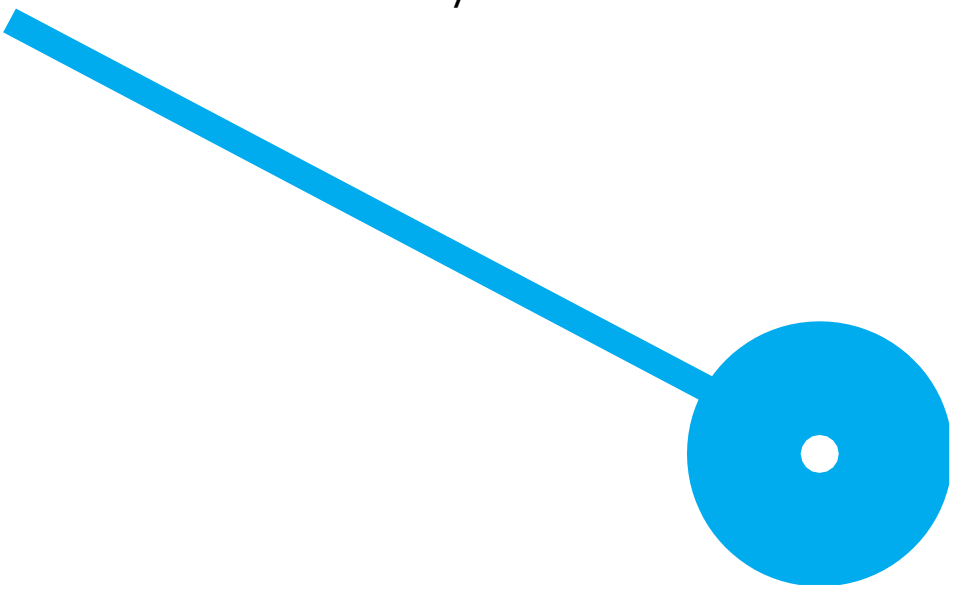
ESCOLA
SUPERIOR
DE
TECNOLOGIA E
GESTÃO -
POLITÉCNICO
DO PORTO

MESTRADO
EM PRÁTICAS JURÍDICO-DIGITAIS

Entre Algoritmos e Direitos: Os desafios da Patenteabilidade das Invenções por Inteligência Artificial.

Luis Jorge Moreira Barbarroxa

Novembro /2025



ESCOLA
SUPERIOR
DE
TECNOLOGIA E
GESTÃO -
POLITÉCNICO
DO PORTO

MESTRADO
EM PRÁTICAS JURÍDICO-DIGITAIS

Entre Algoritmos e Direitos: Os desafios da Patenteabilidade das Invenções por Inteligência Artificial.

Luis Jorge Moreira Barbarroxa
8230780

Orientadores

Professor Doutor Cláudio Renato Nunes Marques Flores

Professor Doutor João Ricardo Martins Ramos

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Práticas Jurídico-digitais pela Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto.

Novembro / 2025

Declaração de Integridade

Eu, Luis Jorge Moreira Barbarroxa, estudante n.º 8230780, do Mestrado em Práticas Jurídico- digitais da Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto, declaro que não fiz plágio nem auto-plágio, pelo que o trabalho intitulado “Entre Algoritmos e Direitos: Os desafios da Patenteabilidade das Invenções por Inteligência Artificial” é original e da minha autoria, não tendo sido usado previamente para qualquer outro fim. Mais declaro que todas as fontes usadas estão citadas, no texto e na bibliografia final, segundo as regras de referenciação adotadas na instituição.

Agradecimentos

Acima de tudo, aos meus filhos, Guilherme Delfim e Cláudia Alexandra, e ainda a toda a minha família e amigos… pela compreensão e paciência nas minhas ausências, necessárias para a realização deste projeto, no tempo que seria nosso.

Aos meus Orientadores Professor Doutor Cláudio Renato Nunes Marques Flores e Professor Doutor João Ricardo Martins Ramos, por todas as inestimáveis sugestões, pela disponibilidade, sabedoria, ajuda e incentivo.

Ao Professor Doutor Pedro Miguel Dias Venâncio e Professora Doutora Mercília Pereira Gonçalves, que durante o Mestrado também contribuíram, em revisões de matéria implicada neste estudo, com preciosas indicações.

E, ainda, a todos os restantes Docentes deste Mestrado e, outrossim, aos meus colegas de estudo, dos quais a partilha de conhecimentos e vivências concorreu nesta empreitada, sendo-lhes merecido dar esta nota.

A todos, um sincero e incomensurável obrigado.

Resumo e palavras-chave

Este estudo aborda a interação da Inteligência Artificial (IA) no processo inventivo e subsequente elegibilidade para patenteabilidade das criações daquela proveniente. A maioria dos regimes jurídicos dos diversos países e organizações internacionais não reconhece nem a pertinência nem a atribuição de autoria de uma invenção a uma IA. Dada a complexidade das atuais IAs e, ainda, a dificuldade latente na compreensão e na reprodutibilidade da invenção gerada com intervenção destas, a patenteabilidade fica comprometida (e, bem assim, todo o processo de reconhecimento e reivindicação de direitos associados). A tendência internacional, nessa vertente formal da patenteabilidade, incide sobretudo na capacidade e necessidade de se possuir o conhecimento, suficientemente claro e aceitável, do modo de funcionamento da IA. De como, e porquê, esta última, chegou a esse resultado nessa mesma invenção. O Direito alcança a sua verdadeira utilidade no presente e futuro de uma Sociedade, e esta não estagna à espera de clarificações exatas que tarde, ou, até mesmo nunca, ocorrem. Poderá a solução passar pela determinação de uma dimensão aceitável de dúvida a valorar? Equaciona-se a ponderação sobre o facto de, a existir a intervenção da IA, esta relevar na aceitação da atribuição de autoria num processo de patenteabilidade de uma invenção. Quer afastando, pura e simplesmente, a titulação da invenção nos termos legais em vigor (apontando para um regime *sui generis*); quer desconsiderando essa contribuição e outorgando a autoria ao inventor natural ou, ainda; quer atribuindo parte da autoria à (s) pessoa (s) humana (s) mantendo, contudo, uma obrigação de indicação que essa invenção teve uma intervenção relevante de IA para ser escrutinada de forma diferenciada, tendo em conta essa mesma componente.

Esta situação configura uma preocupação a nível mundial pois acaba por afetar e delongar a inovação tecnológica e, por conseguinte, fragiliza a própria Economia.

Almeja-se, com o presente estudo, contribuir para esta temática, através do levantamento das principais posições e orientações emergentes na área, culminando na apresentação das nossas próprias perspetivas, com o propósito de mitigar o referido dilema.

Palavras-chave: Propriedade Industrial; Propriedade Intelectual; Patenteabilidade; Inteligência Artificial; Invenção; Autoria; *Black-box*.

Abstract and keywords

This study deals with the interaction of Artificial Intelligence (AI) in the inventive process and the subsequent eligibility for patentability of creations arising from it. Most of the legal systems in the various countries and international organisations do not recognise either the relevance or the attribution of authorship of an invention to an AI. Given the complexity of today's AIs and the latent difficulty in understanding and reproducing the invention generated with their intervention, patentability is compromised (as well as the whole process of recognising and claiming associated rights). The international trend in this formal aspect of patentability focuses on the ability and need to have sufficiently clear and acceptable knowledge of how AI works. How, and why, the latter arrived at that result in that very invention. Law achieves its true usefulness in the present and future of a Society, and Society does not stagnate waiting for exact clarifications that are late, or even never, in coming. Could the solution be to determine an acceptable level of doubt to be valued? We are considering whether, if AI intervenes, it will be relevant in accepting the attribution of authorship in an invention's patentability process. Either by simply removing the title to the invention under the legal terms in force (pointing to a *sui generis* regime); or by disregarding this contribution and awarding authorship to the natural inventor; or by awarding part of the authorship to the human person(s) while maintaining an obligation to indicate that this invention had a relevant AI intervention in order to be scrutinised in a differentiated way, taking this same component into account.

This situation is a worldwide concern because it ends up affecting and delaying technological innovation and, consequently, weakening the economy itself.

The aim of this study is to contribute to this issue by surveying the main positions and emerging guidelines in the area, culminating in the presentation of our own perspectives, with the purpose of mitigating this dilemma.

Keywords: Industrial Property; Intellectual Property; Patentability; Artificial Intelligence; Invention; Authorship; *Black-box*.

Siglas e Abreviaturas

DL - "Deep Learning" - Aprendizagem profunda

DNN - "Deep Neural Networks" – Redes Neuronais Profundas

DPI - Direito de Propriedade Industrial

GAM – "Generalized Additive Models" - Modelos Aditivos Generalizados

IA - Inteligência Artificial

ML - "Machine Learning" - Aprendizagem de máquina

MPEP - "*Manual of Patent Examining Procedure"*

NCPI - Novo Código de Propriedade Industrial

PI - Propriedade Industrial

RL - "Reinforcement Learning" - Aprendizagem por Reforço

SL - "Supervised Learning" - Aprendizagem Supervisionada

SSL - "Semi-supervised Learning"- Aprendizagem Semi-Supervisionada

UE - União Europeia

UL - "Unsupervised Learning" - Aprendizagem Não Supervisionada

USPTO - "*US Patent and Trademark Office"*

WSL- Weakly Supervized Learning - Aprendizagem fracamente supervisionada

XAI - "Explainable AI" - IA explicável

Índice Geral

Índice de Tabelas

Introdução

A mente humana, na sua "mágica" e complexa capacidade, divagando sobre "mil e um temas", vai revelando aptidões em produzir soluções para os diversos problemas que surgem ao Homem, nomeadamente na vertente da Indústria e da Agricultura, conservando, contudo, uma visão antropocêntrica destas soluções inovadoras - destas invenções, que cria.

Com o avançar da tecnologia, com o surgimento da Era Digital - ou seja em plena atualidade, pulula exponencialmente o desenvolvimento de ferramentas cada vez mais evoluídas e com capacidades de raciocínio mais elevadas e rápidas, por vezes acima das do próprio ser humano que as inventa - surge a Inteligência Artificial (IA). No seguimento desta evolução, existe a tendência, recorrente no ser humano numa prosopopeia embriagada, de atribuir, empolgada e desenfreadamente, prerrogativas humanas a essas mesmas invenções (*v.g.* a Mitologia grega ou o reconhecimento de personalidade jurídica às Pessoas Coletivas).

No âmbito deste estudo, esse antropomorfismo também se manifesta em determinadas propostas que defendem a possibilidade de considerar a IA como autora de uma invenção. Contudo para, eventualmente, ponderar essa atribuição, exige-se o preenchimento de diversos requisitos específicos para a invenção ser patenteável. Designadamente: (1) a capacidade de se perceber, de forma clara e compreensível, como surgiu essa mesma invenção; (2) perceber, ainda, qual foi, efetivamente, o papel “criador” que a IA desempenhou na invenção em apreço (ou se esta intervenção se limitou a ser instrumental – uma ferramenta tecnológica que fora utilizada auxiliarmente à engenhosidade do autor humano), para posterior análise e reprodução na Indústria ou Agricultura. São estes imperativos próprios da patenteabilidade no âmbito do Direito de Propriedade Industrial (DPI), sobre o qual este nosso estudo incidirá.

Para a realização deste trabalho recorremos à consulta e análise de diversas perspetivas apontadas por algumas das principais referências doutrinais na área e, ainda, a distintas soluções normativas adotadas a nível internacional, no âmbito das invenções e da IA.

Ao longo deste estudo, analisamos os requisitos de patenteabilidade das invenções e examinamos o enquadramento jurídico de países e organizações relevantes no panorama da inteligência artificial (China, Estados Unidos, Japão, Portugal e União Europeia). Refletimos ainda sobre o papel da IA no processo inventivo e sobre a eventual necessidade de lhe atribuir direitos de propriedade intelectual, considerando o contexto jurídico dessas jurisdições e o panorama internacional.

1. Vertente legal

Existe o fito, internacional e consensual, em aprimorar os requisitos que espelham a necessidade da proteção do espírito criativo na área técnico-industrial e, outrossim, da preservação da contrapartida económica latente no desenvolvimento industrial que a invenção[1] possibilita. Esta proteção encontra respaldo na patenteabilidade dessas novidades técnicas designadas de invenções. Achamos, pois, existir interesse em debruçarmo-nos sobre alguns desses requisitos:

1.1 Requisito substancial – IA autora?

A nossa pesquisa centrou-se sobre os regimes jurídicos, no âmbito do Direito de Propriedade Industrial (DPI) e, mais concretamente, o Direito das Patentes[2], de diversos países e

[1]Não se encontra definição taxativa e consensual do conceito de invenção inferiremos, antes, esse conceito, atendendo a uma delimitação negativa, como se consegue depreender do art.º 51.º do Novo Código de Propriedade Industrial (NCPI) (2018). Não serão invenções: as descobertas, as teorias científicas e os métodos matemáticos; as substâncias que já existem na natureza e as matérias nucleares; as criações estéticas; os projetos, os princípios e os métodos do exercício de atividades intelectuais no domínio do jogo ou das atividades económicas, assim como as apresentações de informação; e, ainda, os programas de computadores, como tais, sem qualquer contributo técnico. Remédio Marques [(MARQUES, J.. 2007, p.232), (Dissertação de Doutoramento em Ciências Jurídico-Empresariais). Faculdade de Direito da Universidade de Coimbra; Repositório científico da UC; https://estudogeral.uc.pt/bitstream/10316/116221/1/Biotecnologia%28s%29%20e%20Propriedade%20Intelectual%20Vol%20I_2007.pdf], conforme refere Venâncio (Venâncio, P., 2023, p. 52), convida a tomar por ponto de partida um conceito aberto de invenção definido como uma "solução técnica de um problema técnico, que implica a atuação de regras técnicas (ou de efeitos técnicos) relativamente a um produto ou a uma atividade (um processo), suscetíveis de serem realizáveis ou executáveis de uma maneira constante, tantas vezes quantas as necessárias, por forma a satisfazer necessidades humanas".
Atinente aos programas de computadores, em si, sem qualquer aporte técnico ao "estado da arte", importará destacar o seu afastamento da área do DPI e indicar o encaminhar do seu enquadramento na área dos Direitos de Autor; como bem ilustra Pereira, A. (2001, p. 9 n.º 9, para 5):"[…] objecto dos direitos de autor, que incidem apenas sobre a forma de expressão desse conteúdo ideativo-funcional […]". Este autor relembra, ainda, com as seguintes palavras: "Ora, atento o regime da descompilação, estas considerações sobre a adequação dos direitos de autor para protegerem os processos codificados nos programas não deixaram de ter acolhimento na directiva comunitária, traduzindo-se na instituição de um direito sui generis de utilização das informações tecnológicas para fins diferentes da interoperabilidade." (Pereira, A. (2001), p.10 n.º 10, para 2); o respaldo concedido pela Directiva n.° 91/250/CEE do Conselho, de 14 de Maio de 1991; hodiernamente revogada pela Directiva 2009/24/CE do Parlamento Europeu e do Conselho, de 23 de Abril de 2009, relativas à protecção jurídica dos programas de computador; disponível em: https://eur-lex.europa.eu/legal-content/pt/TXT/?uri=CELEX%3A32009L0024,quanto à imperatividade de uso, para fins de interoperabilidade, das subpartes dos programas de computadores - "os processos codificados" - resultantes da descompilação destes mesmos programas. Acrescenta, outrossim, que se o fim do uso for outro a tendência penderá para a consideração de um direito *sui generis (entre o DPI e os Direitos de Autores)*.

[2]A Patente afigura-se uma titulação jurídica que confere proteção a uma invenção nova, não conhecida do público e que não seja evidente face ao que já se encontra divulgado. Confere ao titular da mesma o direito exclusivo de produzir, utilizar e comercializar uma invenção, tendo como contrapartida a sua divulgação pública. Uma invenção é uma solução técnica para um problema técnico específico. (Instituto Nacional de Propriedade Industrial (*s.d.*). Venâncio alerta, ainda, sobre a pertinência das patentes nos propósitos da Propriedade Industrial, seja o fim geral de tutela da lealdade da concorrência, seja o fim específico de tutela da inovação tecnológica e desenvolvimento económico

organizações com destaque no panorama mundial da Indústria, onde se recorra, no âmbito da criação de invenções, à utilização de sistemas de IA[3], assentando, estes últimos, em algoritmos[4] complexos. Desta análise conseguimos concluir que as doutrinas dominantes, nestes contextos, e, ainda, com as quais Portugal comunga, pressupõem que o autor de uma invenção seja uma pessoa singular (ser humano) ou uma pessoa coletiva (ficção jurídica), *i.e*., sejam sujeitos com personalidade e capacidade jurídicas; conforme resulta ilustrado no Novo Código de Propriedade Industrial (NCPI) (2018) "Art.º 3.º-Âmbito pessoal de aplicação;1 - O presente Código é aplicável a todas as pessoas, singulares ou coletivas, portuguesas ou nacionais dos países que constituem a União Internacional para a Proteção da Propriedade Industrial […] ".

Nos Estados Unidos da América (EUA), mais concretamente no "*Manual of Patent Examining Procedure* (MPEP)[5], o inventor é tido como um indivíduo ou indivíduos (ser humano/s). E ainda na vertente da interação com IA, a *US Patent and Trademark Office* (*USPTO)* emitiu, em 13 de fevereiro de 2024, um esclarecimento técnico - (*Notice)* sobre *Inventorship Guidance for AI-Assisted Inventions*, esclarecendo que um sistema de IA, por si só, não pode ser nomeado como inventor; porém, o *USPTO* reconhece que um ser humano que, trabalhando com uma IA e realizando contribuições significativas para a invenção, pode, por conseguinte, viabilizar esta última a tornar-se elegível para patenteabilidade, mantendo, contudo, o ser humano como autor. Esta dimensão de "contribuição significativa" é aferida confrontando os fatores do Teste *Pannu*, teste este que tem sido usado durante mais de 26

(Venâncio, P. 2023, p.61). De notar que Sara Peixoto comunga, a nosso ver aceitavelmente, da ideia que ousamos a denominar como uma consideração de "contribuição filantrópica e social" da Patente de Luís Manuel Couto Gonçalves, vertida no Manual de Direito Industrial, 9.ª Edição, Coimbra. Edições Almedina (2022, pp. 39 e 40), quando refere: "Atenta a atual modernização dos setores, inovação e tecnologia, podemos perspetivar a patente como uma contrapartida do investimento na criação e inovação intelectuais. Porém, pese embora não se descure os interesses do inventor, a realidade é que a atribuição de um direito de patente cuida, também, dos interesses da comunidade, isto porque se exige ao inventor a divulgação da tecnologia patenteada, de forma descritiva, facilitando-se, findo o período de proteção, o livre acesso à mesma, gerando, assim, desenvolvimento." (Peixoto, S. (2024, p.24)).

[3]Os sistemas de IA considerá-los-emos: "[…] programas de computador compostos por algoritmos computacionais extremamente complexos que requerem uma grande capacidade de processamento do *hardware*, que têm a capacidade de, pelo processamento de quantidades massivas de dados e processos de treinamento, gerar novas funcionalidades e/ou soluções para problemas não especificamente previstos […] " […] "[…] O resultado do seu processamento pode constituir uma solução técnica para um problema técnico e nessa medida matéria patenteável e, portanto, serem reconhecidos como invenção (quer o SIA quer o resultado do seu processamento)." (Venâncio, P., 2023, pp. 49 e 61).

[4]Existem diversas definições de algoritmo consoante a área de trabalho em causa, contudo, os autores: Borruso & Russo & Tiberi (*Informatica per il giurista: dal bit a internet*, p. 208), Brookshear (*Computer science: an overview*, p. 2) e, ainda, Harel,D. & Feldman,Y. (*Algorithmics: the spirit of computing*, p. 16), aludem, como relata Venâncio (Venâncio, P., 2013, p.90): "É, por isso, possível conceber uma noção geral de algoritmo, transversal a todas as ciências, como «o conjunto, sequencial, de todas as regras precisas, inequívocas, analíticas, gerais e abstractas, feitas "ex ante", cuja aplicação estrita e literal, por qualquer pessoa, coloca-o infalivelmente capaz de alcançar o resultado correcto».

[5]*Vide Appendix L Consolidated Patent Laws — January 2025 Update; Part II – Patentability of Inventions and Grant of Patents Chapter 10 - Patentability of Inventions, Sec. 35 U.S.C. 100 Definitions, (f),* (2025, p. L-18)

anos nos EUA.

Por sua vez, na China, na sua "*Patent Law of the People's Republic of China*" ("Lei de Patentes da República Popular da China"), está vincado que o pedido de Patente, como fica esclarecido no Artigo 26.º do Capítulo III, é reservado a uma pessoa que, pretendendo requerer uma patente de invenção ou modelo de utilidade, deverá apresentar os documentos pertinentes.

