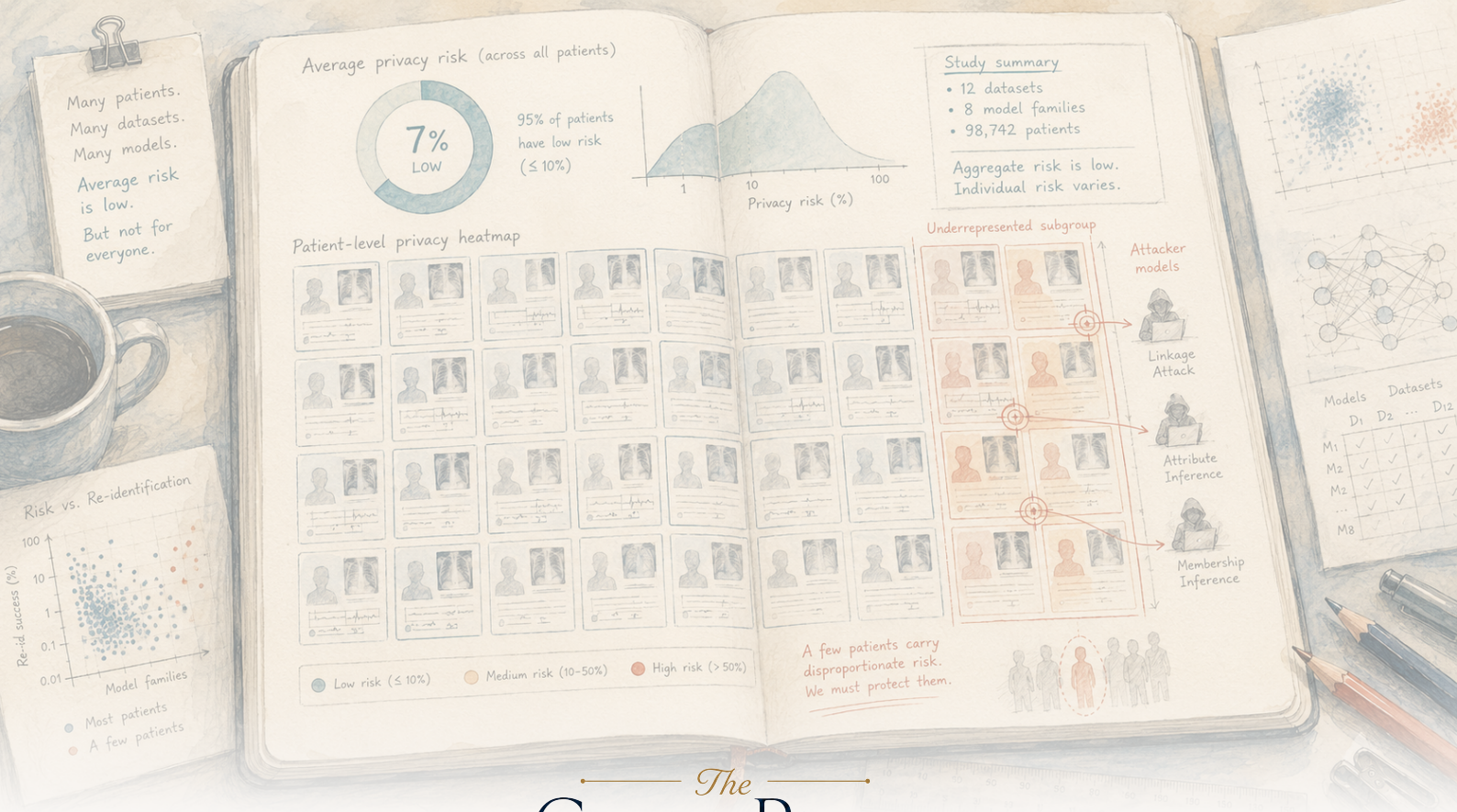




The CLEAN PAPER

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

Medical AI na iya zama private on average amma har yanzu ya expose particular patients — especially underrepresented

