SOCIOLOGIA,
INTEL·LIGÈNCIA ARTIFICIAL
I SALUT MENTAL

CODIFICAR LA SALUT MENTAL

Darrere la predicció i detecció de salut mental amb intel·ligència artificial.

Grau de Sociologia UAB
Rosanna Ramírez Nethersole
Tutoritzat per Enrico Mora

RESUM

La intel·ligència artificial (IA) s'incorpora de forma creixent en sistemes de predicció i detecció a la salut mental, un àmbit sanitari tradicionalment desatès, infrarepresentat i d'accés dificultós. Tot i ser una opció prometedora, l'estudi de la producció de la IA en el camp de la salut mental segueix sent incipient, així com els estudis que es centren en com aquesta tecnologia concep la salut mental. Així, per conèixer la noció de salut mental darrera el procés de creació d'aquests sistemes, aquest treball ha recopil·lat els discursos d'investigadors/es de ciència computacional que treballen per a la detecció i predicció de salut mental amb IA. A través d'aquesta anàlisis qualitativa, hem pogut identificar un posicionament comú respecte els objectius de la IA, la valoració de l'interdisciplinarietat de la salut mental, la noció de subjecte i, sobretot, la concepció de salut mental.

Paraules clau: *salut mental, intel·ligència artificial, detecció i predicció salut mental, sociologia de la salut, producció IA.*

ÍNDEX

I. INTRODUCCIÓ

La intel·ligència artificial (IA), i en concret l'aprenentatge automàtic, s'incorpora de forma creixent en sistemes de predicció i detecció a la salut mental, un àmbit sanitari tradicionalment desatès, infrarepresentat i d'accés dificultós. Tot i ser una opció prometedora, amb nombroses investigacions apuntant a la seva capacitat de millora en la precisió en la diagnosi, presa de decisions clíniques i detecció del risc de suïcidi (Mörch, Gupta & Michara, 2020), l'aportació de la IA en el camp de la salut mental segueix sent incipient, així com els seus potencials, dificultats i reptes.

En aquest context emergeix el creixent interès de les ciències socials per les implicacions d'aquestes innovacions tecnològiques aplicades a la salut mental. Encara que la literatura sociològica amb relació a la IA segueix sent limitada, en els darrers anys s'han desenvolupat diverses investigacions sobre les implicacions ètiques i el marc normatiu que regula la seva aplicació a la salut mental.

Aquesta proposta de tesi de grau, en canvi, pretén aprofundir en el procés de creació d'IA per a la detecció i predicció a la salut mental. Es tracta d'una anàlisi qualitativa dels discursos de els/les investigadors/es de ciència computacional que s'adrecen a la predicció i detecció de salut mental, per tal d'entendre entendre el món social del desenvolupament de la IA per a la salut mental. En concret, s'adreça a respondre el següent interrogant:

> *Quina noció de salut mental tenen els/les investigadors/es de ciència computacional de centres d'investigació públics a escala global en la producció d'IA per a la predicció i detecció de salut mental en els darrers cinc anys?*

De forma genèrica, l'objecte d'anàlisi és el conjunt complex que conforma el món social del desenvolupament de la IA per a la predició/detecció de depressió, ansietat, anorexia i ideació suicida. L'enfocament de la recerca s'emmarca en la branca de la sociologia de la ciència, àrea que analitza les condicions i efectes socials de la ciència, així com les estructures i processos socials de l'activitat científica (Ben-David & Sullivan, 1975).

Concretament, convé mencionar les perspectives complementàries de la Sociologia de la Traducció de Bruno Latour i l'enfocament *Actor-Network Theory* (ANT) de Michel Callon i John Law. Ambdues aborden l'anàlisi de la tecnociència com una xarxa d'interaccions simultàniament pràctiques i significatives, el resultat del qual no es descriu com un procés de negociació social, sinó com una lluita de poders que involucra agents humans i no humans (Domènech & Tirado, 1998).

En efecte, l'estudi integra l'imaginari social de la salut mental dels/les investigadors/es; una anàlisi de les dinàmiques, organització i estructura de producció de coneixement; la valoració i noció de salut mental; la concepció del subjecte; així com el grau de consideració del biaix algorítmic i riscos dels sistemes resultants d'aquest procés de creació.

A la vegada, tenint en compte que la definició de salut mental afecta i impacta la pròpia salut mental, presentant consequències reals sobre l'experiència viscuda d'un trastorn, creiem fortament que els resultats d'aquesta recerca no només permeten entendre la concepció de salut mental, sinó també sobre la seva posterior aplicació en relació a limitacions, oportunitats i riscos.

PERTINENÇA DE LA RECERCA

La IA és competència de la sociologia. La figura del sociòleg s'encarrega de l'anàlisi detallada de l'organització del món social i, tal com ja apuntava Harvey Sacks als anys seixanta, qualsevol tecnologia novella és "made at home in the world that has whatever organization it already has" (Sacks, 1992).

Tanmateix, tot i que la present *quarta revolució industrial* contribueix significativament en la transformació de la vida social, l'estudi de la IA des de les ciències socials segueix sent relativament insignificant. La sociologia contemporània presenta un buit de recerca teòrica i empírica tant en la concepció de la IA com a actor social (no humà) com amb relació al paper que la sociologia podria tenir durant el seu procés de desenvolupament (Mlynar, Alavi, Verma & Cantoni, 2018). En efecte, les condicions de producció dels serveis, sistemes i eines que incorporen IA s'investiguen, de forma predominant, des de les possibilitats tecnològiques, deixant la perspectiva social com a marginal o, en molts àmbits, inexistent.

Concretament, aquesta carència de coneixement social es veu accentuada en l'especificitat de la producció de serveis, sistemes i eines que incorporen IA per a l'atenció a la salut mental. Per aquest motiu, el potencial innovador del projecte en qüestió radica, en gran part, en emmarcar-se en un buit empíric de la sociologia de la ciència.

No només això, sinó que l'enfocament de la recerca suposa una oportunitat per diferenciar-se d'investigacions preexistents i generar coneixement nou. A hores d'ara, es desconeix l'existència d'altres estudis qualitatius que es focalitzin en els/les investigadors/es i les categories que utilitzen per a la creació d'IA per a la salut mental.

En definitiva, considerem que ens situem davant d'una oportunitat per a aplicar la perspectiva sociològica a un objecte d'anàlisi escassament explorat i, tot i les limitacions que suposa un treball de fi de grau, confiem que els seus resultats podran ser d'utilitat, interès i aportar coneixement nou.

II. OBJECTIUS

L'objectiu general és comprendre la noció de salut mental dels/les investigadors/es de ciència computacional que participen o han participat en els darrers cinc anys en el desenvolupament de serveis d'IA que prediuen o detecten la salut mental, oferint així una perspectiva social de la producció d'aquests sistemes. D'aquest propòsit se'n deriven els següents objectius específics:

1. Desenvolupar una recopilació analítica del coneixement sociològic acumulat sobre el desenvolupament i aplicació de la IA per a l'atenció a la salut mental.

2. Recollir informació sobre l'imaginari social de salut mental dels investigadors/es, la noció de subjecte, així com entendre com valoren la figura del/la psicòleg/a o psiquiatra i la dimensió social de la salut (perspectiva de gènere, classe i raça).

3. Inquirir sobre les decisions preses durant el desenvolupament d'IA, fent especial menció al procés d'entrenament d'algoritmes amb etiquetatge manual.

4. Establir patrons en els discursos socials analitzats.

5. Generar una imatge compartida de la concepció de salut mental en el desenvolupament de serveis d'IA per a l'atenció a la salut mental.

6. Reflexionar sobre les implicacions socials dels resultats obtinguts i generar prospectiva per a futures línies d'investigació.

III. APARELL CONCEPTUAL

> **<<** *Tres homes cecs es troben un elefant per primer cop. Per determinar quina mena de criatura és, cada un el toca amb les mans. El primer, que agafa la trompa, exclama que és una serp. El segon toca una cama i afirma que es tracta d'una mena d'arbre. El tercer, acariciant la panxa de l'animal, diu que l'elefant és una paret. >>*
>
> *[Conte anònim provinent de la India, traducció pròpia]*

Un dels aspectes fascinants del camp de la salut mental és que es sorprenentment difícil de definir. Tot i que entenem que una 'bona salut mental' va molt més enllà de l'absència de malestar, englobant diverses dimensions com l'autoestima, la consciència del potencial propi, el benestar psicològic i l'abilitat per mantenir relacions significants, el seu caràcter es notòriament complex i contradictori. Justament per això, la seva conceptualització és complexa i, sovint, ens dona més informació sobre l'observador/a que de la propia salut mental.

En aquest sentit, el/la sociòleg/a s'inquieta pels factors socials i culturals que s'entrellacen en la caracterització de la normalitat i abnormalitat; la relació i causalitat entre l'entorn social i el malestar psicològic; així com el desenvolupament de polítiques d'intervenció (Horwitz & Scheid, 1999). En efecte, la rellevància d'aquesta perspectiva recau en emfatitzar la influència de la societat mitjançant l'estatus social, habitualment derivat de la classe, gènere, edat, raça i etnicitat. D'una forma breu, les condicions de salut mental no es distribueixen de forma igualitària entre la societat, sinó que ocorren més sovint en grups socialment desfavorits (Aneshensel & Phelan, 1999).

