



## **Human–AI collectives make the most accurate medical diagnoses**

**Study highlights the potential for greater patient safety and more equitable healthcare**

**Artificial intelligence (AI) can effectively support doctors in making diagnoses. It makes different mistakes than humans—and this complementarity represents a previously untapped strength. An international team led by the Max Planck Institute for Human Development has now systematically demonstrated for the first time that combining human expertise with AI models leads to the most accurate open-ended diagnoses.**

Diagnostic errors are among the most serious problems in everyday medical practice. AI systems—especially large language models (LLMs) like ChatGPT-4, Gemini, or Claude 3—offer new ways to efficiently support medical diagnoses. Yet these systems also entail considerable risks—for example, they can “hallucinate” and generate false information. In addition, they reproduce existing social or medical biases and make mistakes that are often perplexing to humans.

An international research team, led by the Max Planck Institute for Human Development and in collaboration with partners from the Human Diagnosis Project (San Francisco) and the Institute of Cognitive Sciences and Technologies of the Italian National Research Council (CNR-ISTC Rome), investigated how humans and AI can best collaborate.

The result: hybrid diagnostic collectives—groups consisting of human experts and AI systems—are significantly more accurate than collectives consisting solely of humans or AI. This holds particularly for complex, open-ended diagnostic questions with numerous possible solutions, rather than simple yes/no decisions. "Our results show that cooperation between humans and AI models has great potential to improve patient safety," says lead author Nikolas Zöller, postdoctoral researcher at the Center for Adaptive Rationality of the Max Planck Institute for Human Development.

### **Realistic simulations using more than 2,100 clinical vignettes**

The researchers used data from the Human Diagnosis Project, which provides clinical vignettes—short descriptions of medical case studies—along with the correct diagnoses. Using more than 2,100 of these vignettes, the study compared the diagnoses made by medical professionals with those of five leading AI models. In the central experiment, various diagnostic collectives were simulated: individuals, human collectives, AI models, and mixed human–AI collectives. In total, the researchers analyzed more than 40,000 diagnoses. Each was classified and evaluated according to international medical standards (SNOMED CT).

### **Humans and machines complement each other—even in their errors**

The study shows that combining multiple AI models improved diagnostic quality. On average, the AI collectives outperformed 85% of human diagnosticians. However, there were numerous cases in which humans performed better. Interestingly, when AI failed, humans often knew the correct diagnosis.

The biggest surprise was that combining both worlds led to a significant increase in accuracy. Even adding a single AI model to a group of human diagnosticians—or vice versa—substantially improved the result. The most reliable outcomes came from collective



decisions involving multiple humans and multiple AIs.

The explanation is that humans and AI make systematically different errors. When AI failed, a human professional could compensate for the mistake—and vice versa. This so-called error complementarity makes hybrid collectives so powerful. “It’s not about replacing humans with machines. Rather, we should view artificial intelligence as a complementary tool that unfolds its full potential in collective decision-making,” says co-author Stefan Herzog, Senior Research Scientist at the Max Planck Institute for Human Development.

However, the researchers also emphasize the limitations of their work. The study only considered text-based case vignettes—not actual patients in real clinical settings. Whether the results can be transferred directly to practice remains a questions for future studies to address. Likewise, the study focused solely on diagnosis, not treatment, and a correct diagnosis does not necessarily guarantee an optimal treatment.

It also remains uncertain how AI-based support systems will be accepted in practice by medical staff and patients. The potential risks of bias and discrimination by both AI and humans, particularly in relation to ethnic, social, or gender differences, likewise require further research.

#### **Wide range of applications for hybrid human–AI collectives**

The study is part of the Hybrid Human Artificial Collective Intelligence in Open-Ended Decision Making (HACID) project, funded under Horizon Europe, which aims to promote the development of future clinical decision-support systems through the smart integration of human and machine intelligence. The researchers see particular potential in regions where access to medical care is limited. Hybrid human–AI collectives could make a crucial contribution to greater healthcare equity in such areas.

