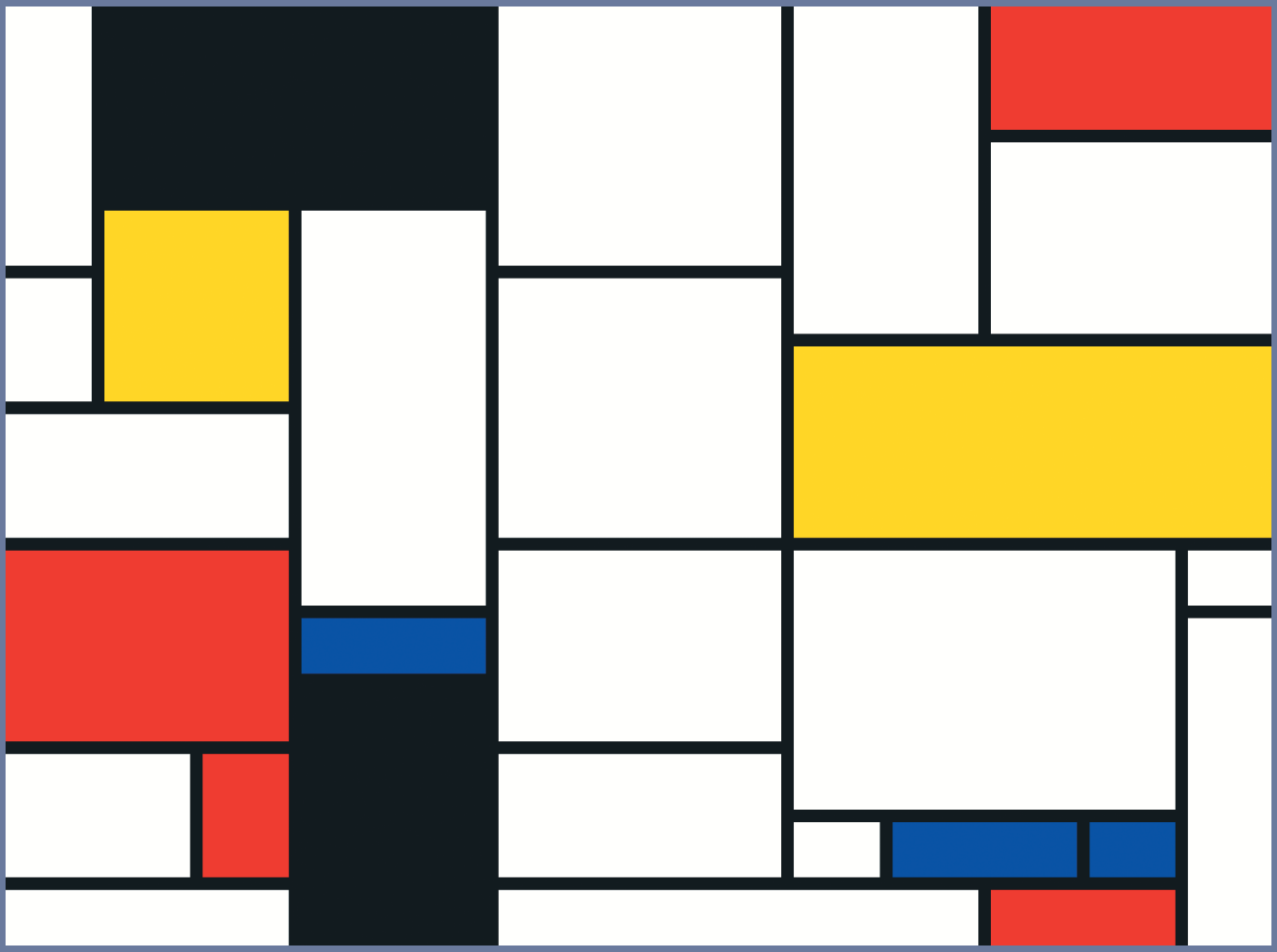


Desafios Sociais e Éticos da Inteligência Artificial no Século XXI

Helena Machado
Susana Silva

Coleção Educação | Ciências Sociais

UMinho Editora

Educação
Ciências Sociais

UMinho Editora

AUTORES

Helena Machado

Susana Silva

COORDENAÇÃO EDITORIAL
Manuela Martins

FOTO CAPA
Imagem concebida em www.mondriangenerator.io

DESIGN
Tiago Rodrigues

PAGINAÇÃO
Carlos Sousa | Talento & Tradição

IMPRESSÃO e ACABAMENTOS
Papelmunde

EDIÇÃO UMinho Editora

LOCAL DE EDIÇÃO Braga 2024

DEPÓSITO LEGAL N.º 540251/24

ISBN impresso: 978-989-9074-51-4

ISBN digital: 978-989-9074-52-1

DOI: https://doi.org/10.21814/uminho.ed.130

Desafios Sociais e Éticos da Inteligência Artificial no Século XXI

Coleção Educação I Ciências Sociais

PREFÁCIO

"Mal vai à obra se lhe requerem prefácio
que a explique, mal vai ao prefácio se
presume de tanto"

José Saramago

No limiar do século XXI, a Inteligência Artificial (IA) emerge como uma força transformadora, que se infiltra tanto nas tarefas mais triviais como nos domínios mais complexos da sociedade. Essa ascensão vertiginosa é, contudo, acompanhada por uma série de desafios sociais e éticos que exigem reflexão e debate aprofundados. É nesse contexto que surge a obra "Desafios Sociais e Éticos da Inteligência Artificial no Século XXI", um compêndio multifacetado que pretende lançar luz sobre as implicações dessa tecnologia revolucionária.

Organizado em seis capítulos abrangentes, o livro traça um panorama meticuloso das questões que envolvem a IA. No primeiro capítulo, Machado e Silva definem de forma clara e acessível o que é a IA, desmistificando mitos e apresentando uma hierarquia dos riscos associados à sua aplicação. O segundo capítulo, por sua vez, mergulha nos princípios éticos universais que orientam o desenvolvimento e a utilização da IA, destacando a importância da participação social e da construção de uma ética de cuidado.

Adotando uma perspetiva sociotécnica, o terceiro capítulo explora a IA como um fenómeno socialmente construído, oferecendo uma análise profunda das dimensões que a envolvem, desde os contextos em que se insere até aos discursos que a moldam. Essa abordagem singular abre caminho para o quarto capítulo, dedicado às aplicações da IA na educação. Aqui, as autoras examinam os temas emergentes nesse campo, as ferramentas disponíveis e os desafios éticos e sociais que precisam ser superados para garantir uma educação justa e equitativa.

O quinto capítulo debruça-se sobre a área da saúde, mapeando as diversas aplicações da IA nesse sector e debatendo os princípios éticos que devem orientá-las. Os desafios sociais e éticos específicos da saúde são explorados em detalhes, com foco na ação e supervisão humanas, na solidez técnica e segurança dos sistemas, na privacidade e governança dos dados, na transparência, na diversidade, na equidade, no bem-estar social e ambiental e na responsabilização.

Por fim, o sexto capítulo centra-se no sistema de justiça, analisando as aplicações da IA em diferentes áreas, como o sistema judicial, o policiamento, as prisões, o controlo

de fronteiras e a cooperação transnacional. As autoras discutem os desafios sociais e éticos que surgem com essas aplicações, convidando o leitor a um debate reflexivo sobre o futuro da justiça na era da IA.

Ao longo dos seis capítulos, "Desafios Sociais e Éticos da Inteligência Artificial no Século XXI" consolida-se como uma obra essencial para todos aqueles que procuram compreender as implicações dessa tecnologia disruptiva. Através de uma análise crítica e multidisciplinar, o livro oferece subsídios valiosos para a construção de um futuro responsável e ético da IA, onde os benefícios dessa tecnologia se traduzam em progresso social e humano para todos.

Este livro é, portanto, um convite à ação, uma chamada para que nos envolvamos ativamente na construção de um futuro onde a IA seja utilizada para o bem da humanidade. Através do diálogo aberto e da reflexão crítica, podemos garantir que essa tecnologia revolucionária contribua para a construção de uma sociedade mais justa, próspera e sustentável para todos.

Eugénio Manuel Faria Campos Ferreira

Introdução

Recentes avanços tecnológicos e científicos têm possibilitado uma aceleração vertiginosa na aplicação e usos de processos cognitivos em máquinas, ou seja, no campo da Inteligência Artificial (IA). Com a massiva disponibilidade de dados digitais e assinaláveis avanços no desenvolvimento das Ciências da Computação, desencadeou-se uma onda crescente de aplicações da IA, abrangendo um espectro vasto de campos – desde a vida quotidiana ao trabalho, saúde, educação e justiça, entre outros. A IA deixou de ser do domínio exclusivo de cientistas e de entusiastas e amantes da ficção científica: Esta tecnologia fala agora à imaginação e às vidas de públicos mais vastos e perspetivam-se efeitos irrevogáveis para a sociedade.

Assistimos hoje a uma "corrida" global de potências mundiais em torno do pioneirismo no desenvolvimento de tecnologias de IA, passíveis de aplicação em diversas áreas. Várias histórias de sucesso da IA têm circulado nos meios de comunicação social, povoando o espaço público de discursos sobre a "revolução" da IA e as grandes transformações sociais, políticas e económicas que estão em curso ou que se avizinham. Referimos, de seguida, alguns dos exemplos mais frequentemente citados por empresas que desenvolvem e promovem tecnologias de IA e que são das aplicações mais conhecidas do público em geral.

A *Amazon* explora o poder da IA para fornecer recomendações personalizadas de produtos. Ao analisar os comportamentos do cliente, o histórico de compras e as suas preferências, o motor de recomendação da *Amazon* sugere produtos a utilizadores individuais. Procedimentos similares são usados em plataformas como a *Netflix* (serviço de subscrição de vídeos, que distribui filmes originais ou comprados) e o *Spotify* (música, podcasts e vídeos) para fornecer recomendações de conteúdos personalizados aos seus subscritores.

Os veículos da *Tesla* podem ter uma condução autónoma graças aos algoritmos de IA e aos dados recolhidos das condições reais de percursos e estradas. A utilização de veículos autónomos tem o potencial de transformar os transportes.

O *IBM Watson*, um sistema de computação cognitiva alimentado por IA, analisa grandes quantidades de dados médicos, incluindo registos de pacientes, documentos de investigação e diretrizes clínicas, para ajudar os médicos a diagnosticar doenças complexas, sugerir planos de tratamento e fornecer informações personalizadas.

O *Facebook* com sistemas de IA[1] aplicados a reconhecimento facial, treinando redes neuronais profundas[2] em conjuntos de dados massivos, alcançou uma elevada precisão no reconhecimento de rostos, permitindo aos utilizadores assinalar e reconhecer rapidamente pessoas nas fotografias.

O *Xiaoice* da *Microsoft* é um *chatbot*[3] alimentado por tecnologias de IA, como o processamento de linguagem natural e a análise de sentimentos. O *Xiaoice* participa em conversas com uma grande base de utilizadores, e oferece apoio emocional, companhia e entretenimento, redefinindo as interações humanos-computador.

A *Siri* é uma assistente virtual da *Apple* alimentada por IA que responde aos comandos do utilizador através do reconhecimento de voz e do processamento de linguagem natural, permitindo aos utilizadores controlar equipamentos domésticos, definir lembretes, enviar mensagens ou fazer chamadas.