Ou seja, a maioria das posições, aludidas supra e que encontram reflexo na normatividade jurídica do Direito das Patentes nos seus países e organizações, reúne consensualidade, ao não reconhecer o preenchimento deste requisito, de âmbito substancial, *i.e.* ao não sufragar a pertinência na atribuição da atividade inventiva à IA *per se*.[6]

Valendo, todavia, referir, em contraponto, o único reconhecimento da autoria de uma Patente, a um sistema de IA, relativamente a um vasilhame (Caso DABUS - Thaler, S.L.), conferido pela África do Sul, conforme publicação no Patent Journal Including Trade Marks, Designs and Copyright in Cinematograph Films (2021).

Ainda na perspetiva do reconhecimento de personalidade jurídica e respetivos direitos inerentes, Barbosa, M. (2017, p.1482) afasta a possibilidade da equiparação dos sistemas de IA à personalidade jurídica das pessoas físicas (para além da evidente diferença fisiológica), aludindo que a autonomia subjacente à atividade desses sistemas é uma autonomia tecnológica fundada nas potencialidades da combinação algorítmica que é fornecida ao *software*. Denota a ausência, nas tomadas de decisões da IA, da pressuposição ética, da preocupação como o outro ser humano, premissa esta que, em muitas situações, poderá *inclusive* conflituar com a eficiência que está na base da programação computacional. A mesma autora remata dizendo que falta ao sistema de IA a dimensão espiritual e da alma; e, portanto, o reconhecimento de personalidade jurídica aos sistemas de IA poderia, eventualmente, ser equacionado pela equiparação com o disposto em relação às pessoas coletivas - uma ficção do meio jurídico. Todavia, centra, aqui também a sua relutância, que acompanhamos, na equiparação de regime e as suas ilações, no substrato humano subjacente às pessoas coletivas – fulcro desta figura jurídica. A autora quanto a isto refere que:

> " A personalidade coletiva não resulta de uma necessidade axiológica de reconhecimento, em nome da

[6] *Vide*, para além do já mencionado no NCPI português, e a título de exemplo, o caso do Japão: "*Chapter II Patents and Patent Applications;(Requirements for Patentability), Article 29 (1) A person that invents an invention with industrial applicability may obtain a patent for that invention […] "* - Lei das Patentes- Japão- *Patent Act-Japan* – (1959).

> dignidade que lhes subjaz; é atribuída em função de determinados interesses das pessoas que estão na base da sua constituição. […]
>
> […].Ora, é precisamente este fim, central para inúmeros aspetos da disciplina das pessoas coletivas, que justifica a atribuição da personalidade jurídica a estes entes. Trata-se, portanto, de uma personalidade jurídica funcionalizada à prossecução de determinados interesses humanos coletivos ou comuns ou, e dito de outro modo, de um expediente técnico que permite que os sujeitos (pessoas físicas) prossigam determinados interesses de modo diverso e mais consentâneo com a sua natureza.“ (Barbosa, M. (2017, p.1486)).

Abbott, R. (2019, para 9-11), antagonicamente, refere uma defesa de posição a favor do reconhecimento do sistema de IA em ser designado como inventor, quando se verificar que está efetivamente “criando” uma invenção, prendendo-se, não com o facto de atribuir direitos às máquinas (a IA não seria detentora de uma patente), mas antes com o facto de proteger os inventores humanos nos seus direitos. Evitaria que alguém assumisse o crédito por uma invenção que não tenha realizado, mas onde estivesse envolvida uma IA, originando a desvalorização da atividade inventiva humana em si. Alguém que somente utiliza a IA para solucionar um problema, ou parte dele, poderá ser, “injustamente”, valorado da mesma forma como aquele que inventa, de facto, alguma coisa nova. O autor continua mencionando que a legislação, em vigor, não atribui aos sistemas de IA direitos jurídicos nem morais, tais como o direito à propriedade, reconhecendo, ainda, que seria oneroso alguma alteração nesse sentido e não lhe reconhece pertinência, o que também perfilhamos. Abbott, R. (2019) acredita, ainda, que ao titular de um sistema de IA deva ser atribuída a titularidade de qualquer patente sobre as invenções geradas pela mesma, nos termos das regras de PI e da proteção de segredos comerciais.

Hilty, R. & Hoffmann, J. & Scheuerer, S. (2020) apontam à consideração da participação da IA na atividade criativa ressalvando, todavia, a presença de uma, ainda considerável, dimensão da intervenção (*Input*) humana na invenção relegando a atribuição de autoria à mesma.

Stierle M. (2021), que, alerta para o debate sobre a intervenção da IA na área das invenções, reflete sobre a hipótese de reformulação do próprio sistema de patenteabilidade das invenções envolvendo IA, referindo algumas eventuais e hipotéticas orientações, tais como:

- A abolição do sistema existente e criação de incentivos à inovação e criação;
- A criação de um sistema s*ui generis* para invenções IA, em paralelo e fora da

normatividade existente para as outras Patentes;

- A eventual necessidade de dotar a IA de personalidade jurídica;
- A declinação da responsabilidade da Sociedade relativamente às invenções "envolvendo/criadas" por sistemas de IA, colocando o foco da preocupação social no desenvolvimento e comercialização da invenção em si.

Ainda, Abbott, R. (2018), no âmbito da verificação dos diversos requisitos da invenção (que abordaremos mais adiante), nomeadamente da condicionante de novidade e, também, na constatação da obviedade (no papel de pessoa perita especializada na área), perfilha a realização dessas tarefas numa estreita comunhão entre IA e seres humanos; quiçá*, inclusive* a longo prazo, total e somente por IA, como resulta ilustrado na Tabela 1. Nesta, Abbott avança um prognóstico, no decurso do tempo, em que faz um paralelismo entre o "ente" que gera a invenção ("*Inventors*") e a capacidade desse mesmo criador ("*Skilled Standard*"). O inventor humano irá, segundo ele, sendo substituído por um "ente artificial", pela IA, devido à fulgurante e vasta capacidade evidenciada pela mesma; até se alcançar o ponto de total suplantação do ser humano pela IA superinteligente, quer seja no âmbito da geração quer na verificação dos requisitos inerentes à patenteabilidade.

Tabela 1: Evolution of Machine Invention

Phase	**Inventors**	**Skilled Standard**	**Timeframe**
I	Human	Person	Past
II	Human > SAI	Augmented Person	Present
III	Human ~ SAI	Augmented Person ~ SAI	Short Term
IV	SAI ~ AGI > Human	Augmented AGI	Medium Term
V	ASI	ASI	Long Term
SAI = Specific Artificial Intelligence; AGI = Artificial General Intelligence; ASI = Artificial Superintelligence; ~ = competing; > = outcompeting			

(Fonte: Abbott, R.. (2018, p.27). *Everything is Obvious*. UCLA - University of California, Los Angeles - Law School. Law Review n.º 2; disponível em: http://dx.doi.org/10.2139/ssrn.3056915)

1.2 Demais requisitos substanciais da invenção

Para que uma invenção possa obter a sua patenteabilidade, havemos, ainda, a realçar a consideração de distintos requisitos exigidos à mesma. Estes requisitos estão presentes na legislação dos países aludidos supra. Estas premissas, corporizam-se de forma muito semelhante, na generalidade dos mesmos, e como tal poderemos resumi-las tendo como referência o que se exige, nomeadamente, em Portugal e na UE, na conformidade prevista no NCPI (2018), conjugado com o Decreto n.º 52/91, de 30 de agosto (1991), que ratifica a Convenção de Munique sobre a Patente Europeia de 1973, a saber:

- O requisito da novidade[7] técnica - deve tratar-se de uma inovação técnica, algo que, quer como produto quer como processo, ainda não esteja já contemplada no "estado da arte", no que já se encontra devidamente patenteado ou, ademais, se encontra presentemente em uso na Indústria ou Agricultura;
- O requisito da não obviedade técnica - não deve surgir do simples decurso nos/dos inventos atualmente em uso, cujo resultado tivesse sido, naturalmente, alcançado mesmo sem a manifestação do potencial autor desse invento. A observação deste requisito é aferida por alguém com conhecimentos na área de aplicação, uma "pessoa qualificada na arte"[8], um perito na especialidade;
- O requisito da aplicabilidade industrial da invenção técnica - deve proporcionar uma

[7]Fernandes, R. (2012, pp. 81-82) refere diversidade na aceção do conceito de novidade de uma invenção, nomeadamente: "[…] são atribuídos vários sentidos à expressão "novidade" no quadro do direito de patente: novidade relativa - exige-se, nesta perspectiva, que a invenção, embora já tenha sido divulgada (por exemplo, por escrito ou na internet), não tenha sido ainda objeto de comercialização (no país ou em qualquer parte do planeta). Neste caso, trata-se de uma novidade mercadológica […], novidade absoluta - requer-se que a invenção não tenha sido divulgada, por qualquer meio, em qualquer lugar do planeta, de forma a colocar os peritos na especialidade na posição de a executar […], novidade nacional - neste enfoque, exige-se apenas que a invenção não tenha sido divulgada, por qualquer meio, no Estado para cujo território se pede proteção […], novidade planetária - neste caso, exige-se que a invenção não tenha sido divulgada, por qualquer meio, em qualquer local do planeta."

[8]Fazendo apelo à definição do "European Patent Office - EPO. (2025). *Guidelines for Examination in the European Patent Office*, Parte G, Capítulo VII (disponível em: https://www.epo.org/en/legal/guidelines-epc):

" […] 3. Pessoa qualificada na arte

Presume-se que o "técnico na área", ou "pessoa qualificada", seja um profissional qualificado na área tecnológica relevante, com conhecimento e habilidade médios (pessoa qualificada média). Esse profissional qualificado está ciente do que era conhecimento geral comum na área na data relevante […]. Presume-se também que o profissional qualificado tenha tido acesso a tudo o que há no "estado da técnica", […], e que tenha tido os meios e a capacidade para o trabalho de rotina e a experimentação, que são normais para a área tecnológica em questão. Se o problema levar o profissional qualificado a buscar sua solução em outra área técnica, o especialista nessa área é a pessoa qualificada para resolver o problema. O profissional qualificado está envolvido em constante desenvolvimento na área técnica relevante […]. Pode-se esperar que o técnico especializado procure sugestões em campos técnicos vizinhos e gerais […] ou mesmo em campos técnicos remotos, se solicitado a fazê-lo. "

solução técnica para um problema técnico [acompanhamos, o referido por Gonçalves, M. (2020, p.440):“ 4. Suscetibilidade de aplicação industrial. À luz do art.º 54.º, n.º 4 do CPI (e art.º 57.º da CPE), o invento é suscetível de aplicação industrial se o seu objeto puder ser fabricado ou utilizado em qualquer tipo de indústria ou na agricultura. A técnica intrínseca à invenção tem de ser executável na prática, isto é, tem de ser reprodutível, o resultado inventivo não pode depender da álea, do destino ou de condições não controláveis tecnicamente pelas pessoas. ”]. No caso deste requisito em específico os EUA reivindicam apenas que a invenção seja novidade e seja útil - “[…] *new and useful […]” ,* como referido no “Manual of Patent Examining Procedure (MPEP)”[9];

- O requisito da reprodutibilidade da invenção - deve ser passível de ser reproduzida, para poder ser disponibilizada, aos interessados na sua utilização. Esperando-se, esta disponibilização, sem qualquer entrave na perceção do modo de operar do sistema de IA e, bem assim, da perceção da maneira como a IA inferiu as inúmeras decisões, que tiveram que ser tomadas, para atingir os objetivos projetados na sua conclusividade funcional;
- E, ainda e sobretudo – o da atividade inventiva envolvida na sua génese *i.e.* ter havido génio, impulso e labor na vontade de alcançar algo novo e diferente.

Para que as Entidades, incumbidas de analisar esses requisitos, se consigam pronunciar rigorosamente revela-se necessário, acima de tudo, que a explicação (em todas as suas facetas - descrição, reprodutibilidade) sobre a invenção proposta, seja suficientemente clara e interpretável pelo ser humano (*in casu*, perito na especialidade). Pelo retratado, abordaremos a perspetiva adotada em diversos contextos:

1.2.1 Clarificação na UE:

A UE emitiu, em 08 de abril de 2019, um Relatório – “*Report / Study- Ethics guidelines for trustworthy AI”*, onde esta Edilidade formulou as Diretrizes Éticas para Inteligência Artificial Confiável. Estas incorreram sobre toda a envolvência dos modelos de aprendizagem dos sistemas de IA*, i.e.,* de forma muito simplista, sobre os modos como a IA assimila os conhecimentos e os interrelaciona para, de seguida, poder fornecer respostas às solicitações que lhe são colocadas.

Nesta perspetiva objetivaremos, pois, a nossa atenção para dois modos de assimilação

[9] *Vide Appendix L Consolidated Patent Laws — January 2025 Update; Part II – Patentability of Inventions and Grant of Patents Chapter 10 - Patentability of Inventions, Sec. 35 U.S.C. 101 Inventions patentable, p. L-19*.

hodiernamente em destaque, para o *Machine Learning* e para o *Deep Learning*.

Quanto ao *Machine Learning* (*ML)* - Aprendizagem de máquina - Man-Cho So, A. (2020, p. 1) enuncia-o como sendo um subcampo da IA que se foca em detetar automaticamente padrões relevantes num universo de dados e utilizar esses mesmos padrões para realizar determinadas tarefas, com o mínimo de intervenção humana. De forma genérica, recorrendo ao uso de um algoritmo, o *ML* utiliza dados, oriundos do método de treino que sofreu e que se fundam em conhecimentos ou experiências anteriores, como *Inputs* – dados de entrada, e, posteriormente, gerando *Outputs* – informações que serão, por sua vez também sujeitos a outro algoritmo, com o intuito de concretizar tarefas como tomadas de decisões ou, ainda, previsões. Este autor refere, ainda, que no *ML* a qualidade do resultado está intrinsecamente ligada ao material com que este é treinado. Os dados utilizados na sua aprendizagem, nas aplicações de maior relevância, têm uma qualidade, dimensão e complexidade avultadas, evidenciando ser desadequado o seu processamento pelo ser humano. Pelo que, é feito o recurso à capacidade da IA para a identificação desses padrões e subsequente extração de informações.

Para Man-Cho So, A. (2020, p.2-10) os métodos de aprendizagem poderão ser classificados em 3 tipos: *Supervised Learning (SL)* - Aprendizagem Supervisionada; *Unsupervised Learning (UL)* - Aprendizagem Não Supervisionada e *Reinforcement Learning (RL)* - Aprendizagem por Reforço. Ilustrando, ainda, que:

No *SL* os dados de treino contêm determinadas informações, respostas associadas, denominados de rótulos, diferentemente dos dados que serão usados nos testes a que será sujeito a IA. O propósito será da IA usar o conhecimento aprendido, com o treino com as amostras dos aludidos rótulos, para obter uma regra, um padrão, para prever as informações que não estão presentes nesses novos dados que lhe são apresentados. A precisão da regra, a qualidade do padrão é avaliado por comparação entre os rótulos previstos e os rótulos realmente presentes nos dados fornecidos para análise/teste.

No *UL* os dados utilizados no treino não têm rótulos, pretende-se, genericamente, descobrir a estrutura ocula nos mesmos. Existe a perceção que os dados criados por processos físicos não são aleatórios mas que contêm informações sobre a própria estrutura dos processos que os originaram. Como os dados não são rotulados procura-se encontrar padrões sem a presença prévia de um conjunto de treino pré-definido. Man-Cho So, A. (2020, p.10) ainda abraça o mencionado por Geoffrey Hinton [em: Pam Frost Groder, ‘Neural Networks Show New Promise for Machine Vision’ (2006) 8 (6) Computing in Science & Engineering 4.], que sublinha que o *UL* embora não possuindo um objetivo tão delineado como o *SL*, é, no entanto,

o que mais se assemelha à aprendizagem humana, pois, e acabando por citar este último: "[…] *When we're learning to see, nobody's telling us what the right answers are—we just look.*([…])".Outrossim o *UL*, por não implicar dados rotulados, mais caros de obter e assaz limitados devido aos substanciais esforços humanos inerentes às suas criações, pode ser aplicado a uma variedade mais extensa de situações.

E, por fim, Man-Cho So, debruça-se, ainda, sobre o *RL*, ressalvando que, neste, o sistema de IA aprende interagindo com o ambiente, ao longo do tempo, para atingir determinado objetivo. Esta interação condiciona o sistema de IA a tomar decisões que alteram o estado do ambiente, recebendo o *feedback* dessa interação sob a forma de recompensas ou de penalizações do próprio ambiente. O auge nesta aprendizagem é conseguir maximizar as recompensas totais. Sempre que o sistema de IA interage com o ambiente há uma nova configuração desse mesmo ambiente, pelo que poderá haver várias maneiras diferentes de definir a recompensa de uma ação. Esta aprendizagem que atende às interações entre a IA e o ambiente, resulta muitas vezes impraticável, pela falta de obtenção de dados de treino, rotulados, que concluam as ações que devam ser consideradas "corretas", e, bem assim, em qual estado do ambiente.

De ressalvar que, existe literatura que, alude, ainda, a outro tipo de aprendizagem, o *Semi-supervised Learning* (*SSL)* – Aprendizagem Semi-Supervisionada. Este é referido por Chapelle, O. & Schölkopf, B. & Zien, A. (2006, p.2), como estando num meio-termo entre o *SL* e o *UL,* pois nele o algoritmo recebe, mas não de forma geral, algumas informações de supervisão mas também dados não rotulados. Esta configuração padrão de informações somente incidirá sobre alguns alvos. Comumente, os dados sujeitos ao *SSL* são divididos em duas partes, uma parte que ostentará rótulos fornecidos e a outra não. Outra forma referida pelos autores, (proposta por Abu-Mostafa, S.. (1995). *Machines that learn from hints*. Scientific American, Vol. 272, n.º4, pp.64-69), poderá passar por estabelecer restrições tais como: "estes pontos têm (ou não têm) o mesmo alvo".

Por sua vez no *Deep Learning* [*DL*] – Aprendizagem Profunda - Goodfellow, I. & Bengio,Y. & Courville, A. (2016, p.5-8), entendem que este permite a um sistema de IA, ao computador, construir conceitos complexos a partir de outros mais simples. Existe uma aprendizagem organizada em várias etapas, em camadas, executando múltiplas instruções em sequência. Este método oferece, desta forma, maior otimização na aprendizagem pois instruções posteriores podem referir-se a resultados anteriores. O *DL* permite aos sistemas computacionais aprimorarem-se através dos dados e da experiência. Apresenta uma

hierarquia de conceitos estruturada, sendo cada conceito definido em relação a conceitos mais simples e, ainda, representações mais abstratas estruturadas com base noutras de menor abstração.

As Diretrizes comunitárias, elencadas no Relatório supracitado, refletem sete requisitos principais para que os sistemas de IA possam ser considerados confiáveis:

– Agência e supervisão humana - o ser humano deve conseguir tomar decisões informadas e conservadoras dos seus direitos fundamentais; devendo-lhe ser possibilitado o acesso a mecanismos de supervisão adequados para assegurar essa mesma garantia, nomeadamente, através de abordagens do tipo "*human-in-the-loop – (HITL)*"[10], "*human-on-the-loop – (HOTL)*"[11] e "*human-in-command – (HIC)*"[12];

– Solidez técnica e segurança - os sistemas de IA precisam ser precisos, confiáveis e reprodutíveis e, ainda, resilientes e seguros, garantindo um plano de recuperação em caso do surgimento de alguma anomalia;

– Privacidade e gestão dos dados - além da garantia de proteção da privacidade e dos dados envolvidos, também estes devem estar acessíveis para a sua melhor e adequada gestão;

– Transparência - os sistemas de IA e as suas decisões devem ser explicados de forma adaptada às partes interessadas. As pessoas têm que saber que estão a interagir com um sistema de IA e devem ser informados sobre as capacidades e limitações desse mesmo sistema;

– Diversidade, não discriminação e justiça - os sistemas de IA devem ser acessíveis a todos. Deve haver a preocupação em evitar viés injustos a fim de mitigar as possíveis e múltiplas implicações negativas, *v.g.* desde a marginalização de grupos vulneráveis até à exacerbação do preconceito e da discriminação;

– Bem-estar ambiental e social – é pretensão que os sistemas de IA possam beneficiar todos os seres humanos, tanto as gerações presentes como as futuras. Importa, para tal, garantir que sejam sustentáveis e ecologicamente corretos;

---

[10]"HITL" - abordagem no campo do desenvolvimento dos sistemas de IA em que o ser humano supervisiona e controla todo o processo de *ML*. *In* Credo AI Glossary. *Explore to learn all the must-know definitions of Responsible AI & AI Governance*; disponível em: https://www.credo.ai/glossary .

