentazione simbolica accurata, infatti, si potrebbe inquadrare quali tra i fattori individuati si dovrebbero mantenere o eliminare per intervenire su situazioni socialmente critiche (Mazzi Floridi, 2023). Altre applicazioni di IA spiegabili per gli OSS sono, ad esempio nell'ambito della gestione dell'energia, l'ottimizzazione del consumo energetico attraverso l'analisi dettagliata dei dati di utilizzo, con l'identificazione di inefficienze e il suggerimento di interventi mirati. Nel settore agricolo, l'applicazione di IA spiegabile permette di monitorare le condizioni del suolo e delle colture, fornendo raccomandazioni precise per l'irrigazione e l'uso di fertilizzanti e riducendo dunque lo spreco di risorse e l'impatto ambientale. Inoltre, nell'industria manifatturiera, sistemi di IA predittiva spiegabile facilitano la manutenzione dei macchinari, prevenendo guasti e prolungando la vita utile delle attrezzature, con conseguente diminuzione dei rifiuti industriali (Lakshmi Tiwari Dharanaj, 2024).

La ricerca su modelli di ML che siano spiegabili e al contempo ecologicamente sostenibili ha avuto uno sviluppo importante soprattutto negli ultimi anni (Lakshmi Tiwari Dhanaraj, 2024). In particolare, i requisiti di "sustainability" e "fairness" sono stati attenzionati dalle regolazioni nel campo dell'IA attualmente al vaglio (Giudici Raffinetti, 2024). É importante notare, a tale proposito, che in letteratura il termine "sostenibilità algoritmica" assume anche la valenza di "robustness with respect to anomalous data" (Giudici Raffinetti, 2024, 1). In questo senso, l'essere sostenibile include una sfumatura di "resilienza". Tale concetto in letteratura viene usato con vari significati, ovvero "inerzia", "elasticità" e "plasticità" (Den Hartigh Hill, 2022). Inerzia è un termine inteso spesso proprio come sinonimo di robustezza, ovvero "the ability to resist change when subjected to a disturbing force" (Miller Albarracin Pitliya, 2022, 2); "elasticità" indica "the ability to flexibly return to good states following a perturbation" (Miller Albarracin Pitliya, 2022, 2); e infine plasticità "the ability to expand the repertoire of good states – and courses of action – in the face of a changing environment" (Miller Albarracin Pitliya, 2022, 2). In modo intuitivo, un agente è considerato resiliente quando possiede la capacità di affrontare e superare con successo situazioni di stress o perturbazioni, riuscendo in particolare a ripristinare il proprio funzionamento ottimale dopo aver subito un impatto o

un cambiamento significativo. La resilienza implica, infatti, non solo la capacità di assorbire gli shock e resistervi, ma anche quella di adattarsi e ritornare a uno stato di equilibrio o di buon funzionamento. Conseguentemente, i significati di inerzia, plasticità ed elasticità sono legati: l'inerzia si riferisce alla capacità di resistere a un intervento esterno, evitando modifiche al proprio stato attuale, ma non implica necessariamente la possibilità di recupero delle proprie facoltà precedenti, o di ritorno a una condizione di buon funzionamento una volta che tale intervento ha avuto luogo. La plasticità, invece, rappresenta la capacità di adattarsi e modificarsi in risposta ai cambiamenti, aumentando la tolleranza e la flessibilità rispetto a nuove condizioni. La resilienza, quindi, può essere vista come un concetto che integra in parte la resistenza (inerzia) e l'adattamento (plasticità), includendo la capacità di ristabilire una condizione ottimale anche dopo eventi perturbativi (elasticità) (Miller Albarracin Pitliya, 2022).

La robustezza come resilienza, dunque, viene considerata un elemento chiave tanto della spiegabilità quanto della sostenibilità ecologica degli algoritmi di ML, poiché consente a questi sistemi di mantenere prestazioni affidabili anche in presenza di variazioni o perturbazioni nei dati di input o nelle condizioni operative. La robustezza rende gli algoritmi meno suscettibili a errori o a comportamenti imprevisti, migliorando così la loro interpretabilità: un sistema che reagisce in modo prevedibile e stabile è infatti più facile da analizzare e comprendere. Inoltre, un algoritmo robusto riduce la necessità di risorse per frequenti riaddestramenti o correzioni, contribuendo a una maggiore sostenibilità operativa ed energetica (Chang Wang Wang, 2023). La combinazione di robustezza e resilienza favorisce l'adozione di modelli che siano non solo efficienti ma anche etici e trasparenti, permettendo agli sviluppatori di implementare soluzioni che non solo rispondono a criteri di performance, ma che rispettano anche principi di responsabilità e minimizzazione dell'impatto ambientale.

La robustezza di un algoritmo di ML non si limita alla capacità di mantenere prestazioni stabili di fronte a variazioni nei dati o nelle condizioni operative, ma si estende anche alla resistenza contro attacchi esterni. Nel contesto della *cybersecurity* la robustezza di un sistema implica la capacità di proteggersi da attacchi malevoli, in cui i malintenzionati manipolano i dati in input per indurre l'algoritmo a compiere errori significativi o generare risultati dannosi (Biggio Roli, 2018; Carlini Wagner, 2017). Tali attacchi possono sfruttare la sensibilità dell'algoritmo a piccole perturbazioni nei dati in

input, spesso impercettibili per l'occhio umano, ma in grado di confondere il modello e alterarne le prestazioni. Per esempio, una piccola alterazione in un'immagine potrebbe portare un sistema di riconoscimento facciale a identificare erroneamente una persona o un oggetto (Sharif Bhagavatula Bauer, 2016). Aumentare la robustezza di un sistema rispetto agli attacchi esterni comporta quindi l'integrazione di misure di sicurezza e resilienza che riducano la vulnerabilità e permettano al sistema di operare in modo affidabile anche sotto pressione o in situazioni avverse (Madry Makelov Schmidt, 2017).

Migliorare la robustezza contro attacchi esterni diventa cruciale per garantire non solo la sicurezza, ma anche la fiducia nell'utilizzo dell'IA in settori critici, come finanza, sanità e sorveglianza, dove errori indotti da manipolazioni potrebbero avere conseguenze gravi. Inoltre, la robustezza di un sistema di IA contribuisce a una maggiore sostenibilità di esso, in quanto favorisce la riduzione degli addestramenti, che, come si è visto, sono energeticamente dispendiosi e richiedono grandi quantità di dati: un sistema robusto che opera bene anche con input variabili o in condizioni avverse può mantenere le proprie prestazioni senza dover essere frequentemente modificato o ottimizzato. La robustezza riduce inoltre la probabilità di malfunzionamenti o errori significativi, aumentando l'affidabilità e la stabilità dell'algoritmo nel lungo periodo (Patterson Gonzalez Le, 2021; Radosavovic Kosaraju Girshick, 2020). Un sistema in grado di resistere agli attacchi è anche garanzia di maggiore sicurezza a lungo termine, poiché permette di evitare costose e complesse operazioni di recupero o risarcimento in caso di compromissione (Carlini Wagner, 2017). La robustezza consente poi di creare sistemi con una vita operativa più lunga, che non devono essere sostituiti frequentemente a causa di vulnerabilità o problemi di prestazione in situazioni inaspettate (Strubell Ganesh McCallum, 2019). Inoltre, riguardo all'aspetto della sostenibilità sociale, bisogna considerare che un sistema robusto può adattarsi a diversi contesti e servire un'ampia gamma di utenti senza introdurre *bias* o discriminazioni. Ad esempio, un algoritmo di riconoscimento facciale robusto non deve essere riaddestrato per diversi gruppi etnici o condizioni ambientali (Hendrycks Dietterich, 2019).

Da quanto detto finora, risulta evidente come gli aspetti della spiegabilità e della robustezza degli algoritmi, fondamentali per garantire un'IA etica e responsabile, siano strettamente legati e in parte sovrapponibili agli obiettivi di sostenibilità sociale e ambientale. Un algoritmo spiegabile è caratterizzato da una trasparenza operativa che non

solo facilita la comprensione dei suoi processi interni, ma promuove anche un utilizzo socialmente inclusivo della tecnologia. La spiegabilità, infatti, consente agli utenti esterni, non solo agli sviluppatori, di comprendere e valutare le decisioni dell'IA, rendendo possibile una maggiore fiducia e accettazione di questi sistemi e favorendo un approccio democratico alla tecnologia. Allo stesso tempo, la robustezza – intesa come capacità di un algoritmo di mantenere le proprie prestazioni nonostante variazioni o perturbazioni negli input – garantisce anche efficienza energetica. Un algoritmo che combina spiegabilità e robustezza non solo soddisfa criteri etici e di inclusività sociale, ma si allinea anche con i principi di sostenibilità ecologica, esibendo un funzionamento stabile e meno dispendioso dal punto di vista computazionale.

Considerata l'esigenza di modelli di IA spiegabile e la constatazione della perdita di spiegabilità come conseguenza dell'allontanamento da architetture di tipo cognitivo, una via che vale la pena provare a perseguire nella ricerca nel campo di IA sostenibili è il ritorno a CA, che possano essere interpretabili e robuste.

2.3 Verso modelli cognitivi sostenibili

Come visto nei paragrafi precedenti, il successo del DL, sebbene incentrato su obiettivi ingegneristici, riapre la possibilità di un dialogo tra scienze cognitive e IA (Perconti Plebe, 2020), che può essere utile per la progettazione di architetture spiegabili. A questo proposito, è utile riprendere il concetto di CA delineato da Newell (1980), che idealmente puntava a rappresentare le varie forma di intelligenza umana in modo tale da integrarle nella cosiddetta AGI, che ha alimentato innumerevoli scenari fantascientifici.

Come visto, storicamente si è avuta una tendenza di sviluppo dell'IA che replica più che altro singoli aspetti della cognizione umana (Lieto Bhatt Oltremari, 2018). Tuttavia, con l'avvento degli LLM, si è assistito a uno spostamento dell'ingegnerizzazione verso architetture che, pur non essendo esplicitamente progettate per emulare le capacità cognitive umane complete, le acquisiscono attraverso un addestramento e un adattamento estesi. Tali sistemi cognitivi non progettati si possono definire "undesigned cognitive architectures" (UCA) (Acciai Angius Perconti, di prossima pubblicazione). A differenza delle CA tradizionali, che sono progettate per simulare funzioni cognitive specifiche, le UCA si evolvono naturalmente attraverso processi di apprendimento profondo e interazioni con vasti insiemi di dati. Questa

evoluzione rispecchia la natura adattiva e dinamica della cognizione umana, dove l'apprendimento di concetti nuovi e l'adattamento a situazioni mai sperimentate prima avvengono continuamente in risposta alle interazioni ambientali. Le UCA sono modelli generici in grado di gestire un'ampia gamma di applicazioni – dalla generazione di testi alla traduzione – senza richiedere un diverso addestramento esteso per ogni nuovo compito. Tale flessibilità riduce la necessità di modelli specializzati e permette il risparmio di risorse computazionali e umane. Le capacità emergenti degli LLM consentono loro di scalare tra i vari domini, rendendoli strumenti preziosi per la risoluzione di problemi generali e consentendo di mantenere un livello di efficienza delle risorse in linea con gli obiettivi di sostenibilità. Una volta addestrati, gli LLM operano in modo efficiente su diversi compiti, con un fabbisogno minimo di energia aggiuntiva, in contrasto con l'elevato costo del riaddestramento continuo che è necessario per l'affinamento di modelli specifici per ogni compito (Acciai Angius Perconti, di prossima pubblicazione).

La flessibilità delle CA ha quindi un forte impatto sulla gestione delle risorse, sia dal punto di vista della sostenibilità ambientale sia della sostenibilità sociale. Tuttavia, è essenziale considerare come, benché vi siano analogie tra i processi cognitivi umani e alcuni processi algoritmici, che fanno ben sperare nell'efficacia delle CA, sussistono anche differenze sostanziali. Il processo decisionale delle macchine, ad esempio, differisce radicalmente da quello umano. Come evidenziato da Teppo Felin e Matthias Holweg (2024), i processi decisionali umani si basano su un ragionamento causale teorico orientato al futuro – "forward-looking" (Felin Holweg, 2024, 13) – che, cioè, funziona esplorando ipotesi controfattuali e sperimentando. Tale approccio consente di formulare teorie e generare nuove conoscenze, a differenza dell'IA che funziona attraverso processamento di dati e previsione probabilistica, restando ancorata ai dati passati. Le teorie basate sul ragionamento causale tipiche della cognizione umana, invece, consentono di superare i limiti imposti dai dati disponibili e di generare vere novità.

Tale differenza è particolarmente evidente nel contesto della spiegabilità. Come visto, la spiegabilità è fondamentale per permettere agli esseri umani di comprendere e valutare il processo decisionale algoritmico, specialmente nei contesti in cui sono richieste capacità creative e innovazione (Doshi-Velez Kim, 2017). È proprio attraverso lo sviluppo di sistemi spiegabili, che traducono l'opacità delle decisioni algoritmiche in

rappresentazioni comprensibili e verificabili, che può essere ridotto il divario tra cognizione umana e processi algoritmici. Un sistema di IA spiegabile non solo facilita il controllo umano, ma stimola anche una collaborazione più efficace tra esseri umani e macchine, unendo il ragionamento creativo umano all'efficienza computazionale dei modelli algoritmici.

Nel contesto del *decision-making* (DM) automatico, in particolare rispetto alle decisioni ad alta incertezza, è necessario esplorare soluzioni nuove e generare dati innovativi, più che esplorare informazioni già esistenti (Felin Holweg, 2024). Inoltre, il DM umano è un processo che integra considerazioni etiche, intuizioni ed esperienza ed è contestualizzato, ovvero, tiene conto delle specificità di ogni situazione. Il DM automatico, d'altra parte, è guidato prevalentemente da dati e metriche quantitative predefinite. Ciò lo rende altamente efficace in termini di velocità e precisione, ma potenzialmente cieco di fronte a implicazioni sociali ed etiche. I sistemi di IA, dunque, devono essere progettati con criteri etici espliciti e interpretabili per mitigare eventuali disparità causate dai dati di addestramento (Kirat Tambou Do Alexis, 2022).

La spiegabilità, dunque, diventa essenziale per comprendere il processo decisionale degli algoritmi e per identificare i punti in cui possono annidarsi *bias* che influenzano i risultati. La trasparenza dei modelli, infatti, consente di identificarne i punti vulnerabili, offrendo la possibilità di correggerli. La spiegabilità, inoltre, aumenta la fiducia degli utenti nei confronti dei sistemi IA, rafforzando così il legame tra cognizione umana e processi algoritmici (Rosenfeld Richardson, 2019). In questo contesto, l'introduzione di metriche di *fairness* si rivela fondamentale per monitorare l'equità di un sistema e fornire una base per l'implementazione di interventi correttivi in tempo reale. Queste metriche consentono di interrogare il sistema su risultati specifici e di apportare modifiche mirate, riducendo così l'impatto negativo dei *bias* preesistenti (Kirat Tambou Do, 2022). Tuttavia, mentre gli esseri umani hanno la capacità di acquisire consapevolezza dei propri pregiudizi, il processo di modificazione dei *bias* artificiali è spesso lento e soggetto a errori, in quanto gli algoritmi mancano di questa consapevolezza. La spiegabilità, dunque, assume un ruolo cruciale, poiché permette agli esseri umani di supervisionare le decisioni algoritmiche e di intervenire efficacemente quando necessario (Kirat Tambou Do, 2022).

Se il DM umano eccelle nella gestione creativa delle risorse, quello automatico richiede un rigoroso controllo delle variabili per ridurre al minimo errori e dispendio

energetico. La capacità di rendere spiegabili e trasparenti le decisioni prese da un sistema di IA, quindi, diventa non solo un obiettivo etico, ma anche un requisito tecnico per garantire efficienza e sostenibilità ecologica.

In questo contesto emerge l'importanza di modelli spiegabili che siano probabilistici ed esibiscano una struttura utile per comprendere la cognizione su più livelli di spiegazione. Giuseppe Boccignone e Roberto Cordeschi (2007) propongono i modelli bayesiani come strumenti ideali per tale scopo. Questi modelli utilizzano il teorema di Bayes, aggiornando la stima di probabilità in base a nuove evidenze o dati. Essi combinano credenze iniziali su un fenomeno con informazioni osservate per calcolare credenze aggiornate, risultando particolarmente efficaci nella rappresentazione di ragionamenti e decisioni presi in condizioni di incertezza (Koski Noble, 2009). I modelli bayesiani trovano applicazione in diversi ambiti della cognizione: nella percezione, descrivono come gli esseri umani integrano informazioni sensoriali ambigue; nel DM rappresentano i processi di ragionamento probabilistico; nell'apprendimento, spiegano come l'incertezza si riduca nel tempo grazie all'acquisizione di nuovi dati. La loro forza risiede nella possibilità di integrare diversi livelli di rappresentazione, aspetto cruciale per lo sviluppo di IA spiegabili e funzionali.

A supporto di questa prospettiva, è utile richiamare la distinzione proposta da David Marr (1982) tra i tre livelli di analisi necessari per la comprensione di un agente impegnato nell'esecuzione di un compito. Il primo livello è quello computazionale (*what/why*), che si occupa di definire lo scopo del processo, la sua logica e le ragioni della sua adeguatezza. Il secondo è il livello algoritmico (*how*), che analizza come la teoria computazionale venga rappresentata e tradotta in un modello. Infine, il terzo livello è quello implementativo, che riguarda il modo in cui la rappresentazione algoritmica viene realizzata fisicamente. Questa tripartizione riflette l'approccio epistemologico dei sistemi complessi, riconoscendo che i sistemi biologici sono costituiti da molteplici livelli di organizzazione. Il comportamento di un sistema complesso, come un organismo, può essere compreso considerando i vari livelli, come quello biochimico, cellulare, neurologico e psicologico. I modelli bayesiani possono essere utilizzati per formalizzare la triplice gerarchia di Marr, riducendola a due soli livelli principali, quello computazionale e quello implementativo, semplificando così l'analisi senza perdere la capacità di rappresentare la complessità (Chater Tenenbaum Yuille, 2006; Knill Kersten

Yuille, 1996; Boccignone Cordeschi, 2007). Pertanto, le simulazioni bayesiane possono fungere da ponte tra la comprensione teorica della cognizione umana e l'implementazione di sistemi di IA capaci di apprendere e adattarsi. I modelli bayesiani, per la loro natura probabilistica e trasparente, offrono un grande vantaggio per lo sviluppo di IA spiegabili, in quanto consentono di tracciare come e perché determinate decisioni siano state prese, migliorando la fiducia e l'interpretabilità dei sistemi. Tali modelli, dunque, permettono di gestire l'incertezza e di ottimizzare le risorse, offrendo un quadro teorico coerente con gli obiettivi della sostenibilità, come verrà approfondito nel capitolo successivo.

Dunque, alla luce di quanto discusso finora, l'obiettivo di questo capitolo è sostenere che l'adozione di modelli cognitivi specifici può consentire lo sviluppo di IA sostenibili, capaci di bilanciare efficienza, interpretabilità e adattabilità. Per tale scopo è necessario ristabilire un legame stretto tra l'IA e scienze cognitive.

Un esempio significativo del divario tra i due campi emerge nel confronto tra la proposta del neuroscienziato Karl Friston, relativa all'attività predittiva fondamentale del cervello e al principio di energia libera (*Free Energy Principle*, FEP), e gli sviluppi del DL. Il FEP, ampiamente noto e discusso nell'ambito delle scienze cognitive, offre una formulazione matematica dell'energia libera, basata sull'inferenza variazionale bayesiana (Friston Stephan, 2007; Friston Kiebel, 2009; Friston, 2010). Sul fronte del DL, un avanzamento analogo è stato ottenuto con l'introduzione di un'architettura chiamata "variational autoencoder", sviluppata indipendentemente da Diederik Kingma e Max Welling (2014) e da Danilo Jimenez Rezende e colleghi (2014). Nonostante le somiglianze matematiche, i collegamenti tra i due approcci non sono stati inizialmente riconosciuti, né dagli autori né dalla comunità del DL. Solo successivamente, André Ofner e Sebastian Stober (2018) hanno evidenziato la connessione (Perconti Plebe, 2020).

Questo esempio illustra chiaramente la necessità di un maggiore dialogo e integrazione tra le scienze cognitive e l'IA per promuovere lo sviluppo di modelli più sostenibili e teoricamente fondati. Il FEP, in particolare. rappresenta una proposta teorica avanzata che unisce sostenibilità ecologica e spiegabilità, offrendo una base per lo sviluppo di sistemi cognitivi artificiali capaci di rispondere dinamicamente alle esigenze ambientali e operative. Nei prossimi capitoli si discuterà in che modo questa teoria può essere integrata in modelli di IA sostenibile.

CAPITOLO III

IL PRINCIPIO DELL'ENERGIA LIBERA E L'INFERENZA ATTIVA

3.1 Un principio unificatore

Nel dibattito sui sistemi cognitivi, il FEP ha assunto di recente una rilevanza crescente ed è stato indicato da filosofi e neuroscienziati come un principio unificatore di carattere universale, applicabile a svariati campi disciplinari, dalla termodinamica alla biologia, dallo studio della mente fino alle scienze sociali (Veissière Constant Ramstead, 2020).

In termodinamica, la formula base dell'energia libera relativa a un sistema è

$$F = E - TS \tag{1}$$

dove *F* rappresenta l'energia libera (nota anche come energia libera di Helmholtz), ovvero l'energia disponibile per compiere lavoro utile in un sistema termodinamico a temperatura e volume costanti. Questa quantità è l'effettiva porzione di energia complessiva di un sistema che può essere utilizzata per un processo. *E* indica l'energia interna del sistema, ovvero l'energia totale che comprende l'energia cinetica, l'energia potenziale delle particelle e l'energia associata alle interazioni interne, come i legami chimici e le forze intermolecolari. *T* è la temperatura assoluta, che misura l'agitazione termica delle particelle nel sistema, mentre *S* rappresenta l'entropia, ovvero la misura del disordine o della quantità di energia non utilizzabile in un sistema: un aumento dell'entropia implica che una minore quantità di energia può essere convertita in lavoro utile. La formula (1), dunque, descrive l'energia libera rimanente, tenuto conto delle perdite dovute all'entropia, che può essere convertita in lavoro.

Tale formula è stata utilizzata anche per descrivere sistemi sociali complessi (Chen, 2009), dove *E* rappresenta la quantità di risorse disponibili per il consumo umano e i

cambiamenti di entropia riflettono le variazioni nel disordine. Un aumento delle possibilità legate al caso corrisponde a un incremento dell'energia di un sistema. Considerando l'entropia come la tendenza naturale di un sistema ad ampliarsi, è possibile stabilire un parallelismo con il desiderio umano di espansione. In questo senso, l'entropia può essere interpretata come una misura della libertà all'interno di una società umana (Chen, 2009). In un sistema fisico, che può includere organismi biologici, l'energia libera si muove spontaneamente verso il minimo: quando E è fissata, T e S tendono a crescere finché il prodotto TS raggiunge il massimo. Analogamente, in un sistema sociale, ciò potrebbe essere inteso come l'idea che le persone sono sempre alla ricerca del maggior livello possibile di espansione e libertà (Chen, 2009).

Il concetto di energia libera è trasposto nel FEP e applicato dal neuroscienziato Karl Friston allo studio del cervello. Come osservato da Matteo Colombo e Cory Wright (2021), il FEP è stato inizialmente proposto per spiegare come la corteccia sensoriale inferisca le cause dei propri input e apprenda regolarità causali. Tuttavia, mentre la formula (1) è concettualmente utile, la sua applicazione al FEP richiede adattamenti, poiché i termini *E*, *T* e *S* assumono significati più astratti: *E* rappresenta l'energia o risorse totali in un contesto più generalizzato, mentre *S* riflette l'incertezza epistemica del sistema piuttosto che il disordine termodinamico. Questo rende il FEP un'estensione della termodinamica che va oltre il suo ambito originario, in quanto viene applicato per chiarire le funzioni di azione, percezione e attenzione, nonché per spiegare i processi di evoluzione e sviluppo di un organismo (Friston, 2003; 2009; 2010; 2013; Friston Kilner Harrison, 2006; Friston Stephan, 2007). Inoltre, il FEP è stato presentato come uno strumento utile a caratterizzare e prevedere il comportamento adattivo degli organismi viventi, e come una proprietà oggettiva dei sistemi analizzati. Esso infatti si applica "to any biological system […] from single-cell organisms to social networks" (Friston, 2009, 293).

Secondo una delle prime formulazioni di Friston "the free energy principle says that any self-organizing system that is at equilibrium with the environment must minimize its free energy" (Friston, 2010, 127). Si tratta, sostanzialmente, di una formula che spiega come i sistemi adattivi, ovvero i sistemi biologici, come gli animali o il cervello, resistono a una naturale tendenza al disordine. L'idea sottostante al FEP è che gli organismi che esistono lo fanno perché sono in grado di persistere nel tempo, mantenendo il proprio

equilibrio e l'equilibrio con l'ambiente esterno tramite la minimizzazione dell'energia libera. L'energia libera è

> an upper bound on surprise, which means that if agents minimize free energy, they implicitly minimize surprise. Crucially, free energy can be evaluated because it is a function of two things to which the agent has access: its sensory states and a recognition density that is encoded by its internal states (for example, neuronal activity and connection strengths). The recognition density[37] is a probabilistic representation of what caused a particular sensation (Friston, 2010, 128).

Il numero di stati fisiologici e sensoriali in cui un organismo può trovarsi è naturalmente limitato, il che implica che la probabilità degli stati sensoriali deve avere un'entropia bassa. L'entropia, infatti, rappresenta una misura di incertezza, ossia di sorpresa[38] e la sopravvivenza di un organismo dipende dalla minimizzazione di tale sorpresa. Da un punto di vista matematico, ciò equivale a ridurre la somma delle probabilità logaritmiche negative degli esiti attesi. Inoltre, l'energia libera nel FEP è formalizzata anche nei seguenti termini:

$$a, \mu, m \;=\; arg\ min\ F\ (s\sim, \mu \mid m) \qquad (2)$$

La formula (2) indica la minimizzazione dell'energia libera F delle sensazioni e la rappresentazione delle loro cause all'interno di un sistema. Ovvero, si ha come obiettivo la minimizzazione della funzione $F\ (s\sim, \mu \mid m)$, dove s rappresenta uno stato osservato o un insieme di rappresentazioni sensoriali, μ rappresenta variabili latenti o stati interni dell'agente e m rappresenta il modello generativo, ovvero la struttura probabilistica interna del sistema, utilizzata per inferire le cause degli input sensoriali (s) in base alle rappresentazioni latenti (μ). Dunque, minimizzare F equivale a ridurre la sorpresa attesa o l'energia libera, il che consente all'agente di migliorare la coerenza tra il proprio modello interno (μ, m) e le osservazioni esterne (s).

Inoltre, è possibile definire l'energia libera anche come:

[37] La densità di riconoscimento è una distribuzione di probabilità approssimata delle cause dei dati (come input sensoriali), ed è il risultato dell'inferenza o dell'inversione di un modello generativo (Friston, 2010).
[38] La sorpresa (o auto-informazione) è il logaritmo negativo della probabilità di un risultato. Un risultato improbabile (ad esempio, dell'acqua che scorre verso l'alto) risulta infatti sorprendente (Friston, 2010).

$$F = -log\ P(s|m) + D_{KL}\ [Q(\mu)\ ||\ P\ (\mu|s)] \quad (3)$$

In (3) $-log\ P(s|m)$ indica l'informazione associata a specifici stati (cioè la sorpresa), dove $P(s|m)$ è la probabilità delle osservazioni s dato il modello m. $D_{KL}\ [Q(\mu)\ ||\ P\ (\mu|s)]$ è la divergenza di Kullback-Leibler (KL)[39] tra la distribuzione variazionale $Q(\mu)$, ovvero la distribuzione che approssima la distribuzione vera, e la distribuzione condizionale $P\ (\mu|s)$, ovvero degli stati latenti μ dato s. In equilibrio, ovvero quando l'approssimazione è perfetta, cioè $Q(\mu) = P\ (\mu|s)$, la divergenza KL è pari a zero, e l'energia libera si riduce alla sorpresa: $F = -log\ P(s|m)$. Inoltre, la divergenza KL di P rispetto a Q rappresenta la sorpresa in eccesso attesa derivante dall'utilizzo di Q come modello quando la distribuzione effettiva è P (Friston Da Costa Sajid, 2023).

In termini di inferenza, quanto esposto sopra viene interpretato come un aggiornamento bayesiano delle credenze alla luce di nuove informazioni. Ciò avviene attraverso il calcolo della distribuzione posteriore, che combina una credenza iniziale (*prior*) con le nuove evidenze osservate (Hoff, 2009). Applicato al FEP, il teorema di Bayes si può esprimere come:

$$P\ (stato|osservazione) = \frac{P(osservazione|stato)\ P\ (stato)}{P\ (osservazione)}$$

In questa formula, $P\ (stato|osservazione)$ rappresenta la probabilità a posteriori, ovvero la probabilità dello stato data un'osservazione. Questo valore esprime la credenza aggiornata sullo stato del sistema alla luce delle nuove informazioni. $P\ (osservazione|stato)$, invece, è la verosimiglianza (*likelihood*), che indica quanto sia probabile osservare una certa evidenza dato uno stato ipotetico. $P\ (stato)$ è la probabilità a priori, ovvero la credenza iniziale o preesistente su quale sia lo stato prima di considerare l'osservazione, mentre $P\ (osservazione)$ è una costante di normalizzazione che garantisce che la somma delle probabilità a posteriori sia pari a 1. Questa rappresenta

[39] La divergenza KL (o divergenza dell'informazione, guadagno di informazione o entropia incrociata) è una una misura non commutativa della differenza non negativa tra due distribuzioni di probabilità (Friston, 2010).

la probabilità marginale delle osservazioni complessive e si ottiene considerando tutte le possibili combinazioni di stati. In tale contesto, gli stati interni rappresentano il modello che un sistema (o agente) ha del mondo. Gli stati esterni sono le caratteristiche dell'ambiente reale che il sistema cerca di inferire. Gli stati interni vengono aggiornati sulla base delle osservazioni fatte dal sistema. Questo aggiornamento è analogo al calcolo di una distribuzione posteriore bayesiana $P\ (stati\ esterni|osservazione)$. Il sistema, dunque, utilizza le osservazioni ricevute per aggiornare il proprio modello del mondo (stati interni), ottenendo così una rappresentazione più adeguata degli stati esterni. Inoltre, gli stati interni codificano credenze condizionali o posteriori bayesiane sugli stati esterni, ovvero, il sistema tiene traccia delle proprie ipotesi probabilistiche sugli stati esterni. Tali ipotesi vengono continuamente aggiornate, minimizzando la discrepanza tra ciò che il sistema si aspetta (modello interno) e ciò che osserva nel mondo esterno (Friston Da Costa Sajid, 2023).

Friston ha recentemente affermato che il FEP non dovrebbe essere considerato una teoria, bensì uno strumento o una metodologia. Come si è visto, esso parte dal presupposto che gli agenti esistenti siano in grado di persistere nel tempo mantenendosi attraverso la minimizzazione dell'energia libera, facendo affidamento esclusivamente su componenti prodotte internamente. Questo aspetto richiama il concetto di autopoiesi sviluppato da Varela e Maturana (1987), già discusso nel Capitolo II. In quell'approccio, la cognizione venne concepita per la prima volta non come un mero processo rappresentazionale, ma come azione e interazione con l'ambiente. Essa si configurò così come un processo di costruzione attiva che coinvolge simultaneamente l'organismo e l'ambiente circostante, evidenziando una visione innovativa e dinamica della relazione tra il vivente e il suo contesto (Cappuccio, 2009). Allo stesso modo, l'aggiornamento bayesiano nel FEP non è solo passivo, ma implica che il sistema possa agire sull'ambiente per ridurre l'incertezza sulle proprie credenze.

Una realizzazione pratica del FEP è il meccanismo del *predictive processing* (PP), o *predictive coding*, che descrive il funzionamento del cervello come macchina predittiva che minimizza l'energia libera o la sorpresa. Ciò significa che il cervello minimizza gli errori predittivi, ovvero le misure della discrepanza tra le previsioni *top-down* di una variabile casuale e il suo valore effettivo. Le previsioni *top-down* sono ipotesi che partono dalle strutture cognitive superiori (come aspettative o credenze) e guidano

l'interpretazione delle informazioni sensoriali che arrivano dal mondo esterno. Se ad esempio si considera un individuo che cammina in una foresta e sente un rumore (informazione sensoriale), il cervello potrebbe predire che si tratta di un animale basandosi su un modello interno costruito dall'esperienza. In altre parole, il cervello cerca di ridurre la mancata corrispondenza tra una specifica previsione che riguarda un input sensoriale e l'input effettivamente ricevuto. Contemporaneamente, cerca di stimare e rendere più precisa l'affidabilità del segnale dell'errore predittivo stesso (Friston Mattout Kilner, 2011; Clark, 2015).

La riduzione degli errori predittivi può essere ottenuta in diversi modi: attraverso un'inferenza immediata sugli stati non noti del mondo, che può spiegare la percezione; tramite l'aggiornamento di un modello globale del mondo per compiere predizioni migliori, che può spiegare l'apprendimento; infine, attraverso l'azione per campionare dati dal mondo che corrispondano alle predizioni (Millidge Seth Buckley, 2022). Il PP, dunque, è considerato un unico modello per percezione, azione e cognizione, e può essere descritto come un'inferenza bayesiana approssimata basata su inferenza gaussiana (Millidge Seth Buckley, 2022). Ciò significa che per ridurre l'incertezza rispetto agli stati futuri il cervello – o un sistema in generale – usa le informazioni ottenute tramite le interazioni avute precedentemente con l'ambiente, usando modelli generativi per predire gli input sensoriali e minimizzare l'energia libera (Friston Mattout Kilner, 2011).