Medical-AI model da ake kira “privacy-preserving” yawanci yana dogara da average number guda. Wannan study yana cewa wrong test ne. Ta studying membership inference attacks — wadanda ke reveal whether specific person’s record yana training data, don haka zai iya betray disease status — authors sun measure risk per patient rather than aggregate, across seven medical datasets (imaging, ECG, electronic records) da many models each. Pattern: models da look safe on average na iya still let attacker identify specific individuals almost perfectly (attack AUC ≥ 0.95); exposed systematically from underrepresented groups; kuma gets worse as models grow. Authors ba su ce abandon medical AI ba — suna ce measure privacy per patient, control model access, use differential privacy. Core: “private on average” ba privacy guarantee ba ne.

Written by Lucio Vaglio · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Disparate privacy risks from medical AI*, Moritz Knolle, Georg Kaissis, Daniel Rückert and colleagues (Technical University of Munich; Imperial College London; Hasso Plattner Institute), Nature (2026) · DOI: 10.1038/s41586-026-10688-0

Alkawarin sirri ana yi wa mutum ne, ba wa matsakaici ba

Idan aka ce wani samfurin AI na lafiya yana “kare sirri,” ana yawan jingina wannan da’awa ga adadi guda: idan aka duba duk marasa lafiyar da aka yi amfani da bayanansu wajen horar da samfurin, matsakaicin yiwuwar gano ko bayanan wani mutum sun shiga horon yana da kasa. Wannan yana iya sa a ji kwanciyar hankali. Amma wannan takarda tana nuna wani abu mai sauƙi kuma marar daƙi: ba wannan ne adadin da ya kamata mu kalla ba.

Sirri ba matsakaici ba ne. Alkawari ne da ake yi wa kowane mutum — cewa kasancewar bayanansa cikin rukunin horo ba zai dawo ya tona masa asiri ba. Matsakaici kuma zai iya cika wannan alƙawarin ga kusan kowa, amma ya karya shi gaba ɗaya ga ’yan kaɗan. Masu binciken sun auna haɗarin sirri ga kowane mara lafiya ɗaya bayan ɗaya, da cikakken matakin da ba a taɓa amfani da shi ba, kuma sun ga haka: samfurin da ya yi kama da mai aminci idan an haɗa kowa wuri guda zai iya bayyana ko bayanan wasu takamaiman mutane sun shiga horo kusan ba tare da kuskure ba — kuma waɗanda wannan ya fi shafa su ne, fiye da kima, mutanen da tuni ba su da isasshen wakilci a bayanansu.

Abin da marubutan suka yi

Kungiyar — farkashin jagorancin Moritz Knolle, tare da Daniel Rückert, dukansu daga Technical University of Munich, da Georg Kaissis daga Hasso Plattner Institute, da abokan aiki a Imperial College London — ta ɗauki wata sananniyar barazanar sirri da ake kira **membership inference attack**, wato ƙoƙarin gano ko bayanan mutum sun kasance cikin rukunin horon samfurin, ta sauya tambayar da ake yi. Maimakon “a matsakaici, sau nawa wannan harin yake yin nasara a cikin ma’ajin bayanai?”, suka tambaya: “ga wannan mara lafiya takamaimai, yaya girman haɗarinsa?”

Sun yi wannan nazari na mutum-da-mutum a kan **ma’ajin bayanansu lafiya bakwai** da aka riga aka kafa, masu nau’ikan bayanai daban-daban — hotunan likita, electrocardiograms da kundin lafiya na lantarki — kuma suka gwada samfurori 200 a kowane. Ga kowane mara lafiya a kowane samfurin, sun kiyasta yadda mai hari zai iya tantance ko kundin wannan mutumin ya shiga bayanansu horo, sannan suka raba sakamakon bisa rukuni: ko mutum yana da cutar da ake dubawa, ƙabilar da ya bayyana da kansa, inshora, jinsi da tsarin ɗaukar hoto.

Menene membership inference attack a zahiri?