D'altra banda, hi manca informació sobre com la IA entèn la salut mental. Per tal d'entendre el posicionament de l'esfera de producció de sistemes d'IA (interrogant clau de la nostra recerca) cal primerament revisar la literatura entorn la pròpia IA i la noció de subjecte en l'atenció psicológica.

En poques paraules, el terme d'IA implica l'ús d'un ordinador per a modelar comportaments intel·ligents amb intervenció humana mínima (Hamet & Tremblay, 2017). En aquest sentit, parteix de la premissa que 'tot aspecte de l'aprenentatge, així com qualsevol altre àmbit de la nostra intel·ligència, pot ser definida amb tanta precisió que una màquina és capaç de simular-la' (McCarthy a Boyd & Holton, 2019).

Es tracta d'un pressupòsit que ha desencadenat rèpliques des de diverses disciplines i en els darrers anys és notable l'augment d'aportacions que, des de camps tan dispars com són la sociologia, la psicologia o l'antropologia, s'interessen per abordar la dimensió social i simbòlica de la IA.

En aquest sentit, l'antropòloga Genevieve Bell (2015) proposa entendre la IA com a un conjunt desordenat format pels processos de *Big Data*, algoritmes, aprenentatge automàtic i sensors, entrellaçats per una lògica i racionalitat comuna (Bell a Holton & Boyd, 2019). D'aquesta aproximació volem incidir en la definició d'algoritme i aprenentatge automàtic:

En primer lloc, per algoritme fem referència a una sèrie d'instruccions que, rigides per normes, es manifesten a la nostra vida quotidiana, com ara als programes de rentatge dels rentavaixelles (Holton & Boyd, 2019). D'altra banda, l'aprenentatge automàtic fa referència al fet que, tal com indica el seu nom, les màquines que poden aprendre a desenvolupar els seus propis judicis i normes, en comptes d'operar simplement processos normatius desenvolupats per persones (Holton & Boyd, 2019).

NOCIÓ DE SUBJECTE / PACIENT

De la mateixa manera que la salut mental és sotmesa a diferents definicions, la noció d'individu també depèn de l'enfocament teòric i professional, així com dels judicis subjectius i patrons culturals. No obstant aquesta diversitat, en l'esfera de la salut (inclosa la salut mental), prevalen dues postures contrastades: la del subjecte i la del pacient.

El terme pacient es caracteritza per dos elements: la passivitat i l'espera (Menéndez a Perrotta, 2008). Per un costat, el pacient és passiu perquè el seu coneixement s'ubica en el professional i, per l'altre, espera del professional la resposta que resol el seu patiment. Així, es tracta d'un esquema de relació on el flux de coneixement és unidireccional i el pacient es limita a ser atès. L'especialista conforma el saber per excel·lència i el discurs mèdic exclou la singularitat del pacient: es parla de casos, històries clíniques i òrgans (Perrotta, 2011).

Tornant a l'especificitat del camp de la salut mental, concebre l'individu com a pacient repercuteix en l'atenció i experiència viscuda del trastorn. Així ho demostra la recerca de Martínez (2012), on l'atenció al pacient es caracteritza per un diagnòstic precipitat i estigmatitzant; absència d'anàlisi d'històries de vida; discursos sustentats en l'esfera biològica, cerebral i genètica; protocols estrictes i infantilitzadors; així com al·locucions mèdiques i paramèdiques moralistes i disciplinaris basades en el 'sentit comú' (Martínez a Perrotta, 2011).

Aquesta objectivació de l'individu es recolza i reafirma en la pretensió d'una objectivitat ateòrica de la salut mental. Els aspectes diferencials es neutralitzen, s'anul·la la diversitat i el component social esdevé trivial. D'aquesta premissa en surten els models de pensament basats en la classificació d'allò normal i allò patològic, tal com el reconegut *Manual estadístic i diagnòstic de malalties mentals* (DSM).

D'acord amb el que Hacking (1995) anomena 'la invenció o construcció de persones', les pràctiques classificatòries de trastorns mentals creen noves possibilitats

d'elecció i acció per a l'individu, de qui o que és i de què pot fer. Més enllà, aquests canvis en la conducta derivats de la classificació influencien directament les mateixes categories de classificació, el que Hacking descriu com 'efecte bucle'[1]. En efecte, la classificació produeix efectes sobre les persones i aquests canvis influencien la pràctica classificatòria.

Com 'la dimensió subjectiva del sofriment esdevé opaca amb la dimensió objectiva del pacient' (Margaril a Rivas & Pimentel, 1996), la noció de subjecte és indefugiblement contrària a la del pacient. Concebre l'individu com a subjecte comporta la dissolució de les dicotomies (sa/malalt, normal/patològic, etc.), entendre la salut mental des de la complexitat, com a un procés social subjecte a canvis i transformacions.

Tenint en compte aquesta complexitat, és coherent que la conceptualització de subjecte sigui sotmesa a debat. Concretament, trobem posicions distinguides amb relació al grau d'autonomia que otorgen al subjecte en la producció de sentit.

Així, diversos autors aposten per al terme subjectivitat en comptes de subjecte, o més aviat, subjectivitats, perquè aquesta és múltiple, dispar, fragmentada i heterogènia, ja que existeixen tantes subjectivitats com situacions i moments (Guatarri a Moreschi, 2013). Segons aquesta perspectiva, les identitats són punts de adhesió temporals a les posicions subjectives que construeixen les nostres pràctiques discursives (Hall a Moreschi, 2013).

De forma paral·lela, altres contemplen que afirmar-se com a subjecte significa ser capaç de reflexionar sobre un mateix i poder ubicar-se en la seva vida, una vida que controla però a la vegada ve imposada per naixement (Touraine a Moreschi, 2013).

En síntesi, no obstant els matissos teòrics, podem concebre el subjecte com a un agent transformador i productor de significats que responen a estrucutres socials més o menys determinants (Moreschi, 2013).

IV. ESTAT DE L'ART

Abans d'iniciar el recorregut pel corpus textual[2] que conforma l'estat de la qüestió de la IA aplicada a la salut mental, creiem necessari mencionar dues consideracions prèvies. En primer lloc, donada l'abundància de formes d'encarar la revisió documental de la IA- centrada en la recepció, teràpia, marc normatiu que regula la seva implementació, etc. -convé reiterar que la present revisió s'adequa a l'objecte d'anàlisi propi; les interaccions entre investigadors/es i els processos tecnològics durant el procés de creació de sistemes de detecció de salut mental. Volem tornar a incidir en aquesta selecció d'informació per explicar l'omissió d'una mirada essencial; la que prové de les persones que reben els serveis d'atenció a la salut mental.

[1] Així ho expressa Hacking en les seves pròpies paraules a l'entrevista amb Álvarez: <<*Faconner les gens es de hecho un intento por expresar en francés no 'Hacer gente' (Making People) sino inventar/construir gente (Making up people) (...) Por encima de todo, me interesa: (i) cómo nuevas clasificaciones de personas crean nuevas posibilidades de elección y de acción, de quién o qué es uno y qué puede uno hacer; (ii) lo que las nuevas clasificaciones les hacen a las personas clasificadas, y como también cambian por ser así clasificadas; (iii) cómo esos mismos cambios en las personas cambian nuestras teorías de las clasificaciones. Esto es lo que llamo un efecto de bucle.* >>

[2] Per a una informació més detallada sobre el corpus textual cal veure els criteris d'elegibilitat (Annex 1) i la bibliografia per al recull de la literatura consultada.

D'altra banda, un altre aspecte a considerar és que el panorama empíric en la nostra àrea d'estudi continua sent escàs i, conseqüentment, ens situem davant d'una bibliografia fragmentària, dispar i poc consensuada. L'objecte d'aquest apartat és servir com a mirada general i interdisciplinària, distant de la discussió i crítica, per constatar el que hi ha fet, el que hi manca fer i, sobretot, quin és l'estat de la qüestió a nivell empíric.

Detecció i predicció amb aprenentatge automàtic

L'alta prevalença de condicions de salut mental i la insuficiència d'atenció efectiva, combinada amb els recents avenços de la IA, ha donat peu a una creixent exploració de com l'aprenentatge automàtic[3] pot assistir en la salut mental (Thieme, Belgrave & Doherty, 2020). Certament, les àrees de recerca són diverses – detecció i diagnosis; tractament, suport i teràpia; salut pública; recerca i administració – però dins d'aquest conjunt heterogeni, ens adrecem al fragment que comprèn l'aprenentatge automàtic per a la detecció i predicció de la salut mental.

Concretament, aquest àmbit de recerca empírica se subdivideix en dues branques predominants: investigacions que busquen (i) comprendre, detectar i diagnosticar els símptomes i resultats de la salut mental; i estudis que (ii) avaluen les relacions entre pacients i professionals, buscant formes de millorar l'atenció tradicional (Thieme, Belgrave & Doherty, 2020). Tenint en compte el nostre objecte d'anàlisi, així com la inexistència d'estudis centrats en el mateix procés de creació de sistemes, hem analitzat l'estat de la qüestió del primer bloc esmentat.