Dunque, secondo il PP, tutti i fenomeni cognitivi provengono dallo stesso processo, ovvero la riduzione degli errori predittivi. Per tale ragione, Andy Clark sottolinea che "another area in which these models are suggestive of deep facts about the nature and construction of human experience concerns the character of perception and the relations between perception and imagery/visual imagination" (Clark, 2013a, 17). Infatti, nel cervello gli schemi guidati da meccanismi di predizione all'interno di sistemi gerarchici, come è il modello del PP, apprendono modelli generativi[40] probabilistici. In tale visione, il cervello è organizzato in una gerarchia di livelli, dove ogni livello elabora le informazioni a una scala diversa di astrazione (ed è per tale motivo che si parla di previsioni *top-down*). A ciascun livello, le popolazioni neurali cercano di generare predizioni sugli stati del livello sottostante, basandosi su un modello probabilistico. Ad

[40] Si ricorda che un modello generativo è una rappresentazione interna che descrive come il cervello si aspetta che i dati (gli input sensoriali) siano prodotti.

esempio, un livello più alto potrebbe predire il movimento di un oggetto, mentre il livello inferiore analizza i dati grezzi del movimento (come il cambiamento di posizione). Se il livello inferiore produce dati che non corrispondono alla predizione superiore, si genera appunto un errore predittivo. Tale errore viene inviato ai livelli più alti della gerarchia per aggiornare il modello generativo e migliorare le predizioni future. Come visto, dunque, il sistema corregge i propri modelli interni, consentendo un'elaborazione efficiente e adattiva, riducendo l'incertezza e ottimizzando il processo di apprendimento. Analogamente a quanto avviene nell'inferenza bayesiana, l'informazione presente è aggiornata costantemente con nuove informazioni e al contempo cresce l'abilità di generalizzare a situazioni nuove.

L'ipotesi di Clark, dunque, è che tale tipo di modello possa essere utile anche per la comprensione dei processi immaginativi, in quanto egli ritiene che il meccanismo di percezione e immaginazione siano analoghi, e quindi i percipienti siano soggetti immaginanti: "they are creatures poised to explore and experience their world not just by means of perception and gross physical action but also by means of imagery, dreams, and (in some cases) deliberate mental stimulations" (Clark, 2016, 84). Secondo Clark, la principale fonte di informazione di tutti i contenuti dell'esperienza per un essere umano è costituita dalle aspettative, anche considerando che alcuni contenuti vengono costantemente controllati, sfumati e selezionati dai segnali degli errori predittivi, e che quanto viene percepito è il mondo stesso, come rivelato dalla migliore ipotesi su di esso (Clark, 2013b). Inoltre, seguendo il PP, i percipienti possono servirsi della conoscenza immagazzinata per generare una sorta di rappresentazione interna simulata ("multilevel virtual analogue", Clark 2016, 85) del segnale sensoriale in arrivo, mentre questo si sviluppa attraverso i diversi livelli e tipi di elaborazione. Da ciò consegue l'immaginazione, in quanto il sistema guida un modello generativo capace di ricostruire il segnale sensoriale usando la conoscenza e le cause interagenti nel mondo. Così, emerge una forma di cognizione:

> it is a form in which perception, imagination, understanding and memory come as a kind of cognitive package deal – a package deal that locates the present where it experientially belongs, at the productive meeting point between past influence and informed future choice (Clark, 2016, 85).

Sulla base del PP come descritto da Clark, ovvero come meccanismo che include una profonda unità tra percezione e immaginazione, Michael Kirchhoff (2018) fa un passo ancora in avanti, proponendo la cosiddetta *Deep Unity Theory* (DUT), per la quale percezione e immaginazione si sovrappongono completamente in virtù della riduzione negli errori di predizione.

Dunque, il nodo che lega percezione, azione, attività cognitiva e immaginazione è la riduzione dell'energia libera, che avviene attraverso il processo dell'inferenza attiva (*active inference*, AIF), cioè il campionamento degli input sensoriali attesi per aumentare l'accuratezza delle previsioni. Il processo di AIF garantisce un costante aggiornamento delle informazioni acquisite dall'ambiente, portando a previsioni accurate degli stati futuri. In altre parole, attraverso l'AIF, un sistema – un organismo, o un agente – minimizza l'energia libera aggiornando il proprio modello del mondo attraverso l'osservazione e le conseguenti inferenze sugli stati del mondo stesso, oltre che attraverso l'azione. Ciò significa che un agente modifica attivamente il proprio ambiente o il proprio comportamento, che è definito dalle azioni, per rendere l'ambiente – e il futuro – più prevedibili. In questo modo, avvengono cambiamenti di stato autogenerati, nonchè la soppressione degli errori predittivi, attraverso le informazioni derivate dalla storia delle precedenti interazioni con l'ambiente (Friston Mattout Kilner, 2011; Kirchoff Parr Palacios, 2018). Il processo di AIF viene intuitivamente spiegata da Friston come: "feeling our way in darkness: we anticipate what we might touch next and then try to confirm those expectations" (Friston, 2010, 3).

In sintesi, si può dunque dire che il FEP è un quadro teorico generale che descrive come gli organismi individuali o sistemi più complessi minimizzino l'incertezza per mantenere i propri stati vitali. Il PP rappresenta il meccanismo specifico attraverso il quale il cervello riduce tale incertezza e può essere considerato anche un'implementazione operativa del FEP. L'AIF, infine, è il processo attraverso cui l'organismo o il sistema agisce per ridurre l'incertezza, integrando percezione, azione e apprendimento.

Nell'ambito del FEP e dell'AIF, Friston ha introdotto il concetto di coperte di Markov (*Markov blankets*, MB), sul quale vale la pena soffermarsi per evidenziare le ambiguità nella formulazione del FEP. Il concetto di MB deriva dalle scienze statistiche ed è stato introdotto per la prima volta da Judea Pearl (1988) nel contesto delle reti

bayesiane. Come accennato nel §2.3, una rete bayesiana è un modello statistico multivariato la cui struttura grafica permette di rappresentare e ragionare su un dominio incerto: ogni nodo è associato a una variabile del dominio e i collegamenti diretti tra i nodi rappresentano relazioni informative o di causa-effetto. Tali dipendenze sono quantificate da distribuzioni di probabilità condizionali, ovvero la misura della probabilità che un nodo sia influenzato dagli altri. Una MB è un sottoinsieme di una rete bayesiana: nello specifico, la MB di una variabile *X* è l'insieme costituito dai genitori di *X*, dai figli di *X* e dalle variabili che condividono un figlio con *X* (Koski Noble, 2009). Ciò significa che i nodi che formano la MB di *X* semplificano molto il calcolo del valore di *X*, e dunque la MB di una variabile random è l'unica conoscenza di cui si ha bisogno per prevedere il comportamento della variabile stessa (Facchin, 2021) (*Fig. 1*).

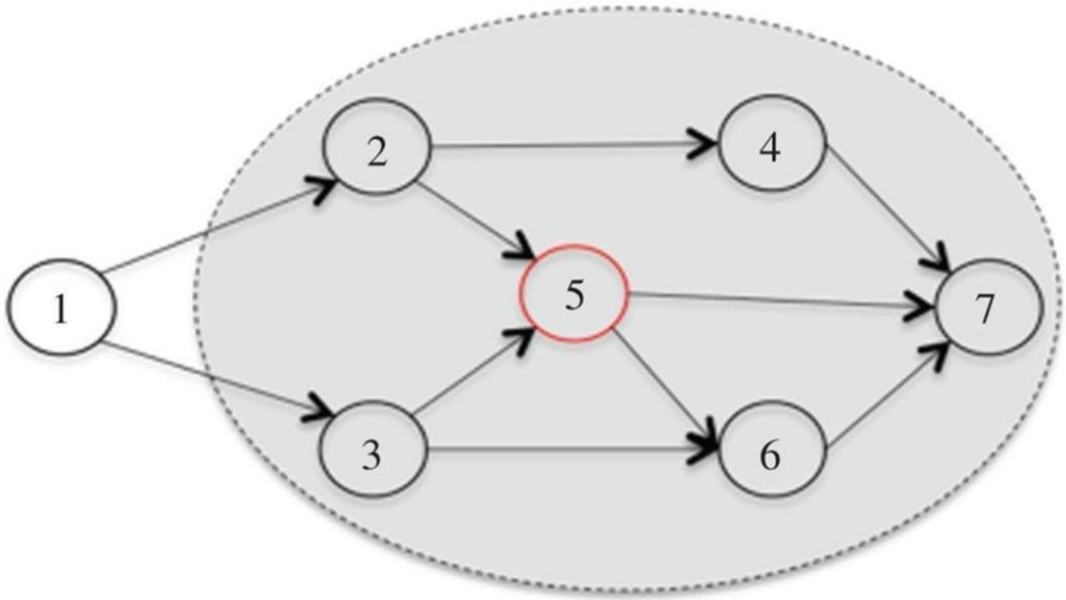

Fig. 1. Il grafo (nodi da 1 a 7) rappresenta una rete bayesiana, dove la variabile di riferimento X è il nodo 5 e le frecce (archi) che congiungono i vari nodi rappresentano i rapporti di causalità, o le relazioni informative, tra le diverse variabili. La MB del nodo 5 è indicata dall'insieme grigio, che racchiude i genitori di 5 (nodi 2 e 3), i figli di 5 (nodi 6 e 7) e le variabili che condividono un figlio con 5 (nodo 4).

D'altra parte, le MB sono descritte da Friston come uno strumento per esemplificare una forma specifica di indipendenza condizionale tra un sistema dinamico e il suo ambiente e sono considerate veri e propri confini dei sistemi viventi (Friston Mattout Kilner, 2011; Kirchhoff Parr Palacios, 2018). Infatti, esse sono associate ai confini fisici che circondano le cellule (cioè, la loro membrana), attraverso i quali vengono mediate

tutte le influenze tra gli spazi intracellulari ed extracellulari. Inoltre, le MB sono utilizzate per spiegare le reti neurali: le cellule piramidali superficiali e profonde della corteccia cerebrale, ad esempio, svolgono il ruolo di una MB, attraverso cui avvengono le interazioni tra le diverse colonne corticali. Questo può essere applicato anche su scala più ampia ai sistemi biologici, considerando i muscoli e i recettori sensoriali come MB di un organismo, e specifici organismi che agiscono come confini per separare gruppi di altri organismi. Ciò che accomuna questi diversi casi è che tutti rappresentano una separazione del mondo in due diversi insiemi di stati, che interagiscono solo attraverso la MB (Rubin Parr Da Costa, 2020). In questo contesto si verifica la cosiddetta indipendenza condizionale: se lo stato del confine organismo/ambiente, cioè la MB, è fisso, ciò che accade da un lato del confine non ha alcuna influenza su ciò che accade dall'altro lato. Ciò significa che tutte le informazioni necessarie per spiegare il comportamento del sistema interno sono date dallo stato della MB (Hohwy, 2019; Palacios Razi Parr, 2020).

Considerando quanto detto sopra, la natura delle MB può risultare non chiara. Infatti, esse sembrano avere due anime: modelli statistici da un lato, oggetti reali dall'altro. In particolare, le MB considerate come confini reali, come vengono descritte da Friston, sono state criticate, tra gli altri, da Jelle Bruineberg e colleghi (2021), che sottolineano come tale approccio costituisca "an example of reification fallacy: treating something abstract as something concrete" (Bruineberg Dolega Dewhurst, 2021, 19). Bruineberg sottolinea la differenza tra le "Pearl blankets" e le "Friston blankets". Le prime descrivono una proprietà dei modelli statistici, poiché catturano le relazioni di indipendenza condizionale tra le variabili del modello. Si calcola la probabilità che le indipendenze identificate in una determinata struttura causale siano valide. A partire da queste probabilità, è possibile formulare ipotesi specifiche di interesse, come ad esempio "*X* è causa di *Y*?", valutando tali ipotesi rispetto a tutte le possibili strutture causali compatibili con i dati disponibili (Spirtes Glymour Scheines, 2000). Le "Friston blankets", invece, sono una particolare interpretazione di questa proprietà per lo studio dei sistemi agente-ambiente. In altre parole, le "Pearl blankets" sono strumenti formali che vengono utilizzati per compiere inferenze su un sistema, utilizzando un modello di quel sistema, mentre le "Friston blankets" abbracciano un'interpretazione sensomotoria del modello e assumono che il sistema di interesse stia esso stesso compiendo inferenze (Bruineberg Dolega Dewhurst, 2021).

L’argomentazione di Bruineberg è stata contestata da Maxwell Ramstead (2022), che sembra non trovare alcuna differenza matematica significativa tra i due oggetti. Tuttavia, l’ambiguità ontologica del concetto rimane ancora argomento di discussione in letteratura. Nell’indagine sulle MB, occorre dunque fare molta attenzione alle assunzioni di base da considerare: da un lato, le MB possono essere utilizzate strumentalmente nella loro forma statistica originale, cioè come costrutto matematico formale per inferenze su un modello generativo, ad esempio una rete bayesiana. D’altra parte, questo utilizzo non può derivare dalle assunzioni forti su cui si basa il FEP come principio unificatore: infatti, le MB come costrutto ontologico che definisce i confini reali del mondo non possono essere giustificate solo con la matematica “tradizionale” di Pearl, e necessitano tecniche aggiuntive che in letteratura sembrano ancora mancare (Bruineberg Dolega Dewhurst, 2021). La consapevolezza degli aspetti critici delle MB, la cui definizione è ancora in corso e che si riflettono sul FEP, è essenziale nell’indagine sul rapporto tra i diversi campi della statistica, delle neuroscienze e delle scienze cognitive (Raffa, 2023).

Al di là di tali ambiguità, rimane notevole come FEP, AIF e MB offrano un quadro generale e onnicomprensivo per affrontare il problema della sopravvivenza a lungo termine e dell’uso ottimale delle risorse per qualsiasi entità e su qualsiasi scala temporale. Infatti, nel contesto della minimizzazione degli errori predittivi, si può considerare che gli agenti siano costantemente impegnati a ottimizzare le condizioni necessarie per evitare di entrare in stati di alta sorpresa, attraverso l’individuazione del giusto sottoinsieme di stati che consenta loro di mantenersi entro limiti di sopravvivenza accettabili. L’obiettivo dell’organismo è regolare le proprie interazioni con l’ambiente per rimanere all’interno di questi limiti, nonostante non disponga di conoscenza esplicita sulle proprie condizioni di vitalità. Come visto precedentemente, la sorpresa è qui intesa come una misura di incertezza, riferita a una distribuzione di probabilità non nota, poiché un organismo non ha modo di valutare direttamente la sorpresa. Secondo la definizione del FEP, l’energia libera è sempre maggiore o uguale alla sorpresa. Minimizzando l’energia libera, l’organismo riduce implicitamente anche la sorpresa e, di conseguenza, rimane entro limiti vitali accettabili (Bruineberg Kiverstein Rietveld, 2018). Si potrebbe dunque affermare che una strategia sostenibile per qualsiasi organismo – dove per “sostenibile” si intende mantenere il bilanciamento di tutte le componenti nel miglior modo possibile,

senza spreco di risorse – consiste nell'affidarsi ad attività predittive basate sull'esperienza passata, riducendo così la propensione ad azioni imprevedibili.

3.2 Inferenza attiva in IA

In quanto strategia sostenibile, la modellazione basata sul FEP e in particolare sull'AIF si rivela estremamente promettente e trova applicazione in numerosi ambiti.

Tra questi, spiccano la ricerca sull'introspezione, sull'auto-modellamento e sull'auto-accesso, con un contributo significativo allo sviluppo di una teoria della coscienza (Ramstead Albarracin Kiefer, 2023). L'introspezione, intesa come la capacità di accedere e valutare i propri stati mentali, pensieri ed esperienze, riveste un ruolo centrale nella consapevolezza di sé, nell'apprendimento e nel processo decisionale. Queste capacità rappresentano elementi fondamentali della coscienza umana, fornendo le basi per un'autovalutazione critica e una maggiore adattabilità agli stimoli ambientali (Limanowski Friston, 2018). Nello specifico, Mahault Albarracin e colleghi affermano che l'auto-modellamento e l'auto-accesso possono essere definiti come "interconnected processes that contribute to the development of self-awareness and to the capacity for introspection" (Albarracin Hipólito Tremblay, 2023, 7). Ciò significa che:

> self-modeling involves the creation of internal representations of oneself, while self-access refers to the ability to access and engage with these representations for self-improvement and learning […]. These processes, in conjunction with introspection, form a complex dynamic system that enriches our understanding of consciousness and the self – and indeed, may arguably form the causal basis of our capacity to understand ourselves and others (Albarracin Hipólito Tremblay, 2023, 7).

L'idea di partenza per un modello della coscienza basato sull'AIF è che, affinché un sistema sia in grado di riportare e valutare le proprie conclusioni, deve implementare una qualche forma di auto-accesso, vale a dire che alcune componenti del sistema possono utilizzare l'output di altre componenti come proprio input per ulteriori elaborazioni. Ciò è strettamente legato al concetto secondo cui alcuni processi cognitivi risultano trasparenti, ovvero permettono di accedere ad altri contenuti senza essere percepibili in sé. Al contrario, altri processi cognitivi sono opachi, nel senso che possono essere valutati direttamente, come accade nell'autoconsapevolezza introspettiva (ad esempio, essere

consapevoli di guardare un albero anziché semplicemente vedere un albero). I processi introspettivi, dunque, rendono accessibili al sistema altri processi cognitivi in quanto tali, conferendo loro opacità. L'AIF è stata utilizzata da Albarracin e colleghi per sviluppare un modello generativo a tre livelli, che descrive l'auto-accesso e le abilità introspettive in termini di processi che regolano la trasparenza e l'opacità a livello fenomenologico e la selezione attentiva a livello psicologico (Albarracin Hipólito Tremblay, 2023; Sandved-Smith Hesp Mattout, 2021).

Al di là della modellazione della coscienza, come trattato nel precedente paragrafo, il FEP e l'AIF offrono un quadro teorico generale per l'ottimizzazione delle risorse. Di conseguenza, essi possono essere applicati a quasi tutti i tipi di organismi e organizzazioni in grado di autosostenersi riducendo al minimo lo spreco di risorse, attraverso la strategia della minimizzazione degli errori di predizione. Come visto, questo concetto si basa sugli stessi principi dell'inferenza bayesiana, e le strutture bayesiane possono trovare impiego anche nell'ambito dell'IA generativa.

Nel Capitolo II è stato accennato come nell'apprendimento automatico molte delle applicazioni più recenti si basino su modelli neurali costruiti su architetture *feedforward*, allenate tramite *backpropagation*[41]. Questi modelli imparano a mappare un input di dimensione fissa (ad esempio, un'immagine) a un output anch'esso di dimensione fissa (come una probabilità associata a diverse categorie) (LeCun Bengio Hinton, 2015). Sebbene tali modelli siano estremamente potenti nella generazione di output rapidi, spesso si costituiscono in processi di *black box*, difficili da spiegare: l'input è noto, ma il percorso che conduce all'output rimane poco chiaro.

Per affrontare il problema della spiegabilità degli algoritmi, i modelli generativi risultano più affidabili (Derks de Waal, 2020). Essi sono essenzialmente reti bayesiane, che tracciano la propagazione degli eventi attraverso multipli nodi ambigui, in cui l'evento si dirama probabilisticamente lungo diversi percorsi. In ogni punto della rete, la probabilità che un nodo venga visitato dipende dalla probabilità congiunta dei nodi precedenti. Questo significa che i processi possono essere tracciati e compresi: a differenza delle reti neurali, infatti, le reti bayesiane consentono di seguire il

[41] Le architetture *feedforward* sono reti neurali in cui le informazioni fluiscono in una sola direzione, dagli strati di input verso gli strati di output, senza retroazioni o cicli. La *backpropagation*, invece, è un algoritmo di apprendimento supervisionato che consente di aggiornare i pesi della rete minimizzando l'errore tra l'output predetto e quello desiderato, attraverso la propagazione all'indietro del gradiente dell'errore.

ragionamento esplicito, pur mantenendo una natura probabilistica. Si è visto che l'idea alla base dei modelli generativi è la stessa della percezione predittiva basata sull'inferenza bayesiana, ovvero la minimizzazione degli errori predittivi: tali algoritmi, infatti, sono in grado di generare segnali sensoriali corrispondenti alle cause predette, poiché apprendono da piccole quantità di dati e generalizzano a nuove situazioni (Seth, 2020).

Tuttavia, considerare le reti bayesiane come modelli della cognizione su cui implementare progetti di IA potrebbe essere rischioso (Lieto, 2021). Questi modelli, infatti, si basano su una teoria decisionale che presuppone un decisore ottimale, capace di calcolare e scegliere, in ogni fase del *problem-solving*, l'azione che massimizza la funzione di utilità, ovvero la rappresentazione quantitativa delle proprie preferenze. Le teorie di massimizzazione dell'utilità attesa e, più in generale, i modelli bayesiani della cognizione ottimali, sono stati criticati per la loro intrattabilità computazionale: essi assumono, infatti, una conoscenza perfetta – hanno quindi scarsa realizzabilità pratica – e non forniscono prove empiriche sulla stabilità delle funzioni di utilità nel tempo (Friedman Isaac James, 2014; Gigerenzer, 2019).

Nonostante tali criticità, i modelli generativi stanno diventando sempre più popolari grazie ai loro vantaggi in termini di spiegabilità. Essi trovano applicazione per l'esecuzione di numerosi compiti, come la generazione di immagini, la predizione di testi, la modellazione di video e la previsione delle dinamiche di sistema (ad esempio, ambientali). Inoltre, vengono studiati per attività di controllo, esplorazione e rilevamento delle anomalie (Mazzaglia Verbelen Çatal, 2022).

Prima di considerare più nel dettaglio le applicazioni dell'AIF in IA, è utile esplorare le relazioni tra l'AIF e uno delle strutture più consolidate per l'apprendimento automatico: l'apprendimento per rinforzo (*reinforcement learning*, RL). Infatti, entrambi condividono l'obiettivo di ottimizzare il comportamento di un agente attraverso le interazioni con l'ambiente, ma si differenziano significativamente nelle loro assunzioni di base e nei metodi di implementazione. Un'analisi delle analogie e differenze tra AIF e RL è fondamentale per comprendere come questi due approcci possano convergere o divergere nella progettazione di IA, soprattutto in termini di gestione delle risorse e spiegabilità.

3.2.1 Inferenza attiva e apprendimento per rinforzo: analogie e differenze

Il RL si fonda sul principio di massimizzare una funzione di ricompensa cumulativa (*reward*). In questo approccio, un agente apprende tramite prove ed errori, esplorando un ambiente e identificando progressivamente le azioni che garantiscono il massimo guadagno in termini di ricompensa, anziché seguire istruzioni esplicite come accade in altri tipi di apprendimento (Sutton Barto, 2018).

L'integrazione del RL con il modello di inferenza bayesiana nel DM ha portato allo sviluppo del RL bayesiano. Questo approccio si rivela particolarmente efficace nei contesti in cui esistono limiti cognitivi, modellati attraverso un "collo di bottiglia" informativo. Tale vincolo implica che un agente debba compiere scelte ottimali con risorse informative limitate, avvicinandosi così al comportamento umano (Arumugam Ho Goodman, 2024).

Un ulteriore avanzamento è rappresentato dal RL profondo (*deep RL*), che integra il RL con reti neurali per consentire l'elaborazione di ambienti ad alta dimensionalità e la gestione di problemi complessi. Negli ultimi dieci anni, il RL profondo ha rappresentato una delle aree di ricerca più fertili nell'ambito dell'IA, raggiungendo prestazioni straordinarie in numerosi domini, tra cui i videogiochi, il poker, le competizioni con giocatori multipli e giochi da tavolo particolarmente complessi, come gli scacchi e il Go (Botvinick Ritter Wang, 2019), di cui si è accennato nel Capitolo II.

Il RL profondo riveste un interesse particolare per le neuroscienze e la psicologia. Infatti, i meccanismi che guidano l'apprendimento nel RL sono originariamente ispirati dalla ricerca sul condizionamento operante negli animali (Sutton Barto, 1981), e si ritiene che siano strettamente correlati ai meccanismi neurali di apprendimento basato sulla ricompensa, con un ruolo centrale della dopamina. La dopamina, infatti, è un neurotrasmettitore che segnala la discrepanza tra aspettativa e risultato, fondamentale per il rinforzo dell'apprendimento: quando un'azione porta a un esito migliore del previsto, la dopamina rinforza i circuiti neurali responsabili di quell'azione, promuovendo comportamenti futuri simili. Nel RL, l'errore di predizione della ricompensa svolge un ruolo analogo alla funzione della dopamina nel cervello, in quanto l'agente apprende aggiornando le proprie strategie di azione (*policy*) in risposta a ricompense ricevute, migliorando gradualmente le proprie decisioni. Una *policy* nel RL è generalmente determinata approssimando la mappatura stato-azione tramite una rete neurale,

consentendo al sistema di esplorare e ottimizzare sequenze di azioni in ambienti complessi. Inoltre, il RL profondo sfrutta le reti neurali per apprendere rappresentazioni che supportano la generalizzazione e il trasferimento, abilità chiave dei cervelli biologici (Schultz Dayan Montague, 1997).

É importante evidenziare che la prima ondata di ricerca nel RL profondo presentava alcune criticità significative, in particolare nel modo in cui questi sistemi apprendono, che risulta molto diverso dall'apprendimento umano. Una delle principali differenze riguarda l'efficienza campionaria, ossia la quantità di dati necessaria affinché un sistema di apprendimento raggiunga un certo livello di prestazioni. In questo senso, i primi sistemi di RL profondo si sono dimostrati notevolmente meno efficienti rispetto agli esseri umani in compiti come i videogiochi o gli scacchi, poiché richiedevano quantità di dati di addestramento superiori di molti ordini di grandezza rispetto a quelle necessarie a esperti umani per ottenere risultati comparabili (Tsividis Pouncy Xu, 2017). Di conseguenza, almeno nelle sue forme iniziali sviluppate a partire dal 2013, il RL profondo appariva troppo lento per rappresentare un modello plausibile dell'apprendimento umano (Botvinick Ritter Wang, 2019). Tuttavia, negli ultimi anni, importanti innovazioni hanno migliorato l'efficienza campionaria dei sistemi di RL profondo, attraverso lo sviluppo di nuovi metodi che riducono significativamente la necessità di enormi quantità di dati per l'addestramento, e accelerano il processo di apprendimento. L'introduzione di queste tecniche avanzate ha rilanciato l'RL profondo come un modello promettente per comprendere l'apprendimento umano, oltre a offrire nuove intuizioni per le scienze cognitive e le neuroscienze (Botvinick Ritter Wang, 2019).

Come l'AIF, dunque, il RL è un approccio orientato alla massimizzazione di un obiettivo attraverso azioni mirate. L'AIF e il RL sono due approcci differenti, ma simili nella modellazione dei processi decisionali adattivi. Infatti, entrambi possono essere formalizzati come processi decisionali markoviani parzialmente osservati (*partially observable Markov decision processes*, POMPD), in cui un agente deve scegliere le azioni ottimali in un ambiente incerto e dinamico, dove lo stato reale del sistema non è completamente osservabile[42]. Lo stato del sistema deve quindi essere inferito sulla base

[42] I POMPD si distinguono dai processi decisionali markoviani (*Markov decision processes*, MDP), dove invece l'agente conosce lo stato esatto del sistema (Lázaro-Gredilla Ku Murphy, 2024).

di osservazioni parziali e stocastiche (Kaelbling Littman Cassandra, 1998). Tuttavia, mentre il RL si concentra sulla massimizzazione della somma cumulativa delle ricompense, come visto l'AIF propone un quadro alternativo in cui le azioni sono selezionate per massimizzare l'evidenza di un modello generativo che riflette le preferenze dell'agente, mantenendo le osservazioni entro un intervallo predetto o desiderato (Tschantz Millidge Seth, 2020).

In generale, i due approcci si differenziano significativamente per i loro principi teorici, le modalità operative e le applicazioni tipiche. La prima differenza cruciale riguarda il bilanciamento tra esplorazione e sfruttamento. Nel RL, questo equilibrio viene implementato con meccanismi ad hoc, come *epsilon-greedy* o *softmax* con temperatura, che separano artificialmente l'acquisizione di informazioni (esplorazione) dalla massimizzazione della ricompensa (sfruttamento). Al contrario, nell'AIF tale equilibrio è intrinseco alla minimizzazione dell'energia libera attesa, che incorpora sia l'obiettivo di raccogliere informazioni riducendo l'incertezza (azioni epistemiche), sia quello di realizzare stati desiderati (azioni pragmatiche). La flessibilità dell'AIF emerge anche dalla sua capacità di specificare preferenze come distribuzioni probabilistiche a priori, offrendo un approccio più naturale e adattabile nella definizione degli obiettivi (Tschantz Millidge Seth, 2020, 2020).

Inoltre, l'AIF e il RL adottano strategie differenti in termini di rappresentazione del modello. Nell'AIF, l'agente dispone di un modello generativo che rappresenta sia l'ambiente esterno sia i propri stati interni. Questo modello genera osservazioni predette e consente all'agente di pianificare azioni che riducano l'incertezza predittiva, attraverso un aggiornamento bayesiano delle credenze. Nel RL, invece, l'approccio può essere *model-free* (senza un modello esplicito dell'ambiente) o *model-based* (con un modello che predice le conseguenze delle azioni) (Botvinick Ritter Wang, 2019). Il RL si focalizza principalmente sul valore delle azioni e degli stati, ottimizzando politiche che massimizzano la funzione di ricompensa attesa, ad esempio attraverso algoritmi come *Q-learning* o *policy gradient* (Sutton Barto, 2018).

Un'altra distinzione significativa emerge nel processo di pianificazione delle azioni. Come trattato precedentemente, nell'AIF le azioni sono selezionate al fine di ridurre l'incertezza predittiva, combinando in un unico processo l'inferenza probabilistica e la selezione delle azioni. Questo approccio non richiede necessariamente una funzione

di ricompensa esplicita, poiché l'obiettivo è mantenere l'evidenza del modello generativo allineata con le preferenze dell'agente. Al contrario, nel RL le azioni sono scelte per massimizzare direttamente la ricompensa cumulativa, con la pianificazione che avviene ottimizzando politiche in grado di garantire il miglior ritorno a lungo termine. Inoltre, si è visto che nel RL la mappatura stato-azione per orientare le politiche avviene attraverso una rete neurale, mentre nell'AIF le *policy* corrispondono a sequenze di azioni fissate (nell'approccio discreto) o a flussi continui di azioni calcolati dinamicamente per regolare il comportamento dell'agente in tempo reale (nell'approccio continuo) che vengono selezionati in base alla minimizzazione dell'energia libera.

Dal punto di vista delle applicazioni, infine, l'AIF è ampiamente utilizzata nelle neuroscienze e nelle scienze cognitive per modellare processi biologici, adattivi e integrati, come la percezione, il controllo motorio e l'interazione con l'ambiente (Friston Parr de Vries, 2017). Particolarmente rilevante è ad esempio il suo utilizzo nella robotica adattiva, dove è essenziale integrare percezione e azione in modo dinamico e resiliente (Sandved-Smith Hesp Mattout, 2021). Al contrario, l'RL è la struttura predominante nell'ambito dell'IA, con applicazioni che spaziano dai giochi e sistemi di controllo dei robot alla gestione delle risorse e ai sistemi di raccomandazione. Grazie alla sua capacità di ottimizzare politiche in ambienti complessi e dinamici, l'RL ha trovato ampio utilizzo in contesti industriali e operativi (Sutton Barton, 2018; Albrecht Christianos Schäfer, 2024).

Dal punto di vista teorico, comunque, si evidenzia una crescente convergenza tra i due approcci. L'adozione di metodi probabilistici, come il controllo basato sulla divergenza KL nel RL (Levine, 2018) e lo sviluppo del RL bayesiano, che bilancia esplorazione, sfruttamento e costi informativi, mostrano un avvicinamento concettuale verso l'AIF. Algoritmi come il *Blahut-Arimoto Thompson Sampling* (BLASTS) evidenziano il tentativo di modellare decisioni ottimali tenendo conto dei vincoli cognitivi e computazionali (Arumugam Ho Goodman, 2024).

Per riassumere, dunque, AIF e RL sono entrambi approcci orientati all'ottimizzazione di un obiettivo attraverso azioni mirate, ma differiscono per i principi sottostanti. L'RL si basa sulla massimizzazione della ricompensa e sull'ottimizzazione delle politiche, mentre l'AIF minimizza l'energia libera, integrando azione e inferenza bayesiana in un processo unificato che enfatizza la riduzione dell'incertezza. In ottica di

sostenibilità, mentre l'RL tende a focalizzarsi su obiettivi definiti, l'AIF offre un approccio più dinamico e adattativo, ottimizzando l'uso delle risorse e promuovendo soluzioni più flessibili.

Per quanto riguarda altri approcci che possono essere utilizzati per la gestione delle risorse, per certi versi simili all'AIF, è utile menzionare anche i cosiddetti algoritmi genetici (Holland, 1992), utilizzati per risolvere problemi complessi. Tale approccio simula la selezione naturale per individuare soluzioni ottimali nel tempo, esplorando lo spazio delle possibilità attraverso mutazioni e *crossover* – ovvero la combinazione di informazioni da due soluzioni "genitori" a due nuove soluzioni "figli". In particolare, la selezione implica che solo gli individui con una migliore adattabilità sopravvivano e si riproducano. La riproduzione con *crossover* e mutazione, inoltre, prevede che gli algoritmi combinino porzioni di soluzioni già esistenti per generarne nuove, introducendo occasionalmente mutazioni casuali per esplorare nuove possibilità. Sostanzialmente, l'idea innovativa di Holland fu che, invece di progettare un algoritmo per risolvere un problema complesso, è possibile lasciare che esso "evolva", introducendo in modo iterativo soluzioni via via migliori, attraverso generazioni successive, che si autovalutano in base alla loro adattabilità, ovvero a quanto risolvono bene il problema *target* (Holland, 1992).