Abin da ke zubewa a nan ya fi dabara fiye da “samfurin ya fitar da kundinka.” Abin da ake ƙoƙarin gano wa shi ne *membership* — wato kawai ko bayananka sun kasance cikin rukunin horo.

Ka fara da abin da irin waɗannan samfura suke yi a kullum. Ka ba shi hoto — misali X-ray na kirji — sai ya dawo da yiwuwar abubuwa daban-daban: 78% yiwuwar pneumonia, tare da kimomi ga cardiomegaly, oedema ko consolidation. Harin ba ya buƙatar wani abu dabam; wannan amsar bincike ta yau da kullum kaɗai yake amfani da ita.

Raunin da ake amfani da shi shi ne cewa samfurin kan nuna **ƙarin tabbaci kaɗan** kan misalan da aka horar da shi da su fiye da waɗanda bai taɓa gani ba. Ka yi tunanin dalibin da ya ga tambayoyin jarraɓawa a boye tun kafin ranar jarraɓawa: a tambayoyin da ya riga ya yi atisaye da su, amsa kan fito da sauri kuma da ƙarin tabbaci. Gano, daga wannan alamar, ko wani takamaiman kundi yana cikin misalan da samfurin

ya koya shi ne abin da ake nufi da membership inference.

Ga yadda gwajin yake aiki a takaice. Wani yana son sanin ko an yi amfani da hoton wani mutum wajen horar da samfurin. Ba ya tambayar samfurin ya ba shi kundin mutumin — tuni yana da hoton da yake son gwadawa, ko wani kwafi mai kusanci. Sai ya aika hoton kamar tambayar bincike ta yau da kullum, ya lura da yadda samfurin ya nuna tabbaci, sannan ya kwatanta wannan tabbacin da yadda samfurin kan yi idan ya taɓa ganin misali a horo ko bai taɓa gani ba. Don koyon wannan bambanci, mai harin zai iya horar da wani samfurin kwatance nasa a kwamfuta ta yau da kullum. Idan samfurin da ake gwadawa ya nuna tabbaci na ban mamaki kan wannan hoton — kamar yana gane shi — wannan na iya zama shaida mai karfi cewa hoton yana cikin bayanan horo.

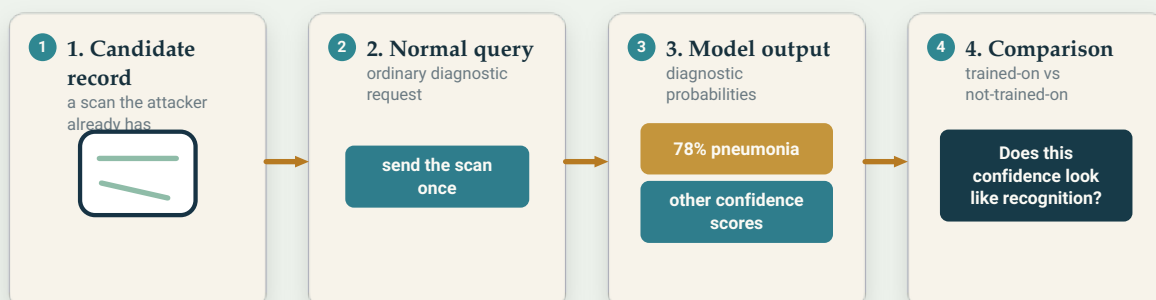
Abin da ke tayar da hankali shi ne cewa bukatar ba ta da wani kamannin hari: tambayar bincike ce irin wadda likita zai iya yi, yayin da alamar sirri take boye cikin hasashen da samfurin ya bayar. Shi ya sa hadarin yake na gaske, ko da yake har yanzu an nuna shi ne a farkashin sharuɗɗan gwaji, ba wai an kama irin wannan hari yana faruwa a asibiti ba.