Estes exemplos da presença da IA na vida quotidiana consolidam a imagem de revolução transformadora provocada pela IA e denotam a crescente ubiquidade destas tecnologias e a ideia de determinismo tecnológico associado à IA (Bareis e Katzenbach, 2022; Héder, 2021). Podemos definir determinismo tecnológico como uma ideia pela qual se acredita em três pressupostos: Em primeiro lugar, que o desenvolvimento tecnológico acarreta sempre benefícios sociais. Em segundo lugar, que os problemas gerados por tecnologias se resolvem com mais desenvolvimento tecnológico e soluções técnicas. Em terceiro lugar, que o progresso tecnológico tem uma lógica interna de eficiência que determinará o desenvolvimento da estrutura social, cultural e política.

Este livro tem como objetivo debater os principais desafios sociais e éticos da IA no século XXI, a partir da perspetiva da Sociologia, começando por desconstruir a ideia de determinismo tecnológico que parece rodear a IA. Fá-lo-emos chamando a atenção para a natureza sociotécnica da IA: A sociedade é afetada pela tecnologia, mas a tecnologia também é afetada pela sociedade (Søraa, 2023). A nossa abordagem reconhece que a IA são se reduz a sistemas tecnológicos: Estes estão inseridos em contextos históricos, sociais, culturais, políticos e económicos. Esta assunção parece ser muito evidente e comumente vemos peritos de diferentes áreas falarem, com à vontade e confiança, sobre o que a IA vai trazer à sociedade. Mas a aparente simplicidade deste tema é enganadora. Na verdade, é altamente complexo e árduo analisar

1 A expressão "sistema de IA" e "tecnologias de IA" são frequentemente usadas de forma intercambiável, mas têm significados diferentes. No âmbito deste livro falaremos em sistema de IA quando nos referirmos a aplicações práticas da IA e a tecnologias de IA para falarmos do conjunto mais amplo de ferramentas e métodos que alimentam o desenvolvimento desses sistemas. Em termos concretos, um sistema de IA refere-se a um conjunto de componentes interconectados (*hardware*, *software*, algoritmos, bases de dados) para realizar tarefas específicas usando IA; enquanto que as tecnologias de IA envolvem um espectro mais alargado de elementos, designadamente métodos, técnicas e abordagens.

2 Consultar o glossário para mais informações.

3 Consultar o glossário para mais informações.

as implicações sociais e éticas[4] da IA devido a, pelo menos, três motivos. Em primeiro lugar, porque as interações entre as tecnologias de IA e os seus contextos (históricos, sociais, culturais, políticos e económicos) são relacionais, isto é, os contextos interferem no desenho e conceção da tecnologia, nos modos de extrair e usar dados para treinar algoritmos, nas decisões sobre como as tecnologias de IA são aplicadas na vida real, nas tentativas de regulação dos usos destas tecnologias e mesmo nos próprios modos como imaginamos o futuro da IA. Em segundo lugar, porque a IA afeta de modo desigual diferentes países, comunidades, grupos sociais e indivíduos – os ricos tendem a beneficiar mais da IA do que os pobres, por exemplo. Em terceiro lugar, porque as vozes de quem tem mais poder são facilmente ouvidas em relação aos riscos e problemas suscitados pela IA, enquanto as opiniões das comunidades marginalizadas e daquelas que são mais atingidas pelos riscos da IA são silenciadas ou não se conseguem fazer ouvir.

Independentemente de visões mais otimistas ou mais céticas em torno da IA, o conhecimento fundamentado sobre as implicações sociais e éticas do desenvolvimento e uso cada vez mais expansivo destas tecnologias é hoje ainda muito incipiente. A ideia de escrever este livro surgiu da constatação não só das limitações existentes na literatura académica, mas também do entusiasmo e curiosidade que as tecnologias de IA têm suscitado junto de diferentes comunidades e do público em geral. Na nossa qualidade de académicas e professoras pudemos constatar o interesse generalizado da parte de estudantes e colegas em debater o presente e o futuro da IA. No entanto, sentimos necessidade de refletir de modo sistematizado sobre as implicações sociais e éticas da IA no século XXI, de modo a podermos avançar no debate para além das meras impressões e opiniões de senso comum, frequentemente influenciadas por narrativas que circulam nos meios de comunicação social, nas redes sociais e na ficção científica. Essa reflexão é feita, neste livro, a partir de uma proposta de Sociologia da IA[5], que permita abordar como é que as estruturas de poder e as desigualdades sociais podem ser exacerbadas, reproduzidas ou mitigadas pelo uso da IA; como é que diferentes grupos sociais são afetados pela IA, considerando questões de acesso, privacidade e justiça social; como é que a IA pode reforçar ou desafiar normas e hierarquias sociais existentes; quais são os valores, interesses e práticas sociais que enformam as narrativas dominantes de progresso e inovação; como é que decisões técnicas aparentemente neutras podem refletir e perpetuar preconceitos sociais; e como é que a participação pública pode contribuir para um desenvolvimento tecnológico mais equitativo e democrático.

4 Embora as implicações sociais e éticas da IA estejam interligadas, geralmente elabora-se uma distinção entre os diferentes aspetos do impacto da tecnologia na sociedade e nos indivíduos. Nesse sentido, em termos simples, podemos dizer que as implicações sociais da IA se referem aos impactos que a tecnologia de IA pode ter na sociedade como um todo; enquanto que as implicações éticas da IA se referem às considerações e princípios morais que orientam o desenvolvimento, implementação e uso das tecnologias de IA. Ambas são fundamentais para garantir que a IA seja desenvolvida e usada de maneira responsável e benéfica para toda a sociedade.

5 Na abordagem sociológica que propomos integramos contributos dos chamados "Estudos Sociais da Ciência e Tecnologia", um campo interdisciplinar que surgiu em finais da década de 1960, para explorar a co-construção da ciência, tecnologia e sociedade (Bijker et al., 1987).

Entre benefícios e riscos: A regulação da Inteligência Artificial

As visões sobre as implicações sociais e éticas da IA são diversas: Desde um puro entusiasmo, pautado pela enfâse nos benefícios da IA para o progresso e bem-estar da humanidade, a perspetivas mais críticas e cautelosas que salientam os danos causados por uma IA orientada para interesses comerciais e de controlo social e político das populações, ou ainda cenários apocalípticos em que as máquinas ultrapassam os humanos na capacidade de "pensar" e de tomar decisões. Em síntese, os benefícios e os riscos geralmente apontados à IA são os seguintes (Tabela 1), sendo este sumário necessariamente redutor. Além de se tratar de um fenómeno muito complexo, uma discussão ampla e aprofundada dos benefícios e riscos da IA será feita ao longo dos diferentes capítulos deste livro.

Tabela 1
Sumário de (alguns) benefícios e riscos da Inteligência Artificial.

Benefícios	Riscos
Automatização: A IA pode lidar com tarefas rotineiras e repetitivas de forma eficiente, libertando os humanos para se concentrarem em atividades mais criativas e estratégicas.	**Desemprego:** A automação impulsionada pela IA pode levar à substituição de muitos empregos, especialmente aqueles que envolvem tarefas repetitivas e previsíveis.
Eficiência e produtividade: Sistemas de IA podem processar grandes volumes de dados muito mais rapidamente do que seres humanos, com ganhos aplicáveis em vários setores de atividade.	**Vieses e discriminação:** Se os dados usados para treinar sistemas de IA forem tendenciosos, os resultados podem refletir preconceitos existentes na sociedade.
Precisão na tomada de decisões: A IA pode analisar grandes conjuntos de dados e identificar padrões que podem ser difíceis de perceber por seres humanos. Isso pode levar a decisões mais informadas e precisas.	**Brechas na privacidade e segurança:** O uso de IA pode levantar preocupações sobre a segurança e a privacidade dos dados, especialmente em setores sensíveis como saúde e finanças.
Personalização e recomendações: A IA proporciona a oferta de produtos e serviços com base no histórico e nas preferências das pessoas.	**Diluição de responsabilidades:** À medida que a IA se torna mais autónoma, surge a questão de quem é responsável por decisões tomadas por sistemas automatizados.
Inovação na indústria: A IA está a impulsionar avanços em áreas como veículos autónomos, robótica ou automação industrial.	**Dependência tecnológica:** Dependência excessiva de sistemas de IA pode tornar a sociedade vulnerável a falhas em larga escala ou ataques cibernéticos.
Acesso a informações e serviços: *Chatbots* e assistentes virtuais usam a IA para fornecer informações e apoio em tempo real numa ampla variedade de contextos.	**Falta de transparência e interpretabilidade:** Algoritmos de IA complexos podem ser difíceis de entender e explicar, levando à falta de transparência e compreensão sobre como as decisões são tomadas.
	Fragilizar a democracia e a paz: A IA levanta questões acutilantes sobre o uso de tecnologias na vigilância com finalidades securitárias, na guerra autónoma, e na manipulação de opiniões e de comportamentos eleitorais, entre outras áreas.

Face a vozes críticas em relação a potenciais riscos da IA na democracia e na distribuição equitativa e justa de recursos e de benefícios, estão em curso tentativas de regulação da aplicação das tecnologias de IA orientadas para concretizar uma IA mais centrada no "humano", "confiável" e "responsável". A ética tem surgido como a resposta para várias controvérsias suscitadas pela IA, designadamente: Viés e discriminação; riscos à privacidade e proteção de dados; incertezas em relação à responsabilidade relacionada com tomadas de decisão autónoma (sem intervenção humana); as implicações ao nível do trabalho advindas da substituição dos humanos por máquinas em várias tarefas; e, mais recentemente, os receios de manipulação de informação e produção de conteúdos falsos (textos, vídeos, imagens) associados a tecnologias de IA ou mesmo a possibilidade de surgirem máquinas superinteligentes que poderão dominar a humanidade.

Organizações internacionais como a Organização para a Cooperação e Desenvolvimento Económico (OCDE), a Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO), o Fórum Económico Mundial e a União Europeia, entre outras instâncias, têm realçado a importância de garantir que as tecnologias de IA estejam alinhadas com valores sociais e orientadas para a criação de um futuro mais justo, equitativo e inclusivo. Importa, para esse efeito, não só refletir acerca das implicações sociais destas tecnologias, como também definir princípios éticos e de inovação responsável relativos à IA. O presente livro tem como intuito responder a alguns dos reptos dessa missão ao analisar os principais desafios que têm enfrentado quer a compreensão das implicações sociais da IA, quer a concretização de uma IA mais ética, em particular:

- Quem é que define o que é "bom" para a sociedade?
- Quais são os valores sociais a prevalecer e o que é um bem coletivo?
- Como é que a IA pode ser projetada e utilizada de forma a beneficiar a sociedade no seu todo?
- Que transformações e mudanças sociais se avizinham?
- Como desenvolver uma atitude prudente e responsável perante os riscos suscitados pela IA e que formas de governação podem contribuir para que a IA seja um pilar de sociedades mais justas e solidárias?
- Como lidar com a ambiguidade dos significados atribuídos à ética associada à IA, entre um sentido geral e impreciso que engloba "tudo" o que tenha a ver com os cidadãos e a sociedade, e um sentido restrito que a considera matéria da competência exclusiva de especialistas em ética?
- Como operacionalizar e converter em práticas concretas os princípios abstratos associados à ideia de uma IA mais ética e compatível com o bem-estar da humanidade?

O debate público sobre Inteligência Artificial

Além dos alertas sobre a necessidade de regulamentações e diretrizes éticas para garantir que a IA seja usada de maneira responsável e segura, especialmente em

áreas sensíveis como a educação, saúde e justiça, vários líderes tecnológicos têm apelado à necessidade de um debate público sobre a IA que esclareça algumas das controvérsias e que defina as preocupações reais a ter em conta em relação a estas tecnologias (em contraponto ao sensacionalismo). Na tabela 2, sintetizamos os termos do atual debate público sobre IA.