D'entrada, la detecció i predicció de condicions de salut mental és un camp incipient però prometedor, amb nombroses investigacions apuntant a la capacitat de millora en la precisió de la diagnosi, presa de decisions clíniques i identificació del risc de suïcidi (Mörch, Gupta & Michara, 2020). A més, es tracta d'una exploració que ha estat fortament accelerada per la recent aparició i proliferació de les xarxes socials, la qual ha facilitat l'accés a grans volums de dades per a l'entrenament dels algoritmes i, en alguns casos, la seva aplicació.

De forma genèrica, l'objectiu general de la literatura consultada és aconseguir una detecció eficaç i precisa de la salut mental. Així mateix, sota aquest propòsit s'articulen diferents objectius específics: identificar els símptomes, caracteritzar malalties i predir el seu desenvolupament.

D'una banda, amb referència a l'anàlisi dels símptomes i caracterització de condicions, destaca sobretot l'estudi de la depressió, però també dels trastorns alimentaris i la ideació suïcida. En aquest marc s'han realitzat investigacions sobre la comprensió de comportaments *desviats*[4] entre comunitats en línia (Chancellor, 2018), la identificació de condicions i caracterització del seu grau de severitat (Ramirez-Cifuentes, D. et al, 2021),

[3] L'aprenentatge automàtic (*machine learning*) es defineix com la branca de la IA que, mitjançant mètodes i estils que poden variar, habilita que l'algoritme "aprengui" (Bzdok, Meyer-Lindenberg, 2018). Planteja que els algoritmes poden millorar automàticament a través de l'experiència i ús de les dades.

[4] Convé comentar l'ús del terme *desviat* de l'article <<Computational methods to understand deviant mental wellness communities>> (Chancellor, 2018), ja que s'explicita la referència a la concepció sociològica de desviació proposada per Akers a *Social Learning and Deviant Behaviour* (1977). Més concretament, Chancellor entén per a comportaments desviats les accions que violen les normes i comportaments socials d'una determinada comunitat (Chancellor, 2018). En aquest sentit, un exemple són les comunitats en línia que mantenen, promouen i reforcen conductes anorèxiques.

així com generar eines d'informació rendibles per a professionals (Gaur, Kursuncu, Alambo, Sheth, Daniulaityte, Thirunarayan, Pathak, 2018).

D'altra banda, els estudis de predicció del risc associat a condicions de salut mental es tendeixen a centrar en el suïcidi. A partir d'interaccions a les xarxes s'identifica i monitoreja persones amb risc de desenvolupar una condició de salut mental (Chen, Sykora, Jackson, Elayan, 2018), es proporciona als departaments d'emergència una eina d'avaluació del risc basada en evidència per preveure intents de suïcidi reiterats (Pestian, Matykiewicz, Grupp- Phelan, 2008) i, principalment, és pretén millorar la prevenció i detecció precoç del suïcidi.

En conjunt, tot i que la metodologia varia, la literatura consultada tendeix a fonamentar-se en una examinació del llenguatge, siguin audios o bé textos compartits de les xarxes. En aquest sentit, la metodologia predominant és *Natural Language Processing,* una examinació i processament del llenguatge que involucra traducció, comprensió semàntica i extracció d'informació (Hirschberg, Manning, 2015).

Paral·lelament, amb relació al procés d'entrenament dels algoritmes, en el camp de la salut mental predomina l'aprenentatge automàtic supervisat. És a dir, les condicions de salut mental són etiquetades (per categories o nivells) per l'investigador/a o un expert (normalment psicòleg o psiquiatre) i, a partir d'aquest etiquetatge l'algoritme aprèn a associar les etiquetes amb un gran volum de dades sociodemogràfiques, biològiques o mesures clíniques (Graham, Depp, Lee, Nebeker & Tu, 2019). En el cas de sistemes de predicció, les dades poden proveir de registres de salut mental, escales de valoració de l'estat d'ànim, imatges cerebrals, sistemes de monitorització com el mòbil o vídeo o, xarxes socials (Thieme, Belgrave & Doherty, 2020).

Dimesió social de la salut mental

És àmpliament reconegut que la salut mental varia segons l'estatus de gènere, raça i classe, a més d'estar relacionada amb una diversitat de factors com són l'ocupació, els moviments socials, l'estructura familiar, el rol adquirit, l'estrès, la identitat, el matrimoni o l'urbanisme (Aneshensel & Phelan, 1999; Avison & Gotlib, 1994; Brown, 2003; Cockerham, 1989; Horwitz & Scheid, 1999; Link & Phelan, 1995; Outlaw, 1993; Wheaton, 2001 a Brown, 2003)

La sociologia de la salut mental se centra en la seva epidemiologia, etiologia, correlacions i conseqüències, en un intent de descriure i explicar la influència de l'estructura social en la salut psicològica d'un individu (Brown, 2003). A la vegada, la sociologia de la ciència, àrea que analitza les condicions i efectes socials de la ciència, així com les estructures i processos socials de l'activitat científica (Ben-David & Sullivan, 1975), és fortament present en l'anàlisi de les interaccions, teorització, procediments i metodologia de l'atenció a la salut mental.

No obstant això, minva exploració social sobre el desenvolupament de la IA i sobretot en l'àmbit de la seva aplicació en la salut mental. Concretament, avui dia, no coneixem cap recerca empírica centrada en els/les investigadors/es que generen sistemes d'aprenentatge automàtic que detecten i prediuen la salut mental. Per aquest motiu, creiem que ens enfrontem davant d'una oportunitat esperançadora, els resultats del qual podrien aportar al buit empíric actual.

Aquesta manca d'investigació, però, no és exclusiva de les ciències socials, ja que els estudis d'IA tampoc analitzen en profunditat l'aspecte social de la salut mental. Tot i que les condicions de salut mental no es distribueixen de forma igualitària entre la

societat, sinó que ocorren més sovint en grups socialment desfavorits (Aneshensel & Phelan, 1999), són pocs els estudis d'IA que incorporen variables de gènere, classe i raça.

De fet, en el camp de la detecció i predicció amb aprenentatge automàtic hi ha relativament poc coneixement sobre la relació entre l'esfera biològica, psicològica i social de la salut mental (Graham, Depp, Lee, Nebeker & Tu, 2019).

Ara bé, aquesta absència no implica necessàriament la negació de l'aspecte social, sinó que més aviat minimitza la seva influència al entendre-la com una peça més del conjunt enrevessat que conformen les condicions analitzades. En aquest sentit, la literatura consultada parteix del fet que, fenòmens com el suïcidi, presenten una etiologia complexa formada per una xarxa de factors psicològics, biològics, "de l'entorn" i econòmics (Fonseka, Bhat & Kennedy, 2019).

Un punt destacable per la seva distància envers la perspectiva sociològica és que els estudis d'IA plantegen, en diferent grau, la individualitat de l'origen de la salut mental. Concretament, s'atribueix causalitat a alteracions neuroquímiques, neuroendocrines i genètiques/epigenètiques (Turecki a Fonseka, Bhat, Kennedy, 2019).

D'aquesta manera, la perspectiva social curteja i es limita a informació bàsica demogràfica, sobretot l'edat i el *gènere*[5]. Aquesta selecció s'explica, en part, per les dificultats d'accés de les dades provinents de xarxes socials i la capacitat d'extreure informació. En el cas de l'edat i el gènere, aquesta informació s'aconsegueix mitjançant lèxics (Mowery, Park, Conway & Bryan, 2016).

Finalment, no podríem finalitzar aquest apartat sense ressaltar alguna recerca d'IA que és distingeix del panorama predominant per integrar, en cert grau, una més complexa perspectiva social en l'anàlisi de la salut mental. En aquest sentit, destaca el model de predicció de depressió per a la població geriàtrica d'un barri marginal de Calculta, Índia, proposat per Sau & Bhakta l'any 2017. Es tracta d'un model que considera variables sociodemogràfiques d'edat, gènere, alfabetització, cònjuge viu, estat de treball, ingressos personals, tipus de família, abús de substàncies, així com altres condicions com problemes visuals, problemes de mobilitat, problemes d'audició i dificultats relacionades amb el son (Sau & Bhakta, 2017).

Així, són molts els aspectes, com hem vist, pendents d'investigació social empírica en l'àrea d'aprenentatge automàtic i salut mental. La necessitat d'anàlisi s'intensifica quan, després de revisar la literatura de la IA, es relaciona amb el limitat coneixement sobre l'etiologia de les condicions que pretenen atendre, així com la poca presència de variables socials més enllà del sexe i l'edat.

Explorar el procés de detecció i predicció de la salut mental, centrant-nos en els individus darrere dels algoritmes, ens permet trencar amb la tendència de la sociologia a limitar-se a estudis teòrics sobre la conceptualització o aplicació de la IA. No només això, sinó que creiem que, entendre els imaginaris socials del procés de creació dels sistemes, ens permet accentuar la distància analítica envers els sistemes de detecció i predicció, facilitant l'objectiu a llarg termini que és oferir una mirada més autèntica de com alleujar l'alta prevalença de condicions de salut mentals actual amb IA.