Me ya sa kasancewa cikin rukunin horo yake da muhimmanci idan kundin kansa bai fito ba? Domin **kasancewa cikin rukunin ma bayani ne**. Idan an horar da wani samfurin ne da bayanan mutanen da suka sami wani maganin rigakafin ciwon daji na musamman, tabbatar da cewa kundinka yana cikin samfurin na iya nuna cewa mai yiwuwa ka yi wannan cutar — irin bayanin da kamfanin inshora ko mai daukar aiki bai kamata ya iya hasashe ba. Abin da ke cikin kundin ya zauna a rufe; amma gaskiyar cewa kana cikin rukunin ce ta fito.

Marubutan sun fayyace sharuɗɗan harin — samun damar ganin hasashen samfurin na yau da kullum, samun kundin da ake son gwadawa, da samun samfurin kwatance na mai hari — ba domin su koyar da yadda ake kai hari ba, sai domin a iya auna barazanar da tunani mai kyau maimakon a yi mata watsi da hannu.

A membership attack can hide inside a normal prediction

The attacker does not ask for the record. They already hold a candidate and query the model once.



The privacy signal is hidden in an ordinary prediction request.

Mechanism only: no pseudocode, no attack threshold, no recipe.

Harin ba ya bukatar wata API ta musamman ta sirri. Tambayar bincike ta yau da kullum za ta iya fitar da alamar

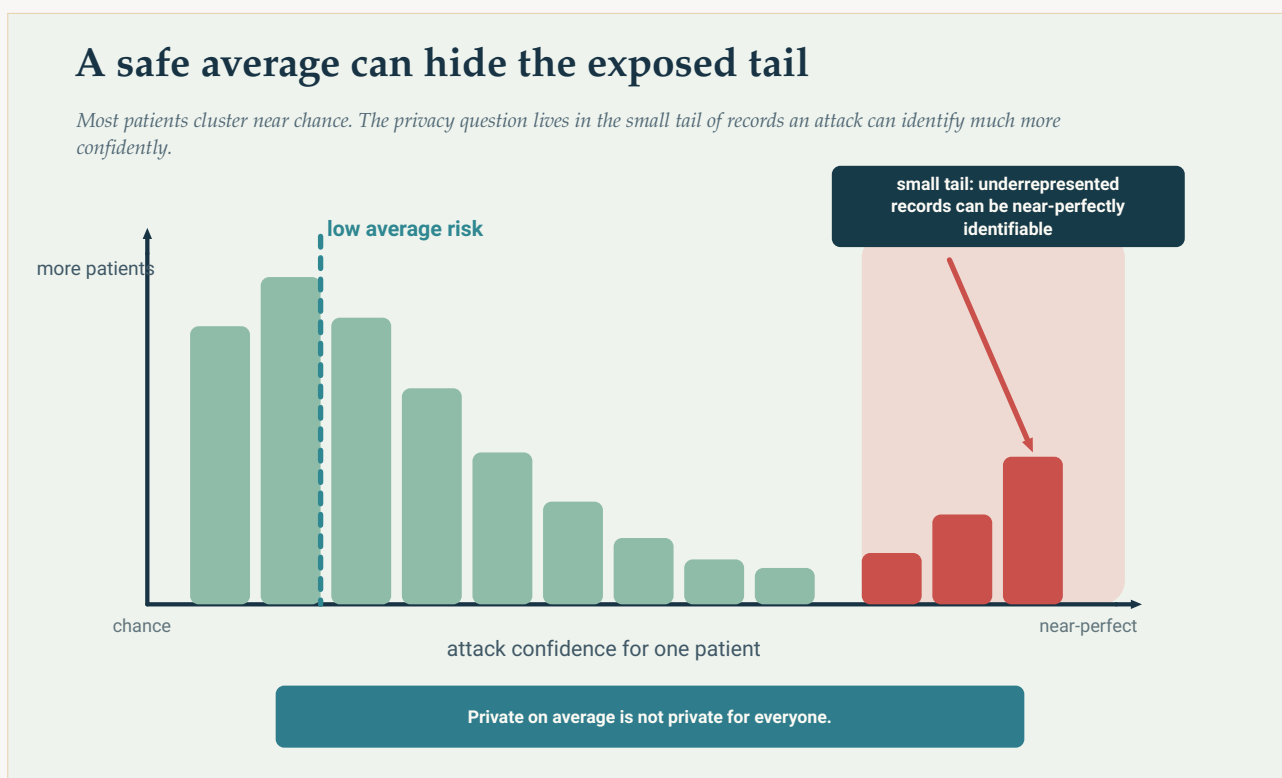
cewa wani kundi yana cikin bayanan horo idan samfuran ya nuna tabbaci na musamman kan abin da ya taba gani.