“The approach can also be transferred to other critical areas—such as the legal system, disaster response, or climate policy—anywhere that complex, high-risk decisions are needed. For example, the HACID project is also developing tools to enhance decision-making in climate adaptation” says Vito Trianni, co-author and coordinator of the HACID project.

#### **In brief:**

- Hybrid diagnostic collectives consisting of humans and AI make significantly more accurate diagnoses than either medical professionals or AI systems alone—because they make systematically different errors that cancel each other out.
- The study analyzed over 40,000 diagnoses made by humans and machines in response to more than 2,100 realistic clinical vignettes.
- Adding an AI model to a human collective—or vice versa—noticeably improved diagnostic quality; hybrid collective decisions made by several humans and machines achieved the best results.
- These findings highlight the potential for greater patient safety and more equitable healthcare, especially in underserved regions. However, further research is needed on practical implementation and ethical considerations.



**Original publication:**

Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., Laskowski, K., Shia, V., Harack, B., Chu, E. A., Trianni, V., Kurvers, R. H. J. M., & Herzog, S. M. (2025). Human-AI collectives most accurately diagnose clinical vignettes. *Proceedings of the National Academy of Sciences of the United States of America*, 122(24), Article e2426153122.  
<https://doi.org/10.1073/pnas.2426153122>

**Max Planck Institute for Human Development**

The Max Planck Institute for Human Development in Berlin was founded in 1963. It is an interdisciplinary research institution dedicated to the study of human development. The Institute belongs to the Max Planck Society for the Advancement of Science, one of the leading organizations for basic research in Europe.

**Contact:**

Max Planck Institute for Human Development  
Public Relations Department

Maria Einhorn  
Phone: +49 (0) 30 82406-211  
[ehorn@mpib-berlin.mpg.de](mailto:ehorn@mpib-berlin.mpg.de)

Nicole Siller  
Phone: +49 (0) 30 82406-284  
E-Mail: [siller@mpib-berlin.mpg.de](mailto:siller@mpib-berlin.mpg.de)

**Further Information:**

[www.mpib-berlin.mpg.de/en](http://www.mpib-berlin.mpg.de/en)  
[www.mpg.de/en](http://www.mpg.de/en)



## Consiglio Nazionale delle Ricerche Unità Ufficio Stampa

COMUNICATO STAMPA 47/2025

### Diagnosi mediche più accurate grazie a team di medici e IA

*Esperti umani e sistemi basati su intelligenza artificiale (IA) possono fare squadra per ridurre gli errori diagnostici, contribuendo ad una maggiore efficacia ed equità dell'assistenza sanitaria. È quanto riporta uno studio internazionale guidato dal Max Planck Institute for Human Development, in collaborazione con lo Human Diagnosis Project (Human Dx) e con l'Istituto di scienze e tecnologie della cognizione del Cnr, dimostrando che la combinazione delle competenze umane con la capacità dei sistemi IA contemporanei porta a diagnosi significativamente più accurate. Lo studio è pubblicato sulla rivista PNAS*

Sappiamo già che l'IA può supportare efficacemente i medici nella formulazione delle diagnosi; la novità consiste nello scoprire che IA ed esperti umani "sbagliano" in maniera differente: una complementarità che rappresenta un punto di forza in precedenza inesplorato. Lo rivela uno studio internazionale guidato dal Max Planck Institute for Human Development, in collaborazione con lo Human Diagnosis Project (Human Dx) e con l'Istituto di scienze e tecnologie della cognizione del Consiglio nazionale delle ricerche (Cnr-Istc): ha dimostrato che la combinazione delle competenze umane con la capacità dei sistemi IA contemporanei porta a diagnosi significativamente più accurate. Lo studio è stato pubblicato sulla prestigiosa rivista PNAS.

Gli errori diagnostici sono tra i problemi più gravi nella pratica medica quotidiana. I sistemi basati su intelligenza artificiale (IA)—in particolare i modelli linguistici di grandi dimensioni (LLM) come ChatGPT-4, Gemini o Claude 3—offrono nuovi modi per supportare efficacemente le diagnosi mediche. Tuttavia, questi sistemi comportano anche rischi considerevoli: ad esempio, possono "allucinare" e generare false informazioni. Inoltre, riproducono pregiudizi osservabili in ambito sia medico che sociale, e a volte commettono errori che lasciano perplessi gli utenti più esperti.