[11]"HOTL" - extensão do HITL, que envolve humanos fornecendo feedback ao sistema de IA para melhorar seu desempenho ao longo do tempo; disponível em: https://www.credo.ai/glossary.

[12]"HIC" - garantem a conformidade ética e legal, colocando os seres humanos no centro dos processos de tomada de decisão, nestes sistemas HIC, são sempre os seres humanos que tomam as decisões finais. *In Optimizing Human-AI Collaboration: A Guide to HITL, HOTL, and HIC Systems*; disponível em: https://www.deepscribe.ai/resources/optimizing-human-ai-collaboration-a-guide-to-hitl-hotl-and-hic-systems).

– Responsabilização – impera a implementação de mecanismos para garantir a responsabilidade e a prestação de contas pelos sistemas de IA e pelos seus resultados.

Deverá ser assegurado a auditabilidade destes sistemas e ainda ser garantida uma reparação adequada e acessível.

Estas Diretrizes, que propunham critérios de avaliação para determinar até que ponto um modelo de IA atenderia a esses requisitos, reverberaram na legislação que surtiu posteriormente sobre este mesmo assunto.

Adjacente a essa necessidade de perceção, não poderíamos deixar de destacar, o pioneirismo mundial, com a adoção a nível da UE, do Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024 (a "Lei da IA - *AI Act*"), onde ficaram firmemente estabelecidas regras harmonizadas sobre inteligência artificial. Adotado pelo Parlamento em março de 2024 e aprovado pelo Conselho em maio de 2024, este Regulamento, que terá total aplicabilidade 24 meses após a sua entrada em vigor, ressalva uma antecipação desta aplicabilidade para algumas das suas vertentes, denotando a prioridade da sua preocupação[13] no vertiginoso estado dinâmico de evolução da atualidade tecnológica. Nomeadamente, quanto às regras sobre sistemas de IA, de uso geral, que precisam, caso queiram operar na UE, cumprir os requisitos de transparência, devendo estar em conformidade 12 meses após a entrada em vigor deste Diploma legislativo.

Ainda, em agosto de 2024, mas desta vez respeitante ao Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016 - Regulamento Geral sobre a Proteção de Dados (RGPD), os autores Retzlaff, C. *et al.* (2024-b), destacam, do ponto de vista jurídico, o forjar de explicações acionáveis, como citado em Hacker, P. & Passoth, J. (2022), Schneeberger, D. & Stoeger, K. & Holzinger, A. (2020) e em Stoeger, K. & Schneeberger, D. & Holzinger, A. (2021), ou seja, explicações orientadas para um destinatário (mitigando a sobrecarga de informações) e, ademais, orientadas para um

[13]De notar a preocupação materializada pela Academia portuguesa em 2020, nomeadamente aquando da Consulta pública, que se ilustra para melhor enquadramento: Godinho, I. & Flores, C., & Marques, N. (2020). Consultation on the white paper on artificial intelligence: a european approach. *ULP Law Review*, *14* (1), 157-167; disponível em https://revistas.ulusofona.pt/index.php/rfdulp/issue/view/769: " II. *Contribuition* […] 7. *Considering machine learning systems in particular, it seems to us that the cumulative criteria for assessing whether an AI application should be considered high-risk, is neither adequate nor sufficient, taking into account the possibility of AI to self-adapt in order to circumvent its classification in the predefined risk categories. Hence, Independently of certification and risk activities classification, human agency and oversight is always necessary for preventing any misuse of AI*. […] 14. *In fact, beyond compliance and ex-post sanctioning (via, e.g., machine liability), criminal enforcement is also to be considered, since there is a real peril of the reorientation of AI technologies to the facilitation or commission of criminal acts (e.g., fraud schemes via Big Data). Considering such aspects is also paramount to achieving an ecosystem of trust, and the White Paper is lacking on specific orientation in this regard. On the other hand, the White Paper also lacks in orientation as to the use of AI in Law Enforcement – vis-à-vis the protection of fundamental rights of citizens and the limits of said use*."

objetivo, com uma explicação, sempre, claramente, vinculada a um objetivo. Resumindo, permitir que o destinatário compreenda uma decisão e consiga exercer os seus direitos inerentes. Retzlaff, C. *et al.* (2024-b), ressaltam o papel ativo, na implementação prática de explicações, sob a perspetiva de uma autoridade reguladora, da Autoridade britânica incumbida da proteção de dados - o "*Information Commissioner's Office- ICO"*, e, ainda, do "*Alan Turing Institute*", com a elaboração de Diretrizes que conseguissem definir vários tipos de explicações [*v.g.,* explicações justificativas (em que, de modo claro e acessível- não técnico, sejam apresentados os motivos que levaram a uma decisão da IA) ou, ainda, explicações de segurança e desempenho (em que seriam apresentadas todas as etapas, assumidas, no projeto e implementação de um sistema de IA)]. Este propósito de maximizar a precisão, confiabilidade, segurança e robustez pode ser revisitado, como citaram os autores, em *Information Commissioner's Office and The Alan Turing Institute* (2022).

Processo C-203/22 – Caso "CK GmbH contra *Magistrat der Stadt Wien*" - Direito a explicação

Outrossim e atinente ao exercício de direitos subjacentes ao conhecimento da matéria, "matéria de facto", envolvida, havemos a recordar que o RGPD foi um dos principais marcos legislativos da União Europeia, permitindo a harmonização em matéria de direito da proteção de dados nos diversos Estados-Membros. Contudo, com o fulgurante desenvolvimento tecnológico, a aplicação dos instrumentos jurídicos nem sempre se revela de fácil conciliação, como ilustra a situação, de reenvio prejudicial, do Tribunal Administrativo de Viena, que decidiu submeter à apreciação do Tribunal de Justiça da UE, dúvidas relacionadas com o direito do pleiteante a obter uma explicação, compreensível, sobre uma decisão, automatizada de um sistema de IA, que o implicou. Trata-se da pronúncia do Tribunal de Justiça da União Europeia (TJUE) – Acórdão de 27 fevereiro 2025, C-65/23; Processo C-203/22 – Caso CK GmbH contra *Magistrat der Stadt Wien*[14], e cuja parte da Decisão se transcreve, como ilustração da vincada preocupação no prevalecimento da elucidação das pessoas que têm interações marcantes com sistemas de IA*:*

" […] 66 Resulta de tudo o que precede que importa responder à primeira e à segunda questão, bem

[14]Pode ser consultada em: https://curia.europa.eu/juris/document/document.jsf?text=&docid=295841&pageIndex=0&doclang=PT&mode=req&dir=&occ=first&part=1&cid=7174952

como à terceira questão, alínea a), que o artigo 15.º, n.º 1, alínea h)[15], do RGPD deve ser interpretado no sentido de que, no caso de decisões automatizadas, incluindo a definição de perfis, na aceção do artigo 22.º, n.º 1,[16] deste regulamento, o titular dos dados pode exigir do responsável pelo tratamento, a título de «informações úteis relativas à lógica subjacente», que este lhe explique, através de informações pertinentes e de forma concisa, transparente, inteligível e de fácil acesso, o procedimento e os princípios concretamente aplicados para explorar, por via automatizada, os dados pessoais relativos a essa pessoa com vista a obter um determinado resultado, como um perfil de solvência. […] ”

1.2.2 Clarificação na OCDE

No seio da OCDE (Organização para a Cooperação e Desenvolvimento Económico) foram, em 2019, redigidos 5 princípios sobre IA, que, abraçando as Diretrizes da UE anteriormente mencionadas, foram dos primeiros padrões intergovernamentais sobre IA. Adotados em 2019 (e atualizados em 2024)[17] continuam a visar a promoção de uma IA inovadora e confiável, com respeito pelos direitos humanos e os valores democráticos, como decorre da leitura dos mesmos: 1. Crescimento inclusivo, desenvolvimento sustentável e bem-estar; 2. Direitos humanos e valores democráticos, incluindo justiça e privacidade; 3. Transparência e explicabilidade; 4. Robustez, segurança e proteção e, ainda, 5. Responsabilidade. Adotados,

[15]Nota do autor:
O artigo 15.º do RGPD, sob a epígrafe «Direito de acesso do titular dos dados», tem a seguinte redação:
“1. O titular dos dados tem o direito de obter do responsável pelo tratamento a confirmação de que os dados pessoais que lhe digam respeito são ou não objeto de tratamento e, se for esse o caso, o direito de aceder aos seus dados pessoais e às seguintes informações:
[…]
h) A existência de decisões automatizadas, incluindo a definição de perfis, referida no artigo 22.º, n.ºs 1 e 4, e, pelo menos nesses casos, informações úteis relativas à lógica subjacente, bem como a importância e as consequências previstas de tal tratamento para o titular dos dados.”Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:02016R0679-20160504 .

[16]Nota do autor:
O artigo 22.º deste regulamento, sob a epígrafe «Decisões individuais automatizadas, incluindo definição de perfis», estabelece:
“1. O titular dos dados tem o direito de não ficar sujeito a nenhuma decisão tomada exclusivamente com base no tratamento automatizado, incluindo a definição de perfis, que produza efeitos na sua esfera jurídica ou que o afete significativamente de forma similar.
2. O n.º 1 não se aplica se a decisão:
[…]
c) For baseada no consentimento explícito do titular dos dados.
3. Nos casos a que se referem o n.º 2, alíneas a) e c), o responsável pelo tratamento aplica medidas adequadas para salvaguardar os direitos e liberdades e legítimos interesses do titular dos dados, designadamente o direito de, pelo menos, obter intervenção humana por parte do responsável, manifestar o seu ponto de vista e contestar a decisão.”Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:02016R0679-20160504.

[17]*Vide* para melhor esclarecimento: OCDE (Organização para a Cooperação e Desenvolvimento Económico). *AI Principles*. (2019); disponível em: https://www.oecd.org/en/topics/ai-principles.html .

até à presente data, por todos os 38 países membros da OCDE, mas também pela UE, e ainda foram acolhidos pela Argentina, pelo Brasil, pelo Egipto, por Malta, pelo Peru, pela Roménia, pela Arábia Saudita, por Singapura, pela Ucrânia e pelo Uruguai. Estes princípios fundam-se na exigência, primacial, de transparência e divulgação responsável dos sistemas de IA para que as pessoas, afetadas pelas decisões dos mesmos, possam compreender o seu resultado e exercer os seus direitos atinentes às mesmas.

1.2.3 Clarificação no Japão

O *Japan Patent Office* (JPO) (Escritório de Patentes do Japão), em março de 2024 reforçou os seus Casos-Exemplos sobre invenções relacionadas com IA, com o intuito de disponibilizar orientações claras sobre a determinação da Etapa Inventiva, do Requisito de Descrição e da Elegibilidade. Este JPO reforça a ideia de esclarecimento essencial para a avaliação da invenção, já contemplada nas suas Diretrizes para o exame a que deverá ser sujeita a invenção. Nomeadamente no requisito de descrição (requisito de habilitação previsto na Secção 1, Capítulo 1 da Parte II das Diretrizes de Exame[18]), que evocando a Lei das Patentes [Patent Act- Act No. 121 of April 13, (1959) -Japan, last version: Act No. 42 of 2021], artigo 36.º, n.º 4, alínea i), vinca expressamente que a declaração da explicação pormenorizada da invenção deve ser clara e suficiente para permitir que uma pessoa, normalmente competente na arte da invenção (o designado de "*skilled-in-the-art*"- perito na especialidade), a possa trabalhar. Ou seja, a descrição deve ser suficientemente esclarecedora para que no caso de uma invenção de um produto - uma pessoa perita na matéria possa produzir e utilizar o produto; no caso de uma invenção de um método - uma pessoa perita na matéria possa utilizar o método e, ainda, no caso de uma invenção de um processo de fabrico de um produto - uma pessoa perita na matéria possa fabricar o produto através desse processo.

O JPO, no sentido do esclarecimento na consideração de patenteabilidade das potenciais invenções e, mais concretamente, para auxiliar os seus peritos na especialidade na averiguação dos requisitos para alcançar a proteção da Patente, sentiu a necessidade de estruturar um compêndio com casos práticos e as suas eventuais resoluções/interpretações que servirão de apoio à decisão na, eventual, atribuição da tão cobiçada titulação. Para melhor ilustração do esforço encetado pelo JPO, apresentam-se, sucintamente vertidos na

[18] *Vide* para melhor esclarecimento: Japan Patent Office. (2020). *Examination Guidelines for Patent and Utility Model in Japan*; disponível em: https://www.jpo.go.jp/e/system/laws/rule/guideline/patent/tukujitu_kijun/index.html .

Tabela 2 (podendo ser aprofundados no *link* junto à Tabela em apreço), a indicação de alguns casos de situações que ocorreram numa das etapas da "vida" e do processo de patenteabilidade de uma invenção. A título de exemplo, no caso-exemplo n.º 3, no âmbito da área da Saúde, ainda na etapa inventiva, elenca-se um caso em que a atividade inventiva é afirmada com base numa diferença no método de *ML* que estima os dados de saída a partir dos dados de entrada através de um processo de aprendizagem automática, em que uma radiografia do corpo humano é introduzida para produzir um parâmetro de ajuste do brilho dessa mesma radiografia. Segundo esta compilação, um perito na especialidade conceberia a alteração da configuração da função de perda para melhorar a exatidão da estimativa do modelo treinado, pelo que se trataria de uma mera modificação da conceção ou de uma questão de escolha da conceção. Relata-se que ao utilizar esse método de *ML* para o processamento de radiografias, na mitigação da função de perda, não se afigura um conhecimento técnico geral comumente usado no momento do registo. Ademais, permitiria suprimir a ocorrência de saturação do valor do *pixel*, proporcionando assim um efeito de aprendizagem eficaz para ajustar o brilho de uma radiografia e melhorar a sua visibilidade, o que revestiria um efeito vantajoso. Pelo ilustrado no caso-exemplo supracitado considerar-se-ia comprovada a presença de atividade inventiva.

Tabela 2: Síntese do enriquecimento adicional de casos-exemplos – *Japan Patent Office*

Caso	Requisito da patente	Nome	Observação
1	Etapa inventiva	Gerador de respostas automáticas para centros de atendimento ao cliente	Caso de sistematização simples de tarefas humanas com recurso à IA generativa (*LLM*s)[19]
2	Etapa inventiva	Método de geração de textos para a introdução em *LLMs*	Caso de caraterísticas (geração de *prompts*)[20] na aplicação da IA generativa (*LLM*s)

[19]Nota do autor:
Stryker, C. & Holdsworth, J. (2024) retratam o *Natural Language Processing* [*NLP*] – Processamento de linguagem natural, como uma ramificação da ciência da computação e da IA; que, com recurso ao *ML*, possibilita aos computadores entenderem e comunicarem com a linguagem humana. O *NLP* viabiliza que computadores e outros dispositivos digitais reconheçam, entendam e consigam produzir texto e, bem assim, comunicação verbal; combinando linguagem informática, modelos baseados em regras da linguagem humana com modelos de estatística, *ML* e *DL* .
Finn, T. & Downie, A. (2025) referem-se aos *Large Language Models* [*LLMs*] – Modelos de Linguagem em Grande Escala, como uma categoria de modelos treinados com uma vastíssima quantidade de dados, tornando-os, ainda e também, capazes de *NLP* - entender e gerar linguagem natural, *v.g.* as interfaces como o ChatGPT-4 da Open AI.
[20]Em nosso entender: *Prompts* – são instruções, perguntas ou frases dadas a um sistema de IA com o intuito de desencadear uma resposta ou ainda uma ação específica.

3	Etapa inventiva	Método de aprendizagem de modelos treinados para ajuste do brilho nas imagens de radiografias (Raio-X)	Caso de um método de aprendizagem de modelos treinados para estimar dados de saída a partir de dados de entrada
4	Etapa inventiva	Dispositivo de processamento de Raios laser	Caso de sistematização simples de tarefas humanas
5	Requisito de habilitação, requisito de suporte	Composto fluorescente	Caso de uma invenção, pela IA, de um produto que se presume ter uma determinada função (Informática dos Materiais)
6	Requisito de apoio	Método para gerar imagens para dados de treino	Caso de geração de dados de formação
7	Requisito de apoio	Aparelho de *ML* para a qualidade da fixação de parafusos	Caso em que a relação de entrada-saída entre vários tipos de dados incluídos nos dados de formação se revela clara ou não
8	Elegibilidade para patente	Dados de treino e método para gerar imagens para dados de treino	Caso dos dados de formação
9	Elegibilidade para patente	Modelo treinado para analisar a reputação das instalações	Caso de um modelo treinado configurado como um conjunto de parâmetros
10	Requisito de clareza	Modelo treinado para produzir o conteúdo do trabalho a ser realizado em resposta a uma avaria	Caso de um modelo treinado em que não é claro tratar-se de um "programa"

[Adaptado da Fonte: Japan Patent Office. (2024). *Examination Standards Office. Explanatory materials for Case Examples pertinent to AI-related technologies (Updated in March 2024). Overview of the Additional Case Enrichment*. (p.9); disponível em: **https://www.jpo.go.jp/e/system/laws/rule/guideline/patent/document/ai_jirei_e/jirei_add2024_e.pdf**].

Ainda no caminho do esclarecimento, segundo Ohno, N. *et al.* (12 de maio de 2025), em 30 de janeiro de 2025, o Tribunal Superior de Propriedade Intelectual do Japão decidiu, em apreciação ao Caso DABUS – "*Device for the Autonomous Bootstrapping of Unified Sentience*" - o sistema de IA desenvolvido pelo Dr. Stephen Thaler, (já por nós aludido no **1.1** deste trabalho), que invenções "geradas por IA" não poderiam receber proteção de patente sob a atual Lei Japonesa sobre Patentes. O Tribunal considerou que a atual Lei de Patentes apenas fornece uma estrutura para a concessão de patentes para invenções feitas por pessoas físicas, tanto em termos de direitos quanto de procedimentos**.**

1.2.4 Clarificação nos EUA

Na senda das diretrizes da UE e da OCDE, e estando plenamente alinhado com as mesmas, nos EUA, em 2022, o "*White House Office of Science and Technology Policy"* (Gabinete de Política Científica e Tecnológica da Casa Branca), propôs um projeto de Declaração de Direitos sobre IA ("*Blueprint for an AI Bill of Rights")*[21], onde identificou cinco princípios que devem orientar o *design*, o uso e a implantação de sistemas automatizados para proteger o público americano na era da inteligência artificial. Avocava a adoção de princípios acautelando: 1. Sistemas seguros e eficazes; 2. Proteções contra discriminação algorítmica; 3. Privacidade de dados; 4. Aviso e Explicação; 5. Alternativas humanas, consideração e *fallback* (previsão do recurso acessível, em caso de falha da IA).

1.2.5 Clarificação na China

Na China, segundo Ronald Y. & Kenneth Y. (2021), as Patentes são avaliadas tendo em consideração métodos de avaliação de PI baseados em métodos alternativos como indicadores de valor da invenção; tais como o custo, o rendimento ou ainda as citações (onde o número de citações que uma patente recebe indica o seu valor relativo). Ainda fazem menção da utilização de uma métrica onde o número de transações de uma licença de Patente é um indicador do seu valor, espelhando mais dados que poderiam ser usados para identificar tendências tecnológicas.

Estes autores referem a introdução, no sistema ordinário da "*China National Intellectual*

[21]*Vide* para melhor esclarecimento: White House Office of Science and Technology Policy. *Blueprint for an AI Bill of Rights, Making Automated Systems Work for the American People*; disponível em: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/ .

Property Administration" – CNIPA, de um tipo particular de patente em que os detentores de patentes informam o Conselho Estatal sobre a sua intenção de disponibilizar uma licença aberta, e que, a partir dessa comunicação, o Conselho Estatal passaria a divulgar publicamente as informações sobre as licenças abertas disponíveis. Enquanto perdurasse a figura de licença aberta, os detentores de Patentes poderiam conceder licenças, mas não únicas ou exclusivas, e em contrapartida as taxas anuais de Patentes, devidas pelo titular destas, seriam reduzidas ou isentas. Esta figura estrutural de disponibilização de licenças abertas forneceria dados reais de preços de mercado, eliminando muitas das suposições inerentes às estimativas metodológicas. Estas Licenças seriam colocadas num género de Bolsa, que estaria sob a alçada e controlo do Estado, com recurso à tecnologia de *Blockchain*, uma vez que este país possui várias Patentes neste domínio, e, ainda com implicação da Moeda Digital Nacional da China (Yuan Digital).