Tabela 2
Temas do debate público sobre Inteligência Artificial.

	Setor privado (líderes tecnológicos)	Setor público e governos	Cidadãos
Função	Desenvolver e comercializar IA.	Regulamentar a IA e criar condições para a transição digital.	Utilizadores de IA, trabalhadores.
Objetivos	Educar as pessoas, esclarecendo-as sobre o que é que o campo da IA alcançou até hoje e qual é a sua trajetória futura esperada, sem sensacionalismos. Como é que planeiam continuar a desenvolver a IA e quais são os reais impactos esperados na sociedade. Reivindicar alocação de recursos e investimento público e uma regulação que não impeça o desenvolvimento tecnológico.	Avaliar as áreas de risco e decidir o que é necessário regulamentar. Lidar com os desafios da transição digital e instigar os cidadãos a serem pró-ativos. Coordenarem-se com o setor privado e outros grupos de interesse para delinear estratégia conjunta.	Serem devidamente informados e educados sobre os benefícios e riscos reais da IA. Mão-de-obra, com necessidade de formação para a aquisição de mais competências digitais. Utilizar produtos e serviços que incorporam IA.

Fonte. As características relativas ao setor privado e ao setor público/governos foram inspiradas numa análise levada a cabo por Sarangi e Sharma (2019, p. 7). Os termos referentes aos cidadãos são inspirados no trabalho de Wilson (2022, pp. 7-8).

O debate público em torno da IA está dominado por três pressupostos que correspondem aos interesses e poderes de líderes tecnológicos, governos e decisores políticos: Em primeiro lugar, a convicção da inevitabilidade e desejabilidade do desenvolvimento ininterrupto da IA (Phan et al., 2022; Steinhoff, 2023). Em segundo lugar, a perceção que os riscos da IA podem ser controlados com regulamentação, legislação e, principalmente, mais intervenções técnicas que assegurem maior privacidade e segurança dos dados e maior transparência dos sistemas de IA (Hagendorff, 2020; Mittelstadt, 2019; Ulnicane et al., 2021a). Em terceiro lugar, a enunciação do valor da participação de diferentes públicos no desenho de decisões e políticas para assegurar uma IA mais "responsável" e "confiável", princípios louváveis que, porém, ficam-se por enunciados abstratos e gerais: Em termos concretos, o enquadramento do papel dos públicos é circunscrito a funções que sirvam dinâmicas económicas e de mercado (Machado et al., 2023).

Ao longo deste livro iremos argumentar porque motivos consideramos que os termos do debate público são limitados e porque é que as circunstâncias de envolvimento dos diferentes públicos devem ser alteradas, de modo a efetivamente incluir as vozes das pessoas e das comunidades, ouvindo, respeitando e integrando nas decisões coletivas a sua visão do mundo e o tipo de sociedade que desejam.

O que este livro é e o que não é

Este livro não é de natureza técnica, nem explica de modo detalhado como é que funcionam as tecnologias de IA. Apenas faculta a informação técnica estritamente necessária para se perceber em que consistem as tecnologias de IA e porque é que estas suscitam interrogações sobre as implicações sociais e éticas do seu desenvolvimento, implementação e utilização.

Este livro também não aborda de modo exaustivo todas as implicações sociais e éticas da IA: Tal missão seria impossível, visto que estamos ainda longe de compreender em profundidade o alcance das transformações provocadas – ou pré-existentes, mas reforçadas – pela IA. De igual modo, nas perspetivas apresentadas não adotamos uma visão otimista ou pessimista em relação ao presente e ao futuro da IA. O que se pode encontrar neste livro é uma abordagem crítica sobre algumas das implicações sociais e éticas da IA no século XXI, ou seja, um posicionamento que suscita questões e que aponta caminhos de reflexão.

Usando uma linguagem acessível, adaptada à compreensão da parte de leitores com diferentes formações disciplinares, procuraremos alcançar dois objetivos principais com esta obra: Por um lado, desvendar os pressupostos subjacentes e as forças mais amplas que sustentam, perturbam ou complicam crenças na capacidade fundamental da IA para resolver os problemas da sociedade. Por outro lado, assinalar que as diferenças contextuais (históricas, culturais, sociais, políticas e económicas) podem influenciar a forma como a IA é percebida e utilizada por comunidades distintas – em suma, o que é benéfico para uns pode tornar-se uma desvantagem e um problema para outros.

Ao longo dos diferentes capítulos apresentaremos um conjunto de questões, de exemplos e de propostas de debate, que têm em vista instigar a reflexão, de modo prático e interativo, junto de públicos diversificados. Poder-se-á dizer que a nossa missão é contribuir para problematizar a ideia da omnipresença da IA enquanto figura estabilizada que encerra questões controversas numa "caixa negra", imagem que tem contribuído para legitimar a expansão das tecnologias de IA (Suchman, 2023). Nas palavras do filósofo e sociólogo francês Bruno Latour, as caixas negras da ciência e tecnologia referem-se à:

> Forma como o trabalho científico e técnico se torna invisível pelo seu próprio sucesso. Quando uma máquina funciona de forma eficiente, quando uma questão de facto é resolvida, é preciso concentrar-nos apenas nos seus *inputs* e *outputs* e não na sua complexidade interna. Assim, paradoxalmente, quanto mais a ciência e a tecnologia forem bem-sucedidas, mais opacas e obscuras estas [caixas negras] se tornarão. (Latour, 1999, p. 304)

Neste livro procuraremos desmontar a caixa negra da IA ao evidenciar alguns enredos que entrelaçam a máquina (as tecnologias de IA) e a sociedade, procurando desvendar como as ramificações aparentemente opacas da caixa negra circulam, de

modo ambíguo e complexo, na sociedade. Trata-se de compreender as implicações sociais e éticas da IA em vários níveis: Desde aquilo que se considera ser o conhecimento e criatividade humanos, aos princípios de justiça, equidade, bem-estar, solidariedade, inclusão, diversidade e não discriminação, e aos modos de organização social, política e cultural. Por outras palavras, a amplitude e os significados mutáveis e plásticos da ideia de IA faz com que uma reflexão sobre as suas implicações sociais e éticas nos conduza a pensar o que somos e o que podemos ser, individual e coletivamente, qual é o futuro projetado e se esse é o futuro desejável.

O presente livro organiza-se do seguinte modo: O capítulo 1 dedica-se à definição de conceitos relacionados com as tecnologias de IA e a um mapeamento das suas aplicações e riscos; o capítulo 2 sistematiza as principais implicações sociais e éticas da IA; e o capítulo 3 apresenta perspetivas teórico-metodológicas da Sociologia aplicáveis ao estudo da IA. Os restantes capítulos deste livro dedicar-se-ão à análise das implicações sociais e éticas da IA em campos concretos da vida social: Na educação (capítulo 4), na saúde (capítulo 5) e na justiça (capítulo 6). A conclusão sintetiza as principais mensagens deste livro, perspetivando agendas de investigação e desafios futuros para o campo emergente da Sociologia da IA.

1. Conceitos, aplicações e riscos da Inteligência Artificial

Introdução

Os primórdios da Inteligência Artificial (IA) remontam à década de 1950, e estão associados a três marcos históricos principais. Primeiro, a publicação do artigo *Computing Machinery and Intelligence*, em 1950, da autoria do matemático e cientista da computação britânico Alan Turing. Este desenvolveu os conceitos de algoritmo e computação por meio da máquina de Turing, que pode ser vista como um protótipo de computador de uso geral. Segundo, o uso da expressão "Inteligência Artificial" pela primeira vez em 1955, num texto relativo a uma proposta de projeto sobre Inteligência Artificial (*Proposal for the Dartmouth summer research project on artificial intelligence*), elaborada por John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon. Terceiro, a oficialização do termo "Inteligência Artificial" em 1956, durante a famosa Escola de Verão na Universidade de Dartmouth, em Hanover, New Hampshire, EUA. Este evento reuniu um grupo de investigadores que estavam interessados em explorar maneiras de fazer com que as máquinas pudessem imitar funções cognitivas humanas, como aprendizagem, raciocínio e resolução de problemas. O termo "Inteligência Artificial", cunhado nessa Escola de Verão, tornou-se a denominação padrão para a área (Wooldrige, 2021).

Desde os primórdios da IA e durante as décadas que se seguiram, em que o investimento público nessa área foi inconstante e variável, até aos dias de hoje, em que já se fala da "era da IA", algo mudou radicalmente. Cientistas e empresários que desenvolvem IA tanto se posicionam de modo entusiasta e laudatório como manifestam preocupações extremas em relação aos riscos da IA. Porquê? Convoquemos exemplos de ações de grandes empreendedores na área da IA, ao tomarem posições públicas sobre a mesma, para compreendermos a complexidade das expectativas sociais em torno deste fenómeno.

Em 22 de março de 2023, mais de 1.000 líderes tecnológicos e investigadores, incluindo grandes empreendedores do sector tecnológico, como Elon Musk e Steve Wozniak, instaram os laboratórios de IA a interromper o desenvolvimento dos sistemas mais avançados. Numa carta aberta tornada pública pela organização *Future of Life Institute* (Instituto do Futuro da Vida), uma organização sem fins lucrativos com o objetivo declarado de reduzir os riscos catastróficos globais e existenciais enfrentados pela humanidade, estes líderes alertaram que a IA apresentaria riscos sérios para a sociedade e a humanidade. De acordo com os seus signatários, quem está a desenvolver IA encontra-se numa corrida descontrolada para desenvolver e usar mentes digitais cada vez mais poderosas que ninguém – nem mesmo os seus criadores – pode compreender, prever ou controlar de forma confiável. As seguintes questões surgiram nesta carta aberta:

> Devemos deixar as máquinas inundarem os nossos canais de informação com propaganda e mentiras? Devemos automatizar todos os trabalhos, inclusive aqueles que são gratificantes? Devemos desenvolver mentes não-humanas que possam vir a superar-nos em número, em inteligência, e tornar-nos obsoletos e substituir-nos? Devemos correr o risco de perder o controlo da nossa

> civilização? Tais decisões não devem ser delegadas em líderes tecnológicos não eleitos. Sistemas poderosos de IA só deverão ser desenvolvidos quando estivermos confiantes de que os seus efeitos serão positivos e os seus riscos serão controláveis.

Apelava-se a uma "pausa" no desenvolvimento de sistemas de IA generativos (ou seja, sistemas de IA capazes de gerar conteúdos), como o *ChatGPT*, mas sem pretensão de recuar no desenvolvimento de uma IA cada vez mais sofisticada. De acordo com a carta:

> Isto não significa uma pausa no desenvolvimento da IA em geral, apenas um retrocesso na corrida perigosa para modelos de caixa negra cada vez maiores e imprevisíveis com capacidades emergentes. A investigação e o desenvolvimento da IA devem ser reorientados para tornar os sistemas poderosos e de última geração mais precisos, seguros, interpretáveis, transparentes, robustos, alinhados, fiáveis e leais.

A posição tomada pelos signatários desta carta aberta apontava para a existência de soluções para lidar com os graves problemas suscitados pelo desenvolvimento desregulado e descontrolado da IA. Apelavam, por exemplo, a que se desenvolvessem protocolos de segurança adequados a sistemas avançados de IA, que pudessem ser "rigorosamente auditados e supervisionados por especialistas externos independentes", e que os criadores de IA pudessem "trabalhar com os decisores políticos para acelerar drasticamente o desenvolvimento de sistemas robustos de governação de IA". Algumas das ações necessárias passariam, de acordo com estes especialistas, pela criação de "autoridades reguladoras novas e capazes dedicadas à IA", pelo desenvolvimento de mecanismos de "supervisão e rastreamento de sistemas de IA dotados de conjuntos de elevada capacidade computacional", pela criação de "um ecossistema robusto de auditoria e certificação", pela definição clara da "responsabilidade por danos causados pela IA", pelo "financiamento público robusto para pesquisas técnicas de segurança em IA" e por dotar as "instituições com bons recursos para lidar com as dramáticas perturbações económicas e políticas (especialmente para a democracia) que a IA irá causar". Em suma, os dilemas e riscos suscitados pela IA deveriam ser resolvidos, nesta perspetiva, com mais intervenções tecnológicas e com maior alocação de recursos (incluindo financiamento público) para o desenvolvimento de sistemas de IA mais confiáveis.

