



The CLEAN PAPER

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

Að svara spurningu er ekki það sama og að reka allt sjúkratilvikið

MIRA vann heil sjúkratilvik inni í hermdu rafrænu sjúkraskrárkerfi: safnaði upplýsingum, pantaði próf, las niðurstöður og setti fram greiningu og fyrmæli. Á 574 afturvirkum tilfellum fékk kerfið háa greiningarnákvæmni, en það var textabundinn sandkassi á fortíðargögnum, pantaði fleiri blóðpróf og var að hluta metið gegn upprunalegri sjúkraskrá. Þetta er sýning á end-to-end getu, ekki sönnun þess að vél sé betri læknir í raunverulegri þjónustu.

Written by Lucio Vaglio · Cross-checked by Vera Rubin · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · 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

Að svara spurningu er ekki það sama og að reka allt sjúkratilvikið

Flestar sögur um gervigreind í læknisfræði snúast um líkan sem **svarar**. Þú gefur því tilfelli eða prófspurningu, það skilar greiningu eða ráðleggingu og fær gott skor. Gagnlegt, en þröngt: snjall ráðgjafi sem snertir aldrei sjúkraskrána.

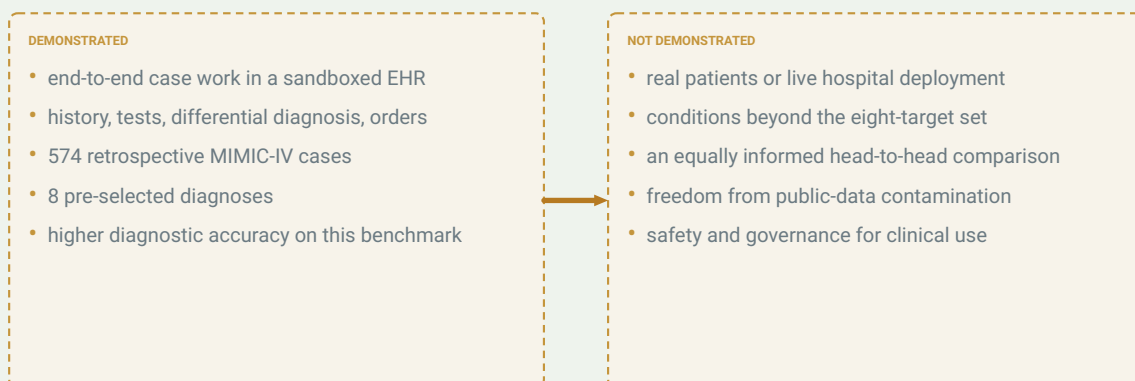
MIRA er byggt til að gera annað, og það er raunverulega nýjungin. Í stað þess að svara einni spurningu vinnur kerfið **allt tilfellið**: les sjúkraskrá, ákveður hvaða sögu vantar, pantar rannsóknarstofupróf, myndgreiningu og örverufræði, les niðurstöðurnar, þrengir mismunagreiningu og skrifar síðan fyrir-mælin sem fylgja — lyfseðla, innlögn eða tilvísun í skurðaðgerð. Það gerir þetta með aðgerðum **inni í rafrænni sjúkraskrá**, eins og læknir, í stað þess að skrifa frjálsan texta sem manneskja þarf svo að færa inn.

Þetta er raunverulegt skref: frá kerfi sem ráðleggur yfir í kerfi sem framkvæmir aðgerðir yfir allt vinnuferlið. En þetta er líka nákvæmlega sú tegund niðurstöðu sem verður auðveldlega að fyrirsögninni „AI slær lækna“. Því skiptir máli að segja nákvæmlega hvað var prófað og við hvaða aðstæður.

Í einni setningu: MIRA vann heil tilfelli sjálfstætt og fékk hærri greiningarnákvæmni en læknar á þessu viðmiði — **en** það gerði það í sandkassa, á afturvirkum sjúkraskrá, yfir átta fyrirfram valdar greiningar, aðeins með textasamskiptum, og stór hluti forskotsins kom í sjúkdómum með skýr prófniðurstöður. Framfaraskrefið er raunverulegt. Stigataflan er ekki sjúkrahúsið.

