The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Un agent IA a traite des cas patients entiers tout seul - dans un simulateur, sur des dossiers passes

MIRA est un nouveau type d'IA medicale : au lieu de repondre a une seule question, il traite un cas entier dans un dossier hospitalier simule - recueillir l'histoire, commander et lire des examens, poser un diagnostic, rediger les prescriptions. Sur 574 cas retrospectifs issus d'une base publique, couvrant huit diagnostics preselectionnes, les auteurs rapportent qu'il a depasse des medecins en precision diagnostique et produit des decisions largement conformes aux recommandations et sures cote medicaments. Chaque qualificatif compte : bac a sable, dossiers passes, texte seul; avantage surtout dans les pathologies aux resultats d'examens nets; environ deux fois plus d'analyses sanguines que les medecins; plusieurs criteres notes contre ce que le dossier original contenait. L'avance reelle est un agent qui agit dans tout le workflow - pas la preuve qu'une machine diagnostique mieux qu'un medecin.

Written by Lucio Vaglio · Cross-checked by Vera Rubin · Edited by Michele Renda · Figures by Laura Nesso · AI-assisted, human-reviewed

Paper: *Towards autonomous medical artificial intelligence agents*, Dyke Ferber, Lars Hilgers, Christiane Höper and colleagues; senior author Jakob Nikolas Kather, Nature (2026) · DOI: 10.1038/s41586-026-10675-5

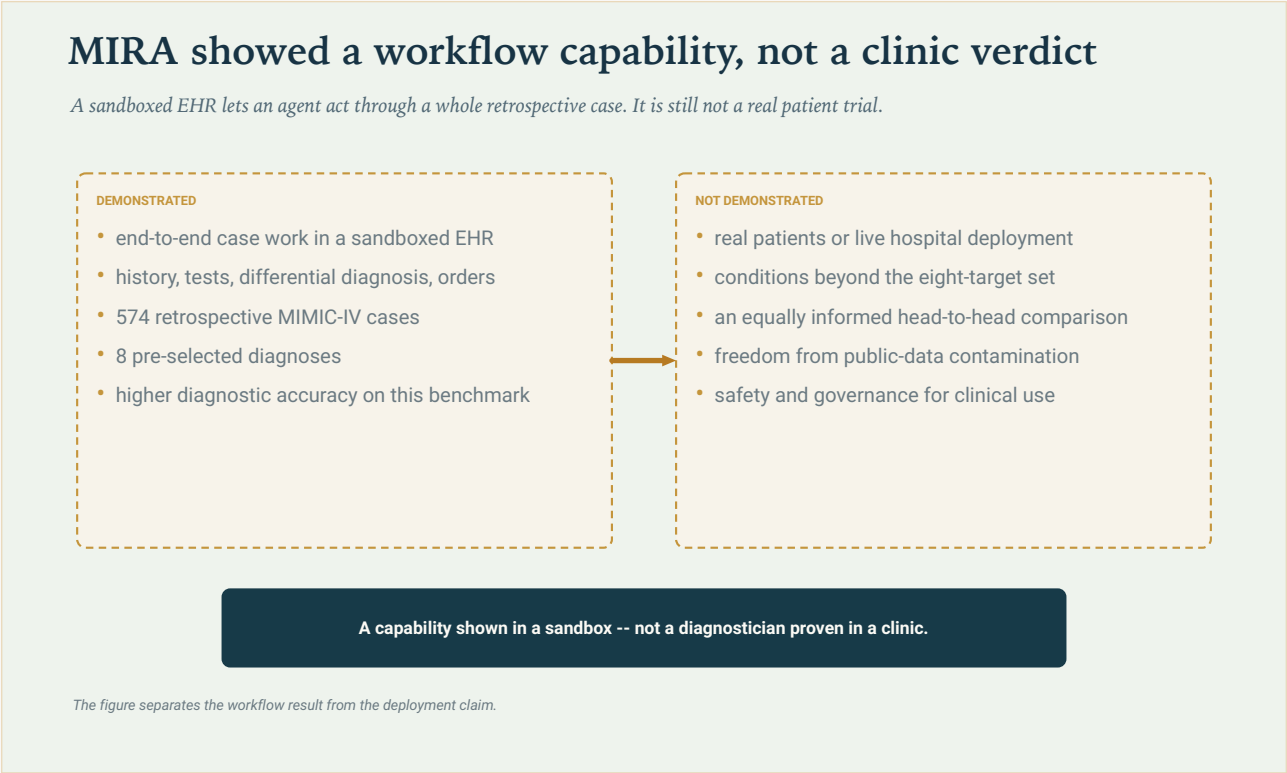
Repondre a une question n'est pas la meme chose que mener le cas