Original Aurora diagram — The Clean Paper CC BY 4.0

Abin da suka gano

Sakamakon uku ne, kuma kowannensu ya fi na baya kaifi.

Matsakaici yana boye waɗanda suka fi fuskantar haɗari. Idan aka auna kowa a haɗe — yadda aka saba yi — yawancin samfuran suna kama da masu kare sirri sosai: ga mafi yawan marasa lafiya, harin bai fi jefa tsabar kuɗi ba. Amma idan aka kalli mutum ɗaya bayan ɗaya, hoton ya canza. Knolle ya bayyana a sanarwar binciken cewa a baya ana auna “matsakaicin haɗari ga duk marasa lafiya” ne kawai; su kuwa sun duba haɗarin kowane mara lafiya a karon farko, kuma sakamakon ya bambanta sosai. Ga wasu mutane, harin yana yin nasara **kusan cikakke** — AUC na harin **0.95 ko fiye**; 0.5 daidai yake da sa’a, 1.0 kuma babu kuskure — duk da cewa matsakaicin duk ma’ajin ya yi kama da rashin fiye da sa’a.



Matsakaicin haɗari zai iya zama kusa da sa’a yayin da wani karamin rukuni na kundin mutane ya kasance da haɗari mai yawa. Babban sakon takardar shi ne cewa dole a duba sirri ga kowane mara lafiya, ba a matsakaici kaɗai ba.

Original Aurora diagram — The Clean Paper CC BY 4.0

Haɗarin ba ya rarrabuwa daidai — kuma ya fi sauka kan waɗanda ba su da isasshen wakilci. Mutanen da suka fi saurin shiga cikin rukunin da harin zai iya gane su kusan cikakke sun kasance, cikin tsari, daga rukunin da

ba su da yawa a bayanai: Kananan kabilu, cututtuka masu wuya, ko siffon hoto da ba a yawan gani. A ma'ajin kundin lafiya na lantarki, marasa lafiya bakaken fata sun bayyana **31% fiye da yadda ake tsammani** a cikin rukunin da ya fi fuskantar hadari. A ma'ajin mammography kuma, hotunan da aka yi wa alamar zargin cutar daji sun yi yawa a wannan rukunin da **+1,179%**. Wadannan misalai biyu ne kawai, ba dukan sakamakon ba; takardar ta ga irin wannan karkata a rukuni da dama. Kuma lambobin suna nuna *karin wakilci a cikin karamin rukunin da ya fi hadari*, ba kaso na dukan bakaken fata ko dukan masu hoton mammography da ke zargin cutar daji da aka fallasa ba. Idan samfurin yana da misalai kadan masu kama da wani mara lafiya, kundinsa zai fi fita dabam — kuma irin wannan bambancin ne harin membership inference yake gano wa. Saboda haka gazawar sirrin ba bazuwar abu ba ce; tana taruwa ne kan wadanda tuni suke a gefen bayanai.

Manyan samfura suna kara matsalar. Yawan marasa lafiya da suka fuskanci harin da ya kusan yin nasara cikakke ya karu sosai yayin da karfin samfurin ya karu. A wani ma'ajin dermatology, kaso ya tashi daga kusan sifili a karamin samfurin zuwa kusan **mutum ɗaya cikin goma** a mafi girma. Yayin da samfuran AI na lafiya suke kara girma da kwarewa — alkiblar da fannin gaba ɗaya yake bi — marubutan suna gargadin cewa wannan takamaiman hadarin yana karuwa, ba raguwa ba.

Rückert ya takaita shi kai tsaye: “Wannan ba hadarin da za a iya yarda da shi ba ne. Bayanai na lafiya suna da matuƙar sirri.”