MIRA showed a workflow capability, not a clinic verdict

A sandboxed EHR lets an agent act through a whole retrospective case. It is still not a real patient trial.



A capability shown in a sandbox -- not a diagnostician proven in a clinic.

The figure separates the workflow result from the deployment claim.

MIRA sýndi end-to-end vinnuferlisgetu inni í afmörkuðu EHR-sandkassakerfi. Það sýndi ekki raunverulega klíniska notkun, viðtæka greiningargetu eða fullkomlega sanngjarnan samanburð við lækna sem höfðu nákvæmlega sömu upplýsingar.

Hvað gerðu höfundarnir?

MIRA — **Medical Intelligence for Reasoning and Action** — er sjálfstæður agent byggður á líkönum OpenAI. Sá hluti sem talar og framkvæmir notar **GPT-4o**, en sérstakt skipulagsskref notar rökflug-sunarlíkanið **o1**. Kerfið situr inni í sérsníðuðum, stöðluðum ramma byggðum á **HL7 FHIR**, staðli sem raunveruleg sjúkrahús nota til að flytja sjúkraskrárgögn, með verkfærakistu sem inniheldur yfir **80.000 mögulegar klínískar aðgerðir**. Inni í sandkassanum getur MIRA sótt sögu sjúklings, pantað og túlkað blóðpróf, myndgreiningu og örverurannsóknir, byggt mismunagreiningu og gert meðferðaráætlun — skrifað lyf, skipulagt skurðaðgerð og innlögn.

Eitt atriði er rétt að nefna strax: sjálfvirku „dómararnir“ sem skoruðu svör MIRA voru einnig GPT-4o — sama líkanafjölskyldan og var til prófunar. Eftir að jafningjarnir benti á þetta bættu höfundarnir við yfirferð læknis með sérfræðileyfi.

Prófunargögnin komu úr **MIMIC-IV**, stórum opinberum og afpersónugreindum EHR-gagnagrunni. Úr um 300.000 sjúklingum sem fengu meðferð á Beth Israel Deaconess Medical Center 2008–2019 völdu höfundarnir **574 tilfelli** sem náðu yfir **átta markgreiningar** — kviðsjúkdóma og bráð innri læknisfræðileg vandamál, meðal annars botnlangabólgu, brisbólgu, lungnabólgu og þvagfærasýkingu. Í hverju tilfelli byrjaði MIRA með takmarkaða mynd og þurfti skref fyrir skref að velja næstu aðgerð. Ferlið líktist þannig raunverulegri uppvinnslu, en var keyrt á sögulegri sjúkraskrá þar sem lokaútkoman var þegar þekkt.

Árangurinn var síðan borinn saman við lækna á sömu tilfellum.

Hvað þýðir „sjálfstæður agent í EHR-sandkassa“ í raun?

Þrjú orð bera mikla merkingu hér.

Agent þýðir að kerfið er ekki bara spjallmenni sem svarar einni fyrirspurn. Það keyrir lykkju: skoðar stöðuna, velur aðgerð — panta þetta próf, spyrja um þetta einkenni — fær niðurstöðu og velur næstu aðgerð þar til það kemst að greiningu og áætlun. „Greindin“ kemur frá mállíkani; **agency**, sjálfstæð aðgerðargeta, kemur frá hugbúnaðarrammanum sem breytir texta líkansins í leyfðar EHR-aðgerðir og sendir niðurstöðurnar aftur inn.

EHR-sandkassi þýðir hermt, lokað eintak af rafrænni sjúkraskrá, ekki lifandi sjúkrahúskerfi. MIRA getur „pantað“ próf og fengið niðurstöðuna sem sjúklingurinn fékk í raun, því tilfellið er sögulegt og niðurstaðan er þegar skráð. Engin aðgerð snertir raunverulegan sjúkling eða raunverulega deild.

Sjálfstætt þýðir að MIRA klárar tilfellið frá upphafi til enda án manneskju í lykkjunni — **inni í sandkassanum**. Það þýðir **ekki** að kerfinu hafi verið leyft að stjórna raunverulegum sjúklingum án eftirlits. Aðeins fyrri fullyrðingin var prófuð.

