Artificial Intelligence: Challenges for Research, Scientific Outcomes, and Academic Publishing

Inteligência Artificial: Desafios para a Investigação, os Resultados Científicos e a Publicação Académica

José Machado[1,2]

Abstract:

Within just a few years, Artificial Intelligence (AI) has become one of the most transformative forces in contemporary science. Its impact reaches far beyond the technological domain, extending to how we conduct research, write, validate results, communicate knowledge, and assess scientific output. The central question of this reflection is how to integrate AI into research and scientific publishing without compromising fundamental values such as truth, rigour, transparency, accountability, and fairness.

Keywords: Artificial Intelligence; Ethics, Research; Generative Artificial Intelligence; Publishing; Scientific Misconduct; Scholarly Communication.

Resumo:

A Inteligência Artificial (IA) tornou-se, em poucos anos, uma das forças mais transformadoras da ciência contemporânea. O seu impacto ultrapassa largamente o domínio tecnológico e estende-se à forma como investigamos, escrevemos, validamos resultados, comunicamos conhecimento e avaliamos a produção científica. A questão central desta reflexão é como integrar a IA na investigação e na publicação científica sem comprometer valores fundamentais como a verdade, o rigor, a transparência, a responsabilidade e a equidade.

Palavras-chave: Comunicação Científica; Ética na Investigação; Inteligência Artificial; Inteligência Artificial Generativa; Má Conduta Científica; Publicação.

A IA como Ferramenta Científica e Desafio Académico

A utilização da Inteligência Artificial (IA) na ciência e na academia tem crescido de forma exponencial (Stokel-Walker & Van Noorden, 2023).[1] Hoje, ferramentas baseadas em IA são utilizadas para analisar grandes volumes de dados, apoiar revisões da literatura, gerar hipóteses, otimizar modelos, escrever textos, traduzir artigos, resumir documentos e até auxiliar processos de revisão científica. Esta realidade cria oportunidades muito relevantes, mas também levanta problemas éticos, técnicos e metodológicos que não podem ser ignorados.

A IA pode ser definida, de forma simples, como o estudo da modelação computacional de funções mentais humanas (Russell & Norvig, 2021).[2] Esta definição ajuda a compreender que a IA não é apenas um conjunto de programas sofisticados, mas uma tentativa de representar, automatizar ou simular capacidades associadas à inteligência: reconhecer padrões, aprender com exemplos, tomar decisões, interpretar linguagem, formular respostas e adaptar-se a contextos diferentes.

Na prática, uma forma simples de compreender a IA é vê-la como a combinação entre dados e algoritmos. Os dados fornecem a matéria-prima. Os algoritmos extraem regularidades, identificam padrões, aprendem relações e produzem previsões ou decisões. Esta visão é particularmente importante porque mostra que a IA não surge do vazio. O seu comportamento depende dos dados com que foi treinada, das arquiteturas utilizadas, dos objetivos definidos e dos contextos em que é aplicada.

A investigação científica vive hoje um momento de transição. Durante décadas, os investigadores utilizaram essencialmente ferramentas estatísticas, bases de dados bibliográficas, processadores de texto e *software* especializado. Atualmente, os sistemas de IA generativa e os grandes modelos de linguagem introduzem uma nova camada de mediação entre o investigador e o conhecimento. Estes sistemas não apenas processam informação: produzem texto, sugerem interpretações, organizam argumentos, simulam raciocínios e intervêm diretamente no processo de criação científica.

Esta mudança obriga a repensar práticas estabelecidas. O que significa escrever um artigo científico com apoio de IA? Como se deve declarar esse apoio? Até que ponto uma hipótese sugerida por um modelo pode ser considerada original? Como garantir que os resultados produzidos por sistemas automáticos são reprodutíveis? Que responsabilidade tem o investigador quando utiliza uma ferramenta que pode gerar informação incorreta, enviesada ou não verificável?

[1]ALGORITMI/LASI
[2]Universidade do Minho, Braga,Portugal

https://doi.org/10.24950/rspmi.2880

Estas perguntas mostram que o problema não é simplesmente técnico. A questão central não é apenas saber se a IA funciona bem, mas saber em que condições pode ser usada de forma responsável, transparente e cientificamente legítima.

Da Aprendizagem Automática à IA Generativa

Para compreender os desafios atuais, é necessário começar pelos fundamentos. Muitos sistemas de IA modernos assentam em métodos de aprendizagem automática, nos quais o sistema aprende a partir de exemplos. Um caso simples é o da regressão linear: a partir de dados sobre a área de casas e o respetivo preço, um modelo pode aprender uma relação aproximada entre essas variáveis e prever o preço de uma nova casa. O exemplo é simples, mas ilustra uma ideia essencial: o modelo não é programado com todas as respostas; aprende uma relação a partir dos dados.