1.2.6 Clarificação no *Five IP Offices* - IP5

Materializando a manifesta preocupação generalizada em conseguir um entendimento, mais eficaz e eficiente, na área da patenteabilidade das invenções onde tenha havido intervenção de IA, destaca-se o nascimento de uma plataforma de cooperação/partilha internacional o "*Five IP Offices* [IP5]"[22]. IP5 é a denominação dada ao fórum, lançado em 2007, entre os cinco maiores Escritórios de Propriedade Intelectual do Mundo, criado para melhorar todo o processo de exame de patentes. Os membros do IP5 são: o Instituto Europeu de Patentes (EPO); o Escritório de Patentes do Japão (JPO); o Escritório Coreano de Propriedade Intelectual (KIPO); Administração Nacional de Propriedade Intelectual da República Popular da China (CNIPA) e o Escritório de Patentes e Marcas dos Estados Unidos (USPTO). Juntos lidam com cerca de 90% dos pedidos de patentes do mundo e 95% de todo o trabalho realizado sob o Tratado de Cooperação em Matéria de Patentes (PCT).

Esta perspetiva de cooperação do IP5 intenta alcançar um futuro sustentável, fomentando a inovação e o crescimento económico *via* um sistema de patentes inclusivo e acessível. Com a partilha e acesso a informações sobre patentes, com a harmonização de práticas e procedimentos preconiza-se a proteção de patentes objetivando conseguir um panorama internacional de patentes eficiente, económico e de utilização acessível. Este esforço é aberto ao contributo da Indústria, da Organização Mundial da Propriedade Intelectual (OMPI) e

[22] *Vide* para melhor esclarecimento: https://www.fiveipoffices.org .

conta, ainda, com o envolvimento relevante dos examinadores dos 5 Escritórios do IP5.

Numa reunião realizada em 2022, os Chefes dos Escritórios IP5 aprovaram o lançamento do projeto "*NET/AI*" relativo à "Recolha de materiais existentes nas práticas de exame dos Institutos do IP5 sobre invenções relacionadas com a IA e a sua publicação no sítio Web do IP5". O escopo deste projeto era a compilação de textos legais relevantes e recursos dos Institutos IP5, incluindo as respetivas seções de diretrizes de exames, manuais de prática, exemplos de casos, etc…; com o intuito de evidenciar os pontos-chave das práticas de exame de cada escritório, aplicáveis ao patenteamento de invenções relacionadas com IA.

A partir da reunião de 2023, os Chefes dos Escritórios IP5 endossaram uma tabela comparativa resumindo as leis, diretrizes de exame, práticas e casos de exame dos Escritórios IP5 em invenções relacionadas com IA. Em junho de 2023[23] desta análise comparativa resultou que:

> " […]
>
> - Todos os cinco institutos têm exemplos de casos relativos à elegibilidade de patentes de invenções relacionadas com a IA. (Ver Q7)
> - O EPO, o JPO e o KIPO têm exemplos de casos relativos a requisitos para descrições de invenções relacionadas com a IA. (Ver Q13)
> - O EPO tem um exemplo de caso relativo à novidade das invenções relacionadas com a IA. (Ver Q16)
> - O EPO, o JPO, o KIPO e a CNIPA têm exemplos de casos relativos à atividade inventiva das invenções relacionadas com a IA. (Ver Q19)
> - Existem algumas abordagens para lidar com as tecnologias de IA. O EPO, o JPO e a CNIPA introduziram secções especializadas ou exemplos de casos sobre invenções relacionadas com a IA nas suas orientações de exame, etc. O KIPO criou diretrizes de exame especializadas no domínio tecnológico da IA. O USPTO publicou uma página Web que compila recursos de patentes relacionados com a IA. (Ver Q2) ".

Por sua vez, mais recentemente, em junho de 2024[24] O IP5 apurou que:

> " […] Nesta altura, designar ou listar a IA como inventor não cumpre os requisitos dos institutos IP5, porque um inventor tem de ser uma pessoa natural na aceção dos sistemas jurídicos dos institutos IP5. O

[23] *Vide* para melhor esclarecimento: IP5. *Examination practices on AI-related inventions*. (June 2023); disponível em: https://link.epo.org/ip5/Chart_Examination%20practices%20on%20AI-related%20inventions.pdf#page=5 .

[24] *Vide* para melhor esclarecimento: IP5. *Inventorship of AI-generated Inventions*. (June 20, 2024);disponível em: https://link.epo.org/ip5/Inventorship_AI-related_inventions_2024 .

EPO, o JPO, o KIPO e o USPTO têm jurisprudência relativa a invenções geradas por IA. ”

Ainda nesse mês de junho de 2024, o IP5 concluiu, a nível da e, respetivamente, que [25]:

“ […]

1 . Elegibilidade

- A CNIPA tem um exemplo de caso sobre “invenções da própria tecnologia de IA”.
- O EPO, o JPO, o KIPO, a CNIPA e o USPTO têm exemplos de casos relacionados com “invenções que aplicam a tecnologia de IA a domínios técnicos específicos”.

2. Etapa inventiva

- O JPO e o KIPO têm exemplos de casos relativos à “mera aplicação de IA”.
- O EPO, o JPO e o KIPO têm exemplos de casos relativos à “modificação dos dados de treino”.
- O JPO e o KIPO têm exemplos de casos relativos à “realização de pré-processamento nos dados de treino”.
- O JPO e o KIPO têm exemplos de casos relativos a “alterar o modelo de aprendizagem”.

3 . Requisito de descrição

- O JPO e o KIPO têm exemplos de casos relativos a “se é presumível que existe uma correlação entre vários tipos de dados”.
- O JPO tem exemplos de casos relativos a "se as correlações, etc., são apoiadas por explicações e informações estatísticas fornecidas na descrição e outras”.
- O JPO tem exemplos de casos relativos a “se as correlações, etc., são apoiadas por avaliações de desempenho dos modelos de IA efetivamente criados”.
- O JPO tem exemplos de casos relativos a “invenções de produtos que a IA presume terem uma determinada função”.
- O KIPO tem exemplos de casos relativos a “se o método de pré-processamento dos dados de entrada é divulgado de forma concreta”.
- O EPO e o KIPO têm exemplos de casos relativos a “se o modelo de aprendizagem ou o método de aprendizagem é divulgado de forma concreta”.

Ou seja, toda esta nebulosidade/densidade técnica leva a uma inconclusividade de avultada

[25] *Vide* para melhor esclarecimento: IP5. *Examination practices on AI-related inventions - Comparison Table for AI cases* – (June 2024); disponível em https://link.epo.org/ip5/exam_pract_AI-related_2024 .

dimensão, precludindo no indeferimento da Patente, por falta de preenchimento dos requisitos, essenciais de patenteabilidade, como previsto nos regimes jurídicos sobre DPI, das jurisdições dos diversos países e organizações supra citados. Em suma, resulta quase unânime, no manancial jurídico do Direito das Patentes, a não atribuição da autoria da invenção a uma IA; posição que também abraçamos.

2. Vertente tecnológica

Neste capítulo importará abordar o estado de autonomia, face ao ser humano, em que se encontram as ferramentas tecnológicas que são usadas no processo de patenteabilidade de uma invenção. Debruçar-nos-emos sobre o seu grau de intervenção nesta novidade técnica e ainda sobre a manifesta necessidade de elucidação do funcionamento dos sistemas de IA na criação de uma invenção.

2.1 Tipos de IA

Importa notar, inferindo ao quadro supra [Abott, R. (2018) - Tabela 1: *Evolution of Machine Invention]*, que têm vindo a ser diferenciadas 3 tipos de IA, como, também, retratam Bahman Z. & Farahnaz B. (2023):

De um 1.º tipo, uma IA denominada de IA estreita *(Narrow AI)*, ou de IA fraca *(Weak AI)*, ou ainda de IA específica, como escreve Abbott (*SAI = Specific Artificial Intelligence*). Esta limita-se a realizar tarefas específicas, tarefas únicas, para que foram criadas. São indicadas para tratar assuntos como analisar com *Big Data,* mapas meteorológicos para especificar padrões climáticos e fazer previsões; ou analisar informações para criar um relatório político, ou ainda, como refere Nico K. (2024), para reconhecer imagens (*v.g.* no *Google Assistant*), traduzir idiomas (*v.g*. no *Google Translator*), conduzir um veículo autónomo (*v.g.* na Tesla), em assistentes virtuais (*v.g*. com a *Siri*) ou, ainda, entre outros, em ferramentas de IA Generativa como o *ChatGPT*.

De um 2.º tipo, uma IA denominada de IA Geral (*AGI = Artificial General Intelligence*), ou de IA Forte *(Strong AI)*, ou ainda de IA profunda *(Deep AI)*. Essas ferramentas poderiam interagir com o ser humano, estariam constantemente a aprender e a evoluir, tendo sido apontado como um exemplo da *AGI* a "*Sophia*", desenvolvida pela *Hansen Robotics*, denotando-se, ainda e contudo, o quão longe ainda estamos, na precocidade do estado de evolução, destes

robots inteligentes se assemelharem aos seres humanos.

Nico K. (2024), acompanha referindo que este tipo de IA teria capacidade de aprender e resolver problemas complexos, tal como os seres humanos. E, em concreto, a *AGI* visaria ensinar as próprias máquinas a fim de poderem tentar pensar e entender as coisas como os seres humanos; tentarem entender as emoções, os credos e o modo de processamento do pensamento humano; em vez de apenas imitá-los.

Sobre este mesmo assunto, Yunji C. *et al.* (2024,Cap. 1.1.1.), diz: “A IA forte, ou IA geral, exibe todos os comportamentos inteligentes de seres humanos com capacidade intelectual equivalente ou superior. Algumas IAs fracas podem facilmente superar os seres humanos em algumas tarefas, como o cálculo de adição e multiplicação, e, portanto, são amplamente adotadas. No entanto, a IA forte não visa apenas alguns problemas específicos, mas resolve todos os problemas que podem ou não ser resolvidos por um humano."

E, por fim, viria a existir um 3.º tipo de IA, uma IA denominada de Super IA (ou, *ASI* = *Artificial Superintelligence*), um estado máximo de IA, muito para além das capacidades do cérebro do ser humano. Perspetivando, ainda, como referem os autores e dando continuidade ao esclarecimento do quadro de Abbott, R. (2018) supramencionado, um exorbitante desenvolvimento da nossa inteligência, com a conexão, numa nuvem, do nosso neocórtex humano a um neocórtex sintético, afigurando-se esta fusão Homem-Máquina na denominada "Singularidade". Esta ligação, *via* nuvem, também extrapolaria com conexões com outras pessoas, possibilitando a descoberta de outros aspetos inexplorados da humanidade.

Ainda retomando a perspetiva do 1.º tipo de IA, denominada de IA estreita *(Narrow AI)*, ou de IA fraca *(Weak AI)*, ou ainda de IA específica, como escreve Abbott, R. (2018) (*SAI* = *Specific Artificial Intelligence*); Ranjan, S. & Konstantinos, I. & Manoj, K. (2025), citando Lu, Y. *et al. (*2024) e Zhang, A. *et al.* (2024), aludem ao sucesso global, em novembro de 2022, da ferramenta de IA Generativa - o *ChatGPT*; que popularizou os Agentes Generativos (estando eles associados aos sistemas de IA generativa, pois operam na mesma, achamos adequado designar estes sistemas de IA para melhor distinção, doravante, de Gen IA), que se definem como sendo, portanto, parte de sistemas baseados em *LLMs* (“*Large Language Models*”- Modelos de Linguagem em Grande Escala) criados para, com base nas solicitações do utilizador, conceber novos resultados tais como texto, imagem, áudio ou ainda código de *software*. A Gen IA teve pronta e efusiva adoção nas mais diversas aplicações; desde e citando, respetivamente Peng, S. *et al*. (2023) e Li, J. *et al.* (2019) e ainda, Jaruga-Rozdolska,

A. (2022), em assistentes de conversação, como por exemplo o "*GitHub Copilot*" e em plataformas de geração de conteúdo, como por exemplo o "*Jasper*" ou até, ainda, em ferramentas criativas, como por exemplo o "*Midjourney*". Os autores esclarecem que esta Gen IA é usualmente concebida, como também ilustra o citado por Deng, Z. *et al.* (2024), como um sistema de entidade única, em que são desenvolvidas tarefas guiadas por objetivos recorrendo a ferramentas externas, aplicando raciocínio sequencial e integrando informação em tempo real para completar funções bem definidas. Continuam narrando que, no final de 2023, o panorama tinha avançado ainda mais para o reino dos sistemas mais complexos de IA. Nesta evolução, em contraste com os sistemas de Gen IA, os sistemas de *AI Agentic* - IA Agêntica (achamos adequado designá-los para melhor distinção, doravante, de IA Ag) representam uma mudança paradigmática marcada pela colaboração multiagente, decomposição dinâmica de tarefas, memória persistente e autonomia orquestrada. Ranjan, S. & Konstantinos, I. & Manoj, K. (2025, p.2, para.7), acompanhando Acharya, D. & Kuppan, K. & Divya, B. (2025), indicam que: " *In contrast, Agentic AI systems are composed of multiple, specialized agents that coordinate, communicate, and dynamically allocate sub-tasks within a broader workflow to achieve a common goal (s*) "; destacando, ademais, que esta diferença estrutural e significativa, com a Gen IA, potencia, significativamente, a IA Ag nas vertentes da escalabilidade, adaptabilidade e no domínio de aplicação desta. Diremos, portanto, que estaremos perante um sistema de IA que manifesta alguma agência, *i.e.*, alguma capacidade de formular e perseguir os seus próprios objetivos, adaptando-se ao longo do tempo com base em experiência, raciocínio e memória.

Por suas vezes, Belcak, P. *et al.* (2025) defendem que os "*Small Language Models*" [SLMs] – Modelos de Linguagem de Pequena Escala, são suficientemente poderosos, inerentemente mais adequados e necessariamente mais económicos para muitas invocações em IA Ag, argumentando depositar elevadas expectativas e confiança no atual nível de capacidades demonstradas por estes modelos de linguagem, assim como nas arquiteturas comuns da IA Ag e, ainda, na crescente evolução da economia de implementação do *ML*. Estes avocam, ainda, que nas situações em que as capacidades de conversação de uso geral se anunciem cruciais, os sistemas de IA Ag heterogéneos (*i.e.*, com agentes que invoquem diversos modelos diferentes) revelar-se-iam a escolha mais apropriada. Ressalvam, outrossim, (embora ainda não tenham obtido uma sustentabilidade consensual no mundo científico) que esta perspetiva de alternância, mesmo que parcial, de *LLMs* para *SLMs,* impactará de modo abismal a Indústria da IA Ag. Mais alguns passos foram dados em direção ao próximo estado

de IA - denominada de IA Geral (*AGI = Artificial General Intelligence*), ou de IA Forte *(Strong AI)*, ou ainda de IA profunda *(Deep AI)*.

Adensa, à conjuntura da dificuldade na patenteabilidade, o facto de não ser muito fácil, e acessível, a compreensão e reprodução da invenção em avaliação, mesmo por pessoa perita na especialidade (como requerem os quadros legais implicados), nomeadamente devido à verificação de insuficiência descritiva da invenção – ressaltando-se, aqui, a teoria denominada de Teoria da "B*lack-box*" (caixa-negra) associada aos algoritmos usados pela IA, pelos mais diversos motivos como bem alude Burrell, J. (2016, Abstract, para.1):" *I draw a distinction between three forms of opacity: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully* […]."

Kissinger, H. & Schmidt, E. & Huttenlocher, D. (2021), alertam, referindo-se à IA, para a presença de um novo ator nas equações de tomada de decisões, diminuindo, por conseguinte, a nossa perspetiva de que somos os principais pensadores e desencadeadores de determinada situação. Quer quando criamos e controlamos a IA, quer quando apenas a utilizamos, haveremos a interagir com esta e poderemos receber respostas ou resultados que não havíamos solicitado. Se por vezes até ficaremos "encantados" com essa produção, noutras, e nomeadamente se disserem respeito a facetas importantes da nossa vida, tais como decisões, que nenhum ser humano consegue explicar, sobre oferta de emprego, decisões sobre empréstimos, decisões judiciais, entre outras; resultará uma sensação kafkiana, de exclusão. Estas tensões, entre explicações racionais e tomadas de decisões opacas, entre pessoas e Entidades, não são novas (*v.g.* certos regimes políticos mais autoritários), porém a novidade está na fonte da decisão ser "algo não humano" e, ainda, a difusão e a escala dessa nova inteligência. Existirão posições extremas de rejeição dessa ingerência nas suas vidas, contudo, a possibilidade de desconexão total será ilusória pois, quanto mais a Sociedade se está a tornar cada vez mais digitalizada, e com a IA mais integrada em serviços, produtos, Empresas, Governos, o alcance desta revela-se quase inevitável.

Na senda do esclarecimento das decisões emanadas pela IA denota-se a necessidade, internacionalmente generalizada, de abordagem da temática da "*XAI- Explainable AI*"- (IA explicável), como alude, *v.g.* Sajid, A. *et al.* (2023, p.1):" […] *the outcomes of many AI models are challenging to comprehend and trust due to their black-box nature. Usually, it is essential to understand the reasoning behind an AI model's decision-making. Thus, the need for eXplainable AI (XAI) methods for improving trust in AI models has arisen*."

Estes autores continuam, ainda, por evidenciar a complexidade técnica envolvida no intento da explicabilidade da IA, devido ao carácter dos dados, atualmente implicados, que são, também eles, de elevada complexidade e, ainda, de carácter "não-linear", tornando a tarefa de processamento para obtenção de esclarecimentos numa empreitada árdua.

Todavia, e para tal propósito, são usadas Redes Neuronais Profundas (*Deep Neural Networks - DNN*), no intuito de conseguir extrair as informações dos conjuntos de dados altamente complexos. Holdsworth, J. & Scapicchio, M. (2024), por suas vezes, mencionam o *DL* como um subcampo do *ML*, que utiliza redes multicamadas, as supramencionadas *DNN, as quais* são sistemas de computação, baseados em algoritmos, que funcionam de forma semelhante, em complexidade e na organização com interligações por nós, aos neurónios do cérebro humano, conseguindo gerir imensas correlações de dados e aprender com estas interações. Nos modelos tradicionais, "não profundos", de *ML* utilizam-se redes neuronais simples com uma ou duas camadas computacionais. Diferentemente, nos modelos de *DL* podem utilizar-se desde três ou mais camadas, sendo comum utilizarem-se centenas ou milhares de camadas para treinar os modelos.

Estas atividades chegam a envolver vários milhões de parâmetros, tornando as representações e o fluxo de dados, emergentes, de elevada dificuldade em serem examinados, ao mesmo tempo que as variáveis, que podem ser aprendidas, aumentam exponencialmente com o aumento da complexidade dos *designs* das redes que surgem.

Estes *designs* são influenciados por uma miríade de fatores, incluindo a função de ativação, tipo e tamanho de entrada, número de camadas, operação de agrupamento, padrão de conectividade, mecanismos de classificação e os resultados de técnicas de aprendizagem compostas. Estas últimas, por sua vez, são, também elas, ainda influenciadas por uma série de funções adicionais, como a normalização/regularização, mecanismos de atualização de peso, funções de custo/perda e o tipo de classificador final usado. Resultando, em suma e conforme estes autores acabariam por ilustrar, que uma decisão de uma *DNN* é deveras difícil de compreender e confiar. (Sajid, A. *et al.*, idem, p.2)

2.2 O caminho da explicabilidade, interpretabilidade, causalidade

Na variada literatura técnica, dirigida ao assunto da tentativa de trazer esclarecimentos sobre o funcionamento dos sistemas de IA, deparamo-nos com a ambiguidade na consensualização de conceitos, nomeadamente entre explicabilidade, interpretabilidade, ou, ainda, causalidade. Iremos, de seguida, modestamente percorrer o tempo e alguns autores para apresentarmos

uma visão da evolução e das perspetivas desses conceitos e, bem assim, estabelecer referências estruturais nesse âmbito.

2.2.1 Pertinência da explicação, da interpretabilidade

Doshi-Velez, F. & Kim, B. (2017, pp.2-4) referem a interpretabilidade como: " a capacidade de explicar ou apresentar em termos compreensíveis para um ser humano"; posição esta que acompanhamos. Estes também aludem a uma definição interessante, embora do campo da psicologia mas que captou a nossa atenção, ao citarem, Lombrozo, T. (2006) que defende que "[…] explicações são [...] a moeda na qual trocamos crenças […]". Estes autores relembram-nos que na Sociedade, nos sistemas com envolvência de IA, com o seu inerente *ML*, nem sempre exigimos a interpretabilidade desses mesmos sistemas, inclusivamente naqueles em que a determinação dos seus resultados não envolve intervenção humana, *v.g.,* os servidores de anúncios, na classificação de códigos postais, ou, ainda, nos sistemas de prevenção de colisões de aeronaves. Relegamos a explicação quer por não se augurar consequências significativas para os resultados inaceitáveis - mas que surtiram, ou quer, pelo facto do problema, ainda que a solução tenha brotado de um sistema com algumas imperfeições, já tenha sido tão amplamente estudado e validado, em aplicações reais, que acabamos por confiar na decisão desse mesmo sistema. Doshi-Velez, F. & Kim, B. (2017), reiteram que a necessidade da interpretabilidade, usualmente, só surge quando existe uma incompletude na formalização dos problemas, originando uma barreira na otimização e avaliação de todo o processo. A incompletude produzirá algum tipo de viés que poderá não ser quantificado, *v.g.* aquando da inclusão do conhecimento do domínio numa seleção de determinado *ML*. As explicações expõem ao ser humano os efeitos das lacunas na formalização dos problemas.