Gli algoritmi genetici sono stati utilizzati con successo in vari ambiti di applicazione, ad esempio per l'ottimizzazione di turbine a getto – riducendo il tempo di sviluppo rispetto a metodi tradizionali – oppure in modellazioni di mercati finanziari, o ancora in giochi strategici. La forza di tale tipo di algoritmi risiede nel fatto che essi hanno la capacità di esplorare varie soluzioni, analizzando e confrontando contemporaneamente più possibilità, e presentano robustezza e flessibilità. D'altra parte, essi presentano anche costi computazionali molto elevati e occorre prestare molta attenzione nella fase della definizione dei parametri, che possono influenzare significativamente le prestazioni dell'algoritmo.

Tuttavia, facendo un confronto con l'AIF, emerge come l'ottimizzazione evolutiva sia meno adattiva su orizzonti temporali brevi, in quanto si basa su miglioramenti graduali delle soluzioni attraverso generazioni progressive, e dunque è meno efficace in ambienti dinamici. Inoltre, gli algoritmi genetici operano su un meccanismo *trial-and-error* – in cui gli errori vengono eliminati e le strategie efficaci rinforzate – senza una rappresentazione esplicita del mondo o della dinamica

ambientale, e ciò non sempre garantisce una risposta rapida ai cambiamenti ambientali. In generale, nonostante anche gli algoritmi basati su AIF siano computazionalmente dispendiosi, gli algoritmi genetici lo sono maggiormente, in quanto caratterizzati da molteplici generazioni di tentativi.

Considerato tutto quanto detto finora, prima di affrontare più nel dettaglio le applicazioni dell'AIF, occorre chiarire la distinzione tra AIF discreta e continua. Tale differenza, infatti, non solo influenza il modo in cui l'AIF viene implementata nei modelli computazionali, ma ne determina anche le applicazioni pratiche in ambito di robotica, neuroscienze e sistemi dinamici complessi.

3.2.2 Applicazioni

Delineare la distinzione tra AIF discreta e continua è cruciale per comprenderne l'applicazione sia in contesti cognitivi sia computazionali. Tale differenza emerge dalla natura dei processi modellizzati, ovvero la selezione tra azioni o politiche discrete e il controllo continuo di variabili dinamiche (Friston Parr de Vries, 2017).

L'AIF discreta è utilizzata per processi decisionali categoriali, come la scelta tra un insieme di alternative possibili. É questo il contesto delineato precedentemente, in cui un agente rappresenta il mondo attraverso variabili non note discrete, come stati e politiche candidate, e ottimizza le proprie azioni selezionando quelle che minimizzano l'energia libera attesa. Tale processo può essere formalizzato come un MDP, in cui le politiche corrispondono a sequenze di azioni che minimizzano l'incertezza predittiva (Friston Parr de Vries, 2017). Questa formulazione è particolarmente utile per modellare comportamenti decisionali e processi cognitivi a livello alto, dove le scelte si sviluppano in domini discreti (Friston Parr de Vries, 2017). Un esempio in biologia è il sistema oculomotorio, ovvero l'indirizzamento dello sguardo scegliendo tra una serie di possibili localizzazioni visive. Alcuni studi di Thomas Parr e Friston (2018) mostrano come nell'encefalo, il collicolo superiore agisca come ponte tra il DM discreto e il controllo motorio continuo, suggerendo un'integrazione naturale tra questi due livelli di rappresentazione. Un'altra applicazione è nel contesto dell'apprendimento di nuove competenze e dell'innovazione strumentale, poiché l'AIF discreta permette agli agenti di sfruttare modelli generativi per esplorare e selezionare strumenti o azioni appropriate in contesti incerti. In questo modo, l'agente può adattarsi dinamicamente e sviluppare

comportamenti adatti alla risoluzione dei problemi, imparando a manipolare l'ambiente e ad anticipare le conseguenze delle proprie azioni (Collis Kinghorn Buckley, 2023).

L'AIF continua, d'altra parte, è impiegata per l'ottimizzazione di variabili dinamiche che evolvono nel tempo, come il controllo motorio e la percezione fluida dell'ambiente. In questo contesto, un agente utilizza un modello generativo continuo per minimizzare l'errore predittivo sensoriale in modo dinamico. Il controllo continuo delle variabili è spesso formalizzato attraverso coordinate generalizzate del moto, che descrivono l'evoluzione temporale delle variabili osservabili e non osservabili (Friston Parr de Vries, 2017). L'approccio continuo è fondamentale nei contesti in cui è richiesto un adattamento preciso e robusto, nonché la capacità di ottimizzare le risposte del sistema in modo accurato e di mantenere la funzionalità nonostante perturbazioni o variazioni impreviste dell'ambiente, come nella robotica adattiva e nel controllo motorio biologico (Priorelli Stoianov Pezzulo, 2024). Ad esempio, un agente robotico può correggere continuamente i propri movimenti in risposta a *feedback* sensoriali in tempo reale, mantenendo la coerenza tra percezione e azione (Pezzulo Donnarumma Iodice, 2017). L'AIF continua è quindi particolarmente efficace per modellare dinamiche sensomotorie in sistemi biologici e artificiali.

Una differenza chiave tra AIF discreta e continua riguarda poi la complessità computazionale dell'informazione. In contesti continui, il guadagno informativo è generalmente difficile da caratterizzare a causa dell'intrattabilità delle misure di teoria dell'informazione in spazi ad alta dimensionalità. Al contrario, negli spazi discreti, misure come il guadagno informativo ammettono spesso soluzioni chiuse, che dunque non necessitano approssimazioni e sono computazionalmente più accessibili. Inoltre, la rappresentazione discreta facilita l'implementazione diretta di tecniche classiche di teoria delle decisioni, come la programmazione dinamica. Questo conferisce un vantaggio significativo nei sistemi di DM, specialmente quando problemi complessi vengono ridotti a sottoproblemi gestibili attraverso la specifica di sotto-obiettivi, come avviene nel comportamento umano (Collis Singh Kinghorn, 2024).

Al contempo, AIF discreta e continua sono integrabili: i modelli discreti, infatti, possono essere utilizzati per vincolare il livello continuo, ad esempio specificando politiche discrete che guidano l'evoluzione dinamica delle variabili sensomotorie. Questo meccanismo consente a un agente di selezionare azioni categoriali (ad esempio

l'esplorazione di una nuova localizzazione visiva) e implementarle attraverso un controllo motorio continuo (ovvero i movimenti oculari). L'integrazione tra i due livelli è particolarmente evidente nei sistemi biologici, dove decisioni discrete vengono tradotte in dinamiche continue attraverso strutture neurali specifiche, come appunto nel sistema oculomotorio (Parr Friston, 2018).

Come si è visto, si riscontrano esempi di AIF discreta nella modellazione del DM e della selezione delle politiche in ambienti dinamici, e di AIF continua nel controllo sensomotorio adattivo e nella robotica. Nello specifico per la robotica, i modelli basati sull'AIF risultano particolarmente efficaci quando la dinamica del robot o del compito è incerta: per l'adattamento dinamico, il controllo adattivo, la tolleranza ai guasti, la pianificazione prospettica e lo sviluppo di competenze cognitive complesse, come la collaborazione uomo-robot e la discriminazione tra sé e gli altri (Lanillos Meo Pezzato, 2021). Pablo Lanillos e Gordon Cheng (2018) hanno sviluppato un modello computazionale basato su PP e AIF che permette a un robot di inferire e aggiornare la propria configurazione corporea. In questo modello, il PP viene utilizzato per implementare una percezione computazionale multisensoriale, che integra informazioni tattili, visive e propriocettive. Grazie a questa capacità, il robot è in grado di stimare e adattare la propria configurazione corporea dinamicamente, correggendo continuamente i movimenti sulla base del *feedback* sensoriale. Precedentemente, anche Jun Tani (2003) e Leo Pio-Lopez e colleghi (2016) avevano implementato modelli basati su approcci simili, utilizzati anche per il controllo di manipolatori industriali (Pezzato Baioumy Corbato, 2020) e per implementare la visione attiva nei robot in ambienti simulati (Van de Maele Verbelen Çatal, 2021). Nel contesto della visione, infatti, Toon Van de Maele e colleghi hanno dimostrato in un ambiente simulato che un robot, muovendosi senza conoscenza preliminare dell'ambiente, è in grado di selezionare la posizione visiva successiva minimizzando l'energia libera attesa. Questo risultato evidenzia come l'AIF possa costituire una soluzione naturale per compiti di visione complessi, nei quali la distribuzione delle informazioni ambientali non è nota a priori. Inoltre, con una capacità computazionale adeguata, come quella fornita da reti neurali generative e GPU ad alte prestazioni, questo approccio può essere esteso al controllo in tempo reale di manipolatori robotici fisici (Van de Maele Verbelen Çatal, 2021).

Sulla medesima scia, per quanto riguarda il DM e la pianificazione di azioni in scenari che richiedono l'anticipazione delle intenzioni altrui, robot basati su AIF sono in grado di predire le intenzioni degli altri agenti per anticiparne i movimenti, migliorando così la sicurezza e l'efficacia dell'interazione. Tale aspetto è cruciale, ad esempio, per i *Socially Assistive Robots* (SARs), come robot assistenti personali, infermieri robotici e robot da compagnia, progettati per l'assistenza a persone anziane o disabili (Alfieri Raffa, 2025). In questi contesti, la selezione delle azioni avviene attraverso la minimizzazione dell'energia libera attesa, consentendo al robot di ridurre rischi e ambiguità e ottimizzare la raccolta di informazioni. Questo comportamento risulta particolarmente vantaggioso in situazioni ad alto rischio e incertezza, tipiche dell'interazione uomo-robot (Da Costa Lanillos Saijd, 2022).

In sintesi, l'AIF continua e discreta costituiscono due aspetti complementari della stessa struttura inferenziale. L'AIF continua si dimostra ideale per la gestione dinamica dei movimenti sensomotori attraverso un'integrazione efficiente di *feedback* multisensoriali in tempo reale. L'AIF discreta, invece, è particolarmente adatta a scenari di pianificazione e DM, dove è essenziale l'anticipazione delle intenzioni e la riduzione dell'incertezza. Entrambi gli approcci rappresentano strumenti avanzati per migliorare le capacità adattive e cognitive dei robot, con applicazioni che spaziano dalla robotica industriale a quella assistiva, integrando sicurezza, flessibilità e adattamento continuo alle esigenze ambientali. L'interazione tra i due livelli offre dunque un quadro teorico potente e flessibile per comprendere e implementare comportamenti adattivi, sia biologici sia artificiali.

Finora sono stati esaminati i concetti di CA orientate alla sostenibilità e di FEP e AIF come esempi di modelli cognitivi applicabili all'IA. É ora, dunque, opportuno iniziare a tirare le fila della ricerca esposta in questa tesi, argomentando in che modo l'AIF può configurarsi come una CA sostenibile, capace di coniugare efficienza, adattabilità e ottimizzazione delle risorse.

3.3 Inferenza attiva come architettura cognitiva sostenibile

Nel §2.1 si è visto come il cognitivismo classico sia stato criticato, soprattutto per via del carattere disincarnato e rappresentazionalista della cognizione. Secondo tale approccio, infatti, la mente elabora simboli e rappresentazioni astratte del mondo in maniera isolata

e indipendente dall'ambiente fisico (Newell Simon, 1976). Tuttavia, approcci più recenti come l'*embodied cognition* e l'enattivismo hanno sottolineato che la cognizione non può essere separata dal corpo e dalle interazioni con l'ambiente (Varela Thompson Rosch, 1991; Gallagher, 2005). É in questo contesto che l'AIF si distingue, costituendo un modello integrato e dinamico che unisce percezione, azione e apprendimento.

Si è visto come nell'AIF, la cognizione non sia intesa come un mero processo simbolico interno, ma in quanto dinamica predittiva basata sulla minimizzazione dell'energia libera e dell'errore di predizione. L'agente non è un sistema passivo che rappresenta il mondo, bensì un soggetto attivo che interagisce con l'ambiente per ridurre l'incertezza sulle proprie percezioni. In questa prospettiva, l'azione diventa parte integrante del processo cognitivo, permettendo all'agente di testare e aggiornare costantemente le proprie aspettative interne.

Confrontando l'AIF con le critiche mosse all'*embodiment* e al corporeismo (Manzotti Chella, 2020), esposte nel §2.1, emerge come il processo riconosca il ruolo cruciale dell'*embodiment* e dell'ambiente, ma non riduca la cognizione a un prodotto dell'interazione fisica tra corpo e contesto. Al contrario, il modello generativo interno su cui si basa l'AIF, permette a un agente di anticipare e inferire le proprie interazioni con l'ambiente, integrando percezione, azione e apprendimento in modo unificato. Questo modello non esclude il ruolo della rappresentazione interna, né si limita a processi puramente reattivi o corporei. Come si è visto, l'AIF consente a un agente di simulare scenari futuri e aggiornare le proprie credenze attraverso un processo di inferenza bayesiana. Tali simulazioni non richiedono una presenza fisica nel mondo reale, ma sono guidate dal modello interno dell'agente (Hesp Tschantz Millidge, 2020), e ciò contrasta con la rigidità del corporeismo, poiché dimostra che la cognizione può emergere anche indipendentemente da un'interazione immediata con il mondo esterno. Le rappresentazioni interne, inoltre, non sono statiche o predefinite, come nel cognitivismo classico, ma funzionano come modelli generativi dinamici che vengono continuamente aggiornati attraverso l'interazione con l'ambiente.

Alla luce della capacità dell'AIF di superare le limitazioni del corporeismo, integrando l'interazione con l'ambiente e simulazioni interne guidate da modelli generativi, è possibile esaminare come tale struttura risponda ai criteri delineati da Newell (1980) per la valutazione delle CA, che sono stati affrontati in §2.1. Il primo criterio è il

comportamento flessibile, che è proprio una delle caratteristiche centrali dell'AIF. Infatti, come è stato visto, un agente, grazie al proprio modello generativo, è in grado di simulare scenari futuri e aggiornare dinamicamente le proprie credenze attraverso l'inferenza bayesiana. Ciò consente di affrontare una vasta gamma di compiti cognitivi, anche senza addestramento specifico per essi, selezionando le azioni che minimizzano l'incertezza predittiva. Tale capacità rende l'AIF particolarmente adatta alla modellazione di comportamenti adattivi sia in contesti biologici sia artificiali. Anche il secondo criterio di Newell, ovvero la performance in tempo reale, si può considerare soddisfatto dall'AIF, in particolare nella sua formulazione continua. Come si è visto, infatti, l'AIF continua è ideale per modellare il controllo motorio e l'integrazione multisensoriale in tempo reale. Dal punto di vista del comportamento adattivo, inoltre – terzo tra i criteri di Newell – l'AIF offre un quadro teorico che enfatizza la capacità degli agenti di adattarsi a un ambiente dinamico e incerto. La minimizzazione dell'energia libera consente di bilanciare in modo efficiente azioni esplorative, necessarie per ridurre l'incertezza, e azioni pragmatiche, mirate al raggiungimento di obiettivi specifici. Questa caratteristica è fondamentale per modellare sistemi capaci di evolvere – quarto criterio – e rispondere alle sfide poste da ambienti complessi e mutevoli.

Il modello generativo dell'AIF permette anche di soddisfare i criteri di avere una vasta conoscenza di base e un comportamento dinamico, in quanto integra un'ampia gamma di conoscenze probabilistiche, rappresentando sia l'ambiente esterno sia lo stato interno dell'agente. Tale integrazione consente agli agenti di affrontare l'incertezza e di adattarsi dinamicamente a un ambiente in evoluzione, prevedendo, pianificando e rispondendo a cambiamenti improvvisi. Inoltre, l'AIF fornisce un quadro probabilistico che consente l'integrazione di processi inferenziali eterogenei, come induzione, deduzione e abduzione, tipici della cognizione umana, dimostrando quindi capacità di integrare diversi tipi di conoscenza. Tramite essa, infatti, l'agente ha una rappresentazione sia dell'ambiente esterno sia del proprio stato interno.

Per quanto riguarda i criteri di linguaggio e apprendimento, l'AIF supporta modelli adattivi di apprendimento che si sviluppano nel tempo. Ad esempio, l'integrazione con modelli gerarchici predittivi consente di gestire processi complessi come la comprensione del linguaggio e l'apprendimento basato su esperienze pregresse. Sistemi avanzati di AIF sono stati utilizzati per simulare apprendimento graduale e sviluppo cognitivo in ambienti

scolastici dinamici, favorendo la crescita di competenze (Di Paolo White Guénin-Carlut, 2024).

L'AIF soddisfa anche il criterio di sviluppo ed evoluzione, in quanto modella la cognizione come un processo dinamico e adattivo. Le capacità dell'agente si sviluppano e si raffinano nel tempo attraverso l'interazione con l'ambiente, riflettendo in questo modo i principi evolutivi che hanno plasmato la cognizione umana. Inoltre, l'allineamento dell'AIF con i principi biologici e neuroscientifici risponde al criterio di mappatura sulle strutture cerebrali. Come visto, infatti, l'AIF trova corrispondenze dirette nei meccanismi neurali di predizione e controllo motorio osservati nel cervello umano, come quelli implementati nella corteccia sensoriale, nel collicolo superiore e in altre strutture deputate alla minimizzazione degli errori predittivi (Friston Parr de Vries, 2017). La connessione con le neuroscienze, dunque, conferma l'AIF come un modello biologicamente plausibile e neurocompatibile.

Infine, l'AIF fornisce un quadro promettente per affrontare la questione della coscienza, l'ultimo criterio suggerito da Newell, come suggerito dagli studi di Albarracin e colleghi (2023) sopracitati. La capacità dell'agente di monitorare e aggiornare il proprio stato interno, oltre che di simulare scenari futuri, infatti, suggerisce una forma di consapevolezza operativa che può essere modellata in sistemi artificiali.

Dunque, l'AIF soddisfa i criteri delineati da Newell per le CA. Inoltre, in virtù della sua flessibilità, adattività e coerenza con i principi biologici, si configura come una CA promettente per lo sviluppo di sistemi artificiali sostenibili e capaci di replicare in modo realistico le funzioni cognitive umane.

Thomas van Es e Inês Hipólito (2020) sottolineano come il FEP venga interpretato sia come una teoria che descrive processi rappresentazionali nei sistemi biologici, sia come uno strumento euristico utile a modellare le dinamiche adattive tra organismo e ambiente. All'interno del dibattito filosofico tra realismo e strumentalismo, il realismo rappresentazionalista sostiene che, come visto, i modelli generativi utilizzati nel FEP abbiano luogo all'interno del cervello, conferendo al sistema la capacità di creare rappresentazioni interne del mondo esterno. Tuttavia, van Es e Hipólito (2020) criticano tale posizione e sottolineano l'assenza di una giustificazione chiara su come le rappresentazioni possano emergere naturalmente nei sistemi biologici. Sostengono invece una visione strumentalista del FEP, secondo la quale esso è un potente strumento

descrittivo per comprendere le dinamiche di auto-organizzazione e adattamento, senza che occorra compiere assunzioni ontologiche sui modelli stessi. In questo senso, il FEP funziona come un quadro teorico euristico che permette di spiegare la riduzione dell'entropia e della sorpresa nei sistemi viventi, come neuroni e organismi complessi, mantenendo la coerenza tra l'approccio scientifico e l'interpretazione filosofica. Questa prospettiva consente di evitare ambiguità concettuali e pone il FEP come una metodologia formale utile a descrivere sistemi adattivi, piuttosto che come una teoria ontologica dei processi cognitivi e biologici. Il FEP, nella sua interpretazione strumentalista, dunque, può essere considerato un quadro teorico euristico utile per descrivere le dinamiche adattive di organismi viventi, senza assumere un realismo ontologico delle rappresentazioni. Se nel realismo rappresentazionalista sembra mancare una chiara giustificazione su come le rappresentazioni possano emergere naturalmente nei sistemi biologici, l'AIF, invece, si configura come un modello in cui le rappresentazioni interne non sono statiche o predefinite, come nel cognitivismo classico, ma dinamiche e adattive, aggiornate costantemente tramite l'interazione con l'ambiente.

L'approccio strumentalista applicato al FEP offre una solida base per comprendere anche il ruolo dell'AIF nella definizione della sostenibilità. Come anticipato nel §2.3, l'AIF, oltre a poter essere considerata una CA che integra percezione, azione e apprendimento per affrontare l'incertezza e adattarsi dinamicamente all'ambiente, si configura anche come un modello sostenibile. In particolare, essa può essere utilizzato per descrivere e promuovere la resilienza, il benessere e la sostenibilità nei sistemi complessi, bilanciando l'ottimizzazione delle risorse con la capacità di adattamento su scale temporali e spaziali estese. É proprio in tale direzione che si muove il lavoro di Albarracin e colleghi (2024), che punta a formalizzare il concetto di sostenibilità attraverso l'AIF. Essa, infatti, non si limita a descrivere il processo di minimizzazione dell'energia libera, ma estende tale logica ai sistemi complessi, all'interno dei quali concetti come resilienza, benessere e autosufficienza sistemica assumono un ruolo chiave (Albarracin Ramstead Pitliya, 2024).

Definire il concetto di sostenibilità attraverso il modello dell'AIF implica considerare "sostenibile" un sistema che sia in grado di mantenere uno stato coeso e stabile nel tempo, soddisfacendo i propri bisogni senza esaurire le risorse essenziali. Questa visione va oltre la sola gestione delle risorse materiali, e include anche reti sociali,

lavoro e conoscenza come fattori chiave per il funzionamento del sistema. Il raggiungimento di uno stato che possa essere considerato sostenibile richiede l'istituzione di un sistema olistico capace di affrontare le principali barriere alla sostenibilità e di mettere in atto pratiche personalizzate nel campo dell'economia, dell'ambiente e della società – ovvero, come visto nel Capitolo I, i tre ambiti che sono considerati pilastri della sostenibilità. Tale approccio, secondo Albarracin e colleghi (2024), include l'adozione di pratiche che promuovono auto-organizzazione e resilienza. In particolare, nell'ambito dell'AIF, la resilienza può essere declinata in tre componenti fondamentali: inerzia (la capacità di resistere ai cambiamenti mantenendo lo stato corrente), plasticità (l'adattabilità strutturale a nuovi contesti e la capacità di apprendere nuovi modelli di comportamento) ed elasticità (il ritorno a stati preferiti dopo perturbazioni) (Miller Albarracin Pitliya, 2022). Tuttavia, la resilienza, se considerata da sola, è insufficiente per garantire la sostenibilità all'interno di un sistema. Essa deve infatti manifestarsi su scale temporali estese e lungo tutti i livelli della gerarchia che costituisce il sistema stesso, coinvolgendo tutte le interdipendenze tra i diversi elementi.

La resilienza è una componente fondamentale del benessere. Attraverso la lente dell'AIF, il benessere può essere definito come la capacità di un sistema di adattare il proprio modello interno all'ambiente in modo tale da minimizzare l'errore presente e prevenirlo in futuro. In questo senso, le misurazioni del benessere possono fungere da indicatori della sostenibilità delle diverse componenti che costituiscono il sistema. Si può considerare il benessere di un sistema nei termini della sua capacità di compiere azioni a diversi livelli, che assicurino la sostenibilità dell'intero sistema. Il benessere non riguarda esclusivamente la stabilità, ma anche il mantenimento di un equilibrio che permetta adattamento ed evoluzione. Infatti, come visto, ogni individuo o entità possiede un modello generativo delle cause non note dei dati sensoriali, il quale è orientato verso risultati favorevoli. L'energia libera corrisponde all'evidenza dei dati secondo questo modello generativo: se l'energia libera è bassa, l'evidenza è alta e i dati sono altamente conformi alle preferenze, e viceversa.

Attraverso il FEP, le pratiche che favoriscono il rinnovamento delle risorse, gli scambi armoniosi tra un sistema e le sue componenti e l'adozione di processi di auto-organizzazione e resilienza assumono dunque un ruolo cruciale come vie privilegiate per il raggiungimento della sostenibilità.

Alla luce di queste considerazioni, si può considerare la sostenibilità definita tramite il modello dell'AIF come la capacità di un sistema di raggiungere e mantenere uno stato coeso e stabile attraverso diverse scale spaziali e temporali. In questo quadro, la resilienza delle singole componenti della gerarchia promuove la resilienza complessiva dell'intero sistema. Per raggiungere tale equilibrio, è necessario promuovere il rinnovamento delle risorse e il loro utilizzo armonioso, l'auto-organizzazione dinamica del sistema e le interconnessioni del sistema stesso, in quanto esse rafforzano la stabilità del sistema (Albarracin Ramstead Pitliya, 2024).

Per modellare e quantificare le interdipendenze tra le risorse e il loro impatto sulla sostenibilità complessiva del sistema, Albarracin e colleghi (2024) suggeriscono di ricorrere all'applicazione della teoria delle reti e della teoria dei sistemi dinamici. Servendosi della teoria delle reti, infatti, è possibile rappresentare le risorse come nodi interconnessi in un sistema dinamico: la rimozione di un nodo rappresenta l'esaurimento di una risorsa, simulando l'impatto sull'intero sistema. La teoria dei sistemi dinamici consente poi di rappresentare e prevedere il comportamento di un sistema, tenendo conto di certi parametri chiave, come risorse disponibili e interazioni. Tali strumenti permettono di descrivere formalmente la complessità dei sistemi adattivi, evidenziando come la gestione delle risorse e le dinamiche interne possano sostenere la stabilità e l'evoluzione del sistema nel lungo periodo (Albarracin Ramstead Pitliya, 2024).

Se si considera un sistema dinamico composto da più agenti che interagiscono tra di loro, ciascun agente raggiunge la sostenibilità minimizzando l'energia libera attesa, bilanciando la riduzione del rischio – attraverso il soddisfacimento delle preferenze – con l'ambiguità – attraverso il miglioramento del proprio modello interno grazie alle nuove informazioni che trae dall'interazione con l'ambiente. In tale direzione di promozione della sostenibilità, bisogna adottare politiche che ottimizzino la libertà energetica di tutti gli agenti all'interno del sistema e che favoriscano interazioni simbiotiche e processi di *feedback* che promuovano resilienza e adattabilità.

D'altra parte, la minimizzazione dell'energia libera non è un procedimento sempre lineare o immediato. Per comprenderlo, è necessario considerare gli obiettivi e i vincoli del sistema. In alcuni casi, tali obiettivi possono portare a comportamenti che non si allineano perfettamente con la minimizzazione immediata dell'energia libera. A causa di vincoli nella loro struttura o funzione, infatti, alcuni sistemi non sono in grado di

minimizzare efficacemente l'energia libera, e questo li rende parzialmente non sostenibili. Ad esempio, alcune condizioni patologiche possono compromettere questa capacità, portando a comportamenti maladattivi o a stati che si discostano significativamente da ciò che il FEP predirebbe. Un esempio è rappresentato da alcuni disturbi neurologici, come la schizofrenia, in cui la capacità del cervello di minimizzare l'energia libera è compromessa. In tali condizioni, si verificano alterazioni della connettività cerebrale che possono distorcere la percezione e il pensiero, rendendoli meno correlati al mondo esterno. Una persona con schizofrenia può, quindi, avere difficoltà a ridurre l'incertezza rispetto al proprio ambiente, sviluppando comportamenti maladattivi e stati che non ottimizzano il funzionamento del sistema (Friston Brown Siemerkus, 2016).

Al netto dei vincoli di ciascun sistema particolare, per essere considerato sostenibile, un sistema deve inoltre mantenere un equilibrio tra ridondanza e complessità. Con ridondanza si intende la presenza di risorse, elementi o funzioni duplicati all'interno di un sistema, che possono svolgere ruoli simili o identici. La ridondanza agisce come margine di sicurezza: in caso di perturbazione o fallimento di una parte del sistema, le componenti ridondanti possono compensare e garantire il funzionamento continuativo del sistema. Ad esempio, in un ecosistema naturale la ridondanza si osserva quando più specie svolgono ruoli simili; in un procedimento tecnico, un processo di *backup* ridondante garantisce il funzionamento complessivo se una risorsa primaria viene meno. La ridondanza è quindi fondamentale per la resilienza del sistema, ma se eccessiva può portare a inefficienze, sprechi o sovraccarico. La complessità, d'altra parte, descrive il grado di interconnessione, varietà e interdipendenza tra le componenti del sistema stesso. I sistemi complessi sono adattivi, in grado di autoregolarsi e rispondere dinamicamente a cambiamenti o perturbazioni. Tuttavia, un alto livello di complessità può anche rendere il sistema fragile, poiché piccole perturbazioni possono propagarsi rapidamente. Un ecosistema, ad esempio, è un sistema complesso in cui le specie interagiscono in una rete intricata di relazioni.

La ridondanza, quindi, fornisce stabilità, fungendo da "rete di sicurezza" contro fallimenti o perturbazioni, mentre la complessità è garante di adattabilità e capacità di risposta dinamica alle sfide ambientali o interne. Il mantenimento dell'equilibrio tra ridondanza e complessità, dunque, deve avvenire attraverso la promozione dell'abbondanza, evitando al contempo l'eccessiva complessità, che può comportare il

degrado delle risorse ed eccessiva instabilità. In tale contesto, sostenibilità, resilienza e benessere costituiscono proprietà dei sistemi complessi, e l'utilizzo dell'AIF consente di ottimizzare l'uso delle risorse e promuovere politiche sostenibili basate su dinamiche di auto-organizzazione e interdipendenze tra agenti e risorse.

Per esplicitare i meccanismi esposti finora, Albarracin e colleghi (2024) propongono l'esempio di un individuo (un agente) che vive in una foresta. Si può facilmente immaginare come tale individuo osservi il mondo attorno a sè e inferisca come funzioni la foresta. Dunque, l'individuo costruisce una rappresentazione del mondo, che viene usata per orientarsi nell'ambiente e acquisire da esso ciò che è necessario per la sopravvivenza. Il benessere dell'individuo può essere misurato sulla base di quanto il suo modello interno si allinea allo stato reale della foresta. Se l'individuo impara a predire in modo accurato e ad adattarsi ai cambiamenti circostanti, come il trascorrere delle stagioni o le variazioni nella disponibilità delle risorse, sarà in grado di sperimentare un livello maggiore di benessere. Tale aspetto può essere quantificato usando l'informazione mutuale: nello specifico, l'informazione mutuale tra il modello interno dell'individuo e lo stato attuale della foresta rappresenta l'abilità nel predire e adattarsi all'ambiente. Un'informazione mutuale tra il modello interno dell'individuo e lo stato reale della foresta indica un modello interno ben calibrato e, conseguentemente, benessere maggiore. L'individuo, dunque, si sforza di mantenere un modello accurato e parsimonioso della foresta minimizzando la complessità e al contempo assicurandosi che il modello sia conforme alle osservazioni, quindi massimizzando l'accuratezza. Una minore energia libera corrisponde a un livello più alto di benessere, poiché indica che l'individuo sta navigando con successo nel proprio ambiente e sta minimizzando lo stress. L'energia libera attesa, inoltre, estende questo concetto agli stati futuri, consentendo all'individuo di anticipare il proprio benessere futuro stimando la divergenza tra il proprio modello interno e le osservazioni future attese. É proprio minimizzando l'energia libera attesa che l'individuo bilancia la spinta a soddisfare le proprie preferenze (ad esempio, trovare cibo e riparo) con la necessità di ridurre l'incertezza sull'ambiente circostante (ad esempio, esplorare nuove aree della foresta) (Albarracin Ramstead Pitliya, 2024).

Dunque, i concetti di benessere, resilienza e sostenibilità sono interconnessi e dipendono dalla capacità di un sistema di evitare la dissipazione e prosperare mantenendo un certo livello di stabilità. Per raggiungere la sostenibilità, è necessario favorire ambienti

complessi e adattabili, ottimizzare la ridondanza, ridurre l'impatto di eventi estremi e coordinare sottogruppi all'interno del sistema per migliorare l'utilizzo delle risorse. Inoltre, migliorare la conoscenza preliminare degli agenti permette loro di comprendere meglio il sistema e ridurre gli errori predittivi, promuovendo resilienza e auto-organizzazione.

Un esempio possibile di applicazione proposto da Albarracin e colleghi (2024) riguarda la pianificazione urbana, dove, grazie all'AIF, potrebbe essere possibile simulare e prevedere gli esiti di diverse strategie di sviluppo, come il miglioramento dei trasporti pubblici, la creazione di spazi verdi o la gestione delle risorse naturali. Attraverso modelli che incorporino dati demografici, economici e ambientali, sarebbe possibile analizzare le relazioni tra variabili e prevedere effetti sull'uso delle risorse, sull'inquinamento e sul benessere dei residenti, aiutando i decisori a trovare uno stato stabile e sostenibile del sistema.

Promuovere la sostenibilità richiede quindi un approccio globale e sistemico che consideri l'interconnessione tra componenti e livelli del sistema, protegga le risorse naturali, incoraggi pratiche sostenibili e coltivi resilienza e benessere nel lungo periodo. Questo approccio, supportato dall'aggiornamento continuo delle credenze sull'ambiente e dalla riduzione degli errori predittivi, apre la strada a un'etica basata sui dati, dove l'azione più etica è quella che minimizza l'energia libera attesa e massimizza il benessere collettivo, la resilienza e la sostenibilità su scale temporali e spaziali estese (Albarracin Ramstead Pitliya, 2024).