Me ya sa “aminci a matsakaici” ba shi ne gwajin da ya dace ba

Yana da sauƙi a ga karamin matsakaicin hadari a matsayin shaidar cewa komai lafiya. Babban darasin takardar shi ne cewa a nan yin matsakaici ba kawai rashin cikakken bayani ba ne — tambayar da yake amsawa ma ba ita ce tambayar sirri ba.

Alkawarin sirri yana da ma'ana ne kawai idan ya tsaya ga mutumin da ya fi fuskantar hadari, ba mutumin da yake tsakiyar rarraba ba. Idan mutane 999 cikin 1,000 ba za a iya gane su ba amma mutum guda za a iya gane shi kusan babu kuskure, ba daidai ba ne a kira tsarin “99.9% mai sirri” ta hanyar da za ta taimaka wa wannan mutum ɗaya. Kuma idan wannan mutumin ana iya hasashen cewa mai cuta mai wuya ne ko daga karamar kabila, matsalar ma ba rashin cikar ma'auni kadai ba ce; ma'aunin yana boye bambancin da yake cutar da wasu rukuni. Yana auna aminci ga mafiya yawa sannan ya kira shi aminci ga kowa.

Wannan shi ne sauyin tambayar da takardar take tilastawa: daga *yaya sirrin samfurin yake a matsakaici?* zuwa *wane mara lafiya ne ya fi fallasa, kuma wane irin mutum ne shi?* Tambayoyi ne biyu dabam, kuma ta biyu ce ainihin tambayar sirri.

Abin da wannan bai tabbatar ba — kuma takardar ba ta yi da'awarsa ba

- **Ba ta ce a daina amfani da AI na lafiya ba.** Abin da marubutan ke nema shi ne rage hadari: a auna shi a gyara, ba a daina gina fasahar ba.
- **Ba ta nufin kowane samfurin lafiya yana zubar da bayanai ba,** ko cewa an riga an kai hari ga wani samfurin da ake amfani da shi. Ta nuna cewa *hadarin yana wanzuwa kuma ba ya rarrabuwa daidai*, kuma ma'aunan matsakaici na yau da kullum ba sa ganinsa.
- **Ba ta nufin kundinka ya riga ya fallasa ba.** Harin yana buƙatar wasu sharudda — damar amfani da samfurin, samun kundin da ake son gwadawa, da kayan aikin mai hari — ba wani abu da kowa zai yi ba tare da shirye-shirye ba.
- **Ba ta nuna cewa abin da ke cikin kundin mara lafiya ne yake fitowa ba.** Abin da zai iya zubewa shi ne kasancewar kundin cikin bayanin horo. Wannan ma hadari ne, amma dalilinsa ya bambanta da fitar da dukan kundin.
- **Ba ta rage bambancin zuwa lamba guda ba.** Sakamakon da ya fi ƙarfi kuma aka maimaita shi shi ne tsarin da aka gani: ma'aunin matsakaici yana rage girman hadarin mutum ɗaya, kuma waɗanda ba su da isasshen wakilci su ne suka fi ɗaukar ragowar hadarin — a ma'ajin bayanai da samfura da yawa.

Yaya ƙarfin shaidar yake?

Wannan sakamako ne na gwaji da kuma hanyar aunawa, kuma yana da ƙarfi. Tsarin — cewa ma'aunin sirri na haɗe yana rage haɗarin da wasu mutane ke fuskanta, cewa ragowar haɗarin ya fi taruwa kan rukunin da ba su da isasshen wakilci, kuma yana ƙaruwa da ƙarfin samfurin — ya bayyana a **ma'ajin bayanai bakwai masu nau'ikan bayanai daban-daban** da kuma yawan samfura a kowannensu. Wannan faɗin gwaji ne yake sa sakamakon ya fi kama da matsalar aunawa ta gaske maimakon wani abu na ma'aji guda.

