cejfe

Administració de justícia

Intel·ligència artificial i administració de justícia

Perspectives d'implantació i problemàtiques jurídiques i pràctiques

Ajut a la investigació 2021

Guillem Soler Solé

Any 2022

Generalitat de Catalunya
Centre d'Estudis Jurídics
i Formació Especialitzada

El Centre d'Estudis Jurídics i Formació Especialitzada ha editat aquesta recerca respectant el text original dels autors, que en són responsables de la correcció lingüística.

Les idees i opinions expressades en la recerca són de responsabilitat exclusiva dels autors, i no s'identifiquen necessàriament amb les del Centre d'Estudis Jurídics i Formació Especialitzada.

Avís legal

Resum

L'objectiu de la recerca és l'abordatge, necessàriament hipotètic, de quines són les perspectives d'implantació i les problemàtiques jurídiques i pràctiques que tindria una eventual inserció d'eines d'intel·ligència artificial en l'àmbit de l'Administració de Justícia. Es tracta d'una tecnologia que s'està desenvolupant significativament en molts àmbits (inclosa l'administració pública en general), però no, de moment, en l'Administració de Justícia. Atesa l'evident sensibilitat del camp judicial en termes de drets potencialment afectats, la investigació se centra a identificar quines són les modalitats d'intel·ligència artificial existents, quines podrien ser emprades a l'Administració de Justícia i quines problemàtiques generarien en termes d'independència judicial, de dret defensa o d'eventuals discriminacions. Per abordar aquesta investigació s'han analitzat quines són les concretes aplicacions d'eines d'intel·ligència artificial que ja estan, en altres països, com a mínim projectades. La conclusió principal seria aquesta: si es descarten els extrems d'un optimisme o un escepticisme tecnològics ingenus o injustificats, sembla factible, en principi, plantejar la viabilitat d'inserir certes eines d'automatització a determinades fases del procediment judicial. No necessàriament a la decisió final, a la sentència, sinó també, i principalment, a fases de tràmit molt voluminoses i repetitives. Sempre, és clar, amb un posterior control humà. L'objectiu de la investigació no deixa de ser, però, merament exploratori. Així ho imposa una tecnologia que canvia a gran velocitat.

Resumen

El objetivo de la investigación es el abordaje, necesariamente hipotético, de cuáles son las perspectivas de implantación y las problemáticas jurídicas y prácticas que tendría una eventual inserción de herramientas de inteligencia artificial en el ámbito de la Administración de Justicia. Se trata de una tecnología que ha vivido un desarrollo significativo en muchos ámbitos (incluida la administración pública en general), pero no, de momento, en la Administración de Justicia. Dada la evidente sensibilidad del campo judicial en términos de derechos potencialmente afectados, la investigación se centra en identificar cuáles son las modalidades de inteligencia artificial existentes, cuáles podrían ser empleadas en la Administración de Justicia y qué problemáticas generarían en términos de independencia judicial, de derecho de defensa o de eventuales discriminaciones. Para abordar esta investigación se han analizado cuáles son las concretas herramientas de inteligencia

artificial que ya están como mínimo proyectadas en otros países. La conclusión principal sería esta: si se descartan los extremos de un optimismo o un escepticismo tecnológicos ingenuos o injustificados, parece factible, en principio, plantear la viabilidad de insertar ciertas herramientas de automatización a determinadas fases del procedimiento judicial. No necesariamente a la decisión final, a la sentencia, sino también, y principalmente, a fases de trámite muy voluminosas y repetitivas. Siempre, claro está, con un posterior control humano. El objetivo de la investigación no deja de ser, de todos modos, meramente exploratorio. Así lo impone una tecnología que cambia a gran velocidad.

Summary

The goal of this research is to address, of course, hypothetically, which are the implementation perspectives and the legal and practical issues that an eventual insertion of artificial intelligence tools in the Judiciary would have and arise. It is a technology that has undergone significant development in many areas (including public administration in general), but not, for the moment, in the Judiciary. Given the evident sensitivity of the judicial field in terms of potentially affected rights, the research focuses on identifying which are the existing artificial intelligence modalities, which could be used in the Judiciary and what problems they would generate in terms of judicial independence, right of defense or eventual discriminations. To address this research, the specific applications of artificial intelligence tools that are at least already projected in other countries have been analysed. The main conclusion would be the following: if the extremes of naive or unjustified technological optimism or scepticism are ruled out, it seems feasible, in principle, to insert certain automation tools at specific stages of the judicial procedure; not necessarily to the final decision, to the ruling, but also, and mainly, to other voluminous and repetitive initial or intermediate procedural phases.– always, of course, with subsequent human control. The purpose of the research is, however, merely exploratory, due to the ever changing nature of this technology.

Descriptors

administració de justícia, algoritme, aprenentatge automatitzat, aprenentatge no supervisat, aprenentatge profund, aprenentatge supervisat, automatització, biaix algorítmic, biaix d'automatització, caixes negres, contestabilitat algorítmica, digitalització judicial, discriminació, dret de defensa, explicabilitat algorítmica, IA, intel·ligència artificial, justícia

predictiva, jutge-robot, processament del llenguatge natural, prova electrònica, resolució en línia de litigis, sistemes experts, transparència algorítmica.

Descriptores

administración de justicia, algoritmo, aprendizaje automatizado, aprendizaje no supervisado

aprendizaje profundo, aprendizaje supervisado, automatización, cajas negras, contestabilidad algorítmica, derecho de defensa, digitalización judicial, discriminación, explicabilidad algorítmica, IA, inteligencia artificial, juez-robot, justicia predictiva, procesamiento del lenguaje natural, prueba electrónica, resolución en línea de litigios, sesgo de automatización, sesgos algorítmicos, sistemas expertos, transparencia algorítmica.

Descriptors

AI, Algorithm, Algorithmic Bias, Algorithmic Explainability, Algorithmic Transparency, Artificial Intelligence, Automated Bias, Automation, Big Data, Black Boxes, Contestability, Deep Learning, E-discovery, Expert Systems, Judicial digitization, Judiciary, LegalTech Machine Learning, Natural Language processing, NLP, ODR, Online Dispute Resolution, Optical Character Recognition, Predictive Justice, Right of Defense, Robot-judge, Supervised Learning, Unsupervised Learning.

Índex de contingut

1. Premisses de la recerca

1.1. Pols de la recerca: administració de justícia i intel·ligència artificial

L'objecte d'aquesta recerca gira al voltant de dos pols: l'*administració de justícia* i la *intel·ligència artificial* (en endavant, IA). El primer ens és molt familiar: les societats humanes, des del seu mateix naixement, han hagut de resoldre els conflictes socials que ineludiblement generen i ho han fet de maneres molt diverses, fins a arribar a l'actual sistema, que ens sembla el millor o més acabat i que acostumem a definir com estat de dret en la seva versió constitucionalitzada. En qualsevol cas, tots els sistemes utilitzats fins el moment poden incloure's, en un sentit ampli, dins de la gran tasca *humana* d'administrar justícia. Per això la coneixem (o creiem conèixer-la) tan bé. I una de les notes principals d'aquesta tasca és, o ha estat fins recentment, el fet de ser duta a terme de manera íntegra per persones, sigui per un pretor, un senyor feudal, un rei omnipotent, un delegat d'aquest rei o un tribunal més semblant als actuals.

Per contra, el segon pol de la recerca, la IA, ens és molt més desconegut i ens remet, ja d'entrada, a una idea que sembla contradir la concepció tradicional de la tasca d'administrar justícia: aquesta podria deixar de ser un acte íntegrament humà. Podria introduir-s'hi, en algun punt o fase, una certa automatització implementada per un programari (*software*) i dirigida per un algoritme (una successió d'ordres codificades en el programari). Lògicament, l'eina que operaria aquesta eventual automatització seria, ella mateixa, una creació humana, un artefacte tecnològic creat amb intervenció humana. La novetat residiria més aviat, per tant, en el tipus i grau de control humà que es prevegi sobre l'ús d'aquestes eines. Com veurem, no es tracta, en absolut, d'haver d'escollir, binàriament, entre un control humà absolut o una automatització completa. Les eleccions seran més complexes i matisades. Però en tot cas podran tenir implicacions que contradiguin, en part, la concepció *tradicional* (íntegrament humana) de l'acte d'administrar justícia. Vegem-ne algunes:

a) En primer lloc, en la creació d'eines d'IA judicial hi intervindran, com a mínim parcialment, persones no expertes en dret, com ara tècnics informàtics o experts en computació, entre altres. Ja ho fan, de fet, a dia d'avui, en la creació d'eines digitals com

els sistemes de gestió processal, però es tracta, com veurem, d'eines *accessòries*, *externes* o *passives*, que no incideixen *materialment* en l'acte d'administrar justícia. Per contra, les eines d'IA sí que podrien tenir, en algun grau, aquesta incidència, per la qual cosa el fet que les confeccionin, ni que sigui en part, experts en matèries no jurídiques planteja ineludiblement certs interrogants.

b) En segon lloc, la *creació* de l'eina d'IA judicial seria, per definició, *anterior* a l'acte d'administrar justícia, mentre que aquest consistiria, en tot o en part, en *executar* de manera *automatitzada* aquesta eina. Per tant, la *creació* de contingut jurídic inherent a l'acte d'administrar justícia ja no seria íntegrament humana, sinó, totalment o en part, automatitzada (deshumanitzada?). Tot plegat sense perjudici, evidentment, d'una convenient (i necessària) intervenció humana posterior, de confirmació, modificació o rebuig.

c) En tercer lloc, mentre que unes modalitats d'IA, abans de la seva implantació real sobre el terreny, *entrenen* i ajusten el seu algoritme i en *proven* la seva eficàcia amb dades prèviament *etiquetades* per éssers humans amb la *resposta correcta*, d'altres aborden aquesta tasca d'entrenament i prova amb dades no etiquetades: busquen la resposta correcta per sí mateixes fins i tot en la fase prèvia d'ajustament i prova de l'algoritme. Es tracta, respectivament, de la IA *supervisada* i la *no supervisada*. Com és lògic, els temors o dubtes que pot generar la seva implantació a l'Administració de Justícia són majors en el segon cas.

d) En quart lloc, certes modalitats avançades d'IA (les anomenades d'*aprenentatge automatitzat* o *machine learning*) tenen una fase de creació estesa en el temps, en el sentit que al primer moment de creació *humana* de l'eina d'IA (construcció del model i ajustament de l'algoritme) li segueix un segon moment, en principi sense fi, de *modificació* i *actualització automatitzada continuada*. És a dir, l'eina anirà aprenent (s'anirà *reajustant*), per si mateixa, sense intervenció humana, a partir dels resultats dels seus usos anteriors, per tal d'obtenir, en principi, la major precisió possible.

De nou, hi pot haver (hi hauria d'haver, sense dubte, en el cas de l'Administració de Justícia) una supervisió humana de l'actualització automàtica, però el sol fet que aquesta última es pugui produir més enllà de la fase de *creació controlada* genera desafiaments jurídics afegits.

e) En cinquè i últim lloc, les modalitats més avançades d'IA (les anomenades d'*aprenentatge profund* o *deep learning*) es caracteritzen no només pel seu caràcter automatitzat (propi de qualsevol eina d'IA) i d'aprenentatge continu autònom (propi de les eines de l'apartat anterior, les de *machine learning*), sinó pel fet d'operar amb unes estructures tan complexes que fan que sigui pràcticament impossible (o directament impossible) obtenir una explicació del per què han donat una determinada resposta o resultat. Són més que evidents, en aquest cas, les problemàtiques que aquesta limitació informativa pot gener en l'àmbit de l'Administració de Justícia.

Veiem, en definitiva, que la IA presenta com a mínim cinc notes que semblen desafiar la noció tradicional que podríem consensuar sobre què significa administrar justícia. Ja tenim detectades, havent tot just començat, cinc possibles tensions entre els dos pols de la recerca. Caldrà analitzar si es tracta de tensions insuperables o si es pot trobar algun tipus d'equilibri raonable. Per fer-ho en condicions, s'explica a continuació fins a quin punt és imprescindible una prèvia tasca d'acotament de l'objecte d'estudi.

1.2. Imprescindible acotament de l'objecte de la recerca

Més endavant s'abordarà la tasca (ja de per sí complexa) d'intentar definir què és exactament la IA. El que cal destacar, però, des d'un inici, és que es tracta d'una creació tecnològica humana d'un camp d'aplicació potencial extraordinàriament ampli. De fet, són poques les activitats humanes respecte de les quals no s'apliqui, ja, o no es pretengui aplicar, en un futur més o menys proper, alguna eina o recurs tecnològic que pugui ser qualificat d'IA.

Si es fa una cerca per la xarxa sobre la matèria, veurem que pràcticament no se n'escapa cap àmbit. I a algun d'ells hi acostumem a dedicar força hores de la nostra vida: les xarxes socials (*Facebook*, *Twitter*, etc.), els sistemes comercials de recomanació de productes audiovisuals (*Netflix*, *HBO*, *Prime Video*, etc.), la *Internet de les coses* (electrodomèstics, com una nevera, equipats amb sensors que transmeten informació en temps real connectada a altres serveis), la salut (diagnòstic, agilització del triatge en urgències, càpsules endoscòpiques, assistència a persones grans, etc.), l'educació (selecció dels grups escolars, reconeixement facial abans d'entrar a l'escola, correcció d'exàmens, etc.), l'àmbit laboral (selecció de personal segons els gestos facials, predicció de les baixes laborals, etc.), els vehicles de conducció automatitzada, la

robòtica, les ciutats intel·ligents (*smart cities*), els sistemes automatitzats d'atorgament d'ajudes públiques, la ciberseguretat, la racionalització dels fluxos de trànsit, la predicció policial en el control de fronteres, el combat contra el blanqueig de diners i el frau fiscal, els filtres de correu brossa (*spam*), les aplicacions de seguiment de la població per controlar els brots epidèmics o, fins i tot, la creació artística (el 2018 es va vendre a *Christie's* per 432.000$ un quadre, *Portrait of Edmond de Belamy*, creat amb IA[1]). La llista podria continuar gairebé indefinidament.

Per tant, el primer risc que ha d'afrontar la recerca és el del seu possible *desbordament* i *dispersió*. Caldrà disposar estratègicament els dics i acotaments que siguin necessaris. La millor manera de fer-ho serà descartant, lògicament, l'estudi de les aplicacions particulars d'IA a camps aliens a l'Administració de Justícia. Cal advertir, però, que a vegades les fronteres no seran tan clares, ja que aplicacions concretes pensades per a un àmbit en principi aliè a la justícia podrien tenir, això no obstant, una possible aplicació judicial[2].

La segona mesura per evitar el desbordament de la recerca consistirà en distingir, de manera precisa, les *qüestions generals* que planteja la IA, potencialment rellevants per qualsevol aplicació concreta que se'n vulgui fer, de les *qüestions particulars* específicament d'interès pel camp concret d'estudi, el de l'Administració de Justícia. Les segones seran el nucli ineludible, i obvi, de la recerca. Caldrà buscar respostes a preguntes com les següents: com es pot veure afectat el dret a un *judici just* quan s'hi aplica, en algun moment del procediment (no necessàriament en la decisió final), un recurs d'IA? Per exemple, per valorar el risc de reincidència delictiva abans d'acordar, o no, una llibertat provisional o un permís penitenciari? Quina informació es donarà a les parts? Serà suficient perquè puguin recórrer en condicions la decisió judicial? Serà possible endinsar-se fins l'interior de l'algoritme o de les dades que l'alimenten per esbrinar si genera algun tipus de resultat discriminatori? Es veuran afectats els principis d'independència i d'imparcialitat judicials, tenint en compte que la decisió judicial ja no

[1] https://www.bbc.com/news/technology-45980863.

[2] Per exemple, en salut, els sistemes automatitzats de detecció de l'exageració del dolor per part del pacient (un programari analitza l'historial de proves i diagnòstics) podrien ser útils en el marc d'un procés judicial en el qual s'hi reclami una indemnització per lesions personals derivades d'un accident de circulació, especialment quan la lesió sigui de les que no acostumen a deixar un rastre objectivat en proves mèdiques (ressonàncies magnètiques, radiografies, TAC, etc.) i es basi principalment en les referències subjectives que doni el mateix pacient, ara actor en un procediment judicial.

serà plenament *humana*? Augmentaran o disminuiran les condicions d'*accés a la justícia*?

Al mateix temps, però, abans d'abordar aquests aspectes de la IA estretament vinculats a l'Administració de Justícia (i que clarament han de formar part de l'objecte de la recerca), caldrà analitzar un conjunt de *qüestions generals* que planteja la IA, com les següents: ètica i IA, principis de la IA, marc de governança, projectes de regulació legal, *Big Data*, protecció de dades i privacitat, restriccions per propietat intel·lectual o industrial, etc. Sense aquest abordatge no disposaríem de l'imprescindible marc teòric i pràctic sobre el qual s'ha d'edificar la recerca. Serà en aquest apartat general on toparem amb el segon gran risc de desbordament, ja que algunes qüestions generals, que poden requerir pel seu tractament adequat un estudi de certa profunditat i detall, tenen un interès només relatiu per a les aplicacions potencials de la IA a l'Administració de Justícia. Per tant, es tendirà a tractar i abordar amb una certa profunditat només aquelles qüestions generals que presentin una relació suficientment intensa amb l'objectiu pràctic final de la recerca.

1.3. Ni un optimisme tecnològic ingenu, ni un escepticisme excessiu

S'intentarà en tot moment adoptar en la recerca una posició el més asèptica possible, que no parteixi, d'una manera esbiaixada i irreflexiva, de la necessitat de trobar, com sigui, eines d'IA potencialment aplicables a l'Administració de Justícia. Es buscarà la màxima equidistància possible entre dos pols. En un hi hauria una espècie de *partidisme tecnològic cec,* que ens porti a acceptar ingènuament i acrítica qualsevol avenç tecnològic pel sol fet de ser el més *avançat*, de ser qualificat d'intel·ligent o *smart* o de semblar que pot aportar una major eficiència a l'Administració de Justícia. A l'altre hi trobaríem un excessiu *escepticisme tecnològic* que, ancorat en concepcions tradicionals que no han demostrat ser infal·libles ni especialment eficaces, tanqui les portes de manera sistemàtica i injustificada, abans d'estudiar-les com mereixen, possibles aplicacions tecnològiques a l'Administració de Justícia basades en la IA. On es troba el punt exacte de l'equilibri és difícil de precisar, però sí que podem tenir clars els extrems als quals no ens volem apropar.

Si després d'una dura tasca d'investigació s'arriba a la conclusió que, des de la perspectiva de l'Administració de Justícia, no és viable tècnicament o, en un altre cas,

no és convenient per motius de fons cap eina d'IA i que, per contra, potser el que cal és un cert aprofundiment en les eines de mera digitalització ja existents, això no implicarà, en absolut, el fracàs de la recerca: per la seva mateixa naturalesa i objecte, es tracta de l'*exploració* d'un terreny molt poc conegut. Del que es tracta és de localitzar possibles eines de millora del funcionament del sistema a la llum d'una tecnologia innovadora. Si no se'n troba cap, aquest serà el resultat de la recerca.

És obvi que caldrà anar amb molta cura, tenint en compte la naturalesa de l'àmbit afectat (*Justícia*) i la possible vulneració dels drets fonamentals que s'hi veuen implicats. Al mateix temps, però, es prioritzarà fixar l'atenció en les tasques judicials més estandarditzades i senzilles en les quals l'eventual afectació de drets sigui potencialment menor. A més, són aquestes les que, si es simplifiquen (si s'automatitzen), podran generar un major estalvi de temps i una millor optimització dels recursos, per poder destinar-los a altres esferes judicials.

No hauríem d'oblidar, tampoc, els baixos nivells de confiança que té la ciutadania en l'Administració de Justícia tal com funciona avui dia. Per tant, l'exploració de vies de millora és més necessària que mai. O tan necessària com sempre. És cert que un dels factors que poden explicar aquest dèficit de confiança pot ser el baix nivell d'inversió pública en justícia. Però fins i tot en cas de ser això així, podem predir que aquests nivells baixos d'inversió continuaran en el temps (no acostumen a ser una prioritat política), per la qual cosa segueixen sent convenients exploracions com la present.

Podríem preguntar-nos, en aquest punt, quina relació té la implantació d'eines d'IA judicial i una possible millora de la confiança ciutadana en l'administració de justícia. Perquè podria semblar, per contra, que aquesta implantació generaria una llunyania i una incomprensió encara majors. Això no obstant, la IA pot ser una oportunitat de millora, per exemple, en l'accés dels ciutadans a la justícia a través de sistemes de resolució en línia de disputes (*ODR*) que estiguin directament vinculats, en cas de no arribar-se a tancar un acord, als tribunals de justícia. I aquests *ODR* podrien utilitzar certes eines d'IA per facilitar o *fomentar* aquests acords, com veurem.

1.4. Una exploració empírica i casuística de futur incert

D'aquesta recerca, més que les conclusions, el que més pot interessar és el camí seguit per conèixer unes eines noves que ja s'apliquen a altres àmbits i que necessitem esbrinar si poden o no ser d'aplicació, també, a l'Administració de Justícia. En iniciar la recerca, desconeixem quin en serà el resultat. I, a més, desconeixem si les conclusions a les quals es pugui arribar es concretaran en la realitat, o no, en un futur més o menys proper. El futur és per definició incert i més encara si la projecció l'hem de fer en un camp, el de la IA, en el qual el progrés tecnològic és constant i l'estat de coses pot canviar substancialment en qüestió de pocs anys, si no de mesos. En definitiva, del que es tracta, a més de poder arribar a una sèrie de conclusions o propostes, és de conèixer quins són, dins de les possibilitats que ofereix amb caràcter general la IA, els factors i condicionants rellevants per plantejar-se la seva eventual implementació a l'Administració de Justícia. I, atesa la gran varietat que existeix de tipus diferents d'IA, aquesta exploració s'ha de fer *sobre el terreny* i seguint un mètode en part *empíric* i *casuístic*: caldrà partir de certes categories generals per orientar-nos, però l'estudi haurà de ser particularitzat, referit a concretes eines d'IA i a concretes tasques judicials.

Podríem dir, fent una analogia, que la recerca actuarà, ella mateixa, com una eina d'aprenentatge que explora un terreny que li és nou i que, en el mateix procés de conèixer-lo, va millorant la seva manera de desplaçar-s'hi. L'*algoritme* que la guiarà serien les premisses que ara estem fixant i el resultat final, totalment desconegut.

1.5. Desmitificant la IA

Una bona manera d'iniciar l'exploració és una dosi controlada de *desmitificació*, aplicada tant a la nova eina que potser serà implantada (la IA) com a la funció ancestral en la qual serà, potser, inserida (la funció jurisdiccional). Sempre són saludables les teràpies de desmitificació. Ajuden a clarificar i desbrossar el terreny pel qual hem de circular.

Començant per la desmitificació de la IA, cal dir que no es tracta, en absolut, d'un camp nou o recent. S'acostuma a fixar els orígens de la IA a l'any 50, quan Alan Turing, després de preguntar-se si una màquina és capaç de pensar, va idear el *joc de la imitació,* que consistia que un interrogador mantingués una comunicació basada en text simultàniament amb un humà i una màquina que imités al primer i que l'interrogador no

els pogués diferenciar (Kirkpatrick i Klingner, 2004). Més endavant hi ha hagut diversos progressos (amb fases d'eufòria posteriorment no confirmada) i estancaments. Des d'una perspectiva més mediàtica, certes eines d'IA han guanyat al campió mundial d'escacs[3], al programa *Jeopardy*[4] o al *Go*[5]. El que és innegable, en tot cas, és que no es tracta d'una realitat recent.

El que sí que és més recent és la disponibilitat de grans quantitats de dades: 2,5 quintilions de bits diaris (Bartoletti, 2020). Hauríem passat, en termes absoluts, de 33 zetabytes el 2018 (un zetabyte són mil milions de terabits i un terabit són mil gigabits, els *GB* que ens són més coneguts) a una previsió de 175 zetabytes pel 2025[6].

També és una novetat el desenvolupament, degut en part, precisament, a la disponibilitat de grans quantitats de dades, de tipologies d'IA més avançades, com les d'*aprenentatge automatitzat* (aprenen i es modifiquen per si mateixes) o, especialment, les d'*aprenentatge profund* (són tan complexes que és difícil, fins i tot pels seus mateixos creadors, arribar a saber com funcionen o per què generen un determinat resultat).

En qualsevol cas, però, com ens recorda Chomsky[7], no ens trobem, estrictament, davant d'una ciència amb la qual s'intenti *comprendre* el significat de les coses o respondre certes qüestions teòriques. Es tracta, més aviat (de moment), d'una "*enginyeria que fa certes coses útils*".

La idea nuclear seria, per tant, la d'*automatització eficient* d'una tasca. En aquesta línia, caldria abandonar, des de l'inici de la recerca, la noció de *jutge* o *jutgessa robot* que sovint és objecte de notícies o articles als mitjans de comunicació (acompanyats, per cridar-nos l'atenció, d'imatges o vídeos més propis de la ciència ficció) i centrar-nos més en la idea de programari (*software*) que pot realitzar certes tasques de suport o complement de la tasca judicial. La vinculació entre IA i robòtica és, certament, molt rellevant, però més aviat en el camp industrial. És en aquest sector on opera la IA que es basa en l'aprenentatge *per reforç*, que seria el tercer tipus d'aprenentatge de la IA,

[3] https://www.theguardian.com/sport/2021/feb/12/deep-blue-computer-beats-kasparov-chess-1996.
[4] https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html.
[5] https://www.bbc.com/news/technology-50573071.
[6] *Llibre blanc sobre la intel·ligència artificial de la Comissió Europea, un enfocament europeu orientat a l'excel·lència i la confiança*, Brussel·les, 19.2.2020, COM(2020) 65 final, p.5.
[7] Lex Fridman Podcast #53 (2019, 29 de novembre), minut 23, *Noam Chomsky: Language, Cognition, and Deep Learning*|; https://www.youtube.com/watch?v=cMscNuSUy0I.

diferenciat dels ja referits, *supervisat* i *no supervisat.* Consisteix en què un agent (un robot) es desplaça per un entorn (real o virtual) i, per mitjà de sensors que li permeten interactuar amb aquest entorn, va aprenent per sí mateix (a desplaçar-se o a fer una determinada tasca) a base de mecanismes de prova i error i maximitzant les recompenses que el propi model li proporciona en cas de comportament exitós. Doncs bé, aquest no és el tipus d'aprenentatge propi de les eines d'IA que ens poden interessar per a l'Administració de Justícia. Aquestes ens remeten més, com dèiem, a la idea d'un programari inserit en un sistema de gestió processal. Potser és una imatge menys poètica, però, probablement, també més realista.

De fet, el filtre desmitificador que s'està proposant ens podria portar a un canvi en la mateixa terminologia. Podríem parlar, més que d'IA, d'*automatització* de certs processos. Ja hem vist que el camp d'aplicacions de la IA és extraordinàriament ampli. A algunes els poden anar bé certes denominacions grandiloqüents. No sembla que sigui el cas, però, de l'Administració de Justícia.

Per últim, quan un comença a estudiar el món de la IA, ben aviat se sorprèn que en la majoria de casos, els algoritmes que integren el nucli o cor d'aquestes eines no es creen o programen de nou, sinó que, simplement, ja estan creats (*built-in*) i poden ser descarregats gratuïtament (quan es tracta de codi obert, força freqüent en aquest sector) en certes plataformes de programació. És a dir, que la complexitat (i màgia) que podríem associar a la programació (codificació) de l'algoritme d'una eina d'IA sovint queda reduïda a la introducció en llenguatge informàtic comú (per exemple *Pyton*) d'una ordre tan senzilla com "*import*": per exemple, "*import LinearRegression*", essent *LinearRegression* l'algoritme que estem *cridant* (Géron, 2020, p. 99). Aquesta simple línia de codi implica, ella sola, la incorporació en el model que s'està creant de l'algoritme que en serà el motor.

Ens trobem, per tant, davant d'una paradoxa de la IA: l'element que s'acostuma a concebre com el seu nucli o essència (el seu fetitxe, podríem dir), l'algoritme, sovint és al que menys elaboració o treball s'hi dedica: la creació de l'eina d'IA acostuma a consistir, més aviat, en la concreció de la tasca que es vol realitzar, la recopilació de les dades, la seva *neteja, depuració* o *preparació*, el seu *etiquetatge* per a la fase d'ajustament i prova de l'eina (incorporació a una part de les dades de la resposta *correcta* per a la pregunta en què consisteix la tasca que es voldrà dur a terme més

endavant respecte de dades noves no etiquetades), la prova de diferents algoritmes (sovint *built-in*, o ja creats, com dèiem), l'elecció d'un d'ells i la implementació de l'eina, amb el corresponent seguiment posterior.

Això ens porta a la segona paradoxa de la IA: tot i consistir, en essència, en instruments per automatitzar certes tasques, requereix, això no obstant, una elevada intervenció i activitat humana, especialment en la fase de creació i implementació, si bé també en la de seguiment (anomenada també de *monitorització*). De fet, una de les tasques que més treball humà requereix és la d'etiquetatge, necessari en els sistemes *supervisats*, com hem vist. Així, en la IA el treball humà segueix sent molt necessari i és, de fet, voluminós. Segurament canviaran els tipus de treballs requerits, però seguiran essent necessaris. A més, aquests nous tipus de treballs no seran, necessàriament, de *qualitat* (treballadors de les empreses *tech*), sinó que també poden acabar creant un submón de certa explotació laboral per tasques com les apuntades d'*etiquetatge* (Bartoletti, 2020, p. 25). És probable que en el camp de l'aplicació judicial de la IA aquesta problemàtica no es presenti, però tampoc no en tenim la certesa, especialment si s'acaben utilitzant productes d'origen total o parcialment privat dels quals se'n desconeixin les condicions laborals en les quals han estat elaborats.

Les dues paradoxes de la IA apuntades operen en sentits oposats: mentre que la primera (molts algoritmes ja estan *prefabricats*) sembla apuntar a una simplificació de la matèria, la segona (hi ha moltes tasques humanes a realitzar, més enllà de l'elaboració o elecció de l'algoritme) ho fa en sentit contrari: crear una eina d'IA i decidir aplicar-la en qualsevol àmbit (especialment si es tracta d'un servei públic com l'Administració de Justícia) és una tasca extraordinàriament complexa i exigent.

1.6. *Desmitificant*, també, la funció jurisdiccional

Una vegada desmitificada, en part, la IA, cal fer el mateix, ara, respecte de l'altre pol de la recerca: la funció d'administrar justícia. No es tracta, en absolut, de reduir-ne la importància. És una de les tres funcions principals en què es fracciona el poder públic en un estat constitucional, democràtic, social i de dret, juntament amb la de legislar i la de govern o d'administrar. Quan parlem de *desmitificar* la funció judicial, es tracta, simplement, de posar de manifest que, sense perjudici de la transcendència jurídica i

social que presenta qualsevol de les seves actuacions (fins i tot les relatives a reclamacions aparentment menors), no sempre les tasques que exigeix són complexes.

De fet, en funció de les jurisdiccions, no és excessivament difícil trobar certes demandes, reclamacions o casos molt freqüents en termes absoluts, estructuralment senzills i als quals els tribunals acostumen a donar respostes uniformes i estandarditzades. Quasi bé mecàniques. En correlació, la creació (redacció) d'aquestes respostes acostuma a requerir comprovacions prèvies també senzilles i estandarditzades. Aquesta actuació judicial *repetitiva* no és, en si mateixa, reprovable o indesitjable. La provoca, de fet, la mateixa realitat social que s'aborda i l'ordenament jurídic que vincula als tribunals: si un conflicte és jurídicament estandarditzat, és raonable (fins i tot recomanable per raons de seguretat jurídica) donar-li una resposta estandarditzada.

Dit això, no estem parlant, necessàriament, només de decisions finals (sentències que abordin el fons del conflicte), sinó també, eventualment, de resolucions processals inicials o intermèdies que requereixen algun tipus d'examen jurídic i que presenten les dites notes de simplificació i estandardització. De fet, quan es planteja l'eventual aplicació d'eines d'IA a l'Administració de Justícia, se sol pensar en programaris que dictin sentències sobre el fons. Més enllà de les evidents problemàtiques jurídiques que aquesta eventualitat pot generar (i que abordarem a fons en el capítol 7), no hem d'oblidar que el procediment judicial està conformat de moltes etapes processals. Moltes d'elles són molt rellevants. D'altres, no tant. En algunes s'hi poden veure afectats els drets de les parts (per exemple, el de defensa o el d'igualtat d'armes). En altres, no. I, especialment, moltes d'aquestes tasques *inicials* o *intermèdies* (no *finals* o *decisòries*) són molt freqüents i voluminoses en termes absoluts[8].

[8] Pensem, per exemple, en seu civil, en la decisió judicial de despatxar una execució de títol judicial (d'una sentència o fins i tot d'un monitori previ en el qual no hi ha hagut oposició). Es tracta, simplement, de constatar el contingut del títol, si hi ha un pronunciament de condemna, si ha estat notificat, si no consta el compliment (pagament al compte del jutjat) i si ha passat el termini legal de compliment voluntari. La tasca no és, en efecte, complexa. Més aviat, senzilla i rutinària. Això no obstant, cal fer-la i acostuma a exigir, en un cas senzill, després de navegar pel sistema de gestió processal, i anant ràpid, un mínim de 2 o 3 minuts. Una vegada el jutge o jutgessa ha arribat a una conclusió (cal despatxar o no), l'haurà de transmetre a la oficina. Ho farà oralment, en una nota en un *post-it*, per correu electrònic o directament a través del sistema de gestió processal. El funcionari haurà de rebre la minuta i introduir-la al sistema. Posteriorment, arribarà la signatura electrònica i, per últim, la notificació telemàtica. Doncs bé, més endavant estudiarem si és possible automatitzar la primera fase d'estudi de la demanda d'execució. Potser a través del corresponent programari i d'eines d'IA de processament del llenguatge natural (*NLP*), el sistema podria desplaçar-se de manera automatitzada pels diferents documents que cal tenir en compte (sentència o decret de finalització del monitori, diligència de notificació de la resolució judicial, demanda executiva i comptes del jutjat) i, en pocs segons,

1.7. Intervenció íntegrament humana o íntegrament automatitzada: una falsa dicotomia?

En l'apartat anterior ja s'ha introduït la idea de la *col·laboració* entre el tribunal i l'eina d'IA. Hem intentat deixar de banda la idea de *jutge-robot* i hem dirigit el focus cap a altres maneres en què les eines d'IA podrien, a través d'una certa *interacció*, ajudar o complementar la tasca estrictament jurisdiccional. Podríem anomenar-ho *sistema de suport en la presa de decisions* (*decision support system*). La intervenció humana consistiria, així doncs, no només en *activar* l'eina, sinó també en comprovar-ne el resultat. Potser no s'arribarà a modificar la proposta o projecte de resolució generada de manera automatitzada, però sí que hi haurà hagut una supervisió humana posterior. El *suport* que aportaria la IA podria, de fet, no materialitzar-se necessàriament en una proposta de resolució: podria limitar-se, simplement, a processar i presentar la informació emmagatzemada que sigui rellevant per a la decisió que s'ha de prendre.

Aquest plantejament *simbiòtic* o *dialèctic,* degudament articulat, pot suposar, de fet, la maximització de totes les capacitats involucrades: la màquina farà el que sap fer millor o més ràpid i la persona farà el que també se li dona millor (la tasca *experta* de supervisió final). El resultat podria ser, en cas d'èxit, una millora global i una major eficiència.

Per altra banda, la proposta *dialèctica* (no *completament* automatitzada) és l'única viable en l'àmbit de l'Administració de Justícia, per raons evidents: en primer lloc, cal mantenir uns estàndards de qualitat mínims en la prestació del servei públic de justícia i aquests requereixen, a dia d'avui, la intervenció humana com a mínim parcial en la presa de decisions, fins i tot les de tràmit. En segon lloc, i de manera especial, la funció jurisdiccional es desenvolupa, necessàriament, en el marc d'un procediment judicial que no només té, per finalitat última, tutelar els drets dels ciutadans, sinó que, a més, no pot infringir, ell mateix (el procediment), aquests drets com a conseqüència de la manera en què es tramita. I això no sembla que es pugui aconseguir si es preveuen fases processals

oferir una resposta (és a dir, presentar una resolució ja redactada). La creació d'aquesta resposta s'hauria automatitzat, cosa que no vol dir que no hi hagués cap tipus d'intervenció humana. El jutge o jutgessa podria donar-la per bona o fer algun tipus de comprovació ulterior. De fet, el mateix programari podria informar del grau de dubtes que hi pot haver en la lectura automatitzada de certes informacions i oferir un enllaç directe per acudir no només al document rellevant, sinó a la part del mateix on hi ha aquesta informació rellevant. Al final, la decisió seria del titular de la funció jurisdiccional. Però és probable que el temps requerit fos menor que en cas de fer totes les comprovacions de manera successiva i individualitzada (fins i tot encara que sigui en un context o entorn altament digitalitzat).

íntegrament automatitzades. Dependrà, lògicament, del tipus de tràmit automatitzat, però el criteri general hauria de ser el negatiu.

Quedaria oberta, certament, la possibilitat de preveure una fase íntegrament automatitzada però amb la possibilitat de recórrer la decisió. En aquest cas, la resolució del recurs no podria estar, de nou, automatitzada. Hauria de ser *humana*. Així, l'actuació processal globalment considerada no estaria *completament* automatitzada. Això no obstant, com a criteri general serà preferible evitar aquests supòsits d'intervenció humana *diferida* i introduir, en tot cas, una fase de supervisió humana abans d'arribar a signar-se qualsevol resolució judicial, final o de tràmit.

Una qüestió diferent seria la necessitat de traslladar als interessats tota la informació que ha estat rellevant per a la decisió que s'ha pres. Només així podran impugnar-la en condicions, especialment pel cas de decisions judicials finals. Es tracta d'un dels temes *estrella* en matèria d'IA: fins a quin punt és materialment factible donar aquesta informació quan els algoritmes que actuen són molt complexos? Més endavant, al capítol 4.3, tractarem aquesta qüestió, però podem deixar apuntat que, probablement, les eines d'IA que eventualment es puguin acabar aplicant en seu judicial no siguin tan complexes com per no poder donar les explicacions i informacions adequades.

1.8. Naturalesa *política* de la decisió d'optar per eines d'IA: la *qüestió zero*

Cal recordar, ara, una obvietat: sense perjudici de la conveniència i necessitat de realitzar recerques com la present, per anar copsant progressivament els condicionants tècnics i les problemàtiques jurídiques que pot suposar la implementació d'eines d'IA en l'àmbit de l'Administració de Justícia, sempre hi haurà un moment determinat, si es constata la viabilitat tècnica i jurídica d'una determinada eina, en el qual caldrà decidir si s'implementa o no i fins a quin punt o respecte de quina tasca. Aquesta decisió és estrictament *política*. No hem de tenir por d'utilitzar l'adjectiu *política* (despullant-lo de les habituals connotacions pejoratives que l'acompanyen) en una matèria eminentment tècnica. La decisió és, de fet, principalment política, sense perjudici que pugui estar condicionada molt intensament per factors tècnics i jurídics. Sempre hi ha un marge discrecional dins del qual els actors polítics competents podran (i hauran) de decidir.

Dit d'una altra manera, si es constata que una determinada eina és tècnicament factible, que aparentment pot aportar una major eficiència a la tasca d'administrar justícia i que, a més, les problemàtiques jurídiques que genera la seva implantació són relativament assumibles, d'això no se'n deriva, en absolut, la *necessitat* d'implantar-la. Seguirà existint un marge decisori que només pot ser omplert amb una decisió *política*. Sempre ha de quedar oberta, especialment en un sector públic com és l'Administració de Justícia, l'anomenada *qüestió zero*: realment volem que una determinada tasca judicial de contingut decisori la realitzi de manera automatitzada un programari? Quina importància donem al fet que un estudi previ descarti problemàtiques jurídiques rellevants, per exemple, de tractament discriminatori o d'insuficient *transparència* o *explicabilitat* del funcionament del programari?

Aquestes qüestions han de quedar sempre obertes perquè han de ser els humans (en aquest cas, els polítics competents) els qui decideixin si una determinada tasca es *delega*, o no, a la IA. I perquè, com és obvi, les conclusions dels estudis previs sobre la inexistència de problemàtiques jurídiques poden no ser infal·libles: els problemes poden acabar sorgint en qualsevol moment durant la implantació de les noves eines. Això és així, especialment, en les eines d'IA d'*aprenentatge automatitzat*, és a dir, les que es modifiquen a sí mateixes durant el seu funcionament en funció dels usos que se'n fa.

En definitiva, la *qüestió zero* és una realitat certament òbvia que, això no obstant, convé no oblidar mai quan s'aborda una recerca com la present. Ens ha de permetre, també, desvetllar o *denunciar* la suposada *objectivitat* i *neutralitat* que a vegades se sosté o es dona per suposat que caracteritzen les eines d'IA, atès el seu caràcter tècnic i el seu funcionament automatitzat. Per contra, és evident que el procés d'implantació d'una eina d'IA està poblat per multitud de decisions humanes (quina tasca volem automatitzar? Quin model algorítmic utilitzarem? Amb quines dades l'entrenarem?, etc.) que descarten, per definició, que es tracti d'una realitat objectiva o neutre. De fet, tot el contrari: les decisions *polítiques* poblaran cadascuna de les fases, també les inicials i les intermèdies, d'implantació d'una eina d'IA. No només l'última, la de decisió final de si s'implementa o no.

Ens recorda Hildebrandt (2019) que un projecte d'aprenentatge automatitzat exigeix, sempre, una *tasca* ben especificada, una *mètrica* de funcionament i una *font* per la fase d'entrenament. També, una sèrie de decisions de *disseny* ineludibles, com ara la tria del

tipus d'entrenament que se seguirà, de la *funció objectiu* que es vol aprendre i la seva *representació*. Aquestes decisions impliquen, ineludiblement, diverses assumpcions, que poden ser vàlides o no. Per exemple, que la distribució entre les dades d'*entrenament* i les dades *futures* és contínua i homogènia. O l'existència d'un (no tan evident) nexe entre la lògica i l'estadística: només acceptant que hi ha una realitat matemàtica subjacent a la vida humana, les eines d'aprenentatge automatitzat tenen sentit. Si no, no en tenen, segons Hildebrandt. El problema és, precisament, que en el sector de l'aprenentatge automatitzat els *errors* o les assumpcions potencialment qüestionables no són fàcils de detectar perquè poden estar amagats en el procés de disseny. I, lògicament, aquests factors poden afectar la *fiabilitat* d'aquestes tècniques. Per tant, si es pretenen aplicar al sector públic, seran imprescindibles sistemes de control institucional de la validesa dels seus resultats.

Ens alerta també Hildebrandt del risc de reduir el problema de l'aplicació de les eines d'aprenentatge automatitzat al fenomen de la *caixa negra* (la impossibilitat de conèixer com funcionen algunes de les seves modalitats avançades, les *xarxes neuronals*) o dels potencials biaixos que poden generar o reforçar. Aquests són problemes rellevants que cal abordar, per descomptat: cal buscar models *transparents* i *explicables*, en un cas, o vies per assegurar el processament just i no esbiaixat de les dades, en l'altre. Però una atenció excessiva a aquests dos problemes *estrella* de la IA ens pot distreure de les nombroses (i discutibles) assumpcions que hi ha darrera del disseny de qualsevol model d'aprenentatge automatitzat. Hi són sempre, fins i tot quan no hi ha problemes de *caixa negre* o de *discriminació*. I no els hem d'oblidar o menystenir, com, de fet, conclou Hildebrandt, acostuma a passar en el sector.

Per tant, qui sostingui que es tracta d'eines neutres i objectives, que no requereixen fer-se cap altra pregunta que no sigui la relativa a la seva eficàcia i viabilitat tècnica o pressupostària, no només estarà, probablement, mentint, sinó que el que segur que estarà fent és posicionar-se *políticament* dins del debat *polític* que ha de generar, necessàriament, la nova realitat de la IA. Podem afirmar, així doncs, que la tecnologia nua ni prendrà les necessàries decisions polítiques que cal prendre ni resoldrà, per si mateixa, els problemes (fins i tot tècnics) que vagin sorgint. Serà un component clau, però no podrà operar sola. Ni ho farà en un entorn neutre.

No hi ha, de fet, cap urgència en implantar eines d'IA. Tampoc no és recomanable inscriure's a cap cursa per veure qui és el primer en aplicar una determinada aplicació a una determinada tasca. Fins i tot és convenient una pausa. Analitzar amb serenor tots els condicionants abans de prendre qualsevol decisió (política). Això és així, especialment, en l'àmbit de l'Administració de Justícia. Al mateix temps, però, cal procurar que un excés de consultes als interrogants que nodreixen la *qüestió zero* de la qual hem parlat no desemboqui en un indesitjat escepticisme tecnològic que bloquegi injustificadament l'avenç tecnològic en l'Administració de Justícia. La virtut està, de nou, en un terme mig mai fàcil de localitzar amb precisió.

1.9. Perspectiva pública

La IA ha estat i està dominada pel sector privat. Les institucions públiques estan fent esforços inicials per acotar-ne la seva evolució, però, de moment, els actors principals són privats, des dels propietaris de les xarxes socials fins a les empreses que ofereixen productes de *legal tech*.

Per contra, l'objecte de la recerca se centra en l'eventual aplicació d'eines d'IA no només a una administració pública (a qualsevol administració pública), sinó, concretament, a l'Administració de Justícia. La perspectiva de l'estudi, per tant, no pot ser una altra que la pública. La de l'interès públic. I és aquí on es genera una certa tensió entre l'evolució i estat del mercat privat que protagonitza l'avenç de l'IA i la recepció que en pugui fer una administració com la de justícia, tradicionalment endarrerida en termes tecnològics i, sobre tot, portadora d'una sèrie de principis i drets *innegociables*, per utilitzar un terme propi del sistema de mercat.

Caldrà disposar, per tant, dics potents contra les fortes inèrcies (i interessos) que impregnen el mercat privat de la IA. Des de la perspectiva de l'administració pública de justícia, ens serà completament aliè, en primer lloc, l'ànim de lucre que justifica la creació de qualsevol empresa privada. En segon lloc, ens començaran a interessar més els paràmetres d'eficàcia (assolir la tasca pretesa) i eficiència (fer-ho amb una despesa econòmica, humana i temporal raonable) que acostumen a tenir en compte aquestes empreses a l'hora de treure un producte al mercat. L'interès estarà, però, molt delimitat. No podrà ser absolut, sinó que vindrà condicionat per la permanent necessitat de mantenir la qualitat en la prestació del servei de justícia i la preservació i tutela dels

principis i drets afectats. En altres paraules, la ponderació que caldrà fer en cada cas serà més complexa: podrà haver-se de renunciar a una eina més eficient en favor d'una altra que tutela en major mesura (o que posa en risc en menor mesura) certs drets potencialment afectats.

Cal tenir present en aquest punt que, de fet, hi pot haver diferents perspectives *públiques* d'abordament de la realitat de la IA i del *Big Data*. És a dir, encara que ens pugui semblar obvi que la correcta sigui la de prioritzar l'interès públic i la preservació dels principis i drets fonamentals, actualment hi ha dos grans models antagònics entre sí (el nord-americà o *privatitzat* i el xinès o *estatal-autoritari*) que s'allunyen, cadascun a la seva manera, d'una altra possible proposta més *garant*, que seria, en sentit ampli, l'europea. Podem distingir, per tant, seguint Bartoletti (2020, p. 96), tres grans enfocaments públics, que si bé es refereixen prioritàriament al fenomen del *Big Data*, també ho fan, per via indirecta, a la IA:

a) El model *nord-americà*, que s'aproxima a aquestes realitats des d'una perspectiva de *mercat* segons la qual la informació de consum ha de ser accessible i el seu processament viable sense entrebancs excessius.

b) El model *xinès*, que confia l'accés i l'explotació de les dades al propi estat, però ho fa no tant per assegurar-se el control de la tutela dels drets de la ciutadania, sinó per instaurar un sistema *pseudo-autoritari* (o *autoritari*) per mitjà del qual controlar la seva població i nodrir el sistema de crèdits socials amb el qual es recompensa o castiga, amb mesures públiques, el comportament social.

c) El model de la *UE* (en procés de formació), més *garant*, que pressuposa que la privacitat és una qüestió vinculada amb la dignitat de l'individu i que són necessàries restriccions en les vies de compartir i manipular la informació.

Sembla evident, en qualsevol cas, que cal establir regles i estàndards imperatius. Preferentment, en una fase inicial, a nivell europeu, de manera que la resta del món s'hi hagi d'ajustar si pretén comerciar amb la UE. De fet, com és notori, un grup molt reduït d'empreses concentren molt poder a nivell global, per la qual cosa el sol fet de no haver-hi regulació en aquest camp fa que aquest poder no pari d'augmentar, amb efectes que és poc probable que reforcin la perspectiva *garant* europea. Tampoc no podem obviar,

com ens recorda Bartoletti (2020, p. 97), que les 9 companyies més grans del sector digital (*Baidu*, *Alibaba*, *Tencent*, *Google*, *Microsoft*, *Amazon*, *Facebook*, *IBM* i *Apple*) són les mateixes que dirigeixen el debat sobre la innovació en IA i sovint ho fan centrant-se més en improbables (però entretingudes) aplicacions robòtiques del futur i no tant en els desafiaments i vertaders riscos que poden plantejar les actuals eines d'IA, més senzilles, certament, però que ja estan en funcionament i són les que més ens han de preocupar.

Podríem concloure, així doncs, que convé prendre la iniciativa en el control, si més no parcial, del debat sobre la IA, i enriquir-lo amb l'aportació de la perspectiva *pública* i de la necessària preservació (real) dels principis i drets fonamentals potencialment afectats.

1.10. Titularitat pública de les eines d'IA judicial

En relació directa amb l'apartat anterior, un dels problemes que pot generar una eina privada d'IA és, a més de les elevades llicències que s'han de pagar, el fet d'estar protegida per drets de propietat intel·lectual o industrial. Si ens estem plantejant la possible implantació d'eines d'IA en qualsevol administració pública, la impossibilitat d'accedir al funcionament intern del sistema automatitzat que s'hagi utilitzat pot anar en contra de la necessària *transparència* amb la qual ha d'actuar tot servei públic. Pot apropar-nos a una certa *arbitrarietat* en l'actuació pública, expressament prohibida per l'ordenament jurídic. No cal dir que aquestes problemàtiques augmenten exponencialment si el servei públic al qual es vol inserir la IA és l'Administració de Justícia: ara es podran veure afectats, a més, drets fonamentals com el de defensa.

En definitiva, sembla evident que cal optar, des d'un inici, per eines d'IA de titularitat pública i, preferentment, amb ús de *codi obert*. Només així s'evitarien problemàtiques greus com les exposades que poden erigir-se, de fet, per sí mateixes, en obstacles insalvables per a la introducció de la IA a l'Administració de Justícia. Dit això, aquesta aposta no ha de ser necessàriament incompatible, quan es consideri convenient, amb la col·laboració, per via de conveni o la forma jurídica corresponent, amb actors del sector privat per desenvolupar i implementar certes aplicacions d'IA. Ni tampoc amb la possibilitat d'importar puntualment recursos generats en el mercat privat i integrar-los a l'entorn públic, sempre, però, que es compleixin totes les exigències i estàndards que siguin aplicables.

Per últim, cal tenir present que si el que es pretén és construir, per a l'Administració de Justícia, un entorn digital complet i integrat on convisquin el sistema de gestió processal en sentit ampli amb certes eines complementàries (algunes d'elles qualificables d'IA), pot donar-se el cas que alguns components no *sensibles* puguin no disposar de codi obert, sense generar problemàtiques particulars, mentre d'altres, pel seu caràcter *sensible*, sí que requereixin aquest codi obert. L'anàlisi ha de ser, de nou, particularitzat.

1.11. Potencial interès per a totes les jurisdiccions i implantació estratègicament progressive

L'autor de la recerca és un magistrat civilista amb un coneixement més profund de la realitat dels jutjats de primera instància, en comparació al que pugui tenir sobre altres jurisdiccions. Per aquest motiu, és probable que de les propostes que es facin de possibles aplicacions d'eines d'IA a la justícia una majoria es refereixi a la jurisdicció civil o que aquestes gaudeixin d'un nivell d'anàlisi més detallat.

Dit això, cal matisar que és probable que alguna de les propostes que es facin, relatives a actes de tramitació, puguin ser aplicables a més d'una jurisdicció, sense perjudici de les modificacions necessàries en funció del règim jurídic aplicable. Pensem, per exemple, en la possible detecció automatitzada de la manca de competència territorial. En qualsevol cas, és evident que, pel cas que en un futur més o menys proper es plantegi la introducció d'eines d'IA a l'Administració de Justícia, seria prudent, i de sentit comú, començar per les jurisdiccions menys *sensibles*. Segurament cap professional admetrà que la seva jurisdicció és poc sensible, però és probable que arribem a un consens segons el qual la jurisdicció penal (en la qual es veu afectat constantment el dret a la llibertat) pot ser la més sensible a efectes d'implantar-hi en primer lloc eines d'IA. Podríem relegar-la, així doncs, a l'últim lloc. Si bé, paradoxalment, és a l'àmbit penal (especialment respecte de la seva fase d'investigació o de les mesures cautelars) on sembla destinar-se més esforços d'investigació en IA. Ho anirem veient.

Dit això, és relativament probable que, en un futur no immediat però segurament tampoc no molt llunyà, s'aprovi una nova LECrim que atorgui la instrucció de les causes a la fiscalia i relegui el jutjat d'instrucció a una funció de garanties. En aquest context, eventual però probable, potser canvia la percepció sobre el grau d'inadequació *a priori* de plantejar la possible aplicació d'eines d'IA a la *investigació penal* en sentit ampli.

Aquesta investigació seria duta a terme per la fiscalia, segons uns paràmetres (salvant, lògicament, totes les distàncies) no tan allunyats de la *investigació policial* dels delictes, en la qual ja estan relativament implantades certes eines d'IA. Evidentment, que estiguin implantades no exclou que no hagin de ser objecte (com qualsevol altra innovació) d'una anàlisi crítica, especialment si en un futur poden ser traslladades, en part, a l'actuació de la fiscalia. Es tracta, simplement, de mantenir un enfocament ampli i obert en tot moment. Més endavant veurem amb més detall totes aquestes qüestions.

En definitiva, el criteri determinant no serà, només (ni, segons els casos, principalment), el tipus de jurisdicció afectada, sinó més aviat la concreta tasca judicial a la qual es pugui aplicar una eina d'IA determinada: podríem trobar una tasca de tràmit quasi de mera gestió en la jurisdicció penal molt més adequada que una altra més complexa i delicada de la jurisdicció civil. L'enfocament haurà de ser, per tant, casuístic. I, en aquesta línia, seria raonable començar amb aplicacions referides a actuacions de tràmit (detecció de casos d'incompetència territorial, admissió a tràmit d'una demanda d'execució, etc.) abans d'inserir eines d'IA en la fase de decisió final. Per últim, no cal recordar que, en tot cas, seria imprescindible abordar una fase pilot de prova en la implantació de qualsevol eina d'IA.

Són, totes elles, qüestions òbvies que poden ajudar, però, a iniciar el camí en una matèria, la IA, i un terreny, el de l'Administració de Justícia, desconeguda i delicat, respectivament.

1.12. Heterogeneïtat d'interessos dels diferents operadors jurídics

Només seran objecte d'anàlisi les eines d'IA que puguin tenir un eventual interès públic des de la perspectiva del tribunal que administra justícia. Per contra, ja no es prestarà el mateix grau d'atenció (potser sí alguna atenció) a altres eines que puguin tenir un interès més específic per altres operadors jurídics, como ara l'advocacia. És una delimitació important, ja que precisament a l'àmbit de l'anomenada *legal tech* ja hi ha, a dia d'avui, a l'àmbit privat, força productes tecnològics que sostenen utilitzar recursos d'IA i que van dirigits principalment a operadors jurídics també privats. Es tracta no només de les bases de dades habituals de legislació o jurisprudència (que poden ser també d'interès pels professionals del tribunal), sinó també d'eines de predicció de sentències futures, fins i

tot amb referència a un tribunal o jutge particular (l'anomenada *justícia predictiva*), o eines de creació d'estratègies jurídiques.

De nou, no es tracta de descartar de manera absoluta aquests instruments, sinó de centrar el focus de la recerca dins d'uns paràmetres d'interès públic-judicial raonables. Al mateix temps, però, veurem que les fronteres no són molt nítides: l'estructura bàsica que hi pot haver en un producte privat de justícia predictiva podria ser (o no) parcialment extrapolable a la predicció de resultats judicials en determinats tipus de casos o contextos que sí que poden tenir un interès públic identificable[9].

En definitiva, caldrà tenir una mirada escèptica i crítica a l'hora d'analitzar els productes privats de *legal tech* destinats a l'advocacia, però no ignorar-los. No hem de descartar, ja d'entrada, qualsevol zona de concurrència d'interessos o usos comuns. Pensem en les eines d'anàlisi de documents, de confecció estandarditzada de documents o l'anomenada prova electrònica o *e-evidence*. El camp és ampli i l'avenç en el sector privat elevat. Caldrà prestar-hi l'atenció deguda, amb consciència, però, que els interessos que hi ha darrera acostumen a ser diferents dels interessos públics de l'Administració de Justícia. De fet, és precisament aquesta disparitat d'interessos el que explica que les entitats privades no acostumin a donar massa informació sobre les especificacions tècniques dels seus productes.

1.13. Accessibilitat, capacitats disminuïdes i bretxa digital

Quan s'aborda qualsevol matèria d'innovació tecnològica, cal tenir present, des d'un inici i en tot moment, la mesura en la qual la nova tecnologia, més enllà de la seva eficiència, pot millorar o empitjorar les condicions d'accessibilitat per a persones amb capacitat reduïda o limitacions econòmiques o socials per fer-ne ús. És una perspectiva ineludible. Es tracta d'un problema general de la IA i de les nostres societats tecnològiques que no està rebent l'atenció deguda: l'anomenada *bretxa digital* provocada per una insuficient

[9] Per exemple, en les reclamacions indemnitzatòries per danys corporals, aquesta informació de predicció estadística podria posar-se al servei de la ciutadania, posem per cas, en el marc d'una negociació prèvia duta a terme en una resolució alternativa de conflictes (*ADR*) o, més concretament, en una resolució en línia de disputes (*ODR*), per tal que puguin tenir una informació aproximada contrastada sobre el que poden o no reclamar o poden o no haver de pagar, per tal de facilitar, en el seu cas, l'assoliment d'un acord. Es tracta només d'una idea que, per cert, ja s'està treballant a França (en el projecte *DataJust*, del qual en parlarem al capítol 10.4.3).

atenció a les limitacions reals existents per a certs col·lectius per accedir a l'ús de les noves tecnologies i que pot acabar creant una ciutadania de primer nivell i una de segon.

Aquesta *bretxa digital* pot consistir, en primer lloc, en no tenir accés al programari o als equips necessaris, o fins i tot a l'ample de banda que exigeixi una aplicació en línia determinada per poder fer un tràmit, presentar un escrit, rebre una notificació o assistir telemàticament a un acte judicial. En segon lloc, pot derivar, simplement, de no tenir les habilitats, coneixements i pràctica suficients per poder utilitzar les noves eines, realitat que exigeix abordar amb caràcter general la deguda *alfabetització digital*. Si, a més, aquestes limitacions estan presents en persones que pertanyen a col·lectius vulnerables, llavors els problemes d'accessibilitat (en aquest cas a la justícia amb possible afectació al dret a la tutela judicial efectiva) poden multiplicar-se exponencialment.

1.14. Tímida regulació normativa i proliferació de cartes i guies ètiques en matèria d'IA

La IA, en les seves modalitats avançades que fan ús del *Big Data*, és un camp força incipient que només recentment ha començat a rebre l'atenció dels legisladors. La reacció normativa és, però, encara molt fragmentària i tímida. És probable que es consolidi en els propers anys, si no mesos. Això explica que, de moment, el que sí que s'ha aprovat són multitud de cartes de principis, guies ètiques o protocols, no vinculants, que haurien de regir la implementació de les eines d'IA. Més endavant les analitzarem amb més detall. El fet que no siguin obligatoris no elimina el seu interès, ja que són gairebé els únics textos amb pretensió pseudo-normativa i mínimament articulats dels quals es disposa ara mateix i que ens poden ajudar a identificar i ordenar quines són les problemàtiques socials i jurídiques que poden generar certes eines d'IA.

Veurem que aquestes cartes ètiques han estat aprovades per operadors tan públics com privats i que acostumen a ser massa genèriques (probablement per la seva pretensió d'aplicar-se a àmbits molt amplis) i a repetir-se. Aquest és el taló d'Aquil·les d'aquestes cartes o guies, ja que cada sector concret, privat o públic, genera una tipologia de problemàtiques específiques en matèria d'IA que no poden ser adequadament abordades amb uns principis definits per poder ser aplicats a qualsevol altre àmbit. Podríem trobar una excepció, precisament, en l'àmbit de la justícia: disposem de la *Carta*

ètica europea sobre l'ús de la IA en el sistema judicial i el seu entorn, aprovada el 2018 per la Comissió Europea per a l'Eficiència de la Justícia (CEPEJ).

Aquests textos ètics ens parlen, en general, dels principis de justícia, no discriminació, responsabilitat, rendició de comptes, transparència, explicabilitat, indemnitat, protecció de dades o seguretat, entre altres. Es tracta, certament, de principis molt generals i respecte dels quals difícilment algú hi estarà en contra. Ens centrarem, lògicament, en allò que sigui rellevant per l'objecte de la recerca, a l'espera que s'aprovin uns imprescindibles textos normatius que, a més de vinculants, siguin més concrets i s'apliquin a camps específics.

A més, caldrà tenir present, com ens recorda Bartoletti (2020, p. 105), que han augmentat els estudis sobre ètica en l'àmbit de la IA redactats per acadèmics patrocinats o promoguts per les empreses privades que operen en el mateix sector de la IA: no es pot ignorar que aquesta efervescència de principis i guies, promocionada en part per les mateixes companyies que comercialitzen aquests productes, pugui respondre a una qüestió d'imatge davant dels clients i potencials compradors. Es pregunta Bartoletti "quantes conferències més necessiten per tornar a proclamar aquests principis, per reconèixer una vegada més el principi de justícia i discutir sobre si els algoritmes tenen prejudicis". Seria hora, segons Bartoletti, que la força de la llei actuï. Hi podria haver un excés d'ètica en la IA. Una tàctica d'*ethics bluewashing* per mitjà de la qual els principis ètics serien utilitzats per justificar i ampliar certes finalitats comercials.

1.15. Protecció de dades

Una matèria vinculada a la IA que sí que disposa de regulació normativa vinculant és la protecció de dades, de la qual cal destacar el Reglament (UE) 2016/679 del Parlament Europeu i del Consell, de 27 d'abril de 2016, relatiu a la protecció de les persones físiques en el tractament de dades personals i en la lliure circulació d'aquestes dades (l'anomenat RGPD), adaptat a l'ordenament jurídic espanyol per la LO 3/2018, de 5 de desembre, de Protecció de Dades Personals i garantia dels drets digitals. Estem parlant, per tant, d'una normativa que ja està en vigor i que regula amb caràcter general tot allò que fa referència al tractament de les dades personals.

Es tracta d'una matèria molt sensible en societats digitalitzades com la nostra[10]. Però la protecció de dades no serà objecte d'una anàlisi detallada en la present recerca. Només hi acudirem de manera puntual i fragmentària. Serà així per raons d'espai i de claredat expositiva[11]. Per exemple, sí que s'abordaran previsions puntuals de la normativa de protecció de dades que es refereixen, de manera específica, a possibles aplicacions d'IA. Principalment, l'art. 22 RGPD que estableix el dret de qualsevol interessat a saber si s'utilitza, o no, en una tasca determinada, un sistema d'IA completament automatitzat i les seves facultats (ja avancem que no absolutes) d'oposar-s'hi. Ho deixem per més endavant (capítol 3.3.1.2).

1.16. Perspectiva interdisciplinària

Del que s'ha exposat fins ara se'n deriva com una obvietat, o un imperatiu, la necessària perspectiva o metodologia *interdisciplinària* que ha de governar tant l'abordament d'un estudi com el de la present recerca com l'eventual implantació, en el terreny judicial real, d'alguna eina d'IA. En aquesta matèria s'hi veuen implicades, com a mínim, quatre esferes o nivells:

a) El de la *política*, que ha de prendre les decisions d'implantar, o no implantar, una determinada eina d'IA a l'àmbit judicial.

b) El *legal* o de coneixement jurídic *expert* (*domain knowledge)*, que col·laborarà en la filtració dels criteris o paràmetres jurídics que siguin rellevants en una determinada tasca judicial i respecte d'una determinada eina d'IA.

c) El de les matemàtiques i l'estadística, ja que probablement s'haurà de fer ús d'equacions i fórmules matemàtiques.

[10] Pensem, per exemple, en l'escàndol del cas *Cambridge Analytica*, que va suposar l'obtenció il·legal de dades de milions de ciutadans exposats a *Facebook* a través d'un concurs o test de personalitat, venudes posteriorment i utilitzades per microdirigir la publicitat aplicant tècniques de les ciències del comportament per influir d'aquesta manera en votacions com la del *Brexit* o en eleccions com les presidencials americanes de 2016.