[5] Entre la literatura consultada el terme empleat és de *gènere*, però dubtem de si és més adient la denominació *sexe*, d'una banda, perquè la categorització tendeix a ser binària i, d'altra banda, per una manca d'informació sobre la concepció de gènere amb la que s'entrena l'algoritme que determina el gènere dels usuaris que no ho expliciten al seu perfil.

V. METODOLOGIA

DISSENY

D'entrada, aquesta recerca qualitativa s'articula des del mètode l'abductiu. El principal atractiu de l'abducció recau en el seu element creatiu d'introduir noves idees en la ciència, a diferència de la inducció que determina un valor i la deducció que desenvolupa les conseqüències necessàries d'una hipòtesi (Nubiola, 1998).

Amb referència a les tècniques d'investigació, en un inici es va realitzar una recopilació documental de l'estat de l'art del tema de recerca i, de forma paral·lela, entrevistes exploratòries amb investigadors/es del camp. Arribat aquest punt, per tal de conèixer amb més profunditat els discursos darrere la predicció i detecció de salut mental amb IA, el nostre objecte d'anàlisi, es van portar a terme entrevistes anònimes, no confidencials i semiestructurades de resposta oberta[6].

Així, en conjunt s'han realitzat set entrevistes, de les quals dos han estat exploratòries amb una psicòloga que treballava amb IA per a la salut mental i una investigadora de ciència computacional que treballa per a l'atenció i seguiment de persones en risc de suicidi. La resta conformen l'objecte d'anàlisis: investigadors/es de ciència computacional de l'àmbit de detecció i predicció de trastorns.

En efecte, la rellevància analítica de l'entrevista recau en construir una representativitat hologramàtica[7] mitjançant estructures de significat i argumentació individuals, fent possible contemplar l'especificitat sense perdre connexió amb la globalitat del fenomen: entendre que la societat està formada per individus però, a la vegada, cada individu reflecteix la societat (Morin a Elorriaga, Lugo & Montero, 2013).

SELECCIÓ INVESTIGADORS/ES

Les persones entrevistades complien tres condicions: presentar una trajectòria acadèmica en ciència computacional o disciplines semblants, treballar a l'actualitat o en els darrers cinc anys a una institució universitària per a la predicció i detecció de SM amb IA, així com presentar com a mínim un article publicat en aquest àmbit d'investigació.

L'accés a les persones entrevistades ha estat per bola de neu i ens ha portat a conversar amb una gran varietat de perfils, tant amb relació a les condicions a les quals s'adreçaven a prediure i/o detectar com amb referència a la seva zona geogràfica.

Concretament, la distribució de condicions ha estat la següent: tres presenten experiència per a la detecció i/o predicció d'anorexia; cinc en el cas de la depressió; dos en l'àrea d'ansietat; i quatre en ideació i risc de suïcidi. D'altra banda, els/les investigadors/es treballen o havien rebut la seva formació en una diversitat de països: Austràlia, Espanya, Argentina, Israel, Panamà i Equador.

Per últim, hem entrevistat a tres investigadores i quatre investigadors. En aquest sentit estem satisfetes perquè la ciència computacional tendeix a ser una disciplina més bé masculinitzada.

---

[6] Les entrevistes han estat verificades pel tutor i la sol·licitud de les transcripcions queda a disposició del tribunal.

[7] El principi hologramàtic proposat per Morin (2006) expressa que les parts constitueixen un "tot", així com a la vegada el "tot" es troba potencialment en cadascuna de les parts. Des d'aquesta perspectiva, es planteja que no es poden estudiar les parts sense entendre el conjunt així com estudiar el conjunt sense entendre les seves parts. La realitat es presenta com a conjunt i parts a la vegada, l'un conté l'altre (Morin a Elorriaga, Lugo & Montero, 2013).

ANÀLISI DE LES DADES

Un cop recollits els discursos, la tècnica empleada ha estat l'anàlisi de contingut, una metodologia que permet aïllar unitats, categoritzar i establir relacions, així com interpretar els significats en context expressiu i comunicatiu (Riba Campos, 2022).

Així, hem realitzat una codificació de les transcripcions d'entrevistes amb el programa *Atlas.ti*. L'elaboració dels codis s'ha dut a terme en funció dels temes i dimensions prèviament identificats a la recopil·lació documental[8].

En conjunt, el treball de camp va generar al voltant de sis hores i mitja d'entrevista, 55 pàgines de transcripció i un total de 244 cites per als 37 codis ordenats en 15 dimensions.

VI. RESULTATS I DISCUSSIÓ

Tot i els matisos entre les persones entrevistades[9], ha estat notable l'analogia d'experiències d'investigació i percepcions entorn la salut mental. Concretament, obtenim resultats sobre el context social de producció, la valoració de la interdisciplinarietat de salut mental així com la noció de subjecte i salut mental.

Abans dels resultats, però, creiem convenient recordar les limitacions de la present recerca. En primer lloc, l'anàlisi se cenyeix a la producció d'IA des d'institucions universitàries. Així, els resultats només al·ludeixen a la particularitat d'aquesta esfera de recerca i, de fet, tal com apunten algunes entrevistes[10], intuïm que el caràcter privat o públic esdevé clau en la determinació del caràcter i formes de producció d'IA.

Un altre punt a recalcar és que dins del mateix àmbit universitari la investigació de ciència computacional i salut mental segueix sent fortament heterogènia. Aquesta diversitat del camp l'hem pogut comprovar a les entrevistes exploratòries, on els estudis destinats al seguiment i atenció es diferencien fortament en les formes de producció d'aquelles investigacions que prediuen i detecten la salut mental.

Ara bé, l'especificitat de l'objecte d'anàlisi no incideix en la rellevància dels resultats. Certament, la predicció i detecció conforma un petit fragment en el conjunt d'IA i salut mental però es tracta d'un àmbit amb una repercussió extraordinària, on tots els/les usuaris/usuàries de les principals plataformes socials són potencials receptors/es.

A més, han estat justament aquestes diferències d'altres grups universitaris de recerca d'IA i salut mental que han facilitat l'identificació de tendències i imaginaris en els discursos del nostre estudi. Així, poder establir converses amb experts/es de salut

[8] La llista de dimensions, temes i codis esta annexada al final del treball, al Annex 2.

[9] Per a una informació més detallada sobre les diferències entre les persones entrevistades cal veure l'apartat de metodologia.

[10] A les entrevistes ens hem trobat amb una varietat de valoracions i diferències envers al sector privat:

> *<< [Sobre la investigació amb IA des del sector privat] (…) La mayoría son hombres que están pensando en el bonus económico del próximo año. De hecho no los obligan a pensar sobre esto nunca y son jóvenes que nunca han asistido a una sola clase de ética. >> E6: B*
>
> *<<Definitivamente es diferente investigar des del sector privado pero nuestro objetivo es que mi trabajo sirva para estas grandes plataformas y se implemente. >> E2: R*

mental i aprenentatge automàtic a les entrevistes exploratòries ha estat crucial en aquest aspecte.

Així doncs, a continuació detallem les principals troballes de l'estudi d'una part del conjunt interdisciplinari i plural que és la IA.

(I) SITUAR LA PREDICCIÓ I DIAGNOSIS

Per tal d'entendre des d'on s'impulsa la recerca, les motivacions i prioritats, hem analitzat el contingut dels discursos que caracteritzen els projectes i la posició dels/les investigadors/es de ciència computacional, així com contrastat ambdós elements.

En aquest sentit, un primer punt a comprendre és el propòsit principal darrere la predicció i diagnosis de salut mental amb IA. Contrari al que anticipàvem, els discursos recollits mostren una clara distinció entre els objectius i l'aplicació dels projectes. D'una banda, els discursos sobre els objectius tendeixen a orientar-se al voltant de la tecnologia:

> *<<El objetivo era predecir la personalidad de la persona que hablaba con la menor cantidad de texto posible. De la forma más eficaz y a través de una conversación natural. >> E1: E*

Una bona evidència d'aquest fet és la distinció entre la funció de la tecnologia i la pròpia capacitat tecnològica és el següent comentari:

> *<<El propósito de la investigación no está en detectar o predecir la salud mental mediante los escritos históricos en redes sociales sino en la detección temprana, en la dimensión temporal. Entonces, no es solo detectar el riesgo, es detectar el riesgo de la forma más temprana.>> E3: D*

En canvi, les finalitats de la seva implementació són sempre les conseqüències sociosanitàries: innovar en una àrea actualment desatesa per alleugerar problemàtiques, ampliar la capacitat d'atenció i millorar l'eficiència mitjançant eines intel·ligents que poden assistir als professionals.