Este princípio está na base de sistemas muito mais complexos. Nas redes neuronais artificiais, a aprendizagem ocorre através da adaptação de pesos e parâmetros. Uma rede neuronal é um modelo computacional inspirado, de forma abstrata, no funcionamento do cérebro humano. É composta por unidades simples, chamadas neurónios artificiais, organizadas em camadas. Cada neurónio recebe valores de entrada, multiplica-os por pesos, soma esses valores, adiciona um termo de ajustamento chamado viés e aplica uma função de ativação. O resultado é transmitido para a camada seguinte (Goodfellow, *et al*, 2016).[3]

Embora esta descrição seja simplificada, ajuda a compreender a lógica básica da aprendizagem. Os pesos representam aquilo que o modelo vai ajustando durante o treino. No início, esses pesos podem ser aleatórios ou pouco informativos. À medida que o modelo processa exemplos e compara as suas previsões com as respostas corretas, calcula o erro cometido. Esse erro é depois propagado para trás através de um processo chamado retropropagação, permitindo ajustar os pesos de forma a reduzir erros futuros. Este ciclo repete-se muitas vezes, por vezes milhões ou milhares de milhões de vezes, até que o modelo adquira uma capacidade razoável de generalização.

Na IA generativa, as redes neuronais aprendem padrões complexos a partir de grandes volumes de dados e utilizam esses padrões para criar novos conteúdos de forma coerente e contextual. Isto significa que o sistema não se limita a classificar, prever ou reconhecer; ele gera texto, imagens, código, música, resumos, explicações e respostas. Esta capacidade é especialmente relevante no contexto académico porque aproxima a IA de atividades tradicionalmente humanas: escrever, argumentar, interpretar e explicar.

No entanto, esta capacidade generativa também cria riscos. Um sistema pode produzir uma resposta fluente, gramaticalmente correta e aparentemente convincente, mas factualmente errada. Pode inventar referências, distorcer conceitos, omitir limitações ou reproduzir preconceitos presentes nos dados de treino. A fluência textual não deve, portanto, ser confundida com verdade científica (Bender *et al*, 2021).[4]

Redes Neuronais, Parâmetros e Conhecimento do Modelo

Um dos conceitos centrais para compreender os sistemas modernos de IA é o de parâmetro. Numa rede neuronal, os parâmetros correspondem essencialmente aos pesos e vieses que são ajustados durante o treino. São estes valores que determinam a forma como o modelo transforma entradas em saídas. Quando se diz que um modelo tem 70 mil milhões de parâmetros, significa que possui 70 mil milhões de números ajustáveis que codificam padrões aprendidos durante o treino.

Esta ideia é importante porque ajuda a perceber como o conhecimento é representado nesses sistemas. O modelo não possui conhecimento da mesma forma que uma pessoa, nem armazena informação como uma base de dados tradicional. O seu conhecimento está distribuído por uma enorme rede de parâmetros. Por isso, quando responde a uma pergunta, não vai simplesmente procurar uma frase guardada num ficheiro; gera uma resposta com base nos padrões estatísticos e semânticos que aprendeu.

Isto tem consequências importantes. Por um lado, permite uma grande flexibilidade. O modelo consegue responder a perguntas variadas, adaptar-se a estilos diferentes, resumir textos, traduzir, explicar conceitos e combinar informação de formas novas. Por outro lado, torna difícil identificar exatamente por que razão o modelo produziu uma determinada resposta. Esta dificuldade está associada ao chamado problema da "caixa negra".

Os sistemas de caixa negra são sistemas cujo funcionamento interno é difícil de compreender ou explicar. O utilizador observa a entrada e a saída, mas não consegue acompanhar de forma transparente todos os passos que conduziram à decisão. Em domínios de baixo risco, isto pode ser aceitável. Mas em domínios regulados, como a medicina, a justiça, a administração pública ou a avaliação científica, a opacidade é problemática. A confiança exige explicação, auditoria e responsabilização.

A Evolução das Arquiteturas: de MLP a *Transformers*

As arquiteturas de redes neuronais evoluíram significativamente ao longo das últimas décadas. As redes MLP, ou *Multilayer Perceptrons*, constituem uma das formas mais clássicas de rede neuronal. São compostas por camadas densamente ligadas, nas quais cada neurónio de uma camada se liga aos neurónios da camada seguinte. São úteis para muitos problemas, especialmente dados tabulares,

mas não têm uma noção natural de ordem ou sequência.

As redes neuronais recorrentes, conhecidas como RNNs, introduziram a ideia de memória. Em vez de processarem cada entrada de forma isolada, mantêm um estado oculto que passa de elemento para elemento numa sequência. Isto tornou-as relevantes para texto, fala e séries temporais. No entanto, as RNNs tinham dificuldades em lidar com dependências longas. Em textos extensos, a informação inicial podia perder-se ao longo do processamento.