[11] Es tracta d'una delimitació negativa força rellevant si tenim en compte que una part molt significativa dels estudis que es fan a l'àmbit de la IA tenen per objecte, precisament, qüestions vinculades amb la protecció de dades.

d) El de la programació informàtica, imprescindible per poder crear i codificar els models.

Sense perjudici de l'autonomia de cadascuna d'aquestes esferes, caldrà interconnectar-les amb una comunicació real i flexible. Només així serà possible la creació d'eines d'IA judicial realment eficients, útils i viables. Caldrà construir ponts de reciprocitat operativa. Aquesta aproximació interdisciplinària, i no perifèrica, s'ha de produir, com recull *A Tale of Cyberjustice* (2019), en un espai de reflexió i experimentació en el qual els juristes puguin interactuar amb els programadors, amb els especialistes en les ciències de la informació i amb els demés actors judicials. Només així, no limitant-nos a un únic plantejament teòric, es podrà capturar la complexitat del fenomen. Caldrà consolidar ponts fluids i eficaços d'enteniment entre els tècnics informàtics creadors de les eines, els tècnics informàtics de l'administració encarregats de la gestió ordinària d'aquests recursos (que és probable que siguin diferents dels primers) i els operadors estrictament jurídics i no especialistes en qüestions tècniques. Caldrà crear un *llenguatge comú*, entenedor, suficientment tècnic, però no massa. Potser serà convenient la creació d'una figura de *punt d'enllaç digital*, ocupada per una persona suficientment experta (fins a un cert nivell) en cadascuna de les esferes indicades i que ajudi a fer real una vertadera comunicació entre els diferents agents implicats. Un cas recent i paradigmàtic d'aquesta aproximació interdisciplinària és *l'European Cyberjustice Network* (*CEPEJ-GT-CYBERJUST*[12]), que per abastar els aspectes tècnics de la *ciberjustícia* i la IA integra experts de camps tan diferents com les tecnologies de la informació i la comunicació (ICT), jutges, personal dels tribunals, personal administratiu de gestió dels tribunals, advocats, funcionaris d'execució i acadèmics, entre altres.

1.17. IA processal i IA com a institució jurídica *material*

Fins ara les premisses de la recerca han girat entorn de possibles aplicacions processals d'eines d'IA. Aquestes seran, de fet, el nucli de la recerca. No hem d'oblidar, però, que fins i tot en el cas que es conclogui que no convé aplicar cap tècnica processal d'IA a la justícia (de la mateixa manera que si es conclou que se n'han d'aplicar moltes), la IA

[12] https://www.coe.int/en/web/cepej/cepej-working-group-cyber-just.

seguirà existint fora del procés judicial, a la realitat social. Tant en el mercat privat com en l'actuació d'altres administracions públiques, que, de fet, ja l'estant utilitzant[13].

Per tant, l'ús per part dels actors privats o públics d'eines d'IA generarà ineludiblement conflictes que, tard o d'hora, acabaran arribant als jutjats perquè una decisió judicial doni una resposta. Per exemple, sobre si el responsable d'una eina d'IA ha de respondre, o no, d'uns danys causats (vehicles autònoms) o si una administració pública ha actuat o no de manera transparent a l'hora d'assignar determinades ajudes.

I quan aquestes demandes (o denúncies) arribin als jutjats, les referències a la IA que s'inclouran a les demandes ho seran com al·legacions de fons, com a elements de les *causes de demanar* d'una determinada pretensió. Llavors haurem de tractar judicialment la IA com a institució jurídica material, no processal.

La paradoxa és que, per poder-la abordar judicialment com a qüestió de fons, caldrà conèixer-la mínimament. I moltes de les qüestions que caldrà saber són, també, objecte d'estudi en la present recerca, encara que el seu objectiu últim sigui el processal.

[13] Pensem, per exemple, en els algoritmes utilitzats per assignar ajudes públiques en matèria d'energia elèctrica (algoritme *BOSCO*) o en les eines emprades per l'administració penitenciària en la gestió dels permisos penitenciaris (en parlarem al capítol 11.20.4).

2. Frontera entre digitalització judicial i IA judicial

2.1. Paradigma analògic passat, digitalització actual i IA judicial future

Una de les disjuntives que cal abordar és la distinció entre *digitalització* de l'Administració de Justícia, per una banda, i la possibilitat d'aplicar-li recursos d'IA, per l'altra. No sembla que hi hagi discussió en el fet que, a dia d'avui, a Espanya i Catalunya ja s'ha arribat, en Justícia, a un grau de digitalització rellevant. Certament, no és comparable al d'altres administracions, com la tributària, però l'avenç és innegable. La possibilitat de presentar escrits i documents telemàtics, la comunicació telemàtica amb les parts i entre les parts o l'expedient judicial electrònic, amb tendència a l'eliminació (mai absoluta) del paper, en serien els tres exemples principals.

Queden lluny els temps en els quals era una novetat la mateixa incorporació d'equips i aplicacions informàtiques d'ofimàtica o la implantació dels sistemes de gestió processal (SGP) que, anant més enllà del simple tractament de textos, permeten una gestió completa i ordenada de la recollida, emmagatzematge, processament i transmissió de la informació de tots els procediments tramitats per les oficines judicials. De fet, els SGP van de la mà de l'expedient judicial electrònic (EJE), que recopila en format digital tota (o gairebé tota) la informació necessària per la tramitació i resolució d'un procediment judicial determinat, tant la generada per la mateixa oficina judicial com l'aportada per les parts o tercers que es relacionen o col·laboren amb l'Administració de Justícia (Delgado, 2020, p. 297)[14]. I, més enllà de l'EJE, cal destacar com a realitats tecnològiques ja existents, palpables i efectivament utilitzades, la presentació telemàtica d'escrits i documents, l'atorgament electrònic de la representació processal, les subhastes electròniques, el compte de dipòsits i consignacions judicials o les seus judicials

[14] El document en paper és, així, substituït per documents electrònics que es gestionen, intercanvien i arxiven a través de sistemes informatitzats. L'EJE tindrà un número d'Identificació General (NIG) i quedarà integrat per diversos elements: el conjunt de documents electrònics corresponents a un procediment judicial, l'índex electrònic (com a garantia de la integritat de l'EJE), la signatura de l'índex electrònic i les metadades. Serà molt rellevant la catalogació documental. És a dir, l'assignació a cada document d'una descripció i tipologia documental adequada. També la garantia que les còpies electròniques siguin idèntiques al document original i que no comportin canvi de format ni de contingut. Només així podran tenir l'eficàcia jurídica pròpia del document electrònic original.

electròniques. Especialment, també, el *Punt Neutre Judicial*, que permet una gran varietat d'actuacions judicials[15].

Certament, als SGP (en plural, perquè n'hi ha diversos dins de l'estat espanyol) els pot quedar un marge rellevant de millora i perfeccionament. Probablement, també, d'interoperabilitat (entre sí i amb la resta d'operadors jurídics i administracions públiques). Però són una realitat plenament consolidada. Són, de fet, la plataforma o entorn dins del qual poden integrar-se, de manera acumulada i harmonitzada, altres avenços tecnològics, com la plena digitalització o, potser, algunes eines d'IA. Estem situats en aquesta cruïlla tecnològica: a dia d'avui, no sembla que s'apliquin en seu judicial, ni a Espanya ni a Catalunya, recursos d'IA en sentit estricte. Un dels objectius de la recerca serà, de fet, constatar si això és així o si ja s'està aplicant alguna eina judicial d'IA. Però no sembla que aquest sigui el cas.

Per tant, hi hauria una nova tensió entre el que ja existeix (digitalització) i el que podria existir en un futur (eines d'IA per a l'Administració de Justícia). Com veurem, però, la tasca d'intentar definir què és una eina d'IA, per diferenciar-la de la mera digitalització, no és en absolut senzilla: en funció de quina noció d'IA sigui la finalment escollida (per exemple, és IA un sistema automatitzat però sense aprenentatge automatitzat?), la distància entre eines d'IA i mera digitalització podrà ampliar-se o escurçar-se. Veurem, per exemple, que un recurs com la *OCR* (reconeixement òptic de caràcters o *optical character recognition*), que podria permetre extreure el contingut de text no només de demandes sinó també de documents, tot i poder semblar una eina de mera digitalització (conversió de text inclòs en un document en text tractable a disposició del tribunal), consisteix, en algunes versions avançades, en una aplicació molt complexa i avançada d'IA (de les anomenades de coneixement profund o *deep learning*).

Anirem veient, en definitiva, fins a quin punt moltes qüestions són merament terminològiques i que les nocions o definicions que escollim seran només relativament rellevants. L'important serà saber en cada moment quina és l'eina analitzada, quins mecanismes de funcionament utilitza i quina afectació pot tenir en els estàndards ineludibles de qualitat i de respecte dels drets que ha de regir el funcionament de

[15] Consultes patrimonials i domiciliàries, a la AEAT, la DGT, al Registre de la Propietat, a les prestacions d'atur, accés a la informació policial del DNI, a la oficina virtual del Cadastre, a la TGSS, al Fons de Garantia Salarial, al CORPME, a la comunicació d'exhorts, les consultes penitenciàries, els embargaments telemàtics de comptes a la vista o l'estadística judicial, entre altres.

l'Administració de Justícia. En altres paraules, caldrà adoptar una perspectiva *material* i no merament *nominal*.

2.2. Mer aprofundiment en la digitalització o automatització decisòria

La perspectiva material i no merament nominal que s'adoptarà augmentarà la complexitat de la recerca, ja que ens obligarà, en certs moments, a endinsar-nos en qüestions eminentment tècniques. No serà suficient identificar pel seu nom les aplicacions i enumerar superficialment els seus usos. Aquest tipus d'informació, tot i que necessària, no permet l'anàlisi crítica de les seves implicacions des de la perspectiva dels estàndards de qualitat de l'Administració de Justícia. Caldrà detallar, en cada cas, les particularitats tecnològiques (tècniques) de cada aplicació. Lògicament, sempre que es disposi de la informació, ja que, com veurem, no sempre és fàcil d'obtenir, especialment si el producte es comercialitza en el mercat privat (en el qual s'acostuma a magnificar les aportacions i la complexitat tecnològica dels productes que es pretén vendre sense concretar-ne massa les seves especificacions tècniques).

Haurem de precisar, en cada cas, si ens trobem merament davant d'un mer aprofundiment en el procés de digitalització en el qual ja estem immersos (per exemple, la transformació d'un element analògic en digital o la modificació d'un element ja digital en una altra forma digitalitzada més pràctica) o, per contra, si es tracta, ja en sentit estricte, d'una eina d'IA en la qual operi algun tipus d'*automatització decisòria*.

Ens trobem, però, davant d'una distinció més subtil del que pot semblar d'entrada. En el primer cas, l'eina no s'insereix en un procediment decisori, ni en la seva fase final ni en una inicial o intermèdia. Simplement opera, prèviament, una transformació passiva d'un determinat element, sigui d'origen analògic per passar a ser digital o d'origen ja digital al qual se li canvia el format o la forma. El procediment decisori s'iniciarà posteriorment. O al marge. Podem incloure dins d'aquest grup les eines de gestió de la informació digital, com serien els sistemes de gestió processal (SGP), que abasten tot el procediment, fins a la seva finalització, i als quals s'hi incrusten, sense arribar a incidir en elles, les diferents decisions judicials (processals o de fons) que van poblant el procediment.

Tots aquests recursos tècnics poden ser qualificats com eines *passives* o *no decisòries*. Això és el que les converteix en un mer aprofundiment en la digitalització. La subtilesa

de la distinció rau en el fet que aquestes eines també operen (lògicament, quan se les activa) de manera automatitzada o automàtica. No requereixen la intervenció humana *mentre actuen*. La seva essència distintiva, per tant, no és el seu caràcter automatitzat (que comparteixen amb les eines d'IA), sinó el fet de no inserir-se, estrictament, en el procediment decisori. No li aporten cap contingut de fons. Només el faciliten o el fan més còmode. Operen abans que s'iniciï o al marge.

Per contra, una eina que puguem qualificar estrictament d'IA ha de ser una tecnologia d'*automatització decisòria*. Podrà presentar, com veurem, diferents graus de complexitat i de problemàtiques jurídiques (podrà aprendre per si mateixa, o no; podrà aportar algun tipus d'explicació del seu funcionament, o no), però caldrà que sigui *activa*, que estigui *inserida* en el procediment de presa de decisió. Que li aporti algun contingut. Aquesta inserció en la presa de decisió podrà ser inicial, intermèdia o final, total o parcial, respecte de la resolució que calgui dictar, però caldrà que hi sigui. És aquesta inserció, aquesta aportació *activa* en la decisió, la que permet qualificar l'eina com a *intel·ligent*, amb tots els matisos que calgui. Si, en sentit contrari, és una eina que, tot i poder ser molt útil, és aliena o col·lateral, prèvia o posterior, a la presa de decisió, llavors probablement faríem millor de qualificar-la com a eina de mer aprofundiment en la digitalització. Les fronteres són en tot cas, com veurem, força difuses.

2.3. Un marc de governança especialment complicat

Des d'un inici convé tenir present que el marc de governança que afecta el servei públic de l'Administració de Justícia és molt especial i heterogeni i, en correlació, especialment complex. Més de l'habitual. S'hi veuen implicats molts actors. Polítics, judicials, corporatius i privats: des de l'estat central en sentit ampli (competent per regular el Poder Judicial i aprovar les reformes processals principals) fins al Ministeri de Justícia (que gestiona l'administració de justícia en les CCAA sense competència), passant per les CCAA amb competència (que han creat, per exemple, els diversos SGP existents). I sense oblidar-nos, lògicament, del CGPJ (òrgan de govern del Poder Judicial), la Fiscalia o els diferents col·legis de l'advocacia i la procuradoria. Fins i tot podríem diferenciar entre el CGPJ en sentit institucional i, en un nivell inferior, el col·lectiu professional de jutges i jutgesses (aquí podríem fer referència a les juntes de jutges) que podran ser els destinataris i usuaris de les eventuals noves eines. Caldrà tenir-los en compte, ja que el

seu posicionament no ha de ser, necessàriament, coincident amb el del CGPJ. El mateix podria dir-se, en un altre nivell, del col·lectiu dels LAJ.

El mapa és, així doncs, molt complex. Segurament per aquest motiu s'acostuma a proposar un enfocament de *cogovernança*, en el qual, sense perjudici de mantenir i exercir cada instància les seves competències pròpies, es busqui en tot moment una coordinació i harmonització de les propostes. El motiu és clar: tots els participants han de sentir-se còmodes amb la implantació de qualsevol novetat en un servei públic amb tantes ramificacions. Si això és així amb caràcter general, encara ho és més si el que ens estem plantejant és la introducció d'eines d'IA, que, segons els casos, poden implicar un cert canvi de paradigma. És en aquest sentit que caldrà tenir sempre present aquest especial marc de (co)governança, ja que la recerca se centrarà no només en la viabilitat tècnica i la conveniència pràctica d'una innovació determinada, sinó també en la seva factibilitat, podríem dir, *política*, atès aquest especial marc de governança.

2.4. Límits competencials: l'*administració* de l'administració de justícia

En estreta relació amb les problemàtiques de governança hi trobem els límits competencials inherents a un estat autonòmic en el qual la competència en matèria de regulació del poder judicial i de les lleis processals és plenament estatal, amb excepcions molt marginals. Això ens porta a la distinció entre les nocions d'*administració de justícia* i d'*administració de l'administració de justícia*.

En alguns casos la distribució competencial serà evident. En altres, com veu-rem, no tant, ja que algunes eines podran incidir principalment en tasques realitzades per l'oficina judicial i altres podran rebre un ús més intensiu per part de LAJ's o jutges i jutgesses. La casuística és molt àmplia. En qualsevol cas, però, el sol fet que respecte d'una possible innovació es plantegin seriosos dubtes competencials no exclourà que sigui analitzada, ja que, en definitiva, l'objecte de la recerca és la IA i l'Administració de Justícia. Simplement es tractarà d'identificar, per alguna de les propostes que es facin, el seu grau de *viabilitat competencial* en el marc de l'actual estat autonòmic. Les problemàtiques competencials d'una IA judicial seran abordades amb més detall a l'annex 2.

2.5. Correlació entre tipus d'eina d'IA i les seves problemàtiques jurídiques

Per poder comprendre quines concretes problemàtiques jurídiques (en termes d'afectació a drets o d'impacte en la qualitat de l'Administració de Justícia) té una determinada aplicació d'IA, cal conèixer, prèviament, la tecnologia i les especificacions tècniques que hi ha darrera. Lògicament, no en un grau de detall molt elevat (objectiu inassolible atesa la formació estrictament jurídica de l'autor de la recerca) però sí en un nivell suficient per poder contextualitzar la complexitat de l'eina i quins desafiaments jurídics planteja. Posarem alguns exemples.

Si ens trobem davant d'un recurs tecnològic que, per molt cridaners que en siguin els resultats, l'acabem qualificant de mer aprofundiment en la digitalització (per no tenir cap tipus d'intervenció en la presa de decisions), llavors seran poques les qüestions a abordar. En primer lloc, si és tècnicament viable; en segon lloc, si realment implica una aportació pràctica interessant i substancial; i, en tercer lloc, si el cost en recursos materials i tècnics i les disfuncions que pot generar la seva implementació queden justificats per l'aportació pràctica que se n'espera. Per contra, no sembla, en principi, que s'hagin de veure implicats altres principis o drets més profunds vinculats amb l'Administració de Justícia (dret a un procés just, dret de defensa, etc.), precisament perquè l'eina no s'insereix en la presa de decisió.

Per contra, si ens desplacem a les eines que acabem qualificant de decisió automatitzada, és a dir, d'IA, els dilemes i problemàtiques jurídiques que es poden generar creixen exponencialment, per motius obvis. Lògicament, el sol fet que un expert no jurista (per exemple, un matemàtic o un informàtic) intervingui en la creació d'un component que ha d'incidir en la presa d'una decisió judicial genera, per si sol, una certa prevenció. Aquesta augmentarà si, a més, l'eina ha estat entrenada i provada amb dades no etiquetades manualment amb la resposta correcta (*aprenentatge no supervisat*) o si aprèn i es modifica a si mateixa durant el seu cicle d'ús (eines d'*aprenentatge automatitzat*). I els dubtes es dispararan si, a més, la complexitat dels mecanismes de funcionament de l'eina és tan elevada que impedeix obtenir una explicació dels motius pels quals ha generat un resultat determinat (és el cas de l'escassa transparència de les eines d'*aprenentatge profund*). Podríem dir que hi haurà una relació directament

proporcional entre la complexitat de l'eina d'IA utilitzada i les cauteles jurídiques amb què caldrà analitzar-la.

Al mateix temps, serà viable, en principi, la coexistència entre eines de mer aprofundiment en la digitalització i les estrictament d'IA. El que caldrà, llavors, és aconseguir una integració plenament satisfactòria de totes elles en el marc del sistema de gestió processal, per magnificar-ne la utilitat i l'eficiència global del sistema.

2.6. Algoritme, tasca, dades i model

De la mateixa manera que l'ús de l'expressió IA, per la seva pròpia generalitat, serà escassament útil pels propòsits de la recerca, també ho seran, en part, les nocions apuntades d'IA *supervisada* o *no supervisada* i d'aprenentatge *automàtic* o aprenentatge *profund*, perquè elles són, també, massa genèriques. Ens remeten a abstraccions utilitzades en el camp de la IA per facilitar la comprensió de la gran heterogeneïtat d'eines o recursos que s'hi utilitzen. Són útils, però necessitem una major comprensió tècnica, ja que, per exemple, dins de l'aprenentatge automatitzat (no profund) hi trobarem eines de diferents nivells de complexitat i transparència, que plantejaran, en correlació, diferents nivells i tipologies de problemàtiques jurídiques. Serà útil, però no suficient, per tant, afirmar que una determinada eina és, per exemple, *supervisada* (en la fase d'ajustament i prova de l'algoritme éssers humans etiqueten manualment les dades amb la resposta correcta) i d'aprenentatge *automatitzat* (l'algoritme aprèn per sí mateix i es modifica, sense intervenció humana directa, durant el seu ús sobre el terreny). Caldrà baixar fins al concret model decisori que hi ha darrera. Haurem de detallar quin tipus d'algoritme, dels molts existents, opera en el cor de l'eina. Haurem de tenir present, a més, que en alguns casos el mateix algoritme pot ser introduït, de manera indistinta, en una eina d'IA *supervisada* o en una *no supervisada* o en una d'aprenentatge merament automatitzat o en una altra d'aprenentatge estrictament *profund*. És a dir, l'elecció de l'algoritme no determina, necessàriament, el tipus d'eina d'IA que s'acabarà utilitzant.

Veiem, així, que, de fet, fins i tot saber el tipus d'algoritme que s'utilitza pot no ser suficient. Hi ha un nivell superior al qual haurem d'acudir per entendre l'eina i tenir la suficient perspectiva: ens haurem de desplaçar, ara cap a dalt, a les nocions de tasca, dades i model: quina és la tasca que hem de resoldre? De quines dades disposem? Quins són els models (instruments entesos en sentit ampli) que utilitzen eines d'IA i que

ens poden resoldre totalment o en part la tasca? Quin tipus d'algoritme (un o més d'un) utilitza cadascun d'aquests models? Sovint es disposarà de diferents models amb diferents algoritmes per poder abordar una determinada tasca i caldrà escollir-ne un. Els criteris d'elecció seran no només la utilitat o eficiència del model, sinó també, en el cas de l'Administració de Justícia, el grau de problemàtiques jurídiques que plantegi: genera resultats parcialment discriminatoris? Podem disposar d'una explicació dels motius pels quals s'ha obtingut un resultat determinat? Les respostes a aquestes preguntes podran obligar-nos, en certs casos, a optar per models menys precisos o eficients però també menys problemàtics.

Es deixa ja apuntada, en definitiva, la complexitat inherent a la comprensió de què significa, realment, una eina d'IA, sovint reduïda, de manera simplificada, a la noció d'algoritme. Cal una visió comprensiva, integral, de tot el procés d'ideació, prova, elecció i implementació de cadascuna de les eines[16].

2.7. Tecnologies potencialment complementàries amb la IA: *blockchain*, *ODR* i videoconferències

L'objecte de la recerca se centra en la IA aplicada a la justícia. Ja hem apuntat, però, que la frontera entre IA i altres mesures de digitalització pot ser força difusa en certs casos. Quelcom similar succeeix amb altres tecnologies que, sent clarament diferents de la IA, hi poden interactuar i complementar-s'hi. Per exemple, el *blockchain*, els sistemes d'*ODR* (resolució en línia de litigis) o les videoconferències en sentit ampli:

a) *Blockchain*: es tracta de sistemes descentralitzats i especialment segurs d'emmagatzemament i transmissió de la informació, que si bé tenen per aplicacions principals les cripto-monedes (*bitcoin*, *ether*, etc.) o els contractes intel·ligents (*smart contracts*), no podem descartar que puguin oferir, en un futur, alguna utilitat en la gestió del tràmit judicial.

[16] Dit això, si bé l'àmbit de l'aprenentatge automatitzat més apassionant és el d'aprenentatge profund (el més complex i opac, el que ens remet a la noció de *caixa negra*), segurament per moltes tasques judicials seran suficients i més adequades solucions d'aprenentatge automatitzat estricte amb tècniques més senzilles i transparents. Al mateix temps, però, l'aprenentatge profund és especialment adequat per problemes complexos que no podem descartar que siguin rellevants per l'objecte de la recerca, com ara el reconeixement del discurs o, especialment, el processament del llenguatge natural (*NLP*).

b) *ODR*: és una derivació moderna i tecnològica dels ADR (*alternative dispute resolution* o resolució de disputes per mètodes alternatius, no judicials). Es nodreixen d'institucions preexistents com la negociació, la mediació o l'arbitratge, però les traslladen a un entorn plenament tecnològic i digital. Sembla evident, per tant, que degudament integrades amb l'administració de justícia, amb plena continuïtat i fluïdesa, els *ODR* podrien aconseguir dinamitzar-la, en termes d'accés a la justícia i eficàcia. En cas de no obtenir-se'n cap resultat (cap acord o solució), seria la porta a la resolució judicial del conflicte o, fins i tot, en cas de sí haver-lo obtingut, la porta a l'execució judicial de l'acord extrajudicial. Una de les seves principals aportacions és la *asincronia*: ja no és necessari que totes les parts implicades acudeixin, físicament, a un lloc determinat un dia i hora també fixats, amb els condicionants logístics, laborals, personals i econòmics que això suposa, sinó que el tràmit va avançant de manera estratificada i cadascuna de les parts pot anar realitzant cada fase en el moment que millor li vagi (lògicament, dins d'uns límits organitzatius preestablerts). Desenvoluparem els *ODR* més endavant.

c) *Videoconferències*: és una tecnologia plenament coneguda. S'ha utilitzat en justícia des de fa temps i amb més intensitat arran de l'actual pandèmia. És indubtable, en qualsevol cas, que pot jugar un paper clau tant en el desenvolupament dels referits *ODR* com en la substanciació estricte del tràmit judicial. El repte serà augmentar-ne i assegurar-ne la qualitat (imatge, so, sincronia, etc.), la seguretat i les vies per poder identificar adequadament els participants.

Es tracta, en definitiva, de tres tecnologies alienes, des d'un punt de vista tècnic, a la IA però que poden ser rellevants en el camí per a integrar de manera harmonitzada el salt tecnològic al qual pot enfrontar-se properament la justícia.

3. Marc normatiu en matèria d'IA

3.1. Introducció

La regulació que expressament aborda la matèria de la IA és, a dia d'avui, realment escassa, tant a nivell europeu com intern. Atesa, però, la progressiva i fins a cert punt imparable implantació social i econòmica d'aquest tipus d'aplicacions, és probable que en els propers mesos o anys presenciem l'aprovació de normes que hi incideixin. Les que hi ha, de moment, són tangencials (protecció de dades, etc.) i no aborden de ple la problemàtica. Cal fer una referència en aquest punt a la Proposta de Reglament de regulació europea general sobre l'ús de la IA. Encara no ha estat, però, aprovat. Més endavant l'analitzarem (capítol 3.4). En qualsevol cas, fins al moment el que s'ha aprovat (de fet, profusament) són guies o principis *ètics* no vinculants. Tant per institucions públiques com privades. De caràcter general o referits a sectors específics, com el legal o judicial. La debilitat que presenten, però, aquests instruments, a més de la més òbvia (el fet de no ser vinculants), és que són excessivament genèrics i, segurament per aquest motiu, massa obvis. Pràcticament ningú no hi pot estar en contra: tothom estarà d'acord que la justícia algorítmica ha de ser *justa*, *no discriminatòria* i *transparent*. O que el fet que s'utilitzin eines d'IA no pot excloure que algú hagi de respondre o ser responsable dels seus eventuals resultats.

No es tracta de negar la rellevància d'aquests documents ètics *principalistes*, sinó de posar de manifest que, vista la seva profusió, caràcter no vinculant, obvietat, generalitat i inviabilitat d'aplicació per manca de concreció, sembla evident que cal superar aquesta fase i passar a una altra estrictament *normativa*, vinculant. Tot plegat, sense perjudici de partir de les reflexions i esquematitzacions que, de ben segur, poden extreure's amb profit de les guies de principis.

Per situar-nos en aquest apartat més *normatiu* de la recerca, podem adoptar una perspectiva àmplia i diferenciar tres blocs:

a) Guies de principis ètics, recomanacions i comunicacions, publicats o aprovats en matèria d'IA, no vinculants, amb una especial menció a aquelles que hagin emès institucions públiques i que es refereixin, específicament, a l'eventual implantació d'eines d'IA a l'àmbit judicial.

b) Normes vinculants en vigor que fan referència directa i expressa a la IA. Són poques i cal destacar-ne, especialment, les relatives a la protecció de dades en el marc de decisions automatitzades.

c) Projectes de normativa, encara no en vigor, expressament referides a la IA. En especial, la ja anunciada Proposta de Reglament pel qual s'estableixen normes harmonitzadores en matèria d'IA, COM(2021) 206 final, presentada per la Comissió Europea el 21 d'abril de 2021 2021/0106 (COD).

3.2. Guies de principis ètics, recomanacions, comunicacions i cartes

En comparació a les (escasses) normes vinculants en vigor sobre IA, és molt més extens el bloc de guies de principis, declaracions o recomanacions en aquesta matèria. Ja s'ha exposat que el seu interès és relatiu. Són expressió, més aviat, de la fase d'infantesa d'un sector nou que està a l'espera que se'l doti d'una normativa vinculant madura. La mateixa tasca d'intentar recopilar quins són aquests documents, què proclamen i en què coincideixen o es diferencien podria ser l'objecte d'una recerca. Aquí només n'esbossarem unes pinzellades sobre aspectes puntuals que poden ser d'interès per eventuals eines d'IA judicial. Veurem que pràcticament cap institució pública rellevant queda al marge de l'interès per la IA.

3.2.1. Parlament Europeu

Podem començar per la Resolució del Parlament Europeu de 2017 sobre les implicacions de la IA en matèria de drets fonamentals. La seva filosofia seria que si bé les oportunitats que ofereix el *BigData* són enormes, només podran ser aprofitades quan la confiança pública en aquestes tecnologies estigui assegurada per una estricta protecció dels drets fonamentals.

Més concretament, a l'octubre de 2020 el Parlament Europeu va adoptar diverses resolucions vinculades amb la IA: Recomanacions a la Comissió Europea sobre el marc relatiu als aspectes ètics de la IA, la robòtica i les tecnologies relacionades (2020/2012, INI), per a l'establiment d'una regulació de responsabilitat civil per a la IA (2020/2014, INI), sobre els drets de propietat intel·lectual en el desenvolupament de les tecnologies d'IA (2020/2015, INI) o sobre la IA en matèria criminal i el seu ús per la policia i les

autoritats judicials (2020/2016, INI). També ha establert un comitè especial sobre IA en l'era digital (2020/2684, RSO).

3.2.2. Comissió Europea

La Comissió Europea, ja a l'abril de 2018, va aprovar l'*Estratègia 'IA per Europa'* i *la Comunicació Cap a un espai digital europeu comú.* El 7 de desembre del mateix any va adoptar un *Pla coordinat sobre IA* (COM(2018), 795 final). Més endavant, a l'abril de 2019, ha adoptat la Comunicació *Construint confiança en una IA centrada en l'home* (COM(2019) 168 final (Annex I, Ref. No. 11)).

És d'especial interès el Grup Expert d'Alt Nivell sobre IA, constituït per la Comissió Europea i integrat per experts acadèmics, de la societat civil i de la indústria. Ha publicat les *Guies Ètiques per una IA fiable*, que es desdoblen en set requeriments clau que hauria de complir qualsevol sistema d'IA:

1) Agència i supervisió humanes.

2) Solidesa tècnica i seguretat.

3) Respecte a la privacitat i governança de dades.

4) Transparència.

5) Diversitat, no discriminació i justícia.

6) Benestar i prosperitat mediambiental i social.

7) Responsabilitat i rendició de comptes.

A partir de les Guies Ètiques s'ha creat un *checklist* destinat als desenvolupadors d'eines d'IA perquè ells mateixos s'assegurin que estan creant una IA *fiable*.

3.2.3. Llibre Blanc sobre IA

Igualment, el 19 de febrer de 2020 la mateixa Comissió Europea ha publicat el *Llibre Blanc sobre IA - Una aproximació europea a l'excel·lència i la confiança*[17]. Es tracta d'un document rellevant, que mereix una anàlisi més detallada. Tot i que el seu àmbit és general, i sembla centrar-se més en els impactes que la IA pugui tenir en el mercat interior, també aborda qüestions referides a la implantació de recursos d'IA a les administracions públiques que poden ser d'interès en el cas d'una eventual IA *judicial*.

Recorda el *Llibre Blanc* que els desenvolupadors i implementadors de la IA ja estan subjectes a la legislació europea en matèria de drets fonamentals (com ara la de protecció de dades, privacitat i no discriminació, entre altres) i de protecció de consumidors (seguretat dels productes, responsabilitat civil, etc.). A l'hora de diferenciar els diferents escenaris de risc, es remet el *Llibre Blanc* al sistema de regulació en cinc nivells proposat pel Comitè alemany sobre ètica en matèria de dades: des de la inexistència de regulació en el cas dels sistemes d'IA més innocus, fins a la prohibició absoluta en els casos més perillosos. Té en compte, per altra banda, que Dinamarca ha creat un segell d'ètica en les dades i Malta, un sistema voluntari de certificació de la IA. Per aquest motiu, hi hauria, segons el *Llibre Blanc*, un risc real de *fragmentació* del mercat interior, que posaria en risc els objectius de confiança i seguretat jurídica.

Caldria centrar la intervenció reguladora en les àrees en les quals hi hagi més probabilitats que sorgeixin riscos. I en cada àrea, diferenciar el risc generat per la concreta eina d'IA: per exemple, qualsevol sistema d'identificació biomètrica remota o de vigilància intrusiva serà *sensible*. Igualment, l'atenció sanitària seria, en general, un sector prioritari, però no tant, per exemple, el sistema que específicament es destini a l'assignació de cites. Un eventual error en el mateix no generarà, probablement, *riscos significatius* que requereixin una intervenció legislativa.

En aquest sentit, els danys que pot provocar la IA serien tant materials (danys personals o materials causats, per exemple, per un vehicle autònom) com immaterials (pèrdua de privacitat, limitacions de la llibertat d'expressió, discriminació en l'accés al treball, etc.). El marc normatiu hauria de centrar-se, per tant, en minimitzar aquests riscos. Un d'ells

[17] https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

seria, per exemple, la capacitat de la IA de rastrejar grans quantitats de dades, desanonimitzar-les i convertir en personals unes dades que, en si mateixes, no ho eren.

Recorda el *Llibre Blanc* que els riscos poden tenir diferents orígens. Es poden generar a la fase de disseny de la tecnologia, referir-se més aviat a la disponibilitat i qualitat de les dades o, simplement, haver-se generat en el mateix procediment d'aprenentatge automatitzat ulterior del sistema. Per això pot ser difícil detectar-los. Serà clau, per tant, que es puguin traçar retrospectivament totes les decisions potencialment problemàtiques que s'hagin pres durant tot el procés, des de la fase de disseny fins a la d'implantació i monitorització.

Caldrà establir obligacions, per tant, en matèria de conservació de registres precisos i de les metodologies sobre la programació dels algoritmes i les dades utilitzades per entrenar els sistemes de risc elevat. Hauria de ser possible saber quines són les característiques principals de les dades i de quina manera va ser escollit el conjunt de dades. I conèixer quines tècniques s'han utilitzat per construir, provar i validar el sistema. Tot plegat sense perjudici que s'adoptin, quan sigui necessari, les mesures necessàries per protegir les informacions confidencials, com ara els secrets comercials.

Aquesta disponibilitat i eventual subministrament d'informació serà la clau per aconseguir un nivell de transparència raonable i un ús responsable de la IA i l'única via que permetrà crear la necessària confiança i garanties de reparació: les persones que hagin pogut sofrir els danys haurien de disposar d'un accés efectiu a les proves necessàries per poder presentar, en el seu cas, una reclamació judicial. En cas contrari, la complexitat de les eines d'IA podria generar el risc que aquests perjudicats tinguin menys probabilitats d'obtenir una reparació efectiva en comparació a les situacions en les quals els danys hagin estat causats per tecnologies tradicionals.

Caldria, fins i tot, promoure pràctiques d'informació proactiva perquè el ciutadà sàpiga, primer, lògicament, quan està interactuant amb un sistema d'IA (i no amb un ésser humà), però també sobre com utilitzar aquestes eines, sobre les seves capacitats i limitacions, les condicions que s'han de donar perquè funcionin adequadament o el seu nivell d'exactitud esperat. El RGPD ja té algunes previsions en aquest sentit (arts. 13, apartat 2, lletra f), però podrien ser insuficients.

L'exigència de solidesa i exactitud dels sistemes d'IA ens remet a la idea de *fiabilitat*. S'han d'haver analitzat i abordat prèviament els riscos potencials. S'han de complir les especificitats tècniques per garantir la *reproductibilitat* dels resultats i la capacitat d'afrontar de manera adequada possibles erros o incoherències durant totes les fases del seu cicle de vida.

Aquí entraria la necessària *agència* o *supervisió humana*: en cap cas els sistemes d'IA poden debilitar l'autonomia humana. La IA ha de ser fiable, ètica i *antropocèntrica* i l'única manera d'aconseguir-ho és garantir una participació adequada de persones en les eines d'IA de risc elevat. El resultat del sistema d'IA no hauria de ser efectiu fins que un ésser humà no l'hagi revisat o validat posteriorment. El *Llibre Blanc* posa l'exemple de la denegació d'una sol·licitud de prestació de seguretat social o de targeta de crèdit.

En matèria ja estrictament judicial, recorda el *Llibre Blanc* que l'ús d'algoritmes per predir el nivell de risc de reincidència delictiva pot operar amb prejudicis racials o de gènere si els seus resultats són diferents, precisament, en funció de si es tracta d'homes o dones o nacionals o estrangers.

En definitiva, el *Llibre Blanc* és un dels documents més rellevants aprovats fins al moment, a l'espera que s'acabi aprovant la normativa europea definitiva i vinculant en matèria d'IA.

3.2.4. Consell de la Unió Europea, OCDE, UNESCO i IEEE

En el marc del Consell de la Unió Europea, que es mostra molt actiu en matèria d'IA i noves tecnologies, el Comitè de Ministres va crear un Comitè *Ad Hoc* sobre IA (*CAHAI*), que té per objecte l'estudi de la viabilitat i les possibilitats d'un marc normatiu per el desenvolupament, el disseny i les aplicacions d'IA, basat en els estàndards del Consell d'Europa en la protecció de drets fonamentals, la democràcia i l'estat de dret.

El Comitè de Ministres del Consell de la Unió Europea va aprovar el 8 d'abril de 2020 unes *Recomanacions sobre l'impacte dels sistemes algorítmics en els drets humans* (CM/Rec(2020)1). També, el 21 d'octubre de 2020, la *Carta de Drets Fonamentals en el context de la IA i el Canvi Digital*. Analitzarem els dos documents quan abordem amb més detall la dicotomia entre IA i drets fonamentals, al capítol 4.6.

Ens són d'un interès especial, per raons evidents, les Conclusions del Consell de Ministres de 8 d'octubre de 2020, *Accés a la justícia - Aprofitant les oportunitats de la digitalització*.

També és rellevant la *Carta Ètica Europea sobre l'Ús de la IA en els Sistemes Judicials i el seu entorn*, adoptada el desembre de 2018 per La Comissió Europea per l'Eficiència de la Justícia del Consell d'Europa (CEPEJ).

Més recentment, el desembre de 2020, el mateix CEPEJ ha emès un estudi sobre la possibilitat o la *Viabilitat de la Introducció d'un mecanisme de certificació d'eines i serveis d'IA en l'esfera de la Justícia i dels tribunals*.

Per la seva banda, també en el marc de la OECD s'han adoptat uns *Principis sobre IA*[18]*,* que recullen els valors del creixement inclusiu i sostenible, l'agència humana i justícia, la transparència i explicabilitat, la solidesa, la seguretat i la responsabilitat. La OECD ha creat, així mateix, un *Observatori sobre Polítiques d'IA*[19]*.*

En un nivell global, també la UNESCO ha iniciat un seguiment de les problemàtiques o les oportunitats que pot generar el desenvolupament de tècniques d'IA[20].

Del sector privat farem menció, per acabar, del destacat paper que ha jugat, fins el moment, l'IEEE (*Institute of Electrical and Electronics Engineers*)[21]. Ha realitzat o està en procés de realitzar projectes vinculats a qüestions d'ètica i discriminacions digitals.

3.3. Normes vinculants en vigor sobre la IA

3.3.1. Protecció de dades

3.3.1.1. Plantejament general

En l'ampli ventall d'àmbits afectats per l'actual procés de digitalització, el de la protecció de dades és el que de moment ha rebut més atenció per part dels reguladors. Així, des del maig de 2018 és d'aplicació a nivell de la UE el Reglament General de Protecció de

[18] https://www.oecd.ai/ai-principles.
[19] https://www.oecd.ai.
[20] https://en.unesco.org/artificial-intelligence.
[21] https://www.ieee.org.

Dades (RGPD)[22]. La protecció de dades és un camp molt ampli que genera una infinitat de problemàtiques. Només en donarem unes pinzellades sobre allò que ens pugui interessar per l'objecte de la recerca. En dos aspectes: per una banda, els principis generals sobre com s'han de tractar les dades. I, per l'altra, la nova regulació dels processos decisoris automatitzats que es basen en el tractament d'aquestes dades.

De fet, per a la IA, especialment pels models d'aprenentatge automatitzat, la possibilitat de recol·lectar i tractar grans quantitats de dades (personals i no personals) és la peça essencial del seu funcionament. Per això es diu a vegades que les dades (el *big data*) és el *petroli* de la IA. El que passa, però, és que és un petroli molt *sensible*, precisament perquè pot consistir en dades *personals*. Si aquest és el cas, passa a ser d'aplicació aquesta normativa, amb la qual es pretenen prevenir o mitigar possibles afectacions a la privacitat. En cas contrari, si les dades no són personals, ja no és d'aplicació[23].

Què implica, però, que sigui aplicable el RGPD? En primer lloc, que caldrà complir amb els principis clàssics de protecció de dades, que podríem sintetitzar així:

a) Licitud en la seva obtenció.

b) Transparència informativa: dret a saber quines dades són utilitzades i, eventualment, a instar la seva rectificació si són errònies o incompletes.

c) Minimització de les dades.

d) Limitació de la finalitat per a la qual són utilitzades.

e) Exactitud: les dades utilitzades han de ser completes.

[22] Reglament (UE) 2016/679 del Parlament Europeu i del Consell, de 27 d'abril de 2016, relatiu a la protecció de les persones físiques pel que fa al tractament de dades personals i la lliure circulació d'aquestes dades.
[23] De fet, és en sí mateix complexa determinar quan serà d'aplicació el RGPD, ja que no sempre és tan evident determinar si ens trobem davant de dades personals o si un sistema d'IA les ha utilitzat: les dades poden ser clarament personals, inicialment personals però *anonimitzades* o, fins i tot, anonimitzades i posteriorment *reidentificades* (quan es desfà el procés d'anonimització). Aquesta possibilitat s'aborda en estudis com el de Rocher et al. (2019). El ventall d'alternatives és ampli i no sempre serà fàcil arribar a saber quines dades han alimentat o entrenat un algoritme o quina proporció de les mateixes és, realment, personal, als efectes de considerar aplicables, o no, les previsions del RGPD.

f) Seguretat: restricció, per raons de confidencialitat, de qui té accés a les dades i eventual encriptació.

g) Limitació del temps de conservació: cal eliminar les dades tan bon punt deixin de ser necessàries.

h) Responsabilitat del responsable del seu tractament: el controlador de les dades ha de complir amb les exigències anteriors i la resta de les previsions legals.

La novetat d'aquesta normativa és que canvia la perspectiva de compliment d'aquests principis, en el sentit que serà primordialment el responsable de la gestió i tractament de les dades qui, de manera proactiva, haurà de respondre'n. Haurà d'adoptar, des d'un inici, un enfocament orientat als riscos i impactes potencials que la seva activitat pugui generar en termes de protecció de dades, amb la finalitat de minimitzar-los. Es tracta de la perspectiva *privacy by design*, que ja preveu, en la mateixa estructura del sistema, a més dels estudis previs i continuats d'impacte en la privacitat, mecanismes de protecció que actuen per defecte. El desafiament és, però, complir amb aquests estàndards no només en el moment inicial de la implantació del sistema de gestió de les dades, sinó en tot el seu cicle de funcionament i evolució, tasca que pot ser especialment complexa en les eines d'IA d'aprenentatge automàtic, que, per definició, es van modificant elles mateixes durant aquest cicle de funcionament.

Aquesta normativa europea general no té un correlatiu als EEUU, on la protecció de dades es tracta a partir d'una gran diversitat de normes estatals i federals que en regulen aspectes concrets. Una figura essencial d'aquesta normativa és, evidentment, la del *consentiment*: el model europeu aposta per un consentiment explícit, informat i limitat a una finalitat determinada. Per contra, als EEUU el consentiment pot ser implícit: es pot deduir del fet de no haver-se demanat explícitament que no es tractin les dades.

3.3.1.2. Protecció de dades i IA judicial: art. 22 RGPD

Arribats a aquest punt, cal indicar que de tota la regulació del RGPD ens interessa especialment, als efectes de la present recerca, l'expressa regulació que efectua el seu art. 22 sobre els drets que té qualsevol ciutadà o ciutadana davant d'una eventual decisió automatitzada que ha fet ús de les seves dades.

Aquest precepte regula les decisions individuals automatitzades. I preveu que tot interessat tindrà dret a no ser objecte d'una *decisió basada únicament en el tractament automatitzat* (inclosa l'elaboració de perfils) que produeixi efectes jurídics en ell o l'afecti de manera *significativa*.

A continuació venen, però, una sèrie d'excepcions:

a) Si es disposa del consentiment explícit de l'interessat.

b) Si la decisió és necessària per la celebració o execució d'un contracte entre l'interessat i el responsable del tractament de les dades.

c) Si la decisió està autoritzada pel dret de la UE o dels Estats Membres.

En els dos primers casos (consentiment i exigència contractual), ja no es necessitarà una expressa autorització legal (més enllà de la que ofereix el mateix RGPD), però caldrà assegurar-se, com a mínim, que l'interessat té dret a obtenir una *intervenció humana* per part del responsable, a expressar el seu punt de vista i a *impugnar* la decisió. Fora d'aquests dos supòsits, caldrà una norma (diferent del RGPD) que doni cobertura legal a les decisions automatitzades, però ja no s'exigeixen, en tot cas, aquestes garanties mínimes de contestabilitat. Dependrà, lògicament, del règim que estableixi cada norma que autoritzi aquest tipus de decisions.

Amb caràcter general, les decisions automatitzades no es podran basar en les *categories especials*, protegides o sensibles, de dades personals que preveu l'art. 9.1 RGPD. És a dir, les que revelin l'origen ètnic o racial, les opinions polítiques, les conviccions religioses o filosòfiques, la afiliació sindical, les dades genètiques o biomètriques dirigides a identificar inequívocament a una persona física, la salut, la vida sexual o la orientació sexual. Hi haurà, però, dues excepcions: de nou, el consentiment exprés (excepte que la normativa de la UE o de l'Estat Membre prohibeixi l'ús fins i tot en cas de consentiment exprés); o quan sigui necessari per raons d'interès públic essencial, de manera proporcional a l'objectiu perseguit i amb respecte a l'essència del dret de protecció de dades.

Dit això, cal fer tres matisos rellevants:

a) El mateix RGPD exclou la seva aplicació al tractament de les dades personals en l'àmbit penal (art. 2.2.d). Hem d'acudir, aquí, a la Directiva (UE) 2016/680 del Parlament Europeu, que, per cert, no ha estat transposada a Espanya fins la LO 7/2021, de 26 de maig.

b) Fins i tot deixant de banda la jurisdicció penal, la regulació del RGPD sembla escassament determinant per la seva eventual aplicació a les altres jurisdiccions. No és una regulació que hagi estat concebuda per referir-se a l'actuació del poder judicial, sinó, més aviat, al sector privat o àmbits d'actuació públics més ordinaris o no tant especials com el judicial. Més aviat sembla remetre's, en aquests casos, a la regulació que cada Estat Membre en pugui fer. A més, per la mateixa naturalesa d'una eventual IA judicial, no es pot concebre aquesta sense una prèvia regulació legal expressa, aprovada per l'autoritat competent, que en aquesta matèria és precisament cada Estat Membre.

c) Sembla evident, per últim, que si s'arriben a aprovar instruments d'IA judicial, els estàndards europeus que exigeix el TEDH respecte del dret de defensa i a un judici just imposaran, pel cas d'hipotètiques decisions judicials automatitzades, el reconeixement no només de certs drets d'informació i d'impugnació (com els previstos al RGPD), sinó d'un conjunt potent de garanties processals que clarament desborden el marc conceptual de la normativa de protecció de dades.

És per tots aquests motius que el marc normatiu de la protecció de dades, si bé és rellevant als efectes d'una eventual IA judicial, només ho és en un sentit parcial o fragmentari. És clarament insuficient. Caldrà una actuació normativa molt més profunda. I serà aquesta la que fixi les condicions d'ús i les garanties que necessàriament hauran d'acompanyar una eventual IA judicial.

3.3.1.3. Decisions automatitzades no judicials

Cal diferenciar, és clar, entre una eventual IA *judicial* (que acabem d'analitzar i que indubtablement requerirà una normativa expressa detallada) i les reclamacions judicials que es puguin articular contra processos automatitzats generats en el sector privat o, fins i tot, en el públic no judicial. En el cas del sector privat, si hi ha consentiment o un context contractual determinat, potser ja no serà necessari que existeixi una normativa que autoritzi expressament les decisions automatitzades, però sí que s'hauran de complir les

exigències d'informació generals establertes pel RGPD, precisament per fer possible que l'afectat pugui, primer, formular directament les queixes o peticions de rectificació corresponents i, després, arribat el cas, formular una demanda judicial si considera que les seves dades no han estat tractades amb compliment dels principis i previsions de la normativa. Per poder preparar i fonamentar amb proves aquestes demandes judicials, caldrà que rebi, ja des d'un bon inici, una informació precisa i clara sobre els seus drets, sobre com s'han tractat, realment, les seves dades i com han incidit en la decisió automatitzada de què es tracti. Només així podrà valorar, *ex ante*, la conveniència d'instar alguna rectificació o d'articular algun tipus d'impugnació. Especialment, podrà estar en condicions de provar els fets rellevants. En cas contrari, si no disposa de tota aquesta informació (i de manera accessible, senzilla i sense necessitat d'assistència legal), podria veure's indirectament afectat, precisament, el seu dret de defensa.

Aquí topem, però, amb un problema rellevant: l'art. 22 RGPD no preveu, expressament, un dret a l'*explicació algorítmica*. Únicament el dret a obtenir una *intervenció humana*, a donar la opinió i a impugnar la decisió. Res més. Es dona el cas, força il·lustratiu, que inicialment sí que estava previst introduir algun tipus de dret a una *explicació*, fins al punt que l'exposició de motius (ap. 71) del RGPD fa referència al dret a rebre una explicació de la decisió. Com que aquest apartat (l'exposició de motius) no és vinculant, la insuficiència de les garanties del RGPD són més que evidents. És cert, però, que l'art. 13.2.f) preveu la obligació d'informar si s'està utilitzant un sistema de decisió automatitzada i, si és el cas, d'oferir, com a mínim, una informació *significativa* sobre la seva *lògica* i les *conseqüències* que pot tenir per l'interessat. Els termes són, en qualsevol cas, força ambigus i queden oberts a diverses interpretacions. Podria entendre's que es refereixen, més aviat, al funcionament general del sistema i no tant als factors concrets que han determinat la decisió automatitzada. O que la referència a les *conseqüències* sí que implica, per contra, una obligació concreta d'explicar la decisió individual automatitzada. Cal tenir en compte, però, que, a més, l'art. 12.7 RGPD preveu que aquesta informació podrà ser transmesa en combinació amb icones normalitzades intel·ligibles i senzilles d'entendre. I aquesta possibilitat, més enllà de poder facilitar la comprensió, és probable que condueixi a explicacions estandarditzades escassament precises que no ofereixin una explicació real i efectiva dels factors que han estat rellevants en una concreta decisió automatitzada, la que de fet acaba afectant l'interessat.

Veiem, en definitiva, que, abans de presentar una demanda amb la qual impugnar una determinada decisió automatitzada, l'afectat podrà sol·licitar aquesta *explicació algorítmica* al responsable del tractament de les dades i aquest l'haurà de donar. Però el nivell de detall de l'explicació dependrà de la interpretació que fem del règim del RGPD. I ja hem vist que hi ha marge interpretatiu. Si la informació facilitada és escassa o insuficient, l'afectat podrà igualment presentar la demanda i intentar obtenir una major *concreció explicativa* ja en el marc del procediment judicial, eventualitat que generarà les seves específiques problemàtiques jurídiques i processals. Sembla evident, però, que presentar una demanda amb aquest grau d'*indefinició probatòria* inicial (amb risc de desestimació i imposició de costes) no respon als millors estàndards en termes de dret de defensa.

3.3.1.4. Implicacions del RGPD en una eventual IA judicial

En aquest apartat ens preguntarem com incidiria una regulació com la dels vigents arts. 12, 13 i 22 RGPD en una hipotètica implantació d'eines estrictes d'IA judicial. És evident que es tracta d'una normativa insuficient i és probable, a més, que properament entrin en vigor altres normatives més exhaustives sobre aquesta qüestió (en especial, el Reglament europeu sobre IA, que és analitzat al capítol 3.4). Però el cert és que encara estan en fase de tràmit, per la qual cosa en una recerca com la present i en un apartat com aquest (marc normatiu), no es pot deixar de banda l'anàlisi indicada.

Començarem amb un matís molt rellevant: l'art. 22 RGPD regula les decisions basades *únicament* en el tractament automatitzat de dades. Sí, *únicament*. Sembla referir-se, per tant, a processos decisoris *completament automatitzats*, en els quals no hi hagi cap tipus d'intervenció humana, ni durant el mateix ni amb posterioritat. Aquest adverbi podria, per tant, deixar fora de la regulació molts supòsits de tractament automatitzat de dades personals. Per exemple, quan hi ha un control posterior o quan un ésser humà pren una decisió però fent ús d'un *input* algorítmic. Les situacions poden ser molt ambigües[24]. Aquesta ambigüitat podria propiciar que els actors privats introdueixin en els seus processos de decisió intervencions humanes aparents o fictícies, no substancials (merament formals), amb la finalitat d'eludir l'estricte règim d'exigències i garanties que

[24] Sara Wachter, que ha ideat les *explicacions contrafàctiques* que abordarem al capítol 4.3.5, posa l'exemple d'una oferta de treball en la qual hi ha 3.000 candidats i un algoritme els ordena prèviament en un rànquing, però és una persona qui decideix qui convidar a l'entrevista. En aquest cas ja no hi hauria un procés de decisió *completament* automatitzat.

preveu la normativa. El risc és evident. I també ho és que la introducció de l'adverbi *únicament* no pot haver estat innocent. Segurament s'ha introduït per limitar els casos en els quals les companyies estaran obligades a oferir informació sobre el seu *codi*. La propietat intel·lectual i els secrets comercials estan darrera d'aquest adverbi. També, segurament, el risc que amb un excés d'explicacions obligatòries es pogués veure afectada la privacitat d'altres persones.

Si traslladem aquesta problemàtica a l'àmbit de la recerca, el de l'Administració de Justícia, hem de tornar a la noció de *col·laboració* entre la IA judicial i les autoritats o funcionaris judicials: sembla evident, atesos els delicats interessos i drets que hi ha en joc en aquest sector, que difícilment seran acceptables eines d'IA íntegrament automatitzades, sense cap tipus d'agència, control o supervisió humana. Potser només seria imaginable en el context de mers tràmits procedimentals sense incidència significativa en els drets dels intervinents. Per tant, aquesta idea de *col·laboració* semblaria no encaixar amb la previsió de l'art. 22 RGPD: la decisió ja no estaria *completament* automatitzada, sinó només de manera *parcial*.

De fet, en l'àmbit de la jurisdicció penal, la ja referida Directiva (UE) 2016/680 (que s'aplica enlloc del RGPD), opta per una solució diferent i en el seu art. 2.2 precisa que s'aplica en casos de tractament de dades total o *parcialment* automatitzat. Així ho recull també la LO 7/2021, de 26 de maig, que la transposa (art. 2.1). Cal matisar, però, que el seu art. 14 només prohibeix les decisions basades *únicament* en un tractament automatitzat, i fins i tot preveu excepcions. És a dir, per fixar l'àmbit d'aplicació de la norma, hi inclou els casos d'automatització *parcial*, però només prohibeix expressament els d'automatització *completa*.

Podem dir, per tant, que quan s'ha d'abordat aquesta matèria des de l'àmbit judicial penal, es tendeix a no restringir l'àmbit d'aplicació de la normativa als casos d'automatització *completa*. S'ha inclòs els d'automatització *parcial*. I és lògic, atesa la naturalesa de la funció jurisdiccional. Tot plegat sense perjudici que només es prohibeixin, de manera expressa, certs casos d'automatització *completa*.

Caldrà veure si passa el mateix amb la futura normativa que abasti la totalitat de la funció jurisdiccional. De moment disposem, només, d'una proposta de normativa. Anticipant-nos a l'anàlisi més detallada que en farem al capítol 3.4, podem avançar que la Proposta

de Reglament pel qual s'estableixen normes harmonitzadores en matèria d'IA, presentada per la Comissió Europea el 21 d'abril de 2021 i que es troba actualment en tràmit d'elaboració, no exigeix que l'automatització sigui *completa* ni, per altra banda, tampoc no precisa que pugui ser *parcial*. Més aviat prescindeix de la noció d'automatització i opta (segurament en una decisió encertada, per la seva major claredat) per fer un llistat d'usos d'IA d'*alt risc* (entre ells, l'Administració de Justícia), sense entrar en si integren processos d'automatització *complets* o *parcials*. En definitiva, supera o obvia la problemàtica que hem exposat fins ara. Cal recordar, però, que es tracta, de moment, només d'un projecte de normativa.

3.3.1.5. Normativa espanyola

A nivell espanyol, s'ha aprovat la LO 3/2018, de 5 de desembre, de protecció de dades personals i garantia dels drets digitals. En destacarem les següents previsions:

a) *Tractament judicial de les dades (art. 2.4)*: el tractament de les dades realitzat amb ocasió de la tramitació pels òrgans judicials dels processos dels quals en siguin competents, i el realitzat dins de la gestió de la oficina judicial, es regirà pel RGPD, aquesta LO i la LOPJ. La recent LO 7/2021, de 26 de maig, ha afegit una referència al règim jurídic del tractament de dades realitzat en ocasió de la tramitació dels processos pel Ministeri Fiscal o la Oficina Fiscal.

b) *Transparència i informació (art. 11)*: l'afectat i interessat podrà obtenir informació relativa, com a mínim, a la identitat del responsable del tractament, la finalitat perseguida i la possibilitat d'exercir, entre altres, els drets de l'art. 22 RGPD (dret d'impugnació). Si les dades es destinen a l'elaboració de perfils, caldrà informar de manera expressa del dret a oposar-se a l'adopció de decisions individuals automatitzades que produeixin efectes jurídics sobre ell o l'afectin de manera significativament similar, si es donen els pressupòsits de l'art. 22 RGPD.

c) *Dret d'oposició* en cas de decisió individual automatitzada (art. 18): es remet a l'art. 22 RGPD, ja analitzat.

Veiem, així doncs, que la LO 3/2018 es remet en bona mesura al RGPD i que aquest, per la seva banda, no aborda una regulació expressa de l'impacte de la nova normativa

sobre protecció de dades en l'àmbit judicial. Si tenim en compte, a més, que la remissió a la LOPJ és, de moment, estèril, atès que encara no ha estat adaptada o actualitzada al RGPD, podem concloure que, a nivell judicial, la regulació de la protecció de dades i dels processos automatitzats és si no nul·la, sí molt escassa.

En efecte, tot i que amb la reforma de la LOPJ de 2015 es va introduir un apartat relatiu a la protecció de les dades de caràcter personal en l'àmbit de l'administració de justícia, es va fer des de la perspectiva tradicional de la mera digitalització de les dades. No s'aborden qüestions referides a l'automatització en el tractament de les dades. Això exigeix, per tant, interpretar la regulació de la LOPJ a la llum del RGPD i de la LO 3/2018. No es tracta, però, d'una tasca fàcil, ja que ens trobem, més aviat, davant d'un buit legal.

Cal reiterar, però, que el mateix RGPD exclou la seva aplicació al tractament de les dades personals en l'àmbit penal (art. 2.2.d). Hem d'acudir, aquí, a la Directiva (UE) 2016/680 del Parlament Europeu, transposada a Espanya per la ja referida LO 7/2021, de 26 de maig, de protecció de les dades personals tractades per finalitats de prevenció, detecció, investigació i enjudiciament d'infraccions penals i d'execució de sancions penals. Aquesta norma prohibeix en el seu art. 14, com hem vist, les decisions basades *únicament* en un tractament automatitzat, inclosa l'elaboració de perfils, que produeixin efectes jurídics negatius per l'interessat o que l'afectin significativament, excepte previsió expressa per una norma amb rang de llei o europea, que haurà de recollir el dret a obtenir la intervenció humana en el procés de revisió de la decisió. Es prohibeix igualment l'elaboració de perfils que generin una discriminació sobre la base de categories especials de dades personals.

3.3.2. Automatització de decisions judicials més enllà del tractament de les dades

Veiem, en definitiva, que la LOPJ només aborda, de moment, qüestions referides al tractament de les dades des de la perspectiva de la protecció de la privacitat, però no, encara, aspectes relatius a la possible automatització del tractament d'aquestes dades. Només ho fa, i molt genèricament (i, també, de manera desfasada), la Llei 18/2011, de 5 de juliol, reguladora de l'ús de les tecnologies de la informació i la comunicació en l'administració de justícia. En concret, a l'art. 19 (amb referència als sistemes de signatura electrònica en actuacions judicials automatitzades), a l'art. 35 (amb referència a un sistema automatitzat de remissió i gestió telemàtica de les comunicacions edictals)

i, especialment, a l'art. 42, relatiu, ja sí en general, a les actuacions judicials automatitzades. El que passa, però, és que només preveu que en cas que existeixin, caldrà, abans, que el Comitè tècnic estatal de l'administració judicial electrònica estableixi la definició de les especificacions, programari, manteniment, supervisió i control de qualitat i, en el seu cas, auditoria del sistema d'informació i del seu codi font. Els sistemes hauran d'incloure els indicadors de la gestió que estableixi la Comissió Nacional de l'Estadística Judicial i el mateix Comitè Tècnic, cadascun en l'àmbit de les seves competències. Per últim, aquesta mateixa norma defineix en el seu annex què hem d'entendre per *actuació judicial automatitzada*: aquella produïda per un sistema d'informació adequadament programat sense necessitat d'intervenció d'una persona física en cada cas singular. Inclou la producció d'actes de tràmit o resolutoris dels procediments, així com també mers actes de comunicació.

3.3.3. Conclusió: una regulació insufficient

Podem concloure, en definitiva, que ha estat, de moment, principalment en el marc de l'aprovació de normes sobre protecció de dades que s'han introduït algunes previsions sobre com s'han de tractar les dades personals judicials. Per altra banda, no queda clar que les previsions, també generals, sobre els processos de decisió automatitzada que acompanyen aquestes normes sobre protecció de dades siguin estrictament aplicables a l'àmbit judicial. En última instància, el mateix art. 22 RGPD es remet, en aquesta matèria, a allò que pugin preveure les normes de la UE o dels Estats Membres. En el cas espanyol, les previsions de la Llei 18/2011 que acabem d'analitzar són clarament insuficients. Sembla evident que fins que no hi hagi una norma estatal interna que expressament, i de manera detallada, reguli els processos automatitzats de presa de decisions judicials, aquests no seran viables. Això és així perquè l'art. 22 RGPD exigeix, com és lògic, aquesta previsió legal. I, a més, perquè en aquesta matèria hi ha una clara reserva de llei: només la llei la pot introduir.

Podríem plantejar-nos si la reserva és merament de llei (de normes processals ordinàries o de normativa relativa a l'ús de les tecnologies en l'administració de justícia) o fins i tot de llei orgànica (si s'entén que la implantació d'eines d'IA judicial afecta directament el mateix estatut dels jutges i jutgesses i les condicions d'independència i imparcialitat amb les quals han d'exercir la funció jurisdiccional). Sigui quina sigui l'alternativa, hi ha un reserva de llei latent. No pot abordar-se la implantació d'aquestes tecnologies,

lògicament, a través d'una mera pràctica judicial, però tampoc, per exemple, pel mer exercici de competències autonòmiques en matèria de recursos informàtics o tecnològics de la oficina judicial. És cert, però, que aquestes limitacions competencials o de previsió normativa es poden matisar en funció del grau d'intervenció o d'*intensitat processal* de les eventuals eines d'IA judicial que es pretengui implantar en un futur en el procés de presa de decisió. Si és molt baix, fragmentari o col·lateral, es podria plantejar la seva adopció sobre la base, per exemple, de meres competències sobre recursos informàtics. La resposta ha de ser, doncs, particularitzada. Al mateix temps, però, atesa la novetat d'aquest tipus de tecnologia i dels interrogants que inevitablement genera la seva hipotètica implementació en l'àmbit judicial, sembla raonable esperar i exigir en tot cas una prèvia regulació legal expressa. La prudència hauria de guiar aquest procés.

3.4. Normativa en tràmit: proposta de Reglament d'harmonització en matèria d'IA

A continuació s'analitzarà una proposta de normativa molt rellevant als efectes de la present recerca. Però cal remarcar que es tracta, només, d'una proposta. Per tant, no només no està en vigor, sinó que no se sap si ho arribarà a estar o, en cas positiu, si serà en els actuals termes o en uns altres modificats. Això no obstant, la importància d'aquest projecte de normativa imposa la seva anàlisi amb un cert detall. No hem d'oblidar, de fet, que precisament la normativa que ja està en vigor és, com hem vist, clarament insuficient. Ni que en l'àmbit de la IA és especialment desitjada i esperada una normativa imperativa que vagi més enllà de previsions genèriques i estèrils.