Esta não foi a primeira vez que figuras proeminentes do campo da tecnologia se pronunciaram sobre o futuro da IA ao longo dos anos. Porém, este tipo de pronunciamento público foi particularmente frequente e incisivo em 2023, ano em que Bill Gates publicou no seu blogue pessoal diversas opiniões sobre IA, afirmando sistematicamente que os riscos reais suscitados pela IA são controláveis desde que haja um debate público sério. A título de exemplo, em julho de 2023, no seu blogue pessoal (GatesNotes, 2023), Gates escreveu o seguinte: "Uma coisa que ficou clara de tudo o que foi escrito até agora sobre os riscos da IA – e muito foi escrito – é que ninguém tem todas as respostas". Acrescentou, ainda, que "Outra coisa que é clara para mim é

que o futuro da IA não é tão sombrio como algumas pessoas pensam ou tão cor-de--rosa como outros pensam".

Na publicação do blogue, Gates citava a forma como a sociedade reagiu a avanços anteriores para defender que os seres humanos se adaptaram a grandes mudanças no passado e que o farão também com a IA. Nas suas palavras: "Por exemplo, a IA terá um grande impacto na educação, mas o mesmo aconteceu com as calculadoras portáteis há algumas décadas e, mais recentemente, com a permissão de computadores na sala de aula". Gates sugeria ainda que o tipo de regulamentação que a tecnologia precisa é de "limites de velocidade e cintos de segurança".

Recuando um pouco no tempo, bem antes do pico da atenção mediática em torno da IA que, na nossa perspetiva, foi em certa medida desencadeado pelo lançamento público do famoso *ChatGPT* em 30 de novembro de 2022, encontramos um exemplo ilustrativo da complexidade da questão relativa aos benefícios e riscos da IA. Referimo-nos à opinião manifestada por Stephen Hawking, renomado físico teórico da Universidade de Cambridge, em relação à IA, em diversos momentos. Numa entrevista à BBC, em 2014, Hawking expressou preocupações acerca dos perigos associados à IA, afirmando que "O desenvolvimento da inteligência artificial total poderá significar o fim da raça humana"[6].

Hawking começou por afirmar que formas básicas de IA desenvolvidas até à altura estar-se-iam a revelar muito úteis, na medida em que ele próprio utilizava uma forma rudimentar de IA para comunicar[7], mas que temia as consequências de criar algo que pudesse igualar ou superar os humanos. O cientista salientou que os esforços para criar máquinas pensantes representariam uma ameaça à existência humana. Nas suas palavras: "Os humanos, que são limitados pela lenta evolução biológica, não poderiam competir e seriam superados". Acrescentou, ainda, que o desenvolvimento de IA mais avançada poderia fazer com que esta se "desenvolvesse por conta própria, redesenhando-se a um ritmo cada vez mais célere".

Estas tomadas de posição pública sobre a IA da parte de empresários da área tecnológica e de cientistas, pela ampla visibilidade mediática que alcançam, têm marcado

6 Estas citações foram traduzidas livremente pelas autoras, socorrendo-se da entrevista de Steven Hawking ao canal televisivo britânico BBC, em 2 de dezembro de 2014, disponível online em https://www.bbc.com/news/technology-30290540. A frase original a que nos referimos nesta tradução é "The development of full artificial intelligence could spell the end of the human race". A expressão "full artificial intelligence" tem um sentido ambíguo, mas atendendo ao contexto geral da entrevista admitimos que Hawking se estivesse a referir à possibilidade teórica, ainda hoje explorada, de desenvolvimento de IA geral ("artificial general intelligence"), que criará máquinas com algumas capacidades cognitivas superiores aos humanos no que diz respeito à capacidade perfeita de recordar, uma base de "conhecimento" (acumulação de dados e informação) vastamente superior, e à capacidade de realizar multitarefas de maneiras não possíveis para entidades humanas.

7 Devido à progressiva esclerose lateral amiotrófica (ELA) que o afetava, Hawking perdeu a capacidade de falar e mover-se e, por isso, utilizava um sistema de comunicação assistida que consistia num *software* que permitia que Hawking selecionasse letras e palavras para formar frases. Além disso, o sistema também aprendia a antecipar as palavras que Hawking poderia querer usar com base nos padrões da sua escrita (vários textos produzidos ao longo dos anos tinham sido arquivados no computador que era usado neste tipo de sistema).

a imaginação coletiva e influenciado muito do que se pensa e fala sobre IA. No entanto, do nosso ponto de vista, é crucial ver para além destas fachadas (Berger, 2001), escrutinando as implicações sociais e éticas da IA para além dos discursos que a definem tanto como a solução para grandes problemas que afetam as sociedades ou como um perigo existencial para a humanidade.

Para iniciarmos este caminho de reflexão sobre as implicações sociais e éticas da IA, temos que primeiro perceber o que é esta tecnologia. Por outras palavras, o que é que se entende por IA? Como é que a tecnologia se desenvolveu e em que ponto estamos atualmente?

Procurando responder a estas questões, na próxima secção exploramos a coexistência de várias definições de IA, desde as propostas da Comissão Europeia e da Organização para a Cooperação e Desenvolvimento Económico (OCDE) a dicionários, apontando para a importância de especificar as tarefas e as competências humanas imitadas ou simuladas pelas tecnologias de IA. Refletimos ainda sobre a distinção entre racionalidade e inteligência ao falar de IA, expondo os processos de aprendizagem envolvidos neste subcampo das Ciências da Computação.

De seguida, desconstruímos os principais mitos em torno da IA e, por fim, apresentamos a hierarquização e tipologia de riscos sugerida no Regulamento da IA da União Europeia, ilustrando-a com exemplos concretos de utilização. Procedemos à análise crítica deste Regulamento, realçando, por um lado, as fragilidades da distinção entre tecnologias de IA que apresentam risco inaceitável, risco elevado e risco não elevado, e, por outro lado, a forma como a institucionalização do risco como uma ferramenta de gestão de controvérsias públicas ao nível das decisões políticas tem contribuído para negligenciar ou invisibilizar preocupações sociais e éticas que estão à margem de agendas políticas e de interesses económicos e comerciais ou cuja "resolução" dificilmente encaixa em soluções técnicas.

Em síntese, neste capítulo pretendemos:

- Identificar várias definições de IA que circulam na sociedade.
- Expor conceitos e mitos relacionados com a IA.
- Discutir as soluções avançadas por líderes tecnológicos contemporâneos para lidar com os problemas suscitados pelo desenvolvimento da IA.
- Apresentar, de forma crítica, a hierarquização e tipologia de riscos enunciada no Regulamento da IA da União Europeia.

1.1. O que é a Inteligência Artificial?

São várias as definições possíveis de IA, o que pode gerar confusão. Definir a IA não é fácil e, em bom rigor, não existe uma definição consensual do conceito (Russel e Norvig, 2020; Sheikh et al., 2023), o que é revelador da própria complexidade deste fenómeno. Comecemos, então, por tentar perceber quais são as várias definições de IA que circulam na sociedade.

Na proposta de Regulamento da IA, cuja primeira versão foi elaborada em abril de 2021, a Comissão Europeia explicitava que a definição de sistema de IA devia "ser o mais possível tecnologicamente neutra e preparada para o futuro, tendo em conta a rápida evolução tecnológica e de mercado no domínio da inteligência artificial" (Comissão Europeia, 2021, p. 14). Em momentos distintos, representantes de órgãos da União Europeia defenderam a ideia que se devia seguir de perto a proposta oficial apresentada pela OCDE e que é a seguinte:

> Um sistema de IA é um sistema baseado em máquinas que pode, para um determinado conjunto de objetivos definidos pelo ser humano, fazer previsões, recomendações ou decisões que influenciam ambientes reais ou virtuais. Os sistemas de IA são concebidos para funcionar com diferentes níveis de autonomia. (OCDE, 2023, p. 7)

A União Europeia acabou por definir, no seu Regulamento em matéria de IA, que um sistema de IA é baseado em máquinas e funciona com níveis de autonomia variáveis, podendo "apresentar capacidade de adaptação após a implantação e que, para objetivos explícitos ou implícitos, e com base nos dados de entrada que recebe, infere a forma de gerar resultados, tais como previsões, conteúdos, recomendações ou decisões que podem influenciar ambientes físicos ou virtuais" (União Europeia, 2024, p. 46).

Outras respostas podem ser encontradas em dicionários ou mesmo em textos científicos. O dicionário *Merriam-Webster*[8] (2023), por exemplo, apresenta duas definições de IA, explicitando que o conceito pode ser entendido em dois sentidos. Num primeiro sentido, que é considerado equivalente ao sentido originalmente usado quando a expressão foi cunhada em 1955 (McCarthy et al., 1955), IA é "a capacidade de sistemas informáticos ou algoritmos imitarem um comportamento humano inteligente". O segundo sentido de IA é formulado nos seguintes termos: "Um ramo da ciência da computação que trata da simulação de comportamentos inteligentes em computadores". Curiosamente, a edição de 2023 do dicionário *Merriam-Webster* sugere a consulta de outro conceito na entrada de definição de IA: "Inteligência Artificial Generativa", equiparando-o à segunda possibilidade de definição de IA, na medida em que propõe o entendimento da IA generativa como "inteligência artificial que é capaz de gerar novos conteúdos (como imagens ou texto) em resposta a uma solicitação apresentada (como uma questão), aprendendo com uma grande base de dados de referência de exemplos".