Akwai iyakoki biyu masu muhimmanci. Na farko, wannan nuna *yiwuwar hadari* ne: marubutan ne suka gudanar da hare-haren da suka tsara domin su auna abin da kwararren mai hari *zai iya* yi a ƙarƙashin sharuɗɗan da aka bayyana. Ba su auna sau nawa irin wannan hari yake faruwa a duniya ta gaske ba. Na biyu, lambobin da suka fi ɗaukar hankali — nasarar hari kusan cikakke ga wasu mutane — an zaɓe su ne daga **mutanen da suka fi fuskantar haɗari**, da gangan. Wannan shi ne ainihin maƙasudin binciken. Ya kamata a karanta su a matsayin “wutsiyar haɗari ta fi yadda matsakaici yake nuna wa nauyi,” ba “yawancin marasa lafiya sun fallasa” ba.

Matsayin da ya dace ba firgici ba ne kuma ba watsi ba. Wannan bincike ne mai hankali da ya yi amfani da ma'ajin bayanai da yawa, yana nuna cewa hanyar da ake yawan amfani da ita wajen tabbatar da sirrin AI na lafiya ba ta ganin mafi munin lokutanta — kuma waɗannan lokutan suna fi sauka kan mutanen da tuni suke da ƙarancin kariya.

Me ya sa yake da muhimmanci

Abubuwa biyu ne suka sa wannan ya wuce ƙaramin bayanin fasaha.

Na farko, yana sauya abin da ya kamata a bari kalmar “mai kare sirri” ta nufa. Idan aka saki samfurin tare da karamin matsakaicin haɗari, wannan adadi zai iya zama daidai amma har yanzu ya boye wani rukuni na marasa lafiya da membership inference attack zai iya gane su kusan babu kuskure. Bukatar takardar a aikace a bayyane take: a auna haɗarin sirri **ga kowane mara lafiya** kafin a saki samfurin, a kula da wanda zai iya samun damar amfani da samfurin, kuma a yi amfani da hanyoyi kamar **differential privacy** — kara karamin hayaniya da aka tsara a hankali yayin horo domin rage damar membership inference. Wannan yana da kuɗinsa ga ingancin samfurin, don haka akwai musayar sirri da amfani; ba kariya ce ta kyauta ba. “Mun duba matsakaici” bai kamata ya ci gaba da nufin “mun duba sirri” ba.

Na biyu, binciken yana haɗa sirri da adalci. Rukuni iri ɗaya da ba su da isasshen wakilci a bayanan lafiya — saboda haka tuni AI na lafiya na iya yi musu aiki kasa da yadda ya kamata — su ne kuma waɗanda wannan AI ta fi kasa kare sirrinsu. Fannin da yake fokarin gyara rashin daidaiton farko ba zai iya ɗaukar na biyu a matsayin matsalar wani sashe dabam ba. Mutane iri ɗaya ne abin ke shafa.

Labarin da yake sa a ji kwanciyar hankali — *mun auna sirri, matsakaici ya yi kasa, komai lafiya* — shi ne abin da wannan takarda take rushewa. Ba domin ta tsoratar da mutane daga fasahar ba, sai domin ta matsar da ma’aunin zuwa inda ya kamata: alkawarin da ba za ka iya cewa ka cika ba sai ka duba mutumin da ya fi yiwuwa ya ji rauni.