Es tracta de la Proposta de Reglament pel qual s'estableixen normes harmonitzadores en matèria d'IA, COM(2021) 206 final, presentada per la Comissió Europea el 21 d'abril de 2021 2021/0106 (COD). Aquest Reglament parteix d'una tripartició estructural dels usos de la IA: els prohibits, els d'alt risc i la resta. Són els següents:

a) Els usos directament *prohibits*: entre altres, l'ús de tècniques subliminals, els que aprofiten la vulnerabilitat de col·lectius específics o alguns sistemes d'identificació biomètrica remota.

b) Els usos d'*alt risc*, permesos però que han de complir obligatòriament amb una sèrie d'exigències (que veurem més endavant): aquí hi trobarem els usos que afecten

l'Administració de Justícia. La Proposta és força abstracta, de moment, en la concreció d'aquests usos: serien els sistemes d'IA d'ajuda a l'autoritat judicial en la investigació i interpretació de fets i de la llei, així com en l'aplicació de la llei a un conjunt concret de fets. També inclou com d'alt risc altres usos de la IA referits a l'*aplicació* de la llei si bé amb innegables implicacions judicials potencials: sistemes per determinar el risc de comissió d'infraccions penals o de reincidència o el risc per les potencials víctimes dels delictes; l'ús de polígrafs o d'eines per detectar l'estat emocional d'una persona física; l'avaluació de la fiabilitat de les proves durant una investigació o l'enjudiciament d'infraccions penals; la predicció de la freqüència o reiteració d'una infracció penal real o potencial en base a l'elaboració de perfils de persones físiques; o la gestió de la migració, l'asil i el control fronterer (detecció de l'estat emocional, avaluació del risc per la seguretat o la salut i verificació de l'autenticitat dels documents).

c) Resta d'usos, pels quals es preveu que la Comissió de la UE i els Estats Membres fomentaran i facilitaran l'elaboració de codis de conducta per promoure l'aplicació voluntària dels requisits previstos pels sistemes d'alt risc.

Veiem que, als efectes de la recerca, els usos que poden afectar l'Administració de Justícia no estarien prohibits però es qualifiquen d'alt risc. Hauran de complir, per tant, amb les exigències que preveu la proposta. Atesa la seva rellevància, les abordem a continuació amb un cert detall:

1) Implantar un sistema de gestió dels riscos, que haurà de consistir en un procés iteratiu continu en funcionament durant tot el cicle de vida del sistema, amb actualitzacions sistemàtiques i periòdiques. Es preveu, però, que es puguin considerar acceptables els riscos residuals associats a cada risc.

2) Si l'eina d'IA implica l'ús de tècniques d'*entrenament* de models amb dades, caldrà que aquestes siguin pertinents, representatives, exemptes d'errors i completes i que compleixin certs criteris de qualitat, tant les dades d'entrenament, les de validació i les de prova. Això afectarà tant a l'elecció del disseny, a la recopilació de les dades, al seu tractament i preparació, a la seva anotació i etiquetatge, a la depuració, enriquiment i agregació, o a l'examen de la seva disponibilitat, quantitat i adequació. També, als seus possibles biaixos.

3) Documentació tècnica adequada i registres automàtics d'esdeveniments mentre estan en funcionament, que permetin una *traçabilitat* de tot el cicle de vida de funcionament del sistema.

4) Tenir un nivell de transparència suficient perquè els usuaris puguin interpretar-los i utilitzar-los correctament.

5) Hauran de poder ser *vigilats* de manera efectiva per persones físiques i disposar d'una interfície *home-màquina* adequada, amb la finalitat de prevenir o reduir al mínim els riscos per a la salut, la seguretat o els drets fonamentals. Aquesta interfície haurà de permetre que s'entenguin completament les capacitats i limitacions del sistema i controlar el seu funcionament, de manera que es puguin detectar indicis d'anomalies, problemes de funcionament i comportaments inesperats, i donar-hi una solució tan aviat com sigui possible. També haurà de permetre a l'usuari ser conscient de la possible tendència a confiar automàticament o en excés en la informació de sortida generada per un sistema d'IA (l'anomenat «*biaix d'automatització*»), especialment en els sistemes que tenen per finalitat aportar informació o recomanacions perquè persones físiques adoptin una decisió. Per altra banda, la interfície ha de permetre la interpretació correcta de la informació de sortida i decidir, en qualsevol situació, no utilitzar el sistema d'IA o desestimar, invalidar o revertir la informació de sortida que generi. I, per últim, ha de permetre intervenir en el funcionament del sistema o interrompre'l accionant un botó específicament destinat a aquesta finalitat o amb un procediment similar.

6) Hauran d'aconseguir durant tot el seu cicle de vida un nivell adequat de precisió, solidesa i ciberseguretat. Hauran de ser resistents als errors, falles o incoherències que puguin sorgir en el mateix sistema o en l'entorn en el qual operin. Aquest objectiu es podrà aconseguir amb recursos de redundància tècnica (còpies de seguretat o plans de prevenció contra errors).

7) Si el sistema és dels que continua aprenent després de la seva posada en servei, haurà de desenvolupar-se de tal manera que els possibles biaixos en la informació de sortida deguts a l'ús d'aquesta com a dades d'entrada en futures operacions («*bucle de retroalimentació*») es reverteixin degudament amb mesures de mitigació.

8) S'hauran de sotmetre al procediment d'avaluació de la conformitat i hauran de conservar els arxius de registre que generin automàticament en el termini i les condicions previstes legalment.

9) Si hi ha motius per creure que el sistema ja implementat no compleix amb els requisits establerts, s'hauran d'implementar immediatament les mesures correctores que siguin necessàries.

10) Caldrà establir un sistema de seguiment posterior que recopili, documenti i analitzi de manera activa i sistemàtica dades pertinents sobre el funcionament del sistema per poder avaluar el compliment dels requisits exigits.

Per altra banda, la proposta preveu una sèrie d'exigències més generals que ja no tenen en compte la distinció entre sistemes d'IA d'alt risc i els que no ho són. Així, tots els sistemes que estiguin destinats a interactuar amb persones físiques hauran de permetre que aquestes persones estiguin informades que estan interactuant amb un sistema d'IA, excepte en aquelles situacions en què això resulti evident per les circumstàncies i el context de l'ús o quan el sistema estigui autoritzat per la llei amb la finalitat de detecció, prevenció, investigació o enjudiciament d'infraccions penals.

4. Principis de la IA judicial

4.1. Qualitat i eficiència

Partint de les premisses proposades i de l'escàs marc normatiu existent, a dia d'avui, en matèria d'IA, ens hem de preguntar, ara, quins principis haurien de governar una hipotètica implantació d'eines d'IA a l'Administració de Justícia. Podríem diferenciar els principis *logístics* o de *gestió*, per una banda, i els principis *materials* o de *fons* (justícia, no discriminació, etc.), per l'altra.

Entre els primers, força obvis, hi hauria el que serien més aviat les finalitats o objectius a perseguir: el principi de preservació o millora de la *qualitat* de la justícia i el de potenciació de l'*eficiència* amb la qual es tramiten els procediments judicial.

Quant a la *qualitat*, caldrà assegurar-se que les eines d'IA copsin adequadament la *complexitat real* de les situacions fàctiques i jurídiques que processen. Aquestes situacions podran referir-se, excepcionalment, a la presa de decisió final (sempre amb agència humana posterior) o, més probablement, al mer tràmit judicial o al mer *suport* a la presa de decisió final. Però caldrà, sempre, que la imprescindible *abstracció* amb la qual opera qualsevol eina d'IA (detecció de *certes* pautes; anàlisi, només, de *certs* factors o paràmetres rellevants, etc.) no impliqui, de fet, una *simplificació excessiva*, inassumible, del context fàctic o jurídic que processa. Ha de tenir, sempre, un grau de *precisió* suficient. En cas contrari, ens trobaríem davant d'una eina que no ofereix les suficients garanties de *qualitat*. Seria, de fet, un factor d'*empobriment jurídic* que justificaria descartar l'eina[25].

En segon lloc, el principi d'*eficiència* ens remet a l'eterna pretensió d'*accelerar*, de fer més ràpida, la prestació del servei públic que és la justícia. Una IA judicial de qualitat en termes de precisió però lenta ja no superarà el primer filtre dels principis de gestió. És inimaginable, de fet (per absurd), que la introducció d'eines d'IA a la justícia impliqui un empitjorament en aquest aspecte. A la mateixa noció d'*automatització* li és inherent la idea d'*estalvi* en termes d'intervenció humana i, en correlació, de guany en temps. Per

[25] S'observa, de fet, la proximitat entre aquesta exigència de qualitat i el principi material o de fons que denominarem justícia de la IA judicial, referit, preferentment, al seu caràcter no discriminatori: una eina d'AI judicial de baixa qualitat (per simplificar en excés i no tenir en compte tots els paràmetres fàctics o jurídics rellevants) podrà ser, alhora, injusta per discriminar injustificadament, excloent-la, una part de la realitat jurídica que hauria de ser tinguda en compte.

tant, habitualment sempre obtindrem, amb les eines d'IA, una millora en termes de celeritat. Del que es tractarà, però, és d'avaluar si aquest guany és suficient per *compensar* altres possibles implicacions (potser menys desitjades) que podrà generar l'eina, sigui en termes de precisió (qualitat) de la resposta judicial o de riscos associats a la inexistència o menor intensitat del control humà. No podrem menystenir, per exemple, la inconscient tendència humana (també dels professionals que treballen als jutjats) de donar per bones les propostes o resultats que ofereixen de manera automatitzada sistemes aparentment objectius i neutres (l'anomenat *biaix d'automatització*). Aquest serà un dels molts riscs o contrafactors que haurà de ponderar-se en cada cas, de manera individualitzada, per valorar si *val la pena*, o no, implementar una eina d'IA determinada que genera, aparentment, una certa *acceleració* del procediment[26].

4.2. Justícia i no discriminació

4.2.1. Intangibilitat dels atributs de l'Administració de Justícia

Sembla una obvietat afirmar que una justícia que utilitzi eines d'IA ha de ser *justa*, no *discriminar*, preservar (no vulnerar) els drets fonamentals (entre ells el de defensa), ser *transparent*, *explicable*, *imparcial* i *responsable* respecte dels efectes (potencialment danyosos) dels seus actes. Si exigim aquestes notes a l'administració *humana* de la justícia, no deixarem de fer-ho respecte d'una administració *artificial*, en tot o en part, de la justícia. De fet, si deixéssim d'exigir aquestes notes, ja no ens trobaríem, probablement, davant d'un acte d'estricte administració de justícia, sinó més aviat davant d'un acte (potser molt eficient i ràpid) de *gestió administrativa* d'un conflicte. Com que aquesta no és una opció (l'administració de justícia ha de seguir sent *Administració de Justícia* encara que li introduïm algunes eines d'IA), hem d'exigir i comprovar o monitoritzar permanentment que les notes que la caracteritzen romanguin intactes. Es

[26] Pot donar-se, també, un guany *en eficiència* que no impliqui, directament, un guany *en temps*: imaginem una eina d'IA judicial que dona suport al tribunal en l'anàlisi dels escrits i els documents de la causa, seleccionant-ne els rellevants o els seus fragments més significatius. O, fins i tot, extraient-ne de manera automatitzada (i conjunta) tot el text perquè sigui el tribunal qui posteriorment faci la selecció, però ja disposant de la transcripció completa. Podria ser que el processament informàtic necessari per executar aquesta funció i la fase posterior de selecció del text rellevant executada pel tribunal sumin, junts, un temps similar al que podria trigar el tribunal efectuant tot el procediment de manera manual (llegint els documents i escrivint directament el resum). Però fins i tot en aquesta hipòtesi, el desgast físic i mental del tribunal, en cas d'emprar la IA, seria segurament menor, ja que s'hauria estalviat tasques mecàniques (tecleig) que no necessàriament s'han de fer manualment. És en aquest sentit, de menor desgast, que podem dir que, tot i no haver-se guanyat *temps*, sí que s'ha guanyat en *eficiència*.

tracta d'una qüestió innegociable, políticament i constitucional, que no requereix major argumentació.

4.2.2. Podem saber, realment, quan la justícia, humana o artificial, és justa o injusta?

Fins aquí, una obvietat. Si baixem el nivell d'abstracció en l'anàlisi, veurem, però, que la qüestió es complica: lògicament exigim a la justícia humana que sigui justa, però realment ho és? Sempre o només a vegades? Disposem, de fet, de paràmetres objectius per identificar els casos de resolucions justes o injustes, més enllà dels supòsits flagrants? Si en tenim, els podem reduir a una màxima o n'hem de construir diverses? Si el cas és el segon, aquestes màximes són sempre compatibles entre sí o ens poden portar a conclusions diferents?

Una resposta honesta a aquestes qüestions, gairebé de filosofia del dret, ens hauria de portar, probablement, a admetre la impossibilitat o extrema dificultat de poder arribar a conclusions fermes en la majoria de casos. Per tant, acte seguit ens hem de fer una altra pregunta: si no estem en condicions, normalment, d'identificar quan un acte d'administració *humana* de justícia és just o injust, podrem fer-ho, per contra, respecte d'un acte d'administració *artificial* de justícia?

Aquí la resposta podria ser doble. Simultàniament negativa i positiva: no sembla, certament, que es produeixi un canvi radical en la impossibilitat general d'identificar, destriar, les resolucions *justes* i les *injustes*. El caràcter inaprehensible d'aquesta qüestió es manté pràcticament intacte. Al mateix temps, però, és innegable que les eines d'IA judicial, per la seva pròpia naturalesa i manera d'operar, poden permetre, fins i tot amb més precisió que en el cas de la justícia *humana*, la identificació de certs factors de potencial *injustícia*. Principalment, pel que fa als *riscos de discriminació*: si una eina d'IA té per nucli un algoritme entrenat a partir de dades, el resultat que generi dependrà d'una sèrie de factors vinculats amb aquestes dades que poden posar de manifest eventuals discriminacions.

Veiem, per tant, que el debat sobre la necessària *justícia* de la IA acaba quedant reduït a la també necessària *prohibició de discriminar* (per raons de gènere, de raça, socials, etc.) aplicable a qualsevol acte de justícia, sigui humana o artificial. En definitiva, caldrà deixar de banda les abstractes filosofies sobre la *justícia* i centrar-nos en el que, de fet,

és el nucli de la mateixa justícia: tractar igual allò que és igual i tractar diferent allò que és diferent només quan hi hagi una raó acceptable per aquest tractament desigual. Quan no es compleix aquesta màxima, ens podrem trobar davant d'una *discriminació* jurídicament inassumible. I, per tant, potser, davant d'una resolució *injusta*.

4.2.3. La discriminació, eina natural del dret

Abans d'entrar en el detall sobre els riscos de discriminació inherents a les eines d'IA judicial, cal partir d'una premissa òbvia però que no se sol tenir massa present en aquesta qüestió: la *discriminació*, entesa com a acte de distinció de les situacions, és, de fet, l'operació habitual, natural, per mitjà de la qual actua el dret en tots els seus nivells. En primer lloc, quan el legislador decideix legislar, o no legislar, un sector determinat, està, ja, discriminant-lo respecte dels sectors no legislats o legislats, respectivament. Per això les normes acostumen a començar el seu articulat fixant el seu àmbit d'aplicació (material i subjectiu). Igualment, quan es legisla un sector, la regulació consistirà bàsicament en la successiva *distinció* de diferents situacions (que es troben dins del seu àmbit d'aplicació) a les quals se'ls aplicarà un règim jurídic específic o un altre. I, així, successivament. Per tant, legislar consisteix, en gran mesura, en discriminar. I les discriminacions legislatives es basaran (i pretendran justificar-se) en la presència, o absència, de certs factors que el legislador ha considerat rellevants (per exemple, si el contracte s'ha negociat, o no; si una lesió s'ubica a l'espatlla o al colze, etc.).

Si ens desplacem al nivell judicial, la funció jurisdiccional consistirà, igualment, en la consideració i valoració de successives discriminacions, amb la particularitat que, segons el principi de legalitat, haurà de partir, necessàriament, de les discriminacions ja realitzades pel legislador. El que passa sovint, però, és que aquestes discriminacions *legislatives* poden no ser suficients per resoldre el cas concret, quan aquest presenta unes especificitats rellevants en les quals no va pensar el legislador: la instància judicial pot haver-ne d'afegir unes altres[27].

[27] Per exemple, el legislador espanyol, a diferència d'altres Estats Membres de la UE, va *ometre* incloure, en la normativa de consum i per delimitar l'àmbit del control d'abusivitat, la distinció entre clàusules contractuals de consum que regulen l'objecte principal del contracte (el preu) i les que afecten a qüestions accessòries. Hi havia un buit legal que generava dubtes interpretatius. Finalment, han actuat els tribunals acollint aquesta discriminació, per concloure que només respecte del segon tipus de clàusula és viable el control d'abusivitat. S'ha introduït una discriminació *judicial*, que pot implicar que finalment no s'acabin aplicant els efectes previstos inicialment a la norma (la nul·litat, per abusiva, de la clàusula que fixa el preu).

Veiem, en definitiva, que la discriminació és la via natural, habitual, amb la qual s'actua jurídicament tant a nivell legislatiu com judicial. D'on ve, per tant, la nota pejorativa que acostumem a vincular a l'acte de discriminar? Ve, lògicament, de l'acceptabilitat, social i jurídica, o manca d'acceptabilitat, dels motius o les raons en base a les quals es decideix discriminar les situacions. L'acte de discriminar és, en tot cas, una realitat ineludible, inherent, a l'instrument jurídic. Hem de discriminar. Necessitem discriminar. Ho fem habitualment. I la clau per determinar si una discriminació és acceptable, o no, rau en les *raons* que utilitzem per fer-la. Això ens porta a les *discriminacions prohibides* i als *biaixos* (conscients o inconscients).

4.2.4. Discriminacions legals prohibides: atributs protegits

Partint de la *naturalitat* amb la qual hem d'acudir a la tasca de *discriminar jurídicament,* el primer que detectem és que el mateix legislador constitucional ha identificat i prohibit certes discriminacions que en cap cas podran ser admeses i que si es materialitzen, ja sigui pel legislador ordinari o per la instància judicial, podran implicar la nul·litat, per inconstitucionalitat, de la norma o de la resolució judicial. Es tracta de les conegudes discriminacions per raó de raça, sexe, religió, opinió o altres circumstàncies personals o socials (art. 14 CE). Serien uns *atributs protegits*. Una tal discriminació reflectiria un *biaix inacceptable* en el tractament jurídic. Cal recordar, però, que en certs àmbits poden ser admeses mesures de *discriminació positiva* o de tractament diferenciat, fins i tot per alguna de les raons indicades, sense que necessàriament es generi la seva nul·litat[28].

Fins aquí ens mouríem dins de les discriminacions i biaixos deliberats i conscients. Expressos. Siguin legislatius o judicials. Acceptables (vàlids) o no assumibles (potencialment invàlids). A la recerca li interessen més, però, els biaixos no deliberats ni conscients. Els no previstos expressament. Els que emergeixen soterradament del funcionament mateix de les eines d'IA judicial. S'ha de dir, però, que d'una manera semblant a la justícia humana, l'algorítmica opera necessàriament (no pot fer-ho d'una altra manera) amb discriminacions: la finalitat principal dels algoritmes de l'aprenentatge automatitzat és categoritzar, classificar i separar. Només d'aquesta manera poden generar els resultats que els demanem. Ens interessa, per tant, analitzar com operen,

[28] Pensem, per exemple, en el cas més extrem, que implica un tractament penal diferent dels delictes de violència contra la dona, en els quals al mateix fet pot aplicar-se-li una pena diferent en funció de si el protagonitza un home o una dona. O, en altres àmbits, les mesures per afavorir persones amb capacitats reduïdes, persones d'edat avançada o el col·lectiu LGTBI.

realment, aquests mecanismes de discriminació algorítmica. La tasca no és fàcil, entre altres motius perquè en l'àmbit de l'aprenentatge automatitzat hi operen les anomenades *caixes negres.*

4.2.5. Biaixos judicials humans inconscients

Parlarem, ara, de discriminacions (de tractaments jurídics diferenciats) potencialment inacceptables per basar-se en raons no compatibles amb l'ordenament jurídic. I intentarem mostrar com aquests problemes de discriminació es concreten de manera diferent en la justícia *humana*, per una banda, i la justícia *artificial*, per l'altra.

Així, en una resolució judicial *humana* potencialment discriminatòria, els factors que generen aquesta discriminació poden ser explícits, conscients i deliberats (com els que s'han examinat anteriorment) o, per contra, inconscients i no deliberadament explicitats en el text de la resolució, però no menys reals que els primers. És en aquest sentit que pels tractaments jurídics deliberadament (conscientment) diferenciats sembla més adequat el terme *discriminació*, mentre que el de *biaix* ho seria pels que són inconscients i no explícits. Aquests segons els haurem de buscar per via indirecta, implícita. En les manifestacions que no es fan o en les ponderacions que no s'aborden. Probablement, en les pressuposicions de les quals es parteix injustificadament. El cert, però, és que sovint no podrem constatar, identificar, de manera precisa i objectiva, on es localitza el factor desencadenant del biaix. Haurem de construir una contra argumentació o un relat alternatiu del cas, que podrà convèncer, o no, però que no serà l'oficialment vinculant (ho serà el de la resolució judicial que precisament s'està criticant). De fet, si fos possible (metafòricament) introduir-nos en la ment del jutge o jutgessa que discrimina injustificadament, seria una recerca probablement estèril, ja que en molts casos les discriminacions humanes operen inconscientment.

Per contra, si ens desplacem al camp de les eines d'IA judicial, augmenten les possibilitats d'una detecció relativament objectiva dels factors *discriminants*. Ho analitzarem a continuació.

4.2.6. Biaix algorítmic: quan la discriminació està a les mateixes dades

El factor *discriminant* més evident en una eina d'IA són les mateixes dades: les utilitzem per *entrenar* l'algoritme i no són il·limitades o infinites. No abasten la totalitat de dades

existents sobre una determinada matèria i en tots els temps. Per contra, són *necessàriament parcials*: seran aquelles que estiguin *disponibles* i que s'hagin obtingut en *determinades* situacions, respecte *determinats* ciutadans o ciutadanes i en un període de temps també *delimitat*. Són una *selecció* de la realitat sobre la qual es vol operar. I no poden ser una altra cosa. Per tant, discriminen la realitat rellevant per a la tasca que es vol dur a terme.

Com que discriminen i no poden deixar de fer-ho, es pot donar el cas que les dades recopilades i utilitzades per crear l'eina d'IA no reflecteixin adequadament la *diversitat* i *complexitat* de la realitat. Per exemple, perquè es disposa de més dades obtingudes de persones de pell blanca en comparació a persones de pell fosca o no tan blanca. O al revés. O, en un altre cas, perquè les dades s'hagin obtingut principalment de sectors de la població de menys recursos perquè són ells, precisament, els que interactuen amb més freqüència amb l'administració i generen, en conseqüència, les dades. Els supòsits poden ser molt diversos[29]. El que ara interessa posar de manifest és que, en principi, aquesta eventualitat (en bona part inevitable) no és, per si mateixa, un problema. Aquest apareix quan el *model algorítmic* (en sentit ampli, més enllà de les dades) que es vol utilitzar per una determinada tasca, i que s'alimenta i *entrena*, precisament, amb aquestes dades, està configurat de tal manera que genera uns resultats que *reflecteixen* el biaix ja contingut a les dades. Serà llavors quan emergiran els problemes de discriminació. Es tracta, per raons evidents, d'una problemàtica molt rellevant per a una hipotètica IA judicial: la prohibició de discriminar de manera no justificada afecta a tots els actors socials, privats i públics, però la màxima negació del valor constitucional de la prohibició de discriminar (per exemple, per raó de sexe o de raça) la causarà una resolució judicial esbiaixada. Ens trobem, per tant, davant d'una línia vermella que cal tenir perfectament identificada i que, en cas de dubte, ens ha de fer descartar la implantació d'eines d'IA judicial potencialment discriminatòries.

4.2.7. Afectació directa o indirecta dels atributs protegits

Un algoritme pot processar directament com a dada rellevant un atribut protegit, com ara el sexe o la raça. En aquest cas, que no acostuma a ser freqüent, el tractament discriminatori és fàcil de localitzar. Cal tenir present, de fet, que l'art. 21 de la Carta de

[29] A l'apartat 6.5.2, que aborda un dels reptes de l'aprenentatge automatitzat, s'entra amb més detall en la problemàtica de les dades no suficientment representatives.

Drets Fonamentals de la UE prohibeix la discriminació basada en diversos paràmetres: sexe, raça, color, ètnia, origen social, atributs genètics, llengua, religió, creences, opinió, pertinença a una minoria nacional, propietat, naixement, discapacitat, edat, orientació sexual o nacionalitat. Quan aquests paràmetres es refereixen a una persona concreta, passen a ser dades personals i reben la protecció legal corresponent, la qual no impossibilita, però, de manera absoluta, que puguin ser utilitzades en eines d'IA. En limita i condiciona l'ús, com hem vist, però no l'impedeix en tots els casos. Per tant, aquestes situacions de *discriminacions algorítmiques directes*, tot i que infreqüents, poden donar-se. I quan ho facin, és molt probable que incorrin en algun tipus d'il·legalitat.

La discriminació, però, acostuma a operar d'una manera més subtil i no necessàriament conscient. És més difícil de detectar i denunciar. És més indirecta. Pot ser útil, a tal efecte, tenir identificats els anomenats *indicadors* (d'etnicitat, etc.) per poder detectar (i corregir) possibles discriminacions en els resultats de l'eina d'IA. Expliquem-ho[30]: sovint els atributs protegits (sexe, raça, etc.), que no estan expressament programats per ser tinguts en compte, tenen, però, una elevada *correlació* amb altres factors o *variables* rellevants, en sí mateixos no directament problemàtics i que sí que han estan expressament programats. La *variable* programada, que serà sotmesa a un tractament estadístic, seria un *indicador* de l'atribut protegit: d'aquesta manera, i per una via indirecta, els resultats obtinguts (l'*output* de l'eina d'IA) poden acabar reflectint un tractament discriminatori *com si el paràmetre realment programat hagués estat el relatiu a l'atribut protegit*.

La discriminació es produiria, per exemple, quan els resultats de l'eina d'IA són sistemàticament diferents respecte d'un grup determinat: quan un membre d'una determinada minoria ètnica té menys opcions de ser escollit en una feina com a conseqüència del fet d'haver estat entrenat l'algoritme amb dades que, respecte d'aquesta minoria, porten a un menor rendiment o un pitjor resultat. A vegades les causes de la discriminació poden ser sorprenents: tenir en compte l'alçada pot ser un indicador de gènere i valorar el codi postal, de classe social o origen ètnic.

[30] Per fer-ho, acudirem a l'estudi "#BigData: Discrimination in data-supported decision making", FRA European Union Agency for Fundamental Rights, 2018, p. 9.

4.2.8. Casos reals de discriminació algorítmica

Si ens desplacem a exemples reals de biaixos algorítmics posats de manifest o abordats judicialment, el més conegut (i citat) és el del programari *Correctional Offender Management Profiling for Alternative Sanctions* (*COMPAS*), utilitzat als jutjats de diversos Estats dels EEUU per predir el risc de fuga o de reincidència a l'hora de decidir sobre la presó preventiva o la llibertat provisional. Uns periodistes de *ProPública* (Larson et al., 2016) van investigar els possibles biaixos racials subjacents a aquesta eina d'IA. En concret, van partir de les bases de dades de reincidència real durant dos anys en una jurisdicció determinada i la van comparar amb la gradació de risc que oferia l'eina COMPAS. La conclusió va ser, per una banda, que els investigats de raça blanca eren erròniament catalogats de *baix risc* amb més freqüència que els de raça negre i, per l'altra, que hi havia més probabilitats que els segons fossin erròniament catalogats d'*alt risc*[31].

En un altre nivell, el policial, podem trobar usos de la IA susceptibles de generar biaixos discriminatoris. Principalment, en l'anomenada *policia predictiva*. Aquestes eines s'utilitzen com a criteris complementaris (dels tradicionals, més intuïtius o basats en l'experiència) per decidir com distribuir els recursos policials existents per a la prevenció i persecució dels delictes. Es basen en les dades estadístiques del passat. Hi ha, per tant, el risc evident de reproduir les pràctiques discriminatòries ja existents. De ser-ne una *càmera de ressonància*: si per criteris humans, s'ha acostumat a acudir més sovint a certs barris (*variable* de localització merament física i *neutre*, però que pot ser un *indicador* de classe social o, en alguns casos, d'etnicitat), serà més probable que s'hagi detectat allà, i no a altres llocs, una taxa més elevada de criminalitat. El programari processarà, per tant, aquesta tendència inicialment *humana* i probablement aconsellarà destinar més efectius en aquell barri. Els riscos d'estigmatització algorítmica són, per tant, evidents.

Per altra banda, si hi ha un elevat percentatge de denúncia i persecució de certes tipologies de crims (delictes contra la propietat) i més baix en altres (violència de gènere, delictes sexuals, etc.), això s'acabarà reflectint en les *propostes logístiques* que ens faci

[31] Cal dir, però, que la mateixa crítica periodística a COMPAS ha estat objecte, a la seva vegada, de discussió per altres investigadors o acadèmics que apuntaven la possibilitat d'aplicar models diferents d'escrutini dels resultats d'aquesta eina predictiva que no reflectien l'indicat biaix. Aquesta disparitat de conclusions posa de manifest, de fet, que no es disposa, de moment, d'estàndards d'avaluació d'aquest tipus d'eines.

l'eina d'IA. No cal oblidar, tampoc, els possibles errors humans en la introducció policial de les dades o en els mateixos biaixos que puguin tenir els agents que les introdueixen.

Canviant de tipologia d'eina d'IA, però seguint en l'àmbit policial, farem referència a les de reconeixement facial: en un cas relatiu a la possible il·legalitat del sistema utilitzat per la policia de Gales del Sud, el tribunal de segona instància, a més de constatar que hi havia massa discreció dels agents a l'hora de decidir qui podia ser introduït a la llista de persones buscades o el lloc on s'ubicaria la tecnologia, va tenir en compte que no s'havia investigat suficientment si el programari exhibia biaix de gènere o raça[32].

4.2.9. Discriminació per associació: privacitat dels grups nous

Per ser subjecte passiu d'una discriminació algorítmica de grup un no ha de formar part, necessàriament, del grup en qüestió. La inclusió en el tractament pot produir-se, simplement, per associació. En aquest cas, el problema pot ser la falta de transparència sobre com actua l'eina d'IA: com puc saber amb qui m'està agrupant l'algoritme? Què està inferint l'algoritme sobre mi? Qui més està en el grup? Com els afecta l'agrupació?

Això ens pot portar a la noció de *privacitat de grup*: els grups nous són diferents als tradicionals grups vulnerables. La IA fa les agrupacions sobre una base nova que no podem anticipar i això genera uns desafiaments desconeguts fins ara. Caldrà pensar en noves formes creatives per abordar les llacunes legals i conceptuals que tenim sobre aquest tipus d'eventualitats. Potser ja existeixen nous grups efectivament discriminats pels algoritmes dels quals encara no en som conscients. Als quals encara no hi hem posat nom.

4.2.10. Es poden detectar els biaixos algorítmics?

Com passa amb els biaixos humans, els algorítmics no són fàcils de detectar. Els mètodes per fer-ho (si és que existeixen) són, però, molt diferents. En parlem en aquest apartat.

El primer problema amb el qual ens podem trobar, especialment quan es tracta de productes privats, és el fet de no tenir, per raons de propietat intel·lectual o secret

[32] Tribunal d'Apel·lació, *Regne Unit (Bridges) v. CC South Wales*, [2020] EWCA Civ 1058, 11 d'agost de 2020.

comercial, accés complet al programari i codi de l'algoritme. El cert, però, és que fins i tot en cas de tenir-lo, podria ser insuficient, perquè les discriminacions no estaran, probablement, expressament codificades. Es posaran de manifest en el funcionament en temps real de l'algoritme. Sí que es podrien extreure més conclusions si es té accés, per exemple, a les dades utilitzades per entrenar l'algoritme. Però aquí hi hauria, potser, problemes de privacitat. És evident, en qualsevol cas, que caldrà buscar mecanismes legals que permetin una auditoria real i efectiva, sense restriccions, però objectiva, externa i imparcial, dels productes d'IA i que al mateix temps preservi els drets i interessos comercials implicats. La Proposta de Reglament pel qual s'estableixen normes harmonitzadores en matèria d'IA, presentada per la Comissió Europea i del que ja n'hem parlat al capítol 3.4, va en aquesta direcció.

La problemàtica dels biaixos algorítmics i la seva eventual i necessària detecció es complica quan adquirim consciència que poden operar en diferents moments:

1) El risc de discriminació racial i de gènere pot generar-se, ja, amb caràcter general, en el mer fet que el sector professional de la IA (des de l'àmbit estrictament tecnològic fins al de la recerca acadèmica) està integrat en un elevat percentatge per homes blancs. Aquesta sola circumstància genera un risc de biaix cultural que pot acabar introduint-se, incrustant-se, fins i tot de manera inconscient, en els productes d'IA. Per aquest motiu són rellevants, també, el criteris de conformació dels equips que elaboren aquestes eines.

2) En el disseny del model o de l'algoritme.

3) En la selecció de les dades.

4) En la *preparació* o *neteja* de les dades (qualitat de les dades).

Veiem, de fet, que es tracta de biaixos potencials completament diferents. I és evident que en cada cas els mecanismes per detectar-los i eradicar-los seran, també, heterogenis. Cal tenir, en definitiva, una visió àmplia i de conjunt, per no incórrer en el risc de creure's que s'han eliminat tots els biaixos pel sol fet d'haver-se actuat en una de les seves múltiples fonts potencials.

Si ens situem en els biaixos *intra-algorítmics*, serà més difícil la seva detecció en funció de quin grau de complexitat i opacitat tingui l'eina. En el *pitjor* dels casos, si es tracta de xarxes neuronals d'*aprenentatge profund* (que treballen amb diferents nivells ocults de relacions i combinacions de múltiples atributs de les dades), pot ser molt difícil determinar si s'està produint un tractament discriminatori sobre la base d'un atribut concret. Aquesta dificultat no exclou, lògicament, la necessitat d'auditar el funcionament de l'algoritme. I si l'auditoria no és possible, haurà de ser tingut en compte com un argument per descartar l'eina en qüestió, especialment en l'àmbit de l'Administració de Justícia (si bé sempre en funció del grau de *sensibilitat* de la tasca judicial de què es tracti).

Un mètode, *analògic*, per constatar el tractament discriminatori laboral consistiria, per exemple, en realitzar l'experiment *real* d'enviar dues sol·licituds de feina fictícies i idèntiques en les quals únicament variï la pertinença a un grup o minoria determinada (canviant el nom del sol·licitant) i rebre'n un resultat diferent. Si es tracta d'eines d'IA en plataformes en línia, es podria imitar el mètode analògic i generar multitud de perfils que seran enviats repetidament i de manera aleatòria per veure si el resultat és diferent en funció de característiques que puguin reflectir algun tipus de discriminació[33].

Més interès tenen els mètodes d'extracció d'informació sobre quines dades han contribuït més en la resposta oferta per l'algoritme. Per exemple, per saber si la diferència en els ingressos ha estat determinant per a la denegació d'un préstec (eventualitat que podria ser raonable). O, per contra, si ho ha estat la pertinença a una minoria determinada (una eventualitat més problemàtica). S'ha proposat, amb aquesta finalitat, el mètode de les *explicacions contrafàctiques*, que no necessiten *obrir* la *caixa negre* en què pot consistir l'algoritme (Wachter et al., 2017). Aquest mètode ens porta a la noció de *sensibilitat algorítmica*, segons la qual, sense disposar d'un coneixement detallat de com ha funcionat l'algoritme (probablement perquè no està disponible), sí que podem constatar, empíricament, si un factor és, o no, determinant, removent-lo, precisament, de les dades utilitzades, i tornant-lo a introduir, per veure si el resultat que se n'obté és diferent.

[33] En un experiment relatiu a fins a quin punt la detecció de gènere funciona igual de bé en diferents grups, es va constatar que els resultats eren pitjors en les dones de pell més fosca en comparació amb els homes de pell blanca. L'estudi "#BigData: Discrimination in data-supported decision making", FRA, 2018, es remet, a la p.6, al treball "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', Proceedings of Machine Learning Research", Buolamwini and Gebru, 2018, Vol. 81, pp. 115, Conference on Fairness, Accountability and Transparency.

4.2.11. Pot ser la IA una eina de reducció dels biaixos humans?

Fins aquí hem analitzat els riscos de les discriminacions algorítmiques *encobertes*. No hem de descartar, però, que es pugui fer un ús *positiu* dels algoritmes per contribuir a detectar i reduir els biaixos de certes actuacions *humanes*: l'anàlisi algorítmica de les dades podria, hipotèticament (i paradoxalment), reduir els biaixos en l'actuació policial, massa centrada, per exemple, en la persecució de certs delictes i no tant en d'altres, com els financers o de *collaret blanc*. O, també, revisar possibles biaixos presents en les antigues bases de dades i dels quals no n'érem prou conscients.

En definitiva, de la mateixa manera que no podem caure en innocents acceptacions acrítiques de les noves eines, emmirallats per la seva aparent objectivitat, neutralitat i asèpsia científica, tampoc no hem de menystenir possibles usos de millora humana que se'n puguin derivar. Tornem, de nou, a la noció de *col·laboració* entre agència humana i IA, de la qual ja n'hem parlat i a la qual tornarem sovint, probablement perquè ella és la resposta, el punt de raonable equilibri, al qual hem de tendir.

Acabem, per tant, amb una nova paradoxa: si bé les eines d'IA poden augmentar, en certs casos, els riscos de discriminació (existents, també, cal recordar, en la justícia humana), al mateix temps podrien oferir alguns mecanismes de detecció objectiva de les seves causes i recursos per mitigar-los o eliminar-los.

4.3. Transparència, interpretabilitat i explicabilitat algorítmiques

4.3.1. Una major precisió algorítmica implica, necessàriament, una menor transparència?

Acostuma a passar, especialment en les xarxes neuronals, que una major *precisió* en els resultats obtinguts ve acompanyada d'una menor *transparència*, *intel·ligibilitat*, *explicabilitat* o *interpretabilitat*. Un exemple interessant per entendre aquesta *contrapartida* o correlació invertida (*trade-off*) el tenim en els dos estudis de justícia predictiva que es van realitzar sobre la jurisprudència del TEDH. El model proposat l'any 2016 va obtenir una precisió del 79 % i el segon, del 82 % (Hildebrandt, 2019). En principi, seria més interessant i fiable el segon, però mentre que el primer permet saber quines característiques o atributs es relacionaven amb la variable objectiu que es perseguia (conèixer quin seria el resultat de la sentència), el segon, que utilitza xarxes neuronals,

opera com una *caixa negre* i ens amaga els atributs potencialment rellevants, tant en les seves interrelacions com en els seus pesos respectius.

Sovint ens trobarem davant d'aquest dilema: si disposem de diverses opcions potencialment vàlides, donarem preferència a aquella que utilitzi un algoritme més precís o a aquella que, tot i tenir una menor precisió, ofereixi una major comprensió de com funciona?[34] No cal dir que aquest dilema és especialment rellevant en l'àmbit legal i judicial, en els quals ha d'haver-hi, sempre, una justificació jurídica de qualsevol decisió. I aquesta exigència sembla incompatible amb l'escenari de *ceguera motivacional* a la qual ens condueixen aparentment certes xarxes neuronals.

En definitiva, l'absència absoluta o gairebé absoluta d'*interpretabilitat* o *explicabilitat* d'un model algorítmic presenta inconvenients tan obvis per a la seva eventual aplicació judicial que gairebé no requereixen major explicació. Caldria diferenciar, és clar, la tasca judicial a la qual es pretén aplicar l'eina: en la mesura que s'aproximi a la de presa de decisió final o de fons, els inconvenients es transformaran en obstacles insalvables. Per contra, si la tasca és merament de tràmit processal o, fins i tot, de mer suport parcial per a la decisió final, llavors l'horitzó no és, potser (només potser), tan *negre*. Però abans de seguir ens hem de fer una pregunta prèvia l'obvietat de la qual potser ha provocat la seva omissió: en què consisteixen, exactament, la *transparència*, la *interpretabilitat* o l'*explicabilitat* d'una eina d'IA. I ens l'hem de fer, aquesta pregunta, perquè la resposta, i les implicacions derivades, són lluny de ser senzilles i monolítiques.

4.3.2. Què vol dir, exactament, que un algoritme és transparent, interpretable o explicable?

Ens endinsem, ara, en un terreny d'una certa densitat però que cal recórrer si volem obtenir un mínim coneixement *real* de com funcionen les eines d'IA i què implica la seva relativa opacitat.

Hem de recordar, però, que quan busquem explicacions sobre com funciona *en general* o com ha funcionat *en concret* un model, no es tracta de trobar les *justificacions* materials

[34] Adverteix, però, Hildebrandt (2019) que tampoc no hem d'assumir, necessàriament, que les xarxes neuronals ofereixen sempre una elevada precisió, ja que precisament la seva manca d'*interpretabilitat* fa més difícil comprovar-ho.

(i menys els arguments jurídics) que ens han portat a un *output* determinat. No els hem de buscar perquè, senzillament, no existeixen: els models predictius no són capaços de cap tipus de raonament, sinó que només estableixen correlacions, principalment estadístiques, entre factors rellevants[35]. Per tant, l'objectiu de la *interpretabilitat* o *explicabilitat* algorítmiques és poder arribar a conèixer, en el màxim detall possible, com s'estableixen, *en general*, aquestes correlacions o com s'han calibrat en el cas *concret* per generar un resultat determinat. És un objectiu limitat, però important. I, en tot cas, l'únic plausible.

La conveniència de disposar d'un cert grau d'*interpretabilitat algorítmica* d'una eventual eina d'IA judicial és, com hem vist, una exigència inherent a l'estat de dret, que no és compatible amb la opacitat absoluta de certes caixes negres. També és necessària per aconseguir un grau acceptable de confiança ciutadana en l'Administració de Justícia, ja que difícilment s'obtindrà si l'actuació judicial no és mínimament transparent i es basa en hipotètiques correlacions estadístiques no jurídiques que, a més, no podem conèixer. Això ens aproparia, més aviat, a una manera d'actuar arbitrària, expressament prohibida per la Constitució (art. 9.3 CE). Hem de poder conèixer, per exemple, com certs atributs influeixen en una determinada predicció i assegurar-nos, per contra, que no són determinants, directament o indirecta, alguns atributs protegits, com la raça i el gènere.

Fins aquí tots estaríem d'acord. Però immediatament hem de buscar mecanismes concrets per determinar, en cada cas, si es disposa d'un nivell raonable d'interpretabilitat algorítmica. Començarem per certs recursos que, si bé ens aporten informació sobre el model, són clarament insuficients en termes d'interpretabilitat material.

4.3.3. Instruments informatius que no generen interpretabilitat

No seran suficients algunes mètriques o representacions gràfiques tradicionals ja disponibles. Per exemple, la fixació, en la fase posterior a l'entrenament, del grau de precisió de l'algoritme (*accuracy*). Es tracta, simplement, d'una fase del procés de creació del model, gestionada pels seus mateixos creadors i en la qual opera, simplement, una

[35] Lipton (2018) ens recorda que els sistemes d'aprenentatge automatitzat no *saben* per què a un input determinat cal donar-li una etiqueta concreta. No estan concebuts per reflectir relacions causals. Només *saben* que certs inputs estan correlacionats amb aquella etiqueta. Ens posa l'exemple d'una base de dades en la qual els únics objectes taronges són pilotes de bàsquet: un model de classificació d'imatges potser aprendrà a classificar tots els objectes taronges com a pilotes de bàsquet.

fórmula matemàtica que determina un determinat percentatge de precisió (encert), del 0 al 100%: després d'haver-se *entrenat* el model amb dades *etiquetades* amb la resposta correcta, es *prova* el model ja ajustat amb dades diferents també etiquetades per veure si encerta la resposta. Aquesta fase de determinació de la precisió del model és imprescindible per a la seva posada en marxa: si es constata un grau de precisió baix, es descartarà i se'n buscarà un altre. Però no aporta informació externa, de contrast mínimament *imparcial*, que ens permeti entendre com funciona el model o que ens generi, amb un mínim de perspectiva, confiança. Si algú pretén implantar un model amb un grau de precisió baix, ens en riurem. Però si el grau de precisió és alt (per exemple, del 99,9%), no necessàriament haurem de *confiar* en el model, ja que, com veurem, hi ha molts factors que poden desvirtuar la viabilitat o utilitat real d'un model que ofereix, en la seva fase de *construcció*, una precisió molt alta.

Tampoc seran suficients les gràfiques *ROC* (*Receiver Operating Characteristics*) i *AUC* (*Area Under the Curve* o àrea sota la corba), que ens ofereixen, en les tasques de classificació, informació probabilística i agregada sobre el comportament del model en totes les situacions possibles i la seva capacitat de diferenciar entre classes. De nou, es tracta d'informació útil, però que, a més de ser ambigua en certs contextos, és, en tot cas, insuficient en termes d'*interpretabilitat* imparcial i objectiva generadora de *confiança*.

Necessitem, en definitiva, instruments complementaris d'*interpretabilitat*. I el primer que caldrà fer és preguntar-nos en què consisteix tenir confiança en un model d'IA judicial. En quin moment podrem afirmar, en termes d'explicabilitat, que tenim prou confiança? Quan tinguem la certesa que el model funciona bé? O serà suficient tenir un cert coneixement del funcionament mecànic del model? On hem de dirigir la mirada? Al codi de l'algoritme? Als paràmetres que té en compte? A un altre lloc? Per intentar respondre a aquestes preguntes haurem d'analitzar les tècniques d'interpretabilitat actualment existents.

4.3.4. Tipus d'interpretabilitat

En l'àmbit acadèmic, quan es parla d'*interpretabilitat*, es fa referència a coses força diverses. Podríem distingir fins a quatre nocions, nivells o perspectives diferents (la

global, la local, l'algorítmica i la *post hoc*), circumstància que posa de manifest, ella sola, la complexitat de la qüestió. Analitzem-les per separat[36].

4.3.4.1. Interpretabilitat global

Es donarà quan sigui possible comprendre, en general, el funcionament de tot el model, incloses totes les seves parts. Hauríem de poder contemplar el model de cop, en la seva integritat, i disposar de la dada d'entrada i els paràmetres rellevants per tal de, en un temps raonable, seguir els càlculs que es fan en cada pas per generar la predicció. El model seria *llegible* per l'usuari, sigui visualment o en text[37].

4.3.4.2. Interpretabilitat local (o descomponibilitat)

La tindrem quan sigui possible validar empíricament, fent simulacions, el funcionament del model respecte d'un resultat específic (d'un *input* concret i petites desviacions del mateix). Aquí ens fixarem, no en el model íntegre ni de cop, sinó en parts del mateix, siguin els input, els paràmetres o els càlculs. Buscarem explicacions intuïtives. La *simulació* amb la qual busquem *explicacions locals* consisteix en la possibilitat que una persona, independentment del *hardware* i obtingut un determinat *output* per un *input* donat, experimenti per *veure què passa* amb l'*output* si s'introdueixen petits canvis en l'*input*.

Per entendre'ns, en el marc d'una assistència a un detingut pel qual s'ha decretat la presó preventiva fent ús d'una eina de predicció algorítmica del risc de reincidència, el seu lletrat pot voler posar en dubte que l'escala de risc obtinguda va ser correctament calculada i, per argumentar-ho, proposa que es determini fins a quin punt petits canvis en els factors concurrents en el seu client poden canviar el resultat final[38].

---

[36] Per fer-ho fusionarem dos treballs: Slack et al. (2019) i Lipton (2018).

[37] Les variables rellevants per la viabilitat de la interpretabilitat global serien, per una banda, la mida del model (per exemple, el nombre de nodes en un arbre de decisió) i, per l'altra, el temps requerit per la computació necessària perquè operin les inferències (per exemple, en un arbre de decisió, el pas de la branca [*root*] a la fulla [*leaf*]).

[38] En això consistiria l'explicabilitat local (Slack et al., 2019). Per tenir-la, necessitem, però, que els mateixos *inputs* siguin interpretables. Haurem de descartar models que utilitzin característiques (*features*) molt processades o despersonalitzades. Serà necessari, també, que cada node d'un arbre de decisió es correspongui amb un text descriptiu (per exemple, "tots els pacients amb pressió sanguínia per sobre de 150") o que els paràmetres d'un model lineal puguin ser descrits com a representacions de la força o intensitat de l'associació entre cada característica (*feature*) i l'etiqueta. Caldrà tenir present, a més, que si bé els pesos en un model lineal poden semblar intuïtius, no són tan asèptics: poden dependre de la prèvia selecció dels

4.3.4.3. Transparència algorítmica

Ja hem vist que un model d'IA té molts components, un dels quals (dels més rellevants, però només un) és l'algoritme mitjançant el qual el model *aprèn* i s'ajusta. És viable, per tant, predicar un determinat grau d'interpretabilitat del model específicament referit a l'algoritme. Així, en els models *lineals*, els més senzills, és possible observar, quasi directament, el procés d'entrenament que porta a determinades solucions i, fins i tot, la forma concreta que adopten certs errors de predicció. Per contra, els algoritmes més avançats d'aprenentatge *profund* operen amb uns potents procediments d'optimització de la xarxa neuronal que, simplement, no comprenem com funcionen, fet que ens pot fer dubtar, justificadament, que funcionin adequadament per a nous problemes.

4.3.4.4. Interpretabilitat post hoc

En el cas de les explicacions *post hoc* ja no pretenem conèixer com funciona *internament* el model. Ni en general, ni localment, ni a nivell algorítmic. Per contra, s'obté *externament* informació generada després de l'actuació del model per entendre millor què ha estat determinant en el resultat. La *interpretació* que s'obté pot ser *informativa* tot i que no ens il·lumini sobre els mecanismes interns del model. Un avantatge d'aquestes explicacions és que operen després que el model hagi actuat i no en comprometen, per tant, l'eficàcia o potencialitat predictiva. La informació *post hoc* pot adoptar diverses formes:

a) Pot tractar-se, directament, d'explicacions en llenguatge natural de com ha funcionat el model. Aquí es donaria la paradoxa que hi hauria dos models: un generaria la predicció i un altre, entrenat per separat (per exemple, una *xarxa neuronal recorrent* que operi amb llenguatge natural), generaria una explicació que transformaria en text l'estat intern del model que genera la predicció. El problema, de nou, és que aquestes explicacions estarien entrenades en experiències passades i oferirien (només poden oferir) una informació de tipus probabilístic sobre el que ha pogut passar a l'*interior* del model.

b) Poden ser explicacions *per exemples*: a més de la predicció concreta que es buscava, també s'identifiquen altres casos que poden haver tingut un tractament algorítmic similar. Per exemple, l'explicació de la predicció que una imatge determinada d'un tumor pot ser

atributs i de la fase de preprocessament. Veiem, en definitiva, que la interpretabilitat local, si bé és factible, ha de superar molts condicionants previs, dels quals cal ser-ne conscient des d'un inici.

maligne seria que s'assembla a altres imatges etiquetades com a malignes i que també són mostrades[39].

c) En tasques vinculades a les imatges, l'explicació pot adoptar la forma d'una *visualització* que representa com ha funcionat el model o què ha tingut més en compte o què ha estat més determinant. En definitiva, què ha après[40].

4.3.4.5. Interacció entre els diferents tipus d'interpretabilitat

Estem ja en condicions, una vegada analitzades les quatre tipologies d'*interpretabilitat* que hem desglossat, d'abordar alguns matisos aparentment contradictoris amb certes assumpcions molt esteses. Ser-ne conscients és important, ja que afecten a decisions que cal prendre en el procés de gestació d'una eina d'IA. Tenir un idea errònia de les potencialitats o les limitacions (en sentit positiu o negatiu) de l'eina que pretenem crear pot portar-nos a prendre decisions equivocades. Hem de conèixer, per tant, la complexitat i extrema *relativitat* del terreny en el qual ens movem.

Per començar, s'acostuma a entendre que, amb caràcter general, els models basats en algoritmes del tipus *arbres de decisió* (annex 1, apartat 2) i *regressions logístiques* (annex 1, apartat 1) són més interpretables que les *xarxes neuronals* pròpies de l'aprenentatge *profund* (capítol 6.2.5). Però, com acostuma a passar, no hi ha regles absolutes, sinó respostes casuístiques: l'afirmació pot ser vàlida al nivell de la *interpretabilitat estrictament algorítmica*. Si ens desplacem, però, a la *interpretabilitat local* (centrada en prediccions concretes i referida als *inputs*, els atributs i els paràmetres) i es tracta d'un model lineal amb atributs intensament processats o amb moltes dimensions, es produirà una pèrdua de la seva potencial *descomponibilitat*. No serà fàcil fer-hi simulacions *significatives*, amb independència que el concret algoritme utilitzat pugui ser, ell mateix, molt *interpretable*.

[39] Això es podria aconseguir, tècnicament, activant certs *layers* ocults de la xarxa per identificar els *k-nearest neighbors* sobre la base de la seva proximitat en l'espai après pel model. Ens remetem, per una major comprensió d'aquesta tècnica, al capítol 7 de l'annex 1.

[40] Un exemple serien les representacions amb *t-SNE* (*t-distributed stochastic neighbor embedding*): una imatge en dos dimensions representa les localitzacions de dades que són més probables d'aparèixer juntes. Alterant l'*input* per mitjà de l'activació de certs nodes seleccionats dels *layers* ocults de la xarxa neuronal, podrem (intentar) conèixer (explicar) què ha après el model. L'anàlisi detallada dels *inputs pertorbats* ens donaria pistes de l'aprenentatge.

Quan s'ha de triar entre models *lineals* o *profunds* (i pot ser que per una tasca determinada tinguem la doble opció), s'afrontarà sovint un dilema (*tradeoff*) entre transparència algorítmica i descomponibilitat, ja que les xarxes neuronals acostumen a operar amb atributs difosos escassament processats. Per tant, *purs* o molt *significatius*, per la qual cosa les informacions *post hoc* que puguem obtenir seran molt *informatives*. Per contra, els models lineals, per aconseguir nivells d'eficàcia similars, necessiten processar manualment els atributs. La implicació serà que, tot i que la *interpretabilitat algorítmica* pot ser major, seran menors les simulacions locals (la *descomponibilitat*) *significatives* que podrem fer. I, també, probablement, les explicacions *post hoc* rellevants que podrem obtenir.

Veiem, per tant, que, en contra del que podríem pensar, les xarxes neuronals poden ser, en certs casos, més adequades en termes d'*interpretabilitat local* o *post hoc*. Aprenen representacions complexes que es poden visualitzar, verbalitzar o utilitzar per fer agrupaments.

Si ens desplacem, ara, com a factor d'*interpretabilitat*, al volum o mida relatius dels models, pot donar-se el cas que una xarxa neuronal petita sigui *més interpretable* que un arbre de decisió molt gran.

Veiem, per tant, que per arribar a concloure si una eina d'IA és *interpretable* o *explicable* en una mesura acceptable o assumible en termes, per exemple, de l'estat de dret, cal partir d'una visió global i integral que abasti tots els components del model. Alguns podran ser molt o bastant interpretables. D'altres, poc o gens. L'anàlisi els haurà d'incloure tots. De fet, podrà tenir en compte, també, altres fons d'interpretabilitat *externes* al model, que no en són components, sinó que adopten la forma d'explicacions *post hoc*, ja analitzades.

4.3.5. Explicacions contrafàctiques

Abordarem, per acabar aquest apartat, una proposta força propera a la interpretabilitat *post hoc.* Es tracta de les *explicacions contrafàctiques* proposades per Wachter et al. (2017). Ens interessen especialment perquè parteix aquesta autora de la perspectiva del dret a ser informats que hi ha un sistema d'IA en ús. Distingeix entre les explicacions

sobre la funcionalitat del sistema (com funciona *ex ante* el sistema) i aquelles que ens informen sobre la racionalitat de la decisió individual, una vegada aquesta ja s'ha produït.

Ja sabem que disposar completament del codi pot ser estèril per comprendre el funcionament general (real, efectiu) del sistema. I, de fet, encara que hipotèticament poguéssim explicar completament un sistema d'IA en el seu funcionament general, això podria ser insuficient per entendre (i poder acceptar, o no) la racionalitat d'una decisió individual, per exemple, relativa a la denegació d'un préstec. Al mateix temps, però, com apunta Wacther, si s'arriba a disposar d'una informació molt completa del funcionament del sistema, tant a nivell general com a nivell de les seves decisions individuals, hi hauria el risc que aquest *excés* de transparència permetés manipular i *enganyar* el sistema, per obtenir-ne la decisió que desitgem en casos futurs. Això generaria interrogants ètics que caldria abordar.

Quan analitzàvem el RGPD al capítol 3.3.1.2, hem vist que l'art. 22 preveu un règim molt succint per a les decisions automatitzades en el qual no recull expressament una obligació d'explicació algorítmica completa (únicament el dret a obtenir una intervenció humana, a donar la opinió i a impugnar la decisió). De fet, inicialment sí que estava prevista, però ha quedat relegada a l'exposició de motius. A nivell vinculant únicament l'art. 13.2.f) RGPD preveu la obligació d'oferir, si s'està utilitzant un sistema de decisió automatitzada, una informació *significativa* sobre la *lògica* i la importància i les conseqüències per l'interessat. Ja hem vist que aquests termes tan ambigus i indeterminats poden deixar molt marge als actors que utilitzin aquestes eines. Seran els tribunals els qui, en última instància, hauran d'establir els cànons raonables d'informació: serà suficient, o no, una mera explicació del funcionament general del sistema, sense necessitat de precisar els factors concrets que han determinat la decisió automatitzada? Quina informació caldrà donar? Les categories de dades utilitzades per crear els perfils? La font de les dades? Quines dades han estat rellevants?

Aquí és on podem introduir la noció d'*explicacions contrafàctiques* com una via que permeti dotar de contingut real al dret a una explicació algorítmica, a la vegada que s'evita la necessitat (sovint inviable) d'obrir les *caixes negres* i s'obvien, també, els problemes de propietat intel·lectual i de secrets comercials o, fins i tot, d'afectació potencial a la privacitat d'altres persones, quan la disposició completa del codi i de les

dades utilitzades per entrenar l'algoritme poguessin ser objecte d'algun tipus de *reversió* tècnica.

Doncs bé, les *explicacions contrafàctiques* series informacions hipotètiques factibles d'obtenir i útils per entendre les decisions, impugnar-les o, en un futur, intentar modificar-les. Descriurien una dependència en els fets externs que han portat a una decisió: si el resultat p es genera com a conseqüència dels valors (*v1*, *v2*, …) de les variables *V*, una explicació contrafàctica consistiria en poder afirmar que si, per contra, les variables *V* haguessin tingut una altres valors (*v1'*, *v2'*, …), s'hauria obtingut un resultat *p'*. Per exemple, en el cas de denegació d'un préstec, l'explicació contrafàctica podria ser: "*li ha estat denegat perquè els seus ingressos anuals són de 50.000 euros. Si haguessin estat de 60.000, se li hauria concedit*".

Es descriuria la dependència respecte de fets externs que determinen la decisió. Aquesta dependència es pot fins i tot traslladar gràficament a imatges que mostren la posició d'una variable i la seva distància respecte de l'espai o llindar que hauria portat a una altra decisió. Aquestes imatges ens permetrien conèixer les *fronteres decisòries* i tenir una o diverses *indicacions* de què s'ha de canviar per poder modificar la decisió, sense necessitat d'explicar com funciona internament la lògica de l'algoritme. Aquest tipus d'explicacions podrien funcionar, de fet, fins i tot en el context de l'aprenentatge profund de les xarxes neuronals.

Sembla, així doncs, que tenen força avantatges. Caldrà veure, és clar, si són sempre factibles, també en el context de la IA judicial. No ho hem de descartar. És una via d'exploració d'un interès innegable, ja que, paradoxalment, el tipus de raonament en què es tradueix l'explicació contrafàctica té una estructura relativament similar a la de certes formes de raonament legal.

També tenen, però, limitacions. No informen sobre si el sistema en general és discriminatori. Només aporten indicacions individuals. Per altra banda, com que els sistemes d'aprenentatge automatitzat es modifiquen a sí mateixos amb l'ús, no hi ha garanties que si es tornés a activar l'algoritme, donaria el mateix resultat.

4.4. Imparcialitat i independència judicials

Una eventual IA judicial només serà viable si els principis d'imparcialitat i independència judicials romanen intactes o no es veuen afectats significativament. Es tracta, de nou, d'una afirmació força òbvia. Però és necessària perquè el mer fet d'introduir en el procés decisori components automatitzats pot generar interrogants al respecte. La posició d'equidistància de la instància judicial i la seva llibertat o autonomia en l'apreciació i ponderació dels elements de judici rellevants per arribar a una decisió final podrien veure's afectats si una eina d'IA fa, en el seu lloc, una part d'aquesta tasca (o tota). Fins i tot en cas de preveure's una ulterior supervisió judicial, els dubtes subsisteixen, atès que és coneguda la ja referida tendència humana (ja sigui per credulitat innocent o per poques ganes de treballar) de donar per bones solucions aparentment objectives que ja venen donades per sistemes automatitzats (l'anomenat *biaix d'automatització*).

En el cas del principi d'*imparcialitat*, el risc consistiria que un possible biaix o decantament parcial, favorable a una de les parts, contingut a l'eina d'IA judicial (ja sigui per discriminació o per qualsevol altre motiu) es traslladaria indirectament a la decisió judicial, potser sense ser-ne conscient el tribunal o sense disposar dels recursos per poder identificar i remoure aquest factor de parcialitat *programada*. És per això que haurien de preveure's mecanismes perquè en tot moment la instància judicial pogués controlar les implicacions, en termes de parcialitat, de les eines d'IA judicial que utilitza, modular-les i, en cas necessari, desactivar-les o prescindir-ne. Ja hem vist els problemes d'interpretabilitat, transparència i explicabilitat (d'opacitat, amb caràcter general) que poden presentar certs productes d'IA. És probable que molts d'ells no permetin aquest tipus de control. En aquest cas disposarem, probablement, d'un argument fort per descartar la seva implementació judicial. Com sempre, és clar, en funció del grau de *sensibilitat* de la tasca judicial de què es tracti.

Pel que fa al principi d'*independència*, el risc consisteix en el fet d'introduir, en el procés decisori, un component elaborat per un *tercer* el funcionament del qual pot ser desconegut i incontrolable pel mateix òrgan jurisdiccional: els factors de decisió ja no serien, només, els que el tribunal consideri rellevants o els que hagin aportat les parts (amb diferents nivells de vinculació pel tribunal segons la jurisdicció en la qual ens trobem). Disposaríem, a més, d'allò aportat per un *tercer*. Tercer que presenta, a la seva vegada, una naturalesa i composició complexa i híbrida: seran programadors informàtics,

tècnics i juristes els que hauran contribuït a crear l'eina d'IA judicial. No es tracta que aquest *tercer* provoqui, amb aquesta aportació, una *pressió* material i concreta sobre quin ha de ser el sentit de la decisió. El risc consisteix, més aviat, que si s'insereixen aquestes eines en el procés decisori en uns termes que no permeten la seva desactivació quan hi hagi motius o, fins i tot, que, permetent-la, aquesta desactivació requereixi un esforç addicional de motivació, llavors sí que pot acabar generant una certa *pressió,* en el sentit d'haver-se de seguir, en principi, allò que *proposa* l'eina. I aquesta certa pressió sí que pot acabar afectant la independència judicial. Per tant, serà imprescindible no només preveure els mecanismes *desactivadors*, sinó també no establir deures de *motivació extra* o *ad hoc* en cas de no seguir-se la proposta algorítmica: el sol fet d'exigir-se aquesta motivació especial generaria la indesitjada pressió de la qual estem parlant.

4.5. Dret de defensa, contestabilitat i motivació de les resolucions

Cal tenir present que si bé els riscos de discriminació algorítmica injustificada són una de les principals disfuncions potencials d'una eventual IA judicial, les limitacions de transparència i d'explicabilitat que caracteritzen el món de l'aprenentatge automatitzat plantegen problemes generals que van més enllà dels biaixos. Afecten, de fet, al nucli del dret de defensa: si en una decisió judicial s'ha utilitzat, de manera íntegra o parcial, en la fase final o en una intermèdia, alguna eina d'IA, el dret de defensa hauria d'exigir, com a qüestió de la màxima obvietat, que els afectats en tinguin coneixement, tant del fet d'haver-se utilitzat com de quin és el seu funcionament. Només així es trobaran en condicions de poder recórrer la decisió. Es tracta, per tant, d'un nou problema. Amb perfils propis. El de la *contestabilitat algorítmica*.