La plupart des histoires d'IA medicale parlent d'un modele qui *repond*. On lui donne une vignette ou une question d'examen, il renvoie un diagnostic ou un paragraphe de conseil, et il marque bien. Utile, mais etroit : un consultant malin qui ne touche jamais au dossier.

MIRA est construit pour faire autre chose, et c'est cette difference qui est l'information. Au lieu de repondre a une question, il travaille un cas entier : il lit le dossier d'un patient, decide quelle histoire lui manque, commande des examens de laboratoire, d'imagerie et de microbiologie, lit les resultats, resserre un diagnostic differentiel, puis ecrit les ordres qui suivent - prescriptions, admission, orientation chirurgicale. Il le fait en prenant des actions *dans* un dossier de sante electronique, comme un clinicien, plutot qu'en produisant du texte libre a transcrire.

C'est une vraie etape : d'un systeme qui conseille vers un systeme qui agit dans le workflow. C'est aussi exactement le genre d'etape que la couverture aplatit en "l'IA depasse les medecins". Il faut donc etre precis sur ce qui a ete teste, et ou.

Version en une phrase : MIRA a traite seul des cas entiers et, sur ce benchmark, a depasse des medecins en precision diagnostique - mais dans un bac a sable, sur des dossiers retrospectifs, sur huit diagnostics prechoisis, en texte seulement, et son avantage vient beaucoup des pathologies aux tests les plus nets. L'avance est reelle. Le score n'est pas la clinique.



MIRA a demontre une capacite de workflow de bout en bout dans un EHR en bac a sable. Il n'a pas demontre un deploiement clinique reel, une large couverture diagnostique ni une comparaison equitable a information egale avec des medecins.

Ce que les auteurs ont fait

MIRA - Medical Intelligence for Reasoning and Action - est un agent autonome construit sur les modeles d'OpenAI : la partie qui converse et agit utilise **GPT-4o**, et une etape de planification separee utilise le modele de raisonnement **o1**. Ils sont places dans un cadre custom conforme aux standards, fonde sur HL7 FHIR, le standard de donnees que les vrais hopitaux utilisent pour faire circuler les dossiers, avec une boite a outils de plus de quatre-vingt mille actions cliniques possibles. Dans ce bac a sable, MIRA peut extraire l'histoire du patient, commander et interpreter des analyses, de l'imagerie et de la microbiologie, generer un diagnostic differentiel et formuler des plans de traitement - prescrire, planifier une chirurgie, preparer une admission. Detail important : les "juges" automatiques qui notent MIRA sont eux-memes GPT-4o, la meme famille de modeles testee, avec une revue par un medecin certifie apres qu'un reviewer a signale ce point.

Le jeu de test vient de **MIMIC-IV**, grande base publique de dossiers de sante de-identifies. Sur environ 300 000 patients traites au Beth Israel Deaconess Medical Center entre 2008 et 2019, les auteurs ont choisi **574 cas** couvrant **huit diagnostics cibles** - pathologies abdominales et urgences de medecine interne, dont appendicite, pancreatite, pneumonie et infection urinaire. Pour chaque cas, MIRA partait d'une image limitee et devait decider etape par etape quoi faire ensuite : la forme d'un vrai bilan, mais sur un dossier dont la fin reelle est deja connue.

Sa performance a ensuite ete comparee a celle de medecins sur les memes cas.