2.2.2 *Trade-off* interpretabilidade *Vs* completude

Por suas vezes, Gilpin, L. *et al.* (2019) retomam que, sendo o objetivo da interpretabilidade descrever os componentes internos de um sistema de uma forma que seja compreensível para os humanos, a explicação, de um sistema tão complexo, não deverá ser tão simplificada assim ao ponto de não poder ser compreendida pelos utilizadores e/ou, pior ainda, de ter sido convenientemente otimizada para ocultar atributos indesejáveis do sistema. As explicações deverão evidenciar um *trade-off* entre interpretabilidade e completude, devendo possibilitar

descrições com maior detalhe e completude com uma possível afetação na interpretabilidade. Gilpin, L. *et al.* (2019), mencionam, ainda, que uma abordagem para a interpretabilidade seria a da criação de sistemas produtores de explicações com arquiteturas projetadas para simplificar a interpretação do seu próprio comportamento. Estas arquiteturas facilitariam a compreensão*, i.e.,* promoveriam uma explicação sobre o processamento (projetando uma resposta a como um *Input* específico levou a este *Output* específico), sobre as representações (projetando uma resposta a quais as informações que a rede encerra) ou, bem assim, de outros aspetos do funcionamento do sistema (definição esta, de explicabilidade, que também perfilhamos).

Para estes autores a avaliação da completude deveria ser de âmbito local, por muito complexo que fosse o sistema de IA, como no caso das *DNN*, deveria ser explicado de forma que fizesse sentido localmente. Estas explicações, do modelo de processamento, tenderiam a reduzir a sua ”complexidade”, fruto da redução da sua extensão, e ainda, pela redução do erro da representação interpretável em relação ao classificador real, por se encontrar mais próximo à instância a ser explicada - alcançar-se-ia uma maior “completude local”. Com a caracterização de partes individuais de uma representação estas podem ser submetidas a apreciação quanto ao seu poder explicativo, avaliando-se se as suas ativações revelam fielmente a existência de um viés específico nessa rede.

Na avaliação de sistemas que produzem explicações pretender-se-á saber a correspondência às expectativas do utilizador. Por exemplo, como aludem Gilpin, L. *et al*. (2019) ao referir Das, A. *et al.* (2017), a atenção em rede pode ser comparada à atenção humana, testando-se representações extraídas/desenleadas com conjuntos de dados sintéticos, onde existem variáveis latentes, e com isso averigua-se se essas mesmas variáveis são denotadas. Gilpin, L. *et al.* (2019) alertam que é comumente percecionado, na comunidade científica, que o nível de interpretabilidade e compreensão teórica necessários para explicações transparentes das *DNN* é tarefa árdua de alcançar; salientando a seguinte dificuldade:

> “[…] since the system itself produces the explanation, evaluations necessarily couple evaluation of the system along with evaluation of the explanation. An explanation that seems unreasonable could indicate either a failure of the system to process information in a reasonable way, or it could indicate the failure of the explanation generator to create a reasonable description. Conversely, an explanation system that is not faithful to the decision making process could produce a reasonable description even if the underlying system is using unreasonable rules to make the decision. An evaluation of explanations based on their

reasonableness alone can miss these distinctions. In [74][26], a number of user-study designs are outlined that can help bridge the gap between the model and the user." (Gilpin, L. *et al.* (2019, p.8, para 7))

2.2.3 Desideratos da interpretabilidade

Carvalho, D. *et al.* (2019), apontam a relevância da interpretabilidade dos sistemas de IA para a exigência da Sociedade e, em sintonia com Doshi-Velez, F. & Kim, B. (2017), ressaltam os desideratos que podem ser alavancados por essa interpretabilidade, a saber:

- Imparcialidade - Aprimorando a garantia da imparcialidade das previsões e a não discriminação, implícita ou explícita, de determinados grupos de pessoas. Ajuda a avaliar se a decisão teve influência de um viés demográfico aprendido, *v.g.* racial;
- Privacidade - Acautelando que as informações confidenciais nos dados estão protegidas;
- Confiabilidade/Robustez - Garantindo que pequenas alterações nos *Inputs* não causem avultadas mudanças na previsão;
- Causalidade - Defendendo que apenas relações causais sejam detetadas.
- Os autores consideram a causalidade como a medida do mapeamento da explicabilidade, (explicação técnica que explique os resultados dos sistemas de IA – conceito de causalidade que acolhe a nossa adesão) com a compreensão humana.
- Confiança – Ampliando esse sentimento humano nos sistemas de IA que expliquem as suas decisões em detrimento dos que apenas as exibem (*Black-box*).

Carvalho, D. *et al.* (2019), acompanhando Kim, B. & Doshi-Velez, F. (2018), agrupam os métodos de interpretabilidade, em consonância com o momento em que estes métodos são aplicáveis, nomeadamente em: antes [(pré-modelo)]; durante [(no modelo) - intrínseco ou ainda por *design*, como classifica Molnar, C. (2019)] ou depois [(pós-modelo) - *Post-hoc*] da construção do modelo de *ML*.

As técnicas de interpretabilidade pré-modelo são independentes do modelo em si; elas visam apenas aos dados em si. Estando, intimamente relacionadas com a interpretabilidade dos dados, configuram-se em técnicas de análise exploratória de dados. São exemplos, do esforço que tem vindo a suceder na criação dessas técnicas, as ferramentas desenvolvidas

[26]De referir que neste n.º [74] Gilpin, L. *et al.* (2019,p.8) faz alusão a Doshi-Velez, F. & Kim, B. (2017).

pelo grupo de Pesquisa em Pessoas e IA (PAIR) do Google: "*Facets Overview"* e "*Facets Dive"* [27].

Quando estamos perante o desenvolvimento da atividade do modelo, referem os autores e mencionando Molnar, C. (2019), evidenciar-se-ão critérios intrínsecos, se a interpretabilidade for alcançada com recurso a restrições impostas à complexidade do modelo de *ML*, ou ainda, critérios de *Post-hoc,* se esta se alcançar pela aplicação de métodos que analisem o modelo após a aprendizagem.

Carvalho, D. *et al.* (2019) aproveitam, ainda, para, citando Rudin, C. (2019), e no âmbito do alcance da interpretabilidade intrínseca, por meio da imposição de restrições ao modelo, referir que estas restrições podem ser, *v.g.* a esparsidade, a monotonicidade, a causalidade ou, outrossim, as restrições físicas que vêm do conhecimento do domínio.

Estes autores, em alusão à argumentação de Lipton, Z. (2018), referem que a interpretabilidade intrínseca também é frequentemente apelidada de transparência e responde à questão: *como funciona o modelo?* e, ainda, por sua vez, a interpretabilidade *Post-hoc* responde à pergunta: *o que mais pode nos dizer o modelo*?

Somos, ainda, alertados pelos autores para a existência de métodos *Post-hoc* que podem ser aplicados a modelos intrinsecamente interpretáveis, pois estes métodos *Post-hoc* são, geralmente, desacoplados do modelo principal.

Os métodos de interpretação podem ser agrupados, segundo os autores, tendo em conta o critério da amplitude que visam alcançar, se visarem um tipo de modelo em específico – específico do modelo, ou a generalidade dos modelos de *ML*- agnóstico de modelo. Conforme mencionado por Carvalho, D. *et al.* (2019), em sintonia com Molnar, C. (2019), nos métodos específico do modelo são analisadas partes do modelo para melhor compreendê-lo. Nos agnósticos de modelo, que são aplicados após o modelo ter sido treinado - *Post-hoc*, ignoramos o que está dentro do modelo e analisamos, apenas, como a saída do modelo sofre alterações devido às mudanças nas entradas dos recursos. Pela sua própria essência, esses métodos não poderão ter acesso ao funcionamento interno do modelo, tal como a pesos ou informações estruturais, pois não se verificaria o desacoplamento, do modelo de *Black-box*, característico destes modelos. Uma outra particularidade reside nos modelos serem interpretados sem sacrificar o seu poder preditivo, pois são aplicados após a aprendizagem - *Post-hoc*, como bem lembram os autores citando Lipton, Z. (2018). Carvalho, D. *et al.* (2019),esclarecem ainda que, embora existam alguns métodos específicos do modelo que

[27]Para melhor ilustração *Vide*: Google People + AI Research (PAIR). *Facets - Visualization for ML Datasets*; disponível em: https://pair-code.github.io/facets/ .

são *Post-hoc*, a maior parte da interpretabilidade específica do modelo é alcançada por *design*, *i.e.*, por meio de modelos intrinsecamente interpretáveis. Molnar, C. (2019) ainda subdivide a classificação dos métodos agnósticos de modelo em métodos locais, que se debruçam sobre explicar previsões individuais, e métodos globais, que se concentram em conjuntos de dados.

2.2.4 Sistemas intrinsecamente interpretáveis

Holzinger, A. *et al*. (2019-a), referindo-se a Holzinger *et al.* (2017) e Holzinger *et al.* (2018), ao realizarem uma abordagem geral de modelos de IA, também retomam a perspetiva de dualidade na classificação dos métodos e modelos de *ML*. Por um lado, em sistemas *Post-hoc* caracterizando-os, contudo, com maior enfoque no fornecimento de explicações locais, para determinada decisão e possibilitar a sua reprodução sob demanda, não se debruçando na explicação de todo o comportamento do sistema de *ML*. E, por outro lado, em sistemas intrínsecos, interpretáveis pelo *design*, que se materializam naquilo que denomina de "*Glass-boxes*" (Caixas de vidro),que se perspetivam possuir a inerente, e desejada, transparência na tomada de decisões; como *v.g.* as árvores de decisão ou ainda os sistemas de inferência *Fuzzy* [que, como salientam os autores e citando Guilhaume, S. (2001), tão já sobejamente utilizados, nas interações ser humano e IA, conseguindo uns bons alicerces para interações entre conhecimento especializado e conhecimento oculto nos dados].

Holzinger, A. *et al.* (2019-a), ainda aludem ao exemplo apresentado, da área da Medicina, por Caruana *et al.* (2015), dos "*GAMs – Generalized Additive Models*"[28] (Modelos Aditivos Generalizados), que evidenciaram alto desempenho com interações em pares e conseguiram produzir modelos inteligíveis, que, por suas vezes, descobriram padrões surpreendentes nos dados que anteriormente impediam que modelos complexos, aprendidos, fossem colocados

[28]Estes modelos remontam a Agosto de 1986, sendo introduzidos por Trevor Hastie e Robert Tibshirani. Como refere de forma mais simplista, Shafi, A. (2021) os *GAMs* são uma adaptação de modelos alusivos a combinações lineares, os designados "*GLMs – Generalized Linear Models"* ("Modelos Lineares Generalizados"), em que são flexibilizadas as restrições inerentes aos modelos de regressão lineares iniciais (modelos que têm o escopo de conseguirem a previsão do objetivo/alvo como uma soma ponderada das entradas de características, em modo de "engenharia inversa" às combinações lineares). A regressão linear, por sua vez, nem sempre irá representar o que vemos da realidade. Daí a emergência desta flexibilidade que é conseguida através de funções, na sua maioria polinomiais que cobrem um pequeno intervalo, denominadas de "*spline*". São funções complexas que possibilitam a modulação de relações não lineares para cada uma das características do modelo. Impera o uso de diversas "*splines*" que acabam por formar o *GAM*; resultando num modelo ultra flexível mas que, ainda assim, mantém grande parte da maior explicabilidade de uma regressão linear.

em campo neste domínio. Demonstraram, ainda, possuírem uma elevada capacidade de escalabilidade potenciando a sua aplicação a grandes conjuntos de dados.

Holzinger, A. *et al.* (2019-a), baseando-se em Pearl, J. (2009), Pearl, J. & Mackenzie, D. (2018) e, ainda, em Peters, J. & Janzing, D. & Schölkopf, B. (2017), apontam para a utilização de modelos causais estruturais, pois, para equiparar a inteligência humana, os sistemas de IA precisam da orientação de um modelo de realidade, semelhante aos usados em tarefas de inferência causal. Os sistemas, que até então eram usados, trabalhavam em modo estatístico ou livre de modelo, com sérias limitações na sua eficácia e desempenho. Estes sistemas não raciocinavam sobre intervenções e retrospeção e por conseguinte não se revelavam adequados.

***"Weakly Supervized Learning (WSL)"* - Aprendizagem fracamente supervisionada**

Ainda no campo, sensível, da Medicina, Holzinger, A. *et al.* (2019-a) destacam a utilização de sistemas de IA que estiveram sujeitos a uma Aprendizagem fracamente supervisionada ["*Weakly Supervized Learning (WSL)*]. Um termo genérico que abarca vários métodos de construção de modelos preditivos através de uma aprendizagem fraca resultante de supervisão incompleta; de supervisão inexata ou ainda de supervisão imprecisa. Isto é, como elucidam os autores:

- Citando Komura, D. & Ishikawa, S. (2018, para 4.2.2), evidenciará ter tido uma aprendizagem fraca se decorrer de uma supervisão incompleta, tal como na incorporação de rótulos insuficientes - fracos, concretamente imagens fracamente supervisionadas, que se utilizam no diagnóstico de determinado tipo de cancro, mais fáceis de obter relativamente à rotulação firmada. Embora estas imagens não acusem a posição exata, da região de interesse, delimitadora do tecido canceroso, é possível extrair, submetendo as imagens a um diagnóstico patológico, informações sobre a presença/ausência deste infortúnio.
- Ainda, na alusão a Komura & Ishikawa (2018, para 3.2), será fruto de uma supervisão inexata, tal como, por exemplo, na *Content Based Image Retrieval (CBI*R) - Recuperação de imagens baseada em conteúdo, que opera na recuperação de imagens semelhantes a imagens de consultas. Estes sistemas de recuperação de imagens, no âmbito da patologia digital, tornaram-se de reconhecida utilidade em

situações de diagnóstico, de pesquisa ou, ainda, para fins educacionais, nomeadamente, possibilitando a estudantes e patologistas iniciantes a oportunidade de recuperar casos relevantes ou imagens histopatológicas de tecidos. Além disso, esses sistemas também são úteis para patologistas profissionais, mormente no diagnóstico de casos raros. A principal vantagem reside, contudo, na *CBIR* não requerer, necessariamente, informações com rótulos, dando, por conseguinte espaço para a utilização da *WSL*.

- Por último, e mencionando a citação, feita pelos autores, ao trabalho de Xu, Y. et al. (2014, Introdução, para. 5), poder-se-á concluir estarmos perante uma supervisão fraca porque imprecisa pois, sendo o propósito de diagnosticar casos de cancro, e recorrendo à análise automática de imagens histopatológicas fracamente supervisionadas, revelou-se muito mais fácil a rotulação desse mesmo tipo de imagens do que a delineação, detalhada, de regiões de interesse em cada uma das imagens.

Os autores, mantendo o acompanhamento a Xu, Y. *et al*. (2014) e ainda a Komura, D. & Ishikawa, S. (2018), mencionam que a *SL* é muito dispendiosa pois é deveras trabalhoso obter informações de supervisão sólidas e rótulos totalmente verdadeiros. Nos estudos em apreço, nos casos de diagnóstico de cancro, a rotulagem de imagens histopatológicas não é somente uma tarefa demorada como também altamente crítica, nomeadamente, na importância de segmentação dos tecidos cancerígenos e do seu devido agrupamento em várias classes. Apontam alguns entraves ao *DL*, tais como: o facto do tamanho das imagens digitais, geralmente, ser muito grande; as imagens estarem rotuladas de forma insuficiente (existindo poucos dados de aprendizagem disponíveis); o tempo despendido pelo patologista (rotulagem cara), entre outros. Daí o surgimento de utilização de sistemas que, com certas características da *SSL*, utilizam uma *WSL* e têm obtido resultados muito satisfatórios; como é possível deduzir pela citação realizada de Xu, Y. *et al.* (2014, p.592, para 1-2):

> ”In the middle of the spectrum is the weakly supervised learning scenario. The idea is to use coarsely-grained annotations to aid automatic exploration of fine-grained information. The weakly supervised learning direction is closely related to semi-supervised learning in machine learning […]. One particular form of weakly supervised learning is multiple instance learning (MIL) […].
>
> In this paper, we develop an integrated framework to classify histopathology images as having cancerous regions or not, segment cancer tissues from a cancer image, and cluster them into different types. This system automatically learns the models from weakly supervised histopathology images using multiple

clustered instance learning (MCIL), derived from MIL. Many previous MIL based approaches have achieved encouraging results in the medical domain such as major adverse cardiac event (MACE) prediction […]".

2.2.5 Importância do conhecimento do domínio

Roscher, R. *et al.* (2020), quando aludem ao defendido por Miller, T. (2019), sugerem uma abordagem ao conceito de explicabilidade, tomando por base a articulação entre os *insights* das ciências sociais e as explicações destinadas a esclarecer a atuação de um sistema de IA. Estes acolhem a estruturação, em três classes, de determinadas questões com pretensões explicativas, a saber: 1.ª- O quê? 2.ª- Como? 3.ª- Porquê? no intuito de trazer contexto e entendimento às dúvidas suscitadas. Os autores ainda comungam da definição de explicação, citando Montavon, G. & Samek, W. & Müller, K. (2018, p.2): " *Definition 2. An explanation is the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression)*."; acrescentando acreditarem que um conjunto de interpretações só constituirá uma explicação se existirem informações contextuais adicionais, provenientes do conhecimento do domínio e relacionadas com o objetivo da análise; posição, esta, que também seguimos. Os autores vincam a importância do conhecimento do domínio como uma parte essencial da explicabilidade, não se afigurando alcançável, unicamente e exclusivamente através de algoritmos, a desejada explicação da tomada de decisões pela IA.

2.2.6 Domínio e restrições práticas:

Sudjianto, A. & Zhang, A. (2021) demonstram, apoiando-se em Ribeiro, M. & Singh, S. & Guestrin, C. (2016) e em Lundberg, S. & Lee, S. (2017), ceticismo em relação aos métodos, *Post-hoc*, agnósticos de modelo, tais como, por exemplo e respetivamente, o modelo interpretável localmente "*LIME"* ("*Local Interpretable Model-Agnostic Explanations"*) e, ainda, o modelo "*SHAP*" ("*SHapley Additive exPlanations*") com explicações aditivas. Segundo os autores, este tipo de modelos só fornecem explicações aproximadas e, bem assim, apresentam potenciais esparrelas, tais como: a deficiente generalização de modelo, as características dependentes, as interações de características ou interpretações causais injustificadas; entre tantas outras elencadas no trabalho citado de Molnar, C. *et. al* .(2020).

Sudjianto, A. & Zhang, A. (2021) mencionam que as suas posições críticas em relação ao uso cego de métodos de explicação agnósticos de modelo são partilhadas por vários outros autores da Academia, de entre os quais os autores citaram: Rudin, C. (2019); Slack, D. *et al.* (2020) ou ainda Kumar, E. *et al.* (2020).

Sudjianto, A. & Zhang, A. (2021) são defensores de *ML* inerentemente interpretável, onde a interpretabilidade inerente significa que o modelo deva ser transparente e autoexplicativo; a interpretabilidade é-lhe intrínseca. Seguem, citando-o, o argumento de Yang, Z. & Zhang, A. & Sudjianto, A. (2021) de que a interpretabilidade inerente de um modelo complexo deve ser induzida por restrições práticas, também já anteriormente denotadas por Carvalho, D. *et al.* (2019) (*v.g.* a aditividade, a esparsidade, a linearidade, a suavidade, a monotonicidade e a quase ortogonalidade, entre demais), que resultarão na tendência a tornar o modelo assaz mais interpretável pelo ser humano. Estas restrições, por suas vezes, também, fornecerão princípios de *design* para o desenvolvimento de modelos interpretáveis de *ML*.

Os autores retomam a posição atinente à compatibilização entre a interpretação do modelo e o conhecimento especializado no domínio da aplicação. Acrescentam a importância da participação de especialistas (aqui somos a levantar a questão, que tipo de especialistas? Seres humanos; IA ou uma simbiose entre ambos?) nos domínios, formados com base na experiência e conhecimento prévios, os quais estarão aptos a fornecer conselhos valiosos para a construção de modelos de *ML* inerentemente interpretáveis.