Ja s'ha apuntat que fins i tot quan superem la opacitat de les *caixes negres* i obtenim algun tipus d'informació sobre com funciona l'algoritme, aquesta explicació no pot ser equiparada, en absolut, amb el que acostumem a entendre com a *motivació judicial*. Es tracta, només, d'informacions descontextualitzades sobre meres correlacions estadístiques, que ens poden ajudar a conèixer, fins a cert punt, quins paràmetres han pogut ser rellevants en la resposta automatitzada, però que estan molt allunyades del tipus d'inferències pròpies del raonament legal que vesteix una motivació judicial estricte. És important, per tant, ser conscients d'aquesta diferència profunda entre *motivació legal* i *explicació algorítmica*, perquè té grans implicacions i no acostuma a ser tinguda suficientment en compte. Així, si ens situem en el context del dret de defensa i de la

informació que ha d'estar a disposició d'una part processal per poder recórrer en condicions, si ho desitja, una decisió judicial, es podrien donar, quatre situacions diferents:

1) La decisió conté una motivació jurídica i, a més, una explicació algorítmica.

2) La decisió conté una motivació jurídica que es remet o inclou l'ús d'una eina d'IA respecte del funcionament de la qual no disposem, però, de cap explicació algorítmica.

3) La decisió conté, únicament, una explicació algorítmica, sense motivació jurídica estricta.

4) La decisió no conté ni motivació jurídica ni una explicació algorítmica.

No necessita massa argumentació concloure que el supòsit 4 no és acceptable en el marc d'un estat de dret en el qual es reconeix el dret de defensa i es prohibeix l'arbitrarietat en l'actuació dels poders públics, inclòs el judicial. La situació seria similar en el supòsit 3: si estem d'acord que l'explicació algorítmica no pot ser equiparada a la motivació judicial, llavors seguirem sense disposar de la segona, encara que hi hagi algun tipus d'informació estadística de com s'ha arribat al resultat final. Es tracta, certament, d'hipòtesis poc plausibles, atès que sempre exigirem (i sempre serà relativament fàcil de subministrar) algun tipus de motivació judicial. Per tant, si s'acaben utilitzant eines d'IA judicial, és probable que ens trobem més sovint en els dos primers escenaris.

En el supòsit 2 hem d'afrontar, primer, la inexistència d'informació sobre com ha funcionat una eina d'IA que ha estat utilitzada per dictar una resolució judicial que, això no obstant, conté, ella mateixa, algun tipus de motivació judicial. La situació seria similar a la del cas *Loomis*, en el qual el denunciat va recórrer la decisió judicial per haver-se utilitzat l'eina *COMPAS* per predir el seu risc de reincidència. Doncs bé, en segona instància el tribunal va concloure que el mer fet de no disposar-se de tota la informació sobre el funcionament del programari utilitzat no afectava el dret de defensa i a un judici

just perquè aquest programari no hauria estat determinant, sinó només un factor més, entre altres, de la decisió judicial final[41].

Hem deixat pel final el *primer escenari*, aquell en el qual disposem tant de *motivació judicial* estricte com d'un nivell d'*explicació algorítmica* raonable. Tot i que és més *desitjable* que el segon, no per això deixa de ser problemàtic. Certament, la pertinença d'un cas concret a un o altre escenari serà graduable. Els límits seran difosos. El que ens interessa, però, posar de manifest és que fins i tot si ens trobem, clarament, en el primer escenari, seguirà havent-hi una problemàtica jurídica: encara que sapiguem com funciona l'algoritme (quins factors té i ha tingut en compte, amb quines dades s'ha *entrenat*, com van ser obtingudes, quines certificacions el validen, etc.), seguirà tractant-se d'un component automatitzat de la resolució judicial, que, per sí sol, podria justificar la seva impugnació per motius d'afectació a la imparcialitat i la independència judicials (ja ho hem vist al capítol 4.4) o per contradir la prohibició de l'arbitrarietat en l'actuació dels poders públics (art. 9.3 CE), en relació amb el dret de defensa entès com la necessitat que qualsevol resolució judicial es basi en la realitat concreta dels fets del cas tal com han estat asseverats i contrastats pel tribunal que la dicta.

En definitiva, podríem concloure que a mesura que anem avançant en els quatre escenaris, s'acumulen progressivament les problemàtiques jurídiques (independència i imparcialitat judicials, dret de defensa, contestabilitat, arbitrarietat en l'actuació dels poders públics i motivació judicial), amb la particularitat que les del primer escenari són, ja, molt rellevants. De nou, segurament dependrà del tipus de tasca judicial en la qual s'utilitzin les eines d'IA (a mesura que ens apropem a la decisió final, les alertes hauran d'intensificar-se) en combinació amb el grau d'opacitat de l'eina d'IA utilitzada (com més *negre* sigui la *caixa*, més problemes tindrem)[42].

[41] És certament discutible aquesta manera d'abordar el problema, perquè el mer fet que la decisió judicial enumeri com un dels elements de judici tinguts en compte l'output d'un algoritme permet concloure, a priori, que alguna incidència haurà tingut. En cas contrari, es podria haver omès la seva referència. A més, cal tenir en compte com n'és de fàcil (i habitual), en les resolucions judicials, complementar l'argument jurídic principal (i determinant) amb altres més estereotipats (i menys determinants) per generar la sensació que hi ha hagut un conjunt de factors concurrents, sense precisar el grau de rellevància de cadascun d'ells.

[42] Apuntem, però, per últim, que no es tracta de paràmetres que es moguin necessàriament en paral·lel: es pot donar el cas que una eina d'IA extremadament opaca sigui utilitzada en la fase inicial d'admissió de la demanda (determinació de la possible manca de competència territorial) i que, per contra, a la fase final de presa de decisió i dictat de la sentència s'emprin eines més senzilles i properes als anomenats sistemes experts, en els quals, a més d'usos instrumentals d'eines d'IA avançades (OCR i PLN), operen regles jurídiques prefixades que es poden conèixer perfectament. Per tant, caldrà analitzar, cas per cas, si són

4.6. IA i drets fonamentals

Hem vist ja, com a principis ineludibles d'una IA judicial, els de qualitat i eficiència, el de justícia o no discriminació, el d'imparcialitat i independència judicials i el dret de defensa en termes de contestabilitat i motivació de les resolucions judicials. Són probablement els principis bàsics d'una IA judicial. Hi podríem afegir, de manera íntimament vinculada, el dret a un judici just (o a la tutela judicial efectiva). Es tracta de valors reconeguts des de fa molt temps i el que ens interessa no és tan la seva proclamació (òbvia) com l'anàlisi de fins a quin punt poden quedar afectats per la introducció d'eines d'IA judicial. L'objectiu és fixar, si és possible, el nivell d'afectació *tolerable*. Amb ells no s'esgoten, però, els principis o drets susceptibles de veure's desafiats per una IA judicial. Abordarem, ara, altres drets fonamentals igualment rellevants que no podem obviar en una recerca com la present.

Ens serà útil, a aquest fi, acudir a les *Recomanacions sobre l'impacte dels sistemes algorítmics en els drets humans* (CM/Rec(2020)1) del Comitè de Ministres del Consell de la Unió Europea, aprovades el 8 d'abril de 2020. Aquestes Recomanacions fan referència als riscos (en termes d'impacte en drets humans) de qualsevol ús de la IA. En destacarem algunes idees interessants:

a) Els drets fonamentals potencialment afectats per l'ús creixent (general, no només *judicial*) dels algoritmes inclouen el dret a un judici just, a la privacitat i protecció de dades, a la llibertat de pensament, de consciència i religió, a la llibertat d'expressió i de reunió, a la igualtat de tracte o els demés drets econòmics i socials. La funcionalitat d'aquests sistemes es basa freqüentment en l'agregació i anàlisi sistemàtiques de dades recol·lectades digitalment a gran escala relatives a la identitat i el comportament *online* i *offline* d'individus i grups. A més de la possible intromissió en la privacitat, els desafiaments a altres possibles drets fonamentals han de ser tinguts en compte al llarg de tot el cicle de vida del sistema algorítmic.

b) Els nivells de *precisió* dels algoritmes no augmenten, ni necessàriament ni de manera automàtica, amb l'augment de la base de dades. Per contra, es pot expandir el nombre

assumibles els nivells respectius, i correlatius, d'opacitat i d'afectació al dret de defensa. Per això convé no adoptar posicionaments (positius o negatius) absoluts, sinó matisats i condicionats a cada context processal.

de persones afectades i la proporció d'errors en forma de *falsos positius* o *falsos negatius*.

c) Els sistemes algorítmics no operen només amb dades personals. També poden fer-ho amb *dades no personals* o *no-observacionals* (simulacions, dades sintètiques o regles generals). Però fins i tot en aquests casos els drets fonamentals poden veure's negativament afectats. Especialment, els dels individus i grups les dades dels quals no han estat processades o que no han estat degudament tinguts en compte.

d) Com que els sistemes algorítmics prioritzen certs valors respecte d'altres, i no sempre ho fan de manera explícita, transparent i controlable, poden generar efectes adversos, especialment per minories i grups marginats.

e) Hi ha un risc més elevat d'impacte a drets fonamentals quan els sistemes algorítmics són utilitzats en la presa de decisions o per serveis públics, especialment si l'individu afectat no té la opció de renunciar al seu ús. Davant d'eventuals eines d'IA *judicial* amb usos d'analítica legal, predicció o determinació individualitzada de riscos, cal adoptar unes prevencions especials i respectar les garanties del dret a un judici just, en els termes que se'n deriven de l'art. 6 CEDH. L'estàndard d'exigència d'*explicabilitat* ha de ser més alt en aquests casos.

f) Pot ser problemàtica la utilització de sistemes algorítmics que no són ni clarament públics ni clarament privats. Per exemple, quan hi ha una externalització de certs serveis públics o quan l'administració pública utilitza directament sistemes desenvolupats en el sector privat.

g) Hi ha d'haver un control regular de l'impacte en termes de drets fonamentals durant tot el cicle de vida del sistema algorítmic, des de la ideació inicial fins a l'avaluació dels seus efectes. El disseny, desenvolupament i implementació del sistema ha de preveure mecanismes perquè els individus estiguin informats per avançat sobre les finalitats i possibles resultats que ofereixi l'eina i sobre les dades processades, amb la possibilitat de controlar-les per vies no passives sinó *interoperables*.

h) La centralització creixent dels sistemes de processament de dades (com ara el processament per *núvol*) i l'eventual manca d'opció en l'elecció de la infraestructura

utilitzada pot debilitar els mecanismes de control de l'impacte dels sistemes algorítmics en els drets fonamentals. Caldria disposar d'infraestructures alternatives i segures per assegurar una alta qualitat en el processament de les dades.

i) Qualsevol experimentació computacional que pugui tenir una afectació significativa en termes de drets fonamentals ha d'estar precedida d'un informe sobre aquest impacte.

j) El disseny, el desenvolupament i la implementació de les eines d'IA han d'incorporar garanties segures i privades de prevenció i mitigació del risc de violacions de drets fonamentals i d'altres efectes adversos. Un sistema de certificació podria ser la solució.

k) L'avaluació i prova dels sistemes algorítmics ha de realitzar-se amb dades que representin adequadament la diversitat de la població. No s'ha de sobrerepresentar ni infrarepresentar cap grup demogràfic rellevant.

l) Cal assegurar que totes les eleccions i decisions preses amb alguna intervenció d'un sistema algorítmic que pot impactar significativament en drets fonamentals siguin identificables i puguin ser objecte d'un seguiment clar.

m) Els individus o grups afectats han de disposar de mitjans efectius per impugnar les decisions preses amb ajuda d'eines d'IA. Se'ls ha d'explicar de manera assequible, clara, imparcial i amb marge de temps quins són els drets o interessos que es poden veure afectats. El remei impugnatori ha de ser efectiu, ràpid, transparent i funcional. Ha d'incloure la possibilitat de ser escoltat i una revisió a fons de la decisió, amb possibilitat d'obtenir una decisió no automatitzada, inclosa la revisió judicial. Aquest dret hauria de ser irrenunciable i el procediment, no costós. Caldria complir amb els estàndards dels arts. 6, 13 i 14 CEDH.

No podem obviar, també en el context del Consell de la Unió Europea, la *Carta de Drets Fonamentals en el context de la IA i el Canvi Digital*, de 21 d'octubre de 2020: ens recorda aquesta Carta que la IA, tot i oferir grans oportunitats, també implica, si no és adequadament utilitzada, riscos pels drets fonamentals, la democràcia i l'estat de dret. Caldrà fer front a desafiaments com la opacitat, la complexitat, els biaixos i un cert grau d'impredictibilitat i autonomia en el seu comportament. Convindria, segons la Carta, fixar uns estàndards tècnics comuns per generar confiança en aquesta tecnologia, en la línia

del que proposa el *Llibre Blanc sobre IA* de la Comissió Europea al qual ja s'ha fet referència en el capítol 3.2.3. Els Estats han de preservar la seva sobirania digital, sense perjudici d'estar oberts a les aportacions de les companyies privades que compleixin amb els estàndards requerits. El nivell de respecte dels drets fonamentals i les normes vigents ha de ser de la mateixa intensitat en el món digital i en el físic. De fet, la tecnologia d'IA pot ser utilitzada per potenciar, per exemple, l'accés als serveis públics. També pot millorar, específicament, els mètodes per localitzar les evidències i proves en les causes penals. Pel que fa referència específica a les noves tecnologies i a la IA en matèria judicial, la Carta fixa com a prioritaris l'accés a la justícia, la transparència, l'explicabilitat, la independència judicial i la seguretat jurídica. La millora en l'accés a la informació legal, la reducció de la duració dels processos judicials i una millora general de l'accés a la justícia haurien de ser les prioritats. Però al mateix temps s'hauria de ser prudent davant de la possible utilització d'algoritmes esbiaixats. Cal preveure remeis o garanties legals efectius per assegurar el dret a un judici just, la presumpció d'innocència i el dret de defensa. En qualsevol cas, recorda la Carta, l'accés no digital a la llei i la justícia ha de seguir essent essencial.

Per la seva banda, la *Carta Ètica Europea sobre l'ús de la IA en sistemes judicials i en el seu context*, aprovada el 3 i 4 de desembre de 2018 per la Comissió Europea per l'Eficiència de la Justícia (CEPEJ), ens recorda que molts drets fonamentals, reconeguts per la Carta Europea i que es poden veure afectats per la IA, no són absoluts i poden sofrir limitacions amb base legal, sempre que responguin a la finalitat de perseguir un interès general o la necessitat de protegir altres drets o llibertats i sempre que es respecti l'essència del dret en qüestió i que les limitacions siguin necessàries i proporcionades. En el cas de l'ús de noves tecnologies, els Estats hauran de trobar l'*equilibri adequat* entre la protecció dels drets fonamentals i el desenvolupament d'aquestes tecnologies[43].

Per concloure, podríem dir que l'ús de tecnologies innovadores com la IA al camp de la justícia haurà de respectar, en tot cas, els valors comuns de la UE, fixats a l'art. 2 TUE

[43] Es fa referència en aquest punt a la STEDH, de 4 de desembre de 2008 (30562/04 i 30566/04), *S. and Marper v. the United Kingdom*, apartat 112, que exposa el següent:

"The Court observes that the protection afforded by Article 8 of the Convention would be unacceptably weakened if the use of modern scientific techniques in the criminal-justice system were allowed at any cost and without carefully balancing the potential benefits of the extensive use of such techniques against important private-life interests. (…) The Court considers that any State claiming a pioneer role in the development of new technologies bears special responsibility for striking the right balance in this regard".

(entre altres, el respecte a la dignitat humana, la igualtat, l'estat de dret, els drets humans, la no discriminació i la justícia), i el principi de *protecció legal efectiva* previst a l'art. 19.1 TUE. No hi ha una altra alternativa. Caldrà abordar molts equilibris. I potser alguns no seran viables.

5. La IA com a superació dels sistemes experts

5.1. Programaris tradicionals amb regles expresses i aprenentatge automatitzat

Els algoritmes existeixen des de fa molt temps, ja sigui en forma analògica o digital. Per tant, els factors tecnològics que realment han possibilitat el desenvolupament actual de la IA els trobem a un altre lloc: un poder de computació exponencialment creixent, la rapidesa de les comunicacions i en l'intercanvi de la informació en temps real, la disponibilitat d'immenses quantitats de dades (no necessàriament de molta qualitat) i una baixada progressiva del cost d'emmagatzematge. Han estat aquests factors els que han propiciat un salt qualitatiu de la IA. Des del seu naixement, als anys 50, les aplicacions d'IA havien estat força rudimentàries, fins i tot merament especulatives. No s'ha començat a implementar processos automatitzats en el món real (en temps real) fins que no s'han produït els referits avenços tecnològics.

Tornem, però, a les dues nocions nuclears de la IA: la tasca a realitzar i l'automatització. Mentre que les activitats humanes són difoses, complexes i imprevisibles quant al seu desenvolupament i resultats, la perspectiva algorítmica de la IA opera per mitjà de la *reducció* dels problemes i de *transformacions segures* per arribar a un resultat. Ens podríem preguntar, ara, si la IA suposa una novetat tant rellevant, ja que de *software* o programaris i d'algoritmes n'hi ha des de fa moltes dècades. No consistiria, la IA, simplement, en la posada en funcionament d'un *software* més o menys tradicional en el marc d'una molt més elevada capacitat de computació i de transmissió de la informació?

No és exactament així. Hi ha una gran diferència, que ens remet a la noció d'*aprenentatge automatitzat*: els programaris tradicionals consisteixen bàsicament en un conjunt de regles *redactades* en un llenguatge de programació determinat (JAVA, SWIFT, C++, etc.), expressament creades pels programadors i explícitament introduïdes en el *software*, per mitjà de les quals, introduïdes unes certes dades, se n'obtindrà una resposta.

Per contra, la IA (l'*aprenentatge automatitzat*) capgira aquest procés d'una manera molt significativa: al programari se li introdueixen primer les respostes i les dades i no és fins després que, una vegada localitzats certs patrons en les dades, en deriven el que podrien ser les regles. De fet, podem deixar de pensar en termes de regles, ja que, si hi són, les

infereix, a partir dels patrons que localitza, la mateixa màquina. I algunes vegades les podrem conèixer, però d'altres, no. És el problema de les *caixes negres* del qual ja n'hem parlat i en parlem en altres apartats de la recerca. Es tracta, per tant, d'un procés *inductiu*. No hi opera cap tipus de raonament abstracte, que sí estava present (tot i que *petrificat* per mitjà de la programació de regles expresses) en els sistemes de *software* tradicionals. El canvi és, per tant, radical.

Es tracta, en definitiva, de dues aproximacions a la solució de problemes completament diferents[44]:

a) Lògica dels *sistemes experts* (o *simbòlics*): amb regles expresses intentem representar el coneixement que ja tenim sobre una matèria. Es tracta, per tant, de models *deterministes*. L'autor del sistema és el seu mateix dissenyador. Amb ells modelarem de manera fidel una realitat (també la jurídica) que hi ha al món real, exterior. S'utilitza, aquí, el raonament computacional *deductiu*, integrat per complexes cadenes amb estructura d'*arbres*.

b) Lògica de l'*aprenentatge automatitzat*: són de naturalesa probabilística i operen per mitjà de la *inducció*. L'algoritme buscarà i potser trobarà certs patrons en les dades que se li introdueixin i n'inferirà per sí mateix regles. Primer ho farà en la fase d'*entrenament* i després en la d'implementació estricta. Podem entendre, de moment, per *regles* els *ajustaments* dels valors o calibratges dels factors rellevants que té en compte el sistema. Els patrons poden ser obvis (els podria detectar a primera vista qualsevol persona) o no tan obvis. Pot operar, a més, amb grans quantitats de dades, més enllà de les capacitats humanes ordinàries. Aprendrà de les dades i, en principi, millorarà amb el temps, en el sentit d'anar adaptant les regles (els *calibratges*) a les noves dades que se li introdueixin.

5.2. Autoria dels sistemes experts i dels models algorítmics

Cadascuna d'aquestes dues aproximacions (*sistemes experts* o *aprenentatge automatitzat*) té els seus avantatges i les seves limitacions. Per exemple, els sistemes experts tenen l'avantatge innegable que estan dissenyats *completament* per especialistes en la matèria, que per elaborar les regles hauran tingut en compte tots els

[44] Per comprendre bé la distinció entre aquestes dues aproximacions és força útil l'estudi tancat el 8 de desembre de 2020 pel CEPEJ del Consell d'Europa sobre la *Viabilitat de la possible introducció de mecanismes de certificació d'eines i serveis d'IA en l'esfera de la Justícia.*

factors rellevants i les situacions que es poden donar. Lògicament, hi col·laboraran experts informàtics per expressar aquestes regles en un llenguatge informàtic, però la creació de les regles la realitzaran *completament* els experts, per exemple, juristes si es tracta d'una eina amb finalitats legals o judicials. Es tracta, però, d'una tasca molt exigent i costosa, en temps i diners. És un dels seus principals inconvenients.

Per contra, en l'eventual creació de sistemes d'*aprenentatge automatitzat* judicials, si bé hi participarien tant experts juristes com informàtics o programadors, les seves tasques ja no estarien tan *compartimentades* com en el cas anterior (creació de regles legals expresses, per una banda, i traducció de les mateixes a un llenguatge informàtic, per l'altra). Hi hauria una major *simbiosi*, ja que no es tractaria de dues tasques tan diferenciades, sinó de la creació d'un sistema *unificat* (un model) que integri els paràmetres jurídics rellevants en el mateix procés d'inferència algorítmica dels patrons, regles o calibratges que acabaran sent determinants per la resposta que se n'obtingui. La conseqüència immediata d'aquesta simbiosi és que el paper dels informàtics o programadors passarà a ser molt més rellevant i a tenir molt més pes que en el cas dels sistemes experts: ja no tindran una posició *passiva* de mera *traducció* de quelcom que els ve donat (les regles creades *completament* pels experts), sinó que hauran de crear, ells mateixos (amb cooperació amb els experts legals, però amb una major capacitat de decisió i de disseny), el model, el programari i l'algoritme que més adequadament s'ajusti a la tasca judicial de què es tracti.

Aquest seria, per tant, un desavantatge: experts informàtics no especialistes en qüestions jurídiques tindrien un poder rellevant en el disseny d'eines que poden acabar sent decisives per algunes tasques judicials. Caldrà tenir, per tant, molta cura i prevenció. I, sobretot, assegurar-se que hi ha el major control estrictament jurídic possible en la mateixa fase de disseny. Això requerirà, sens dubte, que com a mínim una part d'experts jurídics aprenguin també, amb una determinada profunditat (no superficialment), com funciona el món de l'aprenentatge automatitzat. Només així podrà aconseguir-se la imprescindible confiança en els sistemes que es puguin crear. Confiança tant de la ciutadania en general com dels mateixos operadors jurídics.

5.3. Capacitat d'adaptació i evolució

Una limitació força rellevant dels *sistemes experts* és que la cadena de regles en què consisteixen és *rígida* i escassament adaptable. Per entendre aquesta limitació podem pensar en els *arbres de decisió* tradicionals[45], que poden preveure totes les situacions que es poden generar i les alternatives per sortir-ne i seguir avançant fins a arribar a una resposta final. Les bifurcacions podran ser tan complicades i extenses com es vulgui, però sempre quedaran *petrificades* en el punt on s'hagi finalitzat el disseny. El sistema no podrà adaptar-se per sí mateix a les noves situacions que es produeixin.

Certament, per aquests casos (nova situació per la qual no hi ha una regla que ofereixi una resposta o alternativa concreta), es podrà programar una determinada sortida *per defecte* que acabi donant una resposta i eviti el bloqueig del sistema. Però no es tractarà d'una resposta *adequada* o *ajustada* al problema o a les dades introduïdes. Per tant, caldrà *actualitzar* i modificar l'arbre. I només ho podran fer els mateixos experts o especialistes, que són els qui hauran d'introduir noves regles o modificar les existents. Per tant, els costos personals i econòmics seran no només inicials sinó permanents degut a la necessària actualització constant del sistema. I no cal dir que en el marc del dret, que es caracteritza per la seva evolució constant (tant a nivell normatiu com jurisprudencial), aquestes actualitzacions haurien de ser molt freqüents. Massa. Convindria, de fet, que un equip d'experts estigués en permanent disposició per actualitzar el sistema quan fos necessari. Seria un primer gran inconvenient dels *sistemes experts*.

Per contra, en l'*aprenentatge automatitzat* ja no hi hauria, en principi, aquesta necessitat constant d'*actualització humana*. Seria el mateix sistema qui s'adapta a les noves dades que se li introdueixen, en el sentit que, partint dels patrons i calibratges que ja ha fixat en la fase d'*entrenament* i amb els quals ha iniciat la seva *implementació real*, els podrà anar modificant, si així millora les respostes, a partir de la informació que li sigui introduïda amb les noves dades. Aquesta capacitat d'adaptació faria innecessari que els creadors de l'algoritme modifiquin *manualment* el sistema. Ja ho fa per sí mateix. Això no exclou, lògicament, que aquests sistemes requereixin una constant *monitorització*,

[45] Sense ànim de complicar massa aquest punt, hem de matisar que dins de l'*aprenentatge automatitzat* hi ha algoritmes que s'anomenen, també, *arbres de decisió*. Ho veurem a l'apartat 2 de l'annex 1. Ara ens referim, però, als arbres de decisió en un sentit més tradicional, gairebé analògic.

per assegurar-se que segueixen funcionant adequadament. Però seria un control *extern*, de comprovació, sense intervenir, necessàriament, en el mateix *calibratge* del sistema. Si es constata, però, precisament gràcies a aquesta monitorització, que el funcionament no és adequat, llavors sí que potser caldrà algun tipus d'actuació *interna*. Aquesta seria, però, eventual, en cas d'*error* o de *funcionament inadequat*. Podem diferenciar, en definitiva, entre l'actualització necessàriament manual dels *sistemes experts* i la mera monitorització externa amb eventual rectificació manual dels sistemes d'*aprenentatge automatitzat*. Es tracta, certament, d'una petita gran diferència.

5.4. Coneixement i control dels paràmetres d'actuació

La contrapartida, però, a aquesta aparent capacitat d'autoadaptació de l'*aprenentatge automatitzat* és el menor *control* que es té sobre els paràmetres o criteris amb què actua el programari en cada moment: mentre que en els *sistemes experts* el control és total (coneixem *en tot moment* quines són les regles, què diuen, si poden ser discriminatòries i que s'apliquen, només, aquestes regles; són, en definitiva, fàcilment auditables), no passa necessàriament el mateix en el cas de l'*aprenentatge automatitzat*. Aquí hem de diferenciar entre la possibilitat mateixa de *conèixer* les regles o els calibratges algorítmics, per una banda, i, per l'altra, el *control* real i efectiu que puguem tenir sobre aquestes regles o calibratges.

Pel que fa al *coneixement*, el podrem tenir si no ens trobem davant d'una *caixa negre*. S'han analitzat aquestes *caixes* amb més detall quan abordàvem la *transparència algorítmica* i el dret de defensa. Hem vist ja als capítols 4.3 i 4.5 que és una qüestió molt més complexa del que sembla i que presenta una gran varietat de grisos i matisos.

Pel que fa al *control* sobre la fixació de les regles o calibratges, podem concloure que és sempre menor en l'*aprenentatge automatitzat* en comparació als *sistemes experts*. En aquests el control és total. Per contra, en els primers, si bé hi ha un innegable *control* en la fase de disseny, entrenament i implementació inicial, ja hem vist que és en tot cas compartit entre experts juristes i experts informàtics. A més, el control és, fins i tot en aquella fase, relatiu o limitat, ja que es limitarà, com veurem, a aplicar determinats ajustaments en algunes variables que admeten manipulació, amb la finalitat d'acabar obtenint el major grau de precisió (*accuracy*) possible. No serà, en definitiva, un control complet i exhaustiu en la definició de les regles. Si ens desplacem a la fase

d'implementació real de l'eina d'*aprenentatge automatitzat*, el control es veurà encara més debilitat, ja que aquí serà el mateix sistema que aprendrà per si mateix (es calibrarà a si mateix) en funció de les noves dades que se li introdueixin, sense necessitat d'una intervenció *humana* expressa. Sí que hi haurà, és clar, una *monitorització*, però serà externa i només es traduirà en una intervenció (modificació) expressa si es pren aquesta decisió. En cas contrari, el sistema seguirà funcionant *modificat*.

Aquesta capacitat d'*autocalibratge* i de *seguir funcionant* amb uns nivells de precisió en principi elevats (i, fins i tot, amb una possible millora constant) no és, de fet, una limitació de l'*aprenentatge automatitzat*, sinó, més aviat, un dels seus avantatges principals. El problema es presenta, però, quan ens plantegem la seva implantació judicial: en aquest context, el relatiu o escàs *control* que hi ha en la fixació de les regles o els calibratges sí que pot erigir-se eventualment en un problema. Dependrà, de nou, i en una visió de conjunt, del tipus de tasca judicial de què es tracti i del nivell real de *control* que permeti l'eina judicial en qüestió. Però la cautela haurà de presidir l'anàlisi.

5.5. Eventual compatibilitat de les dues aproximacions

Fins al moment hem abordat les dues perspectives per afrontar tecnològicament certes tasques judicials com si fossin incompatibles entre sí. És evident que són molt diferents, però enlloc no està escrit que no es puguin compatibilitzar en alguna mesura. Es tracta, a més, d'una possibilitat que podria ser molt fructífera pel camp judicial.

En última instància, hem d'acudir a la noció de *model*, de sistema. Un sistema pot ser complexa i estar integrat per diferents components. I, de fet, sembla evident que si s'acaben implantant eines d'IA judicial, hauran de tenir un cert nivell de complexitat i de diversitat de components entre els quals caldrà una mínima interoperabilitat. Res no impedeix, d'entrada, que alguns d'aquests components responguin a l'aproximació dels *sistemes experts* i uns altres a la pròpia de l'*aprenentatge automatitzat*. De fet, potser aquí trobarem la via per superar els problemes que s'han apuntat sobre les diferències entre autoria, capacitat d'evolució i control del sistemes: quan sigui imprescindible un control jurídic gairebé complet sobre les regles aplicables i no sigui tant imprescindible una capacitat d'auto-adaptació (pensem, per exemple, en la generació d'esbossos de sentències d'aplanament o en l'automatització de l'admissió a tràmit de certes demandes), podrem acudir a components amb el format dels *sistemes experts*. Per altra

banda, quan ja no sigui tan necessari un control jurídic elevat en els calibratges de les regles i necessitem, per contra, una elevada capacitat d'auto-adaptació sense intervenció manual constant (per exemple, en la transcripció de les declaracions orals o la traducció simultània), podrem acudir a components que operin segons el model de l'*aprenentatge automatitzat*. I res no impedirà, en principi, que ambdues aproximacions *interoperin* de manera adequada en el context del mateix procediment judicial.

6. Una immersió més tècnica en la IA

6.1. Més enllà dels algoritmes: cap a una comprensió tecnològica global del model

Aquest apartat de la recerca abordarà, des d'un vessant més tècnic, en què consisteix la IA. Fins ara la perspectiva ha estat més genèrica i divulgativa. Més imprecisa. Ara ja convé, però, entrar en més detalls. Si no, no s'entendran les constants referències que fem als tipus de *models* que existeixen i el seu grau d'*interpretabilitat* o complexitat, així com els elements o *paràmetres* que, donat un model, poden ser modificats per perseguir una finalitat determinada (per exemple, millorar-ne, precisament, la seva explicabilitat). Intentarem, per tant, definir i entendre, ja amb més concreció, la IA. Diseccionarem els components que la integren. Especialment (però no només) els algoritmes. Adoptarem una visió global que abasti tot el procés que s'ha de seguir des que es planteja, com a idea, la possibilitat d'implementar una eina d'IA judicial fins que la mateixa és efectivament posada en pràctica i ha de ser monitoritzada. Es tracta, per tant, de l'apartat més dens i exigent de la recerca. A la següent nota a peu de pàgina s'inclouen algunes recomanacions de lectura *parcial*[46].

Doncs bé, a la part introductòria de la recerca ja s'ha avançat que, si bé acostumem a associar la IA amb els algoritmes, aquests integren, només, un component dels sistemes utilitzats per realitzar de manera automatitzada certes tasques. En són un element important. Fins i tot, molt important. Però només un. N'hi ha molts altres. I si no els abordem i estudiem, el nostre coneixement real sobre què és la IA serà molt fragmentari. De fet, insuficient. Hem d'analitzar el procés complet d'ideació, gestació, prova, implementació i monitorització d'un projecte (d'un model complet) d'automatització d'una tasca (per exemple, judicial). I veurem que hi intervenen molts factors tecnològics diferents als algoritmes.

[46] Si el lector o lectora no desitja entrar en els detalls tècnics que s'abordaran en aquest capítol, potser li convindrà saltar-se aquest capítol 6 i passar directament al 7, que aborda la controvertida figura del *jutge-robot*. Cal advertir, però, que l'apartat més dens, el relatiu a les diferents tipologies d'algoritmes que existeixen, s'ha ubicat, per raons de claredat i simplificació, a l'annex 1. Sí que seria de lectura més *necessària*, en tot cas, l'apartat 6.6, que ofereix una visió global d'un projecte d'aprenentatge automatitzat. Dit això, si del que es tracta és d'explorar la possible implantació de la IA al sector judicial, no sembla que puguem abordar seriosament el projecte sense aprofundir mínimament, des d'una perspectiva tècnica, en la naturalesa, components, potencialitats i, també, limitacions dels recursos informàtics de què estem parlant.

Comencem, ara, una anàlisi més detallada i tècnica d'aquests sistemes que ens poden *ajudar* a realitzar certes tasques de manera automatitzada. Tasques tan diferents com la classificació d'imatges, la classificació d'articles de notícies, detectar comentaris ofensius o d'odi, resumir textos llargs, gestionar un assistent personal (*chatbot*), predir els ingressos en funció de diverses mètriques de rendiment, reaccionar a ordres de veu, detectar fraus o anomalies, segmentar i agrupar els clients per dirigir-los diferents estratègies de màrqueting, representar un conjunt complex de dades de moltes dimensions en un diagrama clar i més senzill o, per últim, recomanar productes.

Fins aquí, la descripció *no tècnica* de les tasques. L'altra cara de la moneda serien les eines tècniques que realitzen aquestes tasques, els models algorítmics subjacents, que poden ser, també, de molts tipus diferents. Per nombrar-ne alguns: les *regressions lineals* o *polinomials*, les *màquines de vectors suport* (o *SVM*) *de regressió*, els *arbres de decisió*, els *Random Forest de regressió*, les *xarxes neuronals artificials*, les *xarxes neuronals convolucionals* (*RNC*), les *xarxes neuronals recurrents* (*RNR*) o els *transformadors*. Veiem que la matèria es complica. S'utilitzarà un sistema o un altre en funció, precisament, del tipus de tasca de què es tracti. O potser, també, del tipus de dades que estiguin disponibles. O, simplement, del pressupost que es tingui o dels condicionants del sector en el qual s'ha d'aplicar. En el cas de la IA judicial, els condicionants són, lògicament, molt intensos. Ja els hem vist. Però abans d'entrar en més detall en aquestes qüestions més denses (de les quals només se'n volia enunciar un esbós), tornem, per situar-nos millor, a les classificacions clarificadores, començant per la que ens diferencia entre la IA *supervisada* i la *no supervisada.* Podem avançar, per completar la fotografia, que existeix, també, la IA *semisupervisada* i una altra categoria completament diferent, la de l'aprenentatge *per reforç*. Anem, però, per parts[47].

6.2. Tipus d'aprenentatge

6.2.1. IA supervisada

Pertanyen a la IA *supervisada* aquells sistemes d'IA que en la fase d'entrenament (ajustament i prova) utilitzen unes dades que han estat manualment *etiquetades* amb la resposta *correcta*. És a dir, el conjunt de dades d'*entrenament* inclou les solucions desitjades. D'aquí ve la qualificació de *supervisat*. Si es tracta que el sistema ens indiqui,

[47] Per abordar aquesta matèria més tècnica seguirem, en part, a Géron (2020).

en un futur, quan li introduïm fotografies, si hi apareix una poma o un gos, en la fase d'entrenament li haurem subministrat com a dades fotografies que contenen la resposta correcta: la dada no serà, només, la fotografia, sinó també la indicació que es tracta d'una poma o un gos. El sistema s'anirà ajustant a sí mateix de manera autònoma per trobar les correlacions que fan que una imatge sigui, probablement, una poma o un gos, però partirà d'imatges que ja sap que són una cosa o l'altra.

Els supervisats són els models més utilitzats i les tasques típiques que aborden, les següents:

a) De *classificació*: classificació d'imatges o de clients, detecció de frau, diagnòstics o filtres de correu brossa (*spam*) dels correus electrònics. En aquest últim cas, en la fase d'entrenament se li mostren al sistema exemples de correus *dolents*, etiquetats com a tals, perquè en detecti les característiques i les pugui aplicar en un futur a nous correus que arribin.

b) De *predicció* (o *regressió*): l'objectiu és predir un valor numèric, que pot referir-se a la predicció meteorològica, al comportament dels mercats, al creixement de la població o al preu d'un vehicle si s'aporten una sèrie de característiques, anomenades també *predictors.*

c) Ús de la regressió per a la classificació: el valor de sortida pot ser la probabilitat de pertànyer a una categoria determinada. Es combinen, així, les funcions de predicció i classificació.

Aquestes serien les finalitats i els mecanismes de creació dels models *supervisats*. Pel que fa als algoritmes que solen utilitzar són els arbres de decisió, els *Random Forest*, els *KNN* i la regressió logística. Paraules certament estranyes. Les abordarem amb més detall a l'annex 1.

6.2.2. IA no supervisada

En la IA *no supervisada*, per contra, les dades amb les quals s'entrena (es calibra) i es prova el sistema no han estat prèviament etiquetades. Són models menys freqüents que, això no obstant, s'apropen més a l'essència i naturalesa de la IA: com que ja no s'entrenen amb dades etiquetades, el control humà és menor i l'àmbit de decisió

automatitzada és major, ja que abasta tant la identificació dels patrons rellevants com, en certa mesura, els tipus de resposta que es poden donar. El sistema busca les correlacions o similituds (els patrons amagats), podríem dir, *a cegues* o sense professor. Sí que es programa, lògicament, la tipologia i número de respostes que li demanem al sistema que ens doni. Però les dades amb les quals s'alimenta i es calibra no tenen aquestes respostes.

Les tasques més habituals de la IA *no supervisada* són les següents:

1) L'*agrupament per similituds*: les semblances poden ser òbvies o, també, inherents i no aparents. Un exemple prou popular podria ser la plataforma NETFLIX, que utilitza els algoritmes *K-Medias* i *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*) per recomanar grups de pel·lícules similars. Per altra banda, algunes empreses de transport (*Amazon*, *Seur*, etc.) acostumen a utilitzar *DBSCAN* per optimitzar la programació de les rutes dels seus treballadors (Rubiales, 2020).

2) La detecció d'*anomalies* i de *novetats* (per exemple, l'ús fraudulent de targetes de crèdit): a la fase d'entrenament s'alimenta el sistema només amb dades normals (netes, lliures d'anomalies) perquè, després, quan en vegi una de nova pugui analitzar si pot tractar-se, o no, d'una anomalia. Alguns dels algoritmes que poden assumir aquest tipus de tasca són els *SVM* d'una classe o les *Isolation forest*.

3) *Visualització*: es pretén preservar, en una imatge, tanta estructura (provinent de les dades) com sigui possible, ja que sovint es produeixen solapaments que dificulten l'anàlisi d'aquestes dades. Amb una optimització de les tècniques de visualització es poden identificar en les dades patrons insospitats. Es tracta de la *visualització* del Big Data.

4) *Reducció de la dimensionalitat*: quan es tenen dades amb moltes dimensions (variables o atributs), ens pot interessar, per poder gestionar-les o tractar-les més eficientment, simplificar-les sense perdre massa informació. Per exemple, fusionant vàries característiques que estan, de fet, correlacionades (com passa per exemple amb els kilòmetres i l'antiguitat d'un vehicle). Es tracta de la tècnica d'*extracció de*

característiques[48]. La *reducció*, que utilitza algoritmes *no supervisats*, pot aplicar-se, de fet, abans d'introduir les dades en un altre algoritme, que pot ser *supervisat*, per tal que pugui operar amb més rapidesa o requerir menys memòria. Veiem aquí un bon exemple de com poden cooperar, assemblar-se i interconnectar-se eines no supervisades i eines supervisades.

5) *Regles d'associació*: es pretén en aquest cas descobrir relacions interessants entre els diferents atributs de les dades. Estudiar l'estructura subjacent a les dades. Per exemple, potser es detectarà que els clients que compren salsa barbacoa i patates, també acostumen a comprar bistec, per la qual cosa pot ser raonable, des d'un punt de vista de màrqueting, situar a prop tots aquests productes[49].

6.2.3. IA semisupervisada

Aquí es produeix una combinació entre l'aprenentatge *supervisat* i el *no supervisat*. Sovint s'explica perquè disposar o obtenir les dades etiquetades és molt car i requereix força temps. Pot ser útil, en certs casos, utilitzar dades que estiguin etiquetades només en part. Un exemple clarificador seria el programa *Google Photos*, en el qual diverses persones apareixen en diverses fotografies, però el sistema únicament necessita que li diguem qui és cadascuna d'elles en una ocasió o en una fotografia (seria la part *supervisada*). A partir d'aquí, després de trobar les correlacions o els patrons de cadascuna de les persones en les fotografies etiquetades, posteriorment el sistema farà una tasca *no supervisada* d'agrupament de totes les persones que apareixen, sense etiquetar, a la resta de fotografies disponibles i podrà, en principi, posar nom a totes elles.

6.2.4. Aprenentatge per reforç

Més enllà de la IA supervisada i no supervisada (i de la semisupervisada), hi ha una tercera categoria molt diferent, que configura, pràcticament, un món a part: la de l'aprenentatge *per reforç* (*reinforcement learning*).

En aquest cas, l'aprenentatge el protagonitza un *agent* (per entendre'ns, un artefacte o *robot*) que es desplaça, explora i observa l'entorn (per mitjà de sensors), avalua possibles

[48] En aquest cas es poden utilitzar algoritmes de noms tan estranys com els següents: Anàlisi de components principals (*PCA*), Anàlisi de components principals amb *Kernel*, *LLE* (*Locally Linear Embedding*) o la Tècnica *t-SNE* (*t-Distributed Stochastic Neighbor Embedding*).
[49] En aquest cas es pot acudir als algoritmes *Apriori* o *Eclat*.

alternatives, en selecciona una i realitza accions, per les quals rep a canvi recompenses (si han estat encertades) o càstigs en forma de recompenses negatives (si no ho han estat). L'aprenentatge consistirà, en definitiva, en trobar quina és la millor *estratègia* o *política* per maximitzar les recompenses en el menor esforç o temps possible.

En aquest cas es produeix, fins a cert punt, una certa *personificació* de la IA, ja que l'agent existeix materialment, el podem veure i es desplaça per l'entorn físic com ho fa una persona. Ja no seria un mer programari constituït per uns i zeros (bé, no ho seria habitualment, perquè els agents de la IA per reforç també poden existir i actuar en entorns virtuals, simulats).

Els usos d'aquest tipus d'aprenentatge són mols fragmentaris i específics i no tenen, en principi, massa interès per la present recerca. Aquí l'algoritme ajuda un agent, que interacciona amb sensors amb un entorn, a moure's-hi i actuar per tal d'aprendre a maximitzar les recompenses quan té èxit. Ho fa ajustant els pesos dels paràmetres del sistema a partir de la interacció amb l'entorn. A diferència de l'aprenentatge supervisat, aquesta interacció és *avaluativa* i no *instructiva*. Els usos més habituals serien la mobilitat robòtica, els programes de videojoc o la presa de decisió en temps real, entre altres. Un algoritme habitual en aquest tipus d'aprenentatge és l'anomenat *Markov Decision Process*.

6.2.5. Aprenentatge profund

6.2.5.1. Diferències respecte l'aprenentatge merament automatitzat

Podríem dir que la classificació entre aprenentatge supervisat, no supervisat i per reforç abasta la totalitat del camp de la IA. Més enllà d'aquesta tripartició, trobem altres criteris per classificar els sistemes d'IA, que, a la seva vegada, podran ser utilitzats per tasques d'aprenentatge supervisat o no. Aquest és el cas de les *xarxes neuronals artificials*, també referides com a sistemes d'aprenentatge *profund*. No es tracta, però, d'una quarta tipologia d'aprenentatge al mateix nivell que les anteriors, sinó d'una terminologia que s'utilitza per diferenciar certs models més avançats que els tradicionals, que serien els *no profunds* o més *superficials*. És per això que un model *profund* (ara veurem en què consisteix) podrà ser utilitzat per tasques d'aprenentatge supervisat i no supervisat, segons els casos. No hi ha, per tant, cap correlació tancada. De fet, les tasques supervisades (com la classificació d'imatges, el reconeixement de veu o la traducció),

són les més habituals en els models *profunds*. Però, com dèiem, no ha de ser necessàriament així.

Doncs bé, els sistemes d'aprenentatge *profund* són els models més complexos i potents d'aprenentatge automatitzat. Amb ells es busca una aproximació *oberta*, *completa* i *incondicionada* a la solució de determinats problemes. Per contra, en l'aprenentatge automatitzat més *tradicional* s'acostuma a dividir els problemes en parts i a abordar-los de manera fragmentada. Són sistemes (els d'automatització *tradicional*) en tot cas *intel·ligents*: no només automatitzats sinó que, a més, aprenen ells mateixos tant durant la fase de prova i creació com en la posterior de funcionament o implementació ordinària. Les regles que els governen no han estat expressament definides pels seus creadors, sinó que les ha trobat (calibrat) el mateix sistema i, de fet, les segueix calibrant i concretant indefinidament durant el mateix procés de funcionament. No deixa de créixer o madurar. Però tot això ho fa (el sistema *tradicional*) operant amb algoritmes i una computació que podríem definir com *ordinària*, de la qual podem tenir un cert coneixement, més o menys elevat, segons els casos, sobre com funciona i dels motius o factors pels quals el sistema ha donat una resposta determinada.

Per contra, els models d'aprenentatge *profund* tenen totes les notes apuntades inicialment respecte dels models *tradicionals* (automatització, aprenentatge i modificació constant) però, a més, utilitzen una tecnologia més complexa. Necessiten grans quantitats de dades, de les quals n'extreuen una gran quantitat de paràmetres, molts més que amb les tècniques ordinàries d'aprenentatge automatitzat.

A més, la seva major *complexitat* provoca que difícilment podem arribar a obtenir una *explicació* dels factors que han determinat una resposta. De fet, també se'ls anomena *xarxes neuronals* per tenir, en la seva estructura o arquitectura, certes analogies amb el sistema neuronal cerebral de l'ésser humà.

Aquestes limitacions de l'aprenentatge *profund* en termes de transparència el fa, d'entrada i per raons òbvies, escassament atractiu per aplicar-lo a tasques judicials. Si bé dependrà, una vegada més, de la concreta tasca de què es tracti i de quin nivell de *sensibilitat* se li reconegui en termes de dret a un judici just i dret de defensa. Així, en tasques de processament del llenguatge natural, que poden ser molt pràctiques en l'àmbit judicial i que poden no ser especialment problemàtiques o *sensibles* (pensem, per

exemple, en la transcripció de les vistes, sense perjudici de la seva gravació, etc.), l'aprenentatge *profund* pot oferir uns nivells d'eficàcia més alts en comparació a altres models algorítmics. De fet, potser és l'únic model que ofereix les condicions tècniques necessàries per abordar aquesta tasca amb un mínim nivell d'eficàcia.

6.2.5.2. Xarxes Neuronals Artificials (ANN)

L'exemple paradigmàtic dels algoritmes d'aprenentatge *profund* són les *xarxes neuronals artificials* (o *ANN*). Es tracta d'un conjunt de components computacionals que operen en paral·lel i simulen, fins a cert punt, l'estructura neuronal del cervell humà. Una llarga cadena d'unitats (anomenades nodes o neurones) formen un o més nivells (*layers*) i es connecten entre si seguint un determinat patró per permetre les comunicacions entre elles. Cada connexió té associat un *pes* determinat que es projecta sobre la informació que rep com a senyal d'entrada (*input*). Aquest *pes* excita o inhibeix la senyal que s'està comunicant. Per tant, cada neurona tindrà un estat intern *activat* o *desactivat*. En la interacció entre l'*input* i el *pes*, pot generar un senyal de sortida (*output*) que serà enviada a altres neurones.

Si el tipus d'aprenentatge que s'està implementant és *supervisat*, els senyals d'entrada (*inputs*) contindran el valor propi i, també, el valor objectiu (la resposta correcta).

6.2.5.3. Xarxes Neuronals Convolucionals (CNN)

Si les xarxes neuronals *artificials* (*ANN)* suposen un salt en complexitat respecte de l'aprenentatge automatitzat *ordinari*, les xarxes neuronals *convolucionals* en fan un altre respecte de les ANN. Aquestes segones ignoren l'estructura dels senyals d'entrada i converteixen (redueixen) totes les dades en una matriu d'una dimensió. Aquest mètode pot ser adequat per la majoria de dades ordinàries, però no, per exemple, per processar i classificar imatges (que tenen, per definició, dues dimensions). Aquí és on entren les xarxes neuronals *convolucionals* (*CNN*), que tenen en compte l'estructura bidimensional de les imatges per extreure'n les seves propietats específiques. Reben les dades *en brut* de les imatges i n'ofereix una classificació correcte com a *output*.

Des d'un punt de vista arquitectònic, mentre que les ANN s'articulen amb *nivells ocults* connectats per mitjà de les neurones, les CNN disposen les neurones en tres dimensions

(ample, alçada i profunditat) i cadascuna, ubicada en un nivell, està connectada a un petit camí del senyal de sortida del nivell anterior.

6.2.6. Aprenentatge per lots o gradual

Abordem ara, més enllà dels aprenentatges *supervisats* o *no supervisats*, *profunds* o *no profunds*, unes altres classificacions de la IA menys populars o freqüents però tan o més rellevants i que, sobre tot, ens poden ajudar a iniciar aquest camí més dens de comprensió tècnica i no divulgativa de la IA.

Amb aquesta idea, podem diferenciar els sistemes d'IA en funció de si, per entrenar-se, necessiten utilitzar, o no, totes les dades disponibles. Hi hauria dues tipologies, l'aprenentatge *per lots* i el *gradual*. Vegem-les:

a) En l'aprenentatge *per lots,* el sistema s'entrena necessàriament utilitzant totes les dades disponibles. I una vegada inicia la seva execució material (real), deixa d'aprendre. Exigeix una inversió elevada, temps i molts recursos computacionals. Si les dades son voluminoses, pot ser inviable. Si es vol que conegui noves dades, caldrà parar el sistema anterior i entrenar una nova versió des de zero amb el conjunt de dades completes (no només amb les noves, sinó també amb les velles). Aquest procés de renovació es pot automatitzar, però requereix, de nou, temps i pot ser inadequat si es necessita una adaptació ràpida a dades que canvien ràpid (per exemple, en el cas de la predicció del preu de les accions).

b) En l'aprenentatge *gradual* i *online* (sobre la marxa), per contra, l'aprenentatge es produeix gradualment i de manera seqüencial (individualment o per *mini lots*) a partir del flux d'entrada de noves dades. Es pot prescindir de les instàncies de dades ja utilitzades en l'aprenentatge. És adequat per tasques que requereixin una adaptació ràpida a noves dades o quan es disposi d'escassos recursos computacionals.

La *taxa d'aprenentatge* (és a dir, la rapidesa amb què el model s'adapta a les dades entrants) es pot calibrar (modular) de la següent manera:

a) Si la taxa és elevada, la reacció serà ràpida, però el sistema aviat s'oblidarà de les dades antigues, eventualitat que pot no interessar en certes tasques, com la detecció de

correu brossa (no ens interessa que el sistema marqui com *spam* únicament els últims tipus de correu *apresos*).

b) Si és baixa, el sistema tindrà més *inèrcia*: l'aprenentatge serà més lent però també menys sensible al *soroll* (distorsió) que generen les dades noves o els anomenats valors *atípics* (molt poc freqüents i que poden implicar injustificades modificacions dels calibratges).

Per la seva pròpia naturalesa gradual, aquest tipus de sistema exigeix una constant i atenta *monitorització*, en la qual poden utilitzar-se diferents tècniques:

a) Per detectar la introducció de dades *dolentes* que en disminuiran el rendiment (per exemple, algú introdueix *spam* en un motor de cerca per sortir beneficiat en la prelació de la informació que ofereix).

b) Per detectar la introducció de dades anormals o *atípiques* que poden afectar el funcionament del model. Aquí podrien actuar els algoritmes de detecció d'anomalies.

c) Per desactivar l'aprenentatge immediat quan es constata un funcionament inadequat del sistema i revertir-lo a un estat anterior que sí que funcionava.

6.2.7. Com generalitza el sistema: aprenentatge basat en instàncies o en models

Un altre eix conceptual per entendre com funcionen els sistemes d'IA consisteix en analitzar com *generalitzen*. Com fan les prediccions respecte dels exemples nous, els que no han vist abans. Aquí és crucial tenir present que si bé és important que el sistema tingui un bon rendiment (que la *precisió* o la taxa d'èxit siguin elevades) en la fase d'*entrenament* (és a dir, amb les dades utilitzades durant l'entrenament), el que realment ens importa i interessa és que el rendiment sigui bo, després, amb les instàncies *noves*. Aquesta és la finalitat per la qual decidim crear sistemes d'IA. I pot donar-se el cas, com veurem, que un sistema ofereixi un bon rendiment durant l'entrenament però no amb dades noves (en el món *real*). En això consisteix *generalitzar bé*, o no: mantenir (o fins i tot millorar) el rendiment quan es passa de la fase d'entrenament amb dades conegudes a la fase de prova o implementació amb dades noves.

Hi ha, en aquest sentit, dues vies principals de *generalització* o de predicció algorítmica:

a) L'aprenentatge basat *en instàncies*: simplement compara punts de les dades noves i punts de les dades conegudes. Aprèn dels exemples de memòria i després generalitza a nous casos utilitzant una mesura de *similitud*. Un exemple seria, de nou, el sistema de detecció de *spam*, que consisteix, bàsicament, en el recompte de paraules en comú entre els correus ja qualificats com a dolents i els nous.

b) L'aprenentatge basat en *models*: aquí es crea un model a partir d'un conjunt d'exemples i després s'utilitza per fer les prediccions. Es detecten patrons en les dades d'entrenament i es crea posteriorment el *model predictiu*. Un model podria ser, per exemple, assumir que els diners fan més feliç a la gent i establir una correlació entre el PIB i el nivell de satisfacció vital: aquest nivell de satisfacció seria, per tant, una funció lineal del PIB per càpita. En aquests casos necessitem definir (calibrar) el valor dels dos paràmetres perquè el model tingui el millor rendiment possible[50].

Com que és crucial comprendre bé en què consisteix un aprenentatge que generalitza a partir de models, ens hi aturarem. Seguirem l'exemple anterior, pel qual podríem acudir a un algoritme de *regressió lineal* (dels més senzills que existeixen) utilitzat per fer prediccions sobre la correlació ja indicada: entre el grau de satisfacció vital que podran tenir els ciutadans de certes zones del planeta en funció del seu PIB. Li introduirem els exemples d'entrenament i trobarà els paràmetres que faran que el model lineal s'ajusti millor a les dades inicials quan sigui aplicat a dades noves. La clau serà el *calibratge* (la localització dels *valors òptims*) dels *paràmetres.* Doncs bé, aquest *calibratge* és el que es concreta de manera automatitzada en la fase d'entrenament. Probablement el model de regressió lineal ens oferirà una correlació molt estable i regular (de fet, una recta; d'aquí, en part, el nom de *lineal*) entre els dos paràmetres: com més alt sigui el PIB, més alta serà la satisfacció vital, en un progressió relativament estable. Podríem utilitzar, també, un altre tipus d'algoritme igualment vàlid per aquesta tasca, com el *K dels veïns*

[50] Ens podem preguntar, en aquest punt, què entenem per *millor rendiment possible*. Ens remet a la idea de precisió o encert, però cal disposar d'una eina tècnica (matemàtica o algebraica) per poder *mesurar* aquest rendiment. Per poder saber, en definitiva, quin rendiment té un model determinat pel que fa a la seva *capacitat de generalització*. Existeixen dues vies principals per obtenir aquesta mesura del rendiment:

a) Amb una funció d'utilitat (o d'adequació): es mesurarà allò que es considera *bo*.
b) O, inversament, amb una funció de *pèrdua*: es mesurarà allò que es considera *dolent*. És força habitual en les regressions lineals (prediccions) quan es mesura la distància (algebraica, projectable en una gràfica) entre les prediccions del model i els exemples d'entrenament, amb l'objectiu, precisament, de minimitzar aquesta distància.

més propers (KNN), també força senzill, que ens mostraria un agrupament similar de les dades disponibles de PIB i de satisfacció vital.

Imaginem, però, ara, que volem enriquir el model i tenir en compte, a més del PIB, la taxa d'atur, l'estat de salut de la població o el nivell de contaminació. Això ens obligaria, d'entrada, a obtenir més dades d'entrenament i, probablement, de més qualitat. Però també ens obligaria a replantejar-nos si el tipus d'algoritme utilitzat (per exemple, el de regressió lineal) és l'adequat perquè el model aprengui amb les dades que se li ofereixin i, després, pugui fer prediccions amb noves dades desconegudes. La major complexitat de les variables a tenir en compte segurament no podrà ser adequadament *assimilada* per una progressió merament lineal. Potser haurem d'acudir a un model més potent. Un de progressió *polinomial*? Deixem-ho, però, de moment, aquí. Únicament pretenia apuntar-se anticipadament alguna de les decisions tècniques rellevants que clar prendre en la creació d'un model automatitzat.

6.3. L'algoritme, motor del model

Acabem d'abordar els diferents tipus d'aprenentatge que es poden donar en el camp de la IA: supervisat o no supervisat, profund, per lots o gradual i basat en instàncies o en models. Baixarem, ara, una mica més, fins als elements que conformen el *motor* d'aquest aprenentatge: els algoritmes en sentit estricte. Ja hem vist en diverses ocasions que l'algoritme, tot i ser un element clau del model, no sempre és el que més feina genera o la fase respecte de la qual ens hem de fer més preguntes i prendre més decisions. La majoria d'algoritmes ja estan creats i estan disponibles (*built in*) com a programari lliure en les plataformes d'aprenentatge automatitzat. En el marc d'un projecte concret d'IA és habitual provar successivament diferents tipus d'algoritmes, els que *a priori* semblen més adequats. I, si hi ha sort, es podrà acabar disposant, per a una única tasca a realitzar, de diferents models que funcionin amb diferents algoritmes entre els quals poder escollir. En qualsevol cas, les decisions i el temps dedicat estrictament a la creació o ajustament de l'algoritme serà probablement menor al temps dedicat a la selecció i preparació de les dades o a la monitorització del funcionament del sistema. Integra, només, una part de les diverses fases per les quals ha de discórrer tot projecte d'IA.

Dit això, si les dades són l'aliment o el *petroli* de la IA, els algoritmes en segueixen essent l'ànima, el cervell o el motor. Són, indiscutiblement, molt rellevants. Són el repositori

d'ordes o instruccions que permeten al sistema aprendre sense que ningú no el programi expressament per fer-ho.

El primer que cal dir és que cada algoritme d'aprenentatge automatitzat té les seves potencialitats funcionals i raons per ser, o no, aplicat. Dependrà de la tasca a realitzar i de les dades disponibles. No existeix, en general, el *millor* algoritme. Tampoc no cal descartar la possibilitat de combinar diferents algoritmes en el marc d'un únic model creat per a una tasca determinada.

En el marc d'aquesta recerca només farem un breu repàs d'alguns dels algoritmes més habituals o coneguts. Perquè n'hi ha molts, d'algoritmes. Se'n creen constantment. No s'intentarà, en cap cas, analitzar a fons i en detall com funcionen aquests algoritmes: seria un repte excessivament tècnic que excediria notòriament l'objecte de la present recerca, especialment si tenim en compte la fase més que incipient en la qual encara es troba una eventual IA judicial. Només volem apropar-nos una mica més a aquest món dels algoritmes. Fer-nos una idea més exacta sobre què és, exactament, un algoritme i quins tipus d'algoritmes podrien ser eventualment útils en certes tasques d'IA judicial.

Tot i que la ubicació natural d'aquest tema seria aquest mateix capítol, a continuació de les tipologies d'aprenentatge automatitzat, per raons de claredat i simplificació, i atesa la naturalesa més tècnica i àrida de la matèria, s'ha traslladat l'anàlisi dels tipus d'algoritmes a l'annex 1. Per tant, els lectors que vulguin intensificar aquesta immersió tècnica en la IA sense perdre el fil poden acudir-hi en aquest punt i retornar, després, al capítol 6.4. Si no, no han de fer res més que passar directament al següent apartat.

6.4. Autonomia entre tasques i algoritmes: possible combinació dels models

Una vegada finalitzat aquest ràpid trajecte per les principals tipologies d'aprenentatge automatitzat i d'algoritmes (pels lectors que hagin acudit a l'annex 1), anem copsant algunes de les seves particularitats i complexitats. Proposarem, ara, dues conclusions provisionals que en podem extreure.

Per una banda, una idea que és important assimilar des de ben aviat és que no hi ha una correlació estricta o tancada entre les tasques susceptibles de ser realitzades i els algoritmes que les poden fer. És a dir, no està predeterminat que un específic tipus

d'algoritme només pot realitzar una determinada tasca ni, a la seva vegada, que aquesta tasca només pot ser abordada amb un concret tipus d'algoritme. Les possibilitats són molt més àmplies i flexibles. Lògicament, alguns algoritmes no poden fer certes coses, però sovint una tasca la podran fer diversos algoritmes, amb diferents nivells de precisió o d'exigència computacional. O, també, amb diferents nivells de transparència[51]. Caldrà, quan hi hagi diverses opcions *plausibles*, escollir. I en el sector judicial sembla evident que el criteri de la transparència ha de ser rellevant. Tot i que, com sempre, dependrà de la concreta tasca de què es tracti i del grau d'afectació al dret de defensa que generi.

En segon lloc, les tipologies o classificacions dels aprenentatges propis de la IA no són excloents. Poden combinar-se. Per exemple, els filtres *spam* d'última generació utilitzen un aprenentatge supervisat online (gradual) basat en models. És a dir, en un únic sistema convergeixen les notes de l'aprenentatge supervisat (entrenament amb dades etiquetades amb la resposta correcta), gradual (l'entrenament no s'ha de fer amb totes les dades disponibles; pot ser seqüencial i assimilar un flux constant de dades noves) i basat en models (va més enllà de la mera comparació entre la instància nova i les anteriors i pressuposa una correlació més o menys complexa entre diferents paràmetres, que calibra).

6.5. Reptes de l'aprenentatge automatitzat

Pot ser un bon moment, ara, per fer un petit repàs de quins són els principals reptes tècnics de l'aprenentatge automatitzat. Ja podem intuir que les problemàtiques seran molt diferents en funció de la tasca que es vulgui abordar i del model algorítmic que es plantegi implementar. Però amb caràcter general podem apuntar els següents desafiaments[52].

[51] Pot donar-se el cas, per exemple, que una mateixa tasca de classificació o predicció la puguin fer, simultàniament, diversos algoritmes, com ara els *K veïns més propers* (*KNN*), una regressió lineal, una regressió logística, una màquina de vectors suport, un arbre de decisió, un *Random forest* o, en certs casos, fins i tot, una xarxa neuronal. Tots aquests tipus d'algoritme s'analitzen superficialment a l'annex 1.
[52] Seguim, de nou, a Géron (2020).

6.5.1. Quantitat insuficient de dades

La majoria dels algoritmes (no tots) necessiten moltes dades per funcionar bé. Fins i tot per abordar una tasca senzilla caldrà alguns milers de dades, mentre que per altres més complexes com el reconeixement del discurs (llenguatge) o d'imatges, potser en faran falta milions. Sempre es pot plantejar, però, la possibilitat de reutilitzar parts d'un model ja existent.

Un article paradigmàtic, titulat "L'efectivitat irraonable de les dades" (Halevy et al., 2009), va posar de manifest (amb aparell empíric) que algoritmes molt diferents d'aprenentatge automatitzat tenien un rendiment gairebé idèntic en un problema complex de processament del llenguatge natural si se'ls donava les dades suficients. S'apuntava, per tant, la idea que potser no caldria invertir tant temps i diners en el desenvolupament algorítmic i sí, per contra, en el desenvolupament del *corpus* de dades. Doncs bé, és cert que sovint es posa un èmfasi excessiu en el component algorítmic dels sistemes i es descuiden massa altres elements essencials, com és evident que són les dades. Al mateix temps, però, cal tenir present que els conjunts de dades petits o mitjans són molt habituals. No sempre és fàcil ni barat aconseguir noves dades d'entrenament. Per tant, els algoritmes seguiran sent importants.