Eitt atriði um sandkassann er líka auðvelt að misskilja. MIRA má **panta** hvaða próf sem það vill, en kerfið getur aðeins skilað niðurstöðu ef það próf var **raunverulega gert** á sjúklingnum.

Ef það biður um blóðpróf sem sjúklingurinn fékk, skilar kerfið raunverulegu gildunum. Ef það biður um myndrannsókn sem enginn pantaði, kemur „N/A — ekki hægt að framkvæma“, aldrei uppspunnin tala. Sama gildir um sjúkrasögu: ef upplýsingarnar eru ekki í skránni segir hermdi sjúklingurinn einfaldlega að hann viti það ekki. MIRA getur því ekki kallað fram fullkomið próf sem var aldrei gert; það verður að vinna með það sem raunveruleg uppvinnsla skráði. Þessi greinarmunur skiptir máli því orðið **sjálfstætt** er einmitt það sem helst er lesið of vítt.

Hvað fundu þeir?

Á viðmiðinu **sló MIRA lækna í greiningarnákvæmni** — en fyrirsögnin felur þrjár mismunandi tölur og mikilvægt er að halda þeim aðskildum.

Á öllum **574 tilfellum** valdi MIRA rétta greiningu **88,9%** tilvika, miðað við útskriftargreininguna sem var skráð í MIMIC-IV. Í beinum samanburði við menn unnu lækarnir ekki öll 574 tilfelli heldur sameiginlegt **311 tilfella undirmengi**, með sömu skrá og verkfærum og MIRA. Á þessum 311 tilfellum fékk MIRA **87,8%**.

Það var borið saman við tvo læknahópa. Fyrri var **reyndur hópur**: fjórir sérfræðilæknar með 7–11 ára reynslu, sem fengu **78,1%**. Seinni var **blandaður hópur**: aðallega yngri deildarlæknar ásamt tveimur sérfræðingum, nær raunverulegri mönnun bráðamóttöku, sem fékk **71,1%**. Sömu tilfelli og sömu verkfæri — og röðin varð MIRA fyrst, reyndir læknar næst og yngri hópurinn þriðji. Varfærna orðalag greinarinnar er að MIRA hafi verið „stöðugt sambærilegt við og oft yfir“ lækna yfir alla átta sjúkdóma.

Ákvarðanir eftir greiningu stóðust líka nokkuð vel. Höfundarnir segja þær að mestu samræmast leiðbeiningum, vera lyfjafræðilega öruggar og viðeigandi varðandi innlögn. Þeir greina engar alvarlegar lyfjavillur í fimm öryggisflokkum — milliverkunum, nýrnaskömmtnun, ofnæmi, QT-áhættu og óþióíðávísunum — og 467 af 468 lyfjaávísunum voru réttar. Samt taka þeir fram að kerfið hafi **ekki náð 100% áreiðanleika**.

Sé niðurstaðan tekin bókstaflega er hún sláandi: agent sem rekur allt tilfellið og kemur út með betri greiningartölu en læknar. En **lögun forskotsins** skiptir jafn miklu máli og forskotið sjálft.

Óháðir sérfræðingar sem lásu greinina bentu á hvar forskotið varð til. Á sérfræðingaspjaldi Science Media Centre benti Dr Wei Xing við University of Sheffield á að fyrirsagnartalan — AI yfir læknum — væri **að mestu knúin af sjúkdómum með skýrar prófniðurstöður**, svo sem botnlangabólgu og brisbólgu, þar sem afgerandi mynd eða rannsóknarstofugildi leysir málið. Fyrir lungnabólgu og þvagfærasýkingu — tvær mjög algengar ástæður komu á bráðamóttöku — voru **bæði** MIRA og lækarnir verri og bilið milli þeirra minna.