A definição comum de IA, de que se trata de uma tecnologia que permite às máquinas imitarem ou simularem várias capacidades humanas complexas, não nos dá muitas pistas: Se essas competências humanas "inteligentes" não forem especificadas, continua a não ser claro o que é exatamente a IA. Algumas definições de IA

8 A opção por este dicionário decorre da sua versatilidade e carácter interativo, na medida em que não só apresenta definições (e exemplos da correta utilização dos conceitos), como também remete para a origem histórica da palavra, artigos, citações e outros itens de interesse relacionados com a pesquisa realizada.

tentam especificar as competências humanas e as tarefas que a IA consegue "ter" ou "simular", referindo as capacidades de perceber, de prosseguir objetivos, de iniciar ações e de aprender com base em experiências passadas e reações (por exemplo, de programadores ou de utilizadores). Um exemplo deste tipo de definição, que tem por base o tipo de competências humanas a reproduzir ou as tarefas a desenvolver pela IA, é a definição apresentada pelo Grupo de Peritos de Alto Nível em Inteligência Artificial (GPAN IA) nomeado pela Comissão Europeia em 2018, numa versão curta e numa versão mais extensa, respetivamente:

> Sistemas que demonstram um comportamento inteligente, analisando o seu ambiente e tomando medidas – com algum grau de autonomia – para atingir objetivos específicos. Os sistemas baseados em IA podem ser puramente baseados em *software*, atuando no mundo virtual (por exemplo, assistentes de voz, *software* de análise de imagem, motores de busca, sistemas de reconhecimento de voz e rosto) ou a IA pode ser incorporada em dispositivos de *hardware* (por exemplo, robôs avançados, carros autónomos, *drones* ou aplicações da Internet das Coisas[9]). (GPAN IA, 2019a, p. 1)

> Os sistemas de inteligência artificial (IA) são sistemas de *software* (e eventualmente também de *hardware*) concebidos por seres humanos que, tendo em conta um objetivo complexo, atuam na dimensão física ou digital, percebendo o seu ambiente através da aquisição de dados, interpretando os dados estruturados ou não estruturados recolhidos, raciocinando sobre o conhecimento, ou processando a informação, derivada desses dados e decidindo a(s) melhor(es) ação(ões) a tomar para atingir o objetivo dado. Os sistemas de IA podem utilizar regras simbólicas ou aprender um modelo numérico e podem também adaptar o seu comportamento analisando a forma como o ambiente é afetado pelas suas ações anteriores. Enquanto disciplina científica, a IA inclui várias abordagens e técnicas, como a aprendizagem automática (de que a aprendizagem profunda e a aprendizagem por reforço são exemplos específicos), o raciocínio automático (que inclui o planeamento, a programação, a representação e o raciocínio do conhecimento, a pesquisa e a otimização) e a robótica (que inclui o controlo, a perceção, os sensores e os atuadores, bem como a integração de todas as outras técnicas em sistemas ciberfísicos). (GPAN IA, 2019a, p. 6)

Em suma, a IA é um subcampo das Ciências da Computação e abrange duas áreas principais: A aprendizagem automática ou aprendizagem da máquina (*machine learning*[10]); e a aprendizagem profunda (*deep learning*[11]) (Figura 1). O processo de aprendizagem da máquina geralmente ocorre de três formas: Através de aprendizagem supervisionada[12], por via de classificações que são muitas vezes, ainda que não sempre,

9 Consultar o glossário para mais informações.

10 Consultar o glossário para mais informações.

11 Consultar o glossário para mais informações.

12 Consultar o glossário para mais informações.

provenientes de intervenção humana nos conjuntos de dados, e que guiam o processo; por via de aprendizagem não supervisionada[13], na qual a máquina não recebe orientações diretas; e por aprendizagem por reforço (*reinforcement learning*)[14], envolvendo um ambiente dinâmico, onde existe evolução e aprendizagem com base na experiência adquirida, com penalização de erros e recompensas. A IA pode fazer uso de *Big Data*[15], ou seja, de conjuntos de dados extremamente grandes e complexos que desafiam as capacidades tradicionais de processamento de dados. Geralmente, esses conjuntos de dados são caracterizados por três principais aspetos (3 V's): Volume (grandes quantidades de dados gerados); velocidade (com o fluxo de dados em tempo real, os dados podem ser gerados a velocidades incrivelmente elevadas); e variedade (desde dados estruturados tradicionais, como bases de dados, a dados semi-estruturados e não estruturados, como vídeos, textos, etc.).

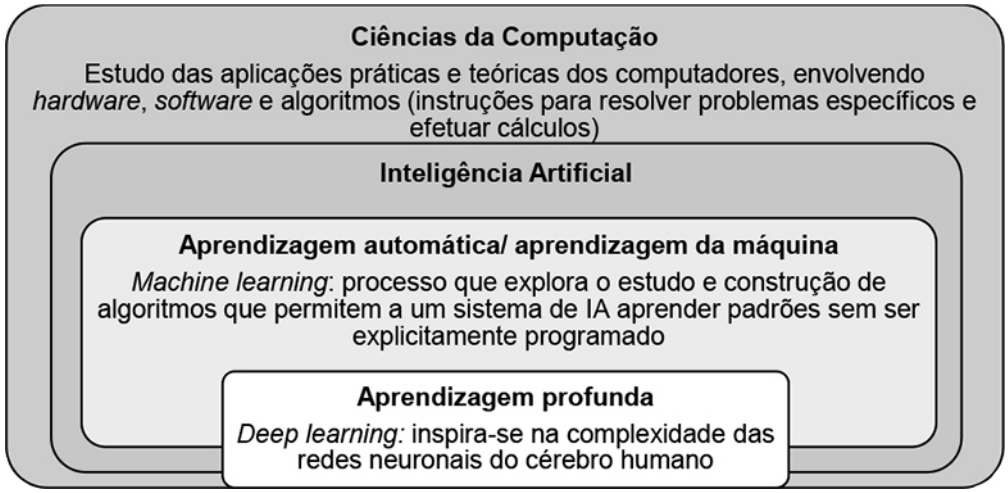


Figura 1
Inteligência Artificial e campos relacionados.

Fonte. Adaptado de Søraa, 2023, p. 6.

Na perspetiva do Grupo de Peritos de Alto Nível em Inteligência Artificial (GPAN IA) da Comissão Europeia, é mais correto falar em racionalidade do que em inteligência quando falamos de IA. De acordo com a definição proposta por este grupo, racionalidade é a "… capacidade de escolher a melhor ação a adotar para atingir um determinado objetivo, tendo em conta certos critérios que se devem otimizar e os recursos disponíveis" (GPAN IA, 2019a, p. 1). Estes peritos acrescentam que uma tecnologia de IA atinge a racionalidade pela perceção do ambiente em que está imersa através de alguns sensores, recolhendo e interpretando dados, processando a informação derivada desses dados, e, através de atuadores[16], decidir qual é a melhor ação e agindo em conformidade, modificando assim possivelmente o ambiente. Esse

13 Consultar o glossário para mais informações.

14 Consultar o glossário para mais informações.

15 Consultar o glossário para mais informações.

16 Os atuadores (do inglês *actuators*) são dispositivos ou componentes que permitem a um sistema interagir com o ambiente. São responsáveis por converter informação digital em ações físicas. Consultar o glossário para mais informações.

raciocínio é feito porque as tecnologias de IA podem utilizar regras simbólicas ou aprender um modelo numérico, além de poderem também adaptar o seu comportamento, analisando a forma como o ambiente é afetado pelas suas ações anteriores (GPAN IA, 2019a).

1.2. Mitos e factos em torno da Inteligência Artificial

Ainda hoje se fala com alguma frequência na distinção entre a "IA fraca" e a "IA forte": A primeira refere-se a tecnologias de IA que podem trabalhar de forma "inteligente" numa tarefa específica; enquanto a segunda aponta para tecnologias de IA que são capazes de desempenhar várias tarefas, de forma integrada. A IA fraca seria, em palavras simples, um método matemático de previsão, e a IA forte seriam máquinas que pensam como humanos (Søraa, 2023, p. 5).

A ideia da superinteligência é frequentemente associada ao filósofo e matemático britânico Irving John Good, que cunhou o termo "explosão de inteligência" (*intelligence explosion*) na década de 1960. Na sua perspetiva, uma vez que uma tecnologia de IA atinja um certo nível de inteligência, ela poderá ser capaz de aprimorar a sua própria inteligência de forma exponencial, levando a um rápido aumento na capacidade intelectual. A superinteligência é um tópico de debate e especulação em ética e filosofia da IA. Muitos teóricos discutem os desafios e as implicações de criar uma IA superinteligente, como a segurança da IA, o controlo sobre os seus comportamentos e a potencial evolução dos seus objetivos e motivações.

A maioria das atuais aplicações de IA são consideradas IA fraca, pois são projetadas para tarefas específicas e limitadas a domínios particulares de atuação. A IA forte corresponderia a uma superinteligência em IA: Isto é, a um nível de IA que ultrapassaria significativamente a capacidade cognitiva humana em praticamente todos os aspetos. É importante notar que a ideia de IA forte é um conceito teórico – implica uma IA superinteligente que seria capaz de realizar tarefas intelectuais com um desempenho muito superior ao dos seres humanos em praticamente todas as áreas, incluindo a resolução de problemas complexos, a tomada de decisões, a criatividade, e a compreensão de *nuances* emocionais e sociais, entre outras.

A construção de uma IA superinteligente é considerada por muitos como uma questão especulativa, mas que envolve uma série de desafios sociais e éticos que precisam ser cuidadosamente considerados desde já. Esta é a posição do Instituto do Futuro da Vida (*The Future of Life Institute*), instituição que reúne vários cientistas que têm vindo a debater os benefícios e riscos reais da IA. Na perspetiva desta instituição, e de acordo com a síntese elaborada por Sarangi e Sharma (2019, p. 8), os principais mitos e factos em torno da IA são os seguintes:

Tabela 3
Mitos e factos em torno da Inteligência Artificial.

Mitos	Factos
A superinteligência em 2100 é impossível.	Pode concretizar-se dentro de décadas, séculos ou nunca.
A superinteligência, por volta de 2100, é inevitável.	Os especialistas estão em desacordo, simplesmente não se sabe.
Apenas os céticos da tecnologia se preocupam com a IA.	Investigadores renomados da área da IA estão preocupados.
A possibilidade da IA se tornar malévola ou consciente é motivo de preocupação.	A possibilidade da IA se tornar competente com objetivos não coincidentes com os nossos é motivo de preocupação.
Os robôs são a principal preocupação.	Uma inteligência não alinhada com os nossos valores e objetivos é a principal preocupação.
A IA não pode controlar os humanos.	A inteligência confere controlo; nós controlamos os animais porque somos mais inteligentes.
As máquinas não podem ter objetivos.	Os mísseis que procuram os alvos através de deteção de calor têm um objetivo.

Fonte. Conn, A. (2015). The top myths about advanced AI. Disponível em: https://futureoflife.org/ai/benefits-risks-of-artificial-intelligence/ [Acesso a 24 de julho de 2024].

Também a União Europeia, na prossecução da tentativa de liderança tecnológica em matéria de IA, tem procurado conjugar o usufruto dos benefícios que as tecnologias de IA podem trazer com a criação de mecanismos legislativos que protejam os cidadãos dos seus riscos e danos. Fê-lo, em particular, ao desencadear um processo de produção da primeira proposta para estabelecer regras harmonizadas em matéria de IA como parte da sua estratégia digital – o Regulamento Inteligência Artificial 2021, cujo preâmbulo refere "… o objetivo da União de estar na vanguarda mundial do desenvolvimento de uma inteligência artificial que seja segura, ética e de confiança" (Comissão Europeia, 2021, p. 20). Na secção relativa às "razões e objetivos da proposta" surge expresso o duplo propósito da posição da União Europeia em relação às tecnologias de IA:

> Os mesmos elementos e técnicas que produzem os benefícios socioeconómicos da IA também podem trazer novos riscos ou consequências negativas para os cidadãos e a sociedade. À luz da velocidade da evolução tecnológica e dos possíveis desafios, a UE está empenhada em alcançar uma abordagem equilibrada. É do interesse da União preservar a liderança tecnológica da UE e assegurar que novas tecnologias, desenvolvidas e exploradas respeitando os valores, os direitos fundamentais e os princípios da União, estejam ao serviço dos cidadãos europeus. (Comissão Europeia, 2021, p. 1)

O primeiro quadro regulamentar da União Europeia para a IA foi proposto em abril de 2021 (Comissão Europeia, 2021). Na sequência de desenvolvimentos posteriores de tecnologias de IA (por exemplo, o lançamento, em novembro de 2022, dos

sistemas de IA generativa, como o *ChatGPT*), os legisladores da União Europeia envolveram-se em negociações para finalizar o novo Regulamento. Este foi publicado no Jornal Oficial da União Europeia a 12 de julho de 2024 (União Europeia, 2024), constituindo, nas palavras da União Europeia, a primeira lei do mundo a regular as tecnologias de IA. Desde a sua versão inicial tornada pública em 2021, o Regulamento foi conhecendo emendas substanciais à proposta inicial da Comissão Europeia, incluindo a revisão da definição de sistemas de IA, a ampliação da lista de sistemas de IA proibidos, e a imposição de obrigações em relação a IA de âmbito geral e em relação a modelos de IA generativos.