6.5.2. Dades d'entrenament no representatives

És obvi que, per poder generalitzar bé (fer prediccions *correctes* amb les noves dades), és imprescindible que les dades d'entrenament siguin *representatives* dels casos nous als quals volem generalitzar. Així, en l'exemple ja apuntat de la correlació entre PIB i nivell de satisfacció vital, si afegim a les dades d'entrenament les d'alguns països rics que, sorprenentment, ja no mostren una correlació estrictament progressiva entre els dos paràmetres, constatarem que potser les dades d'entrenament inicials no eren prou representatives de la resta de països no inclosos en l'entrenament i dels quals en buscarem la predicció. De fet, veurem que les correlacions són molt més complexes que les regressions lineals i les línies rectes i que, probablement, necessitarem un model més complexa. Cal estar alerta, perquè la utilització d'unes dades d'entrenament no suficientment representatives ens pot portar a creure que el problema que volem automatitzar és més senzill (lineal) del que realment és. Ens pot portar a escollir un model algorítmic inadequat. D'escassa qualitat i utilitat: serà poc probable que faci prediccions

exactes. I és millor adonar-se'n aviat que no pas tard. El major drama serà, però, no arribar a adonar-se'n mai.

L'escassa representativitat de les dades d'entrenament ens remet, de fet, al problema ja abordat al capítol 4.2.6 dels biaixos algorítmics. Però sí que interessa afegir, ara, que aquests biaixos o mancances ja en el nivell de les dades pot presentar diverses formes:

a) *Soroll mostral*: la mostra és, simplement, massa petita, fins el punt que les dades utilitzades són massa *decisives*, fan massa *soroll*. Són tan poques que la seva introducció en el model pot haver respost, de fet, al mer atzar.

b) *Biaix mostral*: la mostra pot ser gran, però el mètode de mostreig és defectuós. Per exemple, en una enquesta electoral basada en trucades telefòniques en la qual s'han obtingut els telèfons de subscripcions de revistes o socis de clubs, es produirà un clar condicionant de la tendència ideològica (més conservadora que disruptiva) dels enquestats. Podem afegir, per últim, el biaix de la *no resposta*, quan el nivell de persones que han contestat és baix.

6.5.3. Dades de mala qualitat

Un altre problema obvi però que no pot ser desatès és la possibilitat, en absolut remota, que les dades utilitzades continguin errors. Aquests errors poden traduir-se en valors *atípics* que generen molt *soroll* (una incidència elevada no justificada) i que fan més difícil que el sistema detecti adequadament els patrons subjacents en les dades. Per aquest motiu és convenient, en qualsevol projecte, dedicar temps i esforç a *netejar* les dades. Aquesta tasca pot adoptar diverses formes. Consistirà, amb caràcter general, en descartar, directament, els valors atípics o arreglar manualment els errors. Si el problema (força freqüent) és que a algunes instàncies de les dades els falta alguna característica o atribut, les opcions serien unes altres: ignorar completament per a totes les dades el paràmetre afectat (també respecte de les dades que sí que el tenen); ignorar les instàncies incompletes; reomplir els valors que falten (fent la mitjana dels valors que sí que tenen aquesta dada) o, fins i tot, entrenar dos models diferents, un que integri la característica afectada i un altre que no la tingui en compte. Veiem, en definitiva, que són moltes les decisions que cal prendre fins i tot en la fase inicial de *neteja* de les dades.

6.5.4. Característiques irrellevants: enginyeria de característiques

Un supòsit de mala qualitat de les dades pot consistir que continguin característiques (atributs, informació, en definitiva) irrellevants respecte de la tasca que volem realitzar. Una dita del sector diu que "si introdueixes escombraries en el model, n'obtindràs, també, escombraries". Per evitar-ho hem d'acudir a l'anomenada *enginyera de característiques*, amb la qual podem abordar diverses accions: seleccionar les característiques que ens siguin més útils, *extraure* característiques (combinant característiques existents per produir-ne una de més útil, tasca en la qual poden ser força eficaços els algoritmes de reducció de la *dimensionalitat*, explicats a l'annex 1) o crear noves característiques a partir de les dades noves.

6.5.5. Sobreajustament de les dades d'entrenament

És un problema força habitual. El model té un bon rendiment amb les dades d'entrenament: primer se l'entrena (perquè s'autocalibri) amb una part de les dades conegudes i després es fa la prova de la seva precisió o taxa d'èxit amb una altra part de les dades d'entrenament. En aquesta fase dona uns resultats de precisió molt bons, però, després, quan és aplicat a noves dades o supòsits que no ha conegut durant la fase d'entrenament, no *generalitza* bé. Per això diem que el model *sobreajusta* les dades d'entrenament. Respon de manera massa fidel o propera a aquestes dades, però no localitza uns patrons adequats a les dades noves amb què serà utilitzat a la vida real. Tot i que inicialment, en l'entrenament, generava la sensació de ser un molt bon model, no ens és útil. Més enllà del conjunt de dades d'entrenament, no és eficaç. De fet, un model pot funcionar millor que un altre en la fase d'entrenament però, després, generalitzar pitjor. No hi haurem de confiar.

La causa habitual d'aquests sobreajustaments és que el model ha detectat patrons en el mateix *soroll* de les dades d'entrenament. Per soroll hem d'entendre els atributs o característiques pròpies, per mer atzar, de les dades, sense que responguin als paràmetres habitualment rellevants en les situacions del sector sobre el qual treballem. Aquest risc és més elevat en els models complexos com les xarxes neuronals profundes, que poden detectar els patrons més subtils (inimaginables) en les dades. Alguns podran ser molt útils. O, per contra, poden ser mers derivats de *soroll* no rellevant. El risc augmentarà, igualment, com més petita i sorollosa sigui la mostra d'entrenament.

De fet, pot donar-se el cas que un model sigui massa complexa en relació a la quantitat i soroll de les dades d'entrenament. En aquest supòsit, podrien adoptar-se diverses mesures:

a) Simplificar el model:

1) Reduint-ne els paràmetres o els atributs (passar d'un model polinomial a un de lineal).

2) *Restringint* el model amb les tècniques de *regularització*[53].

b) Reunir més dades d'entrenament.

c) Reduir el *soroll* de les dades d'entrenament: solucionar els errors i eliminar els valors atípics.

6.5.6. Subajustament de les dades d'entrenament

Com es pot intuir, el subajustament és el contrari del sobreajustament: el model és massa simple per aprendre l'estructura subjacent a les dades. Per exemple, si tornem al model que correlaciona el nivell de satisfacció vital amb el PIB d'un país, podem acabar constatant que un model lineal pot no correspondre's amb una realitat més complexa. Potser les respostes merament lineals seran inexactes fins i tot (aquesta és la clau) en els exemples d'entrenament. En aquest cas, el problema és conceptual, de model, fins al punt de no generar expectatives ni tan sols en la fase d'entrenament. Senzillament, el model és massa simple.

Aquí les solucions o sortides són força òbvies:

[53] La *regularització* és una tècnica per mitjà de la qual, tenint un model, per exemple, dos paràmetres, obliguem l'algoritme a mantenir sempre baix el valor d'un d'ells, amb la qual cosa s'aconsegueix un equilibri entre un ajustament adequat a les dades i, al mateix temps, mantenir el model suficientment senzill perquè generalitzi bé. Un *hiperparàmetre* controlarà la quantitat de *regularització* aplicada durant l'aprenentatge. Es tracta d'un *hiperparàmetre* perquè no queda afectat per l'algoritme: s'introdueix abans de l'entrenament i seguirà constant durant tot el procés. Com més alta sigui la regularització, més pla serà el model. Es traduirà en un pendent que tendeix a zero. Segurament no sobreajustarà a les dades d'entrenament, però serà menys probable que trobi bones solucions. Que generalitzi bé. La clau serà, com és habitual, trobar l'equilibri adequat.

a) Seleccionar un model més potent amb més paràmetres.

b) Introduir característiques o paràmetres més ajustats[54].

c) Reduir les *restriccions* que s'hagin aplicat en excés i que han aplanat massa el model. En altres paraules, reduir l'hiperparàmetre de *regularització*[55].

6.6. Visió global d'un projecte d'aprenentatge automatitzat

6.6.1. Introducció

Si hem arribat fins aquí, estarem ja en condicions de construir un mapa complet i exhaustiu del procés en què es tradueix, necessàriament, la implementació d'una eina d'IA. El fragmentarem en un conjunt d'etapes alhora successives i íntimament interrelacionades[56]. Són aquestes:

a) Contextualització del problema.

b) Obtenció i tractament de les dades.

c) Selecció del model i entrenament.

d) Prova del model.

e) Prellançament del model.

f) Desplegament i monitorització.

6.6.2. Contextualització del problema

El primer que haurem de fer és emmarcar el problema i la tasca que cal abordar. Preguntar-nos si caldrà acudir a un tipus d'aprenentatge supervisat (apartat 6.2.1), no supervisat (apartat 6.2.2) o per reforç (apartat 6.2.4). També quina tasca volem

[54] Ens remetem aquí al que ja s'ha dit a l'apartat 6.5.4. sobre la enginyeria de característiques.
[55] Ens remetem a la nota a peu de pàgina 53.
[56] Seguim, de nou, a Géron (2020).

automatitzar: una de classificació o una de regressió i predicció; o d'un altre tipus? La computació serà per lots o gradual (apartat 6.2.7)?

Per exemple, si volem que el sistema ens indiqui per casos futurs si pertanyen a un grup A o B, la tasca serà de classificació. Si el que volem és detectar situacions estranyes o inhabituals, haurem de començar a pensar en algoritmes de detecció d'anomalies en els patrons. Si la resposta haurà de ser una quantitat numèrica d'alguna cosa (en diners o un altre valor), haurem de buscar en els algoritmes de predicció (les regressions). Si volem saber i conèixer com està organitzat un conjunt de dades, probablement ens serà útil una eina d'agrupament o *clustering* (apartat 9 de l'annex 1). Si, per últim, la resposta haurà de ser a la pregunta "Què he de fer a continuació?", no hi ha cap dubte que haurem d'acudir a l'aprenentatge per reforç (apartat 6.2.4).

Ja hem vist que, de fet, les tipologies d'aprenentatge automatitzat es combinen. Així, si pretenem predir el preu mig de les cases per districtes, es tractaria d'un aprenentatge supervisat de regressió múltiple (hi ha diverses característiques) i univariant (volem predir un valor per cada districte).

Hem de tenir clares, però, quines són les *assumpcions inicials* que fem. Un model no és una altra cosa que una versió simplificada de les observacions. Hem de descartar els detalls superflus que tenen poques possibilitats de *generalitzar* a instàncies noves. Hem de decidir, per tant, quines dades descartarem i quines mantindrem. Més concretament, quines dades *buscarem* i quines no. Aquí rau, en definitiva, l'aspecte necessàriament *polític* i *estratègic* de qualsevol projecte d'IA. En el bon, però també crític, sentit de les paraules. No són eines asèpticament objectives i neutres. I no perquè siguin utilitzades amb mala intenció, sinó perquè, per la seva mateixa naturalesa i funció, no poden ser-ho.

Així, un model lineal fa la suposició que les dades (i la realitat que els és subjacent) són principalment lineals i que la distància entre les *instàncies* i la línia recta es només *soroll* que pot ser ignorat. De fet, si no féssim cap suposició sobre les dades, no hi hauria cap raó per preferir un model respecte d'un altre. En última instància, per a un conjunt determinat de dades, el millor model pot ser un lineal; i per un altre, una xarxa neuronal.

No hi ha, en definitiva, un model que tingui garanties *a priori* de funcionar en qualsevol context. L'ideal seria avaluar-los tots. I si no és possible (normalment no ho serà), seria convenient com a mínim partir de suposicions raonables sobre les dades i avaluar uns quants models (no només un) que siguin, també, raonables. Per poder, al final, escollir-ne el millor. O no escollir-ne cap.

Veiem, en tot cas, que en la mateixa fase de mera contextualització del problema, fins i tot abans d'entrar en contacte amb les dades *brutes*, ja anem *equipats* amb un conjunt d'assumpcions (decisions ja preses) molt rellevants i de les quals no sempre se n'és conscient en la resta de fases. Aquesta *ignorància metodològica* és la que fa creure a molta gent en el caràcter objectiu o asèptic de la IA. Res més lluny de la realitat.

6.6.3. Obtenció i tractament de les dades

A partir de les dades disponibles, les haurem de visualitzar i analitzar per maximitzar el rendiment del model. Serà necessari *preparar* les dades perquè puguin ser *processades* per l'algoritme. Potser caldrà plantejar l'ús d'un *pipeline* de dades gràcies al qual diferents components del sistema, que operen de manera relativament independent, processin les dades, aplicant-hi diferents transformacions.

En aquesta fase, més intensa i complexa del que s'acostuma a imaginar, cal adoptar moltes decisions, fins i tot abans d'entrar en el disseny pròpiament algorítmic del model. Les analitzem a continuació. No interessa tant que el lector comprengui i retingui cadascuna de les situacions o eventualitats que ara s'exposaran (un total de 12), sinó, més aviat, que adquireixi consciència de fins a quin punt ens trobem davant d'una fase que, tot i que preliminar, tindrà moltes implicacions en el futur del model que s'estigui creant:

1) Primer haurem de carregar i visualitzar l'estructura de les dades. Es pot utilitzar, amb aquest fi, el programari *Pandas*[57].

[57] *Pandas* ofereix diversos mètodes per visualitzar les dades. Entre ells: el *head* mostra les primeres files superiors; l'*info* en mostra una descripció ràpida: número total de files, la tipologia de cada atribut (si és *float* [numèric] o un objecte [text, etc.]) o el nombre de valors nuls que hi ha); el *value_counts* indica quines categories hi ha i quantes instàncies hi pertanyen; el *describe* fa un resum dels atributs numèrics (mitjana, mínim, màxim o la desviació estàndard, entre altres aspectes); i el *hist* de *Matplotlib* visualitza amb gràfiques un histograma per cada atribut numèric.

2) A continuació crearem un conjunt de *prova*. És una fase molt rellevant, sovint desatesa. Haurem de dividir (*split*) el conjunt de dades disponibles entre les que seran utilitzades per l'*entrenament* estricte i les que seran utilitzades per la fase de *prova* (*test*) per determinar el grau de precisió (*accuracy*) del model. És habitual que el segon grup sigui d'un 20%, excepte quan el conjunt de dades disponible sigui molt gran. Un dels problemes que poden sorgir és que, si s'executa el programa vàries vegades, en cada una d'elles crearà un conjunt de prova diferent i, amb el temps, acabarà *veient* tot el conjunt de dades, una eventualitat que cal evitar per assegurar la solidesa dels resultats. Per assegurar-se'n, existeixen diverses solucions, en les quals no hi entrarem atesa la seva naturalesa eminentment tècnica.

3) En aquesta fase inicial hem de ser conscients dels possibles *biaixos de mostra*, dels quals ja n'hem parlat a l'apartat 6.5.2. El mer fet que el mostreig s'obtingui de manera aleatòria no ens assegura la seva adequació o *representativitat*: si tenim constància que un atribut determinat és important, ens hem d'assegurar que el conjunt de prova és *representatiu* de les diferents categories (o estrats) de l'atribut. En cas contrari, hi hauria el risc de *distorsionar* la importància d'un estrat determinat. El programa *Pandas* ofereix tècniques per abordar aquest risc amb les quals es condiciona de manera *conduïda* el mostreig i s'eviten les distorsions que generaria un mostreig merament aleatori.

4) A l'hora de descobrir, visualitzar i analitzar les dades, ens hem d'assegurar que s'ha apartat ja el conjunt de prova i que només tenim davant el conjunt d'entrenament. Només així podrem evitar que el model es *contamini*.

5) De fet, si el conjunt d'entrenament global és molt gran, fins i tot es pot crear un conjunt d'*exploració* autònom, diferenciat del de *prova* i del d'*entrenament* en sentit estricte. Ens pot ser útil aquest conjunt *exploratori* per visualitzar les dades en mapes de color. Hi buscarem correlacions entre els atributs. Ens ho permetrà el mètode *corr* de *Pandas*, que estableix (de -1 a 1) el coeficient de correlació (positiva o negativa) estàndard entre cada parell d'atributs. Només amida correlacions lineals (si un puja, l'altra també puja o baixa), però no les no lineals[58].

[58] La funció *scatter_matric* de *Pandas* traça per mitjà d'una matriu cada atribut numèric respecte d'un altre atribut numèric i pot revelar, a més de la tendència general, altres línies rectes menys evidents. En certs casos, pot convenir eliminar les dades que les generin per evitar que l'algoritme aprengui a reproduir aquestes *singularitats*.

6) Experimentar amb *combinacions d'atributs*: les dades disponibles tindran una sèrie d'atributs d'entrada, però res no impedeix que els puguem combinar i crear-ne de nous, segons les necessitats de la tasca que es vulgui automatitzar. Una vegada creats, podrem anar a la matriu de correlacions (*corr_matrix*) per veure les noves correlacions entre atributs.

7) La *preparació* de les dades perquè siguin un *aliment* adequat de l'algoritme es pot fer manualment o, també, gestionar per mitjà de funcions. Aquesta segona opció permetrà *transformar* de la mateixa manera qualsevol conjunt de dades nou. També provar amb facilitat diverses transformacions i veure quina combinació de transformacions funciona millor. En aquesta fase es poden separar els *predictors* i les *etiquetes* perquè no els hem d'aplicar, necessàriament, les mateixes transformacions.

8) *Netejar* de les dades: la majoria d'algoritmes no poden funcionar si falten característiques a les dades. Per solucionar-ho, tenim diverses opcions: desfer-nos del sector de dades que no té l'atribut (*dropna*); desfer-nos de l'atribut (*drop*) o establir algun valor per omplir els buits (*fillna*: pot ser un zero o una mitjana, entre altres opcions).

9) Atributs de *text* i *categòrics*: un atribut pot expressar els seus valors en text o en números (per exemple, baix, mitjà o alt, en el primer cas, i 0, 1 i 2, en el segon). En el cas dels textos, si les opcions són limitades (no il·limitades), serà un atribut *categòric*. La majoria d'algoritmes d'aprenentatge automatitzat prefereixen, però, treballar amb números, per la qual cosa acostuma a ser convenient convertir les categories o atributs que s'expressin en text. Ho podem fer amb el mètode *OrdinalEncoder*.

10) Cal tenir present, sempre, el risc que l'algoritme assumeixi que dos valors propers són més similars que dos distants. A vegades pot ser correcte. Per exemple, si la categoria *textual* era *dolent, mig i bo*, la conversió a *0, 1 i 2* conserva una correlació de significat en funció de la proximitat. Per contra, si l'algoritme ha de fer prediccions sobre el preu dels immobles i una categoria és la de la *proximitat al mar*, llavors *0* i *4* seran més similars que *0* i *1*.

11) *Escalat* de característiques: es tracta d'una de les transformacions més importants que acostuma a aplicar-se a les dades. Són convenients perquè els algoritmes d'aprenentatge automatitzat no solen tenir un bon rendiment quan els atributs numèrics

d'entrada tenen escales molt diferents (per exemple, si en un model per predir preus d'immobles, l'atribut de número d'habitacions i el d'ingressos mitjos de la zona tenen una forquilla numèricament asimètrica). El que es buscarà és, dient-ho d'una manera simplificada, reduir l'escala de tots els atributs a valors entre 0 i 1.

12) *Seqüències de transformacions*: hem vist diferents tipus de transformacions a què es poden sotmetre les dades. No són excloents. Un mateix conjunt de dades pot ser sotmès a diverses transformacions en el marc del mateix model. Això ens remet a la noció de *Pipelines* o seqüències de transformació[59].

6.6.4. Selecció del model i entrenament

A continuació seleccionarem un model i l'entrenarem. Comencem a parlar, ara sí, d'algoritmes en sentit estricte. Podrà tractar-se, per exemple, d'una regressió lineal simple, una regressió polinomial (amb diverses dimensions), un arbre de decisions o un *RandomForestRegressor*, entre altres opcions que s'aborden a l'annex 1. En podrà ser un o més d'un, per poder escollir, arribat el cas, el model que ofereixi el millor rendiment.

Ja hem indicat en altres apartats de la recerca que, paradoxalment, la majoria dels algoritmes que s'utilitzen, i que constitueixen el nucli del model (per ser-ne motor, l'emissor de les ordres), ja estan creats. Venen incorporats (*built in*) en les plataformes del sector, de manera que simplement els haurem de cridar per mitjà de la ordre *import*. Això no és necessàriament negatiu, ja que es tracta d'un conjunt molt ampli d'algoritmes lliurement disponibles i molt eficients, que són el resultat de molts anys de perfeccionament i posada en pràctica. Lògicament, sempre en podrem crear de nous o modificar i *personalitzar* els ja existents, però sovint simplement se'ls crida i incorpora al model mitjançant una simple línia de codi. Per això hem dit, en alguna ocasió, que convindria, potser, desmitificar el món de la IA. A vegades és més *simple* del que sembla.

Què significa, però, *entrenar* un model? Si ja hem *carregat*, *netejat* i *transformat* les dades i hem seleccionat i cridat l'algoritme, llavors l'únic que queda és trobar els valors (*ajustaments*) dels *paràmetres* rellevants que té el model per poder funcionar eficaçment amb qualsevol tipus de dades, siguin les ja disponibles o les futures. Per ajustar aquests paràmetres cal posar-lo en funcionament i ho farem (només podem fer-ho) amb les

[59] El mètode *num_pipeline.fit_transform* cridarà successivament les diverses transformacions programades.

dades ja disponibles, que ja hem separat o dividit en dos grups: el d'*entrenament* i el de *prova*[60]. Però la pregunta segueix pendent de resposta: entrenar significa trobar els valors dels paràmetres, però com es troben aquests valors? Doncs a partir d'una *mesura de rendiment*. I en què consisteix una mesura de rendiment? En calcular (per minimitzar-los) els *errors* del model. I com els calculem, els errors? La resposta només ens la poden donar les equacions i les funcions matemàtiques. Caldrà mesurar la distància entre dos vectors, el de les prediccions que fa el model en funcionament i el de valors objectius ja disponibles. Hi ha diversos instruments de mesura del rendiment, però són habituals el RECM (arrel quadràtica mitja, o *mean_squared_error*, molt freqüent en regressions) i el EAM (error absolut mig, adequat si hi ha diverses zones de dades amb valors atípics).

Si la mesura de rendiment de l'entrenament és de 0 errors, la primera impressió podria ser d'eufòria i creure que ja s'ha trobat el model perfecte per automatitzar una tasca determinada. Però és molt probable que no sigui el cas. Pot haver-se produït un *sobreajustament* (una excessiva assimilació) del model a les dades d'entrenament, que faci que tot i la inexistència d'errors en l'entrenament, no sigui adequat per *generalitzar* posteriorment a dades noves. És a dir, per fer prediccions amb dades diferents, futures i desconegudes. Tasca, per cert, per a la qual volem crear el model. Ja n'hem parlat a l'apartat 6.5.5 com un dels desafiaments actuals de l'aprenentatge automatitzat. Una possible solució serà buscar un altre model[61].

Deixant de banda resultats inicials òptims que puguin explicar-se per un sobreajustament o, simplement, per haver-la encertat a la primera, és probable que la mesura de rendiment sigui al principi baixa o no suficientment alta. Caldrà perfeccionar el model. Perfeccionament que es pot abordar des de diferents perspectives:

[60] Hem vist, també, que fins i tot se'n podia fer un tercer, el conjunt d'*exploració,* i aviat veurem una quarta tipologia, la de *validació*.

[61] La cerca pot fer-se amb el mètode de la validació creuada de *K iteracions* (amb el mètode *cross-val_score*). També podríem acudir a un *SVM* amb diferents *Kernels* (apartat 5 de l'annex 1). O, si som més ambiciosos, a una xarxa neuronal. L'objectiu seria seleccionar i anar entrenant d'entre 2 i 5 models prometedors i anar-los guardant (fent ús del mòdul *pickle* de *Python* o de la biblioteca *Joblib*).

1) Modificar els *hiperparàmetres*[62]. No es tracta de paràmetres propis de l'algoritme, sinó generals del model, que seran *permanents* i que no es veuran afectats o modificats, precisament, arran del funcionament de l'algoritme. Més aviat el condicionen[63].

2) Combinar (assemblar) els models que tinguin el millor rendiment per aconseguir, precisament, un millor rendiment global. Per exemple, en el cas dels *RandomForest* (apartat 4 de l'annex 1), que es componen per diferents arbres autònoms, es poden combinar els que tenen un millor rendiment.

3) Detectar la importància relativa de cada atribut[64] i, arribat el cas, eliminar els atributs menys rellevants.

4) Aplicar hiperparàmetres de *regularització* del model per mitjà dels quals s'entrenin, per exemple, 100 models diferents utilitzant 100 valors diferents de l'hiperparàmetre[65].

6.6.5. Prova del model

Una vegada hem entrenat i perfeccionat el model, fins a obtenir-ne un rendiment en principi satisfactori, caldrà llavors avaluar-lo en el conjunt de dades de *prova* (amb els seus *predictors* i les seves *etiquetes*), que prèviament havíem separat o dividit. Aquesta fase no té res d'especial. Consisteix en posar en funcionament el model (mesurar, de nou, el seu rendiment) però amb aquest *altre* conjunt de dades (el de *prova*) i constatar quin és el seu rendiment. Si és necessari, cal aplicar a aquestes dades les mateixes transformacions que hem aplicat al conjunt d'entrenament.

En aquesta fase de prova és important superar la temptació d'ajustar els hiperparàmetres per fer que el rendiment sigui més alt. Si ho fem, serà poc probable que, després, el model *generalitzi* bé amb dades noves.

---

[62] Ja n'hem parlat a la nota a peu de pàgina 53.
[63] Com acostuma a passar, aquesta acció es pot fer a mà o automatitzar per mitjà del mètode *GridSearchCV*, que ens ajuda a decidir amb quins hiperparàmetres volem experimentar i quins valors volem provar.
[64] Mitjançant, precisament, un *RandomForestRegressor* (apartat 4 de l'annex 1).
[65] Com que s'aplicarà moltes vegades a un mateix conjunt de prova, és poc probable que funcioni bé amb dades noves. Això es pot prevenir amb el mètode de validació *hold-out*, amb el qual es reté una part del conjunt d'entrenament per avaluar diferents models candidats i seleccionar el millor (és l'anomenat *conjunt de validació* o de *desenvolupament* o *dev*): s'entrenen diferents models amb diferents hiperperàmetres en el conjunt d'entrenament reduït i se selecciona el que té millor rendiment en el conjunt de validació. Després s'entrena el millor model en el conjunt d'entrenament *complet* (inclòs el de *validació*) i això ens dona el model final, que serà avaluat en el conjunt de *prova*.

6.6.6. Prellançament del model i sandboxes

Abans de llançar el model que ja tenim entrenat i provat satisfactòriament, caldrà reflexionar sobre allò que s'ha après en les fases prèvies, què ha funcionat bé i què no ho ha fet. Quines conjectures s'havien fet inicialment i si s'han confirmat, si s'han refutat o si segueixen obertes. I, en aquest últim cas, si això suposa una problemàtica que desaconselli el desplegament del model. En concret, cal identificar i ser conscients de quines són les limitacions del sistema. Què és el que pot fer i el que no, més enllà que hagi pogut donar molt bons resultats en els rendiments. Caldrà comparar si l'automatització de la tasca obté un resultat de major qualitat que la seva execució manual. També, el temps que s'estalvia realment.

Un context adequat per emmarcar les reflexions en què consisteix la fase de prellançament són les *sandboxes*, els plans pilots de prova en contextos gairebé reals, però de manera controlada i amb totes les garanties i salvaguardes de desactivació immediata del sistema en cas de necessitat. En aquest entorn real però de menor pressió es podran plantejar i intentar respondre les preguntes i dubtes que sorgeixin abans d'un eventual desplegament real definitiu. Si aquestes *sandboxes* són convenients amb caràcter general, podríem dir que en el sector judicial serien imperatives. Seria inimaginable la implantació d'eines d'IA judicial sense un període de prova previ en un context real.

6.6.7. Desplegament i monitorització

Si també se supera satisfactòriament la fase reflexiva (i empírica) del prellançament, només quedarà el desplegament del model. Però la feina no acaba aquí, ja que, després del desplegament, vindrà immediatament la imperiosa necessitat de *monitoritzar* el seu funcionament ja en un context completament real. I això pot requerir més feina que la feta fins el moment.

No cal dir que si s'implementen eines d'IA judicial, la seva monitorització serà tan o més important que la fase de creació i entrenament del model. Així ho imposa el sol fet que el model pugui autocalibrar-se a sí mateix i, per tant, modificar-se respecte de la seva configuració inicial. També ho imposa, però, la necessària preservació de les garanties processals i del dret defensa. Una acceptació acrítica i cega de les eines d'IA judicial basada en el sol fet d'haver superat la fase de prova i la de prellançament (inclòs el

corresponent *sandbox*) no seria acceptable. De fet, és probable que de les eines d'IA judicial que es puguin implantar, moltes necessitin, també, una constant actualització com a conseqüència dels canvis legislatius i jurisprudencials que, indefectiblement, no deixen de produir-se. Caldrà, en tot cas, un equip d'experts legals i informàtics per actualitzar els sistemes. I aquest mateix equip podrà encarregar-se de la necessària monitorització.

Tornant a l'àmbit de la IA general, cal dir que la monitorització pot ser, ella mateixa, automatitzada. Es pot escriure un codi perquè el sistema comprovi per si mateix i en intervals regulars el seu rendiment *en viu.* Es buscaria controlar la taxa d'error en els casos *nous*, amb dades o instàncies mai no vistes abans. És a dir, l'error de *generalització* o error *fora de mostres*. Quan aquest rendiment baixa per sota del llindar que es fixi, s'activaran les corresponents alertes. Això es pot donar quan hi hagi un descens brusc del rendiment (per problemes de *hardware* o d'infraestructura) o quan es produeixen descensos subtils durant molt temps que podrien passar desapercebuts. De fet, el món canvia i el model es va entrenar amb dades del *passat*. Podria ser que ja no s'adapti a les dades més actuals[66].

Una altra tasca de monitorització consisteix en l'avaluació constant de la qualitat de les dades d'entrada. Si el model s'automodifica a si mateix en funció del seu funcionament posterior al llançament, és obvi que una eventual manca de qualitat de les noves dades el poden malmetre. La degradació pot no produir-se immediatament i poden trigar a activar-se les alertes per baix rendiment. Per això pot ser convenient monitoritzar la mateixa *entrada* de les dades, per detectar-ne abans el seu potencial efecte danyós[67].

Ateses aquestes eventualitats, convindrà tenir còpies de seguretat de tots els models i de totes les versions que va adoptant cadascun d'ells, per utilitzar el concret model i la concreta versió que ofereix el millor rendiment. Per poder *tornar-hi*, si és necessari.

[66] Imaginem un sistema creat per detectar fotos de gats i que, amb el temps, al públic comencin a agradar-li noves races que no havien estat presents (o ho havien estat en una mesura molt petita) en les dades que van entrenar el model: potser caldrà actualitzar el conjunt de dades i tornar a entrenar el model.

[67] Així, es poden programar alertes per quan faltin entrades amb una determinada característica o atribut que es considerin rellevants. Per exemple, quan la mitjana d'entrada es desvia massa del conjunt d'entrenament o quan un atribut en principi categòric (amb opcions de valor tancades) comença a tenir noves valoracions no previstes.

També caldrà tenir còpies de seguretat de tots els conjunts de dades, per si cal tornar a un anterior quan el més recent incorpora excessius valors atípics.

6.6.8. Una conclusió sorprenent

Després d'analitzar les fases de contextualització, d'obtenció i tractament de les dades, de selecció, d'entrenament, de prova, de prellançament, de desplegament i de monitorització del model, podem arribar a una conclusió probablement inesperada: una part molt rellevant del treball se centra en la preparació de les dades i la monitorització del model. En l'*abans* i el *després* del treball referit pròpiament al nucli de l'aprenentatge automatitzat, l'algoritme. Els algoritmes són evidentment molt importants, però ho és més estar còmodes amb el procés global de creació del model. Pot ser suficient conèixer bé tres o quatre algoritmes, els més habituals, i centrar el treball en les complexes situacions que es poden donar en les altres fases per les quals necessàriament caldrà passar.

7. L'aparent inviabilitat del jutge-robot

7.1. És viable (i convenient) l'ús de la IA per dictar sentències de fons?

En aquest apartat entrarem, finalment, en el nucli de la recerca, la localització de possibles usos de la IA en l'àmbit de l'Administració de Justícia. I ho farem pel final. Ens situarem en una hipòtesi de màxims i ens preguntarem si és viable utilitzar-la per automatitzar l'acte essencial en què consisteix administrar justícia, la presa de la decisió final (la sentència) sobre els drets i els deures de la ciutadania. Això ens trasllada a la inquietant noció de *jutge-robot* de la qual ja ens hem intentat allunyar a la part introductòria de la recerca[68]. Podem avançar, ja, que la resposta a la pregunta que titula aquest apartat serà negativa i en dos sentits: tant pel que fa a la viabilitat tècnica com a la conveniència jurídica d'aquest tipus d'eines[69].

Recordem, però, que el futur i els possibles desenvolupaments de la IA, ara per ara imprevisibles, podrien refutar aquesta afirmació. En qualsevol cas, el resultat de la recerca s'ha orientat principalment a detectar possibles eines d'IA de *suport* de la oficina judicial i del tribunal en la *tramitació* del procediment. Fins i tot, de *suport parcial* en el dictat de la sentència final. Però un suport, a més de parcial, principalment referit a certes tasques complementàries (*burocratitzades*) que sempre cal realitzar quan es dicta una sentència. No és el mateix, certament, una eina d'IA que pretengui, ella sola, generar de manera íntegra una sentència i una eina d'IA que, per contra, simplement doni un suport en una tasca determinada, parcial, que ajudarà a construir finalment el document de la sentència. De fet, aquesta noció d'aportació i suport *parcials* i de *col·laboració* entre *agència humana* i eina d'IA és la que ens permet reconduir les *potencialitats judicials* de la IA a uns paràmetres raonables en termes de factibilitat tècnica, per una banda, i de

[68] Al capítol 10, relatiu a la IA judicial comparada, veurem algunes experiències properes a la figura del *jutge-robot*, ja en funcionament o només projectades, a països com la Xina o Estònia, entre altres. Veurem que el model xinès, si bé força avançat, no sembla que hagi de ser un mirall en el qual inspirar-se. L'estonià, tot i la profusió de notícies enlluernadores, no ha arribat, de moment, a implementar-se. En un altre nivell, al capítol 8 s'aborda l'anomenada *justícia predictiva*, més desenvolupada però amb força limitacions. Molt allunyada, en definitiva, de la noció de *jutge-robot* que ara ens interessa.

[69] Únicament es deixarà la porta oberta a possibles aplicacions en procediments civils molts específics i acotats legalment, repetits i estandarditzats, habitualment senzills, i sempre amb la posterior supervisió humana. Es tractaria sempre, i en casos molt restringits, de meres propostes o esborranys de resolució, pendents sempre d'una confirmació humana. Ho veurem més endavant a l'apartat 11.18.

necessària preservació dels drets processals implicats i de la *qualitat* de la justícia, per l'altra.

Així, podria ser útil, com veurem, generar inicialment esborranys de sentència que continguin, ja, tota la informació que les eines d'IA puguin extreure del Sistema de Gestió Processal i dels escrits i documents aportats, a l'espera, és clar, del redactat final de l'argumentació jurídica i de la decisió final per part del titular de la funció jurisdiccional. Això no implicaria, però, una automatització íntegra, ni tan sols substancial, del dictat de la sentència. Seria, merament, una eina de suport *parcial*. Ho detallarem més endavant. El que es vol analitzar, en aquest capítol 7, per contra, és la viabilitat tècnica i la conveniència jurídica d'una automatització *íntegra* del dictat de sentències o resolucions definitives. D'una automatització que sí que es podria reconduir a la noció de *jutge-robot*.

Començar l'anàlisi per la tasca judicial més *sensible* (el dictat de la sentència), a la qual se li negarà, amb caràcter general, l'aplicabilitat d'eines d'IA, més enllà de poder transmetre una certa sensació de *fracàs* de la recerca, permet contextualitzar-la adequadament: identificar els motius (tècnics i de fons) que impossibiliten o desaconsellen el dictat de sentències de fons amb eines d'IA ens dotarà d'un quadre de factors condicionants de la mateixa viabilitat d'aquestes eines per altres tasques menys *decisives* però que igualment han de dur a terme els tribunals i que, de fet, poden suposar una part important del volum total de feina. Així, en la mesura en què un acte previ de tràmit (admissió de la demanda, detecció de la possible falta de competència territorial, apreciació de la cosa jutjada, valoració de la concurrència d'un determinat pressupòsit d'una mesura cautelar, etc.) ja no presenti els problemes que sí que existeixen respecte de l'acte de dictar sentències, llavors podrem plantejar-nos seriosament la possibilitat d'aplicar-hi eines d'IA.

Abordem, ja, els motius pels quals no sembla viable, amb caràcter general, aplicar la IA a la tasca de dictar sentències. Són diversos i molt variats i la suma ponderada de tots ells semblen conduir-nos a la resposta negativa que ja hem avançat.

7.2. Superació del model sil·logístic en el dictat de sentències

Començarem amb algunes de les reflexions contingudes a l'article "Judicial Decisions and Artificial Intelligence", que, tot i tenir una certa antiguitat (Taruffo, 1998), van ser emeses per un referent en l'àmbit de l'argumentació jurídica i la valoració de la prova, que ens ha deixat recentment, el 10 de desembre de 2020. Sosté Taruffo que si un considera les notes de complexitat, incertesa, variabilitat, flexibilitat, càrrega de valors i discreció que caracteritzen innegablement el dictat d'una sentència, qualsevol intent d'aproximar-se als raonaments que la conformen a partir de regles i models lògics sembla condemnat al fracàs. La història de les teories *lògiques* del raonament judicial (que el conceben com una cadena de passes sil·logístiques) és en bona part la història de malentesos, manipulacions i derrotes. Si això és així respecte de la concepció lògica, l'escepticisme hauria de ser encara més intens si del que es tracta és d'aplicar lògica computeritzada o eines d'IA. Potser es tracta, simplement, d'una impossibilitat: qualsevol model lògic hauria de deixar de banda atributs importants del raonament inherent a la presa de decisió i seria, per tant, un model fals.

Aquest plantejament inicial no exclou, però, segons Taruffo, la conveniència de seguir quines són les tendències i avenços actuals de la IA en el camp del raonament jurídic. No podem descartar que alguns d'ells estiguin ben orientats. En aquesta línia, hem de superar la falsa i rígida alternativa tradicional entre deducció i irracionalitat, com si tot allò que no sigui deductiu passés a ser, necessàriament, irracional. Per contra, una decisió pot ser racional i fonamentada lògicament sense ser (com, de fet, passa habitualment) deductiva. Per tant, el marge d'exploració (racional) és, d'entrada, molt ampli.

7.3. Models computeritzats: casos senzills i casos difícils

El camp dels *models computeritzats* s'ocupa dels programaris que processen els factors rellevants en un procediment judicial determinat. Exigeixen introduir, en cada cas, les dades específiques del supòsit per tal que es pugui generar un resultat. Aquests models, propers als *sistemes experts* que hem analitzat al capítol 5, presenten diverses limitacions. Per una banda, la necessitat que els casos siguin freqüents i senzills i que no presentin variacions significatives[70]. I, per l'altra, que en cada fase de decisió s'han de preveure i definir per avançat totes les alternatives admeses. Sempre és possible

[70] És possible, certament, localitzar aquest tipus de procediments en l'àmbit de l'administració de justícia, però no és la situació més habitual o normal.

deixar oberta alguna elecció, però llavors es produeix una manca d'eficiència i serà necessària la intervenció *ad hoc* de l'usuari.

Deixant de banda els casos difícils (*hard cases*), fins i tots en els senzills (*easy cases*) poden no donar-se la resta de requisits. Per tant, l'espectre de *candidats* potencials a l'aplicació d'eines d'IA es va reduint. No podem oblidar, tampoc, que l'atribut de *simplicitat* d'un cas determinat és una qüestió de grau i d'eleccions valoratives, fet que dificulta significativament la possibilitat mateixa de fixar per avançat estàndards de simplicitat o complexitat.

7.4. Discrecionalitat forta o dèbil

Per altra banda, si del que es tracta és de racionalitzar decisions discrecionals, caldria diferenciar, segons Taruffo (1998), entre *discrecionalitat forta* i *discrecionalitat dèbil* o *regulada*. En la primera el tribunal és completament lliure de triar la seva decisió dins d'un rang d'alternatives teòricament il·limitat, en funció de les circumstàncies individuals del cas en qüestió. Per contra, en la *discrecionalitat dèbil* o *regulada* la llibertat judicial és només relativa, en el sentit que la decisió s'ha de prendre entre un inventari d'alternatives prèviament fixat per la llei, dins d'un rang quantitatiu prefixat també per la llei entre un mínim i un màxim, o segons uns estàndards o principis específics de la matèria, que la norma imposa. Pensem, per exemple, en la concreció judicial, dins d'un màxim i un mínim, de la pena prevista al codi penal per un delicte determinat, en consideració, també, de certs factors predeterminats, com ara l'edat, les condicions mentals o les condemnes anteriors de l'acusat. Un altre exemple serien les mesures cautelars, tant civils com penals, respecte de les quals la norma preveu quins són els paràmetres o factors a tenir en compte.

Doncs bé, la *discrecionalitat forta* no podria ser reduïda a un esquema lògic: la sentència acabarà acudint a una sèrie de principis, estàndards o cànons adequats, però aquests no poden ser racionalitzats per avançat, *ex ante*. En aquests casos els paràmetres a tenir en compte per derivar-ne la motivació que sempre s'ha de donar es generen necessàriament *a posteriori*. Per contra, semblaria d'entrada que hi ha més espai per a la racionalització lògica prèvia en la *discrecionalitat dèbil*, ja que coneixem, per avançat, els paràmetres que seran rellevants i el context dins del qual es pot moure la decisió.

Això no obstant, fins i tot en els supòsits de *discrecionalitat dèbil*, sempre quedarà necessàriament oberta la discreció de decidir si un cas és similar, o no, a un altre o si entra, o no, dins d'una tipologia de casos. De fet, hi hauria el risc, segons Taruffo (1998), que més que racionalitzar la discrecionalitat, l'estaríem, més aviat, eliminant. I reduir o excloure la discrecionalitat en la tasca de dictar resolucions judicials no és, només, una qüestió de mètode, sinó de *política criminal*.

7.5. Cercle hermenèutic: context fàctic i context legal

La lògica de les eleccions racionals inserides en les sentències opera en un *context* molt peculiar, el de la *funció jurisdiccional*. Aquest context imposa una articulació especial en l'estructura del raonament entre, per una banda, la valoració dels fets (de la prova i les inferències que connecten la prova i les conclusions fàctiques) i, per l'altra, la localització i interpretació de les normes que regulen el cas. Segons Taruffo (1998), el *context fàctic* del raonament podria ser abordat, amb les corresponents adaptacions, amb els conceptes i eines que se sol utilitzar per computeritzar els fluxos d'informació i de coneixement. O, dit d'una altra manera, amb eines de *coneixement cognitiu*. Per la seva banda, el *context legal* s'adaptaria millor a la lògica de l'argumentació racional i a la lògica *deòntica* (relativa a les estipulacions prescriptives). Aquestes dues esferes, la fàctica i la normativa amb què es desdobla el raonament judicial, interaccionen, a més, en unes fronteres difoses i mòbils. La seva és una *textura oberta*. Especialment, quan s'opera amb conceptes jurídics indeterminats (*deguda diligència, bon pare de família, bona fe*, etc.). Però també en la resta de casos. La *obertura* de la textura és permanent i estructural, inherent al raonament jurídic. Segurament la noció que millor capta aquesta peculiaritat és la del *cercle hermenèutic*, segons la qual no és possible, per exemple, fer una distinció nítida entre els fets i les normes, les valoracions fàctiques i les pròpiament legals, ja que el raonament *viatge indefinidament* d'unes a les altres. Sense les primeres, no es poden definir, pel cas concret, les segones. I al revés.

Veiem, en definitiva, que la dinàmica del raonament judicial és especialment *dialèctica* i *oberta*. Hi operen constantment hipòtesis, arguments i contra arguments (fàctics o legals, aportats per les parts o generats pel mateix tribunal) que constantment poden *reconduir* el camí a seguir (i, lògicament, el resultat final). En aquest model (complex però real), diferents tesis normatives s'enfronten successivament i poden arribar a entrar en conflicte. No perquè alguna o algunes d'elles siguin incorrectes en si mateixes, sinó

perquè el seu significat últim o precís, pel cas concret al qual s'apliquen, ve mediatitzat o condicionat, precisament, pel seu contacte amb la realitat.

Així, si ens situem en l'esfera de la *determinació del dret aplicable* pel cas concret, és indubtable que les actuals eines d'IA tenen avui una capacitat de cerca a través de les bases de dades incomparablement més ràpida i precisa que la dels humans. Però, és clar, una cosa és la *localització* de la norma aplicable i una altra força diferent és la concreció del seu sentit (la seva *interpretació*) pel cas concret. Podrà ser necessari qüestionar-se quina és la intenció *implícita* del legislador i indagar, més enllà d'una comprensió literal de la norma, certes ponderacions de tipus més moral o econòmic. És aquí on operen el *cercle hermenèutic* i les difoses fronteres de les esferes fàctica i normativa. I no és, ja, tan evident que les eines d'IA puguin desenvolupar adequadament aquesta tasca. Més aviat el contrari.

En segon lloc, si anem, ara, a l'esfera de la *determinació dels fets* i, en concret, a la de comprensió del llenguatge natural i del comportament humà, necessàries per a la valoració, per exemple, d'una declaració testifical, constatem que l'estat de la IA no sembla que pugui aspirar a equiparar-se a la percepció humana. No ens ha de sorprendre, si tenim en compte que, amb caràcter general, està reconegut a l'ordenament jurídic i per a la majoria de jurisdiccions el principi de *lliure valoració de la prova* (les anomenades regles de la *sana crítica*). Semblaria com si la concreció, comprensió i valoració dels fets del cas exigís, en una mesura major que en la *determinació del dret*, un agent realment *lliure* dotat de la capacitat d'operar amb paràmetres tan indeterminats com els de la *sana crítica*. I això ens allunya, necessàriament (i segurament per sort), de les eines d'IA.

Sempre es podrà contra argumentar, però, que la IA (especialment en les eines d'*aprenentatge profund* de les xarxes neuronals) el que pretén és *imitar* el comportament humà, assimilant com ha actuat en el passat per aplicar-ho a casos nous. Per tant, si aprèn com s'apliquen habitualment les regles de la *sana crítica*, podria fer-ho en un futur sense necessitat que l'eina estigui dotada d'una vertadera *llibertat valorativa*. Veiem, per tant, un conflicte potencial de gruix entre la concepció tradicional de la valoració judicial de la prova i la manera com aquesta tasca seria duta a terme (hipotèticament) per la IA.

7.6. Pot la IA valorar la credibilitat d'una testifical?

El cert és que sí que és concebible una tecnologia que reconegui allò declarat, ho passi de paraula a text i ho relacioni amb els fets rellevants del cas, en el sentit de si allò manifestat per un testimoni reforça, o no, la tesi segons la qual un fet determinat va poder passar, o no. Per contra, no sembla que hi hagi eines mínimament fiables que permetin avaluar, estrictament, la *força probatòria* o la *credibilitat* d'un testimoni: si bé existeixen, en el mercat privat, certes aplicacions que pretenen o afirmen fer-ho, no ofereixen, però, de moment, les garanties necessàries.

A títol de mer exemple, *DeepScore*[71] ha desenvolupat una aplicació mòbil, destinada a asseguradores o entitats financeres que presten diners, que suposadament pot valorar, en temps real i utilitzant tècniques de reconeixement facial i de veu, la credibilitat de les manifestacions dels seus clients. La finalitat seria detectar possibles fraus o mentides. L'eina se centra en els gestos i el to de veu en les respostes a un qüestionari de 10 preguntes. Indubtablement, eines com aquesta no disposen del necessari suport empíric i científic i, de moment, tampoc no han demostrat una eficàcia especialment elevada (la seva taxa d'errors o *falsos negatius* és massa alta, massa propera a una valoració merament aleatòria). Tampoc podem obviar els evidents problemes de privacitat que poden generar respecte de les dades biomètriques que processen. Però és en tot cas una mostra de cap on es dirigeix el mercat privat, que sempre anirà per davant d'eventuals i temptadores implementacions públiques.

Deixant de banda anècdotes comercials com l'exposada, segurament ens hauríem de fer una pregunta prèvia: si, excepte en casos flagrants, és extremadament difícil per un tribunal de carn i ossos valorar la credibilitat d'un testimoni, què ens ha de fer pensar que ho podrà fer (bé o millor) una màquina? I, encara més important: si, com hem vist, existeixen eines d'IA que suposadament poden realitzar aquesta tasca (de fet, l'aplicació *funciona* en el sentit que després d'analitzar unes imatges i una veu, ens dona una resposta), què ens ha de fer confiar en elles? Un nivell d'eficàcia del 80% en la fase de prova? del 90%? Però, aquest percentatge d'eficàcia el fixarem respecte de quina veritat absoluta o contrastable? Existeix aquesta? I què hem de dir dels problemes de privacitat? No queda afectada, de fet, la mateixa dignitat de la persona analitzada? O la de

[71] https://deepscore.ai.

l'Administració de Justícia, que, *derrotada*, delegaria a un programari una de les seves funcions *ancestrals*.

En definitiva, potser hi ha tasques (com la valoració de la credibilitat d'una testifical) necessàriament humanes. Potser ens hauríem de centrar més, per modernitzar realment l'Administració de Justícia, en limitar la pràctica de les proves testificals als casos que realment ho requereixen i, en aquests, aplicar-hi tota l'expertesa humana que sigui possible, amb les seves ineludibles limitacions, que són les que, per altra banda, ens fan humans.

7.7. Naturalesa derrotable del raonament Jurídic

Del que s'ha dit fins ara podem constatar que les eines més adequades per abordar la complexitat subjacent al raonament jurídic no seran, en cap cas, les de la lògica deductiva clàssica, que és *monotònica*, en el sentit que encara que afegim més informació o dades a una inferència, no canviarà la conclusió. Per contra, hi encaixen millor la lògica *no monotònica* (la conclusió pot canviar si afegim noves dades) i la noció de *derrotabilitat* dels arguments (una tesi que, inicialment, o *prima facie*, era la normativament preferible, pot veure's superada o desplaçada per una altra si afegim altres tesis o contra arguments). Així, encara que no s'acostumi a reconèixer, hem d'admetre que les previsions normatives estan subjectes a *excepcions implícites*, a circumstàncies que, tot i no estar-hi expressament enunciades, poden *derrotar* la solució normativa inicialment prevista. La conseqüència és que el contingut normatiu roman, sempre, potencialment *indeterminat*.

Dit d'una altra manera, els preceptes normatius serien condicionals l'antecedent dels quals no és una condició *suficient* per a la derivació de la solució normativa. I, si bé és cert que les lògiques computacionals més avançades estan degudament equipades per expressar aquest tipus de raonament, també sembla evident que aquesta complexitat intrínseca (ineludible, que no s'ha de pretendre eliminar) del raonament jurídic dificulta enormement l'eventual aplicació d'eines d'IA per generar sentències judicials. Com a mínim, fins que no es fusionin satisfactòriament els camps de la lògica computacional jurídica i de la IA. I, tot i que hi ha esforços en aquesta línia, no sembla que encara hagin reeixit.

7.8. Positivisme inclusiu: ponderació de drets i principis i constitucionalitat de les norms

Una clara i transcendental manifestació de l'eventual *derrotabilitat* del raonament jurídic la trobem en la realitat de l'actual estat constitucional, que reconeix una sèrie de principis i valors bàsics (entre altres, els drets fonamentals) que s'integren, no des de l'exterior, sinó com a elements constitutius, en l'ordenament jurídic. D'aquí deriva la noció de *positivisme inclusiu*, que expressa l'evolució del positivisme tradicional i liberal (que concep el sistema jurídic merament com el conjunt de lleis aprovades pel legislador que, a més, han de ser interpretades de manera estricte) fins a una realitat jurídica més moderna i complexa, definida per la interrelació entre aquest ordenament jurídic de rang merament legal i el *context constitucional* que l'emmarca, amb tots els seus valors, principis i drets. I que el condiciona, fins al punt de poder afectar la manera com ha de ser interpretada una norma determinada.

Es pot parlar, així doncs, fins a cert punt, d'una *obertura* del dret i del raonament jurídic a principis morals bàsics, ja que, en bona mesura, els principis i valors constitucionals comparteixen amb aquests un nucli de significat. Un significat que es tradueix en una sèrie de *pautes axiològiques* segons les quals cal obtenir, en cada cas i context, el màxim d'efectivitat dels principis o drets de què es tracti.

Pensem, per exemple, en els casos en què dos drets constitucionals entren en conflicte i cal acudir a la *ponderació*. Es tracta d'un tipus de raonament jurídic (el de la ponderació) que ens és, de fet, d'especial interès perquè presenta unes particularitats que, paradoxalment, semblarien que el fan precisament adequat per a una eventual computació o automatització. En concret, perquè s'articula en diferents fases algunes de les quals impliquen operacions aparentment *traduïbles* a termes aritmètics (cal determinar, entre altres extrems, quin és el pes o la importància respectiva dels drets en conflicte o el seu concret grau d'afectació). De fet, ja s'han utilitzat (fins i tot a l'àmbit judicial, especialment el sud-americà) *formulacions matemàtiques* per expressar els arguments jurídics que hi ha darrera la ponderació. I si acceptem com a viables aquestes formulacions o *reduccions*, també seria factible la seva inserció en una eina d'automatització judicial. Sembla evident, però, que aquestes *reduccions* no són sinó el resultat d'una prèvia apreciació valorativa (per exemple, una que determini en quin grau s'ha vist afectat un dret en un context determinat) que, en sí mateixa, no pot ser

automatitzada, per més que la puguem expressar matemàticament *a posteriori*. No podem afirmar, per tant, que el terreny de la ponderació sigui, per la seva mateixa naturalesa, específicament fèrtil per a possibles temptatives d'automatització. O si ho afirmem, serà al preu de banalitzar injustificadament aquest tipus de raonament judicial.

Una altra manifestació pràctica dels efectes del *positivisme inclusiu* que estem analitzant són aquelles situacions en les quals, una vegada identificada la norma aplicable, sorgeixen dubtes sobre la seva constitucionalitat precisament per poder afectar excessivament un dret o principi constitucional. En aquests supòsits, caldria abordar, també, la disjuntiva de si és possible una interpretació de la norma *conforme a la constitució* (en el sentit de delimitar-ne les seves implicacions de tal manera que el resultat que generi sigui compatible amb el contingut essencial d'un dret constitucional) o si, per contra, això no és possible i cal acudir als mecanismes constitucionalment previstos, com és el plantejament d'una qüestió d'inconstitucionalitat davant del TC.

Es tracta, en definitiva, de nous factors de *complexitat* que clarament dificulten la possibilitat que eines d'IA generin, per sí soles, sentències judicials. I cal tenir present que els dubtes de constitucionalitat d'una norma no es poden delimitar o fixar per endavant, sinó que es manifesten i s'han d'abordar en cada cas concret, quan ja s'ha iniciat el procediment i cal dictar una sentència de fons: una norma pot semblar impecable, en abstracte, en termes de constitucionalitat però generar certs dubtes quan s'enfronta a una nova situació fàctica i jurídica específica. Serà llavors quan caldrà fer-se les preguntes corresponents. No es poden preveure, ni molt menys exhaurir, per avançat, tots els supòsits en els quals es pot produir aquesta eventualitat, fet que, per motius evidents, dificulta enormement qualsevol tipus d'automatització, fins i tot encara que s'acudeixi a eines avançades d'aprenentatge automatitzat.

7.9. Qüestions prejudicials davant del TJUE

Podem completar el quadre de *complexitats potencials* que pot haver d'abordar una sentència judicial amb una ràpida referència a l'eventualitat que la norma interna ja identificada com aplicable al cas presenti, una vegada interpretada, dubtes de si és compatible, o no, amb una altra norma europea igualment aplicable o de la qual en sigui una transposició. La via processal, en cas de confirmar-se aquesta situació, seria la del plantejament d'una qüestió prejudicial davant del TJUE. El que ens interessa, però, posar

de manifest, de manera similar amb el que succeïa amb la qüestió d'inconstitucionalitat, és en quina mesura una eina d'IA podria processar el tipus de raonaments que és necessari abordar en aquests *contextos hermenèutics*[72]. La resposta sembla que ha de ser negativa. I cal tenir present que es tracta de fases *necessàries*, *obligatòries*, del procediment de dictar qualsevol sentència. Que en la majoria de casos no es faci cap menció a aquestes qüestions implica, implícitament, que el tribunal no ha detectat cap problemàtica de constitucionalitat o de compatibilitat europea, però no significa que no hagi existit la fase. Aquesta ha de ser, sempre, possible. Ha de poder donar-se eventualment. I no sembla, com a mínim de moment, que la pugui protagonitzar una eina d'IA.

Context de descobriment i context de justificació

Hem de recordar, també, la tradicional distinció (molt criticada però mai refutada del tot) entre *context de descobriment* (el raonament que ha generat realment la decisió) i el *context de justificació* (l'explicació expressament recollida a la sentència). No sempre coincideixen. I és evident que si una eina d'IA ha de ser utilitzada com a suport pel dictat d'una sentència i el seu algoritme subjacent es basa en les dades extretes de sentències anteriors, només podrà operar, en el millor dels casos, amb el context de *justificació*.

És cert que quan un tribunal *humà* ha de dictar una nova sentència i acudeix a les sentències anteriors sobre casos similars, també acudirà preferentment al text, a la *justificació* expressa (que és la que consta a les bases de dades). En última instància, només qui va dictar les sentències anteriors té la possibilitat d'accedir al context de *descobriment*, que, per aquest sol fet i la seva manca absoluta de publicitat, pot arribar a ser jurídicament irrellevant. Podria ser-li d'interès, només, a la sociologia jurídica.

Dit això, fins i tot respecte del context de *descobriment* serà diferent la posició o perspectiva entre un observador extern humà i un d'artificial: l'eina d'IA ni tan sols podrà plantejar-se l'existència de l'esfera de *descobriment*. Un ésser humà, per contra, si bé

[72] Pensem, per exemple, en les diverses qüestions prejudicials plantejades per tribunals espanyols davant del TJUE per temes relacionats amb el dret de consum i com el legislador, o el mateix TS, han actuat a l'hora de transposar o interpretar certes normes europees, en especial la Directiva 93/13/CEE. En uns casos es plantejava si no permetre al·legar l'abusivitat d'una clàusula de consum en l'execució hipotecària era compatible, o no, amb la Directiva; en uns altres, si ho era reconèixer el dret a recórrer en apel·lació només a una de les parts del contracte; o, per exemple, si protegia adequadament els interessos dels consumidors limitar en el temps els efectes de la nul·litat de la clàusula terra.

tampoc no la pot conèixer (ni en detall ni de manera substancial), sempre podrà però especular, en el cas concret, si es pot haver produït un *distanciament* entre *descobriment* i *justificació*. Fins i tot, a vegades, es poden localitzar en el text de la sentència expressions o subraonaments aparentment il·lògics o inconsistents, innecessaris, que poden ser manifestacions del procés de descobriment que no ha arribat a expressar-se de manera completa o explícita en la sentència: el tribunal *humà* que interpreta la sentència (i que, potser, n'ha de dictar una altra en un cas relativament similar) sí que podrà *especular* sobre aquest tipus de qüestions. I aquestes especulacions podran arribar a influir en la manera d'interpretar la sentència anterior, el *precedent*. Potser per descartar-lo. I, tal vegada, incidirà també en el *nou* raonament jurídic que passarà a ser el nucli de la *nova* sentència que s'ha de dictar.

7.10. Una motivació completa, consistent i coherent

Situem-nos, ara, en l'exigència que tota sentència judicial disposi d'una motivació jurídica suficient. No ha de ser extensa. Pot ser succinta i sintètica. Però ha d'existir. Tant si la sentència la dicta un tribunal humà com una eina d'IA. Es tracta d'un imperatiu no *negociable* de l'estat de dret, el dret de defensa i el principi de seguretat jurídica. Aquesta justificació o *motivació* hauria de reunir, com a mínim i seguint Taruffo (1998), tres atributs:

a) Ser *completa:* que cadascuna de les seves afirmacions rellevants estigui, expressament i adequada, justificada.

b) Ser *consistent:* que els arguments que utilitza no entrin en conflicte entre sí.

c) Ser *coherent*: que els arguments utilitzats es corresponen amb la naturalesa de les qüestions a tractar.

Veient les elevades exigències d'aquests atributs de la motivació judicial, sembla que el desafiament per als informàtics que treballen en el camp de la IA legal és, certament, molt elevat. Especialment perquè, de fet, moltes eines d'IA que podrien ser, hipotèticament, aplicades a la justícia, les anomenades de *justícia predictiva* (amb les quals es pretén anticipar o preparar una resposta judicial), operen amb paràmetres estadístics i no de causalitat. O, fins i tot si els qualifiquem de *pseudo causals*, poden ser

molt difícils de conèixer. O, directament, impossibles d'esbrinar, com en el cas de les *caixes negres* de l'aprenentatge *profund* i les xarxes neuronals (apartat 6.2.5). Per tant, si, directament, no tenim accés a cap justificació o explicació de com ha funcionat l'algoritme, ja no té sentit preguntar-nos si és *completa*, *consistent* i *coherent*. La incompatibilitat amb l'actual concepció del dret a la tutela judicial efectiva sembla, així doncs, més que evident[73].

A més, pels casos en què puguem accedir a una certa informació o coneixement de com ha funcionat l'algoritme i de per què ha arribat a un determinat resultat, del que disposarem serà, merament, d'una certa *explicació estadística* i *causal* de la seva operativa i de les correlacions en què s'ha basat, però no d'una motivació o justificació jurídica o legal equiparable a la idea de *motivació* judicial. Ambdues nocions (*explicació algorítmica* i *motivació jurídica*) operen, en definitiva, en dues esferes o mons diferents[74].

7.11. Aparent inviabilitat de la generalització del jutge-robot

No sembla factible, en definitiva, atès l'estat actual de desenvolupament dels instruments d'IA, automatitzar la creació d'arguments jurídics que siguin *complets*, *consistents* i *coherents*. Que es preguntin, també, si hi ha motius per plantejar una qüestió prejudicial europea o d'inconstitucionalitat. I que, a més d'articular adequadament el *context fàctic* i el *legal*, abordin satisfactòriament els ineludibles problemes de *discrecionalitat* (fins i tot en la *discrecionalitat dèbil*) que es produiran, sigui per diferenciar els casos *senzills* dels *difícils* o per identificar el grau de similitud d'un cas anterior amb el que s'ha de resoldre. Per no parlar de la indeterminació derivada del *cercle hermenèutic* a través del qual es manifesta la *textura oberta* del dret. O de la possibilitat que el sentit inicial d'una norma pot ser *derrotat* per noves dades que s'introdueixin en el raonament, el qual, fins i tot, com hem vist, por arribar a haver-se d'obrir al *debat moral*.

No sembla, certament, que la IA pugui fer front a aquest conjunt de factors de *complexitat*, que poden donar-se, o no, i, si ho fan, fer-ho en diferents intensitats. De fet, la mateixa determinació de si es dona, o no, un d'aquests factors és ja, en sí mateix, un factor de complexitat. No hem d'oblidar que no es tracta de dificultats del raonament

[73] A l'apartat 4.5, quan analitzàvem el dret de defensa a la llum del principi de *contestabilitat algorítmica*, hem abordat amb més detall les situacions de *dèficits motivacionals* que es poden donar quan opera una eina d'IA judicial.
[74] Aquesta idea es desenvolupa a l'apartat 8.2.

jurídic que calgui superar o eliminar, sinó de les seves notes intrínseques. Ens trobem, per tant, davant d'un probable límit insalvable de la IA. Com a mínim, respecte de la majoria o una part significativa de les decisions judicials que es prenen als jutjats. Pretendre, per contra, una eventual aplicació extensiva d'eines d'IA per al dictat de sentències (per a tots els casos en què sigui tècnicament viable) ens abocaria, ineludiblement, a una degradació significativa de la *complexitat real* del cas de què es tracti i al risc d'arribar a una decisió final imprecisa, inadequada i, en definitiva, empobrida i d'escassa qualitat.

La implementació de certes eines d'automatització no pot portar-nos a la *massificació* de resolucions judicials *empobrides*. No hem d'oblidar, a més, que fins i tot en els casos que es prevegi un control humà (del tribunal) posterior a l'automatització de la sentència (el que s'anomena una certa *agència humana*), sempre hi haurà el risc que es consolidi la tendència a acceptar acríticament allò que ens ha generat el sistema. En aquest sentit són interessants estudis com el de Dijkstra (2001) que mostren aquesta tendència (humana?) a deixar-se persuadir pel poder d'aquestes eines atesa la seva suposada major objectivitat o racionalitat.

Per tot plegat, potser acaba intuint Taruffo (1998) que el tipus de procediments judicials susceptibles d'aplicar-hi eines d'IA es trobarien en aquelles àrees (escasses) en les quals l'Administració de Justícia s'aproxima més al que podríem qualificar d'*administració burocràtica*. Més endavant, al capítol 11, veurem quins casos poden ser qualificats de *burocràticament automatitzables*. Però ho farem al final: hem començat per la teulada (dictat de sentències) i més endavant continuarem des de baix, analitzant senzilles tasques de tràmit processal que aparentment són més adequades per aplicar-hi eines d'IA. A mesura que les anem identificant, anirem *pujant* en el grau de complexitat, fins arribar, potser, a algunes sentències amb components repetitius, senzills i delimitats que puguin permetre algun tipus d'automatització. Aquest camí ens porta, però, primer, a l'anàlisi de la *justícia predictiva*.