> *<<Descubrir aquellos que están desarrollando un desorden mental y eso podría ser investigado o tratado por un profesional de la salud. Esta es la parte de intervención y yo no traté este ámbito, yo trato la parte de detectar las "pistas" del lenguaje cuando alguien está desarrollando un desorden mental. Esta sería la practicidad. Poder aplicar esto a una escala mucho más grande de la que se podría con una persona, con la limitación propia humana.>> E1: E*

Assolint el mateix objectiu, un investigador ens comenta com la seva recerca s'orienta al tercer sector:

> *<<We can give this information to social organisations, non profit organisations, we can offer insight to help people and make more accurate interventions with particular populations.>> E4: I*

La rellevància d'aquesta distinció entre objectius i aplicació recau en el fet que la valorització de la IA no es relaciona amb la seva implementació sinó que, més aviat, es postula com una finalitat per si mateixa:

> *<< [Sobre una publicació de predicció de suicidi i trastorns alimenticis amb IA] Este artículo está más orientado a la parte informática y no tanto al aspecto de los psicólogos o psiquiatras>> E2: R*

D'una forma, sembla que en la 'predicció de salut mental' la 'predicció' i la 'salut mental' es postulen com dos termes divisibles, sent el primer càrrec i meta de la IA i el segon la seva conseqüència. Aquesta escissió no només suposa una ruptura amb el pressupòsit que la tecnologia és un mitjà i no una finalitat (Heidegger a Hanks, 2010) sinó que, com veurem, influencia moltes més àrees del disseny d'aquestes tecnologies.

De fet, apareix aquesta distància entre salut mental i IA en els discursos sobre l'experiència personal de l'investigador/a. En primer lloc, la decisió de treballar amb salut mental ve influenciada per un supervisor o l'existència de finançament per a projectes d'IA de caràcter social o sanitari. Així, es tracta d'una presa de decisió on l'interès personal és gairebé marginal i, de fet, entre totes les persones entrevistades no hi ha cap trajectòria acadèmica amb alguna mena de formació en salut mental.

Un cas pràctic d'aquesta situació és l'elecció de condicions analitzades. Tot i que entenem que tota investigació és limitada per les possibilitats reals d'accés a informació, ens sorprèn que l'únic element mencionat per escollir treballar amb trastorns alimentaris, ideació suïcida, depressió o ansietat sigui la disponibilitat de dades:

> *<<Esto viene limitado por la disponibilidad de datasets. La verdad es que no hay muchos datasets, al menos de redes sociales. También porque es muy complejo la gestión de los permisos para hacer este tipo de estudios, los comités de ética… (…) Pero bueno, respondiendo a tu pregunta, es básicamente por la disponibilidad de las colecciones. Los datos que se encontraban disponibles.>>* *E1: E*

De fet, aquesta distància s'accentua en el cas d'un investigador que ens explica que per al seu equip d'investigació l'objecte d'estudi canvia periòdicament:

> *<< El objetivo y el reto es la detección de diferentes condiciones vinculadas a la salud y seguridad, las cuales cada año van cambiando: un año fue el suicidio, otro año fueron los trastornos alimenticios y la ansiedad, otro también fue el alcoholismo.>>* *E3: D*

Arribat aquest punt convé reiterar que amb tot això no volem insinuar un desinterès personal per part dels/les investigadors/es participants sinó descriure l'estructura que guia la recerca. És més, a les entrevistes és notable i recorrent l'interès per discriminacions en l'atenció a la salut mental així com la preocupació per als presents obstacles de la salut mental. De fet, dues investigadores ens expliquen que sovint han de justificar l'interès i necessitat de recerca enfront d'actituds aprehensives dins la seva pròpia disciplina:

> *<<Hay mucha gente que no le veía futuro a este proyecto porque la gente apunta a la parte técnica e informática, no apunta a trabajos que sean interdisciplinares. Yo no creo esto, yo creo que no debería haber este estigma hacia disciplinas más sociales. (…) Por parte de mi supervisor hubo mucha apertura pero otra gente nos decía "deja que la gente que se mate" o "tienen la libertad de suicidarse" o "¿y a ustedes que les interesa?">>* *E2: R*

En aquest sentit és interessant el que ens comenta una investigadora sobre el tipus d'objecte d'estudi ideal de la IA:

> *<< En computer science es como que es más deseable trabajar con algoritmos que te dan una predicción más clara, como la diabetes: si consumes una cantidad*

de azúcar o si fumas, pasa tal. (…) Se te quedan mirando como "¿¡de qué estas hablando!?" >> E5: G

Aquestes darreres cites ens porten a considerar que no només la valorització de la IA és independent de la salut mental, sinó que certs àmbits de la comunitat de ciència computacional consideren la informàtica com a superior. No obstant això, en aquesta investigació no tenim suficient informació sobre aquest fet.

En conjunt, recollint tot el que s'ha esmentat en aquest primer bloc, obtenim que la formulació dels objectius i valoració de la IA és independent a la salut mental, encara quan s'adreça a la predicció i diagnosis. Així, s'estableix una distància congruent envers el camp d'anàlisis que es manifesta en el disseny de les investigacions així com al llarg de l'experiència dels/les investigadors/es. D'aquesta manera, percebem una relació instrumental envers la salut mental.

***(*II) INTERDISCIPLINARIETAT DE LA SALUT MENTAL**

Com la salut mental és inherentment indisciplinar – sent els seus impulsos socials, psicològics, geogràfics, polítics, entre d'altres – sembla coherent que els equips que treballen en aquest camp també ho siguin. No obstant això, en la investigació d'IA per a la predicció i detecció de salut mental, falta un marc que pautin aquestes col·laboracions així com coneixement sobre com es perceben, plantegen i duen a terme.

Per aquest motiu, en aquest apartat ens adrecem a entendre com es posicionen els/les investigadors/es de ciència computacional i conèixer com es reflecteix en la seva pròpia trajectòria de recerca. Concretament, inquirim sobretot en la valoració dels psicòlegs i psiquiatres, però també per com conceben la dimensió social de la salut mental i el rol de la sociologia.

Valoració Psicologia i Psiquiatria

D'entrada, en el camp de la IA i salut mental, el conjunt de persones entrevistades aposten per a una fusió de recerca disciplinària entre la psicologia i l'aprenentatge informàtic a llarg termini.

Tanmateix, apareixen diferències entre els discursos quan es pregunta per l'actualitat o l'experiència personal. De fet, amb referència al rol dels psicòlegs i psiquiatres, les concepcions sobre la forma i límits d'aquesta col·laboració varia fortament. D'una banda, ens trobem amb orientacions que defensen una "informàtica pura":

<<Depende del método que quieras probar, claro, mi tesis es de informática y tenemos que ser puristas: como psicólogo puedes aportar a la psicología pero no a la parte informática. (…) En el segundo artículo no consultamos psicólogos y psiquiatras también por esto, porque quisimos ser más puristas>> E2: R

D'altra banda, altres aposten per una tecnologia produïda sota i per a una mirada experta:

<<Esa es una de las primeras preguntas: quién asume esta competencia. Muchas veces esto es decidido por personas que no tienen la competencia, no técnica, sino política o administrativa. Lo hizo un ingeniero que lo encontró "guay" pero se le olvidó preguntarle a la persona que puede decidir si eso se puede hacer o

no, ¿entiendes?. No siempre la persona que lo hace es la persona que decide.>> *E6: B*

Aquesta pluralitat de posicionaments es reflecteix en les experiències d'investigació, on la tendència és una col·laboració molt limitada o directament inexistent. De fet, com no hi ha cap factor que obligui a la participació d'experts, a vegades la decisió sembla ser arbitrària:

<<No necesitas siempre la figura del experto, sería ideal, pero no necesario, los escritos personales ya son casos útiles donde el riesgo se puede asumir. (…) La decisión que llevó consultar expertos en un artículo y no en el otro fue que el primero es una caracterización de la bulimia, de que consiste y cómo se manifiesta en las redes, mientras que el segundo es una predicción y no interesa tanto cómo se definen estas condiciones.>> *E2: R*

Tot i que en diversos casos no participa cap expert de salut mental, en el cas dels articles on sí que hi ha hagut alguna mena de col·laboració, es tracta d'una participació puntual en la fase primerenca de la definició dels trastorns o durant el procés d'etiquetatge[11]:

<< *Nos ayudaron explicandonos sobre las frases y palabras que podíamos buscar para cada trastorno, nosotros recolectamos ciertas frases de forums y tomamos las frases que se repetian más y las llevamos a los psicólogos y psiquiatras para que las validaran. Ellos dijeron "esta si y esta no". También nos describieron el curso de cada trastorno, de qué va y cómo se manifiesta. Sobre todo su rol fue de etiquetar los datos con el que se entrenó después el algoritmo.>>* *E2: R*

La valoració dels experts, per tant, aparentment contradictòria. D'una banda, es posicionen d'acord amb el pressupòsit d'una investigació interdisciplinari, però, d'altra banda, a la pràctica les col·laboracions esdevenen limitades i escasses.

No només això, sinó que hi ha una desconeixença de la recepció dels sistemes en l'esfera dels psicòlegs i psiquiatres. Quan els hi pregunto, em contesten que no tenen informació sobre la valoració dels/les professionals de salut mental amb relació a aquests sistemes, encara quan, en la majoria de casos, la tecnologia s'adreça a assistir als/les experts/es. Aquesta situació connecta amb la falta de contacte, d'un canal d'informació interdisciplinari, que mencionàvem a l'apartat anterior.