2.2.7 Explicações causais contextualizadas no domínio. Riscos da explicabilidade

Vilone, G. & Longo, L. (2021), apoiam a ideia, quando referem Taehyun, H. & Sangwon, L. & Sangyeon, K. (2018), de que a atribuição causal a eventos está ligado à natureza do ser humano; notando que um sistema de IA que providencia explicações causais no seu processo inferencial é visto como mais “humano”, pois os utilizadores finais adotam a tendência inata da psicologia humana ao antropomorfismo. Estas explicações devem tornar explícitas as relações causais entre os *Inputs* e as previsões do modelo, sobretudo se estas relações não forem intuitivas para os utilizadores. Os autores, citando Lipton, Z. (2018), mencionam que o apuramento de relações causais está dependente do conhecimento prévio, não descartando a ocorrência de algumas associações completamente inesperadas. Reconhecendo não ser tarefa fácil de conseguir atingir a transparência algorítmica em redes neuronais devido à incapacidade atual dos especialistas de entender o processo inferencial desses modelos,

salientam as tentativas de alguns estudiosos, em ultrapassar esse obstáculo, em encontrar métodos para rastrear as previsões de um modelo até às características mais influentes da entrada. Havendo os autores citado, como exemplo desta necessária perseverança, Simonyan, K. & Vedaldi, A. & Zisserman, A. (2014), com a criação de mapas de calor pela retro propagação das previsões de um modelo para o espaço de entrada e ainda originando o destacamento de *pixels* relevantes; ou, ainda, Lou, Y. & Caruana, R. & Gehrke, J. (2012), que, indicaram a substituição dos modelos com *Black-boxes* por Modelos Aditivos Generalizados - *GAMs*, a fim de atender à satisfação das propriedades de simulação e de decomponibilidade de alguns modelos de *ML*.

Contudo e apoiando-se na citação que fazem de Weller, A. (2017), Vilone, G. & Longo, L. (2021), alertam para a ânsia de alcançar a transparência, ressaltando certos perigos latentes, *v.g.*, de exposição de dados pessoais ou até mesmo dados sensíveis; ou, de direcionar um criador a produzir um modelo com características específicas, deteriorando o seu desempenho e grau de generalização; ou, ainda, do refrear da criação de propriedade intelectual, pela exigência imoderada de tornar totalmente visíveis, para o utilizador, os dados e o modelo usados, entre outros; e podendo, inclusivamente até, desacelerar, de alguma maneira, o desenvolvimento de novas tecnologias.

Vilone, G. & Longo, L. (2021), voltam a frisar que as explicações são eficazes quando ajudam os utilizadores finais a terem uma representação mental completa e correta do processo inferencial de um dado modelo. Salientam, contudo, que entre os académicos não existe consenso sobre explicação e, ainda, tão pouco sobre quais as propriedades que devem ser focadas com vista a que os utilizadores finais, pessoas mais comuns e não especificamente profissionais, possam considerar essa explicação como eficaz e compreensível. Segundo eles, e citando os trabalhos de Lawless, W. *et al.* (2019) e Wang, D. *et al.* (2019), a pesquisa deveria ir na direção da exploração do conhecimento e das experiências oriundas da área da interação humano-computador, e, bem assim no desenvolvimento de interfaces explicativas interativas nessa mesma área. Os autores defendem, o que nos parece assertivo, tendo em consideração o *trade-off*, inversamente correlacionado e existente, entre as dimensões de precisão do modelo e sua interpretabilidade/explicabilidade, uma definição de explicabilidade que tenha uma aplicabilidade mais ampla em contextos e domínios de aplicações de sistemas de IA.

Vilone, G. & Longo, L. (2021), ressalvam o papel fundamental dos explicadores (modelos de IA), pois será com estes que os utilizadores finais irão interagir (do centro para fora). O desenvolvimento de explicadores deverá levar em conta os múltiplos atributos e noções que

estão ligados ao constructo psicológico de explicabilidade - estritamente conectado com o ser humano mas também ligado a outros constructos, como confiança, transparência e privacidade.

2.2.8 Características dos modelos de *ML* e auditabilidade

Marcinkevics, R. & Vogt, J. (2023), frisam que, apoiando-se nas citações de Carvalho, D. *et al.* (2019), Doshi-Velez, F. & Kim, B. (2017) e Lipton, Z. (2018), a nível prático a interpretabilidade e a explicabilidade, normalmente, revelam-se mais pertinentes aquando da submissão dos sistemas de *ML* a uma audição na busca dos desideratos auxiliares e, ainda, na confirmação do desempenho preditivo do modelo. Os modelos de *ML* mais apropriados e fortes deverão evidenciar resiliência ao ruído nos *Inputs* e a possíveis mudanças de domínio, e como tal, interpretações e explicações serão potenciadoras na criação de modelos mais confiáveis, robustos, transferíveis e que consigam captar relações do tipo causa – efeito, em vez de associações inesperadas. Os autores, citando Nogueira, A. *et al.* (2022), mencionam que esta visão da interpretabilidade é ambiciosa e obriga a resolução do problema da descoberta causal observacional. A título de exemplos, em que a interpretação foi pedra de toque na área de intervenção dos modelos, os autores indicam, na área da Medicina, citando Caruana, R. *et al. (*2015), a notória facilitação na "depuração" do modelo, fruto da interpretação, e onde, para previsão do risco de pneumonia, foram utilizados *GAMs*, que destacaram e mitigaram a nefasta confusão sentida no conjunto de dados. Outro exemplo, havendo citado Li, L. & Wang, Y. (2019), é o "*Manifold"* - uma ferramenta interna de visualização e depuração para modelos de *ML* desenvolvida na "Uber". Outro exemplo de sucesso, também aludido pelos autores, recordando Larson, J. *et al*. (2016), ocorreu aquando, nos EUA, na aferição da tomada de decisões na área da justiça, recorrendo a métodos de explicação, na auditoria pela "ProPublica" ao modelo de análise de reincidência do "*Correctional Offender Management Profiling for Alternative Sanctions*" (*COMPAS*) revelou que o mesmo propendia a ser racialmente tendencioso.

Marcinkevics, R. & Vogt, J. (2023) indicam que, apesar de terem a noção que não estão ainda definidos padrões uniformes, e bem enraizados, no âmbito da avaliação qualitativa ou quantitativa, provavelmente devido à ausência de uma definição geral de *ML* interpretável e explicável, e, outrossim, à diversidade e subjetividade dos desideratos e princípios comumente investigados na literatura existente; contudo, opinam destacar o trabalho que Nauta, M. *et al*. (2022) realizaram como uma das investigações mais abrangentes nestes

assuntos até à altura.

2.2.9 Abordagens *Ante-hoc* e *Post-hoc*. Explicações contrafactuais, selecionadas e sociais

Retzlaff, C. *et al.* (2024-b), afirmam que uma das abordagens mais comum, na análise e interpretação das decisões tomadas por um sistema de IA, consiste em, ancorando-se em Glanois, C. *et al. (*2021),determinar a importância de diferentes características para um determinado resultado. A relevância dessas características podem ser apuradas através de métodos agnósticos de abordagens *Post-hoc* e, mais concretamente, [*a contrario sensu* do defendido por Sudjianto, A. & Zhang, A. (2021)], os autores avocam, citando Lundberg S. & Lee, S. (2017), a utilização do método "*SHAP"*, que opera com valores de "Shapley", *i.e.*, com base na teoria dos jogos. O "*SHAP"*, valora de forma quantitativa, à contribuição que cada característica teve, num modelo de *ML*, para a previsão de uma decisão específica, calculando, também, a contribuição marginal média de uma característica em todas as combinações possíveis de características. Isto permite-lhe fornecer uma distribuição justa de valores de importância numa contemplação de todas as interações possíveis entre as várias características. Os autores salientam, apoiados na visão de Weerts, H. & Van Ipenburg, W. & Pechenizkiy M. (2019), que este método reduz drasticamente o tempo necessário à instância ser compreendida pelo ser humano, e, bem assim, citando Confalonieri, R. *et al*. (2019), atribuem reconhecimento à consistência evidenciada pelos valores do "*SHAP*", os quais, alinhando-se com a intuição humana, conseguiram maior aproximação com a explicação que um ser humano faria do modelo. Como prova desse alinhamento, em que o "*SHAP*" pode, como mecanismo de seleção de recursos, ajudar na concretização de melhores resultados, Retzlaff, C., *et al.* (2024-b), citaram, estudos realizados no âmbito da previsão na área da Medicina, por: Liu, Y. *et al*. (2022) – no diagnóstico da doença de Parkinson; Marcílio, W. & Eler, D. (2020) – na avaliação genérica de aplicação de recursos, onde o "*SHAP"*, revelou ser capaz de explicar as decisões de um modelo e conseguiu alcançar melhores resultados do que os algoritmos de seleção de recursos mais comumente usados, e, ainda, Lundberg, S. *et al.* (2020) – com a aplicação de um algoritmo de tempo polinomial aos problemas de *ML* médica, tendo como base a teoria dos jogos, com o intuito de calcular as melhores explicações, conseguiram demonstrar que a combinação de muitas explicações locais de alta qualidade permite representar a estrutura global, mantendo a fidelidade local ao modelo original.

Retzlaff, C. *et al.* (2024-b), referem-se às abordagens *Ante-hoc*, com suporte em Holzinger, A. (2018), Holzinger, A. *et al.* (2019-b),e, ainda, Retzlaff, C. *et al.* (2024-a), como sendo também, por vezes, denominadas de: explicabilidade intrínseca, modelos transparentes ou de Caixas de vidro (*Glass-boxes*); e indicam-nas como sendo modelos de *ML* inerentemente interpretáveis, que incorporam técnicas de interpretabilidade diretamente na arquitetura do modelo ou no processo de aprendizagem. Os autores indicam três tipos de exemplos de modelos de *ML* interpretáveis:

- 1.º- As *Decision Tree* (*DT)* - árvores de decisão, como citado por Molnar, C. (2022), representam uma estrutura hierárquica de regras "*if-then-else"* - SE-ENTÃO - SENÃO*",* de fácil visualização, compreensão e interpretação pelo ser humano. São os modelos interpretáveis de *ML* mais utilizados e operam dividindo os dados diversas vezes, em consonância com determinados valores de corte nos recursos, criando diferentes subconjuntos do conjunto inicial de dados, com cada instância a pertencer a um subconjunto. Podem ser interpretados seguindo a estrutura da árvore, começando pelo "nó-raiz", ao longo dos próximos "nós" até alcançar o "nó-folha" onde se encontra o resultado previsto; conforme os autores ilustraram da citação realizada de Safavian, S. & Landgrebe, D. (1991).
- 2.º- Os autores retomam os *GAMs – Generalized Additive Models* - Modelos Aditivos Generalizados, já citados por Caruana, R. *et al.* (2015), que ampliam as capacidades da regressão linear e logística, proporcionando uma abordagem diferente com vista a potenciar a compreensão, particularmente ao considerar termos de baixa dimensão e de fácil leitura pelos seres humanos.
- Retzlaff, C. *et al.* (2024-b), citando Letham, B. *et al.* (2015), referem ainda, como 3.º exemplo de modelos de *ML* interpretáveis - as "*Bayesian Rule Lists*" (*BRL)* - Listas de Regras Bayesianas, que indicam serem precisas, e também interpretáveis, contudo, somente por especialistas humanos, envolvendo listas de decisão com uma série de instruções "*if-Then*" - SE-ENTÃO (por exemplo, SE pressão alta, ENTÃO derrame). Denotando ser um método com características multivariadas de alta dimensão, mas que pode ser transferido para um espaço de decisão de baixa dimensão (*i.e. "*discretizado") e, portanto, interpretável por humanos.

Retzlaff, C. *et al.* (2024-b), retomam o tema das explicações centradas no ser humano e no papel das ciências sociais, acompanhando Miller, T. (2019) e Molnar, C. (2022), revelando que as explicações são contrastivas*, i.e.*, são procuradas em resposta a casos contrafactuais [salientamos que: Molnar, C. (2022, Capítulo 15, para. 1 e 3), escreve no seu livro,

conjuntamente com Susanne Dandl, sobre as explicações contrafactuais, defendendo: “ Uma explicação contrafactual descreve uma situação causal na forma: "Se X não tivesse ocorrido, Y não teria ocorrido". Por exemplo: "Se eu não tivesse tomado um gole deste café quente, não teria queimado a língua". O evento Y é que eu queimei a língua; a causa X é que eu tomei um café quente. Pensar em contrafactuais requer imaginar uma realidade hipotética que contradiga os fatos observados (por exemplo, um mundo em que eu não tenha tomado o café quente), daí o nome "contrafactual". A capacidade de pensar em contrafactuais torna os humanos mais inteligentes em comparação com outros animais.” e, ainda, ”Uma explicação contrafactual de uma previsão descreve a menor alteração nos valores das características que altera a previsão para uma saída predefinida.”]. Ademais, as explicações, são, ainda, selecionadas, *i.e*, não exprimem a causa real e completa do evento, sendo antes causas escolhidas entre as verificadas*. C*ontinuam expondo que, as probabilidades não são tidas em conta, as explicações são sociais*, i.e.*, são relativas às crenças do explicador sobre as crenças do explicado; resultando de cariz contextual, ou seja, é apenas selecionado um pequeno subconjunto de causas relevantes para o contexto em apreço.

2.2.10 ”*Causal Representation Learning* (*CRL)*” – Aprendizagem de representação causal

Gemma M. & Bryon A. (2025), mencionam que, embora percebam que, na busca de entendimento sobre o modo de previsão dos algoritmos de *ML*, haja quem opte, por ferramentas para analisar e interpretar modelos pré-treinados, *i.e.,* métodos de explicabilidade *Post-hoc*, que já forneceram *insights* sobre modelos de imagem e de linguagem, citando Selvaraju, R. *et al. (*2017) e Templeton, A. *et al. (*2024); vêm relembrar, contudo, que, como referiu Rudin, C. (2019), estes métodos podem não espelhar com rigor o que o modelo original processa/ou. Partilham a opinião citada de Bilodeau, B. *et al*. (2024) de que já se verificou que muitos métodos de explicabilidade não vão muito além de serem suposições aleatórias para conjeturar o comportamento do modelo de IA.

Por conseguinte, os autores perfilham a consideração da construção, *ab initio*, de novos modelos que possuam mecanismos internos mais transparentes e interpretáveis, de modelos generativos em que as representações latentes sejam explicitamente parametrizadas. Porém estes modelos generativos, mesmo tendo mecanismos internos completamente diferentes e com parâmetros, também eles, diferentes, ainda assim, poderão produzir estimativas de densidade idênticas, o que, formalmente, indica que a parametrização dos modelos é não

identificável. Tal resultado, na área de estimação de parâmetros, essencial, para a interpretabilidade dos modelos generativos, configura um obstáculo técnico de laboriosa resolução.

Os autores propõem, para obstar a esse desafio técnico, duas soluções utilizadas na modelação de modelos clássicos de aprendizagem de representação: a) restringir a forma do modelo ou b) obter acesso a modalidades de dados mais ricos, provenientes de diferentes ambientes. Quanto à primeira solução, à imposição de restrições, concluem que estas revelam-se, frequentemente, insuficientes para modelar conjuntos de dados complexos, não estruturados e multimodais, que são os que proliferam em aplicações modernas de IA. Nesse tipo de contexto são de considerar modelos mais flexíveis que evitem essas restrições. Por conseguinte restava rumar em direção a modalidades de dados mais ricos, provenientes de diferentes ambientes.

Os autores ressalvam que fora a crescente procura por modelos mais facilmente transferíveis, entre um variadíssimo leque de tarefas subsequentes (diferentemente do comumente verificado nos modelos clássicos*, i.e.*, a especialização em tarefas específicas e bem definidas *a priori*), que motivou o surgimento do campo da *"Causal Representation Learning* (*CRL)*"– Aprendizagem de representação causal.

Gemma M. & Bryon A. (2025) esclarecem que a *CRL* adota uma perspetiva moderna sobre a análise fatorial, impulsionada pelo sucesso da IA generativa e com três princípios orientadores fundamentais:

- A flexibilidade, que é conseguida permitindo as correlações entre fatores latentes e, ainda, permitindo relações não lineares gerais por meio de redes neuronais arbitrariamente complexas e ainda com *DL*;
- A interpretabilidade, que é conseguida com a obrigação da dispersão (para simplificar o modelo) e com o destaque da causalidade (que possibilita uma interpretação intervencionista dos fatores latentes), e, ainda, com recurso ao uso de modelos gráficos causais;
- A transferibilidade, que é conseguida fruto da utilização de modelos causais, sobejamente tidos como geradores de previsões estáveis, invariáveis e transferíveis em contextos e ambientes em constante mudança.

Os autores indicam, ademais, que a distinção entre os objetivos da geração e da compreensão dos modelos generativos de IA tem duas implicações estatísticas relevantes: *Primo*, a geração é notoriamente mais simples do que a compreensão e pode ser conseguida

sem a compreensão; e, *Secundo*, com vista a entender os modelos, os parâmetros devem ser identificáveis para que possam ser estimados de forma consistente.

Gemma M. & Bryon A. (2025), concluem que a *CRL* é uma disciplina emergente com fundações em *ML*, estatística e causalidade. Esta purga numa variedade de domínios para atentar soluções nos importantes desafios práticos na utilização de modelos generativos em aplicações científicas onde sejam essenciais a interpretabilidade e a causalidade. Os autores, fazendo alusão a Peters, J. & Bühlmann, P. & Meinshausen, N. (2016), Huang,B. *et al*. (2020), Perry, R. & Von Kügelgen, J. & Schölkopf, B. (2022) e, ainda, a Mooij, J. & Magliacane, S. & Claassen, T. (2020), salientam o surgimento, com o intuito de melhorar a generalização e/ou identificar estruturas latentes, da utilização de diversos ambientes*, i.e.*, de dados "*non-iid*" (dados não independentes e não distribuídos de forma idêntica) emergentes: do aumento de dados; de mudanças de distribuição; de diferentes locais; de diferentes espaços temporais; de diferentes contextos; entre outros. Os autores destacam o desafio, remetendo para os trabalhos citados de Varici, B. *et al.* (2024) e Chen, T. *et al*. (2024), de perceber como ambientes gerais podem permitir a identificabilidade quando intervenções rigorosas não estão disponíveis.

Gemma M. & Bryon A. (2025), aventam a possibilidade de desenvolver modelos de base pré-treinados com interpretabilidade causal; todavia dão nota dos imensos desafios práticos associados, tais como: conseguir a conexão entre os objetivos da previsão do próximo *token* com os objetivos causais; alcançar a sua utilização baseado em princípios de conjuntos de dados diversos, multimodais e em escala de *Internet*; possibilitar a extração de conceitos interpretáveis; entre outros. A título de exemplo, com progressos verificados, nesse último aspeto, os autores citam Rajendran, G. *et al*. (2024); e noutro objetivo um pouco mais alternativo, o de extrair conceitos interpretáveis *Post-hoc* de *LLMs* pré-treinados usando aprendizagem de dicionário esparso, citaram ilustrativamente Bricken, T. *et al*. (2023) e Templeton, A. *et al.* (2024). Essas abordagens de aprendizagem de dicionário esparso pressupõem que cada *token* é a combinação linear de um pequeno conjunto de vetores de conceitos, selecionados de um dicionário de conceitos supercompleto.

Conclusão

Refletindo sobre a capacidade em aferir a dimensão da intervenção da IA no processo das Patentes do Direito de Propriedade Industrial conseguimos perceber que a tarefa resulta cada vez mais difícil fruto da constante e inopinada evolução dos sistemas de IA.

De ressaltar que não se exalta a potencial atribuição de direitos privativos à IA quando esta tem um papel interventivo nas invenções, ainda que com algum grau de autonomia, com um mínimo de *Input* humano (pois sem qualquer *Input* deste tipo ainda está longe de acontecer), tratando-se, antes, de preservar o impulso da inovação e criação industrial. Importa, pois, manter um equilíbrio entre investimento e retorno económicos, não descurando a adequada atribuição de autoria e direitos respetivos ao seu justo inventor/criador, como também observa Abbott, R. (2019). Contudo, este último tende à defesa, no campo do reconhecimento de autoria, de atribuição desta aos sistemas de IA. Antagonicamente, alinhamos com a consideração, generalizada, da não pertinência em reconhecer um sistema de IA como autor e, pelo contrário, preservar o destino de tal reconhecimento ao ser humano, mesmo havendo recurso a uma ferramenta altamente técnica para auxílio na invenção.