As LSTMs, ou *Long Short-Term Memory networks*, surgiram para resolver parcialmente esse problema (Hochreiter & Schmidhuber, 1997).[5] Introduziram uma célula de memória explícita e mecanismos de controlo, chamados *gates*, que regulam que informação deve ser mantida, esquecida ou atualizada (Gers *et al*, 2000).[6] Durante muitos anos, as LSTMs foram uma arquitetura dominante no processamento de linguagem natural. Ainda assim, continuavam a processar sequências de forma essencialmente linear, palavra a palavra, o que limitava a paralelização e dificultava o tratamento eficiente de textos muito longos.

A grande mudança ocorreu em 2017 com a arquitetura *Transformer*, apresentada no artigo *Attention is All You Need*. O *Transformer* abandonou a recorrência e colocou o mecanismo de atenção no centro do processamento. Em vez de processar palavras uma a uma, permite considerar simultaneamente todas as palavras de uma sequência e aprender quais as relações mais relevantes entre elas, independentemente da distância no texto (Vaswani *et al*, 2017).[7]

Este avanço foi decisivo para o desenvolvimento dos grandes modelos de linguagem. O mecanismo de atenção permite que o modelo identifique relações contextuais complexas: que palavra se relaciona com que conceito, que parte do texto é mais relevante para uma determinada interpretação, que elementos anteriores condicionam a resposta atual. Através deste mecanismo, os modelos conseguem representar o contexto de forma mais rica e produzir texto mais coerente.

Num *Transformer*, cada bloco contém essencialmente dois componentes principais: o mecanismo de atenção e uma rede MLP *feed-forward*. A atenção utiliza estruturas designadas por *Query, Key, Value e Output*. De forma intuitiva, a *Query* representa aquilo que um *token* procura; a *Key* representa aquilo que outros *tokens* oferecem; a *Value* contém a informação a transmitir; e a projeção de saída reorganiza o resultado. Depois da atenção, cada *token* passa por uma camada MLP, que funciona como uma forma de memória associativa.

É relevante notar que, em muitos modelos de grande escala, uma parte substancial dos parâmetros se encontra nas camadas MLP. Isto sugere que grande parte do conhecimento factual e associativo do modelo está codificado nessas estruturas. A atenção organiza o contexto e as relações; as camadas MLP armazenam e transformam padrões aprendidos durante o treino.

Grandes Modelos de Linguagem

Os grandes modelos de linguagem, ou LLMs, são modelos de IA concebidos para compreender e gerar texto semelhante ao humano. São treinados com volumes massivos de texto e possuem, frequentemente, milhares de milhões ou até biliões de parâmetros. Aprendem padrões linguísticos, sintáticos, semânticos e pragmáticos. Com base nesses padrões, conseguem gerar respostas, resumir documentos, traduzir textos, explicar conceitos, responder a perguntas, produzir código e auxiliar tarefas de escrita (Brown *et al*, 2020).[8]

A sua base tecnológica é normalmente a arquitetura *Transformer*. Os LLMs representam o contexto através de vetores, mecanismos de atenção e estados internos. Cada palavra ou fragmento de palavra é transformado num vetor num espaço multidimensional. A partir daí, o modelo calcula relações entre *tokens* e gera a continuação mais provável ou mais adequada de acordo com o contexto.

Nos últimos anos, surgiram vários modelos relevantes. Modelos como Claude, da Anthropic, destacam-se pela capacidade de seguir instruções e produzir respostas longas e coerentes. Modelos como GPT-4o e o3, da OpenAI, combinam capacidades de raciocínio, linguagem e multimodalidade. O Gemini, da Google, tem sido associado a janelas de contexto longas e integração com ecossistemas de informação. Llama, da Meta, representa uma linha de modelos abertos ou semiabertos, com impacto significativo na investigação e desenvolvimento local. O AMALIA, desenvolvido no contexto português, surge como uma aposta estratégica na língua portuguesa europeia e na soberania digital.

A importância dos LLMs para a ciência reside na sua capacidade de lidar com informação textual intensiva. A ciência é, em grande medida, uma prática textual: artigos, relatórios, protocolos, revisões, pareceres, propostas, projetos, atas, comunicações e avaliações. Quando um sistema consegue processar e gerar texto científico, passa a interferir diretamente no ciclo de produção do conhecimento.

Oportunidades da IA na Ciência

A IA oferece oportunidades muito relevantes para a investigação científica. Uma das mais evidentes é a automação parcial da revisão da literatura. A quantidade de artigos publicados anualmente tornou-se difícil de acompanhar. Ferramentas baseadas em IA podem ajudar a identificar trabalhos relevantes, agrupar temas, resumir contributos, comparar abordagens e detectar lacunas. Isto não substitui a leitura crítica do investigador, mas pode acelerar a fase exploratória e tornar o processo mais sistemático.