Aquesta dissonància es trasllada també al debat que emergeix en preguntar sobre l'aplicació dels sistemes i la possibilitat d'automatització, és a dir, funcionar sense la necessitat d'una assistència experta. Les persones entrevistades accentuen com la IA serà una eina d'ajuda per als i les professionals, però les percepcions entorn la substitució, l'automatització absoluta, varien.

En primer lloc, algunes persones consideren poc probable aconseguir una automatització absoluta:

[11]L'etiquetatge és una técnica per reconèixer dades sense processar que poden ser imatges, text, vídeos, etc. Es tracta d'un procés en que s'atribueix una o més etiquetes informatives per determinar a partir d'aquesta informació un model d'aprenentatge automàtic. Un exemple podria ser l'etiquetatge de posts de xarxes socials segons etiquetes de sentiment (positiu, negatiu o neutral).

> *<< A corto término el objetivo es enriquecer y asistir a los expertos de salud mental existentes. A largo término… No sé si se podrá automatizar, es muy difícil automatizar. >> E3: D*

Segonament, altres investigadors/es rebutjen la possibilitat de dissenyar eines que substitueixin als/les experts/es:

> *<<La verdad es que en mi caso no se ha llegado a implementar nada en particular para uso, digamos para uso diario, pero lo que sí que he leído es que hay ciertas reticencias al uso de estas tecnologías muchas veces porque quizás no se entienden y se contemplan como una competencia hasta estos expertos de la salud, y esto no es así…(…) La cosa que yo siempre intento diferenciar es que, nosotros de ninguna manera, o por lo menos yo con mi trabajo, no intento reemplazar los profesionales de la salud sino asistirlos. La idea es ¿cómo asistirlos? bueno, pues una manera es que estos tipos de algoritmos tienen una capacidad de procesar millones de usuarios a la vez y esto permitiría, no sé como se dice en español, pero hacer screening de los usuarios.>> E1: E*

En darrer lloc, altres afirmen tant la seva inclinació per a l'automatització així com la capacitat tecnològica per fer-ho a llarg termini:

> *[Sobre si es posiciona d'acord amb la possibilitat de substituir als experts mitjançant eines d'IA] << Yes, short term it is a tool but long term yes, this is possible.>> E4: I*

> *<< Nuestro reto es traducir el diagnóstico clínico a las redes sociales y capturarlo.>> E2: R*

En conjunt, els discursos sobre la interdisciplinarietat de la salut mental es diferencia fortament de la pràctica. No és habitual que participin psicòlegs o psiquiatres i no hi ha cap element que requereixi la seva intervenció, la incorporació d'un/a expert/a sembla ser una decisió personal, una decisió per enriquir la recerca però no necessària. En els casos que participen, els/les professionals de salut mental, tot i ser els futurs receptors d'aquestes tecnologies, participen de forma secundària, establint col·laboracions puntuals per a la producció dels sistemes. Per tant, no podem pas afirmar l'existència d'equips interdisciplinaris sinó equips de ciència computacional amb col·laboracions secundàries i puntuals.

Salut mental social

Tenint en compte l'estat de contribució dels psicòlegs i psiquiatres, és coherent que el científic social sigui inexistent en la producció d'IA per a la predicció i detecció. Tanmateix, el conjunt de persones entrevistades reconeixen la importància d'introduir, en un futur, la dimensió social en les seves anàlisis, posicionant-se d'acord amb el pressupòsit que la salut mental es concentra en els grups socialment desfavorits.

> *<<Creo que puede ser muy importante incluir estas variables para encontrar los grupos más propensos a un desorden mental. Así nos podríamos concentrar en ciertos sectores de este mundo digital y los algoritmos podrían obtener mejores resultados.>> E3: D*

Ens comenten que el principal obstacle per la introducció de la perspectiva sociològica és la disponibilitat de dades, ja que la informació accessible a les xarxes

socials se cenyeix a informació demogràfica bàsica. En tots els casos, alguns estudis incorporen variables socials, dels quals destaca el sexe i l'edat.

En aquest sentit, ens sembla interessant la perspectiva de gènere que ofereixen. D'una banda, prenen consciència de la importància d'estudiar els trastorns desglossats per sexe, ja que la salut mental s'interconnecta amb desigualtats de gènere, com per exemple els trastorns alimentaris i els cànons de bellesa femenina. D'altra banda, la noció de "gènere" de les investigacions és sempre binària i només una de les persones entrevistades menciona l'existència de gèneres dissidents.

Així, als articles sovint es consideren els termes *sexe* i *gènere* com a equivalents i, fins i tot, en alguns casos, obtenen el gènere d'un usuari mitjançant un sistema que prediu el gènere a partir de la informació bàsica del seu perfil:

> *<<Usamos el género de los usuarios porque empleamos un sistema que detecta directamente el género de los usuarios. Con esta información podemos revisar si hay sesgos de género en el algoritmo.>>* E2: R

Una de les persones entrevistades ens comenta de forma crítica la limitació de l'eina:

> *<<De hecho, ya hay algoritmos que predicen de cierta forma el género pero son limitados porque son binarios y hoy en día puede haber mucha más variedad y entonces se tendría que "re entrenar" este algoritmo para que incluyera más clases.>>* E1: E

Més enllà, al marge de les investigacions, ens sorprenen els comentaris acrítics que apareixen entre algunes persones entrevistades. A tall d'exemple, a continuació adjuntem un discurs d'un investigador amb experiència en la detecció de trastorns alimentaris, depressió i ideació suïcida:

> *<<Por ejemplo, en enfermeras y enfermeros, ¿debería haber paridad de género?. Personalmente yo diría que no, que prefiero que haya más mujeres que hombres, por el tema de la empatía, etc. Entonces no siempre la respuesta social es la que es clara. Esta es mi respuesta personal, pero no sé, si la sociedad se pusiera de acuerdo, no sé cuántos deberían haber como enfermeros... ¿La mitad? ¿Tan solo algunos tipos de enfermeros? No sé…O en salud mental, la gente que cuida a los enfermos mentales, tal vez sea mejor que sean más mujeres, no lo sé >>* E6: B

Convé reiterar que aquesta mena de comentaris, aparentment inofensius i personals, són socialment rellevants perquè són presents en la investigació amb salut mental, però sobretot perquè provenen de persones que dissenyen els mateixos sistemes. Per tant, el risc que repercuteixin aquests imaginaris sobre els/les usuaris/es és alt. A més, cal sumar a aquest punt la dificultat d'identificar un biaix de gènere en sistemes d'informàtica per l'inaccessibilitat als codis, l'educació necessària per poder llegir-los, i, sobretot, pel pressupòsit de neutralitat.

Un exemple que il·lustra aquesta transmissibilitat d'imaginaris és un sistema de predicció de bulímia que presenta biaix de gènere. Concretament, la tecnologia identifica usuaris de la plataforma *Twitter* amb risc de desenvolupar el trastorn alimentari. Després del procés de disseny es troben que, inesperadament, l'algoritme presenta més facilitat per a identificar bulímia entre homes que entre dones. Tal com ens explica la investigadora a càrrec, si ajusten aquest biaix el sistema es torna menys efectiu, menys precís. Per tant, l'equip d'investigació ha de prendre una decisió: un sistema amb menys

biaix de gènere però més eficient o un sistema esbiaixat però més eficient informàticament.

> *<<El algoritmo reprodujo el sesgo del anotador, aunque la mayoría de casos de nuestros datos son de mujeres. Esto es porque las mujeres nos preocupamos más de nuestra apariencia y somos más probables de decir que estamos gordas sin tener bulímia, mientras que si un hombre postea sobre su apariencia es más probable que tenga un trastorno. Entonces es más complejo para el algoritmo ver si una mujer tiene un trastorno o no. >>* *E2: R*

Més endavant, sobre el mateix cas, la investigadora ens comenta:

> *<< Son problemas muy complejos desde la parte informática, para los psicólogos y sociales es más fácil, dicen: "que sea justo y ya" pero para nosotros es muy complejo. (…) Creo que nunca se conseguirá un sistema completamente justo porque siempre hay muchos biases, hay tantos… Siempre hay un tradeoff, por ejemplo, ahora es más justo pero menos accurate o efectivo. Esto siempre pasa. Es muy difícil que mantengas la misma "accuracy" y que tu sistema sea justo.>>*
>
> *E2: R*

En síntesi, es confirma la premissa central de la nostra recerca: conèixer els discursos de ciència computacional ens suposa una petita porta d'accés al món social de la IA, un món que passa per inadvertit pel pressupòsit de neutralitat de la tecnologia. Dit d'una altra forma, la predicció i detecció de salut mental no només emergeix entre uns ideals i imaginaris concrets, sinó que els reprodueix:

> *<<Los algoritmos aprenden de nosotros, reproducen nuestros sesgos >* *E2: R*

*(*III) UNA SALUT MENTAL SENSE SUBJECTE

Tal com hem exposat al llarg de la recopilació documental, la forma d'entendre la salut mental respon ineluctablement a construccions històriques i imaginaris socials. D'aquesta manera, un cop situat des d'on, perquè i de quina manera emergeixen els sistemes de predicció i detecció, en aquesta secció ens proposem perfilar la salut mental des de la IA i, conseqüentment, quin és l'individu resultant.