Sulla base del lavoro sopracitato, emerge dunque come i sistemi devono affrontare shock e stress esterni per mantenere sostenibilità, resilienza e benessere. La resilienza, infatti, consiste nell'assorbire gli shock e gli stress provenienti dall'ambiente, mentre la sostenibilità richiede la capacità duratura di rimanere resilienti. In questo paradigma, le perturbazioni esterne sono centrali per sviluppare strategie migliori volte a mantenere il benessere attraverso i diversi livelli del sistema. Poiché queste perturbazioni possono essere imprevedibili, la temporalità delle strategie può variare: strategie di lungo termine, ad esempio, possono tollerare temporanei aumenti dell'energia libera per raggiungere condizioni più stabili e favorevoli in futuro (Albarracin Hipólito Raffa, 2024).

Dunque, l'AIF rappresenta sia una CA che unisce percezione, azione e apprendimento in modo dinamico, sia un modello applicabile alla sostenibilità, capace di

favorire resilienza, equilibrio e adattamento nei sistemi complessi, offrendo un approccio integrato per affrontare le sfide della gestione delle risorse e della stabilità a lungo termine.

Per validare quanto sostenuto finora, è utile considerare l'implementazione di una simulazione ispirata all'esempio dell'individuo in una foresta sopracitato. Nello specifico, il prossimo capitolo illustra una simulazione che testa l'utilizzo di strategie a lungo termine che comportano temporanei aumenti dell'energia libera all'interno di un sistema che cerca di attuare pratiche sostenibili. Nella simulazione, un agente affamato impara a bilanciare la gratificazione immediata con la sostenibilità a lungo termine, resistendo alla tentazione di consumare subito il cibo disponibile. Attraverso una gestione più oculata delle risorse, l'agente tollera temporanei disagi (aumento dell'energia libera) per mantenere un equilibrio dinamico e raggiungere stabilità nel tempo.

CAPITOLO IV

UN'APPLICAZIONE:

MODELLARE LA SOSTENIBILITÀ TRAMITE LA GESTIONE DELLE RISORSE

4.1 Simulazione

Nei capitoli precedenti è stato illustrato il rapporto tra IA e sostenibilità e si è visto come i modelli cognitivi possono avere un ruolo fruttuoso per IA che siano socialmente sostenibili, ovvero spiegabili. Inoltre, è stato proposto il modello dell'AIF come CA sostenibile in sé, nonché utile per l'implementazione di strategie sostenibili all'interno di un sistema. Si è visto infine come per un organismo si intendano sostenibili quelle strategie che puntano a mantenere l'equilibrio ottimale tra tutte le componenti, evitando sprechi di risorse. Esse consistono nello svolgere attività predittive fondate sull'esperienza accumulata, riducendo di conseguenza la tendenza a intraprendere azioni imprevedibili.

Su tali basi, il presente capitolo illustra una simulazione implementata con Mahault Albarracin e Paul F. Kinghorn[43], che mostra come modellare la sostenibilità attraverso l'AIF consenta di rappresentare ambienti in cambiamento dinamico, tenendo conto dei bisogni immediati di un sistema e incorporando anche obiettivi a lungo termine. L'AIF, infatti, permette agli agenti di adattare i propri comportamenti in risposta alle fluttuazioni

[43] Tale simulazione è stata l'obiettivo del tirocinio da me svolto presso l'azienda *Verses AI* tra febbraio e maggio 2024. Il tirocinio è avvenuto sotto la supervisione della dott.ssa Mahault Albarracin, *Director of Innovation R&D* di *Verses AI*, presso il laboratorio dell'azienda, che ha sede nel dipartimento di *Informatics and Engineering* dell'Università del Sussex (Brighton). Lì sono stata seguita anche dal prof. Christopher L. Buckley, *Machine Learning Lead* di *Verses AI* e dal suo gruppo di ricerca, l'*IM2 Lab*, in particolare dal dott. Paul F. Kinghorn, dalla dott.ssa Poppy Collis e dal dott. Filippo Torresan, in collaborazione dei quali è stata sviluppata la simulazione. L'articolo contenente i risultati di tale studio (https://arxiv.org/abs/2406.07593) è stato presentato per la prima volta presso il *5th International Workshop on Active Inference (IWAI 2024)*, tenutosi presso l'università di Oxford tra il 9 e l'11 settembre 2024.

nella disponibilità delle risorse e per questo risulta particolarmente rilevante per gli studi sulla sostenibilità. Negli scenari contemporanei tratteggiati nel Capitolo I, come il cambiamento climatico e l'esaurimento delle risorse del nostro pianeta, la capacità degli agenti di adattare le proprie strategie sulla base delle predizioni di stati futuri rappresenta un elemento cruciale che è necessario modellare.

Attraverso la simulazione proposta viene sottolineata quindi l'importanza delle strategie adattive, mostrando come un agente possa affrontare le sfide legate alla gestione delle risorse in ambienti soggetti a cambiamenti. Nello specifico, la simulazione consiste nel modellare un agente che si trova all'interno di una stanza con del cibo, di cui deve regolare il consumo in modo da restare in vita il più a lungo possibile lungo un arco temporale definito (Albarracin Hipólito Raffa Kinghorn, 2024). In tale modello, l'agente non è semplicemente addestrato a massimizzare una funzione di ricompensa, come si sarebbe potuto simulare tramite RL, ma sviluppa un modello generativo interno che gli consente di prevedere l'impatto delle proprie azioni sul sistema dinamico nel tempo. Questa capacità di pianificazione e simulazione di scenari futuri è cruciale per l'adozione di un comportamento sostenibile, poiché promuove strategie adattive che evitano l'esaurimento delle risorse e supportano la resilienza del sistema. Inoltre, l'AIF permette di tenere traccia delle incertezze e delle distribuzioni probabilistiche durante il processo decisionale. Questo aspetto è fondamentale in contesti dinamici come la gestione delle risorse, dove l'agente deve continuamente aggiornare la propria conoscenza e predire stati futuri in condizioni di incertezza. In questo caso, quindi, l'AIF risulta vantaggiosa rispetto al RL in quanto monitora non solo le azioni intraprese ma anche il grado di confidenza nelle previsioni effettuate, mentre l'RL funziona su ricompense definite e azioni ottimali.

La simulazione implementata consta di due scenari, i cui dettagli saranno affrontati nel prossimo paragrafo. Nel primo scenario l'ambiente è statico, ovvero, il cibo è sempre disponibile e non avvengono cambiamenti. L'agente si trova a dover decidere se consumare il cibo oppure no. Questo primo scenario costituisce un caso di base che ha lo scopo di verificare se l'agente è in grado di imparare a nutrirsi quando è affamato in un contesto semplice e immutabile. Nel secondo scenario, invece, l'ambiente è dinamico e l'agente deve imparare a moderare il consumo del cibo nel corso di più intervalli temporali. In questo contesto, il cibo si esaurisce quando l'agente mangia, mentre può rigenerarsi se l'agente sceglie di non consumarlo. Viene introdotto un meccanismo di

apprendimento che spinge l'agente a pianificare strategicamente il proprio consumo, così da evitare sia la fame, dovuta a scarso nutrimento, sia l'esaurimento delle risorse dovuto a un consumo eccessivo. Lo scopo della simulazione è duplice: in primo luogo punta a indagare le strategie adattive necessarie per prevedere quando un sistema raggiungerà stabilità e sostenibilità a lungo termine, imparando a bilanciare i bisogni immediati con gli obiettivi futuri. In secondo luogo, risulta utile per individuare le vulnerabilità all'interno del sistema, al fine di prevenire possibili collassi o disfunzioni.

4.1.1 Metodi

La simulazione è stata implementata usando il pacchetto PyMDP, realizzato da Conor Heins e colleghi (2022). Entrambi gli scenari che abbiamo descritto sopra – Caso 1: ambiente statico e Caso 2: ambiente dinamico – sono basati sul medesimo modello generativo, illustrato in *Fig. 2*:

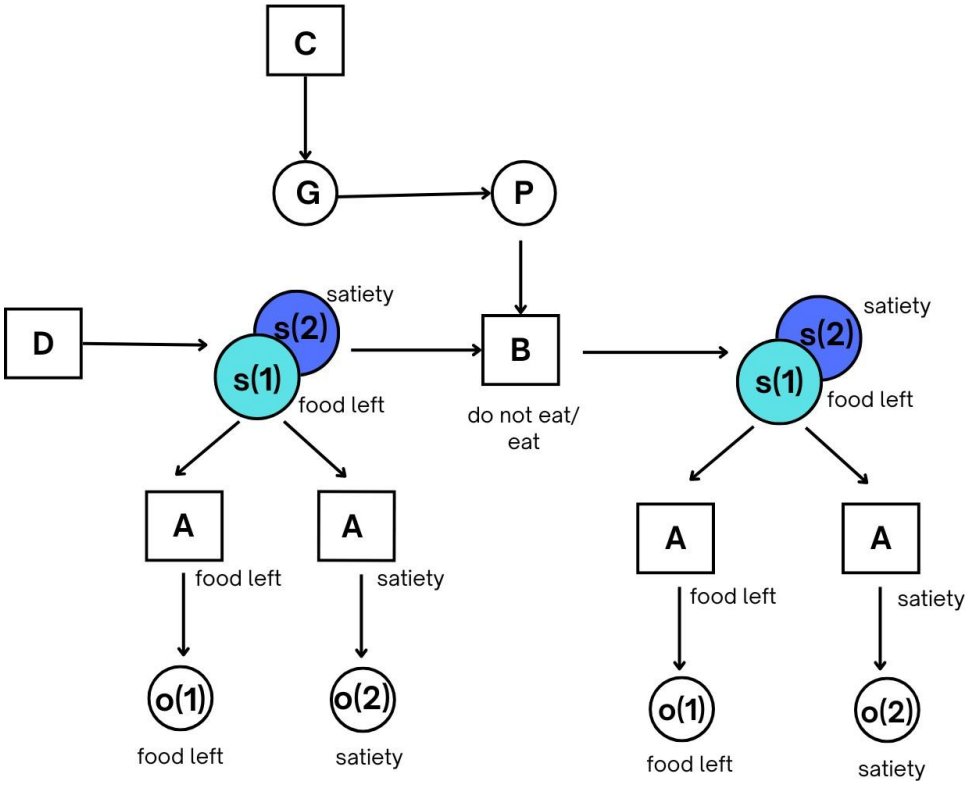

Fig. 2. Grafo che illustra il modello generativo dell'agente per il Caso 1 e il Caso 2. Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 4.

Il modello generativo dell'agente rappresenta la codificazione delle sue credenze riguardo alla struttura causale del mondo e al modo in cui le sue azioni possono influenzare gli stati del mondo. Lo stato reale dell'ambiente è rappresentato da due fattori di stato non noti ("hidden state factors", Albarracin Hipólito Raffa Kinghorn, 2024, 4): la disponibilità del cibo (*s(1)* in *Fig. 2*) e l'essere sazio dell'agente (*s(2)* in *Fig. 2*). L'agente dispone di due modalità di osservazione, che sono analoghe ai fattori di stato non noti: la disponibilità di cibo (*o(1)* in *Fig. 2*) e il proprio livello di sazietà (*o(2)* in *Fig. 2*). A ogni intervallo di tempo, l'agente può scegliere tra due azioni: "mangiare" oppure "non mangiare".

Proseguendo con la descrizione dello schema in *Fig. 2*, *A* è la mappatura di verosimiglianza ("likelihood mapping"), che specifica la probabilità delle osservazioni (ovvero cibo presente/assente, agente sazio/affamato) date le variabili di stato non note (ovvero la presenza effettiva di cibo, lo stato di sazietà effettivo dell'agente). La matrice *A* codifica quindi le percezioni dell'agente. Essa assume un'identità tra gli stati non noti e le osservazioni, ovvero implica che l'agente osserva in modo diretto lo stato effettivo del cibo disponibile e il proprio stato di sazietà.

B in *Fig. 2* è la matrice di transizione ("transition matrix"), che descrive come le azioni dell'agente (mangiare o non mangiare) influenzano gli stati non noti, dunque codifica la pianificazione dell'agente. Nello specifico: quando l'agente mangia, il cibo da presente diventa assente e avviene una transizione dell'agente da affamato a sazio. Quando l'agente non mangia, le dinamiche cambiano a seconda dello scenario preso in considerazione: nel Caso base 1 (ambiente statico) la matrice *B* stabilisce che mangiare porta a sazietà, mentre il cibo resta costantemente disponibile. Nel Caso 2 (ambiente dinamico): la matrice *B* viene aggiornata: mangiare quando il cibo è disponibile porta l'agente alla sazietà, ma il cibo diventa assente, seguendo dinamiche realistiche. Non mangiare, invece, comporta un aumento del cibo disponibile.

C in *Fig. 2* è il vettore delle preferenze prioritarie ("prior preferences vector") che codifica gli obiettivi dell'agente come preferenze rispetto alle osservazioni: l'agente mostra una preferenza neutra rispetto alla presenza di cibo, ma preferisce essere sazio piuttosto che affamato. Tali preferenze sono codificate dal vettore *C* come una distribuzione categorica, dove i valori più alti corrispondono alle osservazioni preferite. Lo scopo dell'agente è massimizzare la probabilità di osservare gli stati preferiti.

Le condizioni iniziali della simulazione sono specificate dal vettore *D*, ovvero la distribuzione relativa allo stato iniziale ("initial state distribution"). *G* in *Fig. 2* rappresenta poi l'energia libera, quindi la discrepanza tra le previsioni dell'agente (il suo modello generativo) e le osservazioni effettive. Infine, *P* indica le strategie ("policies") messe in campo dall'agente per la minimizzazione dell'energia libera attesa, ovvero la sequenza di azioni scelta – l'orizzonte di pianificazione – per la sopravvivenza.

Il processo della simulazione è illustrato nei seguenti diagrammi di flusso (*Fig. 5*):

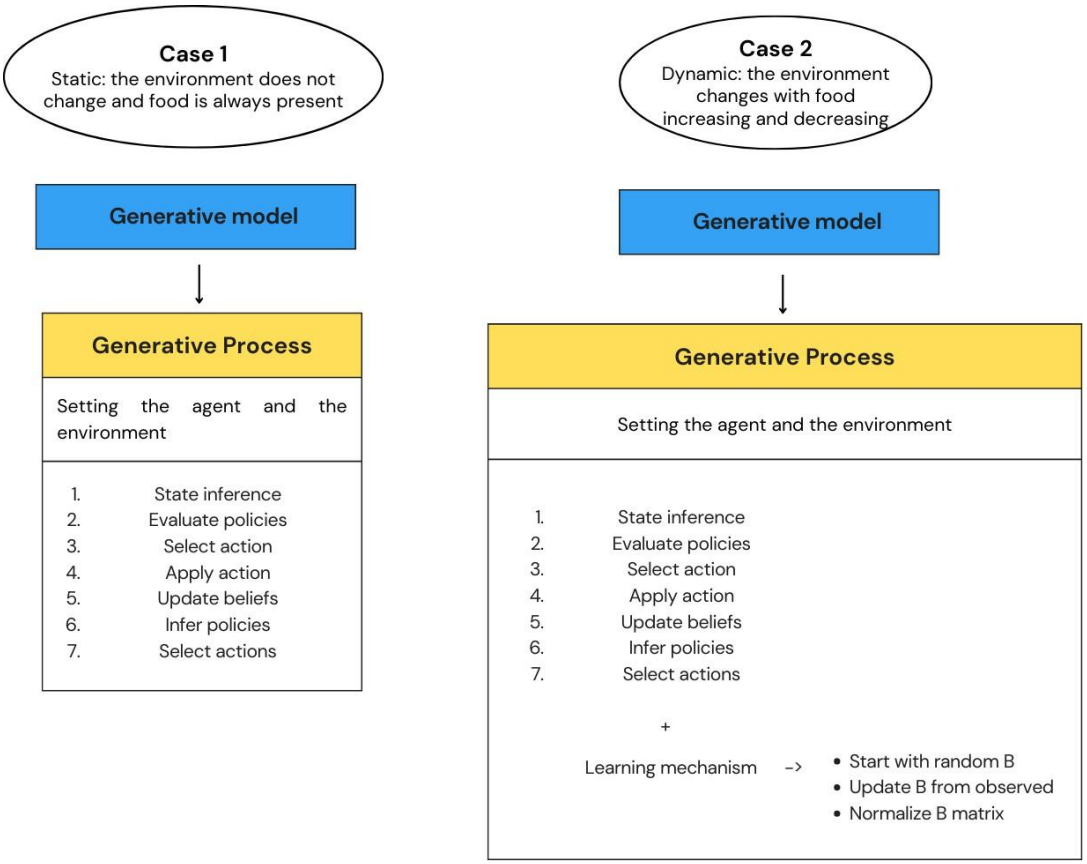

Fig. 3. Diagrammi di flusso che riassumono la configurazione sperimentale per il Caso base 1 (a sinistra, ambiente statico) e il Caso 2 (a destra, ambiente dinamico). Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 5.

Per quanto riguarda il Caso base 1 (*Fig. 3*, sinistra), il cibo e la sazietà, sia come osservazioni sia come stati non noti, possono essere assenti (0) o presenti (1). Le condizioni iniziali (vettore *D*) prevedono che l'agente si trovi in uno stato di fame, quindi sazietà = 0. Dopo aver impostato il modello generativo (le cui componenti sono riassunte

nel dettaglio nella tabella in *Fig. 4*), durante il processo generativo l'agente interagisce con l'ambiente e le sue azioni influenzano le transizioni di stato secondo le dinamiche del processo stesso. Poiché l'agente si trova in un ambiente statico, come si è visto, se sceglie di non mangiare lo stato dell'ambiente rimane immutato, dunque le risorse sono sempre disponibili. Se invece sceglie di mangiare e il cibo è disponibile, l'agente diventa sazio; al contempo, il cibo rimane disponibile a causa della natura statica dell'ambiente.

Component	Values	Description
Hidden States	Food availability: present (1), absent (0)	Represents the true state of food availability in the environment
	Agent satiety: hungry (0), satiated (1)	Represents the true state of the agent's satiety
Observations	Observed food availability	Corresponds directly to the food availability state
	Agent's perceived satiety	Corresponds directly to the agent satiety state
Actions	Eat (1), Don't Eat (0)	The actions available to the agent
Likelihood Matrix (A)	Identity mapping	Assumes the agent directly observes the true states
Transition Matrix (B)	"Eat" (1) leads to satiety (1) when food is present (1)	Specifies the state transitions based on the agent's actions
	"Don't Eat" (0) leads to hunger (0)	
Preference Vector (C)	Strong preference for satiated (1) and food present (1)	Encodes the agent's goals and drives its behavior
Initial State Distribution (D)	Uniform Distribution	Sets the starting conditions for the simulation

Fig. 4. Componenti del modello generativo dell'agente nel Caso base 1 in ambiente statico. Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 6.

A questo punto viene avviato il ciclo della simulazione. L'agente esegue un'inferenza sugli stati basandosi sulle osservazioni presenti; valuta le politiche disponibili per massimizzare l'energia libera attesa e seleziona l'azione che minimizza l'energia libera e si allinea meglio alle proprie preferenze; l'azione selezionata viene applicata all'ambiente, determinando le transizioni di stato e producendo nuove osservazioni; infine, il ciclo si ripete, con l'agente che aggiorna le proprie credenze, inferisce nuove politiche e seleziona ulteriori azioni per raggiungere i propri obiettivi (*Fig. 3*).

Per testare la robustezza del modello è stata introdotta una variazione del Caso base 1. Ovvero, sono state impostate delle matrici *A* e *B* scorrette, dotando quindi l'agente di percezioni e credenze distorte riguardo alle transizioni di stato.

Il Caso base 1 ha lo scopo di verificare che il comportamento dell'agente sia effettivamente basato su una valutazione corretta dell'ambiente. Dopo essersi assicurati di ciò, la complessità del modello aumenta nel secondo scenario simulato. Nel Caso 2, infatti, l'agente si trova in un ambiente dinamico: a differenza dello scenario precedente, dove il cibo era sempre presente nonostante venisse mangiato, in questo caso le azioni dell'agente hanno conseguenze sulla disponibilità del cibo nel corso del tempo. Lo scopo è studiare come l'agente adotti un comportamento sostenibile, bilanciando il proprio bisogno immediato di sazietà con la disponibilità a lungo termine del cibo.

Rispetto al Case base 1, il modello generativo dell'agente nel Caso 2 (le cui componenti sono riassunte nella tabella in *Fig. 5*) è impostato con maggiore granularità

Component	**Values**	**Description**
Hidden States	Food left: none (0), some (1), abundant (2)	Represents the true state of food availability in the environment
	Agent satiety: not satiated (0), somewhat satiated (1), fully satiated (2)	Represents the true state of the agent's satiety
Observations	Observed food availability	Corresponds to the food availability state with some variability
	Agent's perceived satiety	Corresponds to the agent satiety state with some variability
Actions	Eat (1), Don't Eat (0)	The actions available to the agent
Likelihood Matrix (A)	High probability of correct observations, lower for adjacent states	Defines the probability of observations given the true hidden states
Transition Matrix (B)	"Eat" (1): food left decreases, satiety increases	Specifies the state transitions based on the agent's actions and current state
	"Don't Eat" (0): food left increases, satiety decreases	
Preference Vector (C)	Strong preference for satiety. Balances maintaining satiety and sustainable food supply	Encodes the agent's goals and drives its behavior
Initial State Distribution (D)	Uniform Distribution	Sets the starting conditions for the simulation
Policy Length	3 time steps	Allows the agent to plan ahead and consider long-term effects

Fig. 5. Componenti del modello generativo dell'agente nel Caso 2, in ambiente dinamico. Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 8.

negli stati e nelle osservazioni, consentendo una gamma più ampia di variazioni e comportamenti nell'interazione tra agente e ambiente.

Se nel Caso base 1 il cibo e la sazietà erano solo assenti (0) o presenti (1), sia in quanto osservazioni sia in quanto stati non noti, adesso vi sono tre livelli per entrambi. Ovvero cibo disponibile: nessuno (0), in parte (1), abbondante (2); sazietà: assente (0), soddisfatta in parte (1), del tutto soddisfatta (2). Come nel Caso 1, si assume che l'agente osservi direttamente gli stati ambientali reali (identità nella matrice *A*), con alcune variazioni nei diversi livelli di disponibilità del cibo e di sazietà. In questo ambiente, le transizioni dipendono sia dallo stato attuale sia dall'azione intrapresa dall'agente: se l'agente non mangia, la disponibilità di cibo aumenta nel tempo, mentre se l'agente mangia, la disponibilità di cibo diminuisce o si esaurisce del tutto. Per quanto riguarda lo stato di sazietà: se l'agente non mangia, la sazietà diminuisce progressivamente, se invece l'agente mangia, la sazietà aumenta. Le preferenze (codificate dal vettore *C*) sono progettate per trovare un equilibrio tra il mantenimento della sazietà e la garanzia di una fornitura sostenibile di risorse alimentari. Questo incoraggia l'agente a massimizzare la propria sazietà, tenendo però conto della disponibilità a lungo termine delle risorse. Nello specifico, l'agente mostra una forte preferenza per il raggiungimento della sazietà, mentre ha una preferenza neutra rispetto alla quantità di cibo rimanente. Le condizioni iniziali (vettore *D*) prevedono che il cibo sia abbondante (cibo disponibile = 2) e l'agente sazio in parte (sazietà = 1).

Dopo aver impostato il modello generativo, come per il Caso 1, la simulazione prosegue con il processo generativo (*Fig. 3*, destra) e, in una prima fase senza meccanismo di apprendimento. Come visto, in questo secondo scenario, il cibo si esaurisce quando l'agente mangia, mentre può rigenerarsi se l'agente sceglie di non consumarlo. In particolare, esso diminuisce o si rigenera di 1 unità per intervallo temporale. L'agente interagisce con l'ambiente dinamico per 10 intervalli di tempo, aggiornando le proprie credenze e azioni in base agli stati osservati e alla dinamica variabile dell'ambiente. Nel Caso 2 viene utilizzata una lunghezza delle politiche ("policy length") più estesa rispetto al Caso 1, per pianificare su più intervalli di tempo e anticipare le conseguenze future. Mentre nel primo scenario la lunghezza delle politiche era pari a 1 intervallo temporale, adesso è pari a 3 intervalli temporali e l'agente è quindi in grado di bilanciare il consumo immediato con la sostenibilità a lungo termine. La lunghezza

estesa delle politiche, infatti, permette all'agente di prevedere gli stati futuri ed evitare comportamenti miopi o avidi, che potrebbero portare alla fame. Le politiche dell'agente sono inoltre vincolate per garantire azioni coerenti e sostenibili nel corso di tutti gli intervalli temporali e per entrambe le modalità di osservazione (disponibilità di cibo e livello di sazietà).

Nel Caso 2 viene poi introdotto un meccanismo di apprendimento, in modo che l'agente impari a non mangiare anche se non è completamente sazio. La matrice *B* del modello generativo è inizializzata in modo randomico, attraverso una distribuzione Dirichlet che assicura valori di probabilità validi[44]. Essendo inizializzata in tale modo, la matrice non cattura in modo corretto le transizioni degli stati, quindi l'agente non dispone di indicazioni predefinite su come le sue azioni influenzeranno lo stato del mondo e aggiorna la matrice in base all'esperienza derivata dall'interazione con l'ambiente, quindi da ciò che osserva. Come prima, a ogni intervallo di tempo, l'agente osserva, inferisce gli stati, valuta le politiche e seleziona le azioni. Dopo aver eseguito un'azione e ricevuto la nuova osservazione, l'agente aggiorna la propria matrice *B*. In particolare, l'agente registra la transizione dallo stato precedente allo stato attuale in base all'azione intrapresa. Dopo l'aggiornamento, la matrice *B* viene normalizzata per garantire che la somma delle probabilità sia pari a 1, mantenendo così una distribuzione di probabilità valida.

Anche per questo secondo scenario sono state testate delle variazioni, in modo da esplorare il comportamento e le prestazioni dell'agente in condizioni differenti e per testare la robustezza del modello introducendovi errori.

La prima estensione del Caso 2 prevede l'inizializzazione della matrice *B* con valori errati, impostati su estremi (1 e 0, anziché probabilità più basse), che indicano massima certezza (1) o impossibilità (0) che una determinata azione causi una specifica transizione di stato. I valori estremi non riflettono la realtà di un ambiente complesso o dinamico, dove le probabilità sono più sfumate e raramente certe al 100%: nella maggior parte dei casi realistici, le transizioni hanno probabilità intermedie che rappresentano un margine di incertezza, pertanto ci si aspetta che le prestazioni dell'agente peggiorino, dimostrando l'importanza di modelli di transizione accurati.

[44] La distribuzione di Dirichlet è una distribuzione di probabilità utilizzata per generare insiemi di valori probabilistici che devono sommarsi a 1. È spesso utilizzata nei modelli probabilistici quando si lavora con categorie o eventi mutuamente esclusivi (come appunto le probabilità di stati o osservazioni).

Nella seconda estensione viene esaminato il comportamento dell'agente modificando le preferenze prioritarie codificate dal vettore *C*, attribuendo cioè una forte preferenza alla presenza di cibo. In questo caso, l'agente tende a dare priorità ad azioni che garantiscono la presenza di cibo, anche a scapito della propria sazietà. Al contrario, impostando una preferenza sul cibo bassa, si osserva che l'agente riesce ad apprendere il comportamento corretto, in quanto dimostra una chiara preferenza nell'evitare di rimanere senza cibo.

Nella terza estensione, si testa l'agente in un ambiente dove il cibo cresce a una velocità più lenta (0,5 unità per intervallo temporale, rispetto a 1 unità per intervallo temporale, come era nel Caso 2 principale) quando l'agente non mangia e si esaurisce a una velocità più rapida (1 unità per intervallo temporale) quando l'agente mangia. Allo stesso tempo, la sazietà diminuisce più velocemente quando l'agente non mangia (0,2 unità per intervallo temporale, rispetto a 1 del Caso 2 principale) e aumenta più lentamente quando l'agente mangia (0,8 unità per intervallo temporale, rispetto a 1 del Caso 2 principale). In queste condizioni, l'agente deve adattare la propria strategia per tenere conto dei cambiamenti specifici dell'ambiente. Le prestazioni dell'agente potrebbero essere inferiori rispetto al Caso 2 principale, a causa della maggiore difficoltà nel bilanciare i livelli di cibo e sazietà, dato che il cibo si esaurisce più rapidamente quando viene consumato e la sazietà cala più velocemente quando non viene consumato.

La quarta estensione prevede poi la simulazione di dinamiche stagionali per testare la capacità dell'agente di riaddestrarsi e adattarsi ai cambiamenti nelle regole dell'ambiente dopo la fase iniziale di apprendimento. Vengono introdotti dei parametri aggiuntivi per simulare l'estate e l'inverno, aggiornando i tassi di crescita e di esaurimento del cibo in base alla stagione corrente. Viene implementato inoltre un interruttore temporale all'interno del processo generativo, che simula l'alternarsi della stagione tra estate e inverno dopo un determinato numero di intervalli temporali.

Infine, nella quinta estensione viene valutato l'impatto dell'orizzonte di pianificazione sulle prestazioni dell'agente, confrontando più agenti con orizzonti di pianificazione differenti (lunghezza delle politiche di 1 intervallo temporale contro 3 intervalli temporali). Ci si aspetta che gli agenti con un orizzonte di pianificazione più lungo ottengano prestazioni migliori, poiché sono in grado di anticipare meglio gli stati futuri e di prendere decisioni più efficaci per una gestione sostenibile delle risorse.

4.1.2 Risultati

Per quanto riguarda i risultati della simulazione, nel Caso base 1 in ambiente statico, come si è visto, il cibo è sempre disponibile, e lo scopo dell'agente è mantenere lo stato di sazietà. L'agente sceglie costantemente di mangiare a ogni intervallo temporale (*Fig. 6*, prima riga dall'alto), dimostrando di comprendere che il cibo è sempre disponibile (*Fig. 6*, seconda riga) e che mangiare massimizza il suo livello di sazietà (*Fig. 6*, terza riga). La disponibilità di cibo rimane costante durante l'intera simulazione. La sazietà dell'agente aumenta man mano che mangia e si mantiene a livelli elevati, indicando un adattamento ottimale per mantenere il proprio stato interno nel modo migliore (*Fig. 6*, terza riga dal basso). Questo scenario, che funge da base per la simulazione, dimostra la capacità dell'agente di comportarsi in modo ottimale in un ambiente con risorse costanti.

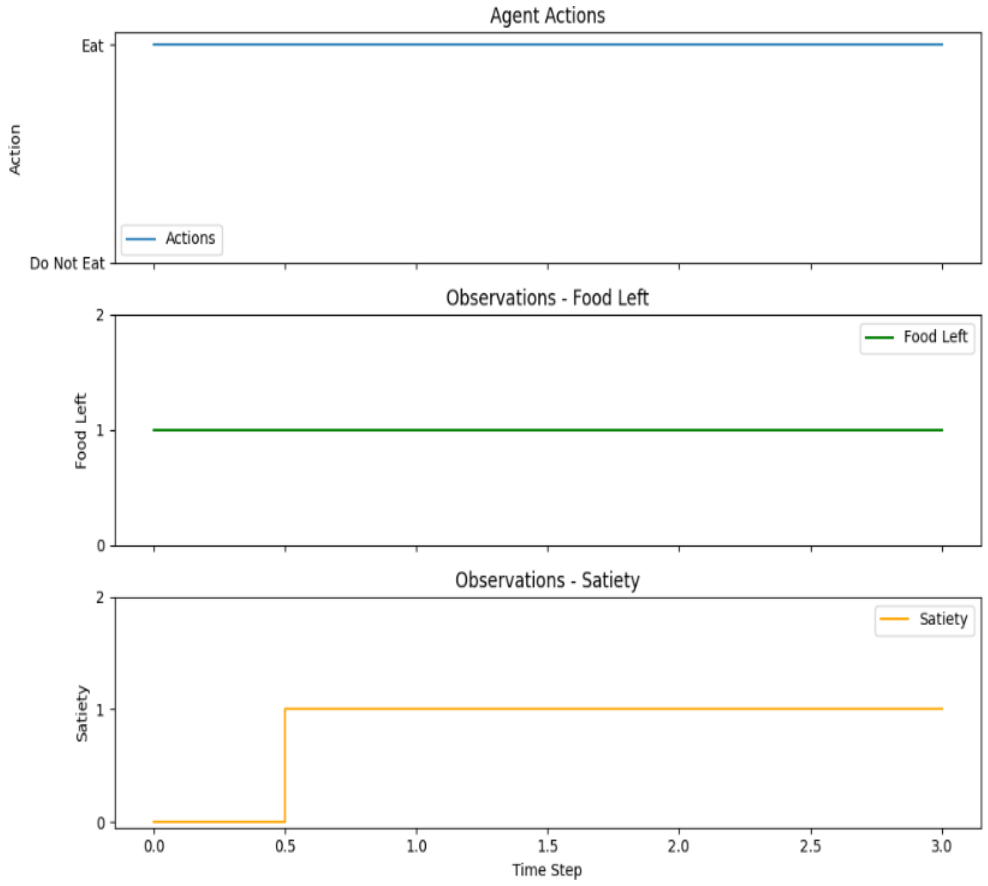

Fig. 6. Risultati del Caso base 1, ambiente statico. Sugli assi delle *x* sono rappresentati gli intervalli temporali di sopravvivenza dell'agente, e sugli assi delle *y*, dal grafico più in alto, rispettivamente le azioni scelte dall'agente, i livelli di cibo disponibile e di sazietà. Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 15.

Nella variazione del Caso base 1, che prevede l'introduzione degli errori nelle matrici *A* e *B*, vengono testate la resilienza e l'adattabilità dell'agente quando il suo modello interno non rappresenta l'ambiente in modo accurato.

Le azioni dell'agente mostrano un andamento più erratico (*Fig. 7*, prima riga dall'alto) e riflettono una certa confusione o incertezza dovuta ai modelli di percezione e pianificazione errati. Nonostante la rappresentazione scorretta, la disponibilità di cibo rimane costante (*Fig. 7*, seconda riga), come nel Caso base 1 principale. Tuttavia, la sazietà dell'agente presenta fluttuazioni maggiori rispetto al Caso base 1 principale, e ciò indica che la capacità dell'agente di mantenere uno stato interno coerente è compromessa dalle percezioni e dai modelli di pianificazione errati. Questa variazione dimostra come deviazioni da modelli ambientali accurati possano influenzare il comportamento e le prestazioni dell'agente, portando a decisioni meno ottimali. Con i risultati del Caso 1, dunque, che funge da test per la simulazione, è stato dimostrato che il modello dell'agente ha un certo grado di validità e che esso è effettivamente in grado di reagire alla qualità del proprio modello ambientale, scegliendo le azioni migliori relative alla propria sopravvivenza.

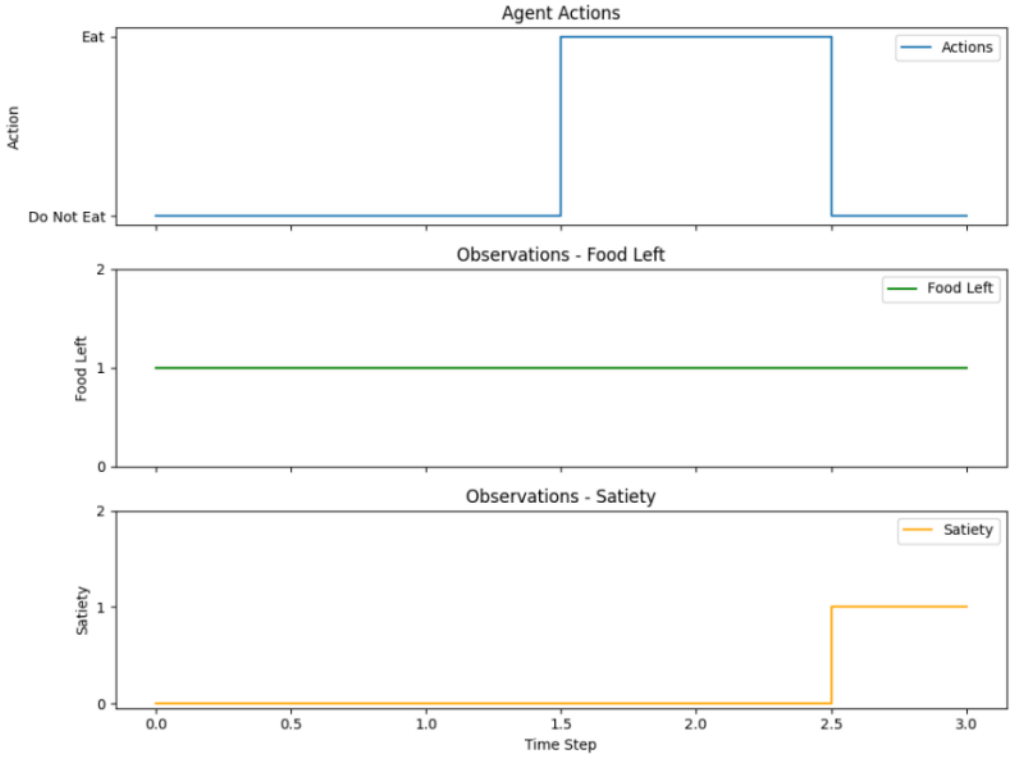

Fig. 7. Risultati dell'estensione del Caso base 1, ambiente statico, con la matrice B impostata con valori estremi. Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 16.

Per quanto riguarda i risultati del Caso 2, come visto, occorre ricordare che l'ambiente è dinamico: il cibo si esaurisce quando viene consumato e si rigenera se non viene mangiato. L'agente deve bilanciare il proprio comportamento alimentare per evitare sia la fame sia l'esaurimento delle risorse. L'agente è dotato di una forte preferenza per uno stato di sazietà pari a 2 e di una preferenza neutra rispetto alla quantità di cibo rimanente. La preferenza neutra è stata impostata per testare la capacità dell'agente di imparare a conservare il cibo, sebbene non gli sia stato esplicitamente detto di farlo, e sopravvivere nel tempo. In teoria, l'agente non dovrebbe preoccuparsi della disponibilità di risorse, ma soltanto del proprio livello di sazietà. In più esecuzioni della simulazione, con un orizzonte di pianificazione (lunghezza della politica) di 3 intervalli temporali e senza meccanismo di apprendimento, l'agente tende a non mangiare, morendo di fame (*Fig. 8*, sinistra). Altrimenti, mangia troppo, esaurendo le risorse e causando ugualmente la propria morte (*Fig. 8*, destra).

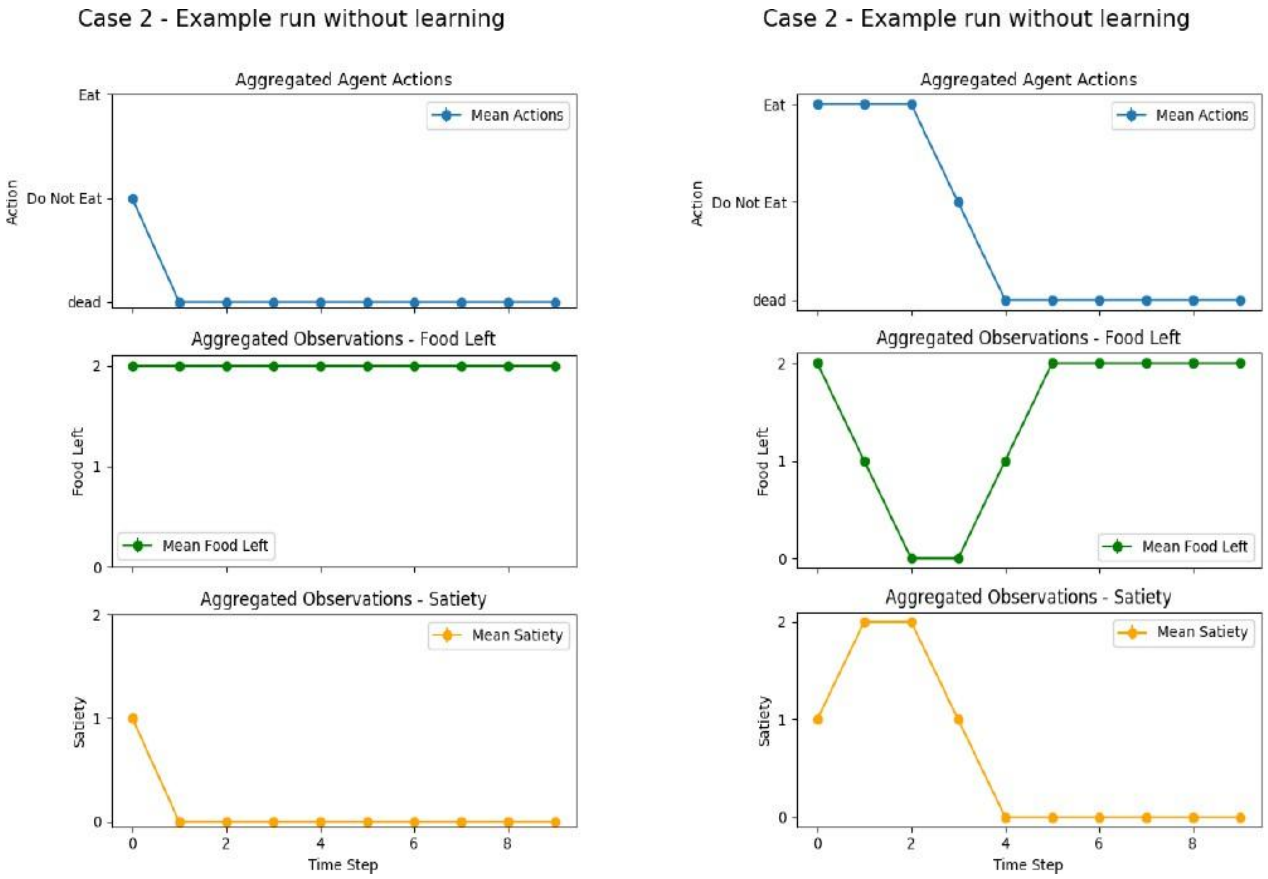

Fig. 8. Risultati del Caso 2, ambiente dinamico, senza meccanismo di apprendimento. Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 18.

É stato visto che nel secondo scenario viene poi aggiunto un meccanismo di apprendimento. In quest'ultimo caso l'agente ha in partenza una matrice *B* inizializzata casualmente e la aggiorna attraverso le interazioni con l'ambiente dinamico. In questa situazione, le azioni dell'agente fluttuano regolarmente tra "mangiare" e "non mangiare" (*Fig. 9*, prima riga dall'alto), suggerendo che abbia appreso una strategia per bilanciare le proprie azioni. Questo gli consente di sopravvivere per tutta la durata della simulazione, mantenendo il livello di sazietà tra 1 e 2 (*Fig. 9*, terza riga).

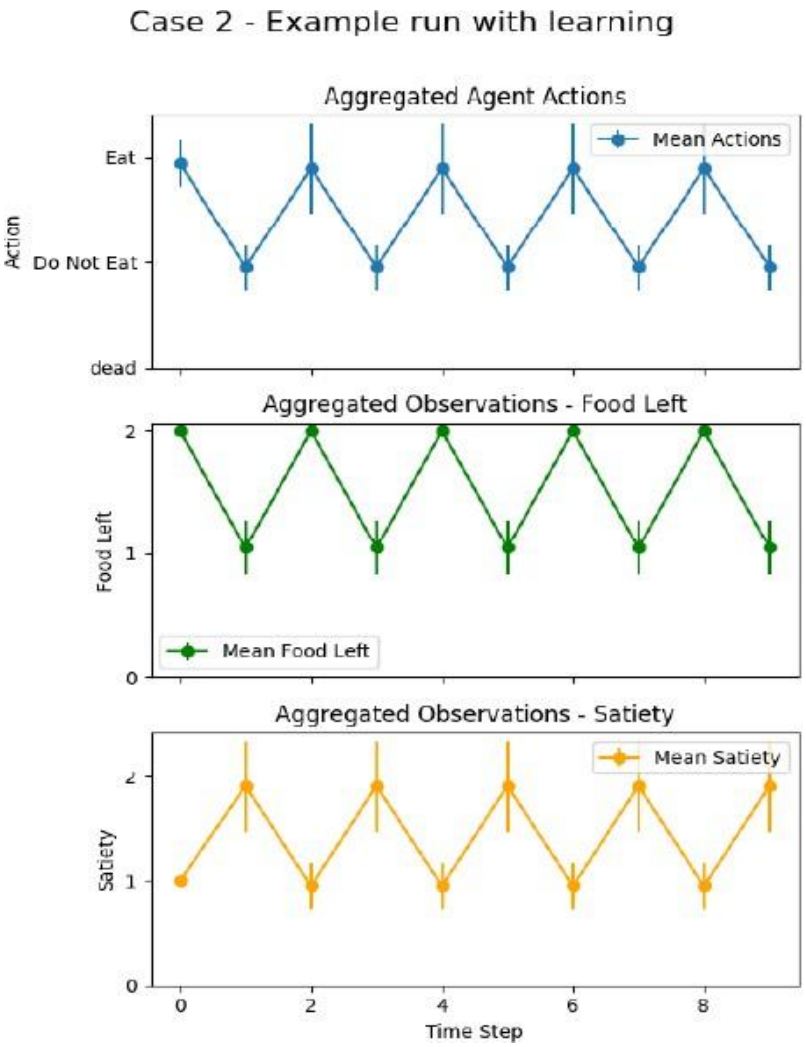

Fig. 9. Risultati del Caso 2, ambiente dinamico, con meccanismo di apprendimento. Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 19.

Osservando il cambiamento della matrice di transizione *B*, per il fattore di stato "cibo disponibile" e l'azione "non mangiare", dalla fase iniziale in cui l'agente non ha ancora avviato il processo di apprendimento fino alla fine di esso, si nota come col passare del tempo l'agente apprenda a compiere l'azione: infatti, nel primo *snapshot* della matrice *B* a sinistra in *Fig. 10*, si osserva come le probabilità di transizione siano più distribuite, suggerendo che l'agente non ha ancora appreso un modello preciso delle conseguenze dell'azione "mangiare". Il cibo può diventare da abbondante a scarso, ma ci sono ancora probabilità relativamente alte che rimanga in stati intermedi. Dopo l'apprendimento (*snapshot* 137-140), alcune transizioni diventano quasi certe (colori scuri), mentre altre sono quasi improbabili (colori chiari). Inoltre, il cibo abbondante diventa scarso con alta probabilità, ovvero, l'agente ha appreso che, se il cibo è presente e viene consumato, la probabilità che il cibo diminuisca è molto alta. Se il cibo è già scarso, d'altra parte, mangiare porta quasi sempre a esaurirlo: la regione inferiore della *heatmap*, infatti, mostra che le transizioni da "cibo presente" a "cibo assente" sono diventate molto probabili.

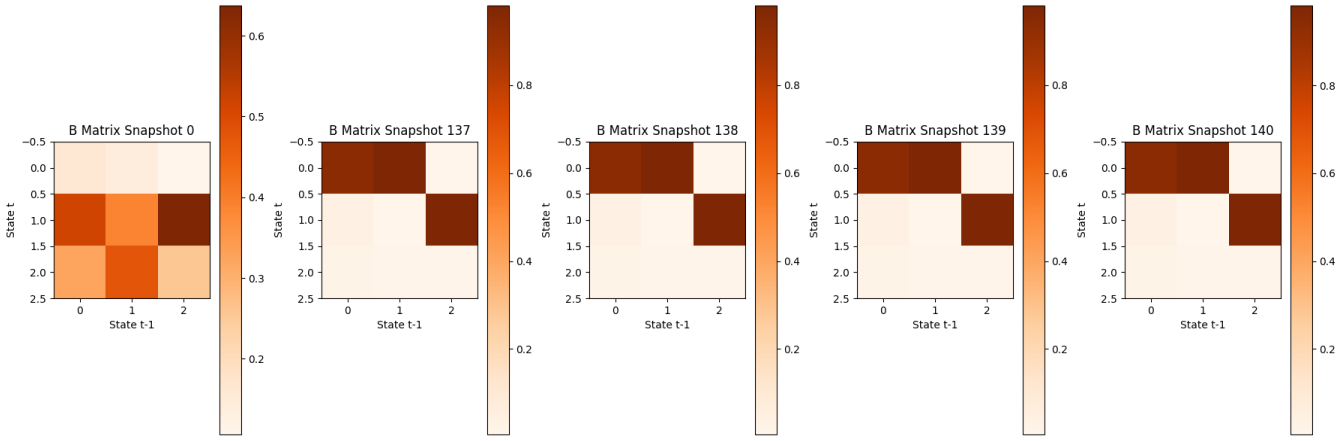

Fig. 10. Evoluzione della matrice di transizione *B* per lo stato "cibo disponibile" quando l'agente sceglie l'azione "mangiare". Sull'asse delle *X* è rappresentato lo stato del cibo al tempo *t*-1, prima dell'azione, mentre sull'asse delle *Y* lo stato del cibo al tempo *t*, dopo l'azione "mangiare". Ogni *snapshot* rappresenta un diverso momento nel processo di apprendimento. All'inizio (sinistra), le probabilità di transizione sono più distribuite, indicando incertezza. Con il tempo (destra), l'agente apprende che mangiare riduce la disponibilità di cibo, rendendo alcune transizioni più probabili (colori più scuri).

Il fattore di stato "sazietà", invece, rimane meno esplorato. In questo caso, infatti, le transizioni rappresentate dalla matrice *B* non vengono apprese in modo perfetto neanche dopo numerose iterazioni. Ciò dipende da diversi fattori legati alla dinamica di apprendimento dell'agente e alla struttura dell'ambiente: ci sono infatti situazioni con

determinati stati e/o azioni in cui l'agente non si imbatte mai, o che sperimenta solo una volta, e quindi non ha abbastanza dati per aggiornare in modo efficace le relative transizioni nella matrice *B*. Questo comporta lacune nell'apprendimento: infatti, se ad esempio l'agente non sperimenta mai la fame estrema – ovvero il livello minimo di sazietà – non avrà modo di apprendere come funzionano le transizioni in quello stato. Il processo di apprendimento, inoltre, segue un principio di soddisfacimento, piuttosto che di ottimizzazione perfetta: una volta che l'agente ha acquisito una conoscenza sufficiente per garantire la propria sopravvivenza, e dato che ha una forte preferenza per essa (espressa nel vettore *C*), non sentirà la necessità di esplorare altre possibilità quando ha già trovato un comportamento efficace. Quindi, rafforzerà costantemente l'apprendimento delle situazioni che incontra frequentemente, mentre ignorerà quelle meno comuni o inesplorate. La matrice *B*, dunque, non converge mai a una rappresentazione completa del mondo, ma a una rappresentazione funzionale che permette all'agente di sopravvivere. Il processo di aggiornamento di *B*, poi, come visto, è di natura probabilistica: se un agente incontra una transizione molto raramente, il suo apprendimento rimarrà incerto e soggetto a rumore. In sintesi, il comportamento dell'agente è guidato dalla necessità di minimizzare la sorpresa e garantire la sopravvivenza, piuttosto che dall'esplorazione esaustiva dell'ambiente. Di conseguenza, la matrice *B* si stabilizza nelle regioni del modello che sono più rilevanti per il raggiungimento degli obiettivi dell'agente, mentre rimane meno definita altrove, risultando in un apprendimento che è funzionalmente sufficiente, ma non perfetto.

Infine, il grafico del tempo di sopravvivenza medio di dieci agenti mostra che l'agente, dopo una fase iniziale di apprendimento, riesce a sopravvivere costantemente per il massimo numero di intervalli temporali, indicando che ha rapidamente imparato una strategia efficace per evitare la fame e mantenere la sopravvivenza (*Fig. 11*, linea rossa). Senza apprendimento, invece, in media gli agenti non sopravvivono oltre la metà del tempo complessivo (*Fig. 11*, linea nera)

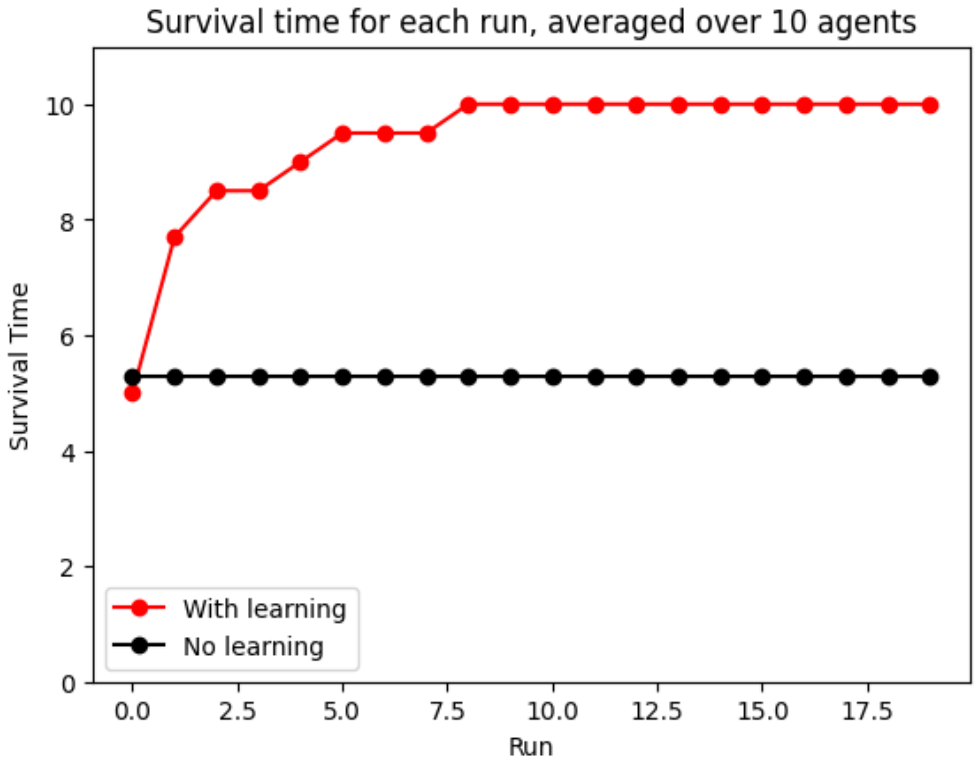

Fig. 11. Grafico del tempo di sopravvivenza medio di 10 agenti per 20 esecuzioni della simulazione (asse delle *x*), su un massimo di 10 intervalli temporali (asse delle *y*) senza meccanismo di apprendimento (linea nera) e con meccanismo di apprendimento (linea nera) e con meccanismo di apprendimento (linea rossa). Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 10.

Il Caso 2 con apprendimento dimostra quindi la capacità dell'agente di imparare dalle interazioni con l'ambiente e di sviluppare strategie più efficaci per la sopravvivenza e la gestione delle risorse.

Le estensioni del Caso 2, come visto, sono state introdotte per esplorare il comportamento dell'agente e le sue prestazioni in varie condizioni. Nella prima estensione, la matrice *B* è stata inizializzata in modo errato, con valori inizialmente impostati su estremi irrealistici (1 e 0, anziché probabilità più basse): in questo caso, senza il meccanismo di apprendimento, l'agente sceglie costantemente di mangiare a ogni intervallo temporale, portando a un comportamento subottimale e, infine, alla morte per fame (*Fig. 12*, grafico in alto a sinistra, prima riga dall'alto). In determinate condizioni, l'agente non è stato in grado di apprendere, rimanendo bloccato in uno stato di inerzia nelle sue transizioni. Tuttavia, nel complesso, con il meccanismo di apprendimento

attivato (*Fig. 12*, tre grafici in alto a destra), l'agente riesce a sopravvivere, uscendo dai valori estremi della matrice *B* e imparando come sopravvivere (*Fig. 12*, grafico in basso), evidenziando il valore della plasticità per superare punti di partenza svantaggiati.

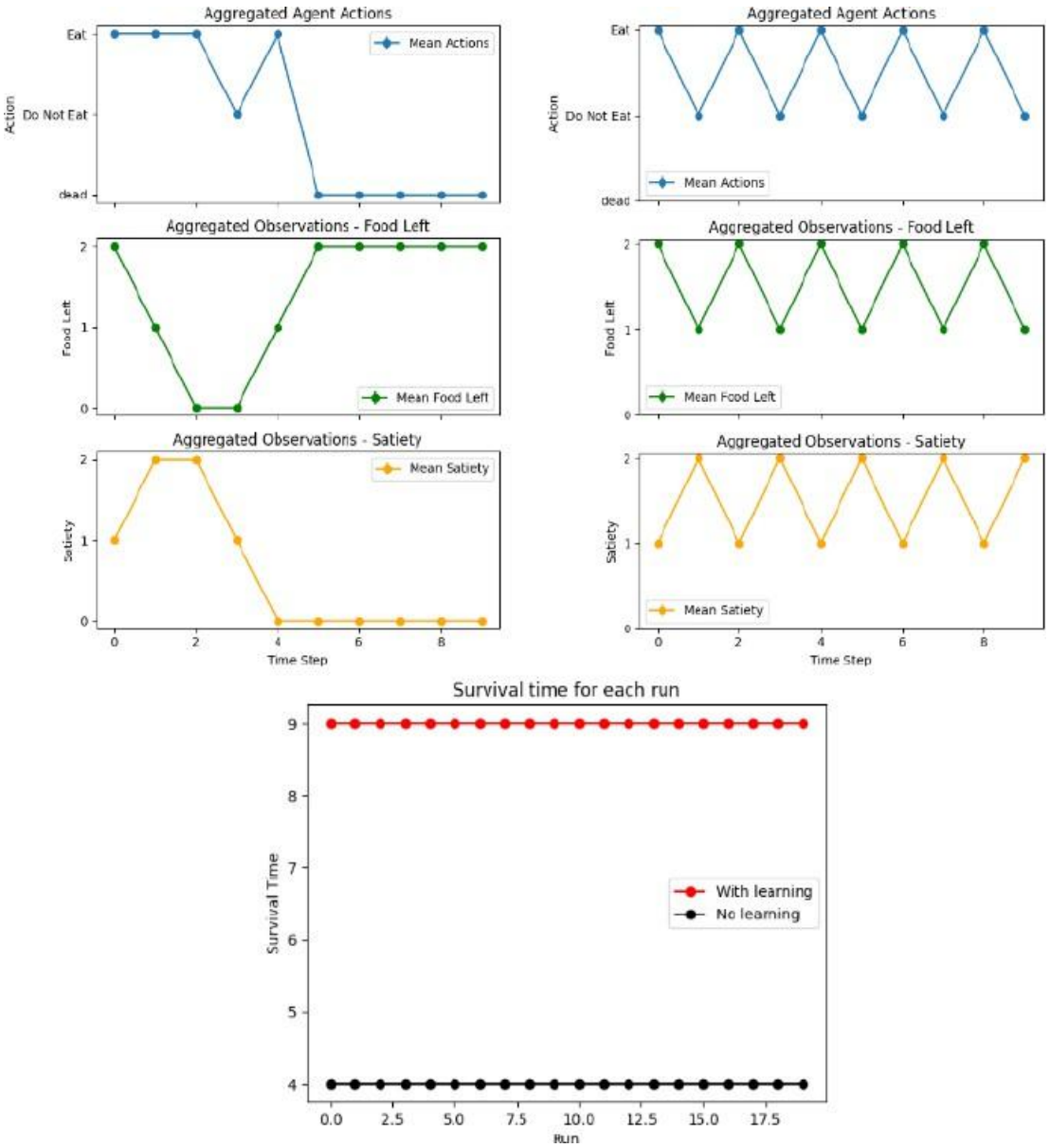

Fig. 12. Risultati della prima estensione del Caso 2, ambiente dinamico, con matrice B impostata con valori estremi. I tre grafici in alto a sinistra illustrano l'andamento della simulazione senza meccanismo di apprendimento, e i tre a destra l'andamento.

Nella seconda estensione del Caso 2, vengono impostate forti preferenze prioritarie su entrambi gli stati (sazietà = 2 e disponibilità di cibo = 2, senza meccanismo di apprendimento). In questa estensione, l'agente inizialmente sceglie di mangiare, dimostrando l'influenza delle forti preferenze sulle sue azioni. Ciò porta l'agente a morire rapidamente nella maggior parte delle simulazioni senza apprendimento (*Fig. 13*, tre

grafici in alto a sinistra e grafico in basso). Con l'apprendimento attivato, l'agente riesce nuovamente a bilanciare le proprie azioni, sopravvivendo più a lungo e trovando un equilibrio tra preferenze e richieste ambientali (*Fig. 13*, tre grafici in alto a destra e grafico in basso). Inoltre, impostando una bassa preferenza per la disponibilità di cibo pari a 0, il comportamento dell'agente rimane pressoché invariato.

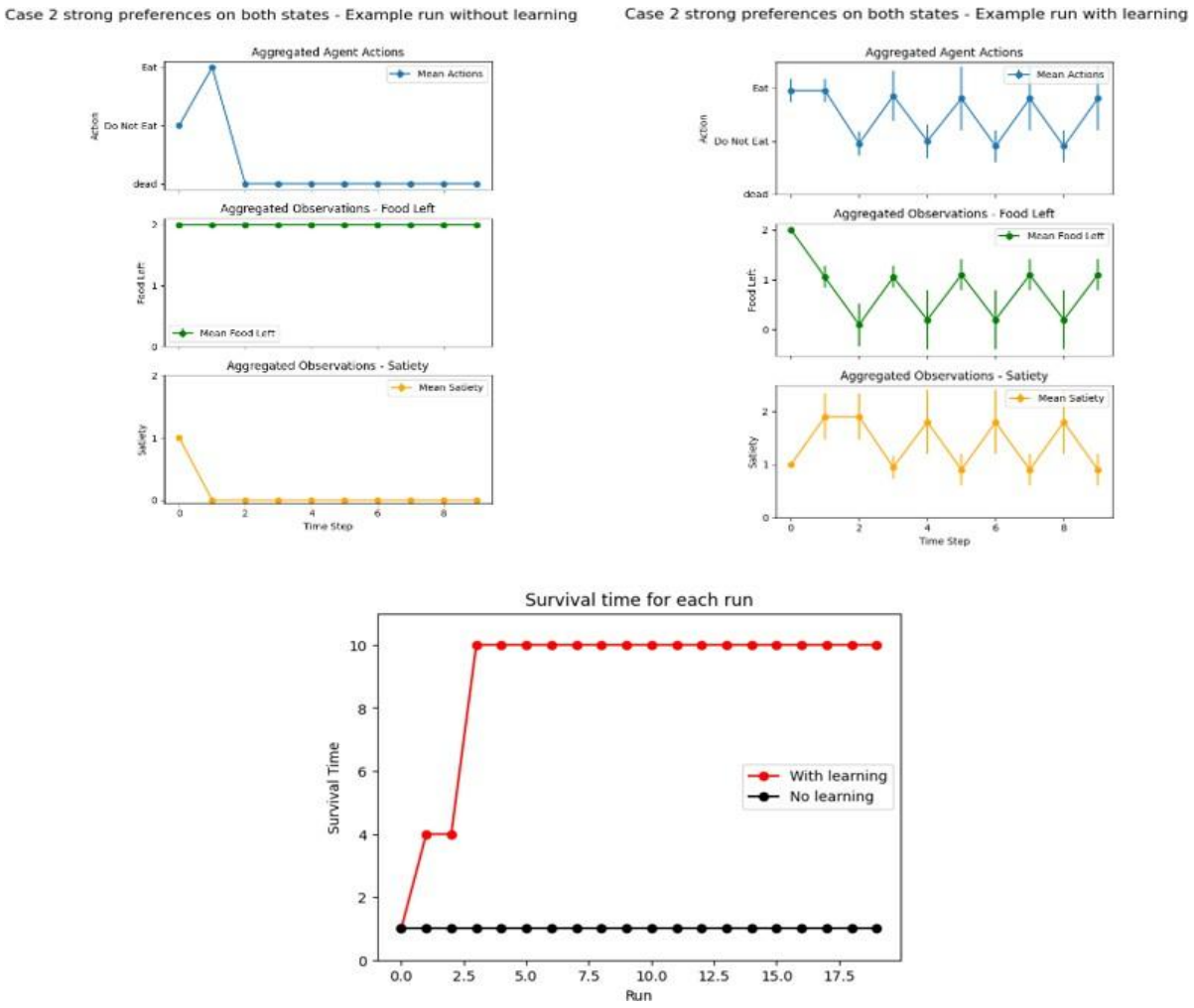

Fig. 13. Risultati della seconda estensione del Caso 2, ambiente dinamico, con valori di preferenze prioritarie alti su entrambi gli stati (cibo disponibile e sazietà). I tre grafici in alto a sinistra illustrano l'andamento della simulazione senza meccanismo di apprendimento, e i tre a destra l'andamento con il meccanismo di apprendimento. Il grafico in basso illustra il tempo medio di sopravvivenza dell'agente per ogni esecuzione, con apprendimento (linea rossa) e senza (linea nera). Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 21.

Nella terza estensione cambia il tasso di variazione dell'ambiente (*Fig. 14*): i livelli di cibo e sazietà aumentano e diminuiscono a velocità diverse rispetto al Caso 2 principale. Qui, le prestazioni dell'agente diminuiscono rispetto al caso principale, ma l'apprendimento fornisce comunque un vantaggio significativo: l'agente, infatti, adatta la propria strategia, mangiando con frequenza minore per conservare le risorse e quindi

mantenendo livelli di cibo più alti in media e gestendo la sazietà in modo più efficace. Questo gli permette di sopravvivere per l'intera durata della simulazione (10 intervalli temporali). Senza il meccanismo di apprendimento, invece, sopravvive solo per circa 3 intervalli temporali. Sebbene le dinamiche ambientali meno regolari aumentino la complessità, costringendo l'agente a pianificare in modo diverso e su orizzonti temporali più lunghi, esso dimostra una notevole capacità di adattare le proprie strategie attraverso l'apprendimento per gestire il nuovo tasso di variazione dei livelli di cibo e sazietà.

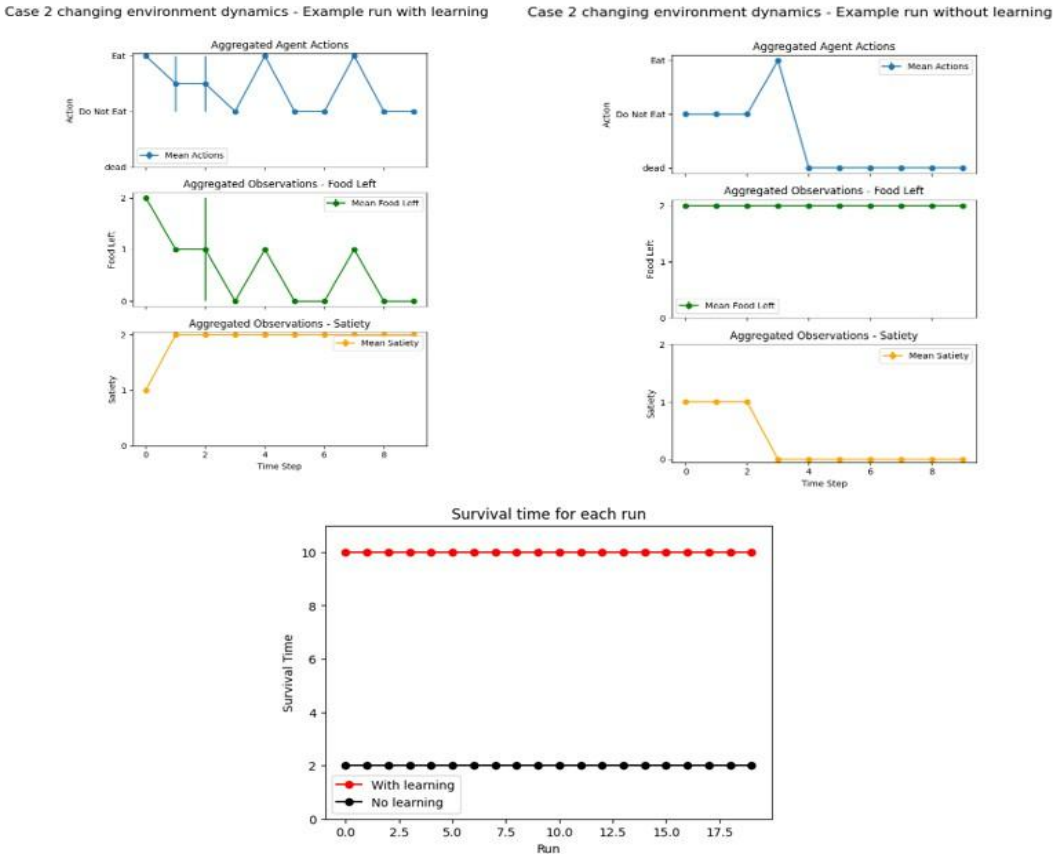

Fig. 14. Risultati dell'estensione del Caso 2, ambiente dinamico, in cui i livelli di cibo e sazietà crescono e diminuiscono con tassi temporali diversi. I tre grafici in alto a sinistra illustrano l'andamento della simulazione senza meccanismo di apprendimento, e i tre a destra l'andamento con il meccanismo di apprendimento. Il grafico in basso illustra il tempo medio di sopravvivenza dell'agente per ogni esecuzione, con apprendimento (linea rossa) e senza (linea nera). Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 22.

Anche la quarta estensione prevede il cambiamento delle dinamiche ambientali. In particolare, vengono simulati cambi stagionali, con il passaggio dall'estate all'inverno. L'agente riesce ad apprendere le dinamiche della prima stagione e ad adeguarvisi per sopravvivere, ma muore dopo il passaggio seconda stagione, poiché non ha abbastanza tempo per apprendere le nuove dinamiche stagionali (*Fig. 15*). In questo caso, sembra che

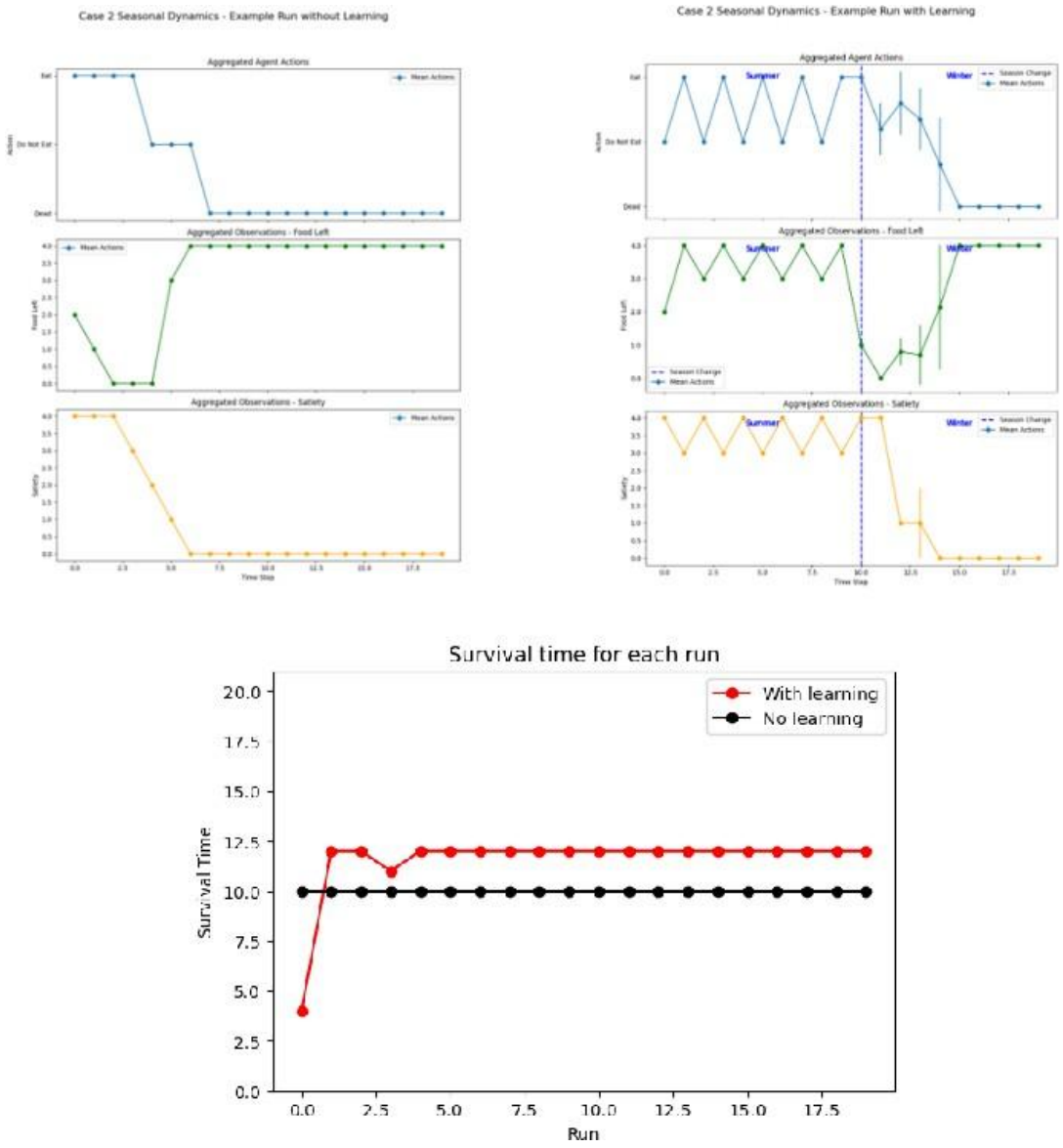

Fig. 15. Risultati dell'estensione del Caso 2, ambiente dinamico, in cui viene simulato l'alternarsi delle stagioni estate e inverno. I tre grafici in alto a sinistra illustrano l'andamento della simulazione senza meccanismo di apprendimento, e i tre a destra l'andamento con il meccanismo di apprendimento. Il grafico in basso illustra il tempo medio di sopravvivenza dell'agente per ogni esecuzione, con apprendimento (linea rossa) e senza (linea nera).

un solo modello non sia sufficiente per gestire due stagioni diverse, poiché per l'agente è difficile apprendere, disimparare e apprendere nuovamente due schemi differenti.

Nell'ultima estensione del Caso 2, infine, è impostato un orizzonte di pianificazione di 1 solo intervallo temporale (mentre nel Caso 2 principale era 3). In questa configurazione, senza meccanismo di apprendimento, l'agente ha prestazioni molto scarse e non sopravvive oltre il primo intervallo temporale. Con il meccanismo di apprendimento attivato, l'agente impiega un po' di tempo per imparare a sopravvivere, ma alla fine ci riesce. Tuttavia, le sue azioni risultano più erratiche e trova una strategia di sopravvivenza, e quindi sostenibile, solo a breve termine (*Fig. 16*). Non riesce quindi

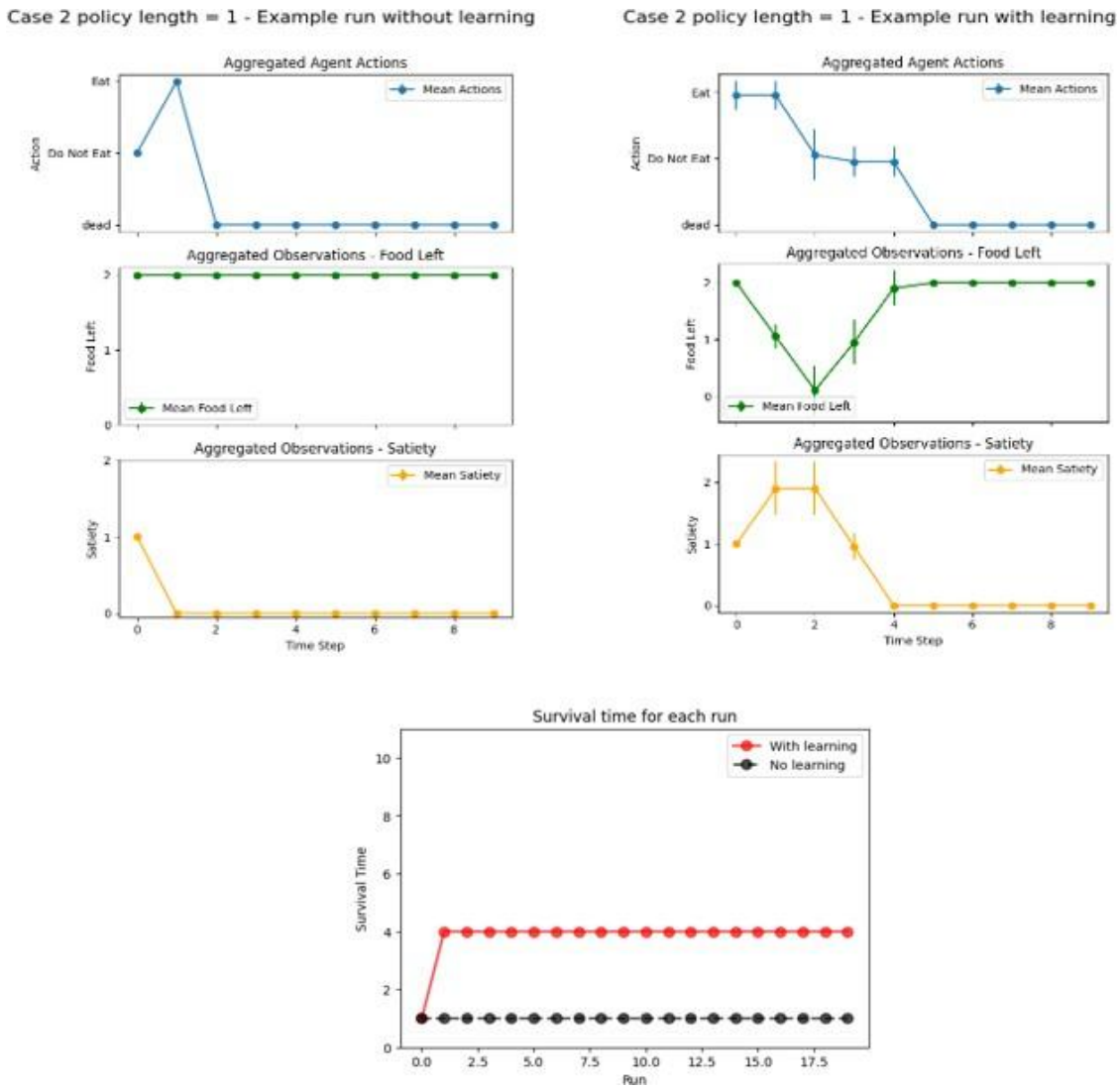

Fig. 16. Risultati dell'estensione del Caso 2, ambiente dinamico, con lunghezza di politiche = 1. I tre grafici in alto a sinistra illustrano l'andamento della simulazione senza meccanismo di apprendimento, e i tre a destra l'andamento con il meccanismo di apprendimento, e i tre a destra l'andamento con il meccanismo di apprendimento. Il grafico in basso illustra il tempo medio di sopravvivenza dell'agente per ogni esecuzione, con apprendimento (linea rossa) e senza (linea nera). Il testo della figura è in inglese in quanto essa è tratta da Albarracin Hipólito Raffa Kinghorn, 2024, 23.

a sopravvivere a lungo, evidenziando la necessità di focalizzarsi su strategie a lungo termine.

Per una panoramica complessiva, nelle tabelle in *Fig. 17* sono riassunti i risultati del Caso base 1 e della relativa estensione (in alto), del Caso 2 principale (in basso a sinistra) e delle relative estensioni (in basso a destra).

Parameter	**Main Case**	**Incorrect A and B matrices**
Agent actions	Consistently chooses to eat at every time step	More erratic actions
Food availability	Remains constant throughout the simulation	Constant
Agent satiety	Increases as it eats and remains at a high level	Fluctuating

Results for Case 1

Parameter	**Main Case**
Agent actions	Initially fluctuates between "eat" and "do not eat". Balances actions with learning
Food availability	Depletes when eaten, replenishes if not consumed. Agent balances food levels with learning
Agent satiety	Initially balances satiety and survival. Stabilizes with learning
Survival time	Agent consistently survives the maximum number of time steps after the initial learning phase

Results for Case 2

1. Incorrect B Matrix		
Initial B matrix setting	Extreme values (0 or 1)	
Agent actions	Consistently chooses to eat, leading to suboptimal behavior and eventual starvation	
Learning	Able to learn and survive by adjusting extreme values with learning	
2. Different preferences		
Preference	Strong preference for both high satiety and high food left	Low preference on no food left (0)
Agent actions	Initially chooses to eat, then stops as food becomes scarce, leading to quick death without learning	Quick death without learning
Learning	Balances actions and survives longer	Balances actions and survives longer
3. Different environment rate of change		
Food rate change	Increases at 0.5 units/step when not eating; decreases at 1 when eating	
Satiety rate change	Decreases at 0.2 unit/step when eating; increases at 0.8 when not eating	
Agent actions	Adaptive with learning; learns to conserve food and maintain satiety	
Survival time	Consistently survives full run time with learning; only around 1/3 time without learning	
3.1 Seasonal simulation		
	Summer	Winter
Food rate change	Increases at 3 when not eating; decreases at 1 when eating	Increases at 1 when not eating; decreases at 3 when eating
Satiety rate change	Increases at 3 when not eating; decreases at 1 when eating	Increases at 1 when eating; decreases at 3 when not eating
Agent actions	Adaptive with learning	
Survival time	Without learning, the agent dies quickly before season shifting. With learning manages to survive throughout summer but quickly dies during winter	
4. Different policy length		
Policy length	Time step = 3 vs Time step = 1	
Agent actions	Performs poorly with 1 time step without learning; more stable but same strategy with learning	
Survival time	Unable to survive long with 1 time step; longer survival with learning	

Results for Case 2 extended

Fig. 17. Tabelle riassuntive dei risultati di tutti gli scenari della simulazione. In alto, Caso base 1, ambiente statico, e relativa estensione con matrici A e B errate. In basso, Caso 2, ambiente dinamico (a sinistra) e relative cinque estensioni (a destra).

4.2 Discussione dei risultati e dei limiti

La simulazione proposta costituisce dunque un modello che dimostra come l'AIF possa essere utile per la gestione sostenibile delle risorse a livello individuale. In essa viene considerata la relazione tra agente e ambiente per sottolineare l'importanza della resilienza, dell'adattabilità e della pianificazione a lungo termine nel raggiungimento di risultati sostenibili. Il comportamento dell'agente si forma tramite l'interazione tra il suo modello del mondo, le preferenze a priori e le dinamiche ambientali. Esso cerca di ottimizzare i bisogni immediati (la fame) e i risultati a lungo termine (la disponibilità costante di cibo), imparando a bilanciare il consumo e il ripristino delle risorse per promuovere la sostenibilità del proprio sistema interno. I risultati sono in linea con la formalizzazione della sostenibilità, resilienza e benessere all'interno dello schema dell'AIF di Albarracin (Ramstead Pitliya, 2024).

Nello specifico, nel Caso base 1, in un ambiente statico l'agente mostra inerzia, poiché mantiene sempre il medesimo modello di consumo senza considerare la disponibilità a lungo termine delle risorse. Sebbene questo comportamento sia adeguato al contesto dato, manca della flessibilità necessaria per raggiungere risultati sostenibili in ambienti più dinamici. Nel Caso 2 è stata quindi introdotta variabilità ambientale, che ha richiesto all'agente di mettere in campo elasticità e plasticità. L'elasticità è esemplificata dalla capacità dell'agente di adattare le proprie abitudini alimentari alla disponibilità variabile di cibo: esso, infatti, è stato in grado di sopportare temporaneamente un aumento dell'energia libera (ovvero la fame) per raggiungere una stabilità a lungo termine. La capacità dell'agente di apprendere e aggiornare il proprio modello del mondo basandosi su nuove informazioni, riflette poi la plasticità e si coniuga con il miglioramento della resilienza di fronte ai cambiamenti ambientali: l'agente, infatti, adatta le proprie azioni per mantenere il benessere in condizioni di disponibilità di risorse variabili. Il legame dinamico tra agente e ambiente nello studio della gestione sostenibile delle risorse si dimostra dunque cruciale a livello di un singolo agente, poiché le azioni dell'agente hanno ottimizzato il suo benessere e hanno contribuito alla resilienza dell'ambiente, prevenendo l'esaurimento completo delle risorse. Questo rapporto reciproco tra agente e ambiente si costituisce quindi come un aspetto fondamentale della sostenibilità.

Nelle estensioni del Caso 2, sono emersi problemi legati all'inerzia con la possibilità che anche gli agenti adattivi rimangano bloccati in politiche difficili da

modificare. Le estensioni relative alle diverse preferenze a priori evidenziano un'interessante area di ricerca sull'origine di tali preferenze. Si può ipotizzare ragionevolmente che gli agenti evolvano sviluppando una forte preferenza per evitare la morte. Tuttavia, non è plausibile che un agente sviluppi una preferenza specifica per la quantità di cibo da lasciare nell'ambiente, poiché questa dipende fortemente dalle caratteristiche dell'ambiente stesso. Sarebbe inefficiente, ad esempio, avere una preferenza a priori per non consumare cibo disponibile se tale risorsa può rigenerarsi rapidamente. Pertanto, la sopravvivenza di un agente dipende dalla sua capacità di adattarsi all'ambiente e di agire in modo sostenibile, rispondendo alle esigenze specifiche del contesto in cui opera. Dalla simulazione emerge dunque che agire in modo sostenibile non è semplicemente una scelta etica o ideologica, ma una strategia necessaria per garantire la propria sopravvivenza.

Occorre sottolineare anche che la simulazione implementata, come visto, è piuttosto semplice e dunque presenta importanti limitazioni. In particolare, la configurazione utilizzata, che prevede un solo agente e una sola risorsa, non è in grado di rappresentare le complessità che caratterizzano i sistemi reali. Nella realtà, infatti, le risorse non sono isolate ma interconnesse, con interdipendenze che influenzano il comportamento del sistema nella sua totalità. Inoltre, i circuiti di *feedback* – ovvero i processi attraverso i quali le azioni di un agente modificano l'ambiente, influenzando poi le scelte future dello stesso agente – sono qui assenti o semplificati e riducono quindi la capacità del modello di rappresentare dinamiche adattive più sofisticate.

Cruciale è anche il fatto che, in un modello individuale come quello descritto, l'agente non può tramandare direttamente le informazioni acquisite, se muore prima di aver completato il processo di apprendimento. Tuttavia, il modello potrebbe prevedere diverse strategie per superare questa limitazione: se si considerassero più agenti successivi, ad esempio, si potrebbe introdurre un meccanismo di eredità delle credenze, dove un nuovo agente inizia con la matrice B appresa dall'agente precedente invece di ripartire da zero, così da avere persistenza delle informazioni. O ancora, con l'implementazione di una sorta di memoria collettiva, più agenti potrebbero condividere conoscenza attraverso un meccanismo di scambio di informazioni sulle probabilità di transizione. Infine, se si riuscisse a migliorare il tempo di sopravvivenza, con un tasso di aggiornamento delle credenze sufficientemente basso, l'agente potrebbe consolidare una conoscenza utile prima della morte. Anche nel caso di morte dell'agente, un

aggiornamento lento della matrice *B* può permettere al sistema di mantenere una coerenza nell'apprendimento nel caso in cui esista un meccanismo di riavvio con eredità, come ad esempio nel caso di un modello evolutivo. In questo caso, ogni nuova generazione potrebbe inizializzare la propria matrice *B* non da zero, ma da una versione che incorpora una piccola frazione delle conoscenze dei predecessori. Questo consente un apprendimento cumulativo senza che gli individui debbano ogni volta ripartire da zero.

Un aggiornamento graduale della matrice *B* risulta vantaggioso anche perché permette una migliore generalizzazione dell'apprendimento, in quanto l'agente può raccogliere più dati nel tempo e ottenere una rappresentazione più accurata delle tendenze globali dell'ambiente. Infatti, se ogni esperienza modifica drasticamente la matrice *B*, l'agente ha difficoltà a sviluppare modelli predittivi robusti. Questo si sovrappone con il rischio di *overfitting*: l'agente, infatti, potrebbe specializzarsi troppo sulle specifiche dinamiche ambientali simulate, imparando a ottimizzare il proprio comportamento solo per quello scenario particolare. L'eccessiva specializzazione potrebbe limitare la sua capacità di adattarsi a nuove situazioni o a condizioni ambientali differenti. In altre parole, un agente troppo adattato a un ambiente simulato potrebbe mostrare difficoltà ad affrontare contesti reali più complessi o variabili, dove le dinamiche delle risorse e le interazioni tra agenti richiedono maggiore flessibilità e capacità di generalizzazione.

Come si è visto, inoltre, le preferenze a priori imprecise possono influenzare in modo significativo il comportamento iniziale dell'agente e la sua traiettoria di apprendimento. Questo implica la necessità di calibrare con attenzione tali preferenzequando il modello viene applicato a scenari reali, per garantire risultati coerenti e affidabili. Inoltre, alcune impostazioni artificiali adottate nelle simulazioni, come l'introduzione deliberata di matrici errate, pur essendo utili per testare la robustezza del modello, potrebbero non rappresentare adeguatamente le sfide che si presentano in contesti più complessi e realistici.

Infine, un altro limite rilevante riguarda il fatto che il modello non è stato concepito per affrontare direttamente il problema del costo computazionale associato all'AIF. Di conseguenza, non fornisce dati concreti sull'efficienza energetica o computazionale dell'AIF rispetto ad approcci alternativi, come l'RL. Questo rappresenta una possibile direzione di ricerca cruciale per comprendere appieno il potenziale sostenibile di AIF.

Ci si propone di affrontare i limiti della simulazione attraverso una serie di sviluppi futuri, dettagliati nel prossimo paragrafo.

4.3 Prossimi sviluppi

Sebbene la semplicità del modello presentato ne limiti l'applicabilità diretta ai sistemi reali, esso fornisce una base per future ricerche che esplorino la gestione sostenibile delle risorse attraverso la lente dell'AIF.

In primo luogo, la ricerca futura punta a esplorare scenari con molteplici agenti che abbiano interessi in competizione e risorse condivise, così come ambienti con risorse interconnesse e dinamiche di rigenerazione più sofisticate. Integrare interazioni tra più agenti con obiettivi e priorità diverse permetterebbe di simulare dinamiche complesse, come cooperazione, conflitti e competizione per risorse condivise. In parallelo, sarebbe necessario introdurre più tipologie di risorse interconnesse, come acqua ed energia, per riflettere meglio la complessità ecologica e socioeconomica dei sistemi reali. Per rappresentare queste interdipendenze, come suggerito da Albarracin (Pitliya Ramstead, 2024) si potrebbe ricorrere alla teoria delle reti, modellando risorse e agenti come nodi connessi all'interno di un sistema dinamico e analizzando le interazioni e i *feedback* attraverso strumenti quantitativi avanzati.

In un ambiente reale, infatti, la gestione delle risorse non avviene in isolamento, ma coinvolge più agenti con interessi diversi e potenzialmente in competizione. Come visto nel paragrafo precedente, un primo problema che potrebbe emergere in un contesto con agenti molteplici riguarda la distribuzione e la condivisione di informazioni. Friston e Christopher Frith (2015a, 2015b) hanno evidenziato come l'apprendimento cooperativo tra agenti può portare a una maggiore efficienza nella gestione delle risorse, riducendo gli errori di previsione e favorendo il coordinamento tra individui. Questo processo può avvenire tramite diversi meccanismi, come la sincronizzazione dei modelli generativi, la memoria collettiva e l'apprendimento cooperativo. La sincronizzazione avviene nel momento in cui tutti gli agenti, ciascuno dotato del medesimo modello generativo, iniziano con una rappresentazione simile dell'ambiente e aggiornano progressivamente le loro credenze tramite interazione reciproca, in modo che il sistema converga verso un comportamento più stabile e coordinato: è questa la cosiddetta sincronizzazione generalizzata tramite AIF, dove agenti che minimizzano il proprio

errore predittivo tendono a sviluppare rappresentazioni condivise della realtà (Friston Frith, 2015b). Il meccanismo della memoria collettiva prevede invece l'integrazione di una memoria comune all'interno del sistema, che potrebbe permettere agli agenti di accumulare esperienza e trasmetterle alle generazioni successive, o ad altri membri del sistema, migliorando così l'adattabilità collettiva. Ciò si allinea con gli studi sull'apprendimento percettivo e sulla plasticità sinaptica, che mostrano come l'accumulo di informazioni nel tempo possa ottimizzare le strategie decisionali degli agenti (Friston Frith, 2015b). L'apprendimento cooperativo, infine, funziona attraverso l'incorporamento di meccanismi di comunicazione implicita, in cui gli agenti si influenzano reciprocamente aggiornando le proprie credenze in modo distribuito, senza necessità di trasmissione esplicita dei dati (Friston Frith, 2015a)

Nel modello attuale, l'agente apprende in modo indipendente, basandosi esclusivamente sulle proprie osservazioni. Nel momento in cui fossero presenti più agenti, bisognerebbe aggiungere un meccanismo di condivisione della conoscenza – ad esempio, appunto, una memoria collettiva o strategie di apprendimento cooperativo – in modo da valutare se il sistema diventi più efficiente o meno nel tempo. Per migliorare la scalabilità del modello, si potrebbe valutare l'effetto di un meccanismo di comunicazione tra agenti per trasmettere informazioni sulle strategie ottimali.

La scalabilità del modello si riflette poi sulla gestione computazionale delle politiche di apprendimento. Infatti, nella simulazione allo stato attuale, l'energia libera è calcolata su un singolo agente, laddove in un sistema più complesso potrebbe aumentare in modo non lineare con l'aggiunta di nuovi agenti e variabili ambientali. Un'estensione naturale della simulazione sarebbe quindi analizzare il rapporto tra numero di agenti, disponibilità delle risorse e stabilità del sistema, per determinare se l'approccio proposto rimanga efficace su scala maggiore. Un'opzione per mitigare il problema del dispendio computazionale potrebbe essere l'introduzione di strategie euristiche o di apprendimento distribuito per ridurre la complessità del calcolo. Poiché in uno scenario realistico la sostenibilità dipende dalla gestione simultanea di più risorse interconnesse, come acqua, energia e materie prime, si potrebbe estendere il modello a più risorse con dinamiche di rigenerazione complesse, verificando se gli agenti possano apprendere strategie di ottimizzazione più generali che non si limitino a un singolo tipo di risorsa.

Inoltre, il modello attuale non considera la possibilità che le risorse si esauriscano in modo permanente, il che richiederebbe di condizionare la sopravvivenza

dell'ambiente sul mantenimento di determinati valori di risorse. In futuro, sarà necessario integrare questo aspetto per comprendere le implicazioni a lungo termine delle strategie di gestione delle risorse.

È importante poi contestualizzare l'approccio della simulazione illustrata nel panorama più ampio della modellazione per la gestione sostenibile delle risorse. Rispetto agli approcci tradizionali, come l'RL basato su modelli o il controllo predittivo basato su modelli, come si è visto l'AIF presenta vantaggi nella sua capacità di bilanciare i bisogni a breve termine con la stabilità a lungo termine. Tuttavia, potrebbe risultare meno efficiente in scenari che richiedono un'ottimizzazione rapida o in ambienti con dinamiche altamente deterministiche, dove modelli più semplici potrebbero essere sufficienti. In futuro si dovrebbero quindi condurre analisi comparative complete per comprendere meglio i punti di forza e i limiti dell'AIF in diversi contesti di sostenibilità, esplorando potenziali approcci ibridi che sfruttino i vantaggi di più paradigmi di modellazione. Implementare il modello in progetti pilota con dati reali sul campo, inoltre, potrebbe permettere di testarne l'applicabilità in situazioni reali, identificando eventuali limiti e opportunità di miglioramento.

Le analisi comparative sarebbero utili anche nell'ottica di un'indagine sul costo dell'AIF, ovvero sul consumo computazionale, sulla scalabilità e sull'efficienza energetica. L'integrazione di metriche standardizzate per misurare il consumo di risorse durante l'addestramento e l'esecuzione dei modelli potrebbe fornire dati quantitativi utili per dimostrare la sostenibilità computazionale dell'AIF. Inoltre, esplorare configurazioni di *hardware* ottimizzati e strategie di apprendimento che riducano la complessità computazionale del modello potrebbe contribuire a dimostrare che l'AIF può essere un paradigma non solo sostenibile in termini di risultati, ma anche di risorse. Ottimizzare la precisione o i tassi di apprendimento potrebbe anche favorire la resilienza elastica e plastica necessaria per garantire sostenibilità e abbondanza a lungo termine.

Inoltre, come è emerso dall'estensione del Caso 2 sulle dinamiche stagionali, un problema dell'AIF, così come implementata, è che l'agente impiega molto tempo per riaddestrarsi e adattarsi quando l'ambiente cambia, soprattutto se il modello è stato appreso per un lungo periodo e presenta un alto grado di certezza. Potrebbe quindi essere interessante esplorare l'apprendimento di modelli discreti differenti per ambienti e stagioni specifici (Collis Singh Kinghorn, 2024; Friston Da Costa Tschantz, 2023).

Potrebbe valere la pena anche di adattare il modello a dati reali, in modo da

validarne l'applicabilità e l'efficacia. Simulando la gestione sostenibile delle risorse in contesti reali, infatti, sarebbe possibile inquadrare meglio gli obiettivi più urgenti e perfezionare il modello per offrire indicazioni operative utili a imprese e decisori politici. Per muoversi in tale direzione, una delle prime operazioni dovrebbe consistere nell'integrare il modello con dataset reali, identificando fonti affidabili e rappresentative, come ad esempio il database globale sugli OSS delle Nazioni Unite[45] o progetti locali, come quelli promossi dai comitati per la sostenibilità. Tali dati, che possono includere informazioni su risorse naturali, consumi energetici, dinamiche economiche e indicatori demografici, necessitano di una preelaborazione accurata per garantire che siano puliti, normalizzati e privi di anomalie che potrebbero distorcere i risultati della simulazione. Una volta ottenuti, i dati reali possono essere utilizzati per aggiornare le matrici del modello, sostituendo le probabilità statiche con distribuzioni derivate direttamente dai dati osservati, rendendo così il modello più rappresentativo delle relazioni causa-effetto presenti nel sistema reale. In secondo luogo, le dinamiche ambientali del modello devono essere rese più realistiche, incorporando ad esempio tassi di rinnovo o esaurimento delle risorse basati su dati empirici, oltre a dinamiche stagionali e impatti del cambiamento climatico. È importante simulare anche shock ambientali, come disastri naturali o crisi delle risorse, per testare la resilienza dell'agente e la sua capacità di adattarsi a scenari imprevedibili. Tali aggiustamenti permetterebbero di osservare il comportamento del modello in condizioni di maggiore complessità, avvicinandolo alle sfide che i sistemi reali affrontano quotidianamente. Come visto, inoltre, le preferenze a priori dell'agente rappresentano un elemento critico per adattare il modello ai dati reali e devono essere calibrate con attenzione per garantire che riflettano comportamenti e priorità realistiche. Per fare ciò, si possono utilizzare studi comportamentali, economici ed ecologici, o ricorrere a un processo iterativo in cui le preferenze vengono affinate attraverso simulazioni ripetute e apprendimento dai dati osservati. Questa calibrazione consentirebbe di assicurare che il modello mantenga coerenza con gli obiettivi reali del sistema e che possa produrre risultati validi e interpretabili. Garantire la trasparenza e l'affidabilità delle decisioni del modello è infatti fondamentale per favorire il

[45] S veda il §1.1, nota 5.

coinvolgimento di *stakeholders* e promuovere l'accettazione sociale. Inoltre, l'impatto delle politiche simulate deve essere valutato attentamente per assicurare che promuovano equità e sostenibilità a lungo termine.

Un altro avanzamento potrebbe essere rappresentato dall'integrazione dell'AIF con gli LLM. Come visto nel §2.3, questi modelli, grazie alla loro capacità di elaborare grandi quantità di dati, possono fornire spunti significativi per ottimizzare le strategie decisionali dell'agente. Inoltre, tecniche di apprendimento avanzate, come il DL, possono migliorare la capacità del modello proposto di generalizzare e adattarsi a dati complessi e dinamici, rafforzando la sua applicabilità in scenari reali. Gli LLM, analogamente ai modelli basati su FEP, utilizzano architetture generative preaddestrate per svolgere compiti come la generazione di testi, la comprensione e la predizione del contesto minimizzando gli errori nella previsione delle sequenze di parole successive (Zu Su He, 2024). Anche se non sono progettati esplicitamente per modellare l'incertezza come i sistemi biologici, il loro comportamento risulta coerente con il PP. Proprio come il cervello adatta le proprie predizioni basandosi sui dati sensoriali in arrivo per ridurre gli errori di previsione, anche gli LLM regolano le previsioni sulle parole successive basandosi su una grande quantità di dati pregressi, per generare output contestualmente appropriati. Essi sono inoltre versatili ed efficienti, in quanto e modelli di uso generale che non richiedono un riaddestramento energeticamente dispendioso per ogni nuovo compito. Tale flessibilità riduce la necessità di modelli specializzati, permettendo il risparmio di risorse computazionali e umane. Grazie alla loro capacità di operare in domini diversi, gli LLM si rivelano strumenti preziosi per la risoluzione di problemi generali, mantenendo al contempo un'efficienza energetica in linea con gli obiettivi di sostenibilità (Raffa Acciai, di prossima pubblicazione).