Il team ha studiato come esperti umani e sistemi IA possano collaborare al meglio: il risultato è che i team ibridi - gruppi composti sia da esperti umani che da sistemi IA - sono significativamente più accurati dei team composti esclusivamente da esperti umani o da sistemi IA. Questo vale in particolare per quesiti diagnostici complessi e aperti con numerose possibili soluzioni, piuttosto che per semplici decisioni binarie. "I nostri risultati dimostrano che la cooperazione tra esperti umani e sistemi IA ha un grande potenziale per migliorare la diagnosi medica, e quindi la salute dei pazienti", afferma l'autore principale Nikolas Zöllner, ricercatore post-dottorato presso il Center for Adaptive Rationality del Max Planck Institute for Human Development.

I ricercatori hanno utilizzato i dati forniti da Human Dx, che associano a vignette cliniche - brevi descrizioni testuali di casi clinici - alle diagnosi corrette. Utilizzando oltre 2.100 di queste vignette, lo studio ha confrontato le diagnosi formulate da professionisti medici con quelle di cinque sistemi IA. Nell'esperimento principale, sono stati simulati diversi team diagnostici: individui, team umani, team di IA e team misti umani-IA. In totale, il team ha analizzato

**Ufficio stampa Cnr:** Francesca Gorini, [francesca.gorini@cnr.it](mailto:francesca.gorini@cnr.it) cell. 329.3178725; **Responsabile:** Emanuele Guerrini, [emanuele.guerrini@cnr.it](mailto:emanuele.guerrini@cnr.it), cell. 339.2108895; **Segreteria:** [ufficiostampa@cnr.it](mailto:ufficiostampa@cnr.it), tel. 06.4993.3383 - P.le Aldo Moro 7, Roma



## Consiglio Nazionale delle Ricerche

### Unità Ufficio Stampa

oltre 40.000 diagnosi: ciascuna è stata classificata e valutata attraverso standard medici internazionali (SNOMED CT).

Lo studio dimostra che la combinazione di più sistemi IA ha migliorato la qualità diagnostica. In media, i team composti da sistemi IA hanno ottenuto risultati migliori rispetto ai team composti da esperti umani nell'85% dei casi. Tuttavia, ci sono stati numerosi casi in cui gli esperti umani hanno ottenuto risultati migliori. È interessante notare che, quando i sistemi IA falliscono, gli esperti umani spesso forniscono la diagnosi corretta. Il risultato più importante è che la combinazione di entrambi i mondi porta ad un aumento significativo dell'accuratezza. Anche l'aggiunta di un singolo sistema IA ad un gruppo di esperti umani - o viceversa - porta a migliorare sostanzialmente l'accuratezza. I risultati più affidabili sono emersi da decisioni collettive che coinvolgono allo stesso tempo più esperti umani e più IA.

La spiegazione è che gli esperti umani e i sistemi IA commettono errori sistematicamente diversi. Quando l'IA fallisce, un esperto umano può compensare l'errore, e viceversa. Questa complementarietà degli errori rende i team ibridi così potenti. "Non si tratta di sostituire gli esseri umani con le macchine. Piuttosto, dovremmo considerare l'intelligenza artificiale come uno strumento complementare che dispiega appieno il suo potenziale nel processo decisionale collettivo", aggiunge il coautore Stefan Herzog, ricercatore senior presso il Max Planck Institute for Human Development.

Tuttavia, i ricercatori sottolineano anche i limiti del loro lavoro. Lo studio ha preso in considerazione solo casi clinici basati su testo, non pazienti reali in contesti clinici reali. Se i risultati possano essere trasferiti direttamente alla pratica clinica rimane una questione che gli studi futuri dovranno affrontare. Inoltre, lo studio si limita alla fase diagnostica senza affrontare le conseguenze terapeutiche, e una diagnosi corretta non garantisce di arrivare ad un trattamento ottimale. Rimane, inoltre, incerto come i sistemi di supporto basati sull'IA saranno accettati nella pratica dal personale medico e dai pazienti. Anche i potenziali rischi di pregiudizi e discriminazioni da parte sia dell'IA che degli esperti umani, in particolare in relazione a differenze etniche, sociali o di genere, richiedono ulteriori ricerche.