Það var líka ósamhverfa í upplýsinganotkuninni. MIRA nýtti um **51%** af tiltækum rannsóknarstofugildum í venjubundinni umönnun, en sérfræðilæknarnir um **28%** — miðgildi var um sjö fleiri blóðgildi á tilfelli. Höfundarnir taka fram að þetta sé samt undir eigin venjubundnu rannsóknarnotkun gagnagrunnsins og að MIRA hafi ekki kerfisbundið aukið dýra þversniðsmyndgreiningu.

Samt geta fleiri upplýsingar einar og sér aukið greiningarnákvæmni. Því er þetta ekki fullkomlega jafnt próf milli jafn-upplýstra aðila. Læknir sem mætti panta öll ódýr blóðpróf án þess að huga að kostnaði, óþægindum eða töf myndi líka líta betur út á pappír.

Af hverju „sló lækna“ þýðir ekki „betri læknir“

Það er auðvelt að lesa of mikið í sigur á viðmiði. Þrjú atriði standa milli „MIRA fékk hærra skor“ og „MIRA er betra“.

Í fyrsta lagi: **hvað var skorað gegn?** Professor Julie Jacko við University of Edinburgh benti á að nokkrir mikilvægar niðurstöður MIRA séu skilgreindar miðað við það sem var skráð í undirliggjandi gagnasafni. Kerfið fær því umbun fyrir að **endurtaka skráða klíniska hegðun**, ekki endilega fyrir að sýna bestu mögulegu umönnun. Sögulega sjúkraskráin verður að svarlykli. Þetta er eðlileg leið til að byggja viðmið, en mælir samræmi við það sem var gert, ekki algildan klínískan sannleik. Þetta á mest við um mælikvarða á meðferðarsamræmi; fyrirsagnarnákvæmni greiningar er borin saman við útskriftargreiningu, sem er nær raunverulegri klínískri niðurstöðu.

Í öðru lagi er **upplýsingaósamhverfan**: MIRA notaði miklu fleiri blóðpróf, um 51% af tiltækum analytes gegn 28%. Hluti forskotsins kann því að vera **keyptur með meiri upplýsingum**, ekki aðeins betri rökhugsun.

Í þriðja lagi: **hvar kerfið keyrði**. Þetta var sandkassi á afturvirkum tilfellum, yfir átta fyrirfram valdar greiningar og með texta einan. Eins og Dr Dominic Oliver við University of Oxford benti á byggir raunverulegt klínískt mat ekki aðeins á því sem sjúklingur segir heldur **hvernig** hann segir það, ásamt líkamsskoðun, hegðun og líkamstjáningu. Textaagent sem les fullbúna sjúkraskrá sér ekkert af þessu. Og sjúklingur með vandamál utan átta valinna greininga er einfaldlega utan þess sem þessi rannsókn getur sagt til um.

Jafningjarýnar nefndu líka hljóðlátari áhyggju: **gagnamengun**. MIMIC-IV er opinbert og mikið skrifað um. Mállíkan þjálfað á opna internetinu gæti hafa séð greinar, tilfellaýsingar eða jafnvel gögnin sjálf. Dr Midhun Parakkal Unni við University of Sheffield benti beint á þetta: ef svör hafa verið í þjálfunargögnum er hluti árangursins minni, ekki klínísk rökhugsun. Aðeins óháð endurtekning getur skilið þar á milli. Höfundarnir sjálfir afgreiða ekki þetta vandamál; þeir skrifa að niðurstöðurnar megi túlka varlega sem möguleg **efri mörk** og geti ofmetið alhæfingu til annarra opinberra tilfella.

Ekkert af þessu gerir MIRA minna áhugavert. Það gerir rétta túlkun nákvæmari: á afturvirku viðmiði sló agent sem gat unnið yfir alla sjúkraskrána lækna í greiningum með skýr próf — að hluta með fleiri prófum og að hluta með samræmi við skráða klíniska ferlið. Það er raunveruleg sýning á getu, ekki úrskurður um að vélar greini nú almennt betur en læknar.

Hvað sýnir þetta ekki?