1.3. Hierarquização e tipologia de riscos da Inteligência Artificial

A União Europeia considera que determinadas características específicas da IA (opacidade, complexidade, dependência de dados e comportamento autónomo) podem afetar negativamente uma série de direitos protegidos pela Carta dos Direitos Fundamentais da União Europeia, bem como a segurança dos utilizadores quando as tecnologias de IA estão incorporadas em produtos e serviços. Constitui matéria altamente sensível o facto de os sistemas de IA poderem comprometer direitos fundamentais, como o direito à não discriminação, liberdade de expressão, dignidade humana, proteção de dados pessoais e privacidade.

Para responder a essas preocupações, o Regulamento da IA segue uma abordagem baseada no risco, ou seja, classifica as diferentes tecnologias de IA em termos dos riscos que possam colocar para os utilizadores, distinguindo entre tecnologias de IA que apresentam risco inaceitável, risco elevado ou risco não elevado. O objetivo é estabelecer, para cada uma destas categorias, requisitos e obrigações diferentes no que respeita o desenvolvimento, colocação no mercado e utilização de tecnologias de IA na União Europeia, adaptando a intervenção jurídica ao nível de risco definido, conforme descrevemos de seguida.

Risco inaceitável

As tecnologias de IA que apresentem riscos "inaceitáveis" serão proibidas. Neste contexto, o Regulamento da União Europeia proíbe todas as tecnologias de IA que possam veicular práticas manipuladoras, exploratórias e de controlo social:

> Essas práticas são particularmente prejudiciais e abusivas e deverão ser proibidas por desrespeitarem valores da União, como a dignidade do ser humano, a liberdade, a igualdade, a democracia e o Estado de direito, bem como os direitos fundamentais consagrados na Carta, nomeadamente o direito à não discriminação, à proteção de dados pessoais e à privacidade, e os direitos das crianças. (União Europeia, 2024, p. 8)

De acordo com esta proposta, será proibido colocar no mercado, colocar em serviços ou utilizar na União Europeia as seguintes tecnologias:

- Tecnologias de IA concebidas para manipular o comportamento humano, sendo passíveis de provocar danos físicos ou psicológicos e de distorcer substancialmente o comportamento de uma pessoa de uma forma que cause, ou seja suscetível de causar, danos a essa ou a outra pessoa. Concretamente, tecnologias de IA que utilizem componentes subliminares que não são detetáveis pelos seres humanos ou que exploram grupos vulneráveis, tais como crianças e adultos com incapacidades físicas ou mentais (União Europeia, 2024, p. 51).
- Tecnologias de IA que usem identificação biométrica remota em "tempo real", como o reconhecimento facial, em espaços acessíveis ao público para efeitos de manutenção da ordem pública, salvaguardando exceções restritas e limitadas no tempo e na duração para propósitos de aplicação da lei na busca direcionada de vítimas (rapto, tráfico, exploração sexual), prevenção de ataques terroristas, e localização ou identificação de suspeitos de crimes específicos (por exemplo, terrorismo, tráfico, exploração sexual, assassinato, sequestro, violação, assalto à mão armada, participação em organização criminal ou crime ambiental):

  > A utilização de sistemas de IA para a identificação biométrica à distância "em tempo real" de pessoas singulares em espaços acessíveis ao público para efeitos de aplicação da lei é particularmente intrusiva para os direitos e as liberdades das pessoas em causa, visto que pode afetar a vida privada de uma grande parte da população, dar origem a uma sensação de vigilância constante e dissuadir indiretamente o exercício da liberdade de reunião e de outros direitos fundamentais. As imprecisões técnicas dos sistemas de IA concebidos para a identificação biométrica à distância de pessoas singulares podem conduzir a resultados enviesados e ter efeitos discriminatórios. Estes possíveis resultados enviesados e efeitos discriminatórios são particularmente relevantes no que diz respeito à idade, etnia, raça, sexo ou deficiência. Além disso, dado o impacto imediato e as oportunidades limitadas para a realização de controlos adicionais ou correções no que respeita à utilização desses sistemas que funcionam em tempo real acarretam riscos acrescidos para os direitos e as liberdades das pessoas em causa no contexto, ou afetadas, pelas autoridades responsáveis pela aplicação da lei. (União Europeia, 2024, p. 9)

- Tecnologias de IA usadas para classificação social (União Europeia, 2024, p. 51). A classificação social refere-se ao processo de avaliação ou classificação de pessoas ou entidades com base em vários fatores sociais e comportamentais (por exemplo, o estatuto socioeconómico ou características de personalidade ou pessoais). Este conceito pode ser aplicado numa variedade de contextos, incluindo finanças, emprego e até mesmo em contextos sociais e políticos. O recurso à IA para classificação social normalmente envolve a utilização de algoritmos de aprendizagem automática (*machine learning*) para analisar grandes

quantidades de dados com o objetivo de fazer previsões ou avaliações sobre o comportamento, a fiabilidade ou a adequação de uma pessoa a determinadas oportunidades ou serviços. Eis alguns exemplos concretos:

Pontuação de crédito. A IA pode analisar o historial financeiro de uma pessoa, os seus hábitos de despesa e outros dados relevantes para prever a sua capacidade de crédito. Esta informação pode ser utilizada pelas instituições financeiras para determinar se aprovam ou não um empréstimo.

Seleção de emprego. A IA pode analisar a presença de uma pessoa nas redes sociais, as atividades *online* e outro tipo de registos de atividade em espaços públicos para fornecer informações sobre o seu comportamento e carácter. Esta informação pode ser utilizada pelos empregadores durante o processo de contratação.

Governo e serviços públicos. A IA pode ser utilizada para avaliar a elegibilidade para benefícios ou serviços sociais com base numa série de fatores, incluindo o rendimento, o estatuto profissional e outros indicadores socioeconómicos.

Cuidados de saúde e seguros. As seguradoras podem utilizar a IA para avaliar o risco de saúde de uma pessoa com base em fatores como o estilo de vida, a genética e o historial médico para determinar os prémios e a cobertura. As instituições de saúde podem usar a IA para cruzar dados clínicos e de estilo de vida dos cidadãos para tomar decisões em relação ao acesso a bens e serviços escassos.

O uso de IA para classificação social suscita vários riscos complexos, designadamente a invasão de privacidade, pela recolha massiva de informações sensíveis ou detalhadas sobre as atividades das pessoas, comportamentos e relações pessoais, assim como o viés e a discriminação, pelo uso de dados que refletem preconceitos existentes (de género, de classe, de raça e etnia), podendo perpetuar ou amplificar desigualdades e discriminação. Acresce quer a falta de transparência e explicabilidade, uma vez que pode ser difícil compreender completamente como é que certos sistemas de IA chegam a uma determinada decisão, o que pode dificultar a explicação de decisões aos afetados, quer a falta de responsabilização, pois se as decisões importantes forem totalmente automatizadas pela IA, pode haver falta de recursos e canais de recurso para pessoas que discordem ou desafiem essas decisões.

De realçar, ainda, a desumanização, quando decisões com impactos importantes na vida das pessoas são tomadas puramente com base em dados e IA, sem levar em consideração a complexidade das experiências humanas, e a manipulação comportamental, pois se as pessoas souberem que estão a ser avaliadas por um sistema de classificação social, podem ser incentivadas a ajustar o seu comportamento para se conformarem aos critérios estabelecidos, o que suprime uma dimensão importante da liberdade de expressão. Também a desigualdade de acesso a tecnologias digitais pode resultar em disparidades na classificação social, agravando as desigualdades existentes.

Risco elevado

As tecnologias de IA consideradas de risco elevado serão autorizadas, mas sujeitas a um conjunto de requisitos e obrigações para obter acesso ao mercado da União Europeia. Os fornecedores destas tecnologias de IA serão obrigados a registá-las numa base de dados à escala da União Europeia, gerida pela Comissão Europeia, antes de as colocarem no mercado ou em utilização efetiva.

Entram nesta classificação as tecnologias de IA que criam impactos adversos substanciais na saúde e na segurança das pessoas ou nos seus direitos fundamentais. A União Europeia (2024, pp. 14-26) distingue entre duas categorias de tecnologias de IA de risco elevado:

- Os sistemas utilizados como componente de segurança de um produto ou abrangidos pela legislação de harmonização em matéria de saúde e segurança da União Europeia (por exemplo, brinquedos, aviação, automóveis e dispositivos médicos).
- Determinados sistemas usados em oito domínios específicos: Identificação biométrica; gestão e funcionamento de infraestruturas críticas; educação e formação profissional; emprego, gestão de trabalhadores e acesso ao emprego por conta própria; acesso e usufruto de serviços privados essenciais e de serviços e prestações públicas essenciais; aplicação da lei; controlo da migração, do asilo e do controlo das fronteiras; administração da justiça e processos democráticos (União Europeia, 2024, pp. 127-129). Eis alguns exemplos:

Gestão e funcionamento de infraestruturas críticas: Tecnologias de IA concebidas para serem utilizadas como componentes de segurança na gestão e no controlo do trânsito rodoviário e das redes de abastecimento de água, gás, aquecimento e eletricidade.

Educação e formação profissional: Tecnologias de IA utilizadas para fins de determinação de acesso a instituições de ensino e de formação ou utilizadas para fins de avaliação de estudantes.

Emprego e gestão de trabalhadores: Tecnologias de IA concebidas para serem utilizadas no recrutamento ou na seleção de pessoas, avaliação de candidatos, ou para serem utilizadas na tomada de decisões sobre promoções ou cessações de relações contratuais de trabalho, na repartição de tarefas e no controlo e avaliação do desempenho.

Acesso a serviços privados e a serviços e prestações públicas essenciais: Tecnologias de IA concebidas para serem utilizadas por autoridades públicas para avaliar a elegibilidade de pessoas singulares quanto a prestações e serviços públicos de assistência; ou para avaliar a capacidade de endividamento de pessoas ou estabelecer a sua classificação de crédito.

Manutenção da ordem pública: Tecnologias de IA concebidas para serem utilizadas por autoridades policiais em avaliações individuais de riscos relativamente a pessoas

singulares, a fim de determinar o risco de uma pessoa cometer infrações, para detetar o estado emocional de uma pessoa; e ainda tecnologias de IA concebidas para serem utilizadas no estudo analítico de crimes, permitindo às autoridades policiais pesquisar grandes conjuntos de dados complexos, disponíveis em diferentes fontes de dados ou em diferentes formatos de dados, no intuito de identificar padrões desconhecidos ou descobrir relações escondidas nos dados.

Risco não elevado

Tecnologias de IA concebidas para interagir com pessoas singulares, ou que geram ou manipulam conteúdos ("falsificações profundas") podem ser consideradas de risco elevado ou não elevado. As pessoas devem ser informadas quando interagem com uma tecnologia de IA ou quando as suas emoções ou características são reconhecidas por meios automatizados através do tratamento dos seus dados biométricos. Estas tecnologias de IA estarão sujeitas a um conjunto limitado de obrigações de transparência.

Todas as tecnologias de IA que forem consideradas como apresentando um "risco não elevado" poderão ser desenvolvidas e utilizadas na União Europeia, prevendo-se a criação de códigos de conduta para incentivar os seus fornecedores a aplicarem voluntariamente os requisitos obrigatórios para tecnologias de IA de risco elevado.