Lo studio fa parte del progetto Hybrid Human Artificial Collective Intelligence in Open-Ended Decision Making (HACID), finanziato nell'ambito di Horizon Europe, che mira a promuovere lo sviluppo di futuri sistemi di supporto alle decisioni cliniche attraverso l'integrazione di intelligenza umana e artificiale. I ricercatori intravedono un potenziale sviluppo per migliorare l'assistenza medica nelle regioni in cui l'accesso ad alte professionalità è limitato. I gruppi ibridi potrebbero dare un contributo cruciale a una maggiore equità nell'assistenza sanitaria in tali aree. "L'approccio può essere trasferito anche ad altri ambiti professionali, come il sistema legale, la risposta alle emergenze da disastri ambientali o le politiche per il clima, ovvero ovunque siano necessarie decisioni complesse e ad alto impatto. Ad esempio, il progetto HACID sta anche sviluppando strumenti per migliorare il processo decisionale nello sviluppo di strategie di adattamento al cambiamento climatico", afferma Vito Trianni, dirigente di ricerca del Cnr-Istc e coordinatore del progetto HACID.

**Ufficio stampa Cnr:** Francesca Gorini, [francesca.gorini@cnr.it](mailto:francesca.gorini@cnr.it) cell. 329.3178725; **Responsabile:** Emanuele Guerrini, [emanuele.guerrini@cnr.it](mailto:emanuele.guerrini@cnr.it), cell. 339.2108895; **Segreteria:** [ufficiostampa@cnr.it](mailto:ufficiostampa@cnr.it), tel. 06.4993.3383 - P.le Aldo Moro 7, Roma



# Consiglio Nazionale delle Ricerche

## Unità Ufficio Stampa

Roma, 20 giugno 2025

### La scheda

**Chi:** Max Planck Institute for Human Development; Human Diagnosis Project (Human Dx); Istituto di scienze e tecnologie della cognizione del Consiglio nazionale delle ricerche di Roma (Cnr-Istc)

**Che cosa :** articolo “Human-AI Collectives Most Accurately Diagnose Clinical Vignettes”, N. Zöllner, J. Berger, I. Lin, N. Fu, J. Komarneni, G. Barabucci, K. Laskowski, V. Shia, B. Harack, E. A. Chu, V. Trianni, R. H.J.M. Kurvers, S. M. Herzog, Human-AI collectives most accurately diagnose clinical vignettes, Proc. Natl. Acad. Sci. U.S.A. 122 (24) e2426153122, <https://doi.org/10.1073/pnas.2426153122> (2025), link [www.pnas.org/doi/10.1073/pnas.2426153122](https://www.pnas.org/doi/10.1073/pnas.2426153122) .

**Per informazioni:** Vito Trianni, Cnr-Istc, [vito.trianni@istc.cnr.it](mailto:vito.trianni@istc.cnr.it), tel. 06.44595277, cell. 348.3925101 (recapiti per uso professionale da non pubblicare)

### Seguici su



ALMANACCO  
DELLA SCIENZA

**Il Cnr ti aspetta anche su WhatsApp! Clicca [qui](#) per seguire il Canale, oppure inquadra il QR CODE**



**Ufficio stampa Cnr:** Francesca Gorini, [francesca.gorini@cnr.it](mailto:francesca.gorini@cnr.it) cell. 329.3178725; **Responsabile:** Emanuele Guerrini, [emanuele.guerrini@cnr.it](mailto:emanuele.guerrini@cnr.it), cell. 339.2108895; **Segreteria:** [ufficiostampa@cnr.it](mailto:ufficiostampa@cnr.it), tel. 06.4993.3383 - P.le Aldo Moro 7, Roma