Outra oportunidade é a geração de hipóteses a partir de grandes volumes de dados. Em áreas como medicina, biologia, ciências sociais, engenharia, economia ou ambiente, existem bases de dados cada vez maiores e mais heterogéneas. A IA pode identificar correlações, padrões emergentes ou relações não triviais que escapariam a uma análise

manual. A partir daí, o investigador pode formular hipóteses, desenhar experiências e validar resultados.

A escrita assistida é também uma área de impacto imediato. A IA pode apoiar a organização de ideias, a melhoria da clareza textual, a tradução, a adaptação do estilo, a criação de resumos, a revisão gramatical e a preparação de versões para públicos diferentes. Para investigadores que escrevem em inglês como segunda língua, estas ferramentas podem reduzir desigualdades comunicativas e melhorar a qualidade formal dos textos.

A IA pode ainda apoiar a análise estatística, a otimização de modelos, a programação, a depuração de código, a construção de gráficos e a interpretação de resultados. Em muitos casos, atua como um assistente técnico que acelera tarefas repetitivas ou ajuda a ultrapassar barreiras de especialização.

Contudo, estas oportunidades não eliminam a responsabilidade do investigador. Pelo contrário, aumentam a necessidade de supervisão crítica. A IA pode sugerir caminhos, mas não deve substituir o juízo científico. Pode propor uma hipótese, mas não a valida. Pode resumir um artigo, mas pode omitir aspectos importantes. Pode gerar código, mas esse código pode conter erros. Pode formular uma explicação, mas essa explicação pode ser superficial ou incorreta.

Desafios Metodológicos

Os desafios metodológicos são talvez os mais críticos para a integridade científica. O primeiro é o problema da caixa negra. Quando um sistema produz uma classificação, uma previsão, uma recomendação ou um texto, pode ser difícil compreender o processo interno que levou a esse resultado. Isto é particularmente problemático quando a IA é usada para apoiar decisões científicas, clínicas ou jurídicas.

O segundo desafio é o viés dos dados de treino. Os modelos aprendem com dados históricos. Se esses dados contêm erros, desigualdades, preconceitos ou lacunas, o modelo pode reproduzi-los ou amplificá-los. Por exemplo, um sistema treinado com literatura científica dominante pode subvalorizar áreas menos representadas, países periféricos, línguas minoritárias ou abordagens alternativas. No contexto da publicação científica, isto pode reforçar assimetrias já existentes.

O terceiro desafio é a reprodutibilidade. A ciência depende da possibilidade de repetir procedimentos e verificar resultados. Quando se utiliza IA, especialmente sistemas proprietários e modelos em constante atualização, a reprodutibilidade torna-se mais difícil. O mesmo pedido feito em momentos diferentes pode produzir respostas distintas. O modelo pode ter sido atualizado. A configuração pode não ser conhecida. Os dados de treino podem ser inacessíveis. Os parâmetros internos são normalmente impossíveis de inspecionar.

O quarto desafio é a ocorrência de falsos positivos e falsos negativos em análises automatizadas. Um sistema pode identificar erradamente uma relação estatística como relevante, ignorar uma relação importante, classificar incorretamente documentos ou gerar interpretações não suportadas pelos dados. Se o investigador aceitar esses resultados sem validação, o erro pode entrar no processo científico e contaminar conclusões.

Estes desafios mostram que a IA deve ser integrada com protocolos claros. É necessário registar que ferramentas foram usadas, em que fases, com que versões, com que parâmetros, para que objetivos e com que tipo de validação humana. Sem esta documentação, o uso de IA pode comprometer a rastreabilidade científica.

Desafios Éticos

Os desafios éticos são igualmente profundos. Um dos primeiros é a autoria. Se uma ferramenta de IA contribui para a redação de um artigo, quem é o autor? A maioria das posições editoriais rejeita a possibilidade de atribuir autoria a sistemas de IA, porque estes não podem assumir responsabilidade científica, responder por erros ou comprometer-se com a integridade do trabalho (Andrade *et al*, 2007).[9] Ainda assim, quando a contribuição da IA é significativa, deve ser declarada.

A questão do plágio também se torna mais complexa. A IA pode gerar texto original no sentido de não copiar diretamente uma fonte específica, mas esse texto resulta de padrões aprendidos a partir de obras existentes. Pode ainda reproduzir frases, ideias ou estruturas sem atribuição clara. O risco não é apenas copiar texto, mas produzir uma aparência de originalidade sem verdadeira contribuição intelectual.

Outro problema é a integridade científica. A IA pode ser usada de forma antiética para gerar resultados falsos, fabricar dados, criar citações inexistentes, produzir revisões superficiais, manipular imagens ou multiplicar publicações de baixa qualidade. A facilidade de produção textual pode contribuir para o aumento de práticas questionáveis, como artigos redundantes, submissões massivas ou revisões automáticas sem profundidade.