Outra preocupação retratada impende sobre os requisitos de novidade, não obviedade e da reprodutibilidade, pois deles depende, grandemente, concluir-se se a invenção é nova, como foi alcançada e se pode ser replicada para aplicação Industrial e Agrícola. Tendo isto em mente, preconizamos que uma eventual solução poderá residir na implicação de IA na avaliação dos requisitos de Patenteabilidade (como também sugeriu Abbott, R. (2018)) e, bem assim, a envolvência de sistemas de computação mais avançados, Supercomputadores ou Computadores Quânticos, para processamentos mais célere dos dados implícitos nessa mesma avaliação. Todavia, face à relativa escassez desses meios de avultada dimensão económica; poderá, eventualmente, emergir a necessidade de colaboração internacional, como por exemplo a aquisição/aluguer, conjunto, destes Computadores mais avançados em apoio ao *Five IP Offices* - IP5.

A essência da patenteabilidade nas invenções com intervenção de IA recaí sobre a necessidade, e capacidade, da posse do conhecimento, suficientemente claro e aceitável, do modo de funcionamento da IA. Para se concretizar esse intento de interpretabilidade, acolhemos as posições de adoção de sistemas de IA ancorados em modelos de *ML*, intrinsecamente/inerentemente interpretáveis (*ab ignitio*, no seu *design*, *i.e.* modelos *Ante-hoc)* que disponibilizem interpretações apoiadas em informações contextuais adicionais,

provenientes do conhecimento do domínio e relacionadas com o objetivo da análise.

Acolhemos de bom grado a visão de Stierle, M. (2021), sobre um eventual regime *sui generis* alusivo à patenteabilidade de invenções com intervenção de IA, ao qual somos a sugerir aditar um período de valência da proteção da invenção menor ao habitual (genericamente 20 anos) dado a galopante evolução tecnológica. Todavia, somos apologistas que uma mudança de política no sistema de patenteamento tomada de forma unilateral, por uma Organização ou Estado, pode revelar-se contraproducente dada a necessidade de universalizar o reconhecimento da invenção e do seu autor, bem como a atribuição de direitos de PI respetivos.

Como mencionado atrás, no campo, sensível, da Medicina, Holzinger, A. *et al.* (2019-a), referindo-se a Xu, Y. *et al*. (2014) e ainda a Komura, D. & Ishikawa, S. (2018), destacam a utilização de sistemas de IA que estiveram sujeitos a uma Aprendizagem fracamente supervisionada ("*Weakly Supervized Learning* (*WSL)*") - um método de construção de modelos preditivos através de uma aprendizagem fraca, *i.e*, fraca fruto de: -supervisão incompleta; - supervisão inexata ou ainda supervisão imprecisa. Ousamos, a sugerir se, em caso de apreciação da intervenção da IA na invenção, não tendo havido total apreensão das informações encerradas nos sistemas de IA, *i.e.* da compreensão do tipo de intervenção e alcance da mesma, seria descabido a estruturação de orientações, específicas para as áreas visadas (v.g. finanças, atividade bancária creditícia, saúde, justiça, entre outras com severas implicâncias na Sociedade), com a determinação de uma dimensão aceitável de dúvida a valorar? Existirá um valor, uma medida, uma percentagem (%) aceitável de dúvida, de não interpretabilidade aceitável? Poderá generalizar-se a utilização e aceitação de modelos *WSL* (com as suas lacunas) por imperatividade da evolução fulgurante da Sociedade e não se "justificar" esperar por esclarecimentos totais e exatos?

Entendemos que as mudanças tecnológicas brotam alucinantemente, contudo, as várias partes interessadas demoram algum tempo a adaptarem-se. As Instituições e os Tribunais demoram a interpretarem e adequarem a legislação e, ainda, as práticas da indústria a consolidarem-se. Estando em jogo interesses económicos, a maioria das vezes, relevantemente diferentes, a solução, com implicação em reformas políticas prematuras, correrá o risco de consequências injustificadas e poderá, inclusive, desconsiderar a vertente de uma certa autorregulação própria dos mercados económicos.

Referências Bibliográficas

Abbott, R.. (2018, pp.27-38). *Everything is Obvious*. UCLA - University of California, Los Angeles - Law School. Law Review n.º 2; consultado em 10-10-2025, disponível em: http://dx.doi.org/10.2139/ssrn.3056915

Abbott, R.. (2019, para. 9-11). *The Artificial Inventor Project*. WIPO MAGAZINE, December 11, 2019; consultado em 10-10-2025, disponível em: https://www.wipo.int/en/web/wipo-magazine/articles/the-artificial-inventor-project-41111

Acharya, D. & Kuppan, K. & Divya, B.. (2025). *Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey.* IEEE Access; consultado em 10-10-2025, disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10849561

Bahman, Z. & Farahnaz, B.. (2023, Cap. 8.2). *Um Manual de Inteligência Artificial na Administração de Medicamentos*; consultado em 10-10-2025, disponível em: https://www.sciencedirect.com/topics/computer-science/artificial-general-intelligence

Barbosa, M.. (2017, pp. 1482 e 1486). *Inteligência Artificial, E-Persons e Direito: Desafios e Perspetivas;* consultado em 10-10-2025, disponível em: https://www.cidp.pt/revistas/rjlb/2017/6/2017_06_1475_1503.pdf

Belcak, P. & Heinrich, G. & Diao, S. & Fu, Y. & Dong, X. & Muralidharan, S. & Lin, Y. & Molchanov, P. . *Small Language Models are the Future of Agentic AI*. (02 de junho 2025). NVIDIA Research & Georgia Institute of Technology; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/2506.02153

Bilodeau, B. & Jaques, N. & Koh, P. & Kim, B.. (2024). *Impossibility theorems for feature attribution.* Proceedings of the National Academy of Sciences, Vol.121, n.º 2:e2304406120 ; consultado em 10-10-2025, disponível em: https://www.pnas.org/doi/10.1073/pnas.2304406120

Burrell,J.. (2016).*How the machine 'thinks': Understanding opacity in machine learning algorithms*. Sage Journal; consultado em 10-10-2025, disponível em: https://doi.org/10.1177/2053951715622512

Bricken, T. & Templeton, A. & Batson, J. & Chen, B. & Jermyn, A. & Conerly, T. & Turner, N. & Anil, C. & Denison, C. & Askell, A. & Lasenby, R. & Wu, Y. & Kravec, S. & Schiefer, N. & Maxwell, T. & Joseph, N. & Hatfield-Dodds, Z. & Tamkin, A. & Nguyen, K. & McLean, B. & Burke, J. & Hume, T. & Carter, S. & Henighan, T. & Olah, C.. (2023). *Towards*

monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread ; consultado em 10-10-2025, disponível em: https://transformer-circuits.pub/2023/monosemantic-features/index.html

Caruana, R. & Lou, Y. & Gehrke, J. & Koch, P. & Sturm, M. & Elhadad, N.. (2015). *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission*. Paper presented at 21th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '15) (pp.1721-1730). ACM; consultado em 10-10-2025, disponível em: https://doi.org/10.1145/2783258.2788613

Carvalho, D. & Pereira, M. & Cardoso, J.. (2019). *Machine Learning Interpretability: A Survey on Methods and Metrics*; consultado em 10-10-2025, disponível em: https://doi.org/10.3390/electronics8080832

Chapelle, O. & Schölkopf, B. & Zien, A.. (2006). *Semi-Supervised Lear*ning. ISBN 978-0-262-03358-9; consultado em 10-10-2025, disponível em: https://www.molgen.mpg.de/3659531/MITPress--SemiSupervised-Learning.pdf

Chen, T. & Bello, K. & Locatello, F. & Aragam, B. & Ravikumar, P.. (2024). *Identifying general mechanism shifts in linear causal representations.* In Neural Information Processing Systems; consultado em 10-10-2025, disponível em: https://arxiv.org/abs/2410.24059

Confalonieri R. & Besold T.R. & Weyde T. & Creel K. & Lombrozo T. & Mueller S.T. & Shafto P.. (2019).*What makes a good explanation? Cognitive dimensions of explaining intelligent machines.* Goel A.K., Seifert C.M., Freksa C. (Eds.), Proceedings of the 41th annual meeting of the cognitive science society, cogsci 2019: creativity + cognition + computation, pp. 25-26; consultado em 10-10-2025, disponível em: https://escholarship.org/content/qt7qd3c6rh/qt7qd3c6rh.pdf

Credo AI Glossary. *Explore to learn all the must-know definitions of Responsible AI & AI Governance*; consultado em 10-10-2025, disponível em: https://www.credo.ai/glossary

Dam, H. & Tran, T. & Ghose, A.. (2018). *Explainable software analytics*. Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, ACM, Gothenburg, Sweden, pp.53-56; consultado em 10-10-2025, disponível em: https://doi.org/10.1145/3183399.3183424

Das, A. & Agrawal, H. & Zitnick, L. & Parikh, D. & Batra, D.. (2017*). Human attention in visual question answering: Do humans and deep networks look at the same regions?.* Computer Vision and Image Understanding, vol. 163, pp. 90–100; consultado em 10-10-2025, disponível em: https://doi.org/10.1016/j.cviu.2017.10.001

Decreto n.º 52/91, de 30 de agosto (1991), Aprova para ratificação a *Convenção de Munique sobre a Patente Europeia de 5 de Outubro de 1973*; consultado em 10-10-2025, disponível em: https://diariodarepublica.pt/dr/detalhe/decreto/52-403723

Deng, Z. & Guo, Y. & Han, C. & Ma, W. & Xiong, J. & Wen, S. & Xiang, Y.. (2024). *Ai agents under threat: A survey of key security challenges and future pathways.* ACM Computing Surveys, vol. 57, no. 7, pp.1-36; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/2406.02630

Dietterich, T. & Lathrop, R. & Lozano-Pérez, T.. (1997). *Solving the multiple instance problem with axis-parallel rectangles*. Artif. Intell. N.º 89, pp.31-71; consultado em 10-10-2025, disponível em: https://www.sciencedirect.com/science/article/pii/S0004370296000343

Doshi-Velez, F. & Kim, B.. (2017, pp.2-4).*Towards A Rigorous Science of Interpretable Machine Learning*; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/1702.08608

European Patent Office - EPO**.** *Guidelines for Examination in the European Patent Office.* April 2025 edition; consultado em 10-10-2025, disponível em: https://www.epo.org/en/legal/guidelines-epc

Fernandes, R.. (2012, pp.81-82). *A Patente Farmacêutica e o Medicamento Genérico. O problema da tensão jurídica entre o direito exclusivo e a livre utilização;* [Tese de Doutoramento em Ciências Jurídicas Especialidade em Ciências Jurídico Privatísticas]. Universidade do Minho - Escola de Direito; consultado em 10-10-2025,disponível em: https://repositorium.uminho.pt/server/api/core/bitstreams/d983a202-1c11-4184-9ca2-f342b1667802/content

Finn, T. & Downie, A. (2025). *Agentic AI vs. generative AI*.; consultado em 10-10-2025, disponível em: https://www.ibm.com/think/topics/agentic-ai-vs-generative-ai

Five IP Offices - IP5; consultado em 10-10-2025, disponível em: https://www.fiveipoffices.org

Gemma M. & Bryon A.. (April 17, 2025). *Towards Interpretable Deep Generative Models*

via Causal Representation Learning. Department of Statistics, Rutgers University ; Booth School of Business, University of Chicago; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/2504.11609v1

Gilpin, L. & Bau, D. & Yuan, B. & Bajwa, A. & Specter, M. & Kagal, L.. (2019, pp.2, 3, 8).*Explaining Explanations: An Overview of Interpretability of Machine Learning*. Computer Science and Artificial Intelligence Laboratory Massachusetts, Institute of Technology Cambridge; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/1806.00069

Glanois C. & Weng P. & Zimmer M. & Li D. & Yang T. & Hao J. & Liu W.. (2021). *A survey on interpretable reinforcement learning*; consultado em 10-10-2025, disponível em: https://arxiv.org/abs/2112.13112

Godinho, I. & Flores, C. & Marques, N.. (2020). *Consultation on the white paper on artificial intelligence: a european approach*. ULP Law Review, Vol. 14,n.º1, pp.157-167; consultado em 10-10-2025, disponível em: https://revistas.ulusofona.pt/index.php/rfdulp/issue/view/769

Gonçalves, M.. (2020). *Requisitos de Patenteabilidade*. Data Vénia - Revista Jurídica Digital. Ano 8, N.º 11, dezembro 2020. ISSN 2182-6242; consultado em 10-10-2025, disponível em: https://www.datavenia.pt/ficheiros/edicao11/datavenia11_p429_446.pdf

Goodfellow, I. & Bengio,Y. & Courville, A.. (2016). *Deep Learnig E-book.* MIT Press book; consultado em 10-10-2025, disponível em: https://www.deeplearningbook.org/

Google People + AI Research (PAIR). *Facets - Visualization for ML Datasets*; consultado em 10-10-2025, disponível em: https://pair-code.github.io/facets/

Guillaume, S.. (2001). Designing *fuzzy inference systems from data: An interpretability-oriented review*. IEEE Transactions on Fuzzy Systems, n.º 9, pp.426-443; consultado em 10-10-2025, disponível em: https://hal.science/hal-01320328v1/file/mo2001-pub00009268.pdf

Hacker, P. & Passoth, J.. (2022).*Varieties of AI explanations under the law.* From the GDPR to the AIA, and beyond. Holzinger, A. & Goebel, R.& Fong, R. & Moon, T. & Müller, K. & Samek, W. (Eds.), XxAI, Springer, pp. 343-373 ; consultado em 10-10-2025, disponível em: https://link.springer.com/chapter/10.1007/978-3-031-04083-2_17?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot&getft_integrator=sciencedirect_contenthosting#citeas

Hilty, R. & Hoffmann, J. & Scheuerer, S.. (February 11, 2020).*Intellectual Property Justification for Artificial Intelligence.* Draft chapter. Forthcoming in: J.-A. Lee, K.C. Liu, R. M. Hilty (eds.), Artificial Intelligence & Intellectual Property, Oxford, Oxford University Press, 2020, Forthcoming, Max Planck Institute for Innovation & Competition Research Paper No. 20-02; consultado em 10-10-2025, disponível em: https://ssrn.com/abstract=3539406

Holdsworth, J. & Scapicchio, M.. (2024). IBM Think. *What is deep learning?;* consultado em 10-10-2025, disponível em: https://www.ibm.com/think/topics/deep-learning

Holzinger, A. & Plass, M. & Holzinger, K. & Crisan, G. & Pintea, C. & Palade, V.. (2017). *A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop*; consultado em 10-10-2025, disponível em*:* https://arxiv.org/pdf/1708.01104

Holzinger, A.. (2018). *Explainable AI (ex-AI).Informatik-Spektrum*, n.º 41, pp.138-143 ; consultado em 10-10-2025, disponível em: https://doi.org/10.1007/s00287-018-1102-5

Holzinger, A. & Plass, M. & Kickmeier-Rust, M. & Holzinger, K. & Crian, G. & Pintea, C. & Palade, V.. (2018). *Interactive machine learning: Experimental evidence for the human in the algorithmic loop.* Applied Intelligence. pp.1-14; consultado em 10-10-2025, disponivel em: https://link.springer.com/content/pdf/10.1007/s10489-018-1361-5.pdf

Holzinger, A. & Langs, G. & Denk, H. & Zatloukal, K. & Müller, H.. (2019-a).*Causability and explainability of artificial intelligence in medicine*; consultado em 10-10-2025, disponível em: https://doi.org/10.1002/widm.1312

Holzinger, A. & Plass, M. & Kickmeier-Rust, M. & Holzinger, K. & Crişan, G. & Pintea, C. & Palade, V.. (2019-b). *Interactive machine learning: experimental evidence for the human in the algorithmic loop*. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, Vol.49, n.º7, pp. 2401-2414; consultado em 10-10-2025, disponível em: https://doi.org/10.1007/s10489-018-1361-5

Huang, B. & Zhang, K. & Zhang, J. & Ramsey, J. & Sanchez-Romero, R. & Glymour, C. & Schölkopf, B.. (2020). *Causal discovery from heterogeneous/nonstationary data.* Journal of Machine Learning Research, Vol.21, n.º89, pp.1-53 ; consultado em 10-10-2025, disponível em: https://www.jmlr.org/papers/volume21/19-232/19-232.pdf

Information Commissioner's Office and The Alan Turing Institute. *Explaining decisions*

made with AI, 1.0.3. (2022) ; consultado em 10-10-2025, disponível em: https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/

Instituto Nacional de Propriedade Industrial (s.d.)., consultado em 10-10-2025, disponível em: https://justica.gov.pt/registos/propriedade-industrial/patente

IP5. *Examination practices on AI-related inventions*. (June 2023); consultado em 10-10-2025, disponível em: https://link.epo.org/ip5/Chart_Examination%20practices%20on%20AI-related%20inventions.pdf#page=5

IP5. *Examination practices on AI-related inventions - Comparison Table for AI cases* – (June 2024); consultado em 10-10-2025, disponível em: https://link.epo.org/ip5/exam_pract_AI-related_2024

IP5. *Inventorship of AI-generated Inventions*. (June 20, 2024); consultado em 10-10-2025, disponível em: https://link.epo.org/ip5/Inventorship_AI-related_inventions_2024

Japan Patent Office. (2020). *Examination Guidelines for Patent and Utility Model in Japan*; consultado em 10-10-2025, disponível em: https://www.jpo.go.jp/e/system/laws/rule/guideline/patent/tukujitu_kijun/index.html

Japan Patent Office. (2024). *Examination Standards Office. Explanatory materials for Case Examples pertinent to AI-related technologies (Updated in March 2024). Overview of the Additional Case Enrichment*. (p.9); consultado em 10-10-2025, disponível em: https://www.jpo.go.jp/e/system/laws/rule/guideline/patent/document/ai_jirei_e/jirei_add2024_e.pdf

Jaruga-Rozdolska, A.. (2022). *Artificial intelligence as part of future practices in the architect's work: Midjourney generative tool as part of a process of creating an architectural form*. Architectus, n.º. 3 (71), pp.95-104; consultado em 10-10-2025, disponível em: https://www.dbc.wroc.pl/dlibra/doccontent?id=121602

Kim, B. & Doshi-Velez, F.. (2018). *Introduction to Interpretable Machine Learning*. In Proceedings of the CVPR 2018 Tutorial on Interpretable Machine Learning for Computer Vision, Salt Lake City, UT, USA, 18 June 2018.]; consultado em 10-10-2025, disponível em: https://beenkim.github.io/slides/DLSS2018Vector_Been.pdf

Kissinger, H. & Schmidt, E. & Huttenloche, D.. (2021). *The Age of AI And Our Human Future*; consultado em 10-10-2025, disponível em: https://pdfcoffee.com/the-age-of-ai-

and-our-human-future-henry-kissinger-eric-schmidt-etc-z-library-pdf-free.html
Komura, D. & Ishikawa, S.. (2018). *Machine learning methods for histopathological image analysis.* Computational and Structural Biotechnology Journal, Vol.16, pp.34-42; consultado em 10-10-2025, disponível em: https://doi.org/10.1016/j.csbj.2018.01.001
Kumar, E. & Venkatasubramanian, S. & Scheidegger, C. & Friedler, S.. (2020). *Problems with Shapley-value-based explanations as feature importance measures*. In International Conference on Machine Learning. PMLR, pp.5491-5500; consultado em 10-10-2025, disponível em: https://proceedings.mlr.press/v119/kumar20e/kumar20e.pdf
Larson, J. & Mattu, S. & Kirchner, L. & Angwin, J.. (2016). *How we analyzed the COMPAS recidivism algorithm*; consultado em 10-10-2025, disponível em: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
Lawless, W. & Mittu, R. & Sofge, D. & Hiatt, L.. (2019). *Artificial intelligence, autonomy, and human-machine teams: Interdependence, context, and explainable AI*. AI Mag., Vol. 40, n.º 3, pp.5-13; consultado em 10-10-2025, disponível em: https://doi.org/10.1609/aimag.v40i3.2866
Letham, B. & Rudin, C. & McCormick, T. & Madigan, D.. (2015*). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model.* The Annals of Applied Statistics, Vol.9, n.º3, pp. 1350-1371 ; consultado em 10-10-2025, disponível em: https://users.cs.duke.edu/~cynthia/docs/LethamRuMcMa15.pdf
Lei das Patentes – Japão - Patent Act- Act No. 121 of April 13, (1959) - Japan, last version: Act No. 42 of 2021; consultado em 10-10-2025, disponível em: https://www.japaneselawtranslation.go.jp/en/laws/view/4617#je_ch2at1
Li, J. & Lavrukhin, V. & Ginsburg, B. & Leary, R. & Kuchaiev, O. & Cohen, J. & Nguyen, H. & Gadde, R.. (2019)., *Jasper: An end-to-end convolutional neural acoustic model*; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/1904.03288
Li, L. & Wang, Y.. (2019). *Manifold: A model-agnostic visual debugging tool for machine learning at Uber*; consultado em 10-10-2025, disponível em: https://www.uber.com/en-PT/blog/manifold/
Lipton, Z.. (2018). *The mythos of model interpretability*. Communications of the ACM, Volume 61, Issue 10, pp. 36-43; consultado em 10-10-2025, disponível em: https://dl.acm.org/doi/pdf/10.1145/3233231
Liu, Q. & Qian, Z. & Marvasty, I. & Rinehart, S. & Voros, S. & Metaxas, D.. (2010). *Lesion-specific coronary artery calcium quantification for predicting cardiac event with multiple*

instance support vector machines. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 484–492; consultado em 10-10-2025, disponível em: https://link.springer.com/chapter/10.1007/978-3-642-15705-9_59