- Það sýnir **ekki** að MIRA virki á raunverulegum sjúklingum. Öll tilfelli voru afturvirk og hermd; enginn sjúklingur var stjórnað af kerfinu.
- Það sýnir **ekki** að kerfið virki utan átta greininga. Ógreind, óskýr og fjölbreytt læknisfræði meirihlutans var ekki prófuð.
- Það sýnir **ekki** að kerfið sé öruggt í notkun. Höfundarnir segja sjálfir að alhæfing, öryggi og stjórnun þurfi framskyggjar raunheimsrannsóknir.
- Það sýnir **ekki** fullkomlega sanngjarnan samanburð við lækna. MIRA notaði mun fleiri blóðpróf, og sum meðferðarútkomumælikvarðar voru að hluta skoraðir gegn sögulegu sjúkraskránni fremur en óháðu „bestu meðferð“ viðmiði.
- Það sýnir **ekki** að niðurstaðan sé laus við minni úr þjálfunargögnum. Opinber MIMIC-IV gögn gætu skarast við þjálfunarlíkanið.
- Það þýðir **ekki** „**AI kemur í stað lækna**“. Kerfið er ákvörðunarstuðningur sem getur líka framkvæmt aðgerðir í sjúkraskrá; raunveruleg notkun myndi krefjast samstarfs við lækna sem halda ábyrgð og bæta við því sem textaskrá getur ekki fangað.

Hversu sterk eru gögnin?

Skipta þarf fullyrðingunni í tvennt vegna þess að sönnunargögnin eru mjög ólík.

Sem sýning á getu — að mállíkanaagent geti unnið heilt klínískt tilfelli í EHR-sandkassa, tengt sögu, próf, greiningu og fyrirmæli frá upphafi til enda — er verkið raunverulega nýtt og nokkuð sannfærandi. Þetta er nýi hlutinn og sá sem vert er að fylgjast með.

Sem yfirburðafulllyrðing — að MIRA sé *betra en lækna* — eru gögnin afmörkuð. Sú niðurstaða gildir á tilteknu afturvirku viðmiði, yfir átta greiningar, með upplýsingamun, sögulegan svarlykil fyrir hluta mælinganna og raunverulegan möguleika á gagnamengun. Það nægir til að segja „agentinn stóð sig mjög vel á þessu viðmiði“. Það nægir ekki til að segja „agentinn er betri greiningarlæknir en læknir“ — og höfundarnir sjálfir setja ekki fram seinni fullyrðinguna.

Gagnlegasta afstaðan er hvorki „AI slær lækna“ né „þetta er bara leikfang“. Hún er: **ný tegund læknisfræðilegrar gervigreindar, sem framkvæmir aðgerðir yfir sjúkraskrá í stað þess að svara einni spurningu, stóð sig vel á erfiðu afturvirku viðmiði og þarf nú að sanna sig þar sem hún hefur ekki enn verið prófuð: á raunverulegum, óflokkuðum sjúklingum, framskyggnt og undir skýrri stjórnun.**

Af hverju skiptir þetta máli?

Undanfarin ár hefur áhugaverða spurningin í læknisfræðilegri gervigreind hljóðlega breyst. Hún var áður „getur líkan fengið rétta greiningu?“ — og líkön fóru að svara já á sífellt hreinni prófasöfnum.

MIRA flytur spurninguna yfir í erfiðara form: **getur líkan unnið verkið — safnað, pantað, túlkað, ákveðið og framkvæmt — yfir allt vinnuferlið?** Það er gagnlegri spurning og heiðarlegri, því bæði erfiðleikarnir og áhættan búa í aðgerðunum.

Þess vegna skiptir umgjörðin máli. Fyrirsagnarviðbragðið — *AI outperforms doctors* — bendir á minnsta nýja og veikast studda hlutann. Raunverulega nýjungin er minni en afdrifaríkari: agent sem getur **fært sig í gegnum alla sjúkraskrána**. Ef sú geta stenst raunheiminn breytir hún vinnuferlinu löngu áður en hún breytir því hver ber ábyrgð. Læknirinn hverfur ekki; pantanir, túlkun, eftirfylgni niðurstaðna og önnur vinnuferli eru fyrsti staðurinn þar sem slíkt tæki lendir.