Existe ainda a questão da desigualdade de acesso. Nem todos os investigadores têm acesso às mesmas ferramentas, aos mesmos modelos, às mesmas infraestruturas computacionais ou às mesmas versões pagas. Isto pode criar novas assimetrias entre instituições, países e áreas científicas. A IA pode democratizar certas capacidades, mas também pode aprofundar desigualdades se o acesso for determinado por recursos económicos.

A privacidade e a proteção de dados são igualmente centrais. Em muitos domínios, especialmente saúde, justiça, educação e administração pública, os dados são sensíveis. A utilização de ferramentas externas pode implicar o envio de informação para servidores fora do controlo institucional. A justificação ética para recolher e usar dados deve estar associada ao benefício das pessoas, à minimização de riscos e ao respeito por direitos fundamentais. A estes desafios acresce a dificuldade do *machine unlearning*:

a capacidade de fazer um modelo esquecer dados específicos após o treino permanece um problema em aberto, sem garantias de remoção verificável nem métodos auditáveis, o que coloca em tensão direta o direito ao esquecimento previsto no RGPD com a arquitetura dos atuais grandes modelos de linguagem (Bourtoule *et al*, 2021).[10]

Usos Éticos e Antiéticos da IA

Uma forma prática de enquadrar a questão ética é distinguir entre usos aceitáveis e usos problemáticos. Na escrita, é eticamente aceitável utilizar IA para discutir ideias, melhorar gramática, clarificar frases ou adaptar linguagem. Já é problemático usar IA para gerar integralmente um livro, artigo ou relatório e apresentá-lo como trabalho humano original sem declaração.

Na programação, é aceitável usar IA para encontrar erros, explicar código, sugerir melhorias ou acelerar tarefas técnicas. É antiético e potencialmente criminoso utilizá-la para criar *malware*, explorar vulnerabilidades ou automatizar ataques.

Na investigação, pode ser útil usar IA para resumir textos, organizar bibliografia, identificar temas ou apoiar a análise exploratória. Mas é claramente antiético utilizá-la para criar citações falsas, inventar resultados, fabricar revisões ou simular conhecimento inexistente.

A tecnologia, em si mesma, pode ser considerada neutra em termos instrumentais. Mas o uso da tecnologia e dos dados nunca é neutro. As decisões sobre objetivos, dados, contextos, interpretação e consequências são humanas e institucionais. Por isso, a responsabilidade não desaparece com a automatização. Pelo contrário, torna-se mais distribuída e exige mecanismos mais claros de governação, incluindo abordagens formais para a tomada de decisão ética em sistemas de IA (Machado *et al*, 2024).[11]

Há ainda uma dimensão que merece particular atenção. Durante séculos, cada nova ferramenta, da imprensa ao computador, passando pelos motores de busca, alargou as capacidades cognitivas humanas. Com os grandes modelos de linguagem, observa-se, pela primeira vez, uma possível inversão desta trajetória: muitos utilizadores deixam de querer compreender como se resolvem os problemas e passam a exigir apenas as soluções. O processo de raciocínio, antes parte essencial da formação intelectual, corre o risco de ser delegado à máquina, com consequências ainda mal compreendidas para a aprendizagem, a autonomia crítica e a própria prática científica. Pela primeira vez na história recente, a tecnologia pode estar a reduzir, e não a ampliar, a inteligência humana.

IA e Publicação Científica

A publicação científica enfrenta desafios específicos. Um deles é a dificuldade em identificar textos gerados por IA. Os detetores automáticos de texto gerado por IA são imperfeitos. Podem produzir falsos positivos, acusando injustamente autores humanos, especialmente não nativos de inglês. Também podem produzir falsos negativos, deixando passar textos gerados ou fortemente assistidos por IA.

Outro desafio é a falta de normalização das orientações editoriais. Algumas revistas exigem declaração explícita do uso de IA. Outras proíbem IA como coautora (*Nature Editorial*, 2023).[12] Algumas aceitam o uso para revisão linguística, mas não para geração de conteúdo científico. Outras ainda estão a desenvolver políticas. Esta diversidade cria incerteza para autores, revisores e editores.

O processo de revisão por pares também é afetado. Um revisor pode usar IA para resumir um artigo ou apoiar a formulação de comentários. Mas pode também delegar indevidamente a avaliação, expondo material confidencial a ferramentas externas ou produzindo revisões genéricas. Além disso, a revisão por pares pode não ser suficiente para detetar fraude automatizada, citações inventadas ou manipulação de resultados.

Existe o risco de normalização de práticas questionáveis. Se a produção científica se tornar demasiado dependente da geração automática de texto, pode aumentar a quantidade de publicações sem correspondente aumento de conhecimento. A pressão para publicar, combinada com ferramentas capazes de gerar rapidamente manuscritos, pode agravar problemas já existentes no sistema científico.