Ainda que a ideia de uma regulamentação adequada da IA seja amplamente reconhecida como necessária e louvável, este Regulamento da União Europeia não clarifica como é que a classificação de risco pode mudar à medida que vão surgindo inovações nas tecnologias de IA. Para além disso, nem sempre é possível distinguir previamente e de modo definitivo entre risco elevado e risco não elevado: Depende da utilização dada a uma determinada tecnologia de IA, do contexto em que esta será aplicada e de quem são as pessoas afetadas por esse uso.

Importa, ao mesmo tempo, reconfigurar a cultura de "institucionalização do risco" (Beck, 1992; Giddens, 1990) que subjaz ao Regulamento da IA da União Europeia, visível na forma como o risco é essencialmente concebido como um assunto técnico e científico objetivamente mensurável e convertido numa ferramenta institucional de gestão de controvérsias públicas ao nível das decisões políticas. Neste contexto, problemas sociais e éticos complexos são limitados apenas ao risco tal qual este é definido pelos especialistas considerados como tendo legitimidade e capacidade política de regular e legislar, incluindo a ciência institucional e a indústria e organizações comerciais, o que suscita debates sobre outros possíveis significados públicos de risco que podem estar a ser negligenciados ou invisibilizados. Por exemplo, preocupações fundamentais cuja "resolução" dificilmente encaixa em soluções técnicas raramente são mencionadas, incluindo temáticas relacionadas com a sustentabilidade, a solidariedade, o cuidado e bem-estar, e a responsabilidade social (Steinhoff, 2023). Ao acionar uma epistemologia de risco de natureza tecnocrática, a governação das tecnologias de IA na União Europeia tem contribuído, na nossa perspetiva, para reproduzir a opacidade de agendas políticas

e de interesses económicos e comerciais por detrás de compromissos normativos (Felt e Wynne, 2007, pp. 16-17).

Veremos, ao longo dos próximos capítulos, como as perspetivas das ciências sociais conferem visibilidade à materialidade, muitas vezes negligenciada (de la Bellacasa, 2011), dos riscos mais problemáticos das tecnologias de IA, situando-os em realidades práticas e enquadrando-os em relações de poder institucional, organizacional, grupal e interpessoal profundamente desiguais e assimétricas (Crawford, 2024 [2021]; Villegas-Galaviz e Martin, 2023).

2. Implicações éticas e sociais da Inteligência Artificial

Introdução

Nos últimos anos, publicaram-se centenas de diretrizes éticas que apelam a uma reflexão ampla, inclusiva, transparente e democrática em torno das implicações éticas e sociais da Inteligência Artificial (IA). Steinhoff (2023), por exemplo, identificou 167 documentos de diretrizes éticas de algum modo relacionadas com IA em todo o mundo disponíveis no início de 2020[17]. Entre estas diretrizes destacam-se, pela visibilidade alcançada tanto em meio industrial como académico, as seguintes:

Primeiro, os 23 "Princípios de Asilomar", desenvolvidos em 2017 pelo Instituto do Futuro da Vida (*Future of Life Institute*), na sequência de um encontro que reuniu investigadores em IA e representantes da indústria, a conhecida Conferência de Asilomar[18]. Estes princípios concentram-se em questões como o impacto da IA na economia, a distribuição de benefícios e a necessidade de evitar corridas ao armamento na área da IA.

Segundo, a "Declaração de Montreal para um Desenvolvimento Responsável da Inteligência Artificial", publicada em 2018 no contexto de uma iniciativa da Universidade de Montreal com o apoio do Instituto Quebec de IA, que reuniu uma equipa de trabalho interuniversitária e multidisciplinar[19]. Apoiada num processo deliberativo inclusivo, que colocou em diálogo cidadãos, especialistas, responsáveis públicos, organizações da sociedade civil, ordens profissionais e partes interessadas no desenvolvimento da IA, a Declaração de Montreal contempla 10 princípios fundamentais, designadamente: Bem-estar; respeito pela autonomia; proteção da privacidade e da vida íntima; solidariedade; participação democrática; equidade; inclusão da diversidade; prudência; responsabilidade; e desenvolvimento sustentável.

Terceiro, as "Orientações Éticas para uma IA de Confiança", publicadas em 2019 como resultado do trabalho de um Grupo de Peritos de Alto Nível sobre a IA (GPAN IA ou, em inglês, AI HLEG), designados pela Comissão Europeia em junho de 2018[20]. Este documento perspetiva uma IA de confiança como legal, ética e sólida, ou seja, uma IA que deve cumprir toda a legislação e regulamentação aplicáveis, garantir a observância de princípios e valores éticos, e ser técnica e socialmente sólida, uma vez que pode causar danos não intencionais. Uma IA ética deverá orientar-se por sete princípios fundamentais: Ação e supervisão humanas; solidez técnica e segurança; privacidade e governação dos dados; transparência; diversidade, não discriminação e equidade; bem-estar societal e ambiental; e responsabilização.

---

17 É desconhecido o número exato de diretrizes éticas sobre IA. As estimativas variam consideravelmente, dependendo do critério de seleção dos documentos.

18 Os Princípios de Asilomar estão disponíveis em https://futureoflife.org/open-letter/ai-principles/ [Acesso a 24 de julho de 2024].

19 A Declaração de Montreal está disponível em https://docs.wixstatic.com/ugd/ebc3a3_d806f109c4104c91a2e719a7bef77ce6.pdf [Acesso a 24 de julho de 2024].

20 As "Orientações éticas para uma IA de confiança" estão disponíveis em https://data.europa.eu/doi/10.2759/2686 [Acesso a 24 de julho de 2024].

Estes documentos partilham o objetivo comum de fornecerem uma base ética para orientar a pesquisa, desenvolvimento e implementação de tecnologias de IA, visando assegurar que estas sejam utilizadas para o benefício da sociedade como um todo, de forma responsável. A importância dos princípios de responsabilidade e de confiança é enfatizada por diversas organizações, como a OCDE (2023), a UNESCO (2021), o Fórum Económico Mundial (2023), e a Comissão Europeia (2020). A participação de diversos públicos na governação da IA é perspetivada, neste contexto, como um elemento fundamental para garantir que as tecnologias de IA estejam alinhadas com valores sociais e orientadas para a criação de um futuro mais justo, equitativo e inclusivo, sustentando uma IA centrada no ser humano (*human-centric AI*)[21] (Sigfrids et al., 2023).

A Recomendação do Conselho da OCDE sobre a IA, por exemplo, realça a importância de capacitar as partes interessadas (*stakeholders*) para se envolverem na governação da IA, considerando que este envolvimento é essencial para sustentar uma IA confiável (OCDE, 2023). Já a Recomendação da UNESCO sobre a Ética da IA salienta que a conscientização e a compreensão públicas das tecnologias de IA e do valor dos dados devem ser promovidas (recomendação 44), e encoraja abordagens inclusivas na governação da IA, onde a participação de diferentes públicos nos processos de decisão é considerada necessária (recomendação 47) (UNESCO, 2021, p. 23). Também o Livro Branco sobre a IA da Comissão Europeia (2020) inclui a consulta pública na abordagem proposta para a confiança e a excelência, em alinhamento com as orientações éticas desenvolvidas pelo Grupo de Peritos de Alto Nível sobre a IA[22], que evidenciam a importância da consulta e da participação públicas no desenho, desenvolvimento e implementação de tecnologias de IA. Em suma, a participação pública afigura-se como um princípio básico consensual plasmado nos principais documentos éticos emanados de instituições internacionais e transnacionais (Ulnicane, 2022; Ulnicane et al., 2021a, 2021b).

Estas iniciativas, à semelhança de outras que propõem princípios universais de ética de IA, têm sido criticadas por instrumentalizarem as ideias de IA de confiança e de IA responsável. E fazem-no, de acordo com estas vozes críticas, por reduzirem a ética a uma forma de capital industrial ou por cooptarem investigadores do domínio da ética como parte de um conjunto de esforços para controlar as narrativas públicas (Hagendorff, 2020; Ochigame, 2019). Subjaz a estes esforços a mobilização simbólica da ética para legitimar a inovação tecnológica e cativar o apoio e a confiança pública,

21 O conceito de IA centrada no humano (*human-centric AI*) visa assegurar que os valores humanos são incorporados no desenho dos algoritmos, que os seres humanos não perdem o controlo sobre os sistemas automatizados, e que a IA é usada em prol da humanidade e do bem-comum para melhorar o bem-estar e os direitos humanos (Sigfrids et al., 2023).

22 Para além do relatório intitulado "Orientações éticas para uma IA de confiança" (GPAN IA, 2019b), os peritos recrutados pela Comissão Europeia produziram um conjunto de 33 recomendações para "orientar uma IA fiável para a sustentabilidade, o crescimento, a competitividade e a inclusão. Ao mesmo tempo, as recomendações capacitarão, beneficiarão e protegerão os cidadãos europeus" (GPAN IA, 2019c), assim como "uma ferramenta prática que traduz as orientações éticas numa lista de verificação de autoavaliação acessível e dinâmica. A lista de verificação pode ser utilizada por programadores e implantadores de IA que pretendam implementar os requisitos-chave" (GPAN IA, 2020).

tranquilizando as críticas ao confiná-las à agenda científica e industrial (Ferretti, 2022; Hagendorff, 2020; Phan et al., 2022; van Maanen, 2022).

A tendência para a cooptação da ética na governação da IA tem-se concretizado através de duas estratégias principais. Primeiro, ao ser mobilizada para projetar a ideia de que uma IA responsável e em harmonia com os valores sociais será alcançada por via do envolvimento e participação pública (Weingart et al., 2021), invisibilizando a reflexão sobre quem serão os públicos efetivamente envolvidos e de forma a definição de categorias que operam distinções entre quem é ou não é envolvido revela relações políticas e de poder (Sieber et al., 2024). Segundo, ao acionar a natureza híbrida e ambivalente da ética enquanto matéria reservada a especialistas em ética, por um lado, e matéria relacionada com o social em sentido amplo e difuso, por outro lado, o que relega a ética para contornos cada vez mais vagos e imprecisos, cooptados para uso político. Estas estratégias são patrocinadas por grandes empresas tecnológicas, políticos e outros grupos de interesse, que se alinham para manter ininterrupto o desenvolvimento de tecnologias de IA (Benkler, 2019; Phan et al., 2022; Steinhoff, 2023).

Uma das iniciativas mais analisadas por comentadores críticos para ilustrar estes processos prende-se com o trabalho do Grupo de Peritos de Alto Nível sobre a IA (GPAN IA), cujos elementos foram recrutados pela Comissão Europeia num processo aberto para prestar aconselhamento sobre a estratégia na elaboração de políticas de IA e para produção de recomendações que pudessem orientar os Estados-membros nessa matéria. Este grupo, composto por 52 peritos (desde académicos a membros de ONGs e antigos funcionários públicos), acabou por ser largamente dominado por representantes da indústria, com uma representação limitada da sociedade civil e do mundo académico (Ulnicane et al., 2021b). Considerando que os representantes da indústria que desenvolve IA falavam sobre ética como uma estratégia para atrasar ou evitar regulamentação vinculativa, as diretrizes éticas desenvolvidas por esta rede de atores mobilizada pela Comissão Europeia foram criticadas como "lavagem ética" (Metzinger, 2019). Uma ética marcada por interesses comerciais tem-se consolidado e legitimado através da contratação, por parte da indústria, de pessoas que trabalham em institutos de investigação, em universidades e outras instituições consideradas independentes para lidar com as implicações éticas e sociais da IA.