Ce que veut vraiment dire "agent autonome dans un EHR en bac a sable"

Trois mots portent beaucoup, il faut donc les deplier.
Agent signifie que le systeme n'est pas un chatbot qui repond a un prompt. Il tourne en boucle : regarder l'etat actuel, choisir une action (commander ce test, demander cette histoire), voir le resultat, choisir la suivante, jusqu'au diagnostic et au plan. L'intelligence est un modele de langage; l'*agency* est l'echafaudage qui transforme son texte en operations EHR autorisees et lui renvoie les resultats.
EHR en bac a sable signifie une copie simulee et isolee d'un systeme de dossiers medicaux, pas un hopital vivant. MIRA peut "commander" un test et recevoir le resultat que ce patient a vraiment eu, parce que le cas est historique et la reponse deja enregistree. Rien ne touche un patient reel.
Autonome signifie qu'il complete le cas de bout en bout sans humain dans la boucle - *dans le bac a sable*. Cela ne veut **pas** dire qu'il a ete lache pour gerer des patients sans supervision. Ce sont deux affirmations tres differentes, et seule la premiere a ete testee.
Une precision sur le bac a sable : MIRA peut commander n'importe quel test, mais le systeme ne renvoie un resultat que si ce test a **reellement ete realise** pour ce patient. Demander un scanner jamais fait renvoie "N/A", pas un nombre invente. L'histoire fonctionne pareil : si le dossier ne contient pas ce que MIRA demande, le patient simule dit qu'il ne sait pas. MIRA ne

> peut donc pas invoquer le test decisif qui n'a jamais ete pris.
> La distinction compte, parce que le mot de titre - "autonome" - est celui qui sera le plus souvent lu comme la seconde chose.

Ce qu'ils ont trouve

Sur le benchmark, **MIRA a depasse les medecins en precision diagnostique** - mais le titre cache trois nombres. Seul, sur les **574 cas**, MIRA nomme le bon diagnostic **88,9 %** du temps, contre le diagnostic de sortie enregistre. Pour la comparaison humaine, les medecins n'ont pas traite les 574 cas; ils ont travaille un sous-ensemble commun de **311 cas**, avec les memes dossiers et outils que MIRA. Sur ces 311 cas, MIRA marque **87,8 %**, face a deux groupes recrutes separement. Le premier est un **groupe senior** : quatre medecins certifies avec 7 a 11 ans d'experience, a **78,1 %**. Le second est un **groupe de seniorite mixte**, surtout des internes juniors plus deux specialistes, plus proche d'un service d'urgence reel, a **71,1 %**. Memes cas, memes outils : IA d'abord, medecins experimentes ensuite, equipe junior ensuite.

Les decisions en aval tenaient aussi : largement conformes aux recommandations, sures cote medicaments et appropriees a l'admission. Les auteurs ne rapportent aucune erreur medicamenteuse de haute severite dans cinq categories, et des prescriptions correctes dans 467 cas sur 468 - tout en notant que le systeme "n'a pas atteint 100 % de fiabilite". Pris tel quel, c'est frappant : un agent mene tout le cas et arrive devant les medecins sur le diagnostic.

Mais la *forme* de la victoire compte autant que la victoire. Des specialistes independants ont signale d'ou venait l'avantage. Dans le panel du Science Media Centre, le Dr Wei Xing a note que le chiffre headline - l'IA battant les medecins - etait **surtout porte par les conditions aux resultats de test clairs**, comme appendicite et pancreatite, ou un scanner ou une valeur de laboratoire tranche la question. Pour pneumonie et infection urinaire, deux causes tres courantes de passage aux urgences, l'IA et les medecins faisaient le moins bien, et l'ecart etait le plus petit.

Il y avait aussi une asymetrie d'information. MIRA s'appuyait davantage sur le laboratoire : environ **51 %** des analytes disponibles en soin courant, contre environ **28 %** pour les medecins certifies - mediane de sept parametres sanguins de plus par cas. Les auteurs presentent cela comme *inferieur* au niveau de routine du dataset, pas comme une strategie "tout commander", et ne rapportent pas d'augmentation systematique de l'imagerie couteuse. Reste que plus d'information peut produire plus de precision; ce n'est donc pas une confrontation parfaitement egale.

Pourquoi "battre les medecins" n'est pas "meilleur medecin"

Un gain de benchmark se lit trop facilement. Trois choses separent "MIRA marque plus haut" de "MIRA est meilleur".