Així, recapitulant a les distincions centrals entre una noció de salut mental de pacient o subjecte, ens centrarem a respondre si la salut mental es formula com una objectivitat ateòrica així com si es donen esquemes de pensament basats en dicotomies.

Un primer element a conèixer són els referents teòrics presents en la fase primerenca d'investigació de definició dels trastorns que s'adrecen a prediure o detectar. En aquest sentit, ens trobem un escenari considerablement heterogeni amb relació als recursos utilitzats però fortament semblant en funció del caràcter i perspectiva d'aquestes fonts. El cas és que els referents són diferents – des d'un investigador que ens confessa no fer ús de cap manual o referència teòrica fins a persones que opten per l'opció de manuals amb afegits d'experts/es – però predomina una mirada dels trastorns articulada des d'una classificació per símptomes, sobretot la cinquena edició del DCM:

> *<<Bueno, de hecho en varios estudios que hemos hecho hemos hemos basado sobretodo en el DCM-5 para la definición de depresión y también para los indicadores que se pueden establecer mediante la interacción digital. Esto también es interesante porque nosotros hemos descubierto, también, otros indicadores que talvez se dan en un comportamiento digital y que no estan*

escritos en el DCM-5 porque justamente es a través de la terapia, bueno, no la terapia, sino de entrevistar a las personas y entender los signos que muestran y, estos signos son diferentes, o pueden ser diferentes, de los de los mundos digitales que aquellos de la "vida real".>> *E1: E*

Tot i les dissimilituds entre respostes, la rellevància recau en el fet que s'estructuren des d'una concepció de la salut mental basada en símptomes. La hegemonia d'aquesta perspectiva s'explica pel fet que la predicció i detecció depèn de l'existència de patrons que permeten una generalització dels indicadors d'un trastorn:

<<Es decir, le dices mira, este es un ejemplo de un problema mental, este es otro ejemplo y no es un problema mental, y así le das muchos ejemplos y el algoritmo aprende de esos ejemplos. Es decir, trata de encontrar patrones de estos ejemplos para generalizar la respuesta.>> *E6: B*

Segons sembla, els sistemes no només parteixen d'una pretensió d'objectivitat de la salut mental, sinó que el seu funcionament i èxit depèn d'aquesta. Coincidint amb Perrotta (2011), les manifestacions d'aquest enfocament doten d'una aparença de neutralitat als aspectes diferencials dels trastorns. La predicció i detecció es concreta en funció dels símptomes o indicadors, sovint amb relació al processament natural del llenguatge de les publicacions a xarxes socials[12].

Així, s'adopta una lògica dicotòmica, on es forma part del que és patològic o normal, evacuant la subjectivitat i la singularitat de la salut mental:

<<Lo puedes pensar como un circuito electrónico: ya están todos los cables unidos y dependiendo del punto de electricidad que le pongas, el resultado será distinto. Te aparecerán unas luces encendidas y en otras ocasiones no, la conexión binaria es si se enciende o si no se enciende. Por ejemplo, predice si tendrá predisposición al suicidio o no. >> *E6: B*

L'individu esdevé un pacient racional, estable i coherent. La salut mental deixa de ser un conjunt complex indesxifrable:

<<Normalmente hay alguna condición detrás, no es algo inmediato, pero los adolescentes se pueden suicidar por algo muy tonto, como que le deja el novio, algo inmediato, pero normalmente hay un trastorno mental detras del suicidio. ¿Qué trastornos causan más suicidios? Pues estos y por eso nos dedicamos a estos casos.>> *E2: R*

De fet, la salut mental no només és explicable sinó també calculable:

<<For example, sometimes the result would come out to be 80%, 80% depressed.>> *E7: L*

El pacient de la IA és, doncs, l'absència de subjecte, d'un subjecte de carn i ossos, amb desigs i impulsos, inconscients i contradictoris. La salut mental es redueix a

[12] La tècnica predominant per a la detecció o predicció de salut mental és el processament del llenguatge natural (Natural language processing). Tal com ens expliquen, es tracta d'un anàlisi micro del llenguatge de les persones diagnosticades amb un trastorn. Un exemple clar és la freqüència d'ús del pronom 'jo' ('I'). Així, aquestes petites manifestacions del llenguatge esdevenen indicadors de salut mental:

<< Yo he trabajado con textos en inglés. En español quizás la historia seria diferente, digamos que se podrían ver otros indicadores. Pero sí, este uso del "I", del pronombre "yo", es característico de la depresión, el hecho de autoreferenciarse con mayor frecuencia. >> *E1: E*

comportament objectivable perquè l'aprenentatge automàtic coincideix i depèn de la pretensió d'objectivitat 'ateòrica' de la salut mental. A més, aquesta posició es reforça pel fet que, com hem vist anteriorment, la IA no és un mitjà sinó un objectiu per si mateixa.

VII. CONCLUSIONS

El present treball ofereix una perspectiva social de la producció de sistemes de IA que prediuen i detecten la salut mental. Concretament, hem analitzat des d'una aproximació qualitativa els discursos sobre la noció de salut mental dels/les investigadors/es de ciència computacional que participen en el desenvolupament de serveis de IA per a la predicció/detecció de depressió, ansietat, anorèxia i ideació suïcida.

En aquest sentit, els resultats evidencien una analogia d'experiències i percepcions del context social de producció, la valoració d'investigacions interdisciplinàries així com la noció de subjecte i salut mental.

D'entrada, trobem que la formulació dels objectius i valoració de la IA és independent a la salut mental, encara quan s'adreça a la predicció i diagnosis. En efecte, la IA es presenta com una finalitat per si mateixa, suposant una ruptura amb el pressupòsit que la tecnologia és un mitjà i no una finalitat (Heidegger a Hanks, 2010). Aquesta distància envers el camp d'anàlisis es manifesta tant en el disseny de les investigacions així com al llarg de l'experiència dels/les investigadors/es, suggerint una relació instrumental envers la salut mental.

Alhora, no podem afirmar l'existència d'equips interdisciplinaris sinó equips de ciència computacional amb col·laboracions secundàries i puntuals amb experts/es de salut mental. No només és la incorporació d'un/a psicòleg/ga o psiquiatre/a en el procés de disseny poc habitual sinó que, quan participen, la decisió sovint sembla arbitrària i el seu rol secundari.

D'altra banda, obtenim evidència de transmissibilitat d'imaginaris de gènere als sistemes, corroborant que la predicció i detecció de salut mental no només emergeix entre uns ideals i imaginaris concrets, sinó que els pot arribar a reproduir.

Finalment, atenint-nos al nostre objectiu principal, trobem que la noció de salut mental en la IA s'articula des d'una lògica dicotòmica, basada en símptomes, deixant al marge la subjectivitat i singularitat dels trastorns. D'aquesta manera, la salut mental deixa de ser un conjunt indesxifrable. De fet, no només esdevé explicable sinó que es pot predir i calcular.

En definitiva, el pacient de la IA és racional, estable i coherent. La salut mental es redueix al comportament objectivable perquè l'aprenentatge automàtic coincideix i depèn de la pretensió d'objectivitat 'ateòrica' de la salut mental. A més, aquesta posició es reforça pel fet que, com hem vist anteriorment, la IA no és un mitjà sinó un propòsit per si mateixa.

Com la noció impacta la mateixa salut mental (Martínez, 2012), esperem que els resultats d'aquesta investigació generin reflexió sobre les limitacions, oportunitats i riscos de l'aplicació d'aquests sistemes. Certament, caldrà explorar noves línies de recerca sobre la recepció dels sistemes, així com continuar explorant la dimensió social de la IA, ja que, com hem identificat a partir de la revisió sistemàtica de l'estat de l'art, es tracta d'una àrea d'investigació amb un fort buit empíric.

BIBLIOGRAFIA

ANESHENSEL, S., PHELAN, C. 1999. Handbook of the sociology of mental health. New York: Kluwer Academic / Plenum.

BEN-DAVID, J., SULLIVAN, T. 1975. Sociology of Science. Annual Reviews.

BROWN, T. 2003. Critical race theory speaks to sociology of mental health: Mental health problems produced by racial stratification. American Sociological Association: Journal of Health and Social Behaviour, XLIV (III).

CHANCELLOR, S. 2018. Computational methods to understand deviant mental wellness communities. Montreal: CHI'18 Extended Abstracts.

CHEN, X., SYKORA, M., JACKSON, T., ELAYAN, S. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. Proceedings of the companion the web conference.

DOMÈNECH, M., TIRADO, J. 1998. Sociología simétrica. Barcelona: Gedisa Editorial.

EDGCOMB, J., ZIMA, B. 2019. Machine learning, natural language processing, and the electronic health record: Innovations in mental health services research. American Psychiatric Publishing: Psychiatric services, LXX (IIII).

EUGENIA MONTERTO, M., ELORRIAGA, K., ELENA LUGO, M. 2012. Nociones acerca de la complejidad y algunas contribuciones al proceso educativo. Telos.

FERNÁNDEZ RIVAS, L., PIMENTEL, M. 1996. El sujeto de la salut mental a fin de siglo. México: UAMX.

FONSEKA, T., BHAT, V., KENNEDY, S. 2019. The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviours. SAGE, I (II).

GAUR, M., KURSUNCU, U., ALAMBO, A., SHETH, A., DANIULAITYTE, R., THIRUNARAYAN, K., PATHAK, J. 2018. "Let me tell you about your mental health!": Contextualized classification of reddit posts to DSM-5 for web-based intervention. ACM.

GRAHAM, S., DEPP, C., LEE, E., NEBEKER, C., TU, X., KIM, H., JESTE, D. 2019. Artificial intelligence of mental health and mental illness: An overview. Springer Science and Business Media.

HANKS, C. 2010. Technology and values: Essential Findings. Sussex: Blackwell.

HIRSCHBERG, J., MANNING, CD. 2015. Advances in natural language processing. Sci Mag 345 (6245).

HOLTON, R., BOYD., R. 2019. 'Where are the people? What are they doing? Why are they doing it' (Mindell) Situating artificial intelligence within a socio-technical framework. SAGE, I (XVII).

MARTÍNEZ, L., CASTRO, X. (2012). Representaciones sociales de los médicos y el personal paramédico respecto a los trastornos de la conducta alimentaria. Tesis de pregrado en Psicología, Universidad Icesi.

MORCH, C., GUPTA, A., MICHARA, B. 2020. Canada protocol: An ethical checklist for the use of artificial Intelligence in suicide prevention and mental health. Montréal: Elsevier.

MORESCHI, A. 2013. La subjetividad a debate. Sociológica (80).

MOWERY, D., PARK, A., CONWAT, M., BRYAN, C. 2016. Towards automatically classifying depressive symptoms from twitter data for population health. Proc Work Comput Model People's Opin Personal Emot Soc Media.

MLYNAR, J., ALAVI, H., VERMA, H., CANTONI, L. 2018. Towards a Sociological Conception of Artificial Intelligence. Prague: Artificial General Intelligence.

MUNTANYOLA, D. 2014. Metodología(s) Perspectivas, prácticas y desafíos. Salamanca: Asociación Contubernio.

NUBIOLA, J. 1998. Walker Percy y Charles S. Peirce: Abducción y lenguaje. Navarra: Analogía Filosófica, XII (I).

PERROTA, G. 2008. Nociones de sujeto: apuntes para el análisis de la concepción de sujeto/paciente para los profesionales de la salud en el abordaje de la sexualidad y la salud reproductiva. Buenos Aires: Memorias de XV Jornadas de Investigación y Cuarto Encuentro de Investigadores del MERCOSUR

PERROTA, G. 2011. Concepciones de sujeto, cuerpo y síntoma en medicina y psicoanálisis. Buenos Aires: III Congreso Internacional de Investigación y Práctica Profesional en Psicología XVIII.

PESTIAN, J., MATYKIEWICZ, P., GRUPP-PHELAN, J. 2008. Using natural language processing to classify suicide notes. ACM.

RAMIREZ-CIFUENTES, D., FREIRE, A., BAEZA-YATES, R., SANZ LAMORA, N., ÁLVAREZ, A., GONZÁLEZ-RODRÍGUEZ, A., ROCHEL, M., LLOBET, R., VELAZQUEZ, D., GONFAUS, J., GONZÀLEZ, J. 2021. Characterization of Anorexia Nervosa on social media: textual, visual, relational, behavioral, and demographical analysis. JMED, XXIII (X).

RIBA, C. Generalitats sobre els mètodes qualitatius: trets bàsics, variants, camps d'ampliació i història. Catalunya: UOC.

SACKS, H. 1992. Lectures on Conversation I-II. Oxford: Blackwell.

SAU, A., ISHITA, B. 2017. Artificial neural network (ANN) model to predict depression among geriatric population at a slum in Kolkata, India. Journal of clinical and diagnostic research, XI (V).

SHATTE, A., HUTCHINSON, D., TEAGUE, S. 2019. Machine learning in mental health: a scoping view of methods and applications. Cambrigde: Psychological Medicine, XLIX (IX).

SRIVIDJA, M., MOHANAVALLI, S., BHALAJI, N. 2018. Behavioral modeling for mental health using machine learning algorithms. Journal of medical systems, XLII (V).

TATE, A., McCABE, R., LARSSON, H., LUNDSTRAM, S., LICHTENSTEIN, P., KUJA-HALKOLA, R., MUMTAZ, W. 2020. Predicting mental health problems in adolescence using machine learning techniques. Public library of science, PloS ONE, XV (IIII).

THIEME, A., BELGRAVE, D., DOHERTY, G. 2020. Machine learning in mental health: A systematic review of HCI literature to support the development of effective and implementable ML systems. ACM Trans, XXVII (V).

ANNEX 1. CRITERIS D'ELEGIBILITAT

D'entrada, l'estratègia de cerca de publicacions ha estat per selecció de base de dades i sobretot per bola de neu. Atès que el mètode abductiu possibilita realitzar el treball de camp paral·lelament amb el procés de recopilació documental, un gruix significatiu de les publicacions revisades han estat obtingudes durant les entrevistes.

Amb referència amb la selecció de base de dades, la llibreria digital utilitzada ha estat *Library Genesis*, una plataforma d'articles acadèmics, tant d'accés gratuït com amb cost associat. Els motius que han portat a escollir aquesta plataforma han estat les presents dificultats amb la base de dades universitària (relacionades amb l'atac informàtic) així com la varietat d'opcions de cribratge que facilita *Library Genesis*.

El següent punt a considerar són els criteris que han guiat la selecció en base de dades dels estudis empírics:

(i) En primer lloc, per tal d'aconseguir una mostra que sigui significativa de l'estat de la qüestió actual, la recerca empírica s'ajusta al període 2018-2022 (ambdós anys inclosos). En cas de dues investigacions fortament similars, s'ha prioritzat el més recent. No obstant això, al llarg del document podem trobar un referències que no compleixen aquest criteri i és important explicar que això és degut al fet que corresponen a la bibliografia teòrica (no empírica) i tan sols serveixen per millorar la comprensió i estructura del document.

(ii) Per a la selecció de base de dades les paraules clau han estat: *artificial intelligence mental health, machine learning mental health, sociology science mental health* i *sociology mental health*.

(iii) Seguidament, s'han cercat els termes clau en l'anglès. Això és degut a la predominança d'aquesta llengua envers les publicacions d'aquest camp. A tall d'exemple, cercant <<Artificial intelligence mental health>> a *Library Genesis* obtenim quaranta-vuit resultats, però si cerquem el mateix en castellà (<<Inteligencia Artificial Salud Mental>>) obtenim només un resultat.

(iiii) El darrer criteri d'elegibilitat correspon a l'alineament i adequació de l'estudi amb la perspectiva i objecte d'anàlisis del nostre treball. Es tracta d'una selecció en funció del títol i resum on es comprova si és un estudi veritablement empíric, des de quina disciplina es du a terme, si presenta informació sobre la detecció o predicció de la salut mental amb aprenentatge automàtic i la seva rellevància envers les altres investigacions prèviament seleccionades.

Recollint tot el que s'ha dit, a continuació s'il·lustra el procés de cribratge per a les publicacions seleccionades en base de dades mitjançant una taula:

	CRIBRATGE

Paraula clau	Total documents	Eliminació duplicats	Anys abarcats (2018-2022)	Atenent al títol i resum	Resultats finals
Artificial intelligence mental health	N=38	N=24	N=19	N=3	N=2
Machine learning mental health	N= 48	N=29	N=20	N=5	N=5
Sociology science mental health	N= 33	N= 11	N= 1	N= 1	N=1
Sociology mental health	N= 501	N= 422	N= 33	N= 2	N=2
Total publicacions per selecció en base de dades					*9*

ANNEX 2. LLISTAT CODIS

TEMA	DIMENSIÓ	CODIS
Salut mental	1Noció Salut mental	101 Valorització SM 102 Definició SM 103 Termes SM
	2Relació investigador/a - SM	201Trajectòria acadèmica 202 Motivacions 203 Contacte amb SM
Predicció / diagnosis amb IA	3Contextualització	301 Justificació 302 Condicions analitzades 302 Projecció futura
	4Relació investigador/a - IA	401 Valorització IA 402 Raonament crític 403 Raonament acrític 404 Neutralitat
	5Resultats	501Achievements 502Feedback
Procés creació IA	6Experts	601 Rol experts 602 Decisió involucrar experts 603 Com valoren als experts
	7Errors	701 Errors 702 Tradeoff 703 Biaix
	8Dades	801Tipologia dades 802 Selecció dades (fonts)

Dimensió social salut mental	11Gènere	111 Noció binària del gènere 112 Imaginaris gènere
	12Raça	121 Imaginaris raça
	13Classe	131 Imaginaris classe
	14Ciències socials	141 Valorització ciències socials en SM 142 Valorització variables gènere, raça, classe IA 143 Presència variables socials