8. Justícia predictive

8.1. Distinció entre jutge-robot i justícia predictive

A l'apartat anterior hem analitzat l'aparentment inviable figura del *jutge-robot*. És a dir, d'un recurs d'IA que realitzi de manera completa la tasca en què consisteix dictar una sentència, d'una manera equiparable a com ho fa un tribunal humà. Per contra, ara abordarem una realitat en part diferent. O més específica: l'existència, ja real, de certs programaris que sostenen utilitzar IA i que prediuen (amb més o menys encert i amb més o menys precisió) allò que resoldrà un tribunal en un cas futur. I ho fan en funció de comportaments anteriors codificats en forma de dades i que el sistema processa estadísticament. Es tracta de l'anomenada *justícia predictiva*.

Intentem, primer, ubicar aquesta *justícia predictiva*. Contextualitzar-la. Hem d'acudir, amb aquesta finalitat, a l'anomenada *legaltech*, que es nodreix, amb caràcter general, de dades legals que poden pertànyer a tres terrenys diferents que tenen, a la vegada, finalitats ben diverses:

a) Documentació referida a la prova (*e-disclosure* o descobriment electrònic).

b) Jurisprudència i legislació en la qual localitzar recursos per al raonament jurídic (*argumentation mining* o extracció i recopilació d'arguments).

c) Jurisprudència i legislació per *predir* el resultat de casos futurs. Aquest seria el terreny estricte de la *justícia predictiva*.

En una espècie d'intuïció del que, en un futur, seria la justícia predictiva, deia el jutge americà Oliver Wendell Holmes, en una coneguda afirmació a *The Path of the Law* (1897), "les profecies del que faran realment els tribunals, i res més pretensiós que això, és el que és, per mi, el dret. Potser l'home d'impremta gòtica és l'adequat avui per l'estudi racional de la llei, però l'home del futur és l'estadístic i el mestre de l'economia".

En aquest capítol abordarem, així doncs, com poden articular-se dret i estadística per fer prediccions jurídiques. Intentarem esbrinar si l'home *estadístic* és, com va predir Holmes, un bon *racionalitzador* de la llei.

8.2. Naturalesa estadística, no jurídica, dels models predictius

Ens recorda però Hildebrandt (2019) que les tecnologies predictives, tot i que proclamen que poden anticipar de manera fiable el sentit de les sentències futures, no actuen en el nivell del contingut estricte (significat jurídic) dels textos que processen. Per contra, es basen en dos tipus de dades, les *sensorials* (obtingudes amb sensors, com ara el to de veu[75]) i les *conductuals* (com ara els vots emesos amb anterioritat). Les segones ens poden interessar més. En cap cas, però, operen d'una manera equiparable a la humana, en termes de percepció, raonament i cognició, amb independència que pretenguin arribar a conclusions similars.

Si ens centrem en la justícia predictiva *conductual*, la que se centra en determinats comportaments passats i presents dels tribunals per predir-ne, precisament, el comportament futur, el primer que cal destacar-ne, com a possible objecció o pressupòsit qüestionable, és que pretén, a més de generalitzar a casos futurs, ser consistent (vàlida) a través del temps i de diferents tribunals. És a dir, pressuposa que una conducta determinada constatada en el passat probablement es reproduirà, si es donen certes condicions, en el futur en el mateix tribunal o en un altre.

Baixem, ara, a una escala més petita: aquestes prediccions diferencien la *variable objectiu* (allò que volem saber, la que es refereix al resultat del cas: que podria ser la revocació o confirmació d'una sentència apel·lada) i les *variables d'entrada* (*feature set*) amb les quals la primera estaria interrelacionada. Hi ha, però, segons Hildebrandt, un condicionant molt rellevant: les preguntes que ens hem de fer per construir el model predictiu s'han de formular, necessàriament, d'una manera que permeti el processament numèric. Quedarien excloses, per tant, formulacions en termes ambigus o més propis de la *textura oberta* tan característica (fins i tot inherent) dels problemes legals. El dret sembla requerir, com a element constituent, una certa *discreció* que, per contra, hauria de ser eliminada en el procés de formalització previ a l'aplicació d'eines *predictives*. Ens trobem, certament, davant d'un problema (una limitació) no pas menor de la IA judicial.

[75] Hildebrandt (2019) posa com exemple l'estudi realitzat per Dietrich, Enos and Sen, *Emotional Arousal Predicts Voting on the U.S. Supreme Court*, en el qual s'utilitzen les dades de potència de veu dels jutges obtingudes durant les vistes per predir el vot i conclouen que com més alta és l'excitació oral respecte d'un lletrat o lletrada en comparació a un altre, més possibilitats té de guanyar el cas.

Com funciona, en un sentit més tècnic i precís, un sistema predictiu? Com busca les correlacions rellevants entre el *feature set* i la variable *objectiu?* Doncs bé, pot utilitzar, entre altres algoritmes, un *Random Forest Classifier*, que seria un conjunt d'arbres estadístics que aprendrien de manera autònoma (cada arbre o *tree*) a realitzar la correlació indicada (a partir de funcions matemàtiques que capturin la correlació tan bé com sigui possible) per, després, obtenir-ne la mitjana de tots ells (el bosc o *forest*[76]).

Per tant, veiem, seguint Hildebrandt, que el model ni *explicarà* per què el vot dins del tribunal ha estat un o un altre (la correlació no és la causa) ni *justificarà* la decisió (no opera amb raonaments jurídics). Només prediu una decisió binària. I, per valorar-ne l'eficàcia o utilitat, se sol comparar el resultat obtingut amb un *model 0* que seria el resultat arbitrari o aleatori de llançar una moneda a l'aire.

8.3. Experiències reals de justícia predictive

En un conegut estudi de justícia predictiva (Alertas et al., 2016) sobre les sentències del TEDH, es van utilitzar tècniques de *Processament del Llenguatge Natural* (*PLN*) amb les quals es va *entrenar* un algoritme amb els textos de les sentències referides a certs articles del CEDH per tal de preparar-lo per predir, en termes binaris (violació o no violació del dret), el resultat de sentències futures. L'estudi va concloure que la predicció va ser correcte (precisa) en un 79 % dels casos. Segons Hildebrandt (2019), l'any 2019 un estudi similar, però amb més dades (sentències) utilitzades (es va passar de 600 a 11.500), va utilitzar xarxes neuronals i sistemes d'aprenentatge no supervisat en la cerca de prediccions més àmplies (també es buscava predir la violació o no d'un dret, però les referències eren a qualsevol article del Conveni) i més precises (a més de predir si s'apreciaria o no la vulneració, es concretaria el grau d'importància del cas en una escala d'1 a 4, en una dada aportada pel mateix tribunal). En l'estudi es van provar quatre tipus diferents de xarxes neuronals, entre elles la coneguda BERT, que processa i analitza, amb nivells de complexitat i estratificació molt elevats, com certes paraules potencialment rellevants estan inserides en els textos. En aquest cas, la precisió obtinguda hauria estat del 82%.

[76] A l'apartat 4 de l'annex 1 s'aborda amb més detall en què consisteixen els algoritmes *Random Forest*.

És també d'interès l'experiència realitzada pels tribunals d'apel·lació de Douai i Rennes de França, que durant uns mesos van fer un seguiment de diversos judicis amb un programa *predictiu* juntament amb un equip de jutges[77].

8.4. Justícia predictiva, explicabilitat algorítmica i motivació judicial

Dit això, cal fer una important precisió que ja avançàvem en part: és cert que alguns models, diferents de les xarxes neuronals, ens permeten, fins a cert punt, *explicar* o *interpretar* la seva manera de funcionar o fins i tot els factors que en major mesura determinen els seus resultats (*output*). Però, com ens recorda Hildebrandt (2019), aquesta *transparència algorítmica* no pot ser equiparada, en absolut, amb una explicació o justificació *jurídica* en termes equiparables a la *motivació* que se sol exigir en un estat de dret a les resolucions judicials com un component del dret a la tutela judicial efectiva: es tracta, més aviat, d'una mera verificació de la relació causal o incidència entre certs atributs que s'introdueixen en el model i el resultat final. Un raonament legal (jurídic) es construeix, per contra (o com a mínim de manera principal), en termes de causalitat entre certs arguments i certes conclusions. Tota decisió jurídica ha de basar-se en el compliment d'una sèrie de condicions legals, en un complex equilibri entre la rellevància respectiva i interrelacionada de les qüestions de dret i les de fet[78]. I una conseqüència ineludible és que els fets no venen *donats*, sinó que s'han de *construir*, de nou, cada vegada que s'aborda un raonament jurídic.

És cert que molts casos que arriben al jutjat presenten notes d'estandardització, tant en els fets com en les normes que els són aplicables, però aquesta realitat no exclou que sempre calgui abordar, encara que sigui de manera inconscient, el *cercle hermenèutic* que connecta els uns amb les altres. I no sembla que els paràmetres de *causalitat estadística* amb els quals operen els principals estudis de justícia predictiva realitzats fins al moment siguin els més adequats per abordar aquesta tasca. I no ho són perquè, seguint Hildebrandt (2019), no poden operar d'una altra manera que no sigui processant els textos legals com a meres *dades*. No poden manejar aquestes dades amb alguna dosi d'empatia o sentit comú, tan essencials en la pràctica judicial. I tampoc no poden aplicar, per definició, el *coneixement jurídic tàcit* tan emprat i del qual se'n parla tan poc,

[77] Experiència citada a la *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment,* aprovada pel CEPEJ.
[78] A l'apartat 7.5 hem analitzat l'aparent inviabilitat de la figura del *jutge-robot*, entre altres motius, per l'ineludible *cercle hermenèutic* entre fets i normes amb el qual opera el raonament jurídic.

precisament per ser tàcit i no estar explicitat ni a les normes ni, sovint, a les resolucions judicials.

De nou, podem introduir un matís i valorar la viabilitat d'aplicar certes eines predictives en funció de la tasca judicial de què es tracti. Així, podríem descartar, sense necessitat de massa argumentació, utilitzar en els jutjats (és a dir, per part dels tribunals) eines com les proposades a l'estudi sobre la jurisprudència del TEDH, per determinar binàriament, per exemple, si en un cas s'ha produït, o no, la vulneració d'un dret fonamental. O si s'ha d'estimar, o no, la demanda. Més interès poden generar, per contra, propostes en curs de desenvolupament com la francesa de *DataJust*, centrada en el càlcul estadístic de la indemnització que per danys corporals podria reclamar una víctima d'un accident. Els dubtes i problemàtiques seguirien existint, certament, però la negativa ja no seria, probablement, tan absoluta o ràpida[79].

De fet, a aquest tipus d'eines podria donar-s'hi un ús no estrictament *judicial*, sinó complementari, *extra* o *prejudicial* (per exemple, com veurem a l'annex 2, en els *ODR*). Potser la ubicació funcional més raonable d'aquest tipus d'eines és aquesta fase preprocessal en la qual pot tenir interès conèixer, encara que sigui de manera aproximada (*estadística*), quina probabilitat d'èxit (o de fracàs) pot tenir un cas davant d'un tribunal. O quin podria ser l'import raonablement esperable.

En un altre ordre de coses, aquestes eines predictives poden ser útils, també, per tenir coneixement de quines són les línies jurisprudencials majoritàries que es van generant en cada moment. I, de fet, que el sistema judicial *sàpiga* que està sent constantment *observat* (encara que sigui per eines estadístiques de predicció) pot generar una *pressió* positiva que faci augmentar la qualitat de la justícia que s'administra.

Convé tenir present, però, que si s'opta per implementar en algun sentit una eina predictiva, caldrà que es nodreixi de resolucions procedents de totes les instàncies (inclosa la primera instància), circumstància que, com és sabut, no s'acostuma a donar a les bases de dades actualment disponibles.

[79] S'analitza amb més detall la proposta del *DataJust* a l'apartat 10.4.3.

Per últim, no pot ser ignorat el *data-snooping* amb el qual operen aquest tipus de sistemes: han de *descartar* (no detectaran) aquelles resolucions que no són *significatives* pels models analítics (els paràmetres) predefinits. És a dir, aquelles que, per la manera d'haver estat redactades (no recullen suficientment el cas en litigi, contenen un raonament jurídic molt sintètic, etc.), no s'hi correlacionen suficientment. En sentit invers, el sistema només seleccionarà les resolucions que sí que són *significatives.* No cal dir que aquesta selecció *tècnica* no ha de correspondre's necessàriament amb criteris de rellevància *jurídica*.

8.5. Perfils individualitzats de jutges

La identificació expressa, a la resolució, dels noms dels jutges o jutgesses que la dicten és un component del principi de publicitat judicial: la imprescindible imparcialitat objectiva que s'ha de preservar seria inviable en un altre cas: l'autor de la sentència ha de poder ser identificat. El fet que aquesta informació es traslladi, més enllà del coneixement de les parts, a les dades obertes (al *Open Data,* a les bases de dades) genera, però, nous interrogants.

Cal diferenciar, en aquest punt, entre la mera possibilitat (ja existent) de fer una cerca de resolucions judicials dictades per un jutge o jutgessa determinat, i la creació de perfils individualitzats de jutges amb finalitats *predictives*. Són els segons els problemàtics: serà viable o és ja viable (o caldrà adoptar mesures en contra) fer cerques de la informació estrictament jurisdiccional disponible a les bases de dades i referida, no a un tribunal concret, sinó a un jutge determinat? És a dir, podran obtenir-se, de manera diferenciada i destacada, les seves tendències interpretatives individuals? Algú podrà plantejar-se creuar aquesta informació amb el contingut que es pugui obtenir d'altres fons, com les xarxes socials, per tal d'identificar el *perfil* ideològic o polític del jutge o jutgessa i derivar-ne, hipotèticament, un possible biaix rellevant per la causa? Els desafiaments i les incògnites són, com veiem, importants[80].

En un altre nivell, caldria abordar una problemàtica afegida: en els tribunals col·legiats dels països en els quals no estigui consolidada la pràctica dels vots particulars o dissidents, pot ser no ajustat a la realitat atribuir a un membre del tribunal el sentit d'una

[80] La llei francesa de 23 de març de 2019 per la reforma de la Justícia prohibeix, en el seu art. 33, la creació de perfils individualitzats de jutges amb fins predictius.

sentència amb la qual, potser, no va estar-hi d'acord durant la deliberació. Caldria diferenciar, per tant, entre les resolucions dels jutjats unipersonals (indiscutiblement atribuïbles a qui les dicta) i les dels tribunals col·legiats (cas en què s'hauria d'informar sobre qui n'és el ponent i qui és, simplement, membre del tribunal).

9. El dret és text: processament del llenguatge natural (NLP)

9.1. Les potencialitats raonables de la IA judicial

Després d'abordar l'aparent inviabilitat de la generalització de la figura del *jutge-robot* i de constatar les elevades limitacions (quant a la seva funcionalitat i els seus possibles usos) de la *justícia predictiva*, ens començarem a centrar en el que podríem qualificar d'usos potencials raonables o estrictament factibles d'una IA judicial. Aquests usos partiran d'una realitat molt òbvia però fonamental: el dret és, essencialment, text. Consisteix bàsicament en text. Tant les normes que s'aproven com les resolucions judicials que es dicten interpretant-les. Tant en els sistemes *continentals* com en els del *common law*. Fins i tot en el cas del costum com a font (residual) del dret, si bé no és estrictament *textual* en la seva gènesi, acabarà sent recollida en forma de text per la jurisprudència que la reconegui o constati.

Així, tots aquests *textos*, a més de ser creats, publicats i llegits per éssers humans, també poden ser *llegits* i *processats* per màquines que operen de manera automatitzada. Per programaris que *processen el llenguatge* (els textos) i que són la base de l'anomenat *processament del llenguatge natural* o *NLP* (*Natural Language Processing*). Un processament automatitzat que, degudament implementat, pot traduir-se, ara sí, en usos judicials de suport a la tasca jurisdiccional.

Per aquest motiu, el dret ha estat sempre un camp atractiu per a les tecnologies del llenguatge. De fet, serà precisament el dret qui portarà al límit les potencialitats del *NLP*. Al mateix temps, si un analitza l'estat actual de la tecnologia digital utilitzada als jutjats, s'adona de fins a quin punt s'estan *desaprofitant* aquestes potencialitats de tractament dels textos amb els quals no deixem de treballar. Perquè hi treballem, de moment, només com a elements *passius*, *inactius* o *no significatius*. Els traslladem o desplacem d'un punt a un altre del sistema de gestió processal i els ubiquem en un apartat determinat de l'expedient judicial electrònic. En una operativa que, tot i ser plenament *digitalitzada*, és eminentment *manual*. I és precisament en aquest punt on sí que sembla que hi podria haver un camp ampli per al desenvolupament tecnològic de suport de la tasca judicial.

Però per adquirir-ne la deguda consciència, caldrà, abans, abordar amb més detall en què consisteix, realment, la *NLP*. De fet, és possible que en alguns casos eines de *NLP*

semblin merament vies de mer aprofundiment en la *digitalització* judicial. Ja hem vist a l'apartat 2.2 que les fronteres entre IA i digitalització no són massa nítides. En qualsevol cas, però, l'important no són els noms o les etiquetes, sinó la utilitat de les eines de què ens dotem per millorar l'Administració de Justícia.

9.2. Usos ja relativament consolidats del NLP en la Legal Tech

Per contextualitzar de què estem parlant exactament quan diem *processament del llenguatge natural* (*NLP)*, farem una breu referència a alguns usos que ja gaudeixen d'una certa consolidació, especialment a l'àmbit privat. Seguint Dale (2018), en podem destacar els següents:

1) Cerques d'informació jurídica general, tant legal com jurisprudencial. Alguns exemples: *LexisNexis*, *Westlaw*, *Wolters Kluwer*, etc. Aquestes eines solen utilitzar tècniques *tradicionals*: variables de cerques binàries (*booleanes*) i índexs creats manualment. Per aquest motiu la qualitat dels resultats obtinguts dependrà en gran mesura de si l'usuari ha fet les *preguntes* adequades. Algunes utilitats (*CaseText* i *CaseMine*) oferirien la possibilitat que l'usuari carregui un passatge rellevant de text o fins i tot un resum sencer que doni el context per a la cerca. Un disseny *UI* (*user interface*) anticiparia allò que l'usuari pot necessitar. Recorda en aquest sentit Dale (2018) que les *big four* (les grans companyies del sector) no han trigat a oferir les seves pròpies solucions reforçades amb IA (*Lexis Analytics*, *WestSearch Plus*, etc.).

2) *Electronic discovery* (*e-discovery*): identificació de la rellevància dels documents emmagatzemats electrònicament en funció de certs criteris de cerca vinculats a una investigació o a la preparació d'una demanda. Es tracta d'eines útils quan s'ha de fer front a milers de documents disponibles en discs durs i calgui diferenciar entre els que són rellevants (*responsive*) d'aquells que no ho siguin. D'una manera similar a les eines de cerca legal de l'apartat anterior, l'*electronic discovery* ha evolucionat des d'aproximacions *tradicionals* (selecció binària prèvia basada en paraules concretes i revisió manual posterior) fins a mètodes que empren ja tècniques d'aprenentatge automatitzat per a la classificació de documents[81]. Però, és clar, la definició de la

[81] Cal destacar en aquest apartat el mètode *TAR* (*technology-assisted review*), amb el qual es pretén trobar tants documents rellevants d'una col·lecció com sigui possible amb un esforç raonable. En una primera fase,

suficiència i la *rellevància* dels documents codificats és una qüestió jurídica complexa que pot ser, ella mateixa, objecte de controvèrsia. Aquí és on operen diferents protocols que pretenen generar confiança en l'ús d'aquestes eines[82].

3) Revisió de contractes: eines per comprovar que un contracte és complet i no genera riscos legals. O que no conté *anomalies*. Són útils en contractes força estandarditzats respecte dels quals és factible predir quin és (i quin ha de ser) el seu contingut. Poden detectar la manca d'una previsió concreta que necessàriament han de contenir.

4) Generació automàtica de documents legals rutinaris.

5) Assessorament legal a través de diàlegs pregunta-resposta.

6) Prova digital: podem definir-la com qualsevol informació o conjunt de dades produïda, emmagatzemada o transmesa per mitjans electrònics que pot ser utilitzada per acreditar fets en el procés (Delgado, 2020). La seva rellevància potencial és més que evident amb l'actual increment de la digitalització de tots els sectors socials, públics i privats. Acostuma a ser necessari traslladar al procés una gran varietat de fets *extra processals*. Si molts d'ells queden reflectits (pàgines web, etc.) o fins i tot *succeeixen* (xarxes socials, etc.) en un marc digital, llavors és evident que la prova adquirirà una naturalesa digital que condicionarà substancialment la manera com hagi de ser tractada processalment[83].

9.3. Desafiaments de l'analítica de textos legals

Com és fàcil de deduir dels usos ja força consolidats del *NLP* legal que s'han esbossat a l'apartat anterior, la clau o premissa, per plantejar-se l'ús d'aquest tipus d'eines, és disposar de documents (textos) jurídics *llegibles* (*interpretables, processables*) pel sistema, es tracti de lleis, sentències, demandes, informes o, en general, documents. Només llavors s'obriran les possibilitats dels processos automatitzats. Si ens plantegem possibles usos judicials, la tecnologia subjacent a les eines que s'acabin desenvolupant

es codifiquen manualment diversos documents com a rellevants (*responsive*) o no. Quan ja es disposa de suficients documents codificats, es passa a la fase de *descoberta*.

[82] El protocol *CAL* es basa en la interacció de dues eines, un sistema de cerca per paraules i un algoritme d'aprenentatge automatitzat. Hi ha altres protocols, amb petites variants, com el *SAL* o el *SPL*. No hi entrarem, per raons d'espai i claredat.

[83] Per exemple, en funció de si tenen un emmagatzemament o allotjament convencional (*smartphone*, *web*) o deslocalitzat (núvol).

serà, probablement, força similar a la ja existent en el mercat privat. Com a mínim, pel que fa a la fase inicial de *lectura* del text i *processament* subsegüent. Diferents hauran de ser, és clar, les garanties suplementàries que calgui introduir atès el context públic i judicial del qual estem parlant.

De fet, el que s'entén per analítica de textos legals té, com diem, una història força llarga (Conrad i Branting, 2018). El que passa és que conferències sobre IA i dret com les celebrades per la *ICAIL* (*International Conference on Artificial Intelligence and Law*) en el marc de la *IAAIL* (*International Association for Artificial Intelligence and Law*) s'han centrat més en l'*argumentació* i la *inferència*, i no tant en una aproximació empírica (basada en el *corpus* de dades disponibles) a l'anàlisi i resolució de problemes.

Els primers sistemes daten dels anys 50 del segle passat i els principals desafiaments que genera aquest camp els trobem en la *formalització* o *representació* de la informació, ja que el coneixement legal no està *ordenat* de manera clara i precisa (Yin i Mok, 2019). Està condicionat, de fet, per l'*experiència*. Per un *coneixement expert* sense el qual el *caos normatiu* que ens rodeja difícilment pot adquirir sentit: el raonament i l'argumentació jurídiques exigeixen una expertesa que va més enllà de la mera memorització d'un gran número de casos o fins i tot de la seva síntesi. Sovint les regles són ambigües o incompletes. Fins i tot, contradictòries. Per això el raonament legal combina en el seu interior diferents tipus de processos argumentatius: entre altres, els basats en *regles*, els basats en *casos anteriors*, els *analògics* o els *hipotètics*[84]. A més, aquest camp se sotmet a un canvi constant en tots els seus nivells (legal, jurisprudencial i doctrinal), per la qual cosa els sistemes que es puguin crear hauran de poder ser modificats i adaptats amb certa facilitat.

En definitiva, el raonament legal, el llenguatge natural i, també, el sentit comú (al qual s'hi ha d'acudir sovint) estan íntimament interrelacionats i són, per definició, difícils de *categoritzar* i *formalitzar* (de representar perquè un programari els pugui *processar*). Els desafiaments de la *NLP* legal són, per tant, diversos i complexos. És en aquest sentit que podem dir que és l'àmbit legal el que porta *al límit* la *NLP* general.

[84] Caldrà afrontar, també, altres dificultats: la representació del *context* imprescindible en qualsevol raonament jurídic; la formalització dels diferents graus d'analogia; la formalització dels raonaments probabilístics, derrotables o no monotònics; o, per últim, la formalització de la valoració de la prova.

9.4. Projectes en curs d'analítica legal de textos

Doncs bé, tot i els desafiaments exposats a l'apartat anterior, l'auge actual de les tecnologies algorítmiques ha obert i enriquit el camp d'exploració de l'analítica de textos legals. En aquest apartat explorarem algunes propostes en vies de desenvolupament que, des de diferents perspectives, aborden un tipus d'analítica legal de textos potencialment d'interès per a una eventual IA judicial. Són les següents:

a) Algunes propostes parteixen d'una perspectiva *en xarxa* dels sistemes basats en textos normatius procedents de diverses autoritats interrelacionades. Sadeghian et al. (2018) aborden la problemàtica tècnica crucial dels *gràfics de citació*: xarxes de previsions normatives i precedents jurisprudencials vinculats (enllaçats) per cites *explícites* (és a dir, aquelles que els textos fan efectivament) [85] . Proposen el desenvolupament d'un *corpus* de cites anotades i una avaluació experimental que mostraria que és possible predir amb precisió la *semàntica categòrica* de la cita (si se segueix o no el precedent citat) fent ús d'un model d'aprenentatge automatitzat entrenat amb els trams de text associats amb les cites.

b) Leibon et al. (2018) exploren un model en xarxa més ampli en el qual els enllaços inclouen no només les cites *explícites*, sinó també les relacions *implícites* derivades de models de *tòpics* amb els quals s'erigeix una espècie de *paisatge legal*. Aquest model habilitaria la cerca en les col·leccions legals per criteris de *rellevància* més que per altres de mera similitud *textual*.

c) També és d'especial interès, en aquest camp, el projecte MIREL[86], que ha creat una xarxa internacional i intersectorial de treballs d'experts per al desenvolupament d'eines d'extracció de sentit i d'argumentació a partir de textos legals. Es tractaria de convertir els textos legals en representacions formals que puguin ser usades per a la cerca de normes o com a suport de la presa de decisions. La complexitat del projecte és difícil d'imaginar. Per començar, requerirà crear un llenguatge formal estàndard que pugui assimilar, entre altres extrems, continguts referents a les propietats temporals de la

[85] Apunta aquest treball que als països del *Common Law* les cites entre els precedents acostumen a indicar expressament la *semàntica* del vincle (si el precedent anterior se segueix, es reverteix o es diferencia del nou cas [l'anomenada tècnica del *distinguishing*]). En el cas de les normes legals, no acostuma a haver-hi, per contra, la indicació d'aquesta *semàntica* del vincle. Es tracta d'una dada clau (la d'identificar la *semàntica* del vincle), perquè serà a partir d'ella que serà determinada (seleccionada) la resta de continguts rellevants.
[86] https://www.mirelproject.eu.

vigència de les normes o als seus efectes (drets que reconeixen, prohibicions que estableixen o les habilitacions que permeten). Es buscaria un accés *ontològic* al coneixement normatiu, que podria ser utilitzat tant per tasques de suport a la presa de decisions com per altres de *compliance*.

d) La *Oficina de Publicacions de la UE* (*OPEU*), que està seguint una decidida estratègia vinculada a les noves tecnologies, promou l'ús de tècniques *semàntiques* per a la representació del coneixement, l'anotació de documents i l'obertura de les dades. L'eina *CELLAR* de la OPEU és el major repositori de dades de la Comissió Europea. El *SeTA* (*Semantic Text Analysis Tool*), per la seva banda, combina específicament el *big data*, l'aprenentatge automatitzat i el processament del llenguatge natural (*NLP*) amb la finalitat d'explorar el coneixement i generar recomanacions de suport. Realitza cerques centralitzades a través de bases de dades tan diverses i voluminoses com la *EUR-Lex*, el *EU Open Data Portal* o la *Wikipedia*, entre altres (Vucheva et al., 2020).

e) Categorització i enginyeria d'ontologies legals: una *ontologia legal* busca transformar textos de llenguatge natural en dades que tinguin un format *tractable* pel sistema. I això es pretén fer-ho a través de la *categorització*, que seria la conceptualització estructurada del camp jurídic de què es tracti per mitjà d'*entitats*, *atributs*, *relacions* i *axiomes*. La *NLP* podria ajudar en la construcció d'aquesta ontologia legal. Ja existeixen, de fet, diverses llibreries de *NLP*[87] de conceptes ja identificats que podrien *activar-se* tant amb fets de la vida real com amb regles i raonaments extrets de casos anteriors (precedents) que siguin d'interès. A tal efecte, treballs recents (Yin i Mok, 2019) proposen diferenciar, d'entre totes les frases o afirmacions que es poden fer en una resolució judicial, fins a sis categories[88].

[87] *NLTK, TextBlob, Stanford CoreNLP, spaCy* o *Gensim*.

[88] Les apuntem a continuació, no per afirmar que és la única classificació correcta o possible en aquesta matèria, sinó amb l'exclusiva intenció de posar de manifest la diversitat (i complexitat) que pot tenir la tasca de crear una *ontologia legal* que pretengui operar amb *NLP*:
1) Frases o afirmacions *fàctiques*: les que contenen els fets rellevants tal com han estat fixats pel tribunal.
2) Frases o afirmacions *legals*: les que contenen una cita legal.
3) Frases o afirmacions relatives als *punts en disputa*: tesis de les parts sobre quina és la correcta aplicació de la norma o la fixació adequada dels fets, amb els corresponents arguments de suport.
4) Frases o afirmacions en les quals el tribunal delimita quins són els *punts rellevants* subjacents a la disputa i com s'interrelacionen.
5) Frases o afirmacions argumentatives de *fonament*: aquelles amb les quals el tribunal aplica directament la llei a la disputa i arriba a una conclusió (*holding sentences*).
6) Frases o afirmacions de *raonament*: aquelles amb les quals el tribunal explica com ha arribat a una conclusió o a una determinada interpretació de la norma o dels fets.

f) *Incrustació*: una tècnica clau en aquest tipus d'usos de l'analítica de textos legals per mitjà del *NLP* podria ser la *incrustació*. Amb ella seria possible superar l'abisme que hi ha entre els textos (*en brut*) i els *vectors* que volem generar. L'objectiu seria articular (*incrustar*) textos *discrets* (concrets, limitats, per la mateixa forma lingüística en què s'expressen) en espais de vectors *continus* en els quals ha d'operar el sistema (Zhong et al., 2018). Alguns projectes pretenen aplicar mètodes d'incrustació ja existents (com el *Word2Vec*) al vocabulari legal. Afronten la complexa tasca d'haver de capturar no només la informació gramatical sinó també el coneixement legal. Es tractaria d'una tasca de modelatge del coneixement o *Knowledge modelling* (Chalkidis i Kampas, 2019). Aquest modelatge permetria construir les *gràfiques* de coneixement legal basades en *xarxes* de conceptes legals. Amb la particularitat (i dificultat) que aquests poden tenir una representació i un significat divergents en diferents països o tradicions[89].

g) *Similar Case Matching*: per acabar farem una referència al mètode de la *correlació* o *emparellament* de *casos similars* (*Similar Case Matching*). Encara que pot ser de més interès per als sistemes del *Common Law* (EUA, Canadà, Índia, etc.), no deixen de tenir-lo per als sistemes continentals, en els quals també acostuma a ser necessària la cerca jurisprudencial. Aquí la semblança semàntica entre els casos es fixa a partir de diverses informacions de granularitat molt desigual. Els mètodes tradicionals d'extracció d'informació (*Information Retrieve* o *IR*) se centren en les semblances a nivell dels termes o paraules i operen amb conceptes legals i models estadístics. Són, de fet, els que s'utilitzen encara avui, majoritàriament, en els sistemes de cerca. Un salt tecnològic pot consistir en operar amb *meta-informació* per captar la semblança semàntica. O, ja en termes estrictes d'aprenentatge automatitzat, pot acudir-se a tècniques de *factorització* o, fins i tot, si l'objectiu és calcular el grau de semblança a un nivell semàntic, a xarxes neuronals[90].

Podríem dir, en resum i seguint Boella et al. (2019) i Chalkidis i Kampas (2019), que els productes de *Legaltech* ja existents que operen amb *NLP* semi-automatitzen tasques

[89] Per tant, els desafiaments per a la construcció d'una gràfica general (occidental?) de coneixement legal són enormes. Això no exclou que els mètodes d'*incrustació* siguin força prometedors, si bé és probable que segueixi sent necessari, de moment, combinar-los amb mètodes *simbòlics*. Seria el cas, en el camp de la predicció de resolucions judicials, del *TopJudge* (Guo et al., 2018).

[90] Tran et al. (2020) proposen, per exemple, un model basat en *CNN* (xarxes neuronals convolucionals) que opera a nivell global dels documents o de les frases que contenen i que es basa en una prèvia síntesi o resum de la resolució. Aquest resum o síntesi del document es codifica i *incrusta* en un espai de vectors continus que serà el que s'utilitzi per la resta de tasques.

com la classificació de documents legals, la identificació de referències creuades, la vinculació de termes legals i definicions o l'extracció d'informació (per exemple, dels punts clau de les previsions legals que siguin rellevants pel cas). Al mateix temps, però, els exemples de *NLP* que pretenen expandir aquestes utilitats amb l'ús d'eines d'aprenentatge *profund* (xarxes neuronals) són, en la seva majoria, només projectes en curs de desenvolupament. No són, encara, realitats palpables. Però sembla que apunten a potencialitats *raonables* de la IA que s'hauran de seguir amb atenció.

9.5. Condicionament processal del format d'entrada dels textos

Ja hem contextualitzat quin és, aproximadament, l'estat actual de les eines legals de *NLP*. Desplacem-nos, ara, al terreny judicial. Un *camp de joc* molt especial. Amb particularitats que poden, però, afavorir la maximització de la utilitat d'aquesta tecnologia: més enllà que, com hem vist, el dret és, essencialment, text, la mateixa existència (necessària) del procediment judicial permet condicionar prèviament, a través de la norma processal, el format amb què les parts presenten les demandes i els documents. I això pot ser clau per permetre una operativitat real i òptima del *NLP*: podria preveure's que el format i contingut dels documents que s'introdueixin per les parts al Sistema de Gestió Processal s'adaptin, ja d'entrada, als paràmetres que l'eina de *NLP* associarà a determinades funcions. Potser caldrà que aquests formats siguin preceptius. O, com a mínim, incentivar-ne l'ús d'alguna manera (reducció de les taxes judicials, acceleració de la tramitació de la causa, etc.). Aquesta *anticipació* respecte del format que han de tenir els documents és possible en el cas de l'Administració de Justícia perquè, precisament, ja podem saber d'entrada quins seran els documents (els textos) rellevants per a la tramitació del procediment. Els tenim perfectament identificats i sabem d'on han de venir: de fonts oficials, com les lleis i la jurisprudència (o, també, la transcripció de la prova practicada durant el procediment); o procedents de les parts, com les demandes, els escrits, els documents o els informes.

Es tracta d'una particularitat del procés judicial: en altres àmbits de la IA, les dades s'han d'anar a buscar, en *brut*, a la realitat. Precisament com que estan en *brut*, cal una important fase de selecció, preparació i *neteja* de les dades, que exigeix força feina manual i pot condicionar (i limitar) les utilitzats del sistema d'IA que es vulgui utilitzar. Per contra, en l'àmbit judicial l'*entrada* dels textos ve molt delimitada per la regulació processal. I això genera un clar avantatge: aquesta mateixa norma processal pot

perfectament condicionar prèviament el format i contingut digitals que hagin de tenir tots aquests textos[91]. Tot plegat amb l'objectiu, és clar, que després els textos puguin ser *llegits* més fàcilment pel sistema de *NLP* en la fase de *tractament* dels textos, d'extracció de la informació rellevant. I sembla evident, en definitiva, que una fase prèvia de *condicionament* del format d'*entrada* dels textos pot facilitar i maximitzar les potencialitats de la següent fase de *processament*, que es perllongarà, de fet, fins a la finalització del procediment judicial.

Posem algun exemple merament indicatiu: sovint en les demandes civils, tal com estan redactades, no queda clar quina és la pretensió principal i quines són, si és que n'hi ha, les subsidiàries o, fins i tot, alternatives. Per això cal aclarir aquestes situacions a l'audiència prèvia. Si es pretén que un sistema de *NLP* processi adequadament i des d'un inici un element del procediment tan rellevant com la demanda inicial, és crucial que comprengui bé, d'entrada, quines són les pretensions principals i subsidiàries. Si el format de la demanda és merament un *pdf* digitalitzat llegible (per exemple, amb tècniques d'*OCR*), l'eina de *NLP* haurà d'operar sobre tot el text en bloc (de fet, en *brut*) per extreure aquesta informació. La tasca serà molt complexa i poblada de riscos que puguin portar a errors. Caldran, a més, les més avançades tècniques d'aprenentatge *profund*, que poden venir associades, com sabem, a una major opacitat i menor transparència. Per contra, si s'exigeix processalment, d'entrada, que la demanda es presenti amb un format determinat (per exemple, un *pdf editable* amb formularis i caselles d'informació digitalment *tractables* i diferenciades per a la pretensió principal i les subsidiàries), és obvi que la tasca indicada se simplifica enormement, en tots els sentits: la tasca d'*extracció* d'informació (amb tècniques d'*OCR*) podrà dirigir-se directament a la casella i formulari específicament i prèviament predefinit. No caldrà, abans, abordar una cerca *en brut* per tot el *pdf*.

El mateix exemple es podria posar respecte de molts altres supòsits (com l'al·legació de pluspetició) que, per manca de claredat dels escrits de part, acostumen a crear confusió i a exigir una dedicació excessiva de *temps judicial*. Es tractaria, en definitiva, de forçar les parts, per via tecnològica, a una major precisió conceptual i concreció material (*desglossament* i *etiquetatge*) dels seus posicionaments. Diferenciant, en les caselles

[91] És probable, certament, que alguns documents (algunes dades o textos) també s'obtinguin en brut de la realitat (per exemple, els documents pretèrits en els quals una part basa la seva pretensió), però sempre es podrà condicionar processalment el format en què la part ha de presentar aquesta documentació.

indicades, per exemple, els següents punts: quina és la pretensió? Quina és la causa de demanar? En quins documents es basa? Quins són els documents principals i quins els secundaris o complementaris? Quina concreta clàusula del contracte se sosté que ha estat incomplerta, que ha de ser activada judicialment o que podria ser nul·la per abusiva? Serà tota aquesta informació, ja delimitada digitalment d'entrada per la mateixa part, la que es traslladaria al sistema de *NLP* perquè hi pugui operar directament, *localment*. No globalment, *a cegues*. La diferència pot ser substancial[92].

Dit això, el que s'acaba d'afirmar respecte dels escrits i documents de part es pot traslladar, amb els matisos necessaris, a les resolucions judicials que es dictin durant el procediment: aquestes resolucions hauran de ser igualment *tractades*, com a textos que són, pel sistema de *NLP* i serà convenient, també, que presentin un format digital igualment desglossat i etiquetat[93].

[92] De l'exposat veiem, de fet, que la introducció d'eines de *NLP* al procediment judicial podria fer convenient i implicar un canvi radical en les maneres de treballar no només de les oficines judicials sinó també dels operadors jurídics. Però es tracta, ben mirat, d'una oportunitat de millora, perquè el que s'acaba de proposar no és sinó un avenç objectiu en termes de claredat i síntesi dels escrits. Atributs, avui, no sempre presents en els escrits que arriben als jutjats.

[93] Aquí també es podria obrir, indirectament, el camí cap a un nou estil de llenguatge judicial. Més breu, clar i concís. Més entenedor per a la ciutadania. Fins i tot podria plantejar-se l'ús d'etiquetes (*tags*) sobre termes clau més tècnics perquè el lector pugui acudir directament i amb facilitat a la seva definició o explicació.

10. La IA judicial en dret comparat

10.1. Hi ha, realment, jutges-robot a la Xina?

Comencem, ara, un repàs de quin és l'estat de desenvolupament actual, en altres països, de la IA judicial. I ho farem, de nou, per la teulada: ens preguntarem si, com sembla desprendre's d'algunes notícies que es difonen i d'alguns articles que s'escriuen, existeixen en altres països *jutges-robot* i, en cas positiu, quin àmbit d'aplicació tenen. Hem abordat al capítol 7 la diversitat de problemàtiques que genera aquesta eventualitat. El que abordarem a continuació és si, tot i aquests entrebancs, algun país ha avançat en aquesta línia. Iniciarem l'anàlisi pel país que sembla haver fet més avenços en aquest àmbit, la Xina. Ja s'ha apuntat a la part introductòria que la Xina està seguint un model *estatal-autoritari* en la implementació de certes noves tecnologies que menysté i relega a un inassumible segon pla la tutela de la privacitat i dels drets fonamentals. Aquesta sola circumstància ens l'ha de fer descartar com a model a imitar. Al mateix temps, però, en una recerca com aquesta li convé, si més no, fer l'intent d'aproximar-se i conèixer (si és possible) les eines tecnològiques que està utilitzant a l'àmbit judicial. No podem descartar, d'entrada, que se'n puguin extreure algunes consideracions pràctiques d'interès.

Però la transparència no caracteritza, certament, la Xina. Tampoc en aquesta matèria. Acudirem a un document que, en principi i amb totes les cauteles, ofereix una certa aparença d'oficialitat. Es tracta del paper blanc *Tribunals xinesos i la Internet judicial*, de 5 de desembre de 2019, presentat pel Tribunal Suprem del Poble Xinès el desembre de 2019. A la data del document hi hauria tres Tribunals *en línia* (a Hangzhou des de 2017, a Beijing des de 2018 i a Guangzhou també des de 2018), dedicats a resoldre casos relacionats amb Internet. I ho fan per mitjà de procediments tramitats principalment *en línia*.

S'estaria construint un sistema de tribunals en línia utilitzant el Big Data, la computació al núvol, la IA i el *blockchain*, en una cadena que pretén preservar l'extracció, la conservació i la traçabilitat de les proves electròniques. De moment, l'àmbit d'aplicació se centraria en el comerç electrònic, la responsabilitat per productes adquirits a Internet,

els préstecs concedits en línia, els pagaments virtuals, les transaccions al núvol, disputes per dominis i els conflictes sobre propietat intel·lectual[94].

El model xinès incorpora un sistema d'ODR (*Online Dispute Resolution*) completament integrat, que inclou la mediació, la presentació d'escrits, el pagament de les taxes, les vistes i l'emissió de la resolució. El sistema intern judicial estaria connectat amb els serveis externs de litigació. S'utilitzarien recursos de reconeixement de veu en les vistes, de verificació automàtica de documents, de generació de documents electrònics i de gestió intel·ligent dels casos.

En concret, el tribunal de *Guangzhou* està realitzant una prova pilot amb accions col·lectives: un cas és seleccionat per ser objecte d'una vista i les parts dels altres casos són *convidades* a seguir en línia la vista, fet que podria potenciar posteriors acords en aquests *altres* procediments.

Pel que fa a la recollida, preservació i autentificació de la prova electrònica, s'està explorant l'aplicació de tècniques de *blockchain* en combinació amb el *Big Data* i l'emmagatzematge al núvol, per garantir la traçabilitat, l'auditació posterior i la prevenció de la manipulació de la prova. D'aquesta manera es pretén construir una connexió segura amb altres plataformes o registres, com les *ODR*, les notaries o els centres forenses. S'ha implementat també, des de 2014, una plataforma digital interdepertamental per a la investigació i embargament patrimonials, que utilitza mètodes de *Big Data* per a la valoració del béns.

La digitalització és, segons el paper blanc, el prerequisit de qualsevol evolució ulterior del procés judicial en termes d'IA. L'eliminació del paper, la presentació telemàtica, l'escaneig digital i la ràpida i clara indexació i categorització dels documents, amb simultània transferència al procediment en qüestió, serien les prioritats. Es pretén igualment l'automatització de la identificació de la complexitat dels casos i la seva categorització, el reconeixement dels arxius de text digitalitzats o la transformació de veu a text. Fins i tot, la detecció automatitzada de conductes inadequades durant les vistes.

[94] Aquest sistema judicial en línia estaria vinculat al sistema de traçabilitat ciutadana i de crèdits socials de premi i castic al qual ja s'ha fet referència. Per exemple, es preveu unes llistes negres judicials de morosos, interconnectades amb tot el sistema públic, i que impliquen, entre altres coses, restriccions per accedir a càrrecs públics o l'augment dels costos de transport. Aquí és on es posa de manifest la naturalesa *autoritària* d'un model, el xinès, que ha de ser, en aquest extrem, absolutament descartat.

També, oferir a les parts serveis de recomanació precisa de lleis aplicables o casos similars. O, per últim, en el que ara més ens interessa, la generació automatitzada i correcció de documents judicials, amb avisos de *desviacions* en els criteris de presa de decisions.

Per la seva banda, el Tribunal Suprem Popular hauria iniciat la plataforma *Faxin* (*Global China Law*) amb la qual, més enllà d'oferir als tribunals informació i recursos legals i jurisprudencials, s'apliquen eines d'anàlisi automatitzada de *Big Data* com l'etiquetatge dels *atributs* de cada cas. El sistema seria capaç d'obtenir en *temps real* dades dels judicis i procediments d'execució seguits a tot el país, amb actualitzacions cada cinc minuts[95].

En conclusió, veiem que fins i tot a la Xina no sembla que s'estigui generalitzant l'ús de *jutges-robot* que dicten, completament, sentències definitives. La informació no és molt precisa, però sembla que aquesta possibilitat es limita a casos molt específics i que ja tenen un origen (contractual o extracontractual) *en línia*. Per contra, el gruix d'eines d'IA s'orienten a tasques *parcials* en la tramitació i gestió del procediment, tant el prejudicial com el judicial. Aquesta sembla, certament, l'alternativa més *raonable*, atès el nivell actual de desenvolupament de la IA.

10.2. Estònia, el model judicial europeu més avançat en IA

Deixem de banda, ara, el model *autoritari-asiàtic* de la Xina, que no ens ha de servir, certament, de mirall preferent, i desplacem-nos a un país més proper i petit, Estònia. Aquesta nació s'ha erigit els últims anys en un referent europeu (fins i tot mundial) en la digitalització dels serveis públics. Inclosa, també, la justícia. Com podem imaginar, però, atesa la seva especificitat, és en la justícia on els passos han estat (estant sent) més lents i primmirats.

[95] En la jurisdicció penal, es pretén la visibilitat, traçabilitat i la monitorització dels procediments policials i de la fiscalia, per mitjà del reconeixement de la imatge, el *NLP*, la identificació de les proves i l'extracció automatitzada de la informació clau de cada cas. Respecte dels casos d'accidents de trànsit, s'ofereixen serveis de valoració dels danys, de determinació de la responsabilitat o de mediació. En els casos d'insolvència, se segueix, per contra, una aproximació més orientada al mercat i que gestioni adequadament l'existència de molts creditors. La plataforma digital en línia ofereix a tots ells una informació actualitzada i habilita les reunions en línia i la valoració dels actius patrimonials, amb intervenció de postors internacionals.

Aquest país va començar ja l'any 2005, amb el llançament de l'*e-File system*, un profund canvi en el sistema judicial per mitjà de la seva digitalització. Hi poden accedir no només els representants processals de les parts sinó també els mateixos ciutadans, amb el DNI i una contrasenya. S'hi pot presentar qualsevol tipus de cas i les dades seran compartides per multitud de tribunals i institucions. Se segueix la política de *només una vegada*, de tal manera que la dada serà introduïda un cop i compartida per totes les institucions: no es permeten les duplicitats en les bases de dades estatals. La plataforma permet als tribunals remetre a la ciutadania diferents tipus de documents, amb certificació garantida de la recepció. Cada document rep un segell temporal i conté una signatura electrònica de seguretat. La informació classificada es pot encriptar pel tribunal per assegurar-se que cap tercer hi accedirà. Aquest sistema ofereix, en definitiva, un panorama general de les diferents fases de tots els procediments penals, civils o administratius, incloses les resolucions judicials.

L'any 2006 es va implantar el *KIS* (el sistema d'informació dels tribunals), destinat al registre de tots els tipus de casos de tots els tribunals. S'hi registren també les vistes i resolucions i, com especialitat, disposa d'una eina de repartiment automàtic de casos als tribunals, crea les notificacions, gestiona la publicació oficial de les resolucions i extrau els *metadata*. El sistema també emet recordatoris i monitoritza el temps dedicat a cada fase del procediment.

Fins aquí podríem dir que els avenços a Estònia en matèria de digitalització judicial i automatització de certes tasques de tràmit són molt destacables i prometedors. Però si ens centrem en la figura que hem anomenat *jutge-robot*, hi va haver, fa uns anys, anuncis de la seva implantació en un determinat tipus de casos. El ministre de justícia estonià va encarregar al seu Director de Gestió de Dades, Velsberg Ott's, que treballés en el disseny d'un *jutge-robot* que dictés sentències en les disputes de menys de 7.000 euros, amb la idea de reduir l'acumulació i retard d'aquest tipus de casos. No sembla, però, com a mínim de moment, que aquest projecte s'estigui materialitzant. Ni que ho hagi de fer en breu[96].

[96] El mateix Secretari General del Ministeri de Justícia ha admès (després de diferenciar, en el camp de la IA judicial, entre eines de suport i eines de presa de decisió) que ara per ara són força escèptics respecte de les segones. Principalment, per problemes de manca de transparència.

En un nivell segurament més realista, actualment s'està treballant en eines de *transcripció* automatitzada de tot tipus de procediments judicials per generar resums de les vistes orals[97]. Aquesta aplicació utilitzaria tècniques d'aprenentatge automatitzat, *NLP* i reconeixement de veu.

10.3. EEUU

A les premisses de la recerca hem diferenciat els tres models existents de desenvolupament del *Big Data*. Un d'ells era l'americà, basat en el mercat. Doncs bé, aquesta orientació *privada* es reflecteix, lògicament, en el tipus d'eines (de *productes*) que els americans estan ja disposats a utilitzar en els seus procediments judicials. En farem un petit esbós. Però cal advertir de l'elevada diversitat de situacions existents a cada Estat Federat. Situacions, per altra banda, molt canviants. Per tant, fer-ne una fotografia alhora actualitzada i completa és pràcticament impossible.

Per exemple, un tribunal d'Ohio, competent en tractaments de joves, va utilitzar el sistema *WATSON* d'*IBM* com a eina de suport per sintetitzar i avaluar la situació personal dels joves subjectes al procediment (Coglianese i Ben Dor, 2020). Per altra banda, en matèria de justícia *predictiva*, un estudi realitzat amb eines d'aprenentatge automatitzat (Hutson, 2017) hauria predit correctament el resultat del 70% de les sentències del TS (d'un total de 28.000, des del 1816 fins al 2015). I, també, el 72% dels vots individuals dels membres del tribunal. Més conegut és, però, l'ús d'eines d'IA en matèria de mesures cautelars personals penals. Ens hi detindrem, atesa la seva rellevància quantitativa i qualitativa.

Es tracta d'eines per donar suport a la decisió d'adoptar, o no, una mesura de presó preventiva abans que no es celebri el judici. També, per la concreció de la pena a imposar. O, per últim, del pla de compliment de la pena. S'anomenen, també, eines de predicció del risc (*risk assessment tools*), ja que és el risc de reincidència delictiva o de fugida (de no comparèixer) el que se sol tenir en compte a l'hora de prendre aquestes decisions. El seu ús està força estès als EEUU. De moment, de manera dispersa i fins i tot caòtica, ja que pot dependre no ja de la normativa de cada Estat Federat, sinó, fins i tot, de la mera dotació de recursos a nivell de cada partit judicial (o *municipal court*). Per

[97] A l'apartat 10.4 analitzarem amb més detall alguns projectes d'innovació tecnològica judicial recollits al *Study on the use of innovative technologies in the justice field – Final Report*, de setembre de 2020.

la seva banda, el Govern Federal central ha anunciat treballs en una eina d'aquesta naturalesa, la *Prisoner Assessment Tool Targeting Estimated Risk and Needs* (*PATTERN*).

Dit això, eines de predicció del risc com les analitzades no necessàriament han d'utilitzar, pel seu funcionament, l'*aprenentatge automatitzat* (Berk, 2017). De fet, es basen més aviat en senzills sistemes d'indexació o en mers models convencionals de regressió (predicció) logística. Acostumen a tenir en compte un número determinat de factors rellevants: edat, tipus de delicte comès, existència d'antecedents penals no cancel·lats, condemnes prèvies, diferenciant la seva gravetat, incompliments de mesures cautelars personals anteriors, consum de substàncies estupefaents, entorn familiar i laboral, etc. Els ponderen en una determinada proporció i fixen un resultat final que seria la probabilitat futura de comparèixer davant del tribunal o de cometre un nou delicte.

Un exemple n'és l'eina predictiva LSI-R (*Level of Service Inventory-Revised*), que també té en compte l'entorn educatiu o l'estat mental de l'investigat. Un altre de més conegut és el cas de la *Correctional Offender Management Profiling for Alternative Sanctions* (*COMPAS*), que ha estat força utilitzada als tribunals americans. Els detalls del codi d'aquest algoritme, tractant-se d'un producte privat, estan protegits pel secret empresarial. Però investigadors de *ProPublica* (Angwin et al., 2016) van revelar el llistat de punts valorats pel sistema[98]. De fet, per fer front a aquesta manca de *transparència*, Estats com Idaho han aprovat normatives que exigeixen que aquest tipus d'eines siguin transparents i permetin la inspecció pública. Amb la possibilitat, per part dels investigats, de tenir accés als càlculs i les dades que han determinat el resultat obtingut[99].

[98] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[99] Es tracta del Codi de procediment criminal d'IDAHO, secció 19-1910, en una reforma de 2019. (https://legislature.idaho.gov/statutesrules/idstat/title19/t19ch19/sect19-1910/). Per l'interès que té (atesa l'escassa normativa existent en la matèria), en reproduirem una part (les cursives són de l'autor de la recerca): "19-1910. pretrial risk assessment tools. (1) All pretrial risk assessment tools shall be transparent, and:

> (a) All documents, data, records, and information used by the builder to build or validate the pretrial risk assessment tool and ongoing documents, data, records, and written policies outlining the usage and validation of the pretrial risk assessment tool *shall be open to public inspection*, *auditing, and testing*;
>
> (b) A party to a criminal case wherein a court has considered, or an expert witness has relied upon, a pretrial risk assessment tool *shall be entitled to review all calculations and data used* to calculate the defendant's own risk score; and
>
> (c) No builder or user of a pretrial risk assessment tool may *assert trade secret or other intellectual property protections* in order to quash discovery of the materials described in paragraph (a) of this subsection in a criminal or civil case".

És rellevant, en aquest sentit, conèixer no només quins són els factors o paràmetres que es tenen en compte, sinó també com opera, en concret, el sistema. Com els pondera per arribar a un resultat de probabilitat determinat. És aquí és on poden aparèixer certs problemes de discriminació: el mateix estudi ja citat de *ProPublica* (Angwin et al., 2016) va posar de manifest que el sistema *COMPAS* sistemàticament adjudicava un major risc de reincidència als investigats de color negre en comparació a altres investigats de color blanc que es trobaven en una situació similar.

Altres estudis han posat de manifest, també, que l'ús d'aquestes eines no necessàriament ha d'implicar la reducció del nombre de presos i de la taxa de reincidència i que, fins i tot, poden ser un factor d'augment de les disparitats racials pel que fa a les taxes de condemnes (Stevenson i Doleac, 2019). Els dubtes i dilemes que generen aquest tipus d'eines predictives són, per tant, nombrosos.

Ens és d'interès, precisament per aquests motius, analitzar amb cert detall un cas en el qual l'investigat va impugnar als tribunals l'ús judicial de *COMPAS*. Es tracta del cas *State v. Loomis*. A la sentència el tribunal va tenir en compte, en part, que havia estat identificat per *COMPAS* com un individu amb un risc alt. A la primera petició impugnatòria (*motion*), *Loomis* va al·legar com a drets processals infringits, entre altres, el de ser jutjat sobre la base d'informació precisa i que la sentència fos individualitzada. El mateix tribunal que va utilitzar l'eina *COMPAS* va desestimar la petició indicant que simplement l'havia utilitzada per corroborar la valoració de la prova que havia fet per una altra via i que hagués imposat la mateixa condemna amb independència de si hagués tingut en compte, o no, el resultat ofert per l'eina predictiva. Un recurs posterior interposat davant el Tribunal Suprem de *Wisconsin* va ser igualment desestimat tenint en compte que les variables utilitzades eren públiques i que la informació es basava o en les mateixes respostes donades pel condemnat o en informació pública sobre els seus antecedents penals. No s'hauria infringit el *dret a un procés degut* perquè hi va haver la possibilitat de comprovar que les preguntes i respostes eren precises[100]. Posteriorment, *Loomis* va presentar un recurs davant del Tribunal Suprem dels EEUU, però sorprenentment va ser inadmès a tràmit.

[100] Per altra banda, sosté aquesta sentència que si bé és cert que aquestes eines impliquen l'ús de dades de grup, com que s'inclouen en una diversitat d'altres paràmetres, s'acaba obtenint una resposta individualitzada. Tampoc no hi hauria problemes de gènere atès que no consta que el tribunal el tingués en compte i que el sistema només valora la diversitat de taxes de reincidència entre homes i dones.

En un altre cas, *Malenchik v. State*, el condemnat va impugnar l'ús de dues eines predictives en la determinació de la pena. Però el Tribunal Suprem d'*Indiana* va concloure que el tribunal inferior hauria tingut en compte altres factors (en concret, els antecedents penals i la negativa a assumir la responsabilitat pels seus actes o a canviar el seu comportament). Per tant, l'eina predictiva no s'hauria utilitzat com un *factor agreujant independent*. Només seria una font d'informació rellevant per determinar si se suspèn, o no, una sentència o per dissenyar un programa de reinserció. Seria viable, en definitiva, que un tribunal complementi i enriqueixi la seva valoració de la prova amb aquest tipus d'informació predictiva.

Acabarem amb el cas *People v. Wakefield*, relatiu no a eines de predicció de risc com les anteriors, sinó a l'ús d'un programari de correlació d'ADN per fonamentar la condemna. Aquesta eina utilitzava, fins a un cert punt, IA. El recurrent va al·legar que la manca d'accés al codi font de l'eina violava el seu dret constitucional a confrontar els testimonis. El tribunal d'apel·lació va concloure, però, que el codi no és un *declarant* i que no hi podia haver, per tant, infracció del dret al·legat. No negava el tribunal que una eina d'IA pugui arribar a ser un testimoni d'un judici, ja que algunes tenen una elevada autonomia, en funció de la distribució cognitiva entre humans i tecnologia que hi hagi darrera. Però l'eina en qüestió operaria amb una intervenció i supervisió humanes tan elevades que es podria afirmar que l'interlocutor real darrera l'informe era, en realitat, l'autor de l'algoritme.

10.4. Altres experiències properes a la IA judicial

10.4.1. Tasques judicials susceptibles de ser innovades

En aquest apartat s'apuntaran de manera molt sintètica algunes experiències d'altres països properes a la IA judicial, amb la finalitat d'obtenir una certa perspectiva de quin és l'estat d'avenç de les tecnologies de la digitalització més avançades aplicades al procediment judicial. Ens és d'especial interès, en aquest sentit, el treball *Study on the use of innovative technologies in the justice field – Final Report*, de setembre de 2020[101]. Aquest estudi diferencia fins a vuit tasques judicials a les quals podria aplicar-s'hi innovacions tecnològiques. Ens serà útil començar per aquesta llista, per copsar la

[101] Estudi encarregat per la Comissió Europea i elaborat per Vucheva et al. (2020).

diversitat d'àmbits i *parts* del procediment a les quals podem plantejar-nos implementar certs avenços:

1) El *processament* de grans quantitats de dades, estructurades i no estructurades, amb la finalitat d'obtenir informació rellevant pel cas, identificar patrons, cercar paraules o casos i classificar o categoritzar. Aquestes tasques (com algunes de les següents) ens remeten a la idea de *justícia predictiva*, a la qual hem dedicat el capítol 8.

2) El processament d'arxius de vídeo, àudio i imatge, per, a través del reconeixement facial o d'objectes, identificar persones o víctimes del delicte en el material disponible (per exemple, a Dinamarca, en les investigacions d'abusos a menors), monitoritzar el comportament, detectar conductes il·legals de presoners o transcriure àudio a text (per exemple, a Croàcia). Aquí seran claus les tècniques de reconeixement de veu i traducció automàtica.

3) Enllaç d'informacions procedents de diferents fons (bases de dades, registres, sistemes, etc.) quan no estiguin centralitzades o connectades per mitjà d'una interfície o punt d'accés. En aquesta tasca poden ser necessaris els enllaços *semàntics* (no només els *literals*).

4) Accés a la Justícia o altres serveis públics vinculats: bases de dades, informació de casos concrets, legislació, navegació pels processos administratius, etc. En aquest punt els *xatbots* podrien millorar l'accés a la Justícia.

5) Compliment de la normativa de protecció de dades, especialment pel que fa a la confecció de les resolucions judicials per fer que sigui compatible aquesta protecció amb la seva publicitat general. La IA podria utilitzar-se per automatitzar l'anonimització (eliminació o modificació) de les dades personals o altres dades sensibles.

6) Processament de dades per preparar un determinat acte processal, com pot ser una vista.

7) Gestió administrativa del tribunal: planificació d'agendes, de vistes, priorització dels casos o distribució de les sales i altres infraestructures.

8) Traçabilitat de l'autenticitat i integritat dels documents i les dades derivades de les actuacions realitzades per tots els actors del procediment judicial. El *blockchain* seria la tècnica adequada per fer front als problemes d'autenticitat, signatura electrònica, traçabilitat i integritat.

Recull aquest informe que el nivell d'ús de les tecnologies d'IA en el camp de la justícia encara era, a data de 2018, relativament baix: només un 7% dels enquestats indicaven que s'utilitzava alguna forma de tecnologia d'IA a la seva organització. A continuació concreta, respecte de diversos països, quins han estat fins el moment els avenços o projectes en curs en IA judicial. O, sent més precisos, en innovació, digitalització i automatització judicial. En els següents apartats destacarem, de manera molt sintetitzada, els casos recollits en aquest informe que ens siguin de més interès (de l'apartat 10.4.2. al 10.4.6) i altres projectes d'IA judicial que s'han pogut identificar en el curs de la recerca (de l'apartat 10.4.7 al 10.4.9).

10.4.2. Àustria

a) Anonimització automatitzada (ja no manual) de les resolucions judicials, amb la idea d'obrir-les al públic, de manera completa i gratuïta. El sistema diferencia les persones físiques i les jurídiques i identifica les adreces i els càrrecs. Busca mantenir la comprensibilitat del text (del cas), tot i les modificacions introduïdes[102].

b) Automatització de la localització i processament dels documents procedents de canals electrònics o de documents escanejats.

c) Identificació de la numeració del cas.

d) Categorització i titulació de documents.

e) Reconeixement de tipus de procediments en les demandes entrants.

f) Suggeriment al tribunal de càrregues i ritmes de treball.

[102] Inicialment es volia utilitzar una eina d'origen privat (*IBM Watson tool*), però el cost de les llicències ha fet que s'optés per una infraestructura de codi obert. Utilitza la *NLP*, biblioteques obertes (com *Stanford NLP333* o *spaCy334*) i l'aprenentatge automatitzat.

g) Suport al tribunal en certes tasques (determinació de les costes processals, etc.).

h) Reconeixement facial d'interns a presó per detectar conductes no habituals.

i) Xatbot: portal de serveis digitals vinculat via mòbil als procediments judicials, per tal que la ciutadania pugui consultar els seus procediments en cada moment i fase.

j) Investigacions criminals: sovint les dades a analitzar en un sol cas són molt voluminoses (de diversos terabits). La fiscalia faria ús d'eines d'aprenentatge *profund*, sistemes experts (basats en regles), *NLP* i visió computeritzada. L'entrenament de l'algoritme seria individualitzat per cada ocasió, en funció del tipus de cas i de les dades disponibles o de la informació que se cerqui. El realitzaria la mateixa fiscalia, ja que és qui coneix millor el cas[103]. En aquest projecte se segueix una aproximació mixta, amb aprenentatge *supervisat* i *no supervisat*[104].

k) IA per analitzar de manera rutinària i automatitzada, abans que intervingui qualsevol funcionari del jutjat, els correus electrònics entrants (inclosos els documents adjunts) a la bústia del jutjat. Es buscaria identificar el procediment al qual es refereixen els documents encara que no incloguin la numeració específica, extreure les metadades dels documents i categoritzar-los a través d'una proposta de descripció i de titulació del document.

l) Visió orientada a objectes (*object-oriented-views*): per exemple, visionar tots els documents aportats per una part.

m) Detecció de correlacions *semàntiques*.

n) Configuració i adaptació dels sistemes de cerques o d'extracció d'informació a *criteris experts* rellevants en cada cas, sense necessitat de modificar o ajustar el codi.

[103] Per exemple, pot interessar-li cercar només les factures disponibles o les que afectin a certes persones, etc. Fins i tot, per exemple, determinar si una cadena de correus electrònics contenen una conversa formal o informal.
[104] Hem analitzat les diferències entre aquests dos tipus d'aprenentatge als apartats 6.2.1 i 6.2.2.