Neste contexto, as revistas e conferências terão de desenvolver políticas mais claras. Essas políticas devem indicar quando a IA pode ser usada, como deve ser declarada, que responsabilidades permanecem nos autores e que limites se aplicam à revisão, à análise de dados e à produção de conteúdo.

Caminhos Possíveis: Transparência, Formação e Políticas Institucionais

A resposta aos desafios da IA não deve ser a sua rejeição pura e simples. A IA é uma ferramenta poderosa e veio para ficar. O verdadeiro desafio é integrá-la de forma ética, crítica e metodologicamente sólida.

O primeiro caminho é a transparência. Os autores devem declarar quando e como utilizaram IA. A declaração deve ser suficientemente específica: não basta dizer que "foi usada IA". É importante indicar se foi usada para tradução, revisão linguística, resumo da literatura, geração de código, análise estatística, formulação de hipóteses ou apoio à escrita.

O segundo caminho é a educação científica. Investigadores, estudantes, revisores e editores precisam de formação sobre as capacidades e limitações da IA. Devem compreender conceitos como alucinação, viés, explicabilidade, privacidade, autoria, propriedade intelectual e reprodutibilidade. A literacia em IA deve tornar-se parte da formação científica.

O terceiro caminho é a colaboração interdisciplinar. Os problemas colocados pela IA não pertencem apenas à informática. Envolvem ética, direito, ciência da informação, sociologia da ciência, medicina, educação, gestão e políticas públicas. A construção de boas práticas exige diálogo entre áreas.

O quarto caminho é a criação de políticas institucionais claras. Universidades, centros de investigação, revistas e agências de financiamento devem definir orientações consistentes. Essas orientações devem evitar tanto a permissividade acrítica como a proibição irrealista, alinhando-se com os quadros regulatórios europeus (Parlamento Europeu & Conselho da União Europeia, 2024).[13] O objetivo deve ser promover o uso responsável, documentado e auditável da IA.

Explicabilidade e XAI como Requisito Fundamental

A IA Explicável, ou XAI, deve ser considerada um requisito central, especialmente em domínios regulados (Doshi-Velez, Kim, 2017).[14] A explicabilidade procura tornar os sistemas de IA mais compreensíveis, permitindo perceber por que razão um determinado conteúdo foi gerado, que fatores influenciaram uma decisão ou que dados sustentaram uma recomendação.

No contexto da IA generativa, a XAI é importante para distinguir inferência de extrapolação. Um modelo pode inferir algo razoável a partir do contexto, mas também pode extrapolar para além da evidência disponível. Sem mecanismos de explicação, o utilizador pode não perceber onde termina a informação suportada e onde começa a especulação (Gala *et al*, 2026).[15]

A XAI também é essencial para detetar enviesamentos e erros. Se um sistema recomenda uma decisão, classifica um artigo, resume um processo clínico ou interpreta um documento jurídico, é necessário saber que elementos foram considerados. A explicabilidade permite auditoria, responsabilização e correção.

Em domínios como medicina e justiça, a explicabilidade não é apenas desejável: é uma condição de legitimidade. Uma decisão apoiada por IA pode afetar direitos, tratamentos, diagnósticos, liberdade, reputação ou acesso a recursos. Nesses casos, não basta que o sistema tenha bom desempenho estatístico. É necessário que a decisão seja compreensível, contestável e auditável.

Na investigação científica, a XAI também é relevante. Se um sistema apoia a geração de hipóteses, a seleção de variáveis, a interpretação de resultados ou a revisão da literatura, deve ser possível compreender o seu contributo. Caso contrário, a IA torna-se uma autoridade opaca, substituindo a argumentação científica por uma aparência de competência.

AMALIA e a Soberania Linguística

Um ponto particularmente importante é a questão da língua e da soberania digital. A maioria dos grandes modelos de linguagem foi treinada predominantemente em inglês e reflete contextos culturais, jurídicos e institucionais fortemente associados ao mundo anglófono. Isto cria desafios para línguas como o português europeu.

O desenvolvimento de modelos como o AMALIA, orientados para a língua portuguesa e para o contexto europeu, tem relevância estratégica. Um modelo soberano em português pode oferecer melhor alinhamento linguístico e cultural, maior adequação a terminologia institucional, melhor tratamento da linguagem jurídica e administrativa, e maior compatibilidade com quadros legais e éticos europeus (Gala *et al*, 2026).[16]

A frase de Fernando Pessoa: "A minha pátria é a língua portuguesa", ganha aqui um significado tecnológico. A língua não é apenas um meio de comunicação; é também uma infraestrutura cultural, científica e institucional. Se a IA passa a mediar o acesso ao conhecimento, à administração, à educação e à justiça, então a capacidade de desenvolver modelos adequados à língua portuguesa torna-se uma questão de autonomia.