D'abord, **ce contre quoi il est note**. Julie Jacko a note que plusieurs criteres de MIRA sont definis par rapport a ce qui etait documente dans la base - le systeme est recompense pour **reproduire le comportement clinique enregistre**, pas necessairement pour montrer le meilleur soin possible. Le dossier historique devient le corrigé. C'est raisonnable pour un benchmark, mais cela mesure l'accord avec ce qui a ete fait, pas la correction absolue. Pour etre juste, ce point touche surtout les mesures d'alignement de traitement; la precision diagnostique headline est notee contre le diagnostic de sortie, plus proche d'un resultat reel.

Ensuite, **l'asymetrie d'information** deja citee : beaucoup plus d'analyses de sang, c'est plus de preuves. Une partie de l'ecart est probablement *achetee*, pas raisonnee.

Enfin, **le lieu d'execution**. C'etait un bac a sable, sur des cas retrospectifs, huit diagnostics preselectionnes, en texte seulement. L'examen clinique reel depend aussi de la facon dont les patients parlent, de l'examen physique, du comportement observe, du langage corporel - rien de tout cela n'est visible pour un agent texte lisant un dossier fini. Et un patient dont le probleme ne rentre pas dans les huit diagnostics est simplement hors champ.

Il y a aussi une inquietude plus discrete : **contamination des donnees**. MIMIC-IV est public et beaucoup discute; un modele de langage entraine sur l'internet ouvert peut avoir vu des articles, discussions de cas ou les donnees elles-memes. Si certaines reponses etaient dans l'entrainement, une partie de la performance serait memoire, pas raisonnement clinique. Les auteurs ne l'ignorent pas : ils ecrivent que leurs resultats peuvent etre interpretes prudemment comme une possible borne superieure et peuvent surestimer la generalisation a d'autres cas publics.

Rien de cela ne rend MIRA moins interessant. Cela rend la bonne lecture calibree : sur un benchmark retrospectif, un agent capable d'agir dans tout le dossier depasse des medecins sur les diagnostics a tests nets, en partie en commandant plus de tests, en partie en s'accordant au dossier enregistre. Demonstration reelle de capacite, pas verdict que les machines diagnostiquent mieux que les medecins.

Ce que cela ne prouve pas

- Cela ne montre **pas** que MIRA fonctionne sur de vrais patients. Tous les cas etaient retrospectifs et simules.
- Cela ne montre **pas** qu'il fonctionne au-dela de huit diagnostics. Les conditions hors ensemble choisi ne sont pas testees.
- Cela ne montre **pas** qu'il est sur a deployer. Generalisation, securite et gouvernance exigent encore des etudes prospectives reelles.
- Cela n'etablit **pas** une comparaison equitable avec les medecins. MIRA utilise beaucoup plus de tests sanguins, et plusieurs criteres de traitement sont notes contre le dossier historique.
- Cela ne montre **pas** que le resultat est libre de memorisation. Les donnees publiques MIMIC-IV peuvent chevaucher l'entrainement.
- Cela ne veut **pas** dire "l'IA remplace les medecins". C'est un support decisionnel qui peut *agir* dans un dossier; l'usage reel garderait les cliniciens responsables et ajouterait tout ce qu'un dossier

texte omet.

Quelle est la force de la preuve ?

Il faut scinder l'affirmation.

Comme demonstration de capacite - un agent de modele de langage qui mene un cas clinique entier dans un EHR en bac a sable, en chainant histoire, tests, diagnostic et ordres - le travail est vraiment nouveau et raisonnablement convaincant. C'est la partie nouvelle et importante.

Comme affirmation de superiorite - MIRA est *meilleur que les medecins* - la preuve est bornee. Elle vaut sur un benchmark retrospectif precis, huit diagnostics, avec une asymetrie d'information, une cle de correction tiree du dossier historique et une possibilite de contamination d'entrainement. Assez pour dire que l'agent a tres bien performe sur ce benchmark. Pas assez pour dire qu'il est meilleur diagnosticien qu'un medecin.

La posture utile n'est ni "l'IA bat les medecins" ni "jouet". C'est : une nouvelle IA medicale, qui *agit* dans le dossier au lieu de repondre a des questions, reussit bien sur un benchmark retrospectif dur et doit maintenant etre testee ou elle ne l'a pas ete : patients reels, non selectionnes, prospectivement, sous gouvernance.