Takaitaccen bayani

Membership inference attack yana fokarin gano ko bayanan wani takamaiman mutum sun kasance cikin bayanan da aka horar da samfurin AI da su. Wannan kansa na iya tona sirri — misali idan kasancewa cikin rukunin horon ya nuna cewa mutum ya yi wata cuta — ko da kundin cikakke bai taɓa fitowa ba. A baya ana yawan auna nasarar irin wannan hari ne *a matsakaici* a duk ma’ajin bayanai. Wannan binciken ya auna shi *ga kowane mara lafiya*, a ma’ajin lafiya bakwai masu hotuna, ECG da kundin lantarki, tare da samfura da yawa a kowane. Ya gano cewa ma’aunin matsakaici yana rage haɗarin da wasu mutane ke fuskanta: ana iya gane wasu kusan cikakke (AUC na hari ≥ 0.95) duk da cewa matsakaicin ma’ajin ya yi kama da aminci; waɗanda suka fi fuskantar haɗarin su ne rukunin da ba su da isasshen wakilci, kamar kananan kabilu, cututtuka masu wuya ko hotuna na musamman; kuma matsalar tana karuwa yayin da samfurin yake kara girma. Marubutan ba su ce a daina AI na lafiya ba. Suna cewa a auna sirri ga mutum ɗaya, a sarrafa damar shiga samfurin, kuma a yi amfani da differential privacy da sauran hanyoyin rage haɗari. Babban darasi: “mai sirri a matsakaici” ba alkawarin sirri ba ne, kuma gazawar wannan matsakaicin ta fi sauka kan mutanen da tuni ba su da kariya sosai.

Bincike ba tare da karin gishiri ba

Abin da takardar ta nuna: A ma’ajin bayanan lafiya bakwai da samfura masu yawa a kowane, haɗarin membership inference ga wasu marasa lafiya ya fi abin da ma’aunin matsakaici yake nuna wa nisa — har zuwa nasarar hari kusan cikakke ($AUC \geq 0.95$). Ragowar haɗarin ya fi taruwa kan rukunin da ba su da isasshen wakilci, kuma yana karuwa da karfin samfurin.

Abin da yake yiwuwa amma ba a tabbatar ba: Cewa masu hari a duniya ta gaske sun riga sun fara amfani da wannan rauni a kan samfuran lafiya da ake amfani da su. Binciken ya nuna *ikon yin harin da yadda haɗarin*

yake rarrabuwa a karkashin sharuɗɗan da aka fayyace; bai auna yawan hare-haren gaske ba.

Abin da bai nuna ba: Ba ya nuna cewa ya kamata a daina AI na lafiya, cewa kowane samfurin yana zubar da bayanai, cewa an kutsa wani takamaiman samfurin da ake amfani da shi, ko cewa abin da ke cikin kundin mara lafiya ne yake fitowa. Haka kuma ba ya rage dukan bambancin rukuni zuwa lamba guda; abin da aka maimaita shi ne tsarin bambancin.

Babban iyaka: An auna mafi munin haɗari ta hare-haren da marubutan suka gudanar, ba abubuwan da aka kama suna faruwa a zahiri ba; lambobin da suka fi ban mamaki suna nufin mutanen da suka fi fallasa da aka zaɓa da gangan; kuma bambancin rukuni ya bayyana a matsayin tsari mai maimaituwa a ma'ajin bayanai daban-daban, ba ma'auni guda da zai takaita kowa ba.

Yaya yawan amincewa ya kamata mai karatu ya yi? Babban amincewa cewa ma'aunin sirri na matsakaici yana rage haɗarin wasu mutane, cewa wannan ragowar haɗarin ya fi sauka kan waɗanda ba su da isasshen wakilci, kuma matsalar tana karuwa da girman samfurin — an nuna hakan a ma'ajin bayanai da samfura da yawa. Matsakaicin amincewa game da yawan irin wannan hari a rayuwa ta gaske, domin binciken bai auna hakan ba. Karatun da ya fi aminci shi ne: ba “AI na lafiya tana zubar da bayananka” ba, kuma ba “an warware sirri” ba, sai “gwajin sirri na yau da kullum ba ya ganin mafi munin lokutansa, kuma waɗanda suka fi rauni su ne suke daukar nauyin — don haka dole a sauya gwajin.”

Majiyoyi

Knolle et al., Disparate privacy risks from medical AI, Nature (2026)

<https://doi.org/10.1038/s41586-026-10688-0>

Technical University of Munich — press release, 'Study reveals privacy risks in medical AI' (2026)

<https://www.tum.de/en/news-and-events/all-news/press-releases/details/study-reveals-privacy-risks-in>

Medical Xpress (2026) — coverage of the study

<https://medicalxpress.com/news/2026-06-patient-groups-vulnerable-privacy-medical.html>