Esta autonomia não deve ser entendida como isolamento. Trata-se antes de garantir que os sistemas utilizados em contextos nacionais e europeus respeitam especificidades linguísticas, valores legais, normas éticas e necessidades institucionais. A soberania digital implica capacidade científica, tecnológica e política para desenvolver, adaptar, auditar e governar modelos de IA.

Importa também considerar a sustentabilidade da IA, não apenas do ponto de vista técnico, mas também ambiental, económico e geopolítico. O treino de modelos de grande escala exige enormes volumes de dados, infraestruturas computacionais muito dispendiosas e um consumo energético elevado, com impacto ambiental significativo. Esta realidade levanta questões éticas sobre a proporcionalidade entre os benefícios obtidos e os recursos consumidos, sobretudo quando muitos destes modelos são usados para fins comerciais, de entretenimento ou de automatização de tarefas de baixo valor social.

Além disso, a concentração destas tecnologias num pequeno número de grandes empresas e países tecnologicamente avançados cria formas de dependência. Os países com menor capacidade económica ou infraestrutural têm maior dificuldade em desenvolver, treinar ou adaptar os seus próprios modelos, ficando dependentes de plataformas externas, de línguas dominantes, de valores culturais incorporados nos dados de treino e de condições comerciais impostas por terceiros. Esta dependência limita a sua soberania digital, a autonomia científica e tecnológica, e até a capacidade de definir políticas públicas independentes em áreas como educação, saúde, justiça, administração pública, defesa e segurança.

Assim, a discussão sobre IA deve incluir não só a inovação e a eficiência, mas também a justiça no acesso, a sustentabilidade energética, a diversidade linguística e cultural,

e a capacidade dos Estados e das comunidades desenvolverem soluções adequadas aos seus próprios contextos. Uma IA verdadeiramente democrática deve ser ambientalmente responsável, tecnicamente acessível e socialmente plural, evitando que a transformação digital aprofunde desigualdades globais já existentes.

Modelos Locais e Privacidade

Outra resposta importante aos desafios da IA é a utilização de modelos locais. Ferramentas como o Ollama permitem executar modelos de linguagem diretamente no computador do utilizador, sem enviar dados para servidores externos. Isto é particularmente relevante para documentos académicos sensíveis, dados clínicos, pareceres confidenciais, propostas de investigação, processos administrativos ou materiais sujeitos a confidencialidade.

A utilização local oferece várias vantagens. Em primeiro lugar, reforça a privacidade, porque os dados não saem do computador. Em segundo lugar, permite trabalhar sem ligação à internet. Em terceiro lugar, dá maior controlo sobre o ambiente de execução. Modelos como Llama, Mistral, DeepSeek-R1 ou Qwen podem ser utilizados localmente, embora com limitações dependentes da capacidade do hardware.

Naturalmente, os modelos locais nem sempre têm o desempenho dos maiores modelos comerciais. Podem exigir memória significativa, por exemplo, 16 GB de RAM como requisito prático mínimo para modelos razoavelmente bons. Ainda assim, em muitos contextos, a troca entre desempenho e privacidade pode justificar a utilização local.

Para instituições académicas, esta possibilidade é especialmente interessante. Permite criar ambientes controlados para uso de IA em investigação, ensino e administração, reduzindo riscos de exposição de dados. Pode também apoiar políticas institucionais de IA responsável, em que certos tipos de documentos só devem ser processados em infraestruturas locais ou certificadas.

Ciência, Responsabilidade e Futuro da Publicação

A IA obriga a ciência a confrontar algumas das suas tensões internas. A publicação científica já enfrentava problemas antes da IA: pressão para publicar, métricas excessivas, revisão por pares sobrecarregada, desigualdades linguísticas, assimetrias de financiamento e práticas questionáveis. A IA não cria todos estes problemas, mas pode amplificá-los.

Ao mesmo tempo, a IA pode ser parte da solução. Pode ajudar a detetar inconsistências, melhorar revisões, apoiar investigadores com menos recursos linguísticos, acelerar análises complexas, promover acesso ao conhecimento e facilitar a colaboração. O seu impacto dependerá das regras, práticas e valores que orientarem a sua utilização.

A questão fundamental é preservar os valores científicos. A verdade continua a exigir validação. O rigor continua a exigir método. A transparência continua a exigir declaração de processos. A equidade continua a exigir atenção às desigualdades. A responsabilidade continua a pertencer aos investigadores, às instituições e às comunidades científicas.

A IA não deve ser tratada como autor moral, nem como substituto do pensamento crítico. Deve ser vista como instrumento poderoso, capaz de ampliar capacidades humanas, mas também de introduzir riscos. A boa ciência não será aquela que rejeita todas as ferramentas de IA, nem aquela que as aceita sem crítica. Será aquela que as utiliza com consciência, documentação, validação e responsabilidade.