10.4.3. França: DataJust

Es tracta d'una eina de predicció o suggeriment de la indemnització màxima reclamable per danys personals, en funció de criteris com l'edat de la víctima, la ubicació del dany o la seva gravetat (segons criteris mèdics). Es destinaria principalment a les víctimes dels danys, però també, a més de a les asseguradores, a l'advocacia o fins i tot a la judicatura com a element orientador. La finalitat seria promoure acords prejudicials. El model analitzaria les dades disponibles i les compararia amb casos similars respecte dels quals ja hi hagi hagut pronunciaments judicials (que serien prèviament anonimitzats).

10.4.4. Alemanya

a) Tecnologia *blockchain* per al Registre de la Propietat.

b) Eines de traducció legal de suport a la fiscalia.

c) Identificació automatitzada d'imatges de pornografia infantil.

d) Identificació de delictes d'odi a les xarxes socials.

e) Sala de vistes moderna amb sistemes de gravació i transcripció de veu a text, de manera que tots els participants rebin en un arxiu la transcripció amb l'àudio incrustat. La sala hauria de permetre, també, la projecció en 3D de l'escena del crim.

10.4.5. Dinamarca

a) Identificació de víctimes del delicte en casos d'abusos a menors. Detecció automatitzada de si el mateix vídeo està present en un altre disc o servidor, encara que hagi pogut ser escurçat o editat. Priorització automatitzada dels vídeos amb casos més greus, per poder començar respecte d'ells l'anàlisi manual.

b) Anonimització de resolucions judicials: aborda el problema de si una resolució criminal ha de ser, o no, publicada.

10.4.6. Itàlia (Tribunal de Gènova)

Es tracta d'un pla pilot aplicat de moment a delictes de violència de gènere i que consisteix en una eina predictiva basada en resolucions del passat respecte de les quals es busquen semblances amb els casos nous[105].

10.4.7. El sistema Prometea argentí i el sistema PretorIA de Colòmbia

D'Argentina ens interessa el sistema PROMETEA, creat amb fons públics pel Laboratori d'Innovació i Intel·ligència Artificial de la Facultat de Dret de la UBA de Buenos Aires, en col·laboració amb el Ministeri Públic Fiscal de la mateixa ciutat.

Es tracta, en termes generals, d'una plataforma *híbrida* (que combina intel·ligència humana i automatització) destinada als funcionaris i que funciona en forma de xatbot. Opera d'una manera relativament semblant a la plataforma SPACE utilitzada a alguns tribunals de la Índia i que veurem a l'apartat 10.4.9. Pot ser activada en un ordinador convencional o en un telèfon mòbil. Les utilitats més rellevants serien, a partir de la introducció del número de l'expedient de què es tracti, el processament automatitzat de tots els textos que l'integren (*NLP*), la interacció amb l'usuari (en forma de formulació de preguntes) i el control dels terminis i dels requisits formals aplicables. Es buscaria, en general, una reducció substancial dels *clics* de ratolí i de les obertures de finestres.

La utilitat de cerca de jurisprudència consisteix en una eina *predictiva* basada en la localització de patrons. En la detecció, en grans volums de precedents, de si existeix, ja, en funció els fets del cas, una tendència jurisprudencial consolidada. Aquesta predicció es complementa amb la generació d'esborranys de resolució (o *projectes de dictamen*) que posteriorment seran completats per qui tingui la funció assignada. També genera enllaços a tots els documents localitzats que siguin rellevants. No integra, però, de moment, la signatura electrònica. El destí potencial de l'eina seria tant l'Administració Pública en general com la de Justícia en particular. Això no obstant, més enllà d'afirmacions genèriques segons les quals es tracta d'un sistema d'aprenentatge

[105] Utilitza codi obert de *TensorFlow*, *Keras*, and *Scikit*. Pretén tenir un nivell d'explicabilitat raonable. Els models que integra són força avançats: *arbres de decisió* (apartat 2 de l'annex 1), *Support Vector Machines* (apartat 4 de l'annex 1) i *Xarxes Neuronals Profundes* (apartat 6.2.5).

supervisat, no ha estat possible localitzar quina és, exactament, la tecnologia, model o tipus d'algoritme que hi ha darrera d'aquesta aplicació.

Es té constància d'una altra experiència d'IA judicial basada en el sistema *Prometea*. Es tracta de *PretorIA*, aplicat pel Tribunal Constitucional de Colòmbia. Seria el primer ús mundial concret de la IA judicial predictiva en un tribunal superior. La funció principal d'aquest sistema és fer una primera gestió dels milers de peticions diàries que rep el tribunal. Ho fa a partir de 33 criteris de *categorització fàctica* delimitats pel mateix tribunal. Els seus responsables afirmen que es tracta d'un sistema d'aprenentatge *supervisat* plenament transparent: seria una *caixa blanca* completament *explicable*, *interpretable* i *traçable*.

10.4.8. Brasil

El Brasil és un país que està fent força avenços en matèria d'IA judicial. Salmão, L. (2020) analitza fins a 72 projectes implementats o en via d'implementar-se. Com veurem, els usos que es pretén donar a la IA judicial brasilera són, de moment, força raonables i propers, en alguns casos, a les propostes que es faran més endavant en aquesta recerca. A continuació se n'enumeren els més rellevants:

1) Comprovar les hipòtesis d'inadmissió preliminar de les demandes, segons els supòsits previstos a la normativa processal.

2) Generació d'esborranys de suggeriment de resolució.

3) Categorització dels casos per matèria. O, fins i tot, agrupació per semblança, als efectes de tenir-ho en compte més endavant (pensem en la tramitació de les demandes col·lectives o massificades).

4) Automatització de tasques repetitives de la oficina judicial, amb eines que prediuen el següent moviment processal i que generen automatitzadament el text pertinent.

5) Gestió de manaments i notificacions: classificació de la seva urgència, naturalesa i complexitat. Geolocalització de les adreces per permetre un seguiment detallat de la seva execució.

6) Gestió automatitzada de l'admissibilitat dels recursos, amb retorn, arribat el cas, també automatitzat, de la causa al tribunal d'origen.

7) Extracció de dades de les vistes.

8) Reconeixement facial: el sistema *Ámon* identifica, a partir de fotografies i per raons de seguretat, els visitants que entren al jutjat.

9) Xatbot per facilitar les tasques de la oficina judicial.

10) Normalització de documents.

11) Transcripció auditiva.

12) Distribució automatitzada de casos.

13) Anàlisi jurisprudencial, per evitar prendre decisions diferents per a casos similars anteriors. La vinculació al precedent se *suggereix* a partir de la *convergència* entre el contingut de la petició inicial d'un procés i una *matriu de comprensió* d'un tema precedent.

L'informe analitza amb detall la plataforma *Athos*, que es va formar amb la *lectura* o processament d'aproximadament 329.000 resolucions judicials d'entre 2015 i 2017 i va indexar més de 2 milions de casos amb 8 milions de peces, cosa que li permetria, ara, agrupar automàticament els casos similars. En concret, la utilitat *Precedent Management Center* (*NUGEP*) identifica processos que tenen la mateixa controvèrsia legal, selecciona la matèria rellevant i detecta l'existència d'eventuals interpretacions convergents o divergents entre diferents sales o seccions d'un mateix tribunal. O, fins i tot, suggereix el possible desviament o superació de precedents qualificats.

Per la seva banda, el sistema *Sòcrates 1.0*, que utilitza el mateix motor d'IA que el sistema *Athos,* fa un seguiment dels procediments, els agrupa i identifica els precedents rellevants. La versió *Sócrates 2.0* se centra en la identificació de controvèrsies idèntiques i l'assignació d'apel·lacions repetitives.

Per acabar, el sistema *Synapses* és una plataforma per al desenvolupament i la disponibilitat, a gran escala, de models d'intel·ligència artificial per part d'altres tribunals

que podran operar-hi de manera independent i autònoma, mitjançant microserveis. Se centra en la classificació de documents i l'extracció de text. És també una interfície per importar conjunts de dades. Segons recull l'informe, utilitza, entre altres tècniques, l'aprenentatge *per reforç* (apartat 6.2.4). També té utilitats per generar textos legals (amb la funció d'autocompletar) i per resumir textos de manera personalitzada segons els paràmetres introduïts per l'usuari. En destacarem tres mòduls: l'*iPrecedente*, que automatitza el procés d'anàlisi dels precedents; l'*iJurisprudència*, que automatitza el procés mateix d'elaboració de la jurisprudència; i l'*iAssistente*, que redacta esborranys per votar.

10.4.9. Índia (SUPACE)

Un altre cas llunyà, però interessant, és la plataforma d'ús judicial *SUPACE* de la Índia. Té una naturalesa *híbrida*, en el sentit que es basa en una simbiosi entre tasca manual i automatització en el marc d'un conjunt heterogeni d'utilitats: gestiona els casos, processa arxius *pdf*, fa tasques de compilació i de cerca de documents i integra un xatbot al qual el mateix tribunal pot plantejar-li qüestions. Per exemple, preguntar-li "*de què va aquest expedient?*", cas en el qual, en uns segons, es generarà un resum complet que integri la demanda inicial i les vicissituds processals per les quals hagi transcorregut. O, simplement, preguntar pel posicionament de la fiscalia o quins drets fonamentals es veuen afectats. També és molt interessant la possibilitat que el sistema generi de manera automatitzada una cronologia completa dels fets rellevants. De fet, el mateix sistema suggereix altres preguntes que se li poden plantejar. I les augmenta a mesura que se'n fa ús: és a dir, aprèn per si mateix a partir no només de les preguntes que plantegen els usuaris, sinó, especialment, de les anotacions o informació que extreuen de les respostes que va donant el sistema. Es tracta, per tant, d'un sistema d'aprenentatge automatitzat en sentit estricte.

En una plataforma com *SUPACE*, deixaria de ser necessari, per tant, saltar d'un document a un altre per obtenir aquest tipus d'informació: la compilació la fa el mateix sistema en funció de la pregunta que se li faci. El mateix sistema informa l'usurari sobre quins documents han estat analitzats per respondre la pregunta (queden ressaltats).

La plataforma integra un *notebook* i un processador de textos amb capacitat d'extreure la informació dels documents associats al cas i de generar esborranys de resolució

(inclosa la referència a la jurisprudència rellevant aplicable), sense necessitat que el tribunal teclegi res. També inclou la funció de dictat oral. Aquestes tasques es poden realitzar en un arxiu *word* genèric o utilitzant, més específicament, arxius amb un format ja predeterminat. Lògicament, a partir de l'esbós el tribunal podrà seguir creant *manualment* la resolució. És per aquest motiu que podem qualificar aquesta plataforma d'*híbrida*, ja que integra en un equilibri força raonable les capacitats humanes les i artificials.

11. Possibles aplicacions judicials de la IA

11.1. Usos parcials i de suport més enllà de la figura del jutge-robot

Abordarem, ara, propostes concretes de possibles aplicacions judicials de la IA. Aquí deixarem de banda, com ja hem fet en altres apartats, la idea de generalitzar l'automatització del dictat de sentències. Ens centrarem en possibles usos de la IA de suport i ajuda, durant tot el procediment, a la oficina judicial i al mateix tribunal. La majoria no s'ubicaran, de fet, en la fase de la resolució del cas o del dictat de la sentència final. I els que ho facin només seran eines de suport parcial i sempre amb control i agència humana preponderants. Ja hem vist al capítol 7 que l'escepticisme amb què cal abordar l'eventual implementació de *jutges-robot* es justifica no només per l'evident afectació a principis elementals de justícia en el context d'un estat de dret (independència i imparcialitat judicials, prohibició de l'actuació arbitrària dels poders públics, dret de defensa, dret a un judici just, etc.), sinó per la seva aparent inviabilitat *tècnica*: l'estat actual de la innovació tecnològica en el camp legal no sembla que permeti, com a mínim segons el resultat d'aquesta recerca, l'automatització de la tasca de dictar sentències de fons en casos que presentin un mínim grau de complexitat. Ni, tampoc, en casos molt senzills que, això no obstant, tinguin la potencialitat de desviar-se de la *tipologia de cas* a la qual pertanyen en principi[106]. Sembla evident, també, que el mer fet de preveure's un control humà posterior no és una garantia (en aquest cas humana) suficient: ja coneixem la tendència humana a donar per bo allò que ens proposa un sistema automatitzat aparentment neutre i objectiu (l'anomenat *biaix d'automatització*).

Per tant, descartada en bona mesura aquesta idea del *jutge-robot*, i centrats en tasques més modestes i menys *sensibles* en termes de drets i garanties (però igualment exigents en termes de dedicació, temps i esforç), podem plantejar-nos més obertament, i amb menys prejudicis, si algunes de les eines que qualifiquem, una mica grandiloqüentment, com IA poden ser aplicades a la feina diària del tribunals. I el cert és que, ben mirat, la

[106] La distinció que s'acaba de fer és rellevant, ja que un cas molt senzill i aparentment pertanyent a una tipologia de casos molt estandarditzada (en principi ideal, per tant, per aplicar-li una tècnica d'automatització), pot tenir la *potencialitat* de presentar una *particularitat* que justifiqui que la sentència es desviï, en alguna mesura (major o menor, no cal que sigui passar de l'estimació a la desestimació o de la condemna a l'absolució), de la solució prevista en principi per a aquest tipus de casos. I, si del que es tracta és d'implementar eines d'IA, caldria que el model fos capaç de detectar, per ell mateix (de manera automatitzada), aquestes particularitats diferenciadores. I no es té constància de l'existència d'una eina d'IA que ofereixi, amb suficients garanties, aquesta funcionalitat.

resposta sembla evident que ha de ser positiva, sense perjudici de tota la cautela i cura amb les quals sempre s'ha d'abordar el binomi tecnologia i justícia.

Al mateix temps, però, aquestes propostes es fan en el marc d'una recerca i en un moment en el qual les tecnologies que serien necessàries no estan del tot desenvolupades o, com a mínim, adaptades al sector judicial. Per tant, les diferents idees que a continuació es desplegaran han de ser enteses en termes completament *provisionals* i *hipotètics*. El desig inicial de la recerca era poder ser molt més concrets o concloents en aquest apartat. Més tècnics, en definitiva. Però l'estat actual de desenvolupament tecnològic de la IA judicial (la captació del qual era també objecte de la recerca) es troba en una fase força verda. Immadura. És innegable que hi ha avenços (els hem vist al capítol 10 quan hem analitzat la IA judicial comparada) i, sobretot, una clara voluntat d'ampliar-ne els usos. Així ho testimonien els diversos projectes i estudis en curs, no acabats encara a la data de tancament d'aquesta recerca.

Cal fer esment, en aquesta línia, al projecte espanyol *JUSTICIA 2030*[107], que pren com a referència la futura elaboració de la Llei d'Eficiència Digital del Servei Públic de Justícia com a marc per a la seguretat jurídica digital i l'ús de la IA a l'Administració de Justícia. Aquesta estratègia de llarg termini aborda, en uns termes, de moment, molt genèrics, qüestions d'analítica legislativa i judicial, amb la incorporació de la intel·ligència artificial a una justícia *orientada a la dada*. O el foment de la traducció automàtica dels documents o de les resolucions judicials a les llengües oficials. Fins i tot, l'eventual integració del *blockchain*, les videoconferències o els *ODR*. No oblida, tampoc, les tècniques d'investigació del delicte (prova científica, tractament automatitzat de les dades i cerques intel·ligents). Un aspecte especialment pràctic serà la futura notificació electrònica, que permetrà remetre les notificacions a les parts per correu certificat a través del *Centro de Impresión y Ensobrado* (CIE) de l'Agència Tributària, que estaria disponible de manera electrònica a la *Carpeta Ciutadana* (aquí s'integrarien la *Notific@*, la gestió d'acusament de rebudes electròniques i el sistema de gestió processal).

Es tracta, però, de previsions de futur que desconeixem quan s'implementaran (si és que s'implementen) i quin contingut o condicionants específics tindran. Comencem, per tant, ara, l'apartat de propostes concretes d'automatització judicial. Un apartat que, per tots

[107] https://www.justicia2030.es.

els motius exposats, cal entomar com una espècia de pluja d'idees *inspirades* en el resultat de la recerca i de les quals eventualment en pot cristal·litzar alguna per motius o factors que ara mateix no podem predir ni controlar.

11.2. Pressupòsits d'una eventual implementació efectiva

Abans de començar, però, convé recordar una sèrie d'exigències tècniques i pràctiques (força òbvies, per altra banda) que necessàriament distancien en el temps l'eventual implementació de qualsevol de les propostes que es faran:

a) *Desenvolupament de la tecnologia necessària*: caldrà, primer, desenvolupar o adaptar les tecnologies que siguin necessàries per a dur a terme de manera automatitzada les tasques que es diran. En cadascuna de les propostes s'intentarà, en la mesura del possible, destacar el tipus de tecnologia que sembla adaptar-se més a cada tasca i s'identificaran els desafiaments tècnics que sigui ja evident que es poden generar.

b) *Interoperabilitat*: les innovacions tecnològiques que seran necessàries no fan referència, només, a les eines d'IA judicial en sentit estricte, enteses individualment, sinó també a la imprescindible *interoperabilitat* dels diferents sistemes que es veuran implicats, en major o menor mesura, en cada cas. Aquest és, de fet, un dels grans desafiaments tècnics que caldrà afrontar. I, a més, des d'un inici, ja que la interoperabilitat és un requisit que convé tenir present des de la mateixa fase de disseny de cadascun dels components que, precisament, han d'interoperar.

c) *Fase de prova*: caldrà seguir en cada cas una fase prèvia de prova en condicions reals controlades (les *sand boxes* que hem vist a l'apartat 6.6.6) per confirmar el bon funcionament del sistema i que no genera més riscos en termes d'afectació de drets i garanties processals que els que s'han previst inicialment.

d) *Certificació i auditories*: Caldrà preveure, també, mecanismes de *certificació* del funcionament adequat de les eines, preferiblement per mitjà d'auditories externes. És d'interès a aquests efectes l'estudi de desembre de 2020 del CEPEJ sobre la possibilitat

o la *Viabilitat de la Introducció d'un mecanisme de certificació d'eines i serveis d'IA en l'esfera de la Justícia i dels tribunals*[108].

e) *Previsió legal*: abans de la implantació de qualsevol d'aquestes idees seria necessària, imprescindible, una expressa previsió legal que l'autoritzi i la reguli. Potser podríem excloure d'aquesta necessària regulació expressa els supòsits d'automatització d'actes processals de mer tràmit sense afectació a drets o garanties i en els quals no hi hagi tractament de dades personals. Actes, per altra banda, difícils d'imaginar. En qualsevol cas, ja hem vist al capítol 3 que la regulació actualment existent és força escassa i es redueix a la normativa de protecció de dades o a la jurisdicció penal. També hem vist, però, a l'apartat 3.4, que l'actual normativa europea sobre IA en curs d'elaboració inclou l'Administració de Justícia en el grup d'àrees d'alt risc i les implicacions que això té. Sí que podem afirmar que la norma que s'acabi aprovant haurà de preveure especials drets d'informació sobre l'ús d'eines d'IA judicial i mecanismes de revisió humana i d'impugnació jurídica efectius contra les decisions que s'hagin automatitzat. Només així serà viable generar un grau mínim de confiança per part dels usuaris i destinataris de l'Administració de Justícia.

f) *Caràcter modulable de l'automatització*: segons els casos i les jurisdiccions afectades, cada proposta podrà tenir un nivell diferent, modulat, d'automatització, que pot consistir des d'una mera alerta o avís del sistema (per modificar, o no, en algun sentit, la tasca humana que s'està realitzant) fins a l'eventual generació d'esbossos de resolucions judicials (complets o parcials i sempre pendents d'una ulterior finalització o revisió humana). De fet, segons els casos, serà possible que a una mateixa tasca judicial se li puguin aplicar diferents nivells d'automatització i que la seva elecció depengui de cada operador jurídic, segons les seves necessitats o conveniències. Sempre, però, informant de manera detallada als demés operadors o parts implicades sobre el grau d'automatització que s'ha aplicat en cada cas concret.

[108] European Commission For The Efficiency Of Justice (CEPEJ), *Possible introduction of a mechanism for certifying artificial intelligence tools and services in the sphere of justice and the judiciary: Feasibility Study*, de 8 de desembre de 2020. https://www.coe.int/en/web/portal/-/cepej-artificial-intelligence-and-cyberjustice-at-the-heart-of-discussions.

11.3. Ordenació temàtica i cronològica de les propostes

Per tal d'aconseguir una major claredat, agruparem les propostes en dos blocs: un de general, concebut preferentment amb referència a la jurisdicció civil però que, per la seva mateixa naturalesa, podrà ser extrapolat en bona mesura a la resta de jurisdiccions (i que inclou de l'apartat 11.4 al 11.19); i un altre relatiu a possibles aplicacions de la IA judicial a la jurisdicció penal (apartat 11.20).

El bloc general s'estructura en funció del moment i context processals en els quals operaria cada proposta (sistemes predictius preprocessals, fase de deganat, admissió de la demanda, tramitació del procediment, esborranys de sentència i execució). De la següent manera[109]:

a) *ODR* i sistemes predictius preprocessals

b) Abans de l'admissió de la demanda (deganat)

   Registre de les demandes

   Qualitat (definició) insuficient de la demanda o dels documents aportats

   Repartiment automatitzat de les causes

c) Admissió de la demanda: qüestions processals

   Capacitat per ser part i capacitat processal

   Detecció de problemàtiques de representació processal

   Detecció de la falta de competència objectiva, funcional o territorial

   Adequació del procediment

   Detecció de possibles defectes en la manera de presentar la demanda

[109] Clicant a sobre dels següents punts se salta directament a l'apartat en qüestió.

Indeguda acumulació d'accions

Cosa jutjada i litispendència

Acumulació de procediments

Litisconsorci passiu necessari

Automatització de l'admissió de la demanda i de l'emplaçament

d) Tramitació del procediment

Consum i clàusules abusives

Monitoris

Taxació de costes

Processos de divisió patrimonial

Prova documental

Cessions de crèdits

Pericials

Reconeixement facial en compareixences o vistes telemàtiques

Transcripció automatitzada d'àudio a text

Traducció automatitzada

e) Generació d'esborranys de sentències

Sentències d'aplanament

Extracció completa de la cronologia dels fets

Avisos d'omissions de fets, jurisprudència o pretensions

Esborranys de sentència complets

f) Execució

Esborranys d'interlocutòries despatxant l'execució

Localització, avaluació i realització dels béns

g) Possibles usos de la IA judicial en la jurisdicció penal

Eines d'IA penal judicial i eines d'IA d'investigació policial

Predicció del risc de reincidència o d'incompareixença

11.4. ODR i sistemes predictius preprocessals

Les anomenades *Online Dispute Resolution* (*ODR*), o Resolució en línia de litigis, no han de contenir, o integrar, necessàriament, eines d'IA. Per tant, no s'hi entrarà en detall en aquesta recerca: només en la mesura que hi hagi possibles àrees de confluència amb la IA[110]. Tampoc són, en si mateixes, entorns d'administració de justícia, però, com veurem, hi poden estar integrats com a fase prèvia[111]. Si aquesta integració s'executa amb encert, podrien suposar, de fet, una millora substancial en l'accés a la justícia: un accés

[110] La idea dels *ODR* no és nova i es remunta als anys 90, quan va iniciar-se el comerç electrònic, on són freqüents transaccions de baix import però que involucren molta distància física entre els actors (les parts). Aquests dos fets, combinats, dificulten per si mateixos l'alternativa d'acudir als tribunals ordinaris. Va ser en el context de plataformes com *eBay* o *Amazon* on va sorgir la idea, ja en els anys 2000, d'oferir als usuaris, en el mateix entorn de la plataforma, eines de negociació, mediació o arbitratge. El model d'*eBay* ha arribat a habilitar la resolució de 60 milions de casos anualment (Dal, 2018), però seguia tractant-se d'una eina privada aplicada a una esfera molt limitada i acotada de disputes. A més, el finançament que exigia i que no podien aportar les parts s'acabava obtenint de tercers, circumstància que en determinats casos podia generar conflictes d'interessos. Una eventualitat, aquesta, directament incompatible amb la idea d'administrar justícia de manera independent i imparcial.

[111] Si inclouen recursos no només de negociació sinó també de mediació o arbitratge, llavors caldrà preservar en el seu disseny, funcionament i composició la màxima objectivitat i imparcialitat possibles, amb uns paràmetres, si no idèntics, sí tendencialment equiparables als propis de l'Administració de Justícia en sentit estricte.

telemàtic, flexible, adaptable i de baix cost pot ser més plausible per l'usuari comú que el tradicional (i costós) accés físic als edificis judicials.

La principal aportació dels ODR és, clarament, la asincronia amb la qual aborden la resolució de conflictes. La idea seria superar la concepció tradicional segons la qual és necessari que qualsevol actuació judicial que pugui ser qualificada de vista o judici hagi de dur-se a terme amb la presència física de les parts, alhora i amb unitat de temps i espai, a l'interior de la sala de justícia[112]. Per altra banda, l'ús dels *ODR* implica la *transferència* a l'usuari de costos que d'una altra manera serien judicials, especialment pel que fa al temps dedicat a la introducció de les dades en el sistema[113]. Fins al moment, algunes aplicacions d'*ODR* han estat exitoses en termes d'accés a la justícia. Per exemple, *PARLe*, una plataforma canadenca d'assistència en la *ODR*, comprèn tres fases: la negociació, la mediació i l'adjudicació judicial[114].

Si ens centrem en els *ODR* que integren eines d'IA, la primera idea que ens ve al cap són els sistemes *predictius* sobre els possibles resultats de la controvèrsia. No cal dir que és una hipòtesi encara no existent. Merament exploratòria. Un dels seus efectes positius podria ser l'augment de l'autonomia i de la informació dels participants. Per diverses vies: anàlisi automatitzada dels contractes; eines avançades de cerca; eines predictives del possible resultat de la controvèrsia o, més concretament, dels marges raonables d'allò que pugui, o no, ser reconegut; agents conversacionals (xatbots); o, entre altres, creació automatitzada de documents. Aquesta tipologia d'eines no es limitaria, així doncs, a potenciar el mer accés a la justícia. Aniria més enllà. Pretendria empoderar jurídicament i de manera directa la ciutadania per poder gestionar els seus problemes jurídics, especialment els d'escassa complexitat o quantia. Els anomenats de

[112] És cert que en el procediment judicial *tradicional* la part afectada no sempre ha d'acudir físicament a totes les actuacions, però sí que ho ha de fer el seu lletrat o lletrada, a qui, lògicament, haurà d'abonar els seus honoraris. Per contra, en les *ODR* és el mateix usuari qui fa, personalment, cada actuació, però remotament, en línia, sense haver de ser-hi present. Això amplia molt la capacitat del sistema d'adaptar-se a les agendes de les parts sense impactar negativament en el procediment.

[113] En sentit invers, un cost afegit seria la necessitat d'ampliar la plantilla judicial amb personal amb perfils nous adaptats a la justícia en línia quan la formació del personal existent no sigui suficient.

[114] El procediment és senzill: qui vol reclamar registra una petició en una plataforma *ODR* certificada per l'estat. Hi exposa els fets i el remei sol·licitat. La plataforma trasllada la petició a la part contrària i li proposa iniciar una sessió i acceptar la proposta o fer una contra-proposta, fins que pugui assolir-se un acord. En cas contrari, se les deriva a la intervenció, també en línia, d'un mediador homologat que ajudarà les parts, mitjançant formularis, fòrums en línia o videoconferències (o, fins i tot, trucades telefòniques), a arribar a l'acord encara no tancat. Si el mediador constata el fracàs de la segona fase, llavors s'acudeix a l'adjudicació judicial, també en línia, amb la particularitat que qui ha de decidir (prèvia pràctica de testificals, o no) obté coneixement directe dels documents i fets ja registrats.

baixa intensitat, que, a més, sovint (no sempre) acostumen a respondre a pautes similars, circumstància que els fa especialment propicis per aplicar-hi tècniques d'IA[115].

En la hipòtesi més ambiciosa, els instruments de justícia *predictiva* susceptibles de ser integrats en *ODR* podrien generar pronòstics sobre quina és la solució més probable al conflicte, amb la finalitat que les parts que hi participen puguin disposar d'una idea aproximada (d'informació, en definitiva) sobre la plausibilitat d'allò que pretenen reclamar. Per exemple, en el cas de danys personals, com en el projecte *DataJust* analitzat a l'apartat 10.4.3. Altres camps potencialment fèrtils serien les compensacions laborals, les divisions o repartiments en processos de família o la mediació familiar.

En una hipòtesi menys problemàtica, les eines predictives podrien limitar-se a donar informació sobre els marges (econòmics) de la resposta que podria obtenir-se en cas d'acabar judicialitzant el problema. O sobre els costos econòmics i de temps que aquesta última alternativa podria generar. Disposar d'aquesta informació podria ajudar, en certs casos, a maximitzar les probabilitats d'arribar a un acord previ a la judicialització.

Al mateix temps, també són evidents les limitacions d'aquests sistemes, que ja han estat abordades al capítol 8 dedicat a la justícia predictiva. No es pot obviar, tampoc, el risc que condicionin de manera distorsionada el comportament dels participants en l'*ODR* si la informació no és acurada i de qualitat. Especialment, la que prediu el possible resultat.

11.5. Abans de l'admissió de la demanda (deganat)

11.5.1. Registre de les demandes

En aquesta fase inicial ja *judicial*, en la qual moltes tasques són especialment rutinàries i estandarditzades, cal plantejar clarament la possible automatització del registre de les demandes. Aquí hi entraran les tècniques d'*OCR* i de *NLP*. Ja n'hem parlat al capítol 9. I, més concretament, a l'apartat 9.5 s'ha apuntat la conveniència de condicionar processalment el format i contingut dels escrits (especialment de les demandes) que les parts presentin al jutjat precisament perquè sigui més àgil (o directament viable) el seu

[115] No es tractaria, en absolut, de relegar a un segon pla l'assistència lletrada professional: no hem d'oblidar, de fet, que dins dels sistemes d'*ODR* s'hi pot integrar la mediació i que aquesta pot ser desenvolupada, amb la homologació corresponent, per membres de l'advocacia. També que caldrà deixar oberta en tot moment la possibilitat d'acudir a l'assessorament legal corresponent i, és clar, a l'adjudicació judicial en cas de manca d'acord.

processament posterior no només com a mer text *brut* sinó com a objectes *simbòlics*, amb un significat jurídic ja predeterminat. Sembla evident que en el cas del registre de les demandes aquest *precondicionament* pot ser clau.

11.5.2. Qualitat (definició) insuficient de la demanda o dels documents aportats

En aquesta fase inicial es podria introduir la detecció automatitzada de la insuficient qualitat i definició de la demanda o dels documents aportats a efectes de ser *llegits* per sistemes d'*OCR* i per poder ser processats per mitjà del *PLN*. És una eventualitat que es dona sovint però que no es posa de manifest fins a una fase més avançada del procediment, amb els problemes d'indefensió i de dilació processal que la seva esmena pot generar. Sembla raonable, per tant, anticipar aquesta detecció per tal que, en cas de donar-se, el mateix sistema requereixi de manera automatitzada la part que ha presentat la demanda o documents deficients perquè ho esmeni en un termini determinat.

11.5.3. Repartiment automatitzat de les causes

Ja s'utilitzen a dia d'avui, en els partits judicials de cert volum, sistemes de repartiment objectiu i automàtic (no manual) de les causes en funció de la seva tipologia. Però són eines força rudimentàries. Del que es tractaria, des del punt de vista de l'automatització basada en la IA, seria que el mateix sistema detectés, per exemple, el tipus de procediment a seguir i el grau probable de complexitat del cas. Seria una dada rellevant per poder precisar a mig i llarg termini un repartiment *equitatiu* entre jutjats de les càrregues de treball *reals*. Aquest repartiment *just* ja es busca, avui dia, però sobre la base d'una categorització dels casos que no sempre és homogènia: dues demandes pertanyents a la mateixa tipologia de cas poden requerir per a la seva tramitació i resolució molta més dedicació i temps (per exemple, per haver-hi parts implicades a l'estranger, per haver-se aportat diverses proves pericials o una d'especialment complexa, etc.).

Si baixem a un nivell més tècnic, es podria pensar en les eines privades que ja existeixen per a l'advocacia per classificar els tipus de contractes quan se n'ha de gestionar un gran volum (n'hem parlat a l'apartat 9.2)[116]. Per altra banda, a l'hora de precisar el grau de

[116] Una certa extrapolació d'aquesta eina a Deganat, lògicament amb les oportunes adaptacions, podria ser una línia interessant d'investigació. Tot i això, pels usos estrictament judicials que estem analitzant, sempre

complexitat i temps requerit pels casos que es registren i s'han de repartir, podria pensar-se en els models d'IA de classificació o *clustering* (apartat 8 de l'annex 1).

11.6. Admissió de la demanda: qüestions processals

Ens desplacem ara, ja, a la seu del jutjat al qual ha estat repartida una determinada demanda des de Deganat. La primera tasca que haurà de realitzar és l'anàlisi de la concurrència dels pressupòsits processals per a l'admissió de la demanda. I el cert és que es tracta d'una fase (molt allunyada, com veiem, de la de dictar sentència sobre el fons) que presenta, com a mínim d'entrada, unes elevades expectatives d'automatització. I, per tant, de millora en eficiència i optimització dels recursos.

Nieva (2018, p. 34 i 35) apunta una sèrie de qüestions processals que podrien ser objecte d'eines d'IA. S'ha partit d'aquesta proposta per desenvolupar i concretar les potencialitats i problemàtiques que pot generar.

11.6.1. Capacitat per ser part i capacitat processal

Pensem en la possibilitat que el mateix sistema, interoperable amb la resta de registres públics o bases de dades administratives, detectés de manera automatitzada (i emetés el corresponent avís) que una de les parts (per exemple, la demandada) pateix algun tipus d'incapacitació i té reconeguda una tutela. En aquest cas, des d'un inici el jutjat podria adoptar totes les mesures necessàries, ja previstes processalment, per tal que aquesta part disposi de la tutela judicial oportuna. Per contra, quan no es disposa d'aquesta informació des del començament, és freqüent que el mateix emplaçament es compliqui i que costi arribar a tenir constància de la situació incapacitant o limitant.

serà preferible una eina creada expressament, amb recursos públics i codi obert, per la tasca de què es tracti. Això no impedeix, però, buscar la *inspiració* tecnològica en altres llocs, encara que no siguin públics.

Aquests factors es traduiran, indefectiblement, en importants dilacions processals i en riscos pel dret de defensa de la part afectada[117].

11.6.2. Detecció de problemàtiques de representació processal

L'anàlisi dels poders aportats al jutjat exigeix molta dedicació i temps. Poden generar-se problemes molt diversos: minoria d'edat, actuació en nom d'una societat determinada, etc. De nou, la lectura automatitzada dels poders i la connexió, també automatitzada, amb els registres públics corresponents (amb generació, en el sistema de gestió processal, dels avisos corresponents advertint de la possible existència de defectes processals), seria un avenç factible força evident. Aquest apartat presenta una major complexitat tècnica que l'anterior. Ja no es tracta, només, de detectar l'existència d'una resolució (administrativa o judicial) que acordi una determinada limitació o capacitat disminuïda (per aquest ús podrien ser suficients tècniques d'*OCR* i *NLP* juntament amb la necessària *interoperabilitat* amb els registres públics corresponents), sinó que caldria introduir aquests recursos en el marc d'un *sistema expert* de regles que canalitzin les diferents situacions que es poden donar. Senzill i rudimentari (ho hem vist a l'apartat 5.5), però necessari.

11.6.3. Detecció de la falta de competència objectiva, funcional o territorial

Aquests supòsits de manca de competència estan regulats amb detall per la llei processal. Els factors que la poden determinar con clars, precisos i taxats. Ens trobem, per tant, davant d'una decisió que podem qualificar, seguint Taruffo (1998), no només de *discrecionalitat dèbil* (els paràmetres rellevants venen clarament predeterminats), sinó altament *burocratitzable* (fet que ens apropa, és clar, a les eines d'IA). Certament, hi pot haver (amb certa freqüència) casos de certa complexitat en què es generen dubtes sobre la deguda competència objectiva, funcional o territorial del jutjat i que, per tant, hauran de ser abordats per una decisió humana. Però l'anàlisi prèvia, la detecció preliminar de possibles problemes de competència, per tal que s'emeti el corresponent *avís*, sí que sembla que pot ser objecte d'automatització. No hem d'oblidar que es tracta d'una tasca senzilla però que cal realitzar respecte de cadascuna de les milers de demandes que es

[117] És cert que la necessària interoperabilitat dels sistemes podria genera problemes en matèria de protecció de dades que caldria abordar. En qualsevol cas, però, els drets que es poden veure afectats justificarien preveure les excepcions legals que es considerin necessàries, sense perjudici d'adoptar tots els mecanismes tecnològics que permetin minimitzar l'exposició de les dades personals (encriptació de la informació, limitació de la informació visible a un avís que identifiqui al tutor, per remetre-li simultàniament la demanda, etc.).

reparteixen cada any a cada jutjat. Aquest suport podria ser, per tant, rellevant, en termes globals.

Caldria, en aquest cas, partir de la prèvia detecció del tipus de procediment que s'hauria produït ja a la fase de deganat (apartat 11.5), per tal d'aplicar, una vegada la demanda ja ha estat repartida, tècniques d'*OCR* i *NLP* en el marc d'un *sistema expert* que reculli les regles legals bàsiques aplicables precisament a cada tipus de procediment (per exemple, que, a efectes de competència territorial, en un monitori cal tenir en compte, només, el domicili de la part demandada, etc.). Podria enriquir-se l'eina amb una *interoperabilitat* i connexió també automatitzada amb les bases de dades disponibles (*Punt Neutre Judicial*, etc.) per constatar, per exemple (sempre que sigui de manera clara i concloent), que el domicili oficial més recent o principal de la part demandada no és l'aportat per la part actora, sinó un altre que pot determinar la manca de competència territorial. En aquesta hipòtesi, podria anticipar-se un pronunciament d'inadmissió i evitar-se la dilació d'intentar, primer, l'emplaçament o requeriment en el domicili aportat, que probablement serà negatiu.

11.6.4. Adequació del procediment

El sistema seria semblant a l'anterior: tècniques d'*OCR* i *NLP* en el marc d'un *sistema expert* que reculli els criteris legals d'adequació dels procediments (per quantia, per matèria, etc.). De nou, no es tracta d'emetre decisions definitives sobre aquestes qüestions processals, sinó merament de generar *avisos* de possibles inadequacions quan es detecti que pot donar-se el cas. Avisos que podrien anar acompanyats, lògicament, de generació d'esbossos de resolució, com veurem.

11.6.5. Detecció de possibles defectes en la manera de presentar la demanda

Es tracta d'una excepció processal més difícil d'automatitzar. De fet, hi pot haver molts tipus diferents de defectes en el mode de presentar una demanda. No estan acotats legalment, fet que ens allunya de la *discrecionalitat dèbil* i de la *burocratitzabilitat* de què parlàvem abans. Sovint seran al·legats per la part demandada, però no es pot descartar que estigui en rebel·lia processal o, fins i tot, que no se n'adoni. I cal tenir present que pot interessar al tribunal aclarir certes obscuritats i possibles incongruències en les pretensions articulades a la pètita de la demanda. Sempre serà viable, és clar, posar-ho de manifest a l'audiència prèvia del judici ordinari o al mateix acte de la vista del judici

verbal (o, si no se celebra vista, en un trasllat previ al dictat de la sentència). Però és preferible que aquest tipus d'inconsistència estructural de la mateixa petició nuclear de la demanda es posi de manifest tan aviat com sigui possible. I és per aquest motiu que es podria plantejar la possible automatització de la detecció d'alguns supòsits senzills, com ara quan hi hagi peticions subsidiàries incongruents o il·lògiques (la subsidiària és d'un import més elevat que la principal, l'import de la pètita no coincideix amb l'import definit com a degut al cos de la demanda, etc.); o quan s'apreciï manca de claredat sobre quins punts (peticions) integra cada pretensió autònoma, entre altres supòsits. Es tractaria, en tots els casos, de generació d'avisos de possibles defectes en la manera d'articular les pretensions perquè acte seguit un funcionari del jutjat constati si aquest és, en efecte, el cas, per donar-li el tràmit corresponent.

Dit això, quan analitzàvem les eines de *NLP* i l'analítica de textos legals, s'ha apuntat ja (apartat 9.5) que seria convenient condicionar processalment (de manera preceptiva, vinculant) el format i contingut que hagin de tenir les demandes inicials. Especialment pel que fa a la pètita. En aquesta hipòtesi seria més fàcil *focalitzar* el control automatitzat de detecció d'inconsistències, amb trasllat (fins i tot sense agència humana) perquè fos esmenada en un termini determinat. Pel cas d'insistir la part demandant en mantenir la pètita en els termes inicials, podria deixar-se la qüestió per més endavant, perquè pugui ser analitzada pel tribunal. Però fins i tot en aquest cas, es visualitzaria l'avís (automatitzat) segons el qual hi pot haver aquesta eventual incongruència.

11.6.6. Indeguda acumulació d'accions

Es tracta d'una qüestió processal de certa complexitat, però que té les regles rellevants clarament predeterminades per la norma processal. Ens movem, per tant, en l'àmbit de la *discrecionalitat dèbil* propera a la *burocratitzabilitat*. Per exemple, la norma preveu que, amb caràcter general, un demandant pot acumular en la demanda totes les accions que tingui contra el mateix demandat (acumulació *objectiva* d'accions), encara que provinguin de diferents títols, amb l'única exigència que no siguin incompatibles entre si. La constatació que totes les accions es dirigeixen contra el mateix demandat és fàcilment automatitzable. Pel que fa a l'eventual incompatibilitat entre les accions, es poden fixar, en un *sistema expert* (apartat 5.5), una sèrie de regles (no necessàriament absolutes) que habitualment identifiquen aquesta incompatibilitat. Per una banda, per confirmar que les peticions no s'articulen de manera subsidiària (fet que exclouria la incompatibilitat).

Per l'altra, per constatar, per exemple, si pel tipus de procediment (judici verbal, etc.), no és legalment viable l'acumulació (seria una prohibició legal) o perquè s'aprecia, en definitiva, una incompatibilitat directa i flagrant (s'insta la resolució del contracte i simultàniament el compliment per l'altra part d'una obligació prevista en el mateix contracte).

En el cas de l'acumulació *subjectiva* (vàries accions contra diversos demandats o a instància de diversos demandants), s'acostuma a exigir un *nexe* entre les accions, que es remet, sovint, a un nexe entre els fets en què es basa cadascuna d'elles. L'automatització pot semblar més complexa, però l'ús d'eines d'associació dels fets de la demanda amb les persones que refereix pot detectar eventualment una aparent autonomia (distanciament *semàntic*) entre un grup de fets i un grup de persones que faci difícilment viable l'acumulació subjectiva. En aquest cas, si se supera un llindar de baixa probabilitat prefixat, es podria genera l'avís perquè un funcionari del jutjat ho analitzi. Sempre estem parlant, per tant, de detecció de defectes processals relativament probables per generar avisos que suscitin una intervenció humana posterior.

Per últim, acostuma a exigir-se, per raons evidents, com a requisit de l'acumulació, que el tribunal al qual es dirigeixi la demanda sigui competent tant per l'acció principal com per l'acumulada. Aquí caldria *incrustar* al sistema que analitza l'acumulabilitat estricta de les accions l'altre eina que ja hem vist a l'apartat 11.6.3 i que se centraria en la constatació de la competència objectiva (ja sigui per la quantia o per la matèria).

11.6.7. Cosa jutjada i litispendència

La cosa jutjada és un defecte processal respecte del qual les eines d'IA poden ser especialment útils. Entre altres motius, perquè, atesa la seva naturalesa d'ordre públic (amb la cosa jutjada es pretén evitar que hi hagi sentències contradictòries), pot ser apreciada d'ofici, sense necessitat que sigui al·legada per les parts. I, com que és freqüent que una de les parts es trobi en rebel·lia, pot donar-se el cas que, existint una demanda anterior amb un contingut idèntic total o parcialment, el tribunal no n'arribi a tenir coneixement perquè ningú no l'al·lega. Per contra, partint de la indicada actuació d'ofici, podria establir-se una fase prèvia inicial de detecció automatitzada de possibles coincidències de la nova demanda amb una anterior ja ferma, per tal que el tribunal en tingui coneixement quan acudeix a la vista principal del judici o al tràmit que correspongui.

Cal precisar en aquest punt que, actualment, pot arribar a detectar-se pel jutjat, a través del sistema de gestió processal, aquest tipus de situacions, però en casos pràcticament aleatoris (per exemple, quan les dues demandes es reparteixen al mateix jutjat o quan un funcionari té una memòria destacable). També és possible que els sistemes de repartiment dels deganats o que el mateix sistema de gestió processal del jutjat pugui arribar a constatar aquesta duplicitat, probablement després que un funcionari faci la deguda comprovació. Però el cert és que només acostumen a donar-se aquests casos quan la similitud entre les parts i els fets és completa i quan, a més, es donen els factors humans o logístics apuntats. Per contra, del que es tractaria és d'introduir un sistema de detecció de possibles coses jutjades que operés de manera *massiva* respecte de totes les resolucions dictades en una jurisdicció determinada. O, fins i tot, en diferents jurisdiccions. També que tingués codificada una certa *flexibilitat* per poder detectar la coincidència dels casos fins i tot encara que no siguin idèntiques les parts (es pot haver produït, per exemple, una successió per causa de mort) o els fets (podria ser que una demanda anterior respecte de la qual ja s'ha dictat una sentència ferma faci referència a tres fets i que la següent demanda només n'inclogui un).

En aquest tipus d'eines operarien, de nou, recursos d'*OCR* i *NLP* en el context d'una *interoperabilitat* massiva amb la totalitat de resolucions judicials dictades a nivell estatal (a la mateixa jurisdicció o en altres) i d'un *sistema expert* de regles expresses que recullin (cal insistir, amb un cert grau de flexibilitat) les tres *identitats* pròpies de la cosa jutjada material (la de persones, la de fets i la de causa de demanar). Aquesta flexibilitat és rellevant perquè, de nou, del que es tracta no és que el sistema dicti per ell mateix una resolució definitiva que arxivi el cas una vegada detectada la cosa jutjada, sinó que activi avisos de la probabilitat de cosa jutjada (probabilitat que fins i tot es podria graduar). Serà a partir d'aquest moment que el tribunal abordi la qüestió. Una qüestió que, cal reiterar, s'ha de poder apreciar d'ofici i que, d'una altra manera, podria no arribar a coneixement del tribunal.

Per últim, fins i tot amb independència que la part demandada estigui o no personada, un cert grau de flexibilitat en la constatació de la concurrència de les tres identitats pot permetre identificar supòsits de cosa jutjada material no negativa (la que exclou un nou pronunciament), sinó positiva (la que si bé no l'exclou, sí que en condiciona el seu contingut, per exemple, respecte de la declaració de part dels fets provats).

Tot el que s'acaba d'afirmar respecte de la cosa jutjada es podria traslladar, amb les adaptacions oportunes al nivell de l'expressió de les regles del sistema expert, a la qüestió processal de la litispendència, que vindria a ser, simplificant, una cosa jutjada referida no a processos ja resolts per sentència definitiva, sinó a processos que encara estan en tràmit.

11.6.8. Acumulació de procediments

Es tracta de nou d'una eventualitat processal basada en raons d'ordre públic que pretenen evitar que es dictin sentències contradictòries. És viable, per tant, l'apreciació d'ofici, per la qual cosa pot ser molt útil, com en el cas de la cosa jutjada o la litispendència, articular mecanismes de detecció automatitzada pels casos en què no hi hagi una part que posi de manifest la duplicitat de procediments. En matèria d'acumulació de procediments acostuma a ser rellevant que una sentència d'un procediment generi efectes *prejudicials* en la sentència de l'altre. Que hi hagi una *connexió* entre els objectes que generi l'indicat risc de pronunciaments contradictoris o incompatibles. Es tracta, és cert, de nocions força genèriques no fàcils d'automatitzar. Però, de nou, poden establir-se una sèrie de regles generals, flexibles, en un *sistema expert* que acoti com a mínim un espai de certa probabilitat de procediments acumulables, per generar un avís i una intervenció humana posterior. En funció, per exemple, de la coincidència parcial entre els fets i les parts, sense necessitat que es donin les similituds més intenses pròpies de la cosa jutjada.

11.6.9. Litisconsorci passiu necessari

Ens trobem en un nou cas d'una excepció processal que si bé és al·legable per la part demandada, pot ser apreciada d'ofici, ja que afecta a la deguda constitució de la relació jurídica processal i a la viabilitat que els pronunciaments de la sentència puguin fer-se efectius, arribat el cas, per haver-se seguit el procediment contra tots els titulars de les relacions jurídiques afectades. Per tant, la IA judicial podria ser, de nou, útil.

De casos de litisconsorci passiu necessari n'hi ha de legals expressos i de creats jurisprudencialment. Per tant, de naturalesa canviant, fet que podria dificultar la seva automatització per mitjà de regles expresses del sistema expert. Al mateix temps, però, la creació de nous supòsits jurisprudencials de litisconsorci passiu necessari no és tan freqüent ni ràpida com per excloure la viabilitat de formular-los expressament en regles

que, quan sigui necessari, es poden anar modificant. Sembla viable, en definitiva, que el sistema generi un avís si en un cas d'una acció que afecta una diversitat de titulars, la demanda no es dirigeix contra tots ells.

11.7. Automatització de l'admissió de la demanda i de l'emplaçament

Per últim, pel cas que el sistema no detecti problemes processals com els indicats (o que, detectats, s'hagin esmenat), ell mateix podria automatitzar l'admissió de la demanda i el trasllat i emplaçament de la part demandada. I l'emplaçament es podria fer, simultàniament, tant al domicili de la demandada aportat a la demanda com a qualsevol altre que aparegui a les bases de dades disponibles i que es detecti de manera automatitzada. Des de la perspectiva del procediment, la clau és que com a mínim un emplaçament sigui positiu, i és evident que és més probable que es doni abans aquesta circumstància si es fa un emplaçament *múltiple* simultani, que no pas que es facin (com acostuma a fer-se) emplaçaments *successius* a mesura que van resultant negatius. Si eventualment es produeix més d'un emplaçament positiu, no hi hauria cap problemàtica processal i simplement caldria computar el termini per contestar des del primer.

11.8. Consum i clàusules abusives

Tant la normativa processal estatal com la jurisprudència del TJUE preveuen la facultat i deure del tribunal, en casos de consum, d'analitzar d'ofici la possible concurrència de clàusules abusives. De fet, la norma processal espanyola ho preveu pel monitori i les execucions de títol no judicial (no pels declaratius). El TJUE, per la seva banda, és més exigent: ha establert que aquest control s'ha d'activar en qualsevol tipus de procediment, sense distincions. Per tant, atès el principi de primacia del dret europeu (art. 4 bis LOPJ), caldria entendre que aquest control d'ofici s'ha de produir sempre, també en els declaratius. Deixant de banda, però, aquesta qüestió, és indiscutible i objectiu, en qualsevol cas, l'enorme esforç i dedicació de temps que cal destinar als cada vegada més abundants monitoris i a les execucions de títol no judicial. Per tant, està plenament justificada l'anàlisi de si és viable, en aquests casos, algun tipus d'automatització, encara que sigui parcial.

Doncs bé, sense ànim d'entrar en excessius detalls, podríem dir, en primer lloc, que és habitual en els contractes estandarditzats de consum (amb condicions generals de la

contractació) que la lletra sigui molt petita, fet que, unit a una qualitat de digitalització dels documents sovint no òptima, fa que la mera lectura de les clàusules contractuals (tasca imprescindible per valorar-ne la seva eventual abusivitat) resulti molt dificultosa o directament impossible. Per tant, es podria introduir una fase inicial automatitzada de *llegibilitat* dels contractes aportats. Perquè, en cas de no poder ser *llegits* pel sistema, es requerís de manera automatitzada la seva esmena[118].

Pressuposant que es pot llegir el contracte de consum, el que cal dir d'entrada és que, òbviament, hi ha una gran varietat de clàusules (interessos de demora, venciment anticipat, comissions, etc.) i que cadascuna d'elles té uns paràmetres propis per determinar si és, o no, abusiva. Podríem dir que cada clàusula és un món. Fet que, certament, dificulta l'automatització d'aquesta tasca: caldria crear un *sistema expert* que integri, en el seu interior, un *subsistema expert* per a cada tipologia de clàusula. La bona notícia és que el TJUE ha declarat en diverses ocasions que per valorar la validesa d'una clàusula només cal tenir en compte la clàusula en qüestió. És a dir, la seva estricta *redacció* en relació amb la resta de clàusules del contracte. No és determinant, per contra, ni rellevant, la *conducta contractual* que hagin pogut tenir les parts al marge de la *literalitat* del contracte[119]. Per tant, el *terreny de joc* es redueix notòriament: al contracte i a les seves previsions textuals. I aquest fet afavoreix, lògicament, l'eventual automatització d'aquesta tasca per mitjà de tècniques d'*OCR* i de *NLP*, degudament integrades, com dèiem, en *subsistemes experts*.

Partint d'aquestes premisses, serà raonable intentar automatitzar només la detecció de l'abusivitat d'aquelles clàusules que, per la seva mateixa naturalesa i complexitat, s'hi adaptin. En posarem algun exemple:

[118] Pensem que és freqüent que un funcionari doni compte al jutge perquè analitzi les clàusules, que aquest iniciï la tasca i que, després de llegir la demanda i obrir alguns dels documents digitalitzats, constati, al cap d'uns minuts (i amb certa frustració), que no pot llegir el contracte. Acte seguit, ho posarà en coneixement del funcionari, perquè aquest remeti un requeriment a la part actora. Doncs bé, aquestes actuacions no requereixen, segurament, ni molt temps ni molt esforç, però es repeteixen amb força freqüència, per la qual cosa pot ser molt eficient automatitzar-les perquè s'esmenin sense que cap funcionari del jutjat hagi hagut de dedicar-hi ni mig minut.

[119] Per exemple, que el creditor hagi acudit materialment al venciment anticipat després de 9 impagaments no ens interessarà. Només haurem d'acudir a si la clàusula del contracte, tal com està redactada, permet acudir-hi amb un únic impagament.

1) Clàusula d'*interessos de demora*: si són abusius, segons el TS, sempre que (i només quan) superin en més de dos punts els interessos ordinaris, sembla evident la viabilitat de reduir aquests paràmetres a un conjunt de regles d'un *subsistema expert*.

2) Clàusula de *comissions*: si el contracte les preveu de manera automàtica, per quan es donin certs supòsits (impagaments, etc.), però sense exigir que s'hagi produït efectivament un servei o una despesa per part del creditor, llavors se'n pot derivar la seva abusivitat. En aquest cas, la regla del *sistema expert* es podria limitar a constatar si al costat de la previsió (a la mateixa frase o clàusula) hi ha alguna actuació o eventualitat que condicioni la generació de la comissió. Si no és el cas, se'n podria derivar l'eventual abusivitat. Igualment, fins i tot si la clàusula fa referència a alguna actuació que justifica la comissió (gestions de reclamació, etc.), el *subsistema expert* podria comprovar si juntament amb la demanda s'ha acompanyat alguna documentació que acrediti que aquesta actuació es va produir *efectivament*.

3) Clàusula de *venciment anticipat*: aquí el sistema hauria de comparar la duració total del contracte (de préstec, etc.) amb el nombre d'impagaments que exigeix el mateix contracte perquè el creditor pugui acudir al venciment anticipat i declarar degut tot allò que s'havia de pagar en un futur durant tota la vida del contracte. Si es detecta una desproporció (que es pot ajustar segons el tipus de contracte), s'activaria un avís de possible abusivitat.

4) Clàusula de *despeses*: el contracte preveu que totes les despeses seran a càrrec del prestatari. O que una despesa que la llei preveu que ha d'assumir una part, l'assumirà una altra. Aquí la detecció automatitzada sembla, de nou, viable.

Els exemples podrien continuar, però no és aquest l'objecte de la recerca. Només s'afegirà que del que es tractaria és, com s'ha avançat, d'activar avisos de probable abusivitat de certes clàusules. Acte seguit seria el tribunal qui emetria el pronunciament definitiu. Però el suport podria traduir-se no només en l'avís sinó també en els corresponents enllaços directes no ja al document que conté el contracte sinó, directament, a les concretes clàusules que siguin rellevants en cada cas.

Per últim, si les regles rellevants per a aquesta eventual abusivitat ja estan prefixades, res no impediria generar un esbós de resolució que podria ser utilitzada pel tribunal

encarregat del cas. Veiem, en definitiva, que certes tasques de suport i de generació d'esbossos poden ser compatibles amb la deguda i necessària agència humana. A tal efecte, i en matèries més delicades com ara la de consum, seria preferible que el subsistema fos *modulable* a nivell del mateix tribunal que ha de resoldre. És a dir, que les regles d'abusivitat, més enllà que puguin fixar-les el TJUE o el TS, puguin ser introduïdes i matisades per cada òrgan judicial, als efectes de fixar amb precisió el *llindar* a partir del qual el sistema activarà els *avisos* corresponents. D'aquesta manera podria combinar-se un grau significatiu d'automatització amb una preservació raonable de la imprescindible independència judicial.

Per últim, podria plantejar-se, ja des d'un inici del procediment, l'automatització d'un requeriment per esmenar la tan freqüent manca de desglossament dels diversos conceptes que es reclamen (principal, interessos ordinaris, interessos de demora, comissions, penalitzacions, etc.). De nou, l'estalvi en temps seria rellevant.

11.9. Monitoris

En aquest cas es podria plantejar l'automatització d'esborranys de diligència d'ordenació que, en peticions de monitori, constatessin la conveniència de donar compte al tribunal per possibles insuficiències documentals, errors aritmètics o conceptes improcedents (que es podrien prefixar en regles d'un *sistema expert*: danys i perjudicis que excedeixen el marc del monitori, despeses no documentades o no reclamables al monitori, etc.).

Avançant en el següent pas, podria automatitzar-se la generació d'esborranys d'interlocutòria basats en l'art. 815 LEC, quan es detecti que algun concepte o import ha de ser exclòs i que cal fer una proposta d'un import alternatiu. Per exemple, en casos de comissions per reclamació no justificades o d'interessos clarament abusius. O, simplement, d'errors aritmètics.

Pel que fa als monitoris de propietat horitzontal (reclamacions per deutes comunitaris), podria automatitzar-se la comprovació de l'aportació de tots els documents exigits legalment (acta de la junta, certificat del deute, etc.) i la coincidència dels imports reclamats i els reflectits en aquests documents. En aquest cas és rellevant aquesta anàlisi prèvia ja que, en matèria de propietat horitzontal, no és imprescindible que el

requeriment de pagament sigui positiu. Per tant, es pot acudir a l'execució sense que el demandat hagi tingut coneixement efectiu de la reclamació.

11.10. Taxació de costs

En matèria de taxació de costes la IA judicial podria ser força útil. Més del que sembla a primera vista: tot i que les normes orientadores que aproven els Col·legis de l'Advocacia són força nombroses i complicades per ser reduïdes a regles d'un *sistema expert*, el cert és que, si ens situem en un context judicial en el qual s'ha de fixar l'import que en concepte de costes d'una part ha d'assumir l'altra part que hi ha estat condemnada, el criteri rellevant és, principalment, el que fixa la jurisprudència del TS. En concret, la *complexitat* material, *real*, que hagi tingut la causa en qüestió. I no tant, o no només, les indicades normes orientadores[120].

Doncs bé, partint d'aquest fet (no sempre tingut suficientment en compte, tot sigui dit), veurem que una eina automatitzada de gestió del procediment i avaluació de la seva complexitat podria aportar elements de judici rellevants per poder determinar amb més detall quines són les costes més ajustades a allò que realment ha succeït durant el procediment. Podria partir de la tipologia en abstracte (inicial) del procediment (ja detectada a la fase de repartiment que realitza el deganat), però afegint-hi totes les vicissituds posteriors que realment s'han produït.

De nou, es tractaria, només, d'un *input* d'informació (que es podria graduar amb diversos nivells de complexitat o dedicació) que estaria a disposició de LAJ i tribunal per tal d'emetre la seva resolució. Res no impediria, de nou, preveure la generació automatitzada d'esborranys de resolució que no només indiquessin l'indicat nivell de complexitat sinó que també concretessin totes les actuacions processals realment realitzades (amb detall, fins i tot, de la seva duració, si es tracta de vistes orals). Veiem, per tant, que en el cas de la taxació de costes, l'aplicació d'eines d'IA podria implicar, de fet, una millora en la qualitat i motivació de les resolucions judicials.

[120] A títol de mer exemple, podem mencionar la ATS, Civil, secció 1, de 4 de febrer de 2020 (ROJ: ATS 1046/2020 - ECLI:ES:TS:2020:1046A).

11.11. Processos de divisió patrimonial

Aquí ens limitarem a indicar que existeixen projectes en curs d'usos d'eines automatitzades per a tasques de divisions de patrimonis en casos de divorci o de successió per causa de mort. El *Conflict Resolution through Equitative Algorithms* (*CREA*), que té el suport de la UE, es basa en una revisió i adaptació al camp legal de l'algoritme *Spliddit* (Corona et al., 2019)[121]. L'algoritme cerca la divisió més equitativa possible i analitza l'interès respectiu dels membres de la parella respecte de concrets béns del patrimoni, en funció de diversos paràmetres, com les obligacions imposades legalment, el règim matrimonial aplicable o la llista de béns (amb el valor de mercat, si pot o no ser compartit o dividit, la quantitat de diner líquid existent o la llista de deutes o responsabilitats existents). Aquest tipus d'eina, centrada més en l'interès real (hipotètic) de les parts que no pas en estrictes divisions matemàtiques, podria ser utilitzat, també, en el context judicial, per fomentar possibles acords. Es tracta, però, d'experiències encara en fases molt preliminars. Caldrà seguir-ne l'evolució[122].

11.12. Prova documental

Ja hem fet referència a l'apartat 9.2 a la *e-evidence* o prova electrònica, especialment pels casos en què s'ha d'analitzar un gran volum de documentació que no és possible abordar *manualment* i amb detall. Aquestes eines es basen en la selecció prèvia dels documents (o arxius) potencialment *rellevants* a partir dels quals s'automatitzaria la recerca, sobre la totalitat de documents disponibles, dels *altres* documents rellevants: només aquests s'introduiran en el procediment i seran valorats pel tribunal.

[121] El projecte *CREA* (*Conflict Resolution Equitative Algorithms*), Romeo ed al., aplica mecanismes algorítmics de la *teoria de jocs* per a la solució de litigis civils interns o transfronterers, en matèries com dret de família, successions, divisions de béns comuns o consum. L'objectiu és facilitar el tancament d'acords entre les parts abans o durant el procediment judicial. Treballen amb diferents algoritmes, en funció de la complexitat del cas. Diferencien entre drets disponibles i normes d'ordre públic dels diferents estats membres de la UE. Es busca un denominador comú europeu. La idea és que el programari que s'està desenvolupant acabi estant disponible a la plataforma *EU ODR* o al portal *e-justice*.

[122] Un altre exemple seria el programa *Split-Up* (Barona, 2021, p. 650). Ofereix consell en casos de repartiments patrimonials en divorcis. Per desenvolupar-lo, experts legals en dret de família australià van identificar i jerarquitzar 35 factors jurídics rellevants. Després es va alimentar el model amb dades de casos passats. I, suposadament, el programa va aprendre com els jutges havien ponderat els factors prèviament identificats. 20 factors s'haurien generat utilitzant instruments d'aprenentatge profund (xarxes neuronals). El programa permet als usuaris mantenir-hi un diàleg sobre situacions hipotètiques que es poden donar. Es tractaria d'adquirir coneixement sobre la fortalesa o debilitat de la posició de la part.

Si baixem al terreny estrictament processal, podríem localitzar una via relativament senzilla per reduir el volum de documents *concretament rellevants*: de la mateixa manera que abans s'ha proposat condicionar, forçosament, el format i contingut de la demanda i dels documents aportats per la part actora, podria fer-se el mateix exigint, ara, a la part demandada que no només concreti, a les caselles corresponents, els motius d'oposició de la demandada (és a dir, la controvèrsia), sinó també que detalli, ja a la contestació, i també en apartats específicament dissenyats a tal efecte (per poder ser processats pel sistema), quins són els documents rellevants vinculats a cada motiu d'oposició. D'aquesta manera es podria reduir substancialment, com dèiem, el volum de documents *realment rellevants*[123].

Farem esment, per últim, a les eines que generen automàticament *resums* de documents. Podria ser interessant, per exemple, en el moment de dictar sentència, quan es va navegant (a vegades *naufragant*) digitalment pels diversos documents aportats (cadenes de correus electrònics, contractes anteriors, etc.), disposar del resum automatitzat (i ràpid) d'un bloc documental determinat. Potser per descartar-ne la seva rellevància o, per contra, per centrar-hi un major interès i dedicació. També podrien ser interessants eines per identificar (i desglossar) les porcions dels textos dels documents que contenen normes legals, arguments legals o mers fets. O les que detecten anomalies. És a dir, porcions que no responen a la tendència homogènia de la resta de documents (apartat 6.2.2).