Og það setur sönnunarbyrðina í rétta átt. Sigur á afturvirku viðmiði er **ástæða til að gera framskyggna rannsókn**, ekki staðgengill hennar. Höfundarnir segja það sjálfir. Heiðarlega túlkunin á MIRA er boð um næstu rannsókn — mælda á sjúklingum, ekki aðeins stigatöflum.

Í stuttu máli

MIRA er sjálfstæður gervigreindaragent sem starfar inni í EHR-sandkassa: hann getur tekið sjúkrasögu, pantað og túlkað rannsóknir, myndgreiningu og örverurannsóknir, komist að greiningu og skrifað meðferðaráætlun. Á **574 afturvirkum MIMIC-IV tilfellum** yfir átta fyrirfram valdar greiningar fékk hann hærri greiningarnákvæmni en læknar — 88,9% í heild; 87,8% á 311 beinum samanburðartilfellum gegn 78,1% hjá sérfræðilæknum — og ákvarðanir voru að mestu samræmdar leiðbeiningum með engum alvarlegum lyfjavillum. En þetta var hermun á sögulegum skrá, aðeins texti, og stór hluti forskotsins kom í sjúkdómum með skýr próf. MIRA notaði miklu fleiri blóðgildi en læknarnir, sum meðferðarviðmið voru skoruð gegn því sem upprunalega sjúkraskráin hafði skráð og opinber MIMIC-IV gagnagrunnurinn skapar raunverulega hættu á þjálfunargagnamengun — sem höfundarnir sjálfir segja geta gert tölurnar að mögulegum efri mörkum. Raunverulega framfaraskrefið er agent sem **framkvæmir vinnuferli**, ekki sönnun þess að AI greini almennt betur en læknar eða sé tilbúið á deild.

Raunhæf yfirferð

Hvað sýnir greinin? Mállíkanaagentinn MIRA getur rekið heilt klínískt tilfelli frá upphafi til enda inni í EHR-sandkassa — saga, próf, greining, fyrirmæli — og fékk á afturvirku viðmiði með 574 tilfellum yfir átta greiningar hærri greiningarnákvæmni en læknar, með 88,9% heildarnákvæmni og 87,8% gegn 78,1% í beinum samanburði við sérfræðilækna, án alvarlegra lyfjavilla í skilgreindu öryggismati.

Hvað er trúverðugt en ósannað? Að þessi geta skili sér í ávinningi fyrir raunverulega, óflokkaða sjúklinga; að nákvæmnisforskiði sé betri klínísk rökhugsun fremur en fleiri pöntuð próf, samræmi við sögulega skrá eða minni úr opinberum gögnum.

Hvað sýnir greinin ekki? Að MIRA virki utan átta greininga; að það sé öruggt í klínískri notkun; að það

slái lækna í fullkomlega jafn-upplýstum samanburði; að það komi í stað lækna; eða að niðurstaðan sé laus við gagnamengun.

Helstu takmarkanir: Afturvirk hermun, aðeins texti, átta fyrirfram valdar greiningar; upplýsinga-munur vegna mun fleiri blóðprófa — um 51% af tiltækum analytes gegn 28% — en ekki kerfisbundið meira myndgreiningarmagn; meðferðarmælikvarðar að hluta gegn sögulegri sjúkraskrá; og opinbert MIMIC-IV viðmið sem líkanið gæti hafa séð í þjálfun.

Hversu mikið traust ætti almennur lesandi að leggja á niðurstöðuna? Mikið traust á að MIRA sé raunveruleg og ný sýning á getu — agent sem **framkvæmir** yfir sjúkraskrá, ekki bara spjallmenni. Lítið traust á stærri fullyrðinguna að kerfið sé almennt **betri greiningarlæknir en læknir**. Sú niðurstaða er bundin við eitt afturvirkt viðmið og ruglast af prófanotkun, sögulegri skorun og mögulegri gagnamengun. Réttu niðurstaðan er sú sem höfundarnir sjálfir setja: þetta þarf framskyggna raunheimsrannsókn áður en það segir nokkuð ákveðið um sjúklingaþjónustu.

Heimildir

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->

News-Medical (2026) — illustrates the 'outperforms doctors' framing

<https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-simulated-EH.aspx>