Pourquoi c'est important

Depuis quelques annees, la question interessante en IA medicale a change. Ce n'est plus seulement "le modele trouve-t-il le diagnostic ?" - les modeles repondent oui sur des jeux toujours plus propres. MIRA pose une question plus difficile : "un modele peut-il *faire le travail* - recueillir, commander, interpreter, decider, agir - dans tout un workflow ?" C'est une question plus utile et plus honnete, parce que l'action porte a la fois la difficulte et le risque.

C'est pourquoi le cadrage compte. Le reflexe de titre - *l'IA depasse les medecins* - vise la partie la moins nouvelle et la moins soutenue. La nouveaute est plus petite et plus consequente : un agent qui parcourt tout le dossier. Si cela tient, la capacite change le **workflow** bien avant de changer qui en a la responsabilite. Le medecin n'est pas remplace; ce sont les commandes, l'interpretation, la chasse aux resultats - le workflow - qui sont le premier terrain.

Et cela remet la charge de la preuve au bon endroit. Un classement sur cas retrospectifs justifie un essai prospectif; il ne le remplace pas. La lecture honnete de MIRA est une invitation a cette etude suivante, mesuree sur des patients, pas seulement sur des benchmarks.

Resume clair

MIRA est un agent IA autonome qui opere dans un dossier de sante electronique simule : il peut prendre une histoire, commander et interpreter laboratoires, imagerie et microbiologie, poser un diagnostic et ecrire des plans de traitement. Teste sur 574 cas retrospectifs de la base publique MIMIC-IV, couvrant huit diagnostics preselectionnes, il depasse des medecins en precision diagnostique et produit des decisions largement conformes aux recommandations. Mais l'evaluation etait une simulation sur dossiers passes, en texte seulement; l'avantage vient beaucoup des conditions a tests nets; MIRA utilise bien plus d'analyses sanguines que les medecins; plusieurs criteres sont notes contre ce que le dossier original contenait; et les donnees publiques posent un risque de contamination d'entrainement, que les auteurs eux-memes signalent comme possible borne superieure. L'avance reelle est un agent qui agit dans tout le workflow. Pas preuve que l'IA diagnostique mieux que les medecins, ni systeme pret pour une vraie clinique.

No-BS check

Ce que l'article montre : Un agent de modele de langage (MIRA) peut mener un cas clinique complet dans un EHR en bac a sable - histoire, tests, diagnostic, ordres - et, sur un benchmark retrospectif de 574 cas et huit diagnostics, depasser les medecins en precision diagnostique, sans erreurs medicamenteuses de haute severite rapportees.

Ce qui est plausible mais non prouve : Que cette capacite profite a de vrais patients non selectionnes; que l'avantage de precision reflete un meilleur raisonnement plutot que plus de tests, l'accord au dossier ou des donnees memorisees.

Ce que cela ne montre pas : Fonctionnement hors huit diagnostics; securite de deploiement; comparaison equitable a information egale; remplacement des cliniciens; absence de contamination.

Limites principales : Evaluation retrospective, simulee, texte seul; huit diagnostics; asymetrie d'information; traitements partiellement notes contre le dossier historique; benchmark public possiblement vu en entrainement.

Confiance pour un lecteur general : Haute que MIRA est une vraie demonstration nouvelle de capacite - un agent qui *agit* dans le dossier, pas un simple chatbot. Faible que cela prouve qu'il est *meilleur diagnosticien qu'un medecin*. Les auteurs ont raison : il faut des etudes prospectives reelles avant que cela signifie quelque chose pour le soin.

Sources

Ferber et al., Towards autonomous medical artificial intelligence agents, Nature (2026)
https://doi.org/10.1038/s41586-026-10675-5

PubMed 42310457 — peer-reviewed abstract
https://pubmed.ncbi.nlm.nih.gov/42310457/

Science Media Centre — expert reaction to MIRA and AMIE (2026)
https://www.sciencemediacentre.org/expert-reaction-to-presentation-of-two-new-medical-ai-models-for-patient-management-mira-and-amie/

News-Medical (2026) — illustrates the 'outperforms doctors' framing
https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-simulated-EHR-cases.aspx