Liu, Y. & Liu, Z. & Luo, X. & Zhao, H.. (2022). *Diagnosis of parkinson's disease based on SHAP value feature selection*. Biocybernetics and Biomedical Engineering, Vol.42, n.º3, pp. 856-869; consultado em 10-10-2025, disponível em: https://doi.org/10.1016/j.bbe.2022.06.007

Lombrozo,T.. (2006). *The structure and function of explanations*. Department of Psychology, University of California at Berkeley, Berkeley, CA 94720, USA. Trends in cognitive sciences, Vol 10, nº 10, pp. 464-470; consultado em 10-10-2025, disponível em : https://fitelson.org/few/few_08/lombrozo_reading.pdf

Lou, Y. & Caruana, R. & Gehrke, J.. (2012). *Intelligible models for classification and regression*. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Beijing, China, pp. 150-158; consultado em 10-10-2025, disponível em: https://www.cs.cornell.edu/~yinlou/papers/lou-kdd12.pdf

Lu, Y. & Aleta, A. & Du, C. & Shi, L. & Moreno, Y.. (2024). *LLMs and generative agent-based models for complex systems research***.** Physics of Life Reviews, Vol.51. (December 2024), pp. 283-293; consultado em 10-10-2025, disponível em: https://www.sciencedirect.com/science/article/pii/S1571064524001386

Lundberg, S. & Lee, S.. (2017). *A unified approach to interpreting model predictions*. In Proceedings of the 31st International Conference on Neural Information Processing Systems. pp.4768–4777; consultado em 10-10-2025, disponível em: https://dl.acm.org/doi/pdf/10.5555/3295222.3295230

Lundberg, S. & Erion, G. & Chen, H. & DeGrave, A. & Prutkin, J. & Nair, B. & Katz, R. & Himmelfarb, J. & Bansal, N. & Lee, S.. (2020). *From local explanations to global understanding with explainable AI for trees*. Nature Machine Intelligence, n.º 2, pp.56-67 ; consultado em 10-10-2025, disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC7326367/

Man-Cho So, A.. (2020).*Technical Elements of Machine Learning for Intellectual Property Law*; consultado em 10-10-2025, disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3635942

Manual of Patent Examining Procedure (MPEP). (2025); *Appendix L Consolidated Patent*

Laws- January 2025 Update, consultado em 10-10-2025, disponível em: https://www.uspto.gov/patents/laws/manual-patent-examining-procedure

Marcílio, W. & Eler, D.. (2020). *From explanations to feature selection: assessing SHAP values as feature selection mechanism*. 33rd SIBGRAPI conference on graphics, patterns and images, pp.340-347; consultado em 10-10-2025, disponível em: http://sibgrapi.sid.inpe.br/col/sid.inpe.br/sibgrapi/2020/09.25.14.27/doc/PID6618233.pdf

Marcinkevics, R., & Vogt, J.. (2023). *Interpretable and explainable machine learning: A methods-centric overview with concrete examples*. WIREs Data Mining and Knowledge Discovery, e1493; consultado em 10-10-2025, disponível em: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1493

Miller, T.. (Feb. 2019, p.6). *Explanation in artificial intelligence: Insights from the social sciences.* Artif. Intell., Vol. 267, pp.1-38; consultado em 10-10-2025, disponível em: https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C.. (2019). *Interpretable Machine Learning*; consultado em 10-10-2025, disponível em: https://christophm.github.io/interpretable-ml-book/

Molnar, C.. (2022). *Interpretable Machine Learning*;(2.nd Edition), consultado em 10-10-2025, disponível em: https://christophm.github.io/interpretable-ml-book/

Molnar,C. & König, G. & Herbinger, J. & Freiesleben, T. & Dandl, S. & Scholbeck, C. & Casalicchio, G. & Grosse-Wentrup, M. & Bischl, B.. (2020). *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models*; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/2007.04131

Montavon, G. & Samek, W. & Müller, K.. (Feb. 2018, p.2).*Methods for interpreting and understanding deep neural networks.* Digit. Signal Process., vol. 73, pp. 1-15; consultado em 10-10-2025, disponível em: https://doi.org/10.1016/j.dsp.2017.10.011

Mooij, J. & Magliacane, S. & Claassen, T.. (2020). *Joint causal inference from multiple contexts*. The Journal of Machine Learning Research, Vol.21, n.º1, pp.3919-4026 ; consultado em 10-10-2025, disponível em: https://jmlr.org/papers/volume21/17-123/17-123.pdf

Nauta, M. & Trienes, J. & Pathak, S. & Nguyen, E. & Peters, M. & Schmitt, Y. & Schlötterer, J. & Van Keulen, M. & Seifert, C.. (2022). *From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI*; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/2201.08164

Nico K.. (February 13, 2024).*The 3 Types of Artificial Intelligence: ANI, AGI, and ASI*. Viso.ai; consultado em 10-10-2025, disponível em: https://viso.ai/deep-learning/artificial-intelligence-types/

Nogueira, A. & Pugnana, A. & Ruggieri, S. & Pedreschi, D. & Gama, J.. (2022). *Methods and tools for causal discovery and causal inference.* WIREs Data Mining and Knowledge Discovery., 12, e1449; consultado em 10-10-2025, disponível em: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1449

Notice by the Patent and Trademark Office on 02/13/2024, *Inventorship Guidance for AI-Assisted Inventions;* consultado em 10-10-2025, disponível em: https://www.federalregister.gov/documents/2024/02/13/2024-02623/inventorship-guidance-for-ai-assisted-inventions

Novo Código da Propriedade Industrial (2018) - Decreto-Lei n.º 110/2018, de 10 de dezembro; consultado em 10-10-2025, disponível em: https://diariodarepublica.pt/dr/detalhe/decreto-lei/110-2018-117279933

OCDE (Organização para a Cooperação e Desenvolvimento Económico). *AI Principles*. (2019); consultado em 10-10-2025, disponível em: https://www.oecd.org/en/topics/ai-principles.html

Ohno, N. & Tsunematsu & Tonomura, K. & Matsuzaki, Y. & Kondo, M. (May 12, 2025). Legal 500. *AI Update – AI Inventorship: IP High Court in Japan Rules AI Cannot Be Listed as Inventor*; consultado em 10-10-2025, disponível: https://www.legal500.com/developments/thought-leadership/ai-update-ai-inventorship-ip-high-court-in-japan-rules-ai-cannot-be-listed-as-inventor/

Optimizing Human-AI Collaboration: A Guide to HITL, HOTL, and HIC Systems; consultado em 10-10-2025, disponível em: https://www.deepscribe.ai/resources/optimizing-human-ai-collaboration-a-guide-to-hitl-hotl-and-hic-systems

Patent Journal Including Trade Marks, Designs and Copyright in Cinematograph Films (2021), Vol.54 n. º7 (p.255). Food *Container and Devices and Methods for Attracting Enhanced Attention*; consultado em 10-10-2025, disponível em: https://iponline.cipc.co.za/Publications/PublishedJournals/E_Journal_July%202021%20Part%202.pdf

Patent Law of the People's Republic of China - Lei de Patentes da República Popular da

China, (China National Intellectual Proprerty Administration); consultado em 10-10-2025, disponível em: https://english.cnipa.gov.cn/transfer/lawpolicy/patentlawsregulations/915574.htm

Pearl, J.. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, England: Cambridge University Press; consultado em 10-10-2025, disponível em https://bayes.cs.ucla.edu/BOOK-2K/

Pearl, J. & Mackenzie, D.. (2018). *The book of why*. New York, NY: Basic Books; consultado em 10-10-2025, disponível em: https://scholar.google.com/scholar_lookup?hl=en&publication_year=2018&author=J.+Pearl&author=D.+Mackenzie&title=The+book+of+why

Peixoto, S.. (2024, p.24). *O requisito da atividade inventiva da invenção patenteável e as invenções geradas por inteligência artificial*. Edições Almedina, S.A.. ISBN 978-989-40-2168-1; consultado em 10-10-2025.

Peng, S. & Kalliamvakou, E. & Cihon, P. & Demirer, M.. (2023).*The impact of ai on developer productivity: Evidence from github copilot*; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/2302.06590

Pereira, A.. (2001, p.9-10), *Patentes de Software: Sobre a Patenteabilidade dos Programas de computador*; consultado em 10-10-2025, disponível em: https://estudogeral.uc.pt/handle/10316/89502

Perry, R. & Von Kügelgen, J. & Schölkopf, B.. (2022). *Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis*. In Neural Information Processing Systems; consultado em 10-10-2025, disponível em: https://arxiv.org/abs/2206.02013

Peters, J. & Bühlmann, P. & Meinshausen, N.. (2016). *Causal inference by using invariant prediction: identification and confidence intervals*. Journal of the Royal Statistical Society. Series B (Statistical Methodology), pp.947-1012 ; consultado em 10-10-2025, disponível em: https://people.math.ethz.ch/~peterbu/Files/Manuscripts/invariant-causal-prediction.pdf

Peters, J. & Janzing, D. & Schölkopf, B.. (2017). *Elements of causal inference: Foundations and learning algorithms*. Cambridge, MA: MIT-Press; consultado em 10-10-2025, disponível em: https://library.oapen.org/bitstream/handle/20.500.12657/26040/11283.pdf?sequ

Rajendran, G. & Buchholz, S. & Aragam, B. & Schölkopf, B. & Ravikumar, P.. (2024). *Learning interpretable concepts: Unifying causal representation learning and foundation models* ; consultado em 10-10-2025, disponível em: https://arxiv.org/abs/2402.09236

Ranjan, S. & Konstantinos, I. & Manoj, K.. (28 de maio 2025). *AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges*. Cornell University, Department of Biological and Environmental Engineering, USA & University of the Peloponnese, Department of Informatics and Telecommunications, Tripoli, Greece ; consultado em 10-10-2025, disponível em: https://doi.org/10.48550/arXiv.2505.10468

Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016 - Regulamento Geral sobre a Proteção de Dados (RGPD). Relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados e que revoga a Diretiva 95/46/CE (Regulamento Geral sobre a Proteção de Dados); consultado em 10-10-2025, disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:02016R0679-20160504

Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que cria regras harmonizadas em matéria de inteligência artificial e que altera os Regulamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e as Diretivas 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (Regulamento da Inteligência Artificial); consultado em 10-10-2025, disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX:32024R1689

Report / Study- Ethics guidelines for trustworthy AI. (2019). Consultado em 10-10-2025, disponível em: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Retzlaff, C. & Das, S. & Wayllace, C. & Mousavi, P. & Afshari, M. & Yang, T. & Saranti, A. & Angerschmid, A. & Taylor M. & Holzinger, A.. (2024-a*). Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities.* Journal of Artificial Intelligence Research (JAIR), Vol.79, n.º1, pp.349-415 ; consultado em 10-10-2025, disponível em: https://jair.org/index.php/jair/article/view/15348

Retzlaff, C. & Angerschmidt, A. & Saranti, A. & Schneeberger, D. & Röttger, R. & Müller, H. & Holzinger, A.. (2024–b); *Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists.* consultado em 10-10-2025, disponível

em: https://doi.org/10.1016/j.cogsys.2024.101243

Ribeiro, M. & Singh, S. & Guestrin, C.. (2016). *Why should I trust you? Explaining the predictions of any classifier*. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135-1144; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/1602.04938

Ronald Y. & Kenneth Y.. (2021, pp. 486–489). *New Changes, New Possibilities: China's Latest Patent Law Amendments*. GRUR International , Volume 70, Edição 5, maio de 2021; consultado em 10-10-2025, disponível em: https://doi.org/10.1093/grurint/ikaa201

Roscher, R. & Bohn, B. & Duarte, M. & Garcke, J.. (2020). *Explainable Machine Learning for Scientific Insights and Discoveries.* IEEE Access , vol. 8, pp.42200-42216. doi: 10.1109/ACCESS.2020.2976199; consultado em 10-10-2025, disponível em: https://ieeexplore.ieee.org/document/9007737

Rudin, C.. (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Machine Intelligence, Vol.1, n.º 5, pp.206-215; consultado em 10-10-2025, disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC9122117/pdf/nihms-1058031.pdf

Safavian, S. & Landgrebe, D.. (1991).*A survey of decision tree classifier methodology*. IEEE Transactions on Systems, Man, and Cybernetics, Vol.21, n.º3, pp.660-674 ; consultado em 10-10-2025, disponível em: https://ntrs.nasa.gov/api/citations/19910014493/downloads/19910014493.pdf

Sajid A.& Tamer A.& Shaker E.& Khan M.& Jose M. Alonso-Moral & Roberto C.& Riccardo G.& Javier Del Seri & Natalia Díaz-Rodríguez & Francisco H.. (2023, pp.1-2). *Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence*; consultado em 10-10-2025, disponível em: https://doi.org/10.1016/j.inffus.2023.101805

Schneeberger, D. & Stoeger, K. & Holzinger, A.. (2020). *The European legal framework for medical AI*. Machine learning and knowledge extraction, CD-MAKE, Lecture notes in computer science LNCS, vol. 12279, Springer International, Cham (2020), pp.209-226 ; consultado em 10-10-2025, disponível em: https://inria.hal.science/hal-03414719/file/497121_1_En_12_Chapter.pdf

Selvaraju, R. & Cogswell, M. & Das, A. & Vedantam, R. & Parikh, D. & Batra, D.. (2017). *Grad-Cam: Visual explanations from deep networks via gradient-based localization*. In IEEE International Conference on Computer Vision ; consultado em 10-10-2025, disponível em: https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf

Shafi, A.. (2021). *What is a Generalised Additive Model?.* Towards Data Science; consultado em 10-10-2025, disponível em: https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a/

Simonyan, K. & Vedaldi, A. & Zisserman, A.. (2014). *Deep inside convolutional networks: Visualising image classification models and saliency maps*. Proceedings of ICLR Workshop, ICLR, Banff, Canada; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/1312.6034

Slack, D. & Hilgard, S. & Jia, E. & Singh,S. & Lakkaraju, H.. (2020). *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods*. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp.180-186; consultado em 10-10-2025, disponível em: https://dl.acm.org/doi/pdf/10.1145/3375627.3375830

Stierle, M.. (2021, pp.115-133). *A De Lege Ferenda Perspective on Artificial Intelligence Systems Designated as Inventors in the European Patent System*; consultado em 10-10-2025, disponível em: https://doi.org/10.1093/grurint/ikaa186

Stoeger, K. & Schneeberger, D. & Holzinger, A.. (2021). *Medical artificial intelligence: The European legal perspective*. Communications of the ACM, Vol.64, n.º 11, pp.34-36; consultado em 10-10-2025, disponível em: https://dl.acm.org/doi/pdf/10.1145/3458652

Sudjianto, A. & Zhang, A.. (2021, p.1). *Designing Inherently Interpretable Machine Learning Models*; consultado em 10-10-2025, disponível em: https://ar5iv.labs.arxiv.org/html/2111.01743

Stryker, C. & Holdsworth, J.. (2024).*What is NLP (natural language processing)?* ; consultado em 10-10-2025, disponível em: https://www.ibm.com/think/topics/natural-language-processing

Taehyun, H. & Sangwon, L. & Sangyeon, K.. (2018). *Designing explainability of an artificial intelligence system.* Proceedings of the Technology, Mind, and Society, ACM, Washington, District of Columbia, USA , p. 14:1; consultado em 10-10-

2025, disponível em : https://doi.org/10.1145/3183654.3183683

Templeton, A. & Conerly, T. & Marcus, J. & Lindsey, J. & Bricken, T. & Chen, B. & Pearce, A. & Citro, C. & Ameisen, E. & Jones, A. & Cunningham, H. & Turner, N. & McDougall, C. & MacDiarmid, M. & Freeman, C. & Sumers, T. & Rees, E. & Batson, J. & Jermyn, A. & Carter, S. & Olah, C. & Henighan, T.. (2024). *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Transformer Circuits Thread* ; consultado em 10-10-2025, disponível em: https://transformer-circuits.pub/2024/scaling-monosemanticity/

Tribunal de Justiça da União Europeia [TJUE] - Acórdão 27 fevereiro 2025, C-65/23; Processo C- 203/22 – Caso CK GmbH contra *Magistrat der Stadt Wien*; consultado em 10-10-2025, disponível em: https://curia.europa.eu/juris/document/document.jsf?text=&docid=295841&pageIndex=0&doclang=PT&mode=req&dir=&occ=first&part=1&cid=7174952

Varici, B. & Acartürk, E. & Shanmugam, K. & Tajer, A.. (2024). *General identifiability and achievability for causal representation learning*. In Artificial Intelligence and Statistics ; consultado em 10-10-2025, disponível em: https://arxiv.org/abs/2310.15450

Venâncio, P.. (2013, p.90). *A Tutela Jurídica do Formato de Ficheiro Electrónico*; [Tese de Doutoramento em Ciências Jurídicas Especialidade em Ciências Jurídico Privatísticas]. Universidade do Minho - Escola de Direito; consultado em 10-10-2025, disponível em: https://repositorium.sdum.uminho.pt/bitstream/1822/35678/1/Pedro%20Dias%20Ven%C3%A2ncio.pdf

Venâncio, P.. (2023, pp. 49,52 e 61). *Questões em Torno da Relação Entre Patentes & Sistemas de Inteligência Artificial*. ULP Law Review, Vol. 17, n.º 2; consultado em 10-10-2025, disponível em: https://doi.org/10.60543/ul-plr-rdul-p.v17i2.9520

Vilone, G. & Longo, L.. (2021). *Notions of explainability and evaluation approaches for explainable artificial intelligence*. Information Fusion, vol. 76, pp.89-106; consultado em 10-10-2025, disponível em: https://doi.org/10.1016/j.inffus.2021.05.009

Wang, D. & Yang, Q. & Abdul, A. & Lim, B.. (2019). *Designing theory-driven user-centric explainable AI.* CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK. ACM, New York, NY, USA, pp. 1-15; consultado em 10-10-2025, disponível em: https://ubiquitous.comp.nus.edu.sg/wp-content/uploads/2019/01/chi2019-reasoned-xai-framework.pdf

White House Office of Science and Technology Policy. *Blueprint for an AI Bill of Rights, Making Automated Systems Work for the American People*; consultado em 10-10-2025, disponível em: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/

Weerts, H. & Van Ipenburg, W. & Pechenizkiy M.. (2019). *A human-grounded evaluation of SHAP for alert processing*; consultado em 10-10-2025, disponível em: https://arxiv.org/pdf/1907.03324

Weller, A.. (2017). *Challenges for transparency. Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*. ICML, Sydney, Australia, pp.55-62; consultado em 10-10-2025, disponível em: https://openreview.net/pdf?id=SJR9L5MQ-

Xu, Y. & Zhu, J. & Chang, E. & Lai, M. & Tu, Z.. (2014). *Weakly supervised histopathology cancer image segmentation and classification.* Medical Image Analysis, n.º 18, pp.591-604; consultado em 10-10-2025, disponível em: https://www.microsoft.com/en-us/research/wp-content/uploads/2014/02/2014SCIMIAWeakly-Supervised-Histopathology-Cancer-Image-Segmentation-and-Classification.pdf

Yang, Z. & Zhang, A. & Sudjianto, A.. (2021). *Enhancing Explainability of Neural Networks Through Architecture Constraints*. IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, n.º 6 (2021), pp.2610-2621; consultado em 10-10-2025, disponível em: DOI: 10.1109/TNNLS.2020.3007259

Yunji, C. & Ling, L. & Wei, L. & Qi, G. & Zidong, D. & Zichen X.. (2024, Cap. 1.1.1.). *AI Computing Systems*; consultado em 10-10-2025, disponível em: https://www.sciencedirect.com/topics/computer-science/artificial-general-intelligence

Zhang, A. & Chen, Y. & Sheng, L. & Wang, X. & Chua, T.. (2024). *On generative agents in recommendation.* In Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval, pp. 1807–1817; consultado em 10-10-2025, disponível em: https://dl.acm.org/doi/pdf/10.1145/3626772.3657844

Zhu, X.. (2008).*Semi-Supervised Learning Literature Survey*. Computer Science TR 1530, University of Wisconsin-Madison; consultado em 10-10-2025, disponível em: https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf