



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI



Università di Cagliari

UNICA IRIS Institutional Research Information System

This is the Author's *accepted* manuscript version of the following contribution:

D. Amoroso, G. Tamburrini, La supervisione umana dell'IA è necessaria. A quali condizioni è possibile?, in *Giornale italiano di psicologia*, 2026, pp. 761-768

The publisher's version is available at:

<https://www.rivisteweb.it/doi/10.1421/121777>

When citing, please refer to the published version.

This full text was downloaded from UNICA IRIS <https://iris.unica.it/>

La supervisione umana dell'IA è necessaria. A quali condizioni è possibile?

Daniele Amoroso

e Guglielmo Tamburrini

14.000 battute spazi inclusi

1. Introduzione

Nel suo articolo bersaglio, Teresa Limata affronta il problema dei bias che possono affliggere sia i sistemi di intelligenza artificiale (IA) ad autonomia crescente sia gli esseri umani che rivestono ruoli significativi nel loro sviluppo o di supervisione del loro comportamento (Limata, 2026). Per mitigare i bias dell'IA e le loro conseguenze indesiderate — ad es. danni a persone e cose, violazione di diritti individuali, discriminazione di gruppi svantaggiati — l'autrice invoca un controllo umano significativo (d'ora in poi CUS per brevità) sui sistemi di IA; identifica ostacoli psicologici all'esercizio del CUS; e auspica un dialogo tra esperti di vari settori — dall'ingegneria alla psicologia, all'etica, al diritto e alle scienze sociali — per mettere a punto i contenuti e le condizioni per lo svolgimento adeguato del CUS.

Lo scopo del nostro commento al lavoro di Limata è duplice. Per prima cosa, cerchiamo di arricchire (nel paragrafo 2) le motivazioni che si possono addurre per imporre il CUS, in aggiunta alla fondamentale questione dei bias esplorata da Limata. Ricorderemo brevemente altre tipologie di errori, che riguardano in particolare le applicazioni conversazionali degli LLM (Large Language Models): allucinazioni, inaccuracy dei processi “Chain of Thought”, nonché il possibile disallineamento tra obiettivi della macchina e quelli degli utenti. La riflessione epistemologica sull'IA e sulle sue limitazioni attuali fornisce un ponte verso l'etica e il diritto dell'IA: il riconoscimento di siffatte limitazioni rafforza infatti le richieste etiche e giuridiche per imporre l'esercizio del CUS.

Il secondo scopo del nostro commento è di illustrare concretamente (nel paragrafo 3) forme di interazione umani-IA che erodono la possibilità di un esercizio genuino del

CUS, di una supervisione che sia davvero significativa e non solo nominale. Come sottolinea a più riprese Limata, sfruttare l'IA per *accelerare* processi percettivi, decisionali e di esecuzione delle azioni può risultare incompatibile con i vincoli che un agente umano deve soddisfare per applicare correttamente il repertorio di euristiche che caratterizzano la razionalità limitata *à la* Simon o i processi decisionali riflessivi conformi ai modelli duali del *decision-making*. Criticità del CUS dovute all'accelerazione mediante IA dei processi decisionali umani sono emerse nell'uso di alcuni sistemi di supporto alle decisioni recentemente impiegati in operazioni belliche.

La lezione appresa da casi concreti in ambito bellico indica più in generale (paragrafo 4) la centralità della psicologia dei processi decisionali e dei bias ad essi collegati per individuare forme di interazione umani-IA che soddisfino le condizioni per esercitare il CUS: dalle diagnosi cliniche allo sviluppo di codice informatico e alla generazione di notizie giornalistiche e ad altri ambiti applicativi per l'IA che sono eticamente e giuridicamente sensibili.

2. Sulla necessità del CUS

Limata rielabora il dibattito sul CUS che si è inizialmente sviluppato in relazione ai sistemi d'arma autonomi e ai problemi che essi pongono per il rispetto del Diritto internazionale umanitario (Amoroso e Tamburrini 2021). Il dibattito si è poi esteso all'esercizio del CUS su sistemi dell'IA ad autonomia crescente in altri ambiti applicativi, nonché a sistemi di supporto alle decisioni basati sull'IA (SSD_IA per brevità). Questi ultimi non godono, per definizione, di completa autonomia operativa e decisionale, essendo soggetti al filtro decisionale degli esseri umani.

Per rafforzare gli argomenti offerti da Limata sulla necessità del CUS — sostanzialmente incentrati sulla questione dei bias algoritmici — è utile ricordare altre limitazioni ed errori che affliggono ancora i sistemi di IA. Ci soffermiamo qui sulle allucinazioni (dette anche confabulazioni), le catene di pensiero (“Chain of Thought”,

CoT) inaccurate, il disallineamento tra gli obiettivi effettivamente appresi dal sistema e gli obiettivi intesi da progettisti e utenti.

Allucinazioni. Un'applicazione conversazionale di LLM può fornire informazioni false, presentate però come aderenti alla realtà. Siffatte “allucinazioni” traggono forse origine dall'obiettivo (posto nella fase iniziale di apprendimento non supervisionato che precede il fine-tuning) di indovinare completamenti plausibili di blocchi di testo, evitando di lasciare inevase le richieste degli utenti. Numerosi e ben noti sono gli esempi di allucinazioni calunniose, di riferimenti sentenze giudiziarie e ad articoli scientifici inesistenti, oppure ancora a romanzi mai pubblicati (Zhang et al. 2025).

Catene di pensiero inaccurate. Un'applicazione chatbot di LLM dotata della funzione CoT, anziché dare solo una risposta immediata e diretta, segue una serie di passaggi intermedi che dovrebbero rispecchiare passi salienti del processo svolto da un essere umano impegnato a risolvere un determinato problema (per es. percettivo, deduttivo o decisionale). Il CoT è stato introdotto anche come fonte di interpretazioni e spiegazioni delle risposte fornite dal sistema, atto a mitigare il problema dell'opacità semantica che affligge diffusamente le elaborazioni condotte dai sistemi di IA. Tuttavia, i passaggi salienti menzionati dal sistema possono rivelarsi come delle razionalizzazioni ex post, scollegate dai reali passaggi sviluppati dal sistema nel corso delle sue elaborazioni. Inoltre, il CoT può essere anche usato dal sistema per fornire giustificazioni a risposte scorrette che esso effettivamente fornisce (Arcuschin et al. 2025).

Disallineamento. È stato rilevato che i sistemi dell'IA possono agire in contrasto con le intenzioni di progetti e utenti. Le applicazioni conversazionali di LLM possono inoltre mentire a proposito dei loro obiettivi divergenti. Il disallineamento tra gli obiettivi effettivamente appresi e perseguiti dal sistema e gli obiettivi intesi da progettisti e utenti può dipendere, in una grande varietà di casi, da specifiche inadeguate del compito da svolgere, come accade nel caso di uno scolaro al quale si

indica lo scopo di ottenere buoni voti, e per raggiungere questo scopo copia dal vicino di banco, in quanto gli insegnanti non avevano non avevano esplicitato il divieto di copiare (Chen et al. 2025).

Una casistica così variegata ma non esaustiva arricchisce le argomentazioni basate sui bias addotte da Limata a sostegno dell'importanza e in molti casi della necessità del CUS; e insieme illustra quanto rapidamente, con lo sviluppo tecnologico, si sia stratificato e sfaccettato il problema dell'esercizio del CUS sui sistemi di IA.

3. Il CUS e l'accelerazione delle interazioni umani-IA

Il CUS appare importante e in molti casi necessario. Bisogna però chiedersi se e in che misura esso sia effettivamente applicabile. Una sfida evidenziata nell'articolo bersaglio riguarda le differenti scale temporali che caratterizzano rispettivamente i processi decisionali umani e dell'IA. L'accelerazione dei processi di elaborazione dell'informazione resa possibile dall'IA può compromettere anche l'esercizio del CUS. Un esempio concreto di tale situazione è stato offerto dall'uso che è stato recentemente fatto di alcuni SSD_IA in ambito bellico. L'incremento di velocità nei processi di decisione consentirebbe di ridurre il tempo necessario per passare dalla percezione della situazione sul campo di battaglia all'individuazione di un bersaglio da colpire e infine alla decisione di applicare la forza contro di esso. Il sistema Maven Smart System, che ingloba anche l'applicazione conversazionale di un LLM, ha consentito di individuare gli obiettivi di circa 1000 missioni di bombardamento eseguite nelle prime 24 dell'operazione Epic Fury iniziata il 28 febbraio 2026 dagli Stati Uniti e da Israele contro l'Iran. "Combattere alla velocità dell'IA" è uno slogan che ispira lo sviluppo di numerosi SSD_IA e di altre applicazioni dell'IA, come i sistemi d'arma autonomi, nel settore della difesa. Ma questa accelerazione della "kill chain" può comportare un prezzo molto alto da pagare in termini di mantenimento del CUS, di rischio di violazioni del Diritto internazionale umanitario e di escalation non volute dei conflitti.

Un'illustrazione concreta di tali rischi riguarda l'uso dei sistemi Habsora e Lavender fatto dall'IDF (Israeli Defense Forces) per individuare obiettivi da colpire nelle prime fasi del recente conflitto a Gaza. Secondo una nota ricostruzione giornalistica (Abraham 2023, Abraham 2024), Habsora e Lavender sarebbero software di supporto alle decisioni basati sull'IA programmati per assistere l'IDF nel processo di *targeting*: in particolare, Habsora avrebbe supportato l'identificazione dei beni qualificabili come obiettivi militari, mentre Lavender avrebbe operato in funzione antipersona, attribuendo agli individui residenti a Gaza un punteggio di probabile appartenenza all'ala militare di Hamas o della Jihad islamica quella degli individui appartenenti all'ala militare di Hamas.

È impressionante l'accelerazione dei processi decisionali degli addetti umani al targeting militare consentita dall'uso questi sistemi. Secondo la testimonianza degli addetti alla divisione targeting dell'IDF, dai 50 possibili obiettivi individuati "a mano" nel corso di un anno intero si è passati a 100 obiettivi al giorno suggeriti dagli SSD_IA. A seguito di questa innovazione tecnologica, gli stessi addetti hanno lamentato di dover lavorare rapidamente, come in una catena di montaggio, senza avere il tempo di approfondire la liceità dei bersagli indicati. In circostanze siffatte, si materializzano le condizioni psicologiche paventate da Limata, che possono indurre, secondo i modelli duali dei processi decisionali, a soppiantare processi decisionali riflessivi nell'esercizio del CUS e a sostituirli con processi euristici carichi di bias (Tamburrini 2025). Per esempio, i suggerimenti degli SSD_IA possono fornire in queste circostanze un'ancora dalla quale gli addetti al CUS stentano a discostarsi, oltre ad attivare eventualmente i bias collegati alle euristiche di disponibilità e WYSIATI. È proprio in questo scarto tra produzione algoritmica dei target e capacità umana di verificarne la liceità che si colloca il rischio di una degradazione delle garanzie poste dai principi di Diritto internazionale umanitario — distinzione, proporzionalità e precauzione — che governano i processi di *targeting* (Mauri 2024).

Gli addetti al targeting dell'IDF hanno anche denunciato in forma anonima i nuovi obiettivi di produttività decisionale definiti dai loro superiori per adeguare la velocità decisionale dei supervisori umani alla velocità dell'IA. Per non rappresentare colli di bottiglia in un combattimento da condurre “alla velocità dell'IA” gli addetti al targeting avrebbero dovuto dedicare 20 secondi alla valutazione di ammissibilità dell'obiettivo umano da colpire, limitandosi a stabilire se si trattava di una persona di sesso maschile o femminile, in quanto Hamas non ammette donne tra i membri della sua ala militare. L'assegnazione di un tempo prefissato — per di più particolarmente breve — per valutare la correttezza della raccomandazione del sistema di IA viola, in generale, le condizioni di applicabilità di algoritmi per la ricerca di soluzioni *satisficing*, i quali contemplano un procedimento iterativo di ricerca che abbassa progressivamente il livello di soddisfazione da perseguire finché non venga individuata una soluzione con il livello corrente di soddisfazione. Anche il criterio generale di basarsi su una sola proprietà per formulare un giudizio è stato esaminato nell'ambito degli studi sulla razionalità limitata à la Simon e sulla ricerca di soluzioni *satisficing* a problemi decisionali. L'uso euristico di una sola proprietà non è di per sé irrazionale, a patto che siano rispettate le condizioni “ecologiche” di contesto per la sua applicazione. Esso, infatti, consente di raggiungere in tempo minore risultati altrettanto buoni di processi decisionali più complessi, ma a condizione che il peso assegnato alla proprietà selezionata sia maggiore della somma di tutti i pesi assegnati alle altre proprietà che si potrebbero prendere in considerazione (Gigerenzer 2021, vedi p. 61 ss.). Questa condizione contestuale ovviamente non vale nel caso della proprietà “maschio” e del tipo di decisione da prendere: una parte certamente non trascurabile degli abitanti di sesso maschile nella striscia di Gaza non è membro dell'ala militare di Hamas. Un controllo così limitato, fondato su una proprietà non decisiva, difficilmente potrebbe assolvere alla funzione di verifica individualizzata richiesta dai principi di distinzione e precauzione, specie a fronte di

sistemi esposti al rischio non trascurabile di falsi positivi (10 %, secondo una fonte anonima dell'IDF, Abraham 2024).

4. Conclusioni

Il caso degli SSD_IA per l'individuazione di obiettivi da colpire nella Striscia di Gaza mostra, tra l'altro, che le motivazioni sociali, economiche, militari e politiche a favore di una maggiore velocità decisionale mediante i sistemi di IA non sono invariabilmente in armonia con l'esercizio genuino del CUS e con le motivazioni etiche e giuridiche che lo sottendono: prevenzione e retribuzione di danni a persone e cose, violazione di diritti individuali, discriminazione di gruppo e così via. Si può facilmente immaginare come la maggiore produttività decisionale abilitata dagli SSD_IA in medicina possa venire incontro alla domanda crescente di cure mediche; la maggiore produttività nello sviluppo di codice informatico grazie agli LLM possa sposarsi con le esigenze delle aziende di rispondere alle commesse in tempi sempre più brevi; e come l'accelerazione nel confezionare notizie possa giocare un ruolo nella competizione tra aziende nel settore dei media. Ma l'accelerazione produttiva in questi e altri settori non deve confliggere con l'esercizio del CUS e la prevenzione delle conseguenze indesiderabili alle quali una supervisione solo nominale può condurre.

L'importanza del CUS trova conferma — pur nella varietà delle formule impiegate — nell'evoluzione dei principali strumenti etici e giuridici in materia di IA. Le *Ethics Guidelines for Trustworthy AI* elaborate dal Gruppo di esperti ad alto livello sull'intelligenza artificiale nominato dalla Commissione europea includono la supervisione umana tra i requisiti essenziali dell'IA affidabile, precisando che essa può sì assumere forme diverse, ma deve in ogni caso consentire l'assunzione di decisioni informate e rispettose dei diritti fondamentali (European Commission (2019)). L'art. 14 dell'AI Act (Regolamento (UE) 2024/1689) traduce questa

indicazione in un requisito di progettazione e utilizzo dei c.d. “sistemi ad alto rischio”. La Convenzione quadro del Consiglio d’Europa sull’IA si muove nella stessa direzione, ancorché in modo meno esplicito. Il testo convenzionale, infatti, non prevede un autonomo requisito di supervisione umana, ma impone garanzie procedurali effettive quando i sistemi di IA incidano in modo significativo sul godimento dei diritti umani e l’Explanatory Report individua proprio nella supervisione umana la forma paradigmatica di attuazione di tali garanzie (Amoroso 2026). Persino nel settore, assai più controverso, dei sistemi d’arma autonomi, sembra addensarsi un consenso trasversale intorno alla necessità di un “context–appropriate human judgement and control”, tale da garantire il rispetto del DIU, e in particolare dei già richiamati principi di distinzione, proporzionalità e precauzione (Amoroso 2025).

Si tratta, a ben vedere, di strumenti che, pur ponendo il requisito della supervisione umana al centro dell’uso dei sistemi di IA, presentano ancora un certo grado di astrattezza. La sfida consiste allora nel dare concretezza applicativa a tale requisito, individuando le condizioni tecniche, psicologiche e organizzative che consentano al supervisore umano di esercitare un controllo effettivo. Da questo punto di vista, il problema del CUS rappresenta un esempio eloquente di come il dialogo auspicato da Limata tra esperti e saperi — ingegneria, psicologia, etica, diritto e scienze sociali — sia fondamentale per “ideare sistemi che siano a supporto sia della collettività sia dell’individuo” (Limata 2026).

Bibliografia

Abraham, Y. (2023, November 30). ‘A mass assassination factory’: Inside Israel’s calculated bombing of Gaza. +972 Magazine. <https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/>

Abraham, Y. (2024, April 3). 'Lavender': The AI machine directing Israel's bombing spree in Gaza. +972 Magazine. <https://www.972mag.com/lavender-ai-israeli-army-gaza/>

Amoroso, D. (2025). La disciplina dei sistemi d'arma autonomi nel diritto internazionale umanitario. Stato dell'arte e prospettive di regolamentazione. *Rivista italiana di informatica e diritto*, 7(2), 33–47. <https://doi.org/10.32091/RIID0237>

Amoroso, D. (2026). Human oversight as a procedural safeguard under the Council of Europe AI Framework Convention. *Sistemi intelligenti*, 38(1), 165 – 179. <https://doi.org/10.1422/120418>

Amoroso, D., & Tamburrini, G. (2021). Toward a normative model of meaningful human control over weapons systems. *Ethics & International Affairs*, 35(2), 245–272. <https://doi.org/10.1017/S0892679421000241>

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., & Conmy, A. (2025). *Chain-of-thought reasoning in the wild is not always faithful*. arXiv. <https://doi.org/10.48550/arXiv.2503.08679>

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning models don't always say what they think*. arXiv. <https://doi.org/10.48550/arXiv.2505.05410>

European Commission, High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Gigerenzer, G. (2021). What is bounded rationality? In R. Viale (Ed.), *Routledge handbook of bounded rationality* (pp. 55 – 69). Routledge. <https://doi.org/10.4324/9781315658353-2>

Limata, T. (2026). Sistemi autonomi e bias: Un dialogo interdisciplinare per comprendere e mitigare i rischi. *Giornale Italiano di Psicologia*, 53.

Mauri, D. (2024). Numeri, persone, umanità: Sistemi di supporto alle decisioni umane in campo militare da parte dell'IDF e diritto internazionale umanitario. *Diritti umani e diritto internazionale*, 18(2), 329–346. <https://doi.org/10.12829/114370>

Tamburrini, G. (2025). The heuristics gap in AI ethics: Impact on green AI policies and beyond. *Journal of Responsible Technology*, 21, Article 100104. <https://doi.org/10.1016/j.jrt.2024.100104>

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4), 1373–1418. <https://doi.org/10.1162/coli.a.16>