Esistono già esempi concreti della fruttuosa relazione tra AIF e LLM: l'AIF è stata utilizzata per migliorare la precisione e la rilevanza delle risposte fornite dagli LLM rispetto al contesto, guidando lo sviluppo di modelli più accurati e pertinenti. Ciò è avvenuto in ambito medico, con terapie supportate da LLM, in cui i principi dell'AIF sono utilizzati per ottimizzare l'efficacia degli interventi. In questi sistemi, un "agente terapeuta" risponde alle domande dei pazienti, mentre un "agente supervisore" ne valuta la veridicità e l'affidabilità. Tale approccio sfrutta l'AIF per ridurre progressivamente gli errori predittivi, migliorando così la qualità dei consigli generati dagli LLM in contesti

complessi, come nel trattamento dell'insonnia (Shusterman Waters O'Neill, 2023). Un altro esempio interessante si trova nel campo dell'istruzione, dove la combinazione di AIF e LLM facilita la simulazione di esperienze di apprendimento più attive. LLM potenziati con AIF sono stati integrati in classi che seguono approcci montessoriani per facilitare le interazioni, aiutare gli studenti a formulare ipotesi, testarle e ridurre gli errori predittivi. Questo approccio ibrido enfatizza l'esplorazione e il coinvolgimento con ambienti materiali, in accordo con lo schema di elaborazione predittiva che può essere considerato un modello dell'apprendimento umano (Di Paolo White Guénin-Carlut, 2024).

Preso atto delle analogie messe in luce finora, potenzialmente significative per l'approfondimento dell'integrazione tra AIF e LLM, bisogna considerare anche le importanti differenze tra le due strutture. Nonostante abbiano come scopo comune la minimizzazione degli errori predittivi, infatti, ciò avviene con modalità differenti: l'AIF, come visto, si basa sull'inferenza bayesiana esplicita, mentre gli LLM utilizzano l'apprendimento statistico. Inoltre, come sottolineato da Giovanni Pezzulo e colleghi (2024), i sistemi basati su AIF esplorano attivamente l'ambiente, aggiornando costantemente le loro predizioni in base alle interazioni con il mondo, mentre gli LLM sono modelli passivi che generano predizioni basate su dati preesistenti, senza interagire con un ambiente dinamico. Ciò rende gli LLM potenti per l'elaborazione del linguaggio, ma privi delle caratteristiche adattive e guidate dall'ambiente proprie dell'AIF. Non si possono inoltre trascurare le sfide alla sostenibilità poste dai LLM, in particolare il significativo consumo energetico che essi comportano durante la fase iniziale di addestramento. Come è stato visto, i sistemi basati sull'inferenza bayesiana – inclusi quelli basati sull'AIF – diventano esponenzialmente più impegnativi dal punto di vista computazionale man mano che il numero di variabili aumenta (Kwisthout van Rooij, 2020). Lo stesso vale per gli LLM, che richiedono risorse computazionali significative. Mitigare la domanda energetica rimane una sfida cruciale per lo sviluppo futuro di un'IA sostenibile.

Tutto ciò considerato, sebbene LLM e AIF differiscano nel loro approccio alla gestione dell'incertezza e all'interazione con l'ambiente, gli LLM mostrano comunque caratteristiche che li rendono candidati validi come CA per IA sostenibili e la loro

integrazione con modelli basati sull'AIF può essere considerata una via che vale la pena esplorare.

Adottando le strategie illustrate, la semplice simulazione basata sull'AIF descritta in questo capitolo può essere ampliata e adattata per affrontare le sfide reali legate alla gestione sostenibile delle risorse. Questo approccio offre strumenti preziosi per supportare decisioni basate sui dati e promuovere soluzioni per un'IA sostenibile ed efficace, con applicazioni che vanno oltre l'ambito accademico, rappresentando anche una risorsa pratica per le aziende impegnate nel raggiungimento dei propri obiettivi di sostenibilità.

CONCLUSIONE

Il lavoro condotto durante il mio percorso dottorale, sintetizzato in questa tesi, si è proposto di rispondere a tre domande di ricerca, affrontandole attraverso le prospettive integrate della filosofia della mente, delle scienze cognitive e dell'IA.

La prima riguarda il modo in cui l'IA può integrarsi con i principi di sostenibilità, un tema di cruciale importanza nel panorama scientifico e sociale attuale. Ciò implica un'esplorazione delle modalità con cui i sistemi di IA possono essere progettati per rispettare la sostenibilità ambientale e sociale, analizzando le implicazioni legate all'uso di queste tecnologie, come i potenziali impatti ecologici dei processi computazionali ad alta intensità energetica e i rischi sociali connessi. Questo interrogativo costituisce la base per delineare i criteri con cui le IA sostenibili devono essere progettate, collegandosi direttamente alle successive domande di ricerca.

La seconda domanda, infatti, si concentra sul ruolo dei modelli cognitivi, con un focus sul FEP e sull'AIF, per tradurre i principi di sostenibilità in criteri operativi e progettuali. Tali modelli, ispirati ai processi biologici, forniscono una solida base teorica per sviluppare sistemi artificiali che siano sostenibili sia dal punto di vista ambientale, ottimizzando il consumo delle risorse computazionali e la velocità dei processi, sia dal punto di vista sociale, rendendo le IA più spiegabili, trasparenti ed eticamente affidabili.

Infine, la terza domanda rappresenta la verifica pratica delle due precedenti, poiché riguarda come validare l'integrazione tra il modello cognitivo dell'AIF e l'IA in contesti concreti, traducendo le soluzioni teoriche in applicazioni. Questo passaggio è cruciale per verificare se i modelli proposti siano effettivamente in grado di affrontare le sfide della sostenibilità quando applicati a scenari dinamici, e valutare quali siano le metodologie più adeguate per testare l'efficacia di tali sistemi.

Le tre domande di ricerca, dunque, interconnesse tra loro, delineano un percorso che parte dalla comprensione dei principi di sostenibilità applicabili all'IA, passa per l'identificazione di modelli cognitivi adeguati e giunge alla loro validazione in contesti concreti. Per affrontare tali temi, la ricerca ha adottato un approccio interdisciplinare, volto non solo a colmare le lacune tra discipline che, pur trattando oggetti di ricerca simili, spesso risultano frammentate e distanti, ma anche a creare un terreno comune di dialogo, dimostrando come linguaggi e metodologie diverse possano intersecarsi e arricchirsi

reciprocamente. In questo contesto, il FEP e l'AIF emergono come modelli teorici unificatori di particolare valore, in grado di fornire una base condivisa per comprendere e sviluppare IA sostenibili, integrando diverse prospettive e promuovendo una visione sistemica e collaborativa.

Gli obiettivi principali di questo lavoro, dunque, sono stati fornire una riflessione critica e interdisciplinare sui limiti e sulle opportunità legati all'utilizzo dei modelli cognitivi per un'IA sostenibile dal punto di vista ambientale e sociale, promuovendo il dialogo tra filosofia della mente, scienze cognitive e IA; esplorare le potenzialità dell'AIF come paradigma per un'IA sostenibile; dimostrare tramite una simulazione la capacità di un agente basato sull'AIF di bilanciare bisogni immediati e sostenibilità a lungo termine in contesti dinamici.

Per perseguire tali obiettivi, si è iniziato delineando il contesto entro cui si colloca la ricerca, esplorando la relazione tra IA e sostenibilità e discutendo il ruolo dell'IA nella promozione di pratiche sostenibili. Questa analisi ha permesso non solo di mettere in evidenza le opportunità offerte dall'IA, ma anche di comprenderne i limiti e le criticità, come l'elevato consumo energetico e le implicazioni sociali.

Successivamente, è stato affrontato il concetto di spiegabilità come mezzo per un'IA socialmente sostenibile, mostrando come esso rappresenti uno strumento fondamentale per garantire trasparenza e interpretabilità nei sistemi intelligenti. Si è poi esplorato il ruolo che i modelli cognitivi possono giocare nella creazione di sistemi più interpretabili e affidabili, sia dal punto di vista sociale sia ambientale. A questo punto, è stato introdotto il FEP come quadro teorico esemplare, che unifica la modellazione di sistemi cognitivi biologici e artificiali che siano sostenibili. Si sono poi approfonditi il FEP e l'AIF, mettendo in luce il loro potenziale per implementazioni di IA sostenibili. È stato visto come questi approcci minimizzino la discrepanza tra aspettative e realtà, ottimizzando il consumo energetico e migliorando l'efficienza delle risorse. Inoltre, un confronto con l'apprendimento per rinforzo ha permesso di evidenziare i vantaggi dell'AIF nell'integrare esigenze immediate e obiettivi a lungo termine.

Infine, i concetti teorici sono stati applicati a una simulazione pratica, dove si è dimostrato che un agente basato sull'AIF è in grado di poter bilanciare i propri bisogni immediati con la sostenibilità a lungo termine in un ambiente dinamico. I risultati hanno validato l'efficacia del modello, pur evidenziandone alcuni limiti, come la semplicità

dell'ambiente simulato e la mancanza di dati reali. Tra le future piste di ricerca, si è suggerito di confrontare i risultati con contesti più realistici, di condurre analisi comparative con alti approcci e di integrare modelli di apprendimento profondo avanzati.

Nel complesso, la ricerca ha dunque progressivamente raggiunto gli obiettivi prefissati, esplorando il potenziale teorico e pratico dell'AIF per un'IA sostenibile, offrendo parallelamente una riflessione interdisciplinare sulle sue opportunità e sui suoi limiti. Inoltre, l'analisi del FEP e dell'AIF condotta attraverso le diverse lenti della filosofia della mente, delle scienze cognitive e dell'IA ha contribuito a chiarire concetti che spesso risultano oscuri per i filosofi e troppo impliciti per gli esperti di IA, sottolineando la necessità di maggiore chiarezza e dialogo interdisciplinare.

Per queste ragioni, la ricerca è di interesse per una pluralità di destinatari, ciascuno con interessi specifici e differenti necessità. Per gli studiosi di filosofia della mente, essa offre una chiarificazione dei concetti di FEP e AIF, spesso trattati in modo ambiguo nella letteratura, con lo scopo di fornire una base teorica solida per esplorare come i modelli cognitivi possano ispirare lo sviluppo di IA etiche e sostenibili, promuovendo così un dialogo tra filosofia e tecnologia. Per gli studiosi di scienze cognitive, l'elaborato apre nuove prospettive sperimentali, inserendosi nel filone emergente che evidenzia il potenziale dell'AIF applicata per comprendere meglio i meccanismi di apprendimento e adattamento umano e per indagare analogie e differenze rispetto ai sistemi artificiali. Gli esperti di IA possono inoltre rintracciare in questo lavoro un approccio alternativo ai modelli tradizionali, che enfatizza la sostenibilità ambientale e sociale come principi fondamentali: l'AIF emerge, in tal senso, come uno strumento efficace per bilanciare bisogni immediati e obiettivi a lungo termine, tramite l'ottimizzazione delle risorse, e come modello artificiale potenzialmente spiegabile rispetto a sistemi più complessi. Infine, per decisori politici e imprenditori, il lavoro propone soluzioni innovative che coniughino progresso tecnologico e responsabilità ambientale e sociale, offrendo simulazioni che possono fungere da base per politiche mirate e pratiche aziendali orientate alla sostenibilità.

Nonostante il tentativo di perseguire gli obiettivi di ricerca nel modo più completo possibile, il lavoro di tesi presenta alcune limitazioni, dettate principalmente dalla scelta consapevole di mantenere la trattazione snella e accessibile, così da renderla fruibile da un pubblico più ampio e diversificato. L'intersezione tra IA e sostenibilità è stata solo

delineata, considerata l'impossibilità di esplorare in modo esaustivo tutte le strategie e applicazioni esistenti, per le quali si rimanda ai riferimenti citati in bibliografia. Le implicazioni della sostenibilità economica, inoltre, non sono state affrontate direttamente, per via della necessità di mantenere il focus sugli aspetti più rilevanti e coerenti con le competenze e gli ambiti esplorati nel presente lavoro. I passaggi storici e filosofici relativi allo sviluppo del concetto di CA sono stati trattati in modo selettivo, concentrandosi sugli snodi principali e funzionali all'argomentazione. Allo stesso modo, il FEP e l'AIF sono stati analizzati ad alto livello, perlopiù evitando dettagli tecnici e matematici al fine di mantenere il focus sui nessi fondamentali che collegano i campi esplorati. Infine, pur rappresentando un valido strumento esplorativo, la simulazione proposta è stata sviluppata in un ambiente virtuale semplificato, caratterizzato dalla presenza di un solo agente e una sola risorsa, benché si sia consapevoli che tale approccio non consenta di catturare la complessità delle dinamiche reali, in cui più agenti competono per risorse limitate e interdipendenti. Inoltre, non sono stati affrontati i costi computazionali legati all'implementazione dell'AIF, né è stato possibile utilizzare dati reali per validare il modello. Questi aspetti, approfonditi nella parte finale del lavoro, costituiscono tuttavia direzioni promettenti per future ricerche, alcune delle quali già in corso di sviluppo.

In sintesi, il presente lavoro di ricerca rappresenta un tentativo di integrare conoscenze teoriche e pratiche provenienti da diversi campi disciplinari per promuovere lo sviluppo di IA che non siano solo performanti, ma anche sostenibili dal punto di vista ambientale e sociale. L'intersezione tra IA e sostenibilità è infatti una sfida tecnologica, filosofica e sociale che richiede un approccio integrato. Il futuro di questo campo risiede nella capacità di creare sinergie tra discipline, attraverso lo sviluppo di soluzioni che rispondano non solo alle esigenze della scienza e della tecnologia, ma anche a quelle della società e del pianeta.

BIBLIOGRAFIA

Acciai A., Angius N., Perconti P., Plebe A., (di prossima pubblicazione) *Undesigned Cognitive Architectures*, in "Advances in Cognitive Systems".

Ackerman E., Koziol M., 2019, *The blood is here: Zipline's medical delivery drones are changing the game in Rwanda*, in "IEEE Spectrum", 56, 5, pp. 24-31, doi: 10.1109/MSPEC.2019.8701196.

Albarracin M., Hipólito I., Tremblay S. E., Fox J. G., René G., Friston K. J., Ramstead M. J. D., 2023, *Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making*, in Buckley, C.L., Cialfi, D., Lanillos, P., Ramstead, M., Sajid, N., Shimazaki, H., Verbelen, T., Wisse, M. (ed.), Active Inference: 4th International Workshop, IWAI 2023, Ghent, Belgium, September 13-15, 2025, Revised Selected Papers, "Communications in Computer and Information Science", vol. 1915, Springer, Cham, https://doi.org/10.1007/978-3-031-47958-8_9.

Albarracin, M., Hipólito, I., Raffa, M., Kinghorn, P., 2024, *Modeling Sustainable Resource Management Using Active Inference*, in Buckley C.L., Cialfi D., Lanillos P., Pitliya R. J., Sajid N., Shimazaki H., Verbelen T., Wisse M. (ed.), Active Inference: 5th International Workshop, IWAI 2024, Oxford, UK, September 9-11, 2024, Revised Selected Papers, "Communications in Computer and Information Science", Springer, Cham, https://doi.org/10.1007/978-3-031-77138-5, pp. 237-259.

Albarracin M., Ramstead M., Pitliya J. R., Hipólito I., Da Costa L., Raffa M., Constant A., Manski S. G., 2024, *Sustainability Under Active Inference*, in "Systems", 12 (5), 163, https://doi.org/10.3390/systems12050163.

Albrecht S. V., Christianos F., Schäfer L. 2024, *Multiagent Reinforcement Learning: Foundations and Modern Approaches*, MIT Press.

Alfieri I., Fleres A., Damiano L., 2022, Workshop *Eco-socio-botics 2022 – Social Robotics for Sustainability* at 14th International Conference, ICSR 2022, Florence, Italy, December 13-16, 2022.

Alfieri I., Raffa M., 2025, *Active Inference for Ethical Decision-Making in Socially Assistive Robotics*, in: Seibt J., Fazekas P., Quick O. (ed.), Social Robots With AI: Prospects, Risks, and Responsible Methods: Proceedings of Robophilosophy 2024, IOS Press, Amsterdam.

Alsharkawi A., Al-Fetyani M., Dawas M., Saadeh H., Alyaman M., 2021, *Poverty Classification Using Machine Learning: The Case of Jordan*, in "Sustainability", 13, 1412, https://doi.org/10.3390/su130314121.

Alshuhri M. S., Al-Musawi S. G., Al-Alwany A. A., Uinarni H., Rasulova I., Rodrigues P., Alkhafaji A. T., Alshanberi A. M., Alawadi A. H., Abbas A. H., 2023, *Artificial intelligence in cancer diagnosis: Opportunities and challenges*, in "Pathology, research and practice", 253, 154996.

Ananny M., Crawford K., 2018, *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability*, in "New Media & Society", 20 (3), pp. 973-989, doi: 10.1177/1461444816676645.

Anderson J. R., Lebiere C., 1998, *The Atomic Components of Thought*, Erlbaum.

Anderson J. R., Lebiere C., 2003, *The Newell test for a theory of cognition*, in "Behavioral and Brain Sciences", 26 (5), pp. 587-601.

Anderson J. R., Bothell D., Byrne M. D., Douglass S., Lebiere C., Qin Y., 2004, *An integrated theory of the mind*, in "Psychological review", 111 (4), 1036.

Angius N., Perconti P., Plebe A., Acciai, A., 2024, *The Simulative Role of Neural Language Models in Brain Language Processing*, in "Philosophies", 9 137, https://doi.org/10.3390/philosophies9050137.

Araiza E., Morris M., Integlia R., 2019, *Using Sustainable Robotics in an Intelligent Robotic Gardening System for Education*, IEEE International Symposium on Technology in Society (ISTAS) Proceedings, doi: 10.1109/ISTAS48451.2019.8937955.

Arumugam D., Ho M.K., Goodman N.D., Van Roy B., 2024, *Bayesian Reinforcement Learning With Limited Cognitive Load*, in "Open Mind", 8, pp. 395-438, doi: 10.1162/opmi_a_00132.

Ashby W.R., 1960, *Design for a Brain*. Springer, Dordrecht, https://doi.org/10.1007/978-94-015-1320-3_8.

Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López, Molina D., Benjamins R., Chatila R., Herrera F., 2020, *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and*

challenges, in “Information Fusion”, 58, pp. 82-115.

Beheshtian N., Moradi S., Ahtinen A., Väänenen K., Kähkonen K., Laine M., 2020, *GreenLife: A Persuasive Social Robot to Enhance the Sustainable Behavior in shared Living Spaces,* in Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, NordiCHI ‘20, October 25-29, Tallinn, Estonia, ACM, NewYork, NY, USA.

Bender E., Gebru T., McMillan-Major A., Shmitchell S., 2021, *On the dangers of stochastic models: Can language models be too big,* in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ‘21). Association for Computing Machinery, New York, NY, USA, pp. 610-623, https://doi.org/10.1145/3442188.3445922.

Bergson H., 1911, *L’evoluzione creatrice*, ed. 2002, a cura di Acerra M., Raffaello Cortina, Milano.

Biggio B., Roli F., 2018, *Wild patterns: Ten years after the rise of adversarial machine learning*, in “Pattern Recognition”, 84, pp. 317-331, https://doi.org/10.1016/j.patcog.2018.07.023.

Block N., 2005, *Review of Alva Noë’s “Action in Perception”*, in “The Journal of Philosophy”, 102 5 pp. 259-272.

Boccignone G., Cordeschi R., 2007, *Bayesian models and simulations in cognitive science,* in “Workshop Models and Simulations” 2, Tillburg, NL.

Boden M., 2016, *AI: Its nature and future,* Oxford University Press, Oxford.

Botvinick M., Ritter S., Wang J. X., Kurth-Nelson Z., Blundell C., Hassabis D., 2019, *Reinforcement Learning, Fast and Slow*, in “Trends in Cognitive Sciences”, 23 (5), https://doi.org/10.1016/j.tics.2019.02.006.

Brette R., 2019, *Is coding a relevant metaphor for the brain?*, in “Behavioral and Brain Sciences”, 42(e215), pp. 1-58.

Bromley J., 2005, *Guidelines for the use of Bayesian networks as a participatory tool for water resource*, Walliford, United Kingdom.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelekantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D. M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D., *Language models are few-shot learners*, in "ArXiv" preprint, arXiv:2005.14165.

Bruineberg J., Kiverstein J., Rietveld E., 2018, *The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective*, in"Synthese", 195, pp. 2417-2444, https://doi.org/10.1007/s11229-016-1239-1.

Bruineberg J., Dolega K., Dewhurst J., 2021, *The Emperor's New Markov Blankets*, in "Behavioral and Brain Sciences", 45, pp. 1-63, doi: 10.1017/S0140525X21002351.

Brundtland G.H., 1987, *Our Common Future: Report of the World Commission on Environment and Development*, Geneva, UN-Dokument A/42/427.

Bugmann G., Siegal M., Burcin R., 2011, *A Role for Robotics in Sustainable Development?* IEEE Africon 2011 - The Falls Resort and Conference Centre, Livingstone, Zambia, 13-15 September 2011.

Burgess N., Donnett J.G., Jeffery, K.J., O-keefe J., 1997, *Robotic and neuronal simulation of the hippocampus and rat navigation*, in "Philosophical Transactions of the Royal Society", Londra, Ser. B Biol. Sci., 352, pp. 1535-1543.

Burrell J., 2016, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, in "Big Data & Society", 3 (1), 205395171562251, https://doi.org/10.1177/2053951 715622512.

Cappuccio M. (ed.), 2009, *Neurofenomenologia: Le scienze della mente e la sfida dell'esperienza cosciente*, Mondadori.

Caradonna J. L., 2014, *Sustainability: A History*, Oxford University Press.

Carlini N., Wagner D., 2017, *Towards Evaluating the Robustness of Neural Networks*, in "IEEE Symposium on Security and Privacy (SP)", pp. 39-57.

Castellano G., De Carolis B., D'Errico F., Macchiarulo N., Rossano V., 2021, *PeppeRecycle: Improving Children's Attitude Toward Recycling by Playing with a*

Social Robot, in "International Journal of Social Robotics" 13, pp. 97-111, https://doi.org/10.1007/s12369-021-00754-0.

Castelvecchi D. 2016, *Can we open the black box of AI?*, in "Nature News" 538.7623, p. 20, doi: 10.1038/538020a.

Chalmers D., 1995, *Facing up to the problem of consciousness*, in "Journal of Consciousness Studies", 2 (3), 1995, pp. 200-19.

Chan A., Okolo C., Terner Z., Wang A., 2021, *The limits of global inclusion in AI development*, https:// arxiv. org/ pdf/ 2102. 01265. Pdf.

Chang Y., Wang X., Wang J.,Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Yu P. S., Yang Q., Xie X., 2023, *A Survey on Evaluation of Large Language Models*, in "ArXiv", arXIv:2307.3109.

Charmet F., Tanuwidjaja H.C., Ayoubi S., Gimenez P., Han Y., Jmila H., Blanc G., · Takahashi T., Zhang Z., 2022, Explainable artificial intelligence for cybersecurity: a literature survey. *Ann. Telecommun.* 77, pp. 789-812, https://doi.org/10.1007/s12243-022-00926-7.

Chater, N., Tenenbaum J. B., Yuille, A., 2006, *Probabilistic models of cognition: Conceptual foundations*, in "Trends in Cognitive Sciences", 10 (7), pp. 287-291.

Chella A., 2022, *Robots and machine consciousness*, in Cangelosi A., Asada M. ed., *Cognitive Robotics*, MIT Press.

Chella A., 2023, *Artificial consciousness:the missing ingredient for ethical AI?*, in "Frontiers in Robotics and AI", 10:1270460, doi: 10.3389/frobt.2023.1270460.

Chen J., 2009, *Understanding Social Systems: A Free Energy Perspective*, in "Journal of Human Thermodynamics", 5, doi: 10.2139/ssrn.1269035.

Chiarabelli C., Stano P., Luisi P. L., 2009, *Chemical approaches to synthetic biology*. *Current Opinion in Biotechnology*, 20 (4), pp. 492–497, https://doi.org/10.1016/j.copbio.2009.08.004, PubMed: 19729295.

Chui M., Harrysson M., Manyika J., Roberts R., Chung R., van Heteren A., Nel P., 2018,

Notes from the AI frontier: Applying AI for Social Good, McKinsey Global Institute.

Clark A., 1999, *An embodied cognitive science?*, in "Trends in Cognitive Science", 3 (9), pp. 345-351.

Clark A., 2013a, *Whatever next? Predictive brains, situated agents, and the future of cognitive science*, in "Behavioral & Brain Sciences", 36 (3), pp. 181-204.

Clark A., 2013b, *Expecting the World: Perception, Prediction, and the Origins of Human Knowledge*, in "The Journal of Philosophy", 110 (9), pp. 469-496.

Clark A., 2015, *Radical predictive processing*, in "The Southern Journal of Philosophy", 53, pp. 3-27.

Clark A., 2016, *Surfing uncertainty Prediction, action and the embodied mind*, Oxford, Oxford University Press.

Coeckelbergh M., 2021, *AI for climate: freedom, justice, and other ethical and political challenges*, in "AI and Ethics", 1 (1), pp. 67-72, https:// doi. org/ 10. 1007/ s43681-020- 00007-2.

Coget J., 2017, *Technophobe vs. Techno-enthusiast: Does the Internet Help or Hinder the the Balance Between Work and Home Life?* in "Academy of Management Perspectives", 25 (1), https://doi.org/10.5465/amp.25.1.95.

Collis P., Kinghorn P. F., Buckley C. L., 2023, *Tool Discovery and Tool Innovation Using Active Inference*, in "ArXiv", arXiv:2311.03893**.**

Collis P., Singh R., Kinghorn P. F., 2024, *Learning in Hybrid Active Inference Models*, in "ArXiv", arXiv:2409.01066.

Colombo M., Wright C., 2021, *First principles in the life sciences: the free-energy principle, organicism, and mechanism*, in "Synthese", 198 (14), pp. 3463-3488.

Commissione Europea, 24 maggio 2024, *Artificial Intelligence Act,* https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

Cordeschi R., 2002, *The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics*, Berlin and New York, Springer, doi:10.1007/ 978-94-015-9870-5.

Cordeschi R., 2008, *Il metodo sintetico: problemi epistemologici nella scienza cognitiva*, in “Sistemi intelligenti”, 20 (2), pp. 167-191.

Cowls J., King T., Taddeo M., Floridi L., 2019, *Designing AI for Social Good: Seven Essential Factors*, http://dx.doi.org/10.2139/ssrn.3388669.

Cowls J., Tsamados A., Taddeo M., Floridi L., 2021, *The AI gambit Leveraging artificial intelligence to combat climate change: opportunities, challenges, and recommendations*, in “SSRN Electronic Journal”, http://dx.doi.org/10.2139/ssrn.3804983.

Crawford K., 2021, *Atlas of AI. Power, Politics and the Planetary Costs of Artificial Intelligence*, Yale University Press, New Heaven and London.

Da Costa L., Lanillos P., Sajid N., Friston K., Khan S., 2022, *How active inference could help revolutionise robotics*, in “Entropy”, 24 (361), doi: 10.3390/ e24030361.

Damiano L., Cañamero L., 2012, *The Frontier of Synthetic Knowledge: Toward a Constructivist Science*, in “World Futures: The Journal of New Paradigm Research”, 68 (3), pp. 171-177.

Damiano L., Hiolle A., Cañamero L., 2011, *Grounding synthetic knowledge*, in “Advances in artificial life”, ECAL 2011, Lenaerts T., Giacobini M., Bersini H., Bourgine P., Dorigo M., Doursat R. (ed.), pp. 200-207, MIT Press, Cambridge MA.

Damiano L., Stano P., 2023, *Explorative Synthetic Biology in AI: Criteria of Relevance and a Taxonomy for Synthetic Models of Living and Cognitive Processes*, in “Artificial life”, 29, pp. 1-21. 10.1162/artl_a_00411.

Dauvergne P., 2020, *AI in the wild: Sustainability in the age of artificial intelligence*, MIT Press.

De Magistris S., Del Bimbo A., 2023, *SDG-based AI Ethics: an analysis of recent Computer Vision research*. in “IEEE MetroXRAINE 2023: International Conference on Metrology for eXtended Reality, Artificial Intelligence, and Neural Engineering Proceedings”.

Den Hartigh R. J., Hill Y., *Conceptualizing and measuring psychological resilience: What can we learn from physics?*, in “New Ideas in Psychology”, 66, 100934, doi: https://doi.org/10.1016/j.newideapsych.2022.100934.

Derks I. P., de Waal A., 2020, *A Taxonomy of Explainable Bayesian Networks*, in: Gerber A. (ed.), *Artificial Intelligence Research*, SACAIR 2020, *Communications in Computer and Information Science*, vol. 1342, Springer, Cham, doi: 10.1007/978-3-030-66151-9_14.

Di Paolo E. A., 2002, *Organismically-Inspired Robotics: Homeostatic Adaptation and Teleology Beyond the Closed Sensory-Motor Loop*, Advance Knowledge International, Adelaide, Australia.

Di Paolo L. D., White B., Guénin-Carlut A., Constant A., Clark A., 2024, *Active inference goes to school: the importance of active learning in the age of large language models*, in "Philosophical Transactions of the Royal Society London B Biological Sciences", 37, doi: 10.1098/rstb.2023.0148.

Doshi-Velez F., Kim B., 2017, *Towards a rigorous science of interpretable machine learning*, in "arXiv", doi: 10.48550/arXiv.1702.08608.

Drake M., Ong J., Hansen M., Peets L., 2023, *EU AI Policy and Regulation: What to look out for in 2023*, in "Inside Privacy", url: https://www.insideprivacy.com/ artificial-intelligence/eu-ai-policy-and-regulation-what-to-look-out-for-in-2023/ (23/10/2024).

Dumouchel P., Damiano L., 2019, *Vivere con i robot. Saggio sull'empatia artificiale*, Raffaello Cortina Editore, Varese.

Durán J. M., 2018, *Computer Simulations in Science and Engineering: Concepts-Practices-Perspectives*, Springer Nature, Cham, Switzerland.

Epstein R., 2016, *The empty brain. Your brain does not process information and it is not a computer*, in "Aeon".

Eubanks V., 2018, *Automating inequality. How high-tech tools profile, police and punish the poor*, St. Martin Press, New York.

Facchin M., 2021, *Extended predictive minds: do Markov Blankets matter?*, in "Review of Philosophy and Psychology", doi: 10.1007/S13164-021-00607-9.

Felin T., Holweg M., 2024, *Theory Is All You Need: AI, Human Cognition and Decision-Making*. Disponibile a SSRN: https://ssrn.com/abstract=4737265.

Ferilli S., Girardi E. (a cura di), 2021, *L'intelligenza artificiale per lo sviluppo sostenibile*, CNR edizioni.

Ferrara E., 2023, *Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models*, in "ArXiv", arXiv:2304.03738.

Floridi L., 2019, *What the Near Future of Artificial Intelligence Could Be*, in "Philosophy & Technology", 32 (1), pp. 1-15, https://doi.org/10.1007/s13347- 019- 00345- y.

Floridi L., 2020, *Il verde e il blu: Idee ingenue per migliorare la politica*, Raffaello Cortina Editore.

Floridi L., 2022, *Etica dell'intelligenza artificiale: Sviluppi, opportunità, sfide*, Raffaello Cortina Editore.

Friedman D., Isaac R. M., James D., 2014, *Risky Curves: On the Empirical Failure of Expected Utility*, New York, Routledge.

Friston K. J., 2003, *Learning and inference in the brain*, in "Neural Networks", 16, pp. 1325-1352.

Friston K. J., 2009, *The free-energy principle: A rough guide to the brain?*, in "Trends in "Cognitive Sciences", 13, pp. 293-301.

Friston K. J., 2010, *The free-energy principle: a unified brain theory?*, in "Nature Reviews Neuroscience", 11 (2), pp. 127-138.

Friston K. J, 2013, *Life as we know it*, in "Journal of the Royal Society Interface", 10, doi: 10.1098/rsif.2013.0475.

Friston K. J., Brown H. R, Siemerkus J., Stephan K. E., 2016, *The dysconnection hypothesis*, in "Schizophrenia Research", 176 (2-3), pp. 83-94.

Friston K. J., Da Costa L., Sajid N., Heins C., Ueltzhöffer K., Pavliotis G. A., Parr T., 2023, *The free energy principle made simpler but not too simple*, in "Physics Reports", 1024, pp. 1-29, https://doi.org/10.1016/j.physrep.2023.07.001.

Friston K. J., Da Costa L., Tschantz A. Kiefer A., Salvatori T., Neacsu V., Koudahl M., Sajid N., Heins C., Markovic D., Parr T., Verbelen T., Buckley, C., 2023, *Supervised structure learning*, in "ArXiv", arXiv:2311.10300, 2023.

Friston K. J., Frith C. D., 2015a, *Active inference, communication and hermeneutics*, in "Cortex", 68, pp. 129-143.

Friston K. J., Frith C. D., 2015b, *A duet for one*, in "Consciousness and Cognition", 36, pp. 390-405.

Friston K. J., Kiebel S., 2009, *Predictive coding under the free-energy principle*, in "Philosophical Transactions of the Royal Society B" 364, pp. 1211-1221.

Friston K. J., Kilner J., Harrison L., 2006, *A free energy principle for the brain*, in "Journal of Physiology", 100, pp. 70-87.

Friston K. J., Mattout J., Kilner J., 2011, *Action understanding and active inference*, in "Biological Cybernetics" 104, pp. 137-160, doi:10.1007/s00422- 011- 0424- .

Friston K. J., Parr T., de Vries B., 2017, *The graphical brain: Belief propagation and active inference*, in "Network Neuroscience", 1 (4), pp. 381-414, doi: 10.1162/NETN_a_00018.

Friston K. J., Stephan K., 2007, *Free-energy and the brain*, in "Synthese", 159, pp. 417-458.

Froese T., Taguchi S., 2019, *The problem of meaning in AI and robotics: still with us after all these years*, in "Philosophies", 4 (14), 1-14.

Gadd M., De Martini D., Pitt L., Tubby W., Towlson M., Prahacs C., Bartlett O., Jackson J., Qi M., Newman P., Hector A., Salguero-Gómez R., Hawes N., 2024, *Watching Grass Grow: Long-term Visual Navigation and Mission Planning for Autonomous Biodiversity Monitoring*, in "ArXiv", arXiv:2404.10446.

Galati R., Mantriota G., Reina G., 2022, *Mobile Robotics for Sustainable Development: Two Case Studies*, in: Quaglia G., Gasparetto A., Petuya V., Carbone G. (ed.), *Proceedings of I4SDG Workshop 2021. I4SDG 2021. Mechanisms and Machine Science*, vol. 108, Springer, Cham.

Gallagher S., 2005, *How the Body Shapes the Mind*, Oxford University Press.

Gallagher S., Nelson, 2003, *Handbook of Psychology: Biological Psychology*, John Wiley & Sons.

Gallagher S., Zahavi D., 2008, *The phenomenological mind: An introduction to philosophy of mind and cognitive science*, Routledge.

Gigerenzer G., 2020, *How to explain behavior?*, in "Topics in cognitive science", 12 (4), pp. 1363-1381.

Giudici P., Raffinetti E., 2023, *Sustainable, Accurate, Fair and Explainable Machine Learning Models*, https://en.wiki.topitalianscientists.org/Sustainable_Accurate_ Fair_and_Explainable_ Machine_Learning_Models.

Goertzel B., 2014, A*rtificial general intelligence: Concept, state of the art, and future prospects*, in "Journal of Artificial General Intelligence", 5, pp. 1-46.

González L. A., Coronado Martín J. A., Vaca-Tapia A. C., Rivas F., 2021, *How Sustainability Is Defined: An Analysis of 100 Theoretical Approximations*, in "Mathematics", 9(11): 1308. https://doi.org/10.3390/math9111308.

Grasso F. W., Consi T.R., Mountain D.C., Atema J., 2000, *Biomimetic robot lobster performs chemo-orientation in turbulence using a pair of spatially separated sensors: Progress and challenges*, in "Robotic and Autonomous Systems", 30, pp. 115-131.

Guest O., Martin A. E., 2023, *On logical inference over brains, behaviour, and artificial neural networks*, in "Computational Brain & Behavior", pp. 1-15, doi: 10.1007/s42113-022-00166-x.

Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., and Pedreschi D., 2018, *A survey of methods for explaining black box models*, in "ACM Computing Surveys" 51.5, pp. 1-42, doi: 10.1145/3236009.

Gunning, D., 2017, *Explainable Artificial Intelligence (XAI)*, in "Defense Science Research Projects Agency 2.2", p. 1, doi: 10.1609/aimag.v40i2.2850.

Ham J., Midden C., 2014, *A Persuasive Robot to Stimulate Energy Conservation: The Influence of Positive and Negative Social Feedback and Task Similarity on Energy-Consumption Behavior*, in "International Journal of Social Robotics", 6, pp. 163-171.

Hartmann F., Baumgartner M., Kaltenbrunner M., 2021, *Becoming Sustainable, The New Frontier in Soft Robotics*, in "Advanced Materials", 33, 2004413, https://doi.org/10.1002/adma.202004413.

Héder M., 2020, *The epistemic opacity of autonomous systems and the ethical consequences*, in "AI & Society", pp. 1-9.

Héder M., 2023, *Explainable AI: A Brief History of the Concept*, in "ERCIM News", 134, pp. 9-10.

Heilinger J., Kempt H., Nagel S., 2023, *Beware of sustainable AI! Uses and abuses of a worthy goal*, in "AI Ethics", pp. 1-12, https://doi.org/10.1007/s43681-023-00259-8.

Heins C., Millidge B., Demekas D., Klein B., Friston K. J., Couzin I., Tschantz, A., 2022, *Pymdp: A python library for active inference in discrete state spaces*, in "ArXv", arXiv:2201.03904.

Henderson P., Hu J., Romoff J., Brunskill E., Jurafski D., Pineau J., 2020, *Towards the systematic reporting of the energy and carbon footprints of machine learning*, in "Journal of Machine Learning Research", 21 (248), pp. 1-43.

Hendrycks D., Dietterich T. G., 2019, *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*, in "ArXiv", arXiv:1903.12261.

Hesp C., Tschantz A., Millidge B., Ramstead M., Friston K. J., Smith R., 2020, *Sophisticated Affective Inference: Simulating Anticipatory Affective Dynamics of Imagining Future Events*, in "IWAI 2020 Proceedings", Springer, pp. 179-186.

High-Level Expert Group on Artificial Intelligence, 2019, *Ethics guidelines for trustworthy AI*, B-1049 Brussels.

Hinton G. E., McClelland J. L., Rumelhart D. E., 1986. *Distributed representations*, in Rumelhart McClelland (ed.), *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*, MIT Press, pp. 77-109.

Hipólito I., 2023, *The Human Roots of Artificial Intelligence*, in "PsyArxiv", doi: 10.31234/osf.io/cseqt.

Hoff P. D., 2009, *A first course in Bayesian statistical methods*, Springer.

Hohwy J., 2019, *Quick'n'Lean or Slow and Rich? Andy Clark on predictive processing and embodied cognition*, in Colombo M., Irvine E., Stapleton M. (ed.), *Andy Clark and His Critics*, pp. 191-205, New York, Oxford University Press.

Holland J., 1992, *Genetic Algorithms*, in "Scientific American", 267 (1), pp. 66-73.

Holzinger A., Saranti A., Molnar C., Biecek P., Samek W., 2022, *Explainable AI Methods - A Brief Overview*, in Holzinger A., Goebel R., Fong R., Moon T., Müller K.R., Samek

W. (ed.), *xxAI - Beyond Explainable AI. xxAI 2020*, in “Lecture Notes in Computer Science”, vol. 13200, Springer, Cham, https://doi.org/10.1007/978-3-031-04083-2_2.

Indurkyha B., Sienkiewicz B., 2024, *Robots and Social Sustainability*, in “ArXiv”, arXiv:2401.03477.

Jean N., Burke M., Xie M., Davis W. M., Lobell D. B., Ermon S., 2016, *Combining Satellite Imagery and Machine Learning to Predict Poverty*, in “Science”, 353, pp. 790-794.

Jimenez Rezende D., Mohamed S., Wierstra D., 2014, *Stochastic backpropagation and approximate inference in deep generative models*, In Xing E. P., JebaraT. (ed.), Proceedings of Machine Learning Research, pp. 1278-1286.

Jones A., 2016, *Brains, tortoises, and octopuses: Postwar interpretations of mechanical intelligence on the BBC*, in “Information & Culture”, 51 (1), University of Texas Press, pp. 81-101, doi:10.7560/IC51104. S2CID 143096137.

Joshi, Naveen, 2018, The Negative Environmental Impact of Robotics, Sitex: https://www.bbntimes.com/environment/the-negative-environmental-impact-of-robotics, 16/06/2022.

Kaelbling L. P., Littman M. L., Cassandra A. R., 1998, *Planning and acting in partially observable stochastic domains*, in “Artificial Intelligence”, 101 (1-2), pp. 99-134, doi: https://doi.org/10.1016/S0004-3702(98)00023-X.

Kar A. K., Choudhary S. K., Singh V. K., 2022, *How can artificial intelligence impact sustainability: A systematic literature review*,in “Journal of Cleaner Production”, Volume 376, 134120, https://doi.org/10.1016/j.jclepro.2022.13412.

Kim J., Jun S., Jang D., Park S., 2018, *Sustainable technology analysis of artificial intelligence using Bayesian and social network models*, in “Sustainability”, 10 (1), p. 115, doi: https://doi.org/10.3390/su10010115.

King, T. C., Aggarwal N., Taddeo M., Floridi L., 2020, *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*, in “Science and Engineering Ethics”, 26 (1), pp. 89-120, https://doi.org/10.1007/s11948-018-00081-0.

Kingma D. P., Welling M., 2014, *Auto-encoding variational Bayes*, in “Proceedings of International Conference on Learning Representations”.

Kirat T., Tambou O., Do V., Tsoukiàs A., 2022, *Fairness and Explainability in Automatic Decision-Making Systems. A challenge for computer science and law*, in “ArXiv”, arXiv: 2206.03226.

Kirchhoff M. D., 2018, P*redictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it?*, in “Philosophical Studies”, 175 (3), pp. 751-767.

Kirchhoff M. D., Parr T., Palacios E., 2018. *The Markov Blankets of Life: Autonomy, Active Inference and the Free Energy Principle*, in “Journal of the Royal Society Interface”, 15 (138).

Kitano H., Hamahashi S., Luke S., 1998, *The perfect* C. elegans *project: An initial report*, in “Artificial Life”, 4, pp. 141-156.

Knill D. C., Kersten D., Yuille A., 1996, *Introduction: A Bayesian formulation of visual perception*, Cambridge University Press, pp. 1-21.

Koski, T., Noble J. M., 2009, *Bayesian Networks: an Introduction*, Chichester, Wiley and Sons.

Kwisthout J., van Rooij I., 2020, *Computational resource demands of a predictive bayesian brain*, in “Computational Brain & Behavior”, 3, pp. 174-188, https://doi.org/10.1007/s42113-019-00032-3.

Lacoste A., Luccioni A., Schmidt V., Dandres T., 2019, *Quantifying the carbon emissions of machine learning*, in “ArXiv”, arXiv:1910.09700.

Laird J. E., 2012, *The Soar Cognitive Architecture*, MIT Press.

Laird J. E., Newell A., Rosenbloom P. S., 1987, *SOAR: An architecture for general intelligence*, in “Artificial Intelligence”, 33 (1), pp. 1-64.

Lakshmi D., Tiwari R. S., Dhanaraj R. K., Kadry S. ed., 2024, *Explainable AI (XAI) for Sustainable Development Trends and Applications*, Chapman & Hall, CRC Press.

Lambrinos D., Möller R., Labhart T., Pfeifer R.,Wehner R., 2000, *A mobile robot employing insect strategies for navigation*, in “Robotics and Autonomous Systems”, 30, pp. 39-64.

Lawrence E., El ShazlyA., Seal S., Chaitanya K. J., Liò P., Singh S., Bender A., Sormanni P., Greenig M., *Understanding Biology in the Age of Artificial Intelligence*, in “ArXIv”, arXiv:2403.04106v1.

Lanillos P., Cheng G., 2018, *Adaptive robot body learning and estimation through predictive coding*, in “IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Proceedings, IEEE”, pp. 4083-4090

Lanillos P., Meo C., Pezzato C., Meera A. A., Baioumy M., Ohata W., Alexander Tschantz A., Millidge B., Wisse M., Buckley C. L., Tani J., 2021, *Active Inference in Robotics and Artificial Agents: Survey and Challenges*, in “ArXiv”, arXiv:2112.01871.

Lázaro-Gredilla M., Ku L. Y., Murphy K. P., George D., 2024, *What type of inference is planning?*, in “ArXiv”, arxiv:2406.17863.

LeCun Y., Bengio Y., Hinton G., 2015, *Deep Learning*, in “Nature”, 52, 7553, pp. 436-444.

Levine S., 2018, *Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review*, in “ArXiv”, arXiv:1805.00909.

Lieto A., Bhatt M., Oltramari A., Vernon D., 2018, *The role of cognitive architectures in general artificial intelligence*, in “Cognitive Systems Research”, 48, pp. 1-3, ttps://doi.org/10.1016/j.cogsys.2017.08.003.

Lieto A., 2021, *Cognitive design for artificial minds*, New York, Routledge.

Limanowski J., Friston K. J., 2018, *‘Seeing the dark’: Grounding phenomenal transparency and opacity in precision estimation for active inference,* in “Frontiers in Psychology”, 9, p. 643. doi: 10.3389/fpsyg.2018.00643.

Loach K., Rowley J., Griffiths J., 2016, *Cultural sustainability as a strategy for the survival of museums and libraries*, “International Journal of Cultural Policy” 23 (2), pp. 186-198, doi: 10.1080/10286632.2016.1184657.

Longo L., Brcic M., Cabitza F., Choi J,, Confalonieri R., Del Ser J., Guidotti R., Hayashi Y., Herrera F., Holzinger A., Jiang R., Khosravi H., Lecue F., Malgieri G., Páez A., Samek W., Schneider J., Speith T., Stumpf S., 2024, *Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions*, “Information Fusion”, 106, https://doi.org/10.1016/j.inffus.2024.102301.

Lo SY., Lai YY., Liu JC., Yeh SL., 2022, *Robots and Sustainability: Robots as Persuaders to Promote Recycling*, in “International Journal of Social Robotics”, 14,

pp. 1261-1272.

Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A., *Towards Deep Learning Models Resistant to Adversarial Attacks*, in "ArXiv", arXiv:1706.06083.

Manzotti R., 2012, *The computational stance is unfit for consciousness*, in "International Journal of Machine Consciousness", 4, 401420.

Manzotti R., 2019, *Embodied AI beyond embodied cognition and enactivism*, in "Philosophies", 4, 115.

Manzotti R., Chella, A., 2020, *Conscious Machines: A Possibility? If So, How?*, in "Journal of Artifcial Intelligence and Consciousness", 7, 2, pp. 1-16, doi 10.1142/S2705078520710022.

Manzotti R., Rossi S., 2023, *Io e Ia: Mente, Cervello e GPT*, Rubbettino Editore.

Marconi D., 2001, *Filosofia e scienza cognitiva*, Editori Laterza, Bari.

Marr D., 1982, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W. H. Freeman, New York.

Marwala T., 2023, *Artificial Intelligence, Game Theory and Mechanism Design in Politics*, Palgrave McMillan, doi: 0.1007/978-981-99-5103-1.

Marwala T., Mbuvha R., Mungwe W. T., 2023, *Hamiltonian Monte Carlo Methods in Machine Learning*, Academic Press.

Mazzaglia P., Verbelen T., Çatal O., 2022, *The Free Energy Principle for Perception and Action: A Deep Learning Perspective*, in "Entropy", 24 (2).

Mazzi F., Floridi L. ed., 2023, *The Ethics of Artificial Intelligence for the Sustainable Development Goals*, Springer.

Mbiti I M., Weil D., 2011, *Mobile Banking: The Impact of M-Pesa in Kenya*, in "Development Economics Economics eJournal", doi: 10.3386/W17129.

Metta G., Natale L., Nori F., Sandini G., Vernon D., Fadiga L., von Hofsten C., Rosander K., Lopes M., Santos-Victor J., Bernardino A., Montesano L., 2010, *The iCub*

humanoid robot: an open-systems platform for research in cognitive development, 23 (8-9), pp. 1125-1134, doi: 10.1016/j.neunet.2010.08.010.

Miller M., Albarracin M., Pitliya R. J., Kiefer A., Mago J., Gorman C., Friston K. J., Ramstead M. J. D., 2022, *Resilience and Active Inference*, in "Frontiers in Psychology", 22, 13: 1059117, doi: 10.3389/fpsyg.2022.1059117.

Miller T., 2019, *Explanation in Artificial Intelligence: Insight from the Social Sciences*, in "Artificial Intelligence", 267 (1), doi: 10.1016/j.artint.2018.07.007.

Millidge B., Seth A., Buckley C., 2022, *Predictive Coding: A Theoretical and Experimental Review*, in "ArXiV", arXiv.2107.12979.

Mitchell M. 2022, *Intelligenza Artificiale: Una guida per esseri umani pensanti*, Einaudi, Torino.

Molnar C., 2018, *A guide for making black box models explainable*, in https://christophm.github.io/interpretable-ml-book.

Morozov E., 2013, *To save everything, click here. The folly of techno- logical solutionism,* New York, Perseus.

Nasir O., Javed R., Gupta S., Vinuesa R., Qadir, J., 2022, *Artificial intelligence and sustainable development goals nexus via four vantage points*, in "Technology in Society", 72. 102171, doi: 10.1016/j.techsoc.2022.102171.

Newell A., 1980, *Physical symbol systems*, in "Cognitive Science", 4, pp. 135-183.

Newell A., 1990, *Unified Theories of Cognition*, Cambridge, MA, Harvard University Press.

Newell A., Simon H. A., 1972, *Human Problem Solving*, Englewood Cliffs, Prentice Hall, NJ, USA.

Newell A., Simon H. A., 1976, *Computer science as empirical inquiry: symbols and search*, in "Communications of the ACM 19", 3, pp. 113–126 https://doi.org/10.1145/360018.360022.

Nunez R., Freeman W. J. (a cura di), 1999, *Reclaming cognition*, Imprint Academic Press, Thorveton.

Ofner A., Stober S., 2018, *Towards bridging human and artificial cognition: Hybrid variational predictive coding of the physical world, the body and the brain*, in "Advances in Neural Information Processing Systems", 27.

Oliveira L. F. P., Moreira A. P., Silva M. F., 2021, *Advances in Agriculture Robotics: A State-of-the-Art Review and Challenges Ahead*, in "Robotics", 10 (2), 52, https://doi.org/10.3390/robotics10020052.

Oltramari A., Lebiere C., 2012, *Pursuing artificial general intelligence by leveraging the knowledge capabilities of act-r*, in "Artificial General Intelligence", Springer, pp. 199-208.

Pachauri, R. K., Meyer L. A. (a cura di), 2014, *Cambiamenti climatici 2014: rapporto di sintesi*, Contributo dei Gruppi di lavoro I, II e III al Quinto Rapporto di Valutazione del Gruppo Intergovernativo sui Cambiamenti Climatici. IPCC, Ginevra, Svizzera, 2014.

Palacios E. R., Razi A., Parr T., 2020, *On Markov Blanket and hierarchical self-organization*, in "Journal of Theoretical Biology", 486, 110089.

Palomares I., Martínez-Cámara E., Montes R., García-Moral P., Chiachio M., Chiachio J., Alonso S., Melero F. J., Molina D., Fernández B., Moral C., Marchena R., Pérez de Vargas J., Herrera F., 2021, *A panoramic view and swot analysis of artificial intelligence for achieving the sustainable development goals by 2030: progress and prospects*, in "Applied Intelligence", 51, pp. 6497-6527, https://doi.org/10.1007/s10489-021-02264-y.

Parr T., Friston K., 2018, *The Discrete and Continuous Brain: From Decisions to Movement - And Back Again*, in "Neural Computation" 30, pp. 2319-2347, doi: 10.1162/neco_a_01102.

Patterson D., Gonzalez J., Le Q., Liang C., Munguia L. M., Rothchild, D. So D., Maud Texier M., Dean, J., 2021, *Carbon emissions and large neural network training*, in "ArXiv", arXiv:2104.10350.

Pearl J., 1988, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kauffman ed., 77-110.

Pedemonte V., 2020, *AI for Sustainability: An overview of AI and the SGDs to contribute*

to the European policy making, https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-sustainability-overview-ai-and-sdgs-contribute-european-policy-making (19/01/2025).

Perconti P., Plebe A., 2020, *Deep learning and cognitive science*, in "Cognition", 203, 104365, doi: 10.1016/j.cognition.2020.104365.

Pezzato C., Baioumy M., Corbato C. H., Hawes N., Wisse M., Ferrari R., 2020, *Active inference for fault tolerant control of robot manipulators with sensory faults*, in "International Workshop on Active Inference Proceedings", Springer, pp. 20-27.

Pezzulo G., Donnarumma F., Iodice P., Maisto D., Stoianov I., 2017, Model-Based Approaches to Active Perception and Control, in "Entropy", 19 266, doi:10.3390/e19060266.

Pezzulo G., Parr T., Cisek P., Clark A., Friston, K., 2024, *Generative meaning: Active inference and the scope and limits of passive AI*, in "Trends in Cognitive Sciences", 28(2), doi: 10.1016/j.tics.2023.10.002.

Pfeifer R., Scheier C., 2000, *Understanding intelligence*, MIT, Cambridge, MA.

Phan T. D., Smart J. C. R., Capon S. J., Hadwen W. L., Sahin O., 2016. *Applications of Bayesian belief networks in water resource management: A systematic review*, in "Environmental Modelling and Software", 85, pp. 98-111, https://doi.org/10.1016/j.envsoft.2016.08.006.

Pickering A., 2010, *The cybernetic brain*, The University of Chicago Press, Londra.

Pio-Lopez L., Nizard A., Friston K. J., Pezzulo G., 2016, *Active inference and robot control: a case study*, in "Journal of The Royal Society Interface", 13 (122).

Plebe A., Grasso G., 2019, *The unbearable shallow understanding of deep learning*, in "Minds and Machines", 29, pp. 515-553.

Priorelli M., Stoianov I., Pezzulo G., 2024, *Learning and embodied decisions in active inference*, in "bioRxiv", https://doi.org/10.1101/2024.08.18.608439.

Purvis B., Mao Y., Robinson D., 2019, *Three pillars of sustainability: in search of conceptual origins*, in "Sustainability Science", 14 (3), pp. 681-695, doi:

10.1007/s11625-018-0627-5.

Radosavovic I., Kosaraju R. P., Girshick R., He K., Dollár P., 2020, *Designing network design spaces*, in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 10428-10436.

Raffa M., 2023, *Markov Blankets for Sustainability*, in: Masci P., Bernardeschi C., Graziani P., Koddenbrock M., Palmieri M. (ed.) Software Engineering and Formal Methods. SEFM 2022 Collocated Workshops. SEFM 2022, "Lecture Notes in Computer Science", vol. 13765, Springer, Cham. https://doi.org/10.1007/978-3-031-26236-4_26.

Raffa M., Acciai A. (di prossima pubblicazione), *Free Energy Principle and Active Inference in Neural Language Models*, in: Bruno A., Pipitone A., Manzotti R., Augello A., Mazzeo P. L., Vella F., Chella A. (ed.), Proceedings of the 2nd Workshop on Artificial Intelligence for Perception and Artificial Consciousness (AIxPAC 2024) in AIxIA 2024, Bolzano, Italia, 28 Novembre, 2024, CEUR Workshop Proceedings.

Ramstead M., 2022, *The empire strikes back: Some responses to Bruineberg and colleagues*, in "Behavioral Brain Sciences", 45, doi: 10.1017/s0140525x22000139.

Ramstead M., Albarracin M., Kiefer A., Klein B., Fields C., Friston K., Safron A., 2023, *The inner screen model of consciousness: applying the free energy principle directly to the study of conscious experience*, in "ArXiv", arXiv:2305.02205.

Redclift M., 1993, *Sustainable Development: Needs, Values, Rights*, in "Environmental Values", 2 (1), pp. 3-20.

Requejo Castro D., 2021, *Data Driven Bayesian Networks modelling to support decision-making: Application to the context of Sustainable Development Goal 6 on water and sanitation*, PhD thesis.

Rosenblatt F., 1957, *The perceptron: A perceiving and recognizing automaton*, Report n. 85-460-01, Cornell Aeronautical Laboratory Inc., Buffalo, New York.

Rosenblueth A., Wiener N., 1945, *The role of models in science* in "Philosophy of Science", 12, pp. 316-321.

Rosenfeld A., Richardson A., 2019, *Explainability in human–agent systems*, in

"Autonomous Agents and Multi-Agent Systems", 33, pp. 673-705, https://doi.org/10.1007/s10458-019-09408-y.

Rubin S., Parr T., Da Costa L., 2020, *Future climates: Markov blankets and active inference in the biosphere*, in "Journal of the Royal Society Interface", 17, http://dx.doi.org/10.1098/rsif.2020.0503.

Ryder D., 2009, *Problems of representation I: nature and role*, in Robins S., Symons J. & Calvo P. (ed.), *The Routledge Companion to Philosophy of Psychology*, New York, NY, Routledge, pp. 233-259.

Sætra H. S., 2021, *AI in Context and the Sustainable Development Goals: Factoring in the Unsustainability of the Sociotechnical System*, in "Sustainability", 13 (4), 1738, https://doi.org/10.3390/su13041738.

Scheutz C., Law T., Scheutz M., 2021, *EnviRobots: How Human-Robot Interaction Can Facilitate Sustainable Behavior*, in "Sustainability", 13, (21), 12283, https://doi.org/10.3390/su132112283.

Shusterman R., Waters A. C., O'Neill S., Luu P., Tucker D. M., 2023, *An active inference strategy for prompting reliable responses from large language models in medical practice*, in "ArXiv", arXiv:2407.21051.

Schultz W., Dayan P., Montague P. R., 1997, *A neural substrate of prediction and reward*, in "Science", 275, pp. 1593-1599.

Seth A., 2020, *The brain as a prediction machine*, in: Mendonça D., Curado M., Gouveia S. S. (ed.), *The Philosophy and Science of Predictive Processing*, Bloomsbury, Londra, pp. XIV-XVII.

Sandved-Smith L., Hesp C., Mattout J., Friston K. J., Lutz A., Ramstead M. J. D., 2021, *Towards a computational phenomenology of mental action: Modelling meta-awareness and attentional control with deep parametric active inference*, in "Neuroscience of Consciousness", doi: 10.1093/nc/ niab018.

Sharif M., Bhagavatula S., Bauer L., Reiter M. K., 2016, *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition*, in "Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security", pp. 1528-1540.

Siegel M., Breazeal C., Norton M. I., 2009, *Persuasive Robotics: The Influence of Robot Gender on Human Behavior*, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, MO, USA, pp. 2563-2568, doi: 10.1109/IROS.2009.5354116.

Sierra L. A., Yepes V., García-Segura T., Pellicer, T. 2018, *Bayesian Network Method for Decision-Making about the Social Sustainability of Infrastructure Projects*, in "Journal of Cleaner Production", 176 (March), pp. 521-534, https://doi.org/10.1016/j.jclepro.2017.12.140.

Silver D., Huang A., Maddison C. J., Guez A., Sifre L., van den Driessche G., Schrittwieser J., Antonoglou I., Panneershelvam V., Lanctot M., Dieleman S., Grewe D., Nham J., Kalchbrenner N., Sutskever I., Lillicrap T., Leach, M., Kavukcuoglu K., Graepel T., Hassabis D., 2016, *Mastering the game of Go with deep neural networks and tree search*, in "Nature", 529, pp. 484-489.

Simon H. A, 1996, *The Sciences of the Artificial*, MIT Press, Cambridge, MA, USA.

Sirmacek B., Gupta S., Mallor F., Azizpour H., Ban Y., Eivazi H., Fang H., Golzar F., Leite I., Melsion G. I., Smith K., Fuso Nerini F., Vinuesa R., 2023, *The Potential of Artificial Intelligence for Achieving Healthy and Sustainable Societies*, in "The Ethics of Artificial Intelligence for the Sustainable Development Goals", Springer.

Spirtes P., Glymour C., Scheines R., 2000, *Causation, prediction, and search*, MIT Press, Cambridge.

Strubell E., Ganesh A., McCallum A., 2019, *Energy and Policy Considerations for Deep Learning in NLP*, in "ArXiv", arXiv.1906.02243.

Sun R., 2005, *The CLARION cognitive architecture: Extending cognitive modeling to social simulation*, in Sun R. ed., Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation. Cambridge University Press, pp. 79-99.

Sutton R. S., Barto A. G., 2018, *Reinforcement Learning: An Introduction*, MIT Press.

Sutton R. S., Barto A. G., 1981, *Toward a modern theory of adaptive networks: expectation and prediction*, in "Psychological Review", 88, 135.

Tan K. H., Lim B. P., 2018, *The artificial intelligence renaissance: Deep learning and the road to human-Level machine intelligence*, in "APSIPA Transactions on Signal and Information Process", 7, e6.

Tani J., 2003, *Learning to generate articulated behavior through the bottom-up and the top-down interaction processes*, in “Neural networks”, 16 (1), pp. 11-23.

Tanni V., 2023, *Dialoghi Artistici con la macchina,* in “Artribune” , 72, maggio-giugno 2023, pp. 45-55.

Thompson E., 2007, *Mind in Life*, in “Biology, Phenomenology, and the Sciences of Mind”, The Belknap Press of the Harvard University Press, Cambridge, MA.

Thórisson K., Helgasson H., 2012, *Cognitive architectures and autonomy: A comparative review*, in “Journal of Artificial General Intelligence”, 3 (2), pp. 1-30.

Tononi G., 2004, *An information integration theory of consciousness*, in “BMC Neuroscience”, 5, 122.

Trimarchi M., 2004, *Regulation, integration and sustainability in the cultural sector*, in “International Journal of Heritage Studies”, 10 (5), pp. 401-415.

Tsividis P., Pouncy T., Xu J. L., Tenenbaum J. B., Gershman S. J., 2017, *Human Learning in Atari*, in “AAAI Spring Symposia”, pp. 643-646.

Tsamados A., Aggarwal N., Cowls J., Morley J., Roberts H., Taddeo M., Floridi L., 2022, *The ethics of algorithms: Key problems and solutions*, in “AI & Society”, 37, pp. 215-230, https://doi.org/10.1007/s00146-021-01154-8.

Tschantz A., Millidge B., Seth A. K., Buckley C. L., 2020, *Reinforcement Learning through Active Inference*, in “ArXiv”, arXiv:2002.12636.

Turing A., 1950, *Computing Machinery and Intelligence*, in “Mind”, pp. 433-460.

Unione Europea, 2024, *The Act Texts. EU Artificial Intelligence,* https://artificialintelligenceact.eu/the-act/, ultimo accesso 25/11/2024.

United Nations General Assembly, 2015, *Transforming our World: The Sustainable Development Agenda to 2030*.

Upreti N. C., Singh V., Nagpal N., 2023 *Towards a Healthier Future: The Trasformative Role of AI in Promoting Good Health and Well-Being (SDG-3)*. AISD 2023: First International Workshop on Artificial Intelligence: Empowering Sustainable

Development, September 4-5, 2023, co-located with International Conference on Artificial Intelligence: Towards Sustainable Intelligence (AI4S-2023), Pune, India, CEUR Workshop Proceedings.

Utama A. B., Wibawa A. P., Handayani A. N., Chuttur M. Y., 2024, *Exploring the Role of Deep Learning in Forecasting for Sustainable Development Goals: A Systematic Literature Review*, in "International Journal of Robotics and Control Systems", 4 (1), pp. 365-400.

Van de Maele T., Verbelen T., Çatal O., 2021, *Active Vision for Robot Manipulators Using the Free Energy Principle*, in "Frontiers in Neurorobotics", 15, doi: 10.3389/fnbot.2021.642780.

Van Es T., Hipólito, I., 2020, *Free-Energy Principle, Computationalism and Realism: a Tragedy*, preprint.

Van Wynsberghe A., 2021, *Sustainable AI: AI for sustainability and the sustainability of AI*, in "AI Ethics", 1 (3), pp. 213-218, https://doi.org/10.1007/s43681-021-00043-6.

Varela F., Maturana H., Uribe R., 1974, *Autopoiesis: The organization of living systems, its characterization and a model*, in "Biosystems", 5 (4), pp. 187-196.

Varela F. J., Thompson E., Rosch E., 1991, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, Cambrige, MA.

Vernon D., 2014, *Artificial cognitive systems: A primer*, MIT Press.

Vernon D., 2017, *Two ways (not) to design a cognitive architecture*, in "Cognitive Robot Architectures", 42.

Vernon D., Metta G., Sandini G., 2007, *The iCub cognitive architecture: Interactive development in a humanoid robot*, in "Development and Learning", 140, ICDL IEEE 6th International Conference, Ieee, pp. 122-127.

Veissière S. P. L., Constant A., Ramstead M., Friston K. J., Kirmayera L. J., 2020, *Thinking Through Other Minds: A Variational Approach to Cognition and Culture*, in "Behavioral and Brain Sciences", 43, pp. 1-90, doi:10.1017/S0140525X19001213.

Vinuesa R., Hossein A., Leite I., 2020, *The Role of Artificial Intelligence in Achieving the Sustainable Development Goals*, in "Nature Communications" 11 (1), p. 233, https://doi.org/10.1038/s41467-019-14108-y.

Vinuesa R., Sirmacek B., 2021, *Interpretable Deep-Learning Models to Help Achieve the Sustainable Development Goals*, in "Nature Machine Intelligence", 3, p. 926, https://doi.org/10.1038/s42256-021- 00414- y.

Webb B., 2002, *Robots in invertebrate neuroscience*, in "Nature", 417, pp. 359-363.

Werner L. C., 2019, *Gordon Pask and the Origins of Design Cybernetics*, in "Design Research Foundations", doi: 10.1007/978-3-030-18557-2_3.

Whittaker M., 2021, *The Steep Cost of Capture*, in "Interactions", 28 (6), pp. 50-55. https://doi.org/10.1145/3488666.

Wiener N., 1948, *Cybernetics or Control and Communication in the Animal and the Machine*, MIT Press, Cambridge, MA, USA.

Winograd T., Flores F., 1986, *Understanding Computers and Cognition: A New Foundation for Design*, Intellect Books.

Winsberg E., 2010, *Science in the Age of Computer Simulation*, Chicago University Press, Chicago, IL, USA.

Zu Y., Su D., He L., Xu L., Yu D., 2024, *Generative Pre-trained Speech Language Model with Efficient Hierarchical Transformer*, in "ArXiv", arXiv:2406.00976.

La borsa di dottorato cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
risorse Fondo Sociale Europeo REACT-EU Azione IV.5 “Dottorati su tematiche Green”