Conclusão

A IA representa uma oportunidade extraordinária para a investigação científica e para a publicação académica. Pode acelerar revisões da literatura, apoiar a escrita, sugerir hipóteses, analisar dados, melhorar traduções, automatizar tarefas e facilitar o acesso ao conhecimento. Contudo, essa oportunidade vem acompanhada de desafios metodológicos, éticos e institucionais significativos.

Entre os principais riscos encontram-se a opacidade dos modelos, os vieses dos dados, a dificuldade de reprodutibilidade, as alucinações, a fabricação de referências, a autoria ambígua, a desigualdade de acesso, a violação de privacidade e a normalização de práticas científicas frágeis. Estes riscos não justificam a rejeição da IA, mas exigem uma resposta madura.

Essa resposta deve assentar em transparência, literacia, explicabilidade, políticas claras, supervisão humana e responsabilidade institucional. A XAI deve ser vista como requisito fundamental, sobretudo em domínios regulados. Modelos soberanos, como o AMALIA, e modelos locais, executados em ambientes controlados, podem desempenhar um papel estratégico na proteção da língua, da privacidade e da autonomia científica.

O desafio, portanto, não é evitar a IA. O desafio é integrá-la eticamente e criticamente, garantindo que os valores fundamentais da ciência permaneçam no centro do processo: verdade, rigor, transparência, responsabilidade e equidade. A IA pode transformar a investigação e a publicação científica, mas essa transformação só será positiva se for acompanhada por uma cultura científica capaz de questionar, auditar e governar as suas próprias ferramentas. ■

Ethical Disclosures

Conflicts of Interest: The authors have no conflicts of interest to declare.

Financial Support: This work has not received any contribution grant or scholarship.

Provenance and Peer Review: Commissioned; without external peer review.

Responsabilidades Éticas

Conflitos de Interesse: Os autores declaram a inexistência de conflitos de interesse.

Apoio Financeiro: Este trabalho não recebeu qualquer subsídio, bolsa ou financiamento.
Proveniência e Revisão por Pares: Solicitado; sem revisão externa por pares.

© 2026 Sociedade Portuguesa de Medicina Interna. This is an open-access article under the CC BY-NC 4.0. Re-use permitted under CC BY-NC 4.0. No commercial re-use.
© 2026 Sociedade Portuguesa de Medicina Interna. Este é um artigo de acesso aberto sob a licença CC BY-NC 4.0. Reutilização permitida de acordo com CC BY-NC 4.0. Nenhuma reutilização comercial.

Correspondence / Correspondência:
José Machado - jmac@di.uminho.pt
Escola de Engenharia – Universidade do Minho, Campus de Azurém
4800-058 Guimarães, Braga, Portugal

Recebido / Received: 03/06/2026
Aceite / Accepted: 12/06/2026
Published Online / Publicado Online: 06/07/2026
Published / Publicado: 30/10/2026

REFERÊNCIAS

1. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature. 2023;614:214-6. doi: 10.1038/d41586-023-00340-6.
2. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 4.ª ed. Yorkshire: Pearson; 2021.
3. Goodfellow I, Bengio Y, Courville A. Deep Learning. Massachusetts:MIT Press;2016.
4. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 610–23.
5. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9:1735–80.
6. Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation. 2000;12:2451–71.
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Informat Process Syst. 2017; 30:5998–6008.
8. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Informat Process Syst. 2020;33:1877–901.
9. Andrade F, Novais P, Machado J, Neves J. Contracting agents: Legal personality and representation. Artif Intell Law. 2007;15.
10. Bourtoule L, Chandrasekaran V, Choquette-Choo CA, Jia H, Travers A, Zhang B, et al. Machine unlearning. In: 2021 IEEE Symposium on Security and Privacy (SP). P. 141–59.
11. Machado J, Sousa R, Peixoto H, Abelha A. Ethical Decision-Making in Artificial Intelligence: A Logic Programming Approach. AI. 2024; 5:2707-24. doi: 10.3390/ai5040130
12. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature. 2023;613:612. doi: 10.1038/d41586-023-00191-1.
13. Parlamento Europeu & Conselho da União Europeia. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que estabelece regras harmonizadas em matéria de inteligência artificial (Regulamento da Inteligência Artificial). J Ofic União Europeia: 2024; L1689.
14. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv, 2017:1702.08608.
15. Gala A, Marcondes F, Rodrigues M, Machado J, Novais P. Semiotic ablation as an interface-level XAI method for analyzing generative uncertainty in large language models. In: Proceedings of the 4th World Conference on eXplainable Artificial Intelligence (XAI 2026). Lecture Notes Comput Sci. 2026 (in press).
16. Gala A, Marcondes F, Machado J, Novais P. Designing a local RAG-based intelligent tutoring system for domain-specific education. In: Proceedings of the 39th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2026). IEEE. 2026 (in press).