11.13. Cessions de credits

Amb la crisi del sector bancari i la seva forta reestructuració, han estat freqüents les cessions de carteres de crèdits entre entitats bancàries o, més freqüentment, a fons d'inversió. Els jutjats han d'analitzar amb detall (o haurien de fer-ho) si els documents aportats són suficients per tenir per acreditada aquesta cessió, com a pressupòsit de la successió processal. Cal constatar, principalment, si el certificat notarial de la cessió del crèdit fa referència al deutor demandat i si la numeració que indica coincideix amb la

[123] Lògicament, la concreció de quins són aquests documents *rellevants* pot ser, en si mateix, una qüestió controvertida. I, de fet, el que s'acabi delimitant tampoc no ha de vincular d'una manera absoluta al tribunal, ateses les regles de la *sana crítica* que han de regir la valoració de la prova. Es tractaria, simplement, d'articular eines de racionalització i reducció de causes innecessàriament voluminoses en termes de prova documental.

numeració del contracte objecte de la demanda inicial (ens sorprendria la freqüència amb què no es dona aquesta segona coincidència). Doncs bé, sembla evident que es tracta de tasques feixugues i mecàniques que consumeixen molt temps i esforç. Per aquest motiu la seva automatització, a més d'aparentment factible, seria especialment desitjable: en cas de detectar-se la no coincidència entre els números, podria generar-se un avís amb enllaços documentals directes tant al número del contracte com al número del certificat notarial. I, acte seguit, pel cas de confirmar-se, manualment, la no coincidència, podria generar-se automàticament un esborrany de resolució judicial que requerís per esmenar o denegués la successió processal, segons els casos. L'eficiència guanyada seria enorme.

11.14. Pericials

Podria automatitzar-se tant l'avaluació del currículum del pèrit com l'eventual concurrència de conflictes d'interessos. A tal efecte, caldria disposar d'una plena *interoperabilitat* amb els registres públics que siguin rellevants.

11.15. Reconeixement facial en compareixences o vistes telemàtiques

Una de les principals problemàtiques que genera la connexió telemàtica és la identificació dels intervinents. La normativa processal permet acudir a qualsevol mitjà adequat, però convindria buscar mecanismes d'identificació adaptats al nou context virtual[124]. Una idea interessant seria demanar a l'intervinent, a l'inici de la declaració, que reprodueixi oralment els últims números del codi CSV que conté el document de citació. O, situats ja en termes d'IA judicial, acudir a tècniques de reconeixement facial.

Es tracta d'un dels usos més delicats de la IA. A la normativa en curs a nivell de la UE es preveu de fet la prohibició d'alguns tipus d'eines d'IA de reconeixement facial de tipus biomètric. Dit això, els majors dubtes es generen en matèria d'investigació delictiva i identificació dels autors de possibles delictes. Per contra, la possibilitat que ara s'apunta

[124] En les compareixences físiques acostuma a ser suficient la presentació física del carnet professional corresponent o el DNI. Podria plantejar-se, per tant, la possibilitat que l'intervinent *telemàtic* mostrés a la càmera el seu carnet professional o DNI. El problema és que sovint no es poden veure amb la claredat suficient, degut, precisament, a la insuficient qualitat i precisió de la imatge. Sempre hi ha la possibilitat que totes les parts estiguin conformes que la persona que apareix a la pantalla és qui diu ser. I sempre es podrà tenir en compte, a més, que per poder establir la connexió haurà hagut de tenir accés al correu electrònic o a una altra via de comunicació amb el jutjat on s'haurà remès la invitació per participar en la vista telemàtica. Sembla, però, que caldria alguna cosa més per a la identificació de l'intervinent.

fa referència a potencials usos de tècniques de reconeixement no biomètriques amb la única finalitat de contrastar la identitat de les persones que participen en una vista judicial telemàtica. Es tracta d'una eventualitat de moment força remota però que caldrà anar seguint.

11.16. Transcripció automatitzada d'àudio a text

Pocs dubtes hauria de generar la utilitat d'una eina com aquesta. Pels casos en què hagi desaparegut la tradicional i obsoleta acta de judici (substituïda per la gravació en vídeo), l'eina seguiria sent útil precisament per poder disposar de la transcripció de les declaracions. Per exemple, pel tribunal a l'hora de dictar sentència. De fet, podria fomentar valoracions de la prova més riques i precises.

No hem d'obviar, tampoc, la possibilitat d'aplicar aquestes eines per transcriure els escassos supòsits permesos legalment per dictar sentències *in voce* i que tants problemes generen quant a la freqüent manca de correlació entre el que s'ha resolt *in voce* i el que s'ha recollit posteriorment de manera succinta per escrit.

El pressupòsit serà, és clar, que la qualitat de la transcripció sigui gairebé perfecte i que generi un elevat grau de confiança que exclogui la necessitat d'acudir al vídeo, excepte en casos puntuals. Són precisament les més avançades eines d'aprenentatge *profund* (xarxes neuronals) les que poden generar aquest grau de fiabilitat.

11.17. Traducció automatitzada

En estreta relació amb l'anterior punt, cal apuntar la possibilitat d'introduir eines de traducció automatitzada. Aquestes poden actuar tant a nivell de traducció de documents com de manifestacions orals. De fet, en aquest segon cas, les manifestacions orals serien primer transcrites a text i només posteriorment entraria en funcionament el sistema de traducció en sentit estricte. És òbvia la rellevància que podrien tenir aquest tipus d'eines en partits judicials on hi hagi una elevada diversitat de llengües. No es tractaria, en principi, d'eliminar la figura de l'intèrpret, que seguirà sent necessària per a la celebració de les vistes amb les degudes condicions i garanties. Només quan aquestes eines adquireixin un grau de desenvolupament i precisió gairebé absoluts, seria viable plantejar-se prescindir de la traducció humana en temps real. Per altra banda, i amb caràcter més general, en aquells partits judicials amb dues o més llengües oficials, com

Catalunya, la disponibilitat d'eines de traducció automatitzada podria fomentar la redacció de sentències en la llengua minoritària en el sector judicial, com el català. També, de fet, la redacció *íntegra* de sentències amb aquesta llengua minoritària (pensem en els casos de sentències que es dicten en català però que tenen parts en castellà, segurament perquè les bases de dades de lleis i jurisprudència només estan disponibles en castellà, o perquè les parts han presentat els seus escrits principals també en castellà).

11.18. Generació d'esborranys de sentències

11.18.1. Introducció

Arribem ara a un dels punts més delicats de l'apartat de *propostes*: l'eventual generació d'esborranys de sentències. Semblaria que ens apropem a la figura del *jutge-robot* que hem abordat al capítol 7. Dit això, si seguim el fil de les anteriors propostes (referides a usos parcials, de suport, de certes tasques de tràmit) i ens desplacem, ara, al moment de dictar certes sentències especialment senzilles, potser no ens semblarà tan inviable la proposta. Que, cal reiterar, consisteix, merament, en generar esborranys, esbossos, de sentències, per ser posteriorment completades, confirmades (o no, en tot o en part) i signades pel titular de la funció jurisdiccional.

11.18.2. Sentències d'aplanament

En primer lloc, podria plantejar-se la generació d'esborranys de sentències d'aplanament. Són senzilles. Són molt pocs els casos en els quals no s'accedeix a l'aplanament. De fet, la qüestió més complexa que cal abordar és si s'imposen, o no, les costes processals, en funció de si hi ha hagut requeriments previs amb un determinat contingut. Doncs bé, semblen clarament automatitzables aquestes eventualitats. El sistema podria constatar, per una banda, si hi pot haver, en funció del tipus d'acció exercitada (això requeriria complementar el sistema de *NLP* amb unes mínimes regles de *sistema expert*), afectació als drets d'un tercer o a una matèria d'ordre públic. I, per l'altra, si hi ha hagut requeriments previs. Després de la constatació, podria generar-se, de manera *completa*, l'esborrany de sentència, que quedaria a l'espera de la indicada *agència humana* de comprovació, modificació (o no) i signatura. Podria pensar-se que una tasca tan senzilla no requereix tan temps ni esforç si es fa manualment. Però

difícilment seran menys de cinc minuts i cal tenir en compte, sempre, l'efecte de l'estalvi acumulat.

11.18.3. Extracció completa de la cronologia dels fets

Si ens desplacem a altres tipus de suport parcial per generar esborranys de sentència, un ús que podria ser alhora automatitzable i escassament *sensible* seria l'extracció, de la documentació de la causa, de la cronologia completa dels fets, creuant demanda i contestació (passa sovint que cada part només selecciona els fets que li interessen i és el tribunal qui els ha de *creuar* després manualment, amb un cost de dedicació força elevat). Fins i tot podria afegir-se, quan fos possible i li constés al sistema, la qualificació automatitzada de si cada fet és o no controvertit i, per tant, necessitat de prova.

11.18.4. Avisos d'omissions de fets, jurisprudència o pretensions

També podria explorar-se que el sistema analitzés de manera automatitzada el projecte de sentència ja redactat pel tribunal (però encara no signat) i pogués, després de *navegar* pel procediment complet i les bases de dades disponibles, generar avisos al tribunal sobre la possibilitat que no s'hagi tingut en compte un fet potencialment rellevant o una resolució judicial específicament pertinent d'un òrgan superior (un antecedent jurisprudencial). O, per últim, que s'hagi omès un pronunciament sobre una pretensió degudament articulada. Lògicament, es tractaria, només, d'avisos, que podrien ser tinguts en compte, o no, pel tribunal. La finalitat seria evitar l'emissió de sentències incompletes o incongruents. I, en correlació, evitar peticions d'aclariment o fins i tot recursos d'apel·lació (cal tenir present que alguns casos de meres omissions no són abordables, en seu d'aclariment, pel mateix tribunal que ha dictat la sentència i aboquen a un inevitable recurs d'apel·lació que, tal vegada, s'hauria pogut evitar amb avisos com els proposats).

11.18.5. Esborranys de sentència complets

Abordem, ara ja sí, la possibilitat de generar de manera automatitzada i *completa* o *gairebé completa* esborranys de sentència en casos, podríem dir, ordinaris. Senzills i estandarditzats, però no de mer aplanament. Caldria reduir al màxim els supòsits, evidentment. Però pensem, per exemple, en aquells en els quals s'apliquen poques normes i les que s'apliquen són d'interpretació clara o no controvertida. O, més

concretament, en els quals les diferents interpretacions possibles i els fets rellevants estan perfectament predeterminats. Podríem posar com exemples certs casos de judicis verbals (tan abundants) per danys de filtracions d'aigua o alteracions elèctriques. O fins i tot judicis verbals per danys materials en accidents de trànsit en els quals l'argumentació jurídica de les sentències es repeteix indefinidament fins al punt que és possible que, no ja les parts, sinó els mateixos advocats ja no les llegeixin, perquè ja coneixen els criteris de cada jutjat. En aquestes resolucions només són realment rellevants, podríem dir, els paràgrafs que analitzen la prova, què va passar i les consideracions sobre la responsabilitat i l'abast dels danys. Res no impediria, per tant, generar de manera automatitzada tot els cos de la sentència (antecedents de fet, peticions principals i normativa aplicable) i, també, extraure de demanda, contestació i documentació aportada (inclosos els informes pericials) les valoracions que sosté cadascuna de les parts. A partir d'aquest cos de partida, el tribunal només haurà d'afegir la concreta valoració dels documents i arribar a una conclusió. De fet, l'esborrany de sentència podria ja incloure a la decisió tots els pronunciaments *possibles* aplicant criteris de congruència que serien introduïts com a regles de *sistema expert*. El tribunal només hauria d'afegir la valoració de la prova i *descartar* (esborrar) totes les solucions *proposades* excepte aquella per la qual hagi optat. No hi hauria, per tant, problemes de dèficit de motivació. I si ens preguntem sobre una possible afectació a la *qualitat* de la justícia administrada, només caldrà recordar que les sentències confeccionades manualment en procediments com els que s'estan analitzant no acostumen a destacar, precisament, per la seva qualitat, interès o innovació.

11.19. Execució

11.19.1 Introducció

Més enllà de la rellevància, en els procediments d'execució, dels controls inicials de la concurrència dels pressupòsits processals (especialment en el cas de les execucions de títol no judicial, en les quals no s'ha tramitat un declaratiu previ), la resta d'actuacions pròpies de la referida execució judicial són substancialment de naturalesa *administrativa* o *pseudo-administrativa*, sense perjudici dels *incidents* que puguin sorgir, que requereixen, lògicament, una intervenció jurisdiccional puntual. Això explica que a altres països l'execució la tramitin òrgans administratius. Es tracta d'un matís rellevant perquè, amb caràcter general, podem admetre que és més factible i desitjable l'automatització

de l'actuació administrativa que no pas la judicial. Perquè és més estandarditzada i repetitiva i es remet a criteris més generals (menys casuístics) i, per tant, més susceptibles de ser predefinits. Per aquest motiu, si hem d'analitzar una activitat que, tot i ser *formalment* judicial, des d'un punt de vista *material* la podem gairebé equiparar a l'administrativa, podrem tenir, d'entrada, una actitud més receptiva a l'eventual implementació d'eines *executives* d'IA judicial. Sense perjudici, és clar, de ser molt curosos en cada cas quant a les possibles afectacions de drets i garanties processals que es puguin produir.

11.19.2. Esborranys d'interlocutòries despatxant l'execució

Dit això, no sembla excessivament problemàtica l'eventual automatització de la generació d'esborranys d'interlocutòria de despatx d'execucions de títol judicial, amb anàlisi automatitzada de la concurrència de tot els pressupòsits processals: que existeixi un títol judicial ferm, que coincideixin les parts del declaratiu i de la demanda executiva, que coincideixin, també, els imports de la condemna i els reclamats, que la quantitat instada en concepte d'interessos i costes de l'execució no superi el límit legal (sorprendria l'estadística dels casos en què això succeeix), etc.

En el cas del despatx de les execucions de títol no judicial, la situació és més complexa, perquè ja no s'ha tramitat, per definició, un declaratiu previ amb el dictat d'una sentència respecte de la qual s'hagi de contrastar si s'hi ajusta la demanda d'execució. Per contra, és aquesta demanda d'execució la que, juntament amb els documents aportats, justificarà, o no, el despatx de l'execució. Per tant, l'anàlisi que cal fer és més delicada. I, a priori, menys adequada per a la seva automatització. Dit això, però, sí que sembla factible una tasca automatitzada que consisteixi en contrastar la coincidència de parts i imports entre la demanda i els documents aportats (contracte, certificació del deute, requeriments previs, etc.). A més, si ens trobem en un cas de consum, caldrà acudir al control d'ofici (parcialment automatitzable) de les possibles clàusules abusives, com hem vist a l'apartat 11.8.

11.19.3. Localització, avaluació i realització dels béns

Per últim, una de les tasques que amb més claredat seria viable (i fins i tot desitjable) automatitzar és la localització de béns del demandat. Ja existeixen, és cert, eines de localització ràpida de béns, per mitjà de la centralització de registres i la simplificació de

les cerques. Però no estan dotades d'utilitats estrictament automatitzades. Podria acudir-s'hi, de fet, per a la mateixa realització dels béns. En concret, en tasques com les següents: avaluació de la realitzabilitat de cada bé, determinació de la modalitat de realització més adequada i fixació de la corresponent prelació entre tots els béns localitzats, en funció del benefici que pot aportar a l'executant i el perjudici que genera la seva realització a l'executat, tot plegat en conjunció amb els criteris legals d'inembargabilitat, que serien reduïbles a regles de *sistema expert*. A partir d'una completa *interoperabilitat* amb tots els registres públics i privats patrimonialment rellevants, es podria generar, de manera automatitzada, com proposa Nieva (2018), un pla d'execució del qual se'n donaria trasllat al demandat. Així, amb la implicació del demandat (a qui li pot interessar una determinada modalitat o prelació en la realització dels béns), podria guanyar-se en termes d'eficiència i celeritat. De fet, una major celeritat en la localització dels béns realitzables, unida a la consolidació del referit pla d'execució (amb algunes modalitats de realització automatitzada), podria implicar, a mig o llarg termini, la pèrdua de pes d'una institució tan ancestral com l'embargament.

11.20. Possibles usos de la IA judicial en la jurisdicció penal

11.20.1. Eines d'IA penal judicial i eines d'IA d'investigació policial

En matèria d'eines d'IA amb rellevància penal, hem de fer dues precisions inicials: per una banda, es tracta d'un dels camps més desenvolupats dels usos *públics* de la IA; per l'altra, però, hem de diferenciar entre els usos *policials* o d'*investigació criminal* (predicció, prevenció i investigació dels delictes) i els usos estrictament *judicials* o, com a mínim, amb implicacions directament judicials. Aquests segons no estan, en absolut, tan desenvolupats. I cal recordar que l'objecte d'aquesta recerca són les eines d'IA judicial.

En una zona més intermèdia o indefinida (si bé més policial que judicial) podríem situar eines destinades a la reconstrucció dels fets sobre la base dels vestigis disponibles. Nieva (2018, p. 26) ens parla de sistemes com l'*STEVE* (centrat en la construcció de fets coherents), l'*ECHO* (que genera hipòtesis i estratègies d'acusació i defensa) o l'*ALIBI* (que realitza pronòstics sobre diferents explicacions del comportament de l'investigat). I en una zona més propera a l'àmbit judicial hi tindríem les *autòpsies virtuals*, amb les

quals es generen diferents alternatives de relat del que ha pogut passar en funció de les dades obtingudes.

En qualsevol cas, les eines preponderantment *policials* o d'*investigació* no seran objecte d'una anàlisi detallada, amb independència que, lògicament, una actuació policial d'investigació pot acabar esdevenint una prova en el marc del procediment penal (per exemple, els sistemes automatitzats de correlació de mostres d'ADN[125]).

Feta aquesta important precisió inicial, sí que direm que les eines d'IA judicial utilitzades per prevenir o investigar el delicte presenten, per definició, dos riscos importants: en primer lloc, el de confirmació, reforçament i expansió de possibles biaixos que podien existir en les dades amb les quals s'entrena el model; i, en segon lloc, lògicament, el risc que es normalitzin tècniques de control generalitzat i indiscriminat de la ciutadania (pensem, per exemple, en els sistemes automatitzats de prevenció i detecció del discurs d'odi a les xarxes socials). Dit això, es tracta, clarament, de qüestions que transcendeixen l'objecte de la recerca.

Si retornem al camp penal estrictament *judicial*, el primer que constatem és que, a diferència d'altres àmbits, el risc potencial de vulneracions o afectacions a drets fonamentals és més elevat, atesa, entre altres coses, la permanent posada en qüestió del dret a la llibertat. A aquesta circumstància s'hi afegeix que, des d'un punt de vista estadístic, els delictes són menys freqüents que altres conflictes jurídics, fet que implica un menor (en termes relatius) fons acumulat de casos per fer un tractament i processament realment massiu de dades. Per últim, el fet que en el marc de la jurisdicció penal segurament siguin més rellevants factors personals específicament presents en les parts del procediment (víctima denunciant i denunciat) és un altre factor contrari a la possible automatització (estandardització) de la tasca judicial (Nieva, 2018, p. 36).

Dit això, tampoc cal concloure que en seu penal haguem de topar, necessàriament i per definició, amb una inviabilitat absoluta d'automatitzar qualsevol tasca judicial. Posem alguns exemples merament il·lustratius:

[125] *AI Case Study: Probabilistic Genotyping DNA Tools in Canadian Criminal Courts*, Law Commission of Ontario, juny de 2021.

1) Podríem apuntar, per exemple, dins dels usos estrictament judicials, la possibilitat d'automatitzar, en seu d'execució penal de sentències privatives de llibertat, la determinació del grau o del règim concret que hagi d'aplicar-se (Barona, 2021, p. 678). Seria aparentment viable tenint en compte que es tracta d'un supòsit de *discrecionalitat dèbil* (apartat 7.4), en el qual la norma preveu expressament els factors que cal tenir en compte, alguns dels quals poden reduir-se aritmèticament (transcurs del temps de compliment de la condemna, grau de satisfacció de la responsabilitat civil fixada a la sentència o la bona conducta).

2) Podria plantejar-se, també i per raons similars, l'automatització del control o gestió de l'execució de les penes privatives de drets previstes a l'art. 39 CP (Barona, 2021). En aquest cas caldria disposar d'una adequada interconnexió i interoperabilitat de diversos arxius o registres públics.

3) Per últim, situats ja en seu de beneficis penitenciaris, no sembla forassenyat la possibilitat d'automatitzar (sempre, és clar, amb l'agència humana posterior) la concreció de les condicions materials en què s'han de desenvolupar les penes privatives de llibertat (Barona, 2021, p. 680). En aquests casos caldrà tenir en compte factors com la bona conducta, la participació en activitats de reinserció o l'evolució positiva, tots ells fàcilment computeritzables.

11.20.2. Proposta de Resolució del Parlament Europeu sobre la intel·ligència artificial en dret penal

Per adquirir una adequada perspectiva sobre les implicacions jurídiques i en impacte sobre drets fonamentals que poden tenir les eines d'IA penal, ens és de molt interès la recent Proposta de Resolució del Parlament Europeu *sobre la intel·ligència artificial en dret penal i el seu ús per part de les autoritats policials i judicials en matèria penal*[126].

Ens recorda que el desplegament de la IA en el camp de l'aplicació de la llei i el poder judicial no s'ha de considerar com una mera qüestió de *viabilitat tècnica*, sinó com una decisió *política*. Insisteix en la importància de la *qualitat* de les dades utilitzades i que els *biaixos* poden ser inherents als conjunts de dades subjacents, especialment quan s'utilitzen dades històriques. Hi hauria una tendència de la IA a augmentar gradualment

[126]https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/2016(INI)

i, per tant, perpetuar i amplificar la discriminació existent, en particular per a persones que pertanyen a determinats grups ètnics o comunitats racialitzades: les dades utilitzades per entrenar els algoritmes predictius de policia poden reflectir (de fet, reflecteixen) les prioritats de vigilància existents. Caldria, a més, prestar molta atenció a la *asimetria* existent entre aquells que fan servir tecnologies d'IA i els que hi estan sotmesos. Especialment, pel que fa referència a l'impacte sobre els drets de defensa dels sospitosos. En aquest sentit, caldrà fer front als obstacles per obtenir informació *significativa* sobre el seu funcionament i a la conseqüent dificultat per impugnar judicialment els seus resultats. En aquest sentit, caldria consolidar el dret de les parts a tenir accés al procés de recollida i avaluació de les dades.

Per altra banda, l'informe recorda que les persones confien excessivament en la naturalesa aparentment objectiva i científica de les eines d'IA i no consideren la possibilitat que els seus resultats siguin incorrectes, incomplets, irrellevants o discriminatoris. Es tractaria del *biaix d'automatització* del qual ja hem parlat.

Aquest informe detalla alguns usos judicial penals de la IA: entre altres, donar suport a les decisions sobre la presó preventiva, calcular la probabilitat de reincidència o l'anàlisi i predicció de la jurisprudència. Atesa aquesta diversitat d'eines, serien també diferents els seus graus de fiabilitat, precisió i impacte en la protecció dels drets fonamentals i en la dinàmica dels sistemes de justícia penal. En aquest sentit, seria clau la *traçabilitat* dels sistemes d'IA i disposar d'un reflex del procés de presa de decisions que descrigui les seves funcions, capacitats i limitacions. Caldria, a més, conservar la documentació completa de les dades utilitzades per la creació del sistema, el seu context, finalitat, precisió i efectes secundaris, així com el processament aplicat pels constructors i desenvolupadors dels algoritmes.

Finalment, l'informe constata que diverses ciutats dels Estats Units han deixat d'utilitzar sistemes de policia predictiva després d'algunes auditories. En concret, els departaments de policia de la ciutat de *Nova York* i *Cambridge*, *Massachusetts*, haurien eliminat progressivament els seus programes de policia predictiva a causa de la seva manca d'eficàcia i de l'impacte discriminatori que generen.

11.20.3. Predicció del risc de reincidència o d'incompareixença

Hem parlat ja en diverses ocasions d'aquest tipus d'eines, especialment de *COMPAS*: quan analitzàvem els riscos de biaix (apartat 4.2.8), la contestabilitat algorítmica (apartat 4.5) o les experiències d'IA judicial existents als EEUU (apartat 10.3). Ens hi remetem, per evitar reiteracions. Ara afegirem o recordarem, només, algunes referències a casos judicials que han abordat aquestes eines:

1) Ja hem analitzat el cas *State v. Loomis*, relatiu a *COMPAS*. Dèiem que el tribunal va concloure que no s'havia vulnerat el dret a un procés degut perquè la informació que va generar l'eina no hauria estat *determinant*. S'hauria arribat al mateix resultat si no s'hagués disposat d'aquesta informació. Afegia que les variables tingudes en compte eren públiques i que la informació processada es basava o bé en la mateixes respostes de l'acusat al qüestionari o en els historials també públics sobre la seva activitat criminal. Per últim, tot i que l'eina processa dades de grup, la concreta combinació d'aquests factors grupals amb altres que no ho són generaria, segons el tribunal d'apel·lació, una resolució judicial suficientment *individualitzada*.

2) En el cas *Malenchik v. State*, el Tribunal Suprem d'Indiana va concloure que el tribunal d'instància havia tingut en compte altres elements de judici a més de la informació proporcionada per l'eina predictiva, com ara l'historial criminal. La informació automatitzada no hauria operat, per tant, com un factor agreujant *independent*.

3) En el cas *Brady v. Maryland* s'aborda si la fiscalia té el deure de mostrar i aportar tota la informació obtinguda amb eines d'IA (reconeixement facial, eines de predicció del risc de reincidència, etc.) com a conseqüència de la obligació de mostrar, també, tota la prova *exculpatòria*.

11.20.4. Eines de predicció a Catalunya

Per altra banda, l'informe "Intel·ligència Artificial, Decisions Automatitzades a Catalunya", de l'Autoritat Catalana de Protecció de Dades[127], ens recorda que a Catalunya fa uns deu anys que s'apliquen programes similars al COMPAS per detectar la reincidència criminal en adults i joves. S'apunta que "fins a la data, cap investigació no ha demostrat

[127] https://apdcat.gencat.cat/ca/documentacio/intelligencia_artificial/.

que hi hagi biaixos perjudicials per als interns". L'investigador Carlos Castillo –director del grup d'investigació de Ciència de la Web i Computació Social de la Universitat Pompeu Fabra (UPF)– ha realitzat diverses investigacions sobre aquests sistemes intel·ligents i, segons el seu criteri, funcionen prou bé: "els tècnics que en fan ús, en última instància, valoren individualment els resultats que ha donat la màquina i decideixen la mesura que cal aplicar"[128].

El *RisCanvi*[129], per la seva banda, és un protocol o eina de valoració del risc posat en marxa el 2009 a totes les presons de Catalunya, per estimar (en casos de peticions de permisos, etc.) les possibilitats que una persona torni a delinquir una vegada ha sortit de la presó. La predicció de la criminalitat seria individualitzada i personalitzada. Es basa en 43 variables combinades (inclòs el comportament violent de l'intern a la presó) i és el tècnic qui acaba prenent la decisió. Si el sistema detecta una alta probabilitat de delinquir, alerta al tècnic amb un senyal vermell. En aquests casos no es tracta tant d'excloure l'efectivitat dels permisos com d'activar un seguiment diari, una polsera electrònica o el contacte amb un familiar, segons l'informe.

En matèria de reincidència juvenil, el programa *Structured Assessment of Violence Risk in Youth* (*SAVRY*[130]) funciona amb la mateixa lògica que el *RisCanvi*, si bé té menys factors de valoració. És més manual i la valoració final depèn més del tècnic, perquè, en el cas dels joves, els canvis de comportament serien molt bruscos o van més de pressa que els dels adults.

[128] Carles Castillo forma part del Grup d'investigació de Ciència de la Web i Computació Social. Citat a l'informe "Intel·ligència Artificial, Decisions Automatitzades a Catalunya", p. 43.
[129] CEFJE.
[130] Informe "Valoració del risc de reincidència amb joves infractors", Direcció General d'Execució Penal a la Comunitat i de Justícia Juvenil, Maig de 2012.

Annex 1. Tipus d'algoritmes

Per raons de claredat expositiva, hem relegat a aquest annex 1 una possible relació (perquè n'hi ha moltes de relacions possibles) dels principals tipus d'algoritmes que podem trobar en els models d'IA que més ens interessen. L'objectiu d'aquest apartat no és tant aprendre l'inabastable món dels algoritmes (atesa la seva magnitud i caràcter permanentment canviant) com adquirir consciència de la seva immensa diversitat i dels condicionants, potencialitats i limitacions que hi pot haver darrera d'un algoritme. Els agruparem, quan sigui possible, segons el tipus de funció que acostumen a realitzar (predicció, classificació, etc.). Començarem amb els algoritmes més senzills, els de regressió lineal, i anirem augmentant-ne progressivament la complexitat, fins arribar a nocions tan estranyes com les de *Depth First Search*. Seguirem, a més de Gerón (2020), Brownlee (2019).

1. Algoritmes de regressió (predicció)

La regressió és una de les eines més importants en l'estadística i l'aprenentatge automatitzat. Podríem dir, de fet, que aquest segon comença amb la regressió. És la tècnica paramètrica que ens permet prendre decisions i fer *prediccions* basades en les dades a través de l'aprenentatge de les relacions entre les variables d'*entrada* i les de *sortida*. Les segones són *dependents* de les primeres i la regressió ens ajuda a aprendre com el valor *y* canvia a mesura que canvia *x*.

Un model de regressió *lineal* és el més senzill. El menys complex. És útil en les tasques de predicció amb variables *contínues* (és a dir, quan el valor de la variable pot ser qualsevol o infinit entre dos valors qualsevols, per exemple 1 i 2: 1; 1,13; 1,69; 1,78; 2)[131].

En la regressió lineal el model assumeix que hi ha una relació *lineal*, directa, entre les variables que s'introdueixen (variable *x*, independent) i la *sortida* individual que genera (variable-objectiu *y*, que és la dependent). Aquesta relació pot ser establerta ajustant-li la millor línia possible: la gràfica serà senzilla. És a dir, caldrà buscar l'equació de la línia recta i els significats de les variables. I la mesura de l'*error* del model es calcularà com

[131] Per contra, les variables *discretes* seran aquelles variables (sempre numèriques) que tenen un determinat i tancat número de valors entre dos valors qualsevols (per exemple, entre 1 i 5: 1, 2, 3, 4 i 5).

la diferència entre el valor *real* i el valor *previst*. L'objectiu serà trobar la línia recta que millor s'ajusti als punts dispersos de l'eix XY.

Quan la regressió *lineal* tingui una única variable independent, serà *simple* (per exemple, predir el preu de les cases en funció de la seva mida). Si en té diverses, serà *múltiple*. Sovint passarà, però, que les mostres de dades, atesa la seva major complexitat (en termes de relació entre els punts), no es podran ajustar, una vegada projectades a l'eix XY, per mitjà d'una línia recta amb regressió lineal. Llavors acudirem a la regressió *polinomial* amb la qual abordarem els problemes de *no linealitat*: aquells en els quals les variables no són linealment dependents i cal construir corbes complexes per obtenir representacions més ajustades a les situacions del món real.

Els algoritmes de regressió lineal més coneguts són els següents: *Ordinary Least Squares Regression* (*OLSR*), *Linear Regression*, *Logistic Regression*, *Stepwise Regression*, *Multivariate Adaptive Regression Splines* (*MARS*), *Locally Estimated Scatterplot Smoothing* (*LOESS*).

2. Arbres de decisió

Els arbres de decisió prenen com a *input* un vector amb els valors dels atributs i retornen una decisió en forma de *output* individual. Tant l'*input* com l'*output* poden ser valors discrets o continus. Si els valors són discrets i, a més, la resposta només pot tenir dos valors (vertader o fals, etc.), es tractarà d'una funció de classificació *Booleana:* l'atribut final o resposta serà *vertader* si i només si els atributs de l'input satisfan un dels camins que porten a la fulla (*leaf*) amb valor *vertader*. Cal tenir present, però, que la classificació *Booleana* és, només, una de les modalitats que poden adoptar els arbres de decisió.

L'arbre processarà el conjunt sencer de dades com un camí o tronc (*root*) que, en condicions específiques, comença a dividir-se en branques o nodes (*nodes*) internes i pren una decisió quan genera una sortida en forma de fulla (*leaf*). Prendrà una decisió per mitjà d'una seqüència de tests: cada *node* intern de l'arbre es correspon amb un test del valor d'un dels atributs. Les *branques* que surten del node són etiquetades amb els possibles valors dels atributs. I cada fulla (*leaf node*) especifica el valor que ha de ser retornat per la funció.

Els arbres poden oferir molt bons resultats quan les dades són principalment *categòriques* i depenen de condicions que es poden donar o no. Ja hem vist la distinció entre les variables *contínues* (número infinit de valors numèrics entre dos valors qualsevols) i les *discretes* (número determinat, tancat, de valors numèrics entre dos valors qualsevols). Doncs bé, les variables *categòriques* pròpies dels arbres de decisió són aquelles, no necessàriament numèriques, que integren un número finit de grups o categories. Poden no tenir un ordre lògic. Un exemple seria el sexe (femení, masculí, etc.).

Veiem, així doncs, que els contextos amb dades categòriques i dependents de situacions que es poden donar o no són freqüents en l'àmbit legal. També és cert, però, que només els podrem abordar de manera adequada i sòlida amb un arbre de decisió si el problema és relativament senzill i, sobretot, tancat quant a les condicions rellevants, situació que no sempre es dona en l'àmbit jurídic. De fet, el més freqüent és que no es doni.

En el context dels arbres de decisió s'utilitza una gran varietat d'algoritmes. En destacarem tres, no per analitzar-los a fons, sinó per fer-nos una idea aproximada sobre què és, exactament, un algoritme:

a) *ID3* (*Iterative Dichotomiser 3*): creat l'any 1986. Aquest algoritme genera un arbre que tracta la totalitat del conjunt de dades com un únic node. Itera respecte de cada atribut i divideix les dades en subconjunts per calcular l'*entropia* (grau d'incertesa) o el guany en informació que es produeix per l'atribut en qüestió. Té un risc alt de *sobre ajustar* les dades.

b) *C4.5*: és el successor més avançat del ID3. Ja no opera, necessàriament, amb variables categòriques. Va definint de manera dinàmica un atribut discret (basat en variables numèriques). Després de la divisió, té en compte l'atribut que genera el major guany relatiu en informació. Converteix els arbres entrenats en regles *if-then*, avalua la precisió de cada regla i decideix l'ordre en què ha de ser aplicada. Després de construir l'arbre, el va *podant* (*pruning*) eliminant les branques que tenen menys importància. És a dir, si es constata que la precisió de la regla millora sense una determinada precondició.

c) *CART* (*Classification and Regression Trees*): pot realitzar tasques de classificació i de regressió (predicció). Es basa en la mesura de la dispersió estadística i no tant en el

guany en informació. En la tasca de classificació, indica la puresa de les fulles dels nodes. En la tasca de regressió, busca la millor predicció sobre la base de la funció de costos.

3. Boscos Aleatoris (Random Forest)

Es tracta d'un algoritme de classificació d'aprenentatge *supervisat* que consisteix en una col·lecció o conjunt (un *bosc*) d'arbres de decisió, com els que acabem de veure. Cada arbre dona una classificació i el *bosc* (el *Forest*) escull la millor classificació. Aquest tipus de conjunts d'arbres són adequats per abordar problemes de manca de disponibilitat de certs valors.

4. Màquines de vectors de suport (Support Vector Machine o SVM)

Formen part de l'aprenentatge automatitzat *supervisat*. S'utilitzen principalment per tasques de classificació. Per entendre-les, hem de pensar en una gràfica o espai (de *n* dimensions o atributs) en el qual projectem a les coordenades corresponents, en forma de punts, les mostres de dades en funció dels valors dels atributs respectius. Cada punt serà un *vector suport*. I el *classificador* serà la línia que divideix i classifica les dades (els punts) en diferents grups.

L'objectiu del SVM serà trobar l'*hiperplà* (espai entre dues línies, per entendre'ns) que classifiqui (separi) sobre la gràfica de la manera més distintiva possible els punts de dades. Per separar dos grups de dades, sempre hi haurà molts hiperplans possibles. Es tractarà de trobar el pla que tingui el màxim de marge (espai) o la màxima distància possible entre els punts de dades de les dues classes. Maximitzant aquesta distància, augmentarà les probabilitats que en el futur, quan s'apliqui el model a noves dades (nous punts en la gràfica), la classificació que se'n faci sigui adequada.

Un exemple conegut d'algoritme *SVM* és l'anomenat *Kernel SVM*.

5. Naïves Bayes

És una tècnica classificatòria que segueix el teorema de *Bayes*. És a dir, que assumeix que els *predictors* són independents: que la presència d'un atribut particular en una classe no està relacionada amb la presència d'altres atributs. És senzill de construir i

especialment útil amb grans conjunts de dades. Se li dona bé la categorització de documents. Així, en les notícies, pot detectar i ordenar els temes d'un article, si bé actua amb categories prefixades. Pesa les paraules i fa assignacions.

6. Trobar els Veïns més Propers (Finding Nearest Neighbors)

Pot ser utilitzat, per exemple, en els sistemes de recomanació de pel·lícules. Busca en el conjunt de les dades el punt més proper a la dada d'*input* que se li ha donat.

7. Classificador K-Nearest Neighbors (KNN)

És un dels algoritmes més senzills de l'aprenentatge automatitzat. És útil tant en tasques de classificació com de regressió. Buscarà i trobarà un número determinat (el número *K* que fixarà l'usuari) de mostres d'entrenament que es troben el més properes en distància a la nova dada que ha de ser classificada i que, precisament, obtindrà la seva *etiqueta* de la dels seus mateixos veïns. En concret, serà l'etiqueta que decideixi el *vot majoritari* dels veïns: s'assignarà la classificació que sigui més comuna entre aquests veïns, solució que es mesurarà amb una funció de distància.

En aquests casos, l'algoritme, més que crear regles, treballa directament en exemples ja apresos. Ha tingut èxit en el reconeixement de caràcters o l'anàlisi d'imatges. És un algoritme més car d'implementar que altres algoritmes utilitzats per tasques de classificació, perquè requereix normalitzar les variables i preprocessar les dades per eliminar-ne el *soroll*. Operacions, aquestes, en les quals es pot generar una discriminació o biaix indesitjats.

8. Agrupació (clusterin)

L'algoritme divideix un conjunt d'observacions en subconjunts, anomenats *clusters*, en funció de criteris de similitud i dissimilitud: les observacions del mateix *cluster* són similars entre sí en una mesura que són dissimilars a les dels altres *clusters*.

És una manera eficient d'iniciar l'anàlisi de les dades: dividir-les en grups lògics. Extreure'n valor, especialment quan es tracta de conjunts voluminosos de dades no estructurades. D'aquesta manera podem fer una mirada fins i tot ràpida a les dades per detectar-hi possibles pautes o estructures subjacents abans d'abordar una anàlisi més

focalitzada. Per tant, en alguns casos la partició de les dades serà el resultat final buscat. En altres, un prerequisit per abordar altres tasques automatitzades.

Un dels algoritmes que s'utilitzen per tasques d'agrupament és el *K-Means*: assumeix inicialment que ja coneixem el nombre de *clusters* que necessitem. És un algoritme iteratiu. Fixa el número de *clusters* i escull un punt per cadascun d'ells (el *centroid*). De manera aleatòria assigna cada punt a un *cluster*. Repeteix successivament aquesta operació. En cada iteració s'actualitzen les localitzacions dels *centroids* fins que aquestes arriben a un estat òptim en el qual es produeix una *convergència*. Al final podrem assignar cada punt nou al *centroid* que tingui més a prop.

Lògicament, les dades del món real no s'organitzen *naturalment* en números determinats de *clusters*. Cal intervenir-hi per dibuixar o visualitzar les inferències. Aquesta artificiositat exigeix algun tipus de control de la qualitat i precisió del funcionament dels algoritmes d'agrupament. Amb aquesta finalitat s'utilitza l'eina d'anàlisi *Silhouette*, que amida la distància entre els *clusters*. En concret, com d'a prop està cada punt d'un *cluster* dels punts pertanyents al *cluster* veí.

9. Algoritmes de reducció de la dimensionalitat

Amb aquests algoritmes es pretén resumir o presentar les mateixes dades disponibles però utilitzant menys informació. L'aprenentatge és, en aquest cas, *no supervisat*. La finalitat és visualitzar o simplificar les dades. De fet, una vegada s'ha aplicat aquest mètode no supervisat a les dades, llavors aquestes podran ser utilitzades per altres mètodes sí supervisats. Com veiem, els recursos són combinables. Alguns dels algoritmes més coneguts de reducció de la dimensionalitat són, entre altres: el *PCA* (*Principal Component Analysis*), el *PCR* (*Principal Component Regression*) o el *PLSR* (*Partial Least Squares Regressions*).

10. Processament de Llenguatge Natural (Natural Language Processing o NLP)

És la manera que té la IA de comunicar-se amb sistemes intel·ligents. Ho fa per mitjà de llenguatges naturals, com l'anglès o el català. L'*input* i l'*output* d'un *NLP* pot ser tant text escrit com oral. Es tracta, de fet, d'una de les modalitats algorítmiques més rellevants per l'àmbit del dret, per la qual cosa l'hem analitzada amb més deteniment al capítol 9.

Es tracta de tasques força complexes i modulars, que impliquen diverses fases, des de l'anàlisi lèxica, sintàctica (*Parsing*) o semàntica fins a la *integració* del discurs (el significat d'una sentència pot dependre del significat de la sentència que la precedeix i pot incidir, a la vegada, en el significat de la següent). Això ens porta a l'anàlisi *en el temps* de sèries de dades i als models de *predicció seqüencial*: a la predicció d'allò que vindrà a continuació donat un *input* determinat i sobre la base del que s'ha observat abans. Per aquest motiu les dades s'estructuren en diverses sèries d'intervals particulars. La predicció pot referir-se a qualsevol objecte que vingui a continuació. Des d'un símbol, un número, el temps de demà o, entre altres, el proper terme en una conversa. En aquest sector d'usos, molt ampli, hi podem trobar les prediccions del valor de les accions, les prediccions meteorològiques o, fins i tot, les recomanacions de productes.

11. Cerca heurística en la IA

Mai no podrem estar segurs que la solució que ens dona un algoritme és la correcte. Molts problemes són *exponencials*: ofereixen tantes solucions possibles que no és materialment possible comprovar-les totes. Ens hem de acontentar, sovint, amb la solució més *probable*. En aquest sentit, però, sí que es disposa de certes tècniques *heurístiques* amb les quals reduir, restringir o estimular la cerca i eliminar i descartar les solucions clarament *incorrectes*.

Aquesta cerca pot ser *no informada* o *cega* quan únicament disposem de la definició del problema i no tenim més informació. Aquí els exemples podrien ser el *Breadth First Search* (*BFS*) i el *Depth First Search* (*DFS*).

Per contra, la cerca serà *informada* quan disposem d'informació extra sobre els *estats* que podrem utilitzar per computar les preferències. En aquests casos podem tenir algun tipus de control estratègic. Un exemple seria el *Best First Search* (*BFS*).

12. Processament de la imatge

En aquest cas s'estudia la imatge per transformar-la. Tant l'*input* com l'output són imatges. Les aplicacions són diverses, des de la robòtica (localització d'un robot), la navegació, l'evitació d'obstacles o la interacció de robots amb humans, per exemple, per servir-los. En el cas de la medicina, la classificació i la detecció de malalties (per exemple, de tumors), la reconstrucció d'òrgans humans en tres dimensions o la cirurgia visualment

guiada. En seguretat, els sistemes de control biomètrics (iris, empremta digital o reconeixement facial) o la videovigilància amb detecció d'activitat o comportaments sospitosos. En transport, els vehicles autònoms o la monitorització de la seguretat en la conducció. O en la indústria, la inspecció o detecció de defectes, l'assemblatge, la lectura dels codis i de les etiquetes d'empaquetatge o la comprensió de documents (*Optical Character Recognition* o *OCR*).

13. Visió computeritzada

Una de les tasques que pot emprendre la IA és la replicació o modelització de la visió humana per mitjà de software i hardware. Per exemple, reconstruir o entendre una escena en tres dimensions a partir de la seva imatge en dues dimensions.

Hi ha tres nivells diferents en la visió computeritzada: el baix (amb el qual es pretén merament processar una imatge per extreure'n atributs), l'intermedi (reconeixement d'objectes i interpretació d'escenes en tres dimensions) i l'alt (descripció *conceptual* d'una escena, como ara l'activitat o el comportament que s'hi desenvolupa o la intenció que hi ha darrera).

14. Algoritmes de detecció de regles d'associació

Es tracta d'algoritmes que extreuen les regles que expliquen millor les relacions observades entre les variables de les dades. Són útils si cal abordar una gran base de dades multidimensional. Alguns d'aquests algoritmes són l'*Apriori* i l'*Eclat*.

15. SGD: Stochastic Gradient Descent

Es tracta d'un algoritme propi de models senzills i lineals. Processa les instàncies de dades d'una en una i de manera independent. Entrena de manera estrictament aleatòria (d'aquí el qualificatiu d'estocàstic) en el conjunt complet de dades. Emet una puntuació basada en una funció de decisió. Si se supera un *llindar* determinat, el cas pertanyerà a la classe positiva. Si no, a la negativa. Si elevem el *llindar*, s'incrementa la *precisió* (hi haurà menys falsos positius i més vertaders negatius), però alguns vertaders positius poden convertir-se en falsos negatius (és a dir, es redueix la *sensibilitat*).

La clau resideix, per tant, en on se situa el llindar. Alguns mètodes ens hi poden ajudar: el *cross_val_predict*, que prediu no les prediccions (valgui la redundància) sinó les *puntuacions* de les decisions; el *precision_recall_curve*, que calcula la *precisió* i la *sensibilitat* per tots els llindars possibles; o el *Matplotlib*, que ens indica visualment on comença a baixar la *precisió*, per tal de fixar el llindar just abans.

16. Ensemble Algorithms

Per acabar aquest annex, farem referència als mètodes *combinatoris* d'algoritmes. Es tracta de models compostos d'altres models més febles que han estat entrenats independentment i les prediccions dels quals són combinades d'una manera determinada per tal de fer una predicció de conjunt o conjunta. En aquest cas, els algoritmes més rellevants serien el *Boosting*, el *Bootstrapped Aggregation* (*Bagging*), el *Weighted Average* (*Blending*) o el ja referit del *Random Forest* (annex 1, apartat 3).

Annex 2. Problemàtiques competencials

1. Abast de la competència estatal en matèria d'Administració de justícia

S'han relegat a aquest annex 2 una sèrie de reflexions sobre les probables problemàtiques competencials que poden aflorar davant d'una eventual implantació de la IA judicial per l'administració catalana. La història sobre la distribució de competències en matèria de justícia entre l'estat central, les Comunitats Autònomes i el CGPJ és llarga i complexa. S'hi barregen factors jurídics i polítics i afecta a qüestions tan importants i dispars com la fixació de la demarcació i planta judicial, els cossos de funcionariat, les normes del procediment o la dotació de mitjans. Aquesta història mereixeria, ella sola, ser l'objecte d'una recerca. Aquí ens centrarem, únicament, en com s'ha acabat definint (si és que s'ha acabat definint) l'abast de les competències autonòmiques (en aquest cas de Catalunya) en matèria de dotació de mitjans. I, en concret, de recursos i sistemes informàtics judicials. I la mesura en què l'eventual *inserció* en aquests recursos d'eines d'IA pot generar, o no, disfuncions competencials.

Aquesta problemàtica gira, sense ànim de ser exhaustius, al voltant de diversos pols competencials:

a) El nucli de l'*administració de justícia*: seria la funció estricta de jutjar i fer executar allò jutjat (art. 117.5 CE) i és una competència exclusiva de l'estat (art. 149.1.5 CE).

b) El *govern*, únic, del poder judicial, que correspon al CGPJ (art. 122.2 CE).

c) La *legislació processal*: es tracta d'una competència exclusiva de l'estat (art. 149.1.6 CE), sense perjudici de les necessàries especialitats processals que derivin de les particularitats del dret substantiu de les CCAA.

d) L'*administració de l'administració de justícia*: pot correspondre a aquelles CCAA que assumeixin en els seus EEAA competències en matèria de justícia. És a partir d'aquesta noció, no prevista a la CE i creada pel TC, que s'han generat àmbits competencials autonòmics de certa entitat en matèria de justícia. Es defineix, podríem dir, negativament, o per contraposició, respecte dels altres tres pols competencials: no pot afectar al nucli de l'administració de justícia, ni al govern del poder judicial ni a la fixació de les normes del procediment judicial (excepte el cas de les particularitats del dret substantiu propi, que és una altra problemàtica ben diferent i en la qual no s'hi entrarà).

El quadre és, certament, complicat. Explica, de fet, el dens *marc de governança* que afecta qualsevol qüestió de dotació de mitjans i recursos judicials. Hem de tenir present, però, com a criteri preliminar, que en principi la competència estatal exclusiva abastaria estrictament (o de manera necessària) *només* els tres primers apartats. Així ho assumeixen, per exemple, resolucions inicials del TC com la STC 46/90. Quedaria, aparentment, un ampli marge potencial d'actuació competencial autonòmica, ja que són moltes les competències o funcions públiques que afecten l'administració de justícia sense arribar a incidir en el seu nucli, en el govern del poder judicial o en el procediment judicial en sentit estricte. Aquest seria el camp d'actuació de l'*administració de l'administració de justícia*. El que ha succeït, però, posteriorment, és una progressiva, quirúrgica i casuística delimitació competencial per part del TC que, sense negar l'indicat camp competencial autonòmic, ha estès la competència estatal (ja sigui la legislativa o la de govern del poder judicial) fins a àmbits que van més enllà del que s'havia concebut inicialment com a nucli de l'Administració de Justícia o govern del poder judicial. I aquesta evolució ha assumit una forma molt concreta i específica en el punt que ara ens interessa, el de la dotació de mitjans materials i recursos informàtics. Vegem-ho.

2. *Administració* de l'administració de justícia

La STC 56/1990 identifica com a *administració* de l'administració de justícia un conjunt de mitjans materials i personals al servei de l'administració de justícia que, però, no s'hi integren. L'art. 104 EAC preveu la competència en matèria de mitjans materials de l'administració de justícia, que inclouria, entre altres, la configuració, implantació i el manteniment dels sistemes informàtics i de comunicació, sense perjudici de les competències de coordinació i homologació que corresponen a l'estat per garantir la compatibilitat del sistema.

Ens podem plantejar, per tant, si dins de l'apartat relatiu als sistemes informàtics, s'hi pot incloure, o no, l'eventual inserció d'eines d'IA. La primera resposta sembla òbvia: dependrà de la funció i abast que tinguin aquestes eines. Si, per exemple, es tracta d'un mer suport tècnic, extern al procedimental judicial estricte, de la tasca ordinària, *burocràtica*, del funcionariat autonòmic adscrit als jutjats, segurament no detectarem cap problemàtica competencial. En sentit invers, a mesura que l'eina d'IA es dirigís a les tasques pròpies dels LAJ o dels titulars de la funció jurisdiccional, els dubtes emergiran de manera progressiva.

Hem de partir, però, d'una premissa clau: pel cas que es plantegés la implantació d'eines d'IA vinculades a una determinada tasca estrictament *jurisdiccional* (per exemple, la generació automatitzada d'una interlocutòria declarant la incompetència territorial en un monitori o d'un esbós de sentència en un judici verbal en rebel·lia i sense judici), es tractaria, en tot cas, de meres eines *optatives*, no vinculants. I *de mer suport,* sense automatitzar *completament* la tasca, que tindria, sempre, una determinada *agència humana* posterior. Serien esborranys, projectes, de resolucions judicials. Ho hem vist amb més detall al capítol 11.

Per tant, no es produiria, en principi, una afectació directa o necessàriament condicionant en les normes de procediment en sentit estricte. En seria un complement eventual, optatiu, i de suport, sense perjudici de la necessitat, en tot cas, d'informar a les parts sobre el seu ús, als efectes de disposar de tota la informació rellevant per poder, en el seu cas, preparar en condicions els recursos corresponents contra la decisió judicial presa amb ús d'aquestes eines. El dret de defensa així ho exigeix. Seria més dubtós, certament, si, més enllà de la normativa processal, es podria produir en aquests casos

una afectació a la funció jurisdiccional, reservada a l'estat central (art. 149.1.5 CE). De nou, dependrà de la funció i les condicions d'ús de cada eina d'IA. El que sembla evident és que únicament en la mesura que es limitin a ser recursos optatius, no vinculants i de mer suport, podria plantejar-se la seva viabilitat competencial. Fins i tot en aquest cas, però, els dubtes romandrien fins que no arribés un pronunciament de fons del TC. O una normativa que establís clarament els àmbits competencials respectius.

3. Marc de governança dels sistemes informàtics

Dit això, una via per poder superar les eventuals problemàtiques competencials que generi, en un futur, la implantació *autonòmica* d'eines d'IA judicial podria ser, precisament, l'especial marc de governança previst en matèria de sistemes informàtics judicials i que ja existeix. La seva finalitat és harmonitzar la competència autonòmica amb una garantia mínima de compatibilitat dels diferents sistemes informàtics: per això es reserven a l'estat funcions de coordinació i homologació. Doncs bé, aquesta podria ser una via adequada de *validació.*

El marc és, però, de nou, complex. Fem-ne un esbós:

a) Art. 230.1 LOPJ: "*les instruccions generals o singulars d'ús de les noves tecnologies que el CGPJ o la FGE dirigeixin a jutges i magistrats o a fiscals, respectivament, determinant-ne la seva utilització, seran d'obligat compliment*".

b) Art. 230.6 LOPJ: "*els sistemes informàtics que s'utilitzin en l'Administració de Justícia hauran de ser compatibles entre si per facilitar la seva comunicació i integració, en els termes que determini el Comitè Tècnic Estatal de l'Administració de Justícia Electrònica* [CTEAJE]*. La definició i validació funcional dels programes i aplicacions s'efectuarà pel Comitè Tècnic Estatal de l'Administració de Justícia Electrònica*".

c) Art. 560.1.16ª.l) LOPJ: el CGPJ exerceix la potestat reglamentària en matèria d'establiment de les bases i estàndards de compatibilitat dels sistemes informàtics que s'utilitzin a l'administració de justícia.

d) Art. 42 Llei 18/2011, de 5 de juliol, reguladora de l'ús de les tecnologies de la informació i la comunicació en l'Administració de Justícia, relatiu a les actuacions judicials

automatitzades. Només preveu que en cas que existeixin, caldrà, abans, que el CTEAJE estableixi la definició de les especificacions, programari, manteniment, supervisió i control de qualitat i, en el seu cas, auditoria del sistema d'informació i del seu codi font. Els sistemes hauran d'incloure els indicadors de la gestió que estableixi la Comissió Nacional de l'Estadística Judicial i el mateix Comitè tècnic, cadascun en l'àmbit de les seves competències. En el seu annex defineix què hem d'entendre per actuació judicial automatitzada: aquella produïda per un sistema d'informació adequadament programat sense necessitat d'intervenció d'una persona física en cada cas singular. Inclou la producció d'actes de tràmit o resolutoris dels procediments, així com, també, mers actes de comunicació.

e) Art. 44.2 de la Llei 18/2011: "sense perjudici de les competències del CGPJ com a garant de la compatibilitat de sistemes informàtics, aquest Comitè tindrà les següents funcions:

1) Afavorir la compatibilitat i assegurar la interoperabilitat dels sistemes i aplicacions utilitzats en l'Administració de Justícia.

2) Preparar plans i programes conjunts d'actuació per a impulsar el desenvolupament de l'Administració judicial electrònica, respectant en tot cas les competències autonòmiques relatives als mitjans materials de l'Administració de Justícia.

3) Promoure la cooperació d'altres Administracions públiques amb l'Administració de Justícia per a subministrar als òrgans judicials, a través de les plataformes d'interoperabilitat establertes pel CGPJ i per les Administracions competents en matèria d'Administració de Justícia, la informació que precisin en el curs d'un procés judicial.

4. Una conclusió necessàriament provisional

En definitiva, la competència per garantir la compatibilitat dels sistemes és del CGPJ, si bé al CTEAJE hi participen les CCAA. A més, la naturalesa de la competència és de coordinació o homologació, referida sempre a l'objectiu d'assolir la necessària compatibilitat. No incideix substancialment, per tant, en la competència autonòmica relativa als mitjans materials i a la configuració i implantació dels sistemes informàtics i

de comunicació. Hi hauria, en definitiva, un aparent marge potencial per aquesta implantació i el marc adequat d'homologació i validació seria, probablement, el CTEAJE.

Aquest marge no inclou, probablement, una estricta potestat reglamentària, però aquesta no sembla imprescindible per implantar, simplement, certes eines digitals. En sentit invers, podrien ser un impediment els criteris del TC que, en matèria de gestió d'arxius judicials, acudeix com a paràmetre competencial al fet de si hi ha, o no, *afectació* a la funció jurisdiccional (STC 224/2012 (FJ 6.a)). Al mateix temps, però, no es pot equiparar completament la problemàtica de la gestió i reglamentació dels arxius judicials amb la d'implantar eines d'IA. Potser caldria acudir a les potestats més específicament vinculades als sistemes de gestió processal, si bé aquí també emergeix, de nou, la problemàtica relativa a l'eventual afectació a la potestat jurisdiccional. Aquest serà, probablement, el paràmetre clau.

En definitiva, els dubtes competencials són innegables i no seran aclarits fins que no es produeixi un pronunciament exprés del TC sobre aquestes eines, en cas de ser implantades *autonòmicament* i, acte seguit, impugnades. Mentrestant, el context adequat per abordar-les seria, com s'ha dit, el marc de governança ja existent i perfectament institucionalitzat.

En última instància, potser sí que seria *competencialment* viable, per exemple, una eina per automatitzar projectes d'interlocutòria d'incompetència territorial en monitoris i que no ho fos, per la seva major afectació al nucli de la funció jurisdiccional, una altra que, després d'analitzar tota la documentació disponible, generés esborranys de sentències en judicis verbals en rebel·lia, sense judici i basats en factures, a l'espera de la ulterior validació i examen pel titular de la funció jurisdiccional. No podem descartar que hi hagi, efectivament, un cert marge competencial fins a un nivell determinat de tasques judicials. I que, a més, aquest marge pugui ser, ell mateix, modulable o adaptable amb el temps, en funció, també, del resultat efectiu, a la realitat, de les eines d'IA que s'implantin i el grau real d'afectació que es constati que tenen en el nucli de la funció jurisdiccional. Potser seria menys intens que no ens pensem.

BIBLIOGRAFIA

A. Halevy, P. Norvig i F. Pereira (2009). *The Unreasonable Effectiveness of Data*. IEEE Intelligent Systems, vol. 24, no. 2, pp. 8-12.

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D. i Lampos, V. (2016) *Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective*. PeerJ Computer Science 2:e93 Article.

Angwin, J., Larson, J., Mattu, S. i Kirchner, L. (2016). *Machine Bias, There's software used across the country to predict future criminals. And it's biased against blacks*. Propublica.

Barona, S. (2021). Algoritmización del Derecho y de la Justicia. De la Inteligencia Artificial a la Smart Justice. Madrid. Tirant lo Blanch.

Bartoletti, I. (2020). *An Artificial Revolution, On Power, Politics and AI*. London. The Indigo Press.

Berk, R. (2017). *An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism*, J Exp Criminol **13,** 193–216 (2017). Article.

Boella, G, Di Caro, L. i Leone, V. (2019). *Semi-Automatic Knowledge Population in a Legal Document Management System*. Artif Intell Law 27, 227–251. Article.

Brown, J. (2019). *A Tour of Machine Learning Algorithms*. Mchine Learning Mastery.

Coglianese, C. i Ben, L. (2020). *AI in Adjudication and Administration*. Faculty Scholarship at Penn Law. 2118. Scholarship.

Conrad, J.G. i Branting, L.K. (2018). *Introduction to the special issue on legal text analytics*. *Artif Intell Law* 26**,** 99–102. Article.

Corona, F., Dall'Aglio, M. i Morelli, G. (2019). *The Application of fair division systems in cases involving the judicial division of assets*. Jusletter-IT_the-application-of-f_d434640630_de (DEF).

Cyberjustice Laboratory, A Tale of Cyberjustice: A Modern Approach to Technology in the Canadian Justice System (2019). Montreal, Quadriscan. Cyberjustice.

Chalkidis, I. i Kampas, D. (2019). *Deep learning in law: early adaptation and legal word embeddings trained on large corpora.* Artificial Intelligence and Law, 27(2):171–198. Enllaç Article.

Dal Pubel, L. (2018). *E-BAY dispute resolution and revolution: an investigation on a successful ODR model*. Research Gate.

Dale, R. (2018). *Law and Word Order: NLP in Legal Tech*, Natural Language Engineering, Volume 25, Issue 1, January 2019, pp. 211 – 217. Cambridge University Press. Enllaç Article.

Delgado, J. (2020). *Judicial-Tech, el proceso digital y la transformación tecnológica de la justicia. Obtención, tratamiento y protección de datos en la justicia*. La Ley, Wolters Kluwer, 2020.

Dijkstra, J. (2001). *Legal Knowledge-based Systems: The Blind Leading the Sheep?*, International Review of Computers & Technology, Vol. 15, No. 2, pp. 121-123.

Géron, A. (2020). Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow. Conceptos, herramientas y técnicas para conseguir sistemas inteligentes. Madrid. Anaya Multimedia.

Guo, Z., Zhong, H., Tu, C., Xiao, Ch., Liuy, Z., Sun, M. (2018). *Legal Judgment Prediction via Topological Learning*. D18-1390 Volume: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

Hildebrandt, M. (2019). *Data-driven prediction of judgement. Law's new mode of existence?* Draft Chapter for OUP Collected Courses Volume EUI 2019 Summerschool Mireille Hildebrandt W.

Holmes, O. (1897). *The Path of the Law.* Harvard Law Review 457.

Hutson, M. (2017). *Artificial Intelligence Prevails at Predicting Supreme Court Decisions*, SCIENCE. Science.

Kirkpatrick, B. and Klingner, B. (2004). *Turing's Imitation Game: a discussion with the benefit of hind-sight*, *discussion in the Berkeley Computer Science course, Reading the Classics*.

Larson, J., Mattu, S., Kirchner, L. i Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. Propublica.

Leibon, G., Livermore, M., Harder, R., Riddell, A. I Rockmore, D. (2018). *Bending the Law*: geometric tools for quantifying influence in the multinetwork of legal opinions. Artif Intell Law (2018) 26:145–167. Enllaç Article.

Lipton, Z. (2018). *The Mythos of Model Interpretability. In machine learning, the concept of interpretability is both important and slippery.* Enllaç Article.

Nieva, J. (2018). *Inteligencia artificial y proceso judicial*. Madrid. Marcial Pons.

Rocher, L., Hendrickx, J. M. i de Montjoye Y. (2019). *Estimating the success of re-identifications in incomplete datasets using generative models*, Nature Communications 10, No. 3069.

Romeo, F., Giacalone, M. i Dall'Aglio, M. (2018). *CREA PROJECT – Conflict resolution, equitative algorithms*. Revista Jusletter IT. Febrer de 2018, p. 251-254.

Rubiales, A. (2020). *Clustering con DBSCAN y HDBSCAN con Python y sus hiperparámetros en Sklearn*. https://rubialesalberto.medium.com/clustering-con-dbscan-y-hdbscan-con-python-y-sus-hiperpar%C3%A1metros-en-sklearn-8728283b96ac.

Sadeghian, A., Sundaram, L., Wang, D.Z. et al. (2018). Automatic *Semantic Edge Labeling over Legal Citation Graphs*. Artif Intell Law 26, 127–144 (2018). Enllaç Article.

Salmão, L. F. (2020). *Inteligência artificial, tecnologia aplicada à gestão dos conflitos no âmbito do poder judiciário brasileiro*. CIAPJ.

Slack, D., Friedler, A., Scheidegger, C., Dutta, C. (2019). *Assessing the Local Interpretability of Machine Learning Models*, Association for the Advancement of Artificial Intelligence.

Stevenson, M. i Doleac, J. (2019). *Algorithmic Risk Assessment in the Hands of Humans* 1-6, 36. Working Paper, 2019). Enllaç Article.

Taruffo, M. (1998). *Judicial Decisions and Artificial Intelligence*. Judicial Applications of Artificial Intelligence. Springer, Dordrecht. Enllaç Article.

Tran, V., Le, M. i Satoh, K. (2020). *Building Legal Case Retrieval Systems with Lexical Matching and Summarization using A Pre-Trained Phrase Scoring Model*. Enllaç Article.

Vucheva, M., Rocha, M., Renard, R. i Stasinopolous, D. (2020). *Study on the use of innovative technologies in the justice field – Final Report.* Estudi encarregat per la Comissió Europea.

Wachter, S., Mittelstadt, B. i Russel C. (2017), *Counterfactual explanations without opening the black box: automated decisions and the GDPR*. Harvard Journal of Law & Technology, 2018. Enllaç Article.

Yin, W. i R.Mok, J. (2019). *Classification of Breach of Contract Court Decision Sentences*.

Zhong, H., Xiao, Ch., Tu, C. Zhang, T., Liu, Z. i Sun M. (2018). *How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence*. Enllaç Article.