Neste contexto, importa refletir sobre três questões fundamentais: Que valores éticos e sociais são incorporados nas tentativas de regular, a uma escala global, o desenvolvimento de tecnologias de IA? De que forma é que o apelo ao envolvimento de diversos atores individuais e institucionais, incluindo cidadãos e organizações da sociedade civil, na governação da IA, tem sido vertido na definição de políticas para a regulação e legislação do desenvolvimento e implementação de tecnologias de IA? Quem é que tem sido envolvido na discussão sobre as implicações éticas e sociais da IA, e que tópicos têm dominado o debate?

Neste capítulo procuramos responder a estas questões. Numa primeira parte, detalhamos os principais princípios e valores éticos e sociais plasmados nos requisitos

propostos pelo Grupo de Peritos de Alto Nível sobre a IA (GPAN IA) para orientar uma "IA de confiança". Procedemos a uma reflexão crítica sobre os desafios que esta abordagem principialista (isto é, baseada em princípios gerais, abstratos e prescritivos) em torno das implicações éticas e sociais da IA suscita para a cidadania. Numa segunda parte, exploramos os papéis que têm sido atribuídos ao público em diversas iniciativas regulamentares, assim como em processos de produção de conhecimento e inovação. Com o propósito de contextualizar os conteúdos dos capítulos seguintes, concluímos com a enunciação de uma abordagem alternativa às implicações éticas e sociais da IA, assente numa ética de cuidado, cujos pilares são a diversidade, a inclusão, a solidariedade e o bem-estar social.

Em síntese, neste capítulo pretendemos:

- Apresentar, de forma crítica, os princípios éticos e os requisitos básicos universais a respeitar no âmbito da pesquisa, desenvolvimento e implementação de tecnologias de IA propostos pelo Grupo de Peritos de Alto Nível sobre a IA (GPAN IA).
- Explorar os papéis que são atribuídos ao público em diversas iniciativas regulamentares.
- Analisar a forma como práticas concretas de investigação incorporam e traduzem as perspetivas públicas sobre as implicações éticas e sociais da IA.
- Propor uma abordagem às implicações éticas e sociais da IA assente numa ética de cuidado.

2.1. Princípios éticos universais

Com o principal objetivo de promover uma IA de confiança, ou seja, uma IA legal, ética e sólida, o Grupo de Peritos de Alto Nível sobre a IA (GPAN IA), constituído pela Comissão Europeia em 2018, elaborou um conjunto de orientações éticas enraizadas em direitos fundamentais e em direitos dos cidadãos, como o respeito da dignidade humana, da democracia, da justiça e do Estado de direito, a liberdade do indivíduo, a igualdade, a não discriminação e a solidariedade. Os princípios éticos contemplados nestas diretrizes são os seguintes: 1) *Respeito da autonomia humana*, que inclui o respeito da liberdade, autodeterminação e escolha humanas; 2) *prevenção de danos*, que visa a proteção da dignidade e integridade humanas, do ambiente natural e de todos os seres vivos; 3) *equidade*, que implica uma distribuição equitativa e justa de benefícios e custos, e a inexistência de enviesamentos injustos, discriminação e estigmatização; e 4) *explicabilidade*, que se refere à transparência dos processos e respetiva comunicação aberta, de modo a tornar as decisões explicáveis.

Estes princípios traduzem-se em imperativos éticos que deverão ser respeitados no âmbito da pesquisa, desenvolvimento e implementação de tecnologias de IA. De acordo com a proposta do Grupo de Peritos de Alto Nível sobre a IA (GPAN IA), estes imperativos passam pela centralidade do ser humano, que deverá assumir a supervisão e controlo de tecnologias de IA, assim como pela segurança de ambientes naturais e técnicos e de pessoas, em especial aquelas que se encontram em situação

de maior vulnerabilidade (por exemplo, crianças, pessoas com deficiência e outros grupos historicamente desfavorecidos ou em risco de exclusão) ou que estão envolvidas em relações sociais de poder assimétricas:

> Os sistemas de IA não devem subordinar, coagir, enganar, manipular, condicionar ou arregimentar injustificadamente os seres humanos. Em vez disso, os sistemas de IA devem ser concebidos para aumentar, complementar e capacitar as competências cognitivas, sociais e culturais dos seres humanas. A distribuição de funções entre os seres humanos e os sistemas de IA devem seguir princípios de conceção centrados no ser humano e deixar uma oportunidade significativa para a escolha humana. Isto implica que se garanta a supervisão e o controlo por parte de seres humanos sobre os processos de trabalho dos sistemas de IA. (GPAN IA, 2019b, p. 15)

> Os sistemas de IA e os ambientes em que operam devem ser seguros e protegidos. Devem ser tecnicamente sólidos e deve garantir-se que não estão abertos a utilizações malévolas. As pessoas vulneráveis devem receber maior atenção e ser incluídas no desenvolvimento e na implantação dos sistemas de IA. Há também que prestar especial atenção às situações em que os sistemas de IA podem causar ou agravar impactos negativos devido a assimetrias de poder ou de informação, nomeadamente entre empregadores e trabalhadores, empresas e consumidores ou governos e cidadãos. (GPAN IA, 2019b, p. 15)

Acrescem, ainda, imperativos relacionados com a promoção da igualdade, da proporcionalidade e da possibilidade de contestação, independentemente das escolhas individuais, além da diversificação de medidas que potenciem a explicabilidade de resultados ou decisões:

> Se for possível evitar os enviesamentos, os sistemas de IA podem até aumentar a equidade societal. A igualdade de oportunidades em termos de acesso à educação, aos bens e serviços e à tecnologia deve ser igualmente promovida. (...) A utilização de sistemas de IA nunca deverá levar a que os utilizadores (finais) sejam iludidos ou prejudicados na sua liberdade de escolha. (...) Os profissionais no domínio da IA devem respeitar o princípio da proporcionalidade entre os meios e os fins, e analisar cuidadosamente a forma de equilibrar os interesses e objetivos em causa. A (...) equidade implica uma possibilidade de contestar e procurar vias de recurso eficazes contra as decisões tomadas por sistemas de IA e pelos seres humanos que os utilizam. Para o efeito, a entidade responsável pela decisão deve ser identificável e os processos decisórios explicáveis. (GPAN IA, 2019b, p. 15)

> Nem sempre é possível explicar por que razão um modelo gerou determinado resultado ou decisão (e que combinação de fatores de entrada contribuiu para esse efeito). (...) Nessas circunstâncias, podem ser necessárias outras medidas da explicabilidade (p. ex., a rastreabilidade, a auditabilidade e a comunicação transparente sobre as capacidades do sistema) (...). (GPAN IA, 2019b, p. 16)

Na perspetiva deste grupo de peritos, as tecnologias de IA devem cumprir sete requisitos concretos, que descrevemos de forma sumária na Tabela 4[23]. De modo a garantir a sua aplicação e melhoria sistemática, com recurso a métodos técnicos e não técnicos, apela-se à promoção de investigação e inovação em IA, assim como à reflexão e ao debate alargados, a nível mundial, sobre o enquadramento ético da IA.

Tabela 4
Principais requisitos para uma Inteligência Artificial de confiança.

Requisitos	Alguns indicadores
Ação e supervisão humanas	Apoiar a autonomia e a tomada de decisões informadas de seres humanos.
	Permitir a supervisão humana, que pode ser realizada mediante mecanismos de governação como as abordagens de intervenção humana (*human-in-the-loop*), de fiscalização humana (*human-on-the-loop*), ou de controlo humano (*human-in--command*).
	Possibilitar aos utilizadores a avaliação e/ou a contestação.
Solidez técnica e segurança	Prevenir riscos e danos inaceitáveis.
	Resiliência perante ataques e proteção contra vulnerabilidades.
	Possuir salvaguardas que possibilitem planos de recurso perante problemas.
	Minimizar e prevenir consequências não intencionais e inesperadas (por exemplo, aplicações de dupla utilização ou intervenientes mal-intencionados) e erros.
	Fazer apreciações e previsões corretas (exatidão).
	Fiabilidade e reprodutibilidade dos resultados.
Privacidade e governação dos dados	Garantir a privacidade e a proteção de dados.
	Assegurar a qualidade e a integridade dos conjuntos de dados utilizados.
	Adotar protocolos de governação do acesso aos dados (quem pode aceder aos dados e em que circunstâncias).
Transparência	Transparência dos dados, do sistema e dos modelos de negócio.
	Permitir a rastreabilidade (dos processos de recolha e etiquetagem dos dados, dos algoritmos utilizados, e das decisões tomadas pelo sistema de IA).
	Explicar tanto os processos técnicos como as decisões humanas com eles relacionadas, de modo a que sejam compreendidos pelos utilizadores.
	Informar os utilizadores de que estão a lidar com um sistema de IA.
	Comunicar as capacidades e limitações do sistema de IA às partes interessadas, permitindo-lhes criar expectativas realistas.
	Divulgar os resultados e as questões em aberto junto do público em geral.

23 Esta lista de requisitos não é exaustiva, e encontra-se em permanente atualização. Todos os requisitos têm igual importância e encontram-se interligados. A sua aplicação e avaliação deve acontecer ao longo de todo o ciclo de vida das tecnologias de IA.

Diversidade, não discriminação e equidade	Prevenção de enviesamentos injustos (associados à inclusão de preconceitos já existentes, lacunas e maus modelos de governação, entre outros), que podem originar e/ou reforçar a discriminação, a marginalização e preconceitos dirigidos a determinados grupos ou pessoas em situação de vulnerabilidade.
	Garantir processos de conceção inclusivos e a acessibilidade universal, independentemente das capacidades ou das características das pessoas.
	Envolver todas as partes interessadas em todo o ciclo de vida do sistema de IA.
Bem-estar societal e ambiental	Ser utilizada em benefício de todas as pessoas, incluindo as gerações futuras.
	Respeitar o ambiente, optando por escolhas sustentáveis.
	Considerar os impactos ao nível das relações e normas sociais.
	Avaliar os efeitos nas instituições, na democracia e na sociedade em geral (por exemplo, em situações relacionadas com processos eleitorais).
Responsabilização	Criar mecanismos para garantir a responsabilidade e a responsabilização pelos sistemas de IA e os seus resultados, tanto antes como depois da sua adoção.
	Possibilitar a auditabilidade, ou seja, a avaliação de algoritmos, dados e processos de conceção, sobretudo em aplicações ou situações críticas.
	Identificar, avaliar, comunicar e minimizar potenciais impactos negativos.
	Fundamentar, documentar e rever continuamente as soluções de compromisso adotadas para resolver conflitos.
	Prever mecanismos acessíveis para assegurar vias de recurso adequadas perante a ocorrência de um impacto adverso injusto.

Fonte. Adaptado de GPAN IA, 2019b, pp. 17-25.

Antecipando a existência de eventuais dilemas éticos e/ou conflitos entre diferentes princípios e requisitos, o GPAN IA prevê a respetiva identificação, avaliação, documentação e comunicação contínua. A ideia será alcançar "soluções de compromisso eticamente aceitáveis", que deverão resultar da aplicação de "métodos de deliberação responsável". De acordo com o GPAN IA, esta proposta alinha-se com o ideal de participação política aberta e democrática que caracteriza a União Europeia. Trata-se de procurar um "consenso mundial" enquadrado numa abordagem baseada nos direitos fundamentais:

> Nem a utilização dos sistemas de IA nem o seu impacto conhecem fronteiras nacionais. Por conseguinte, são necessárias soluções a nível mundial para as oportunidades e os desafios globais resultantes da IA. Incentivamos, assim, todas as partes interessadas a trabalharem em prol da criação de um quadro mundial para uma IA de confiança, estabelecendo um consenso internacional, ao mesmo tempo que promovem e defendem a nossa abordagem baseada nos direitos fundamentais. (GPAN IA, 2019b, pp. 6-7)

Se a mobilização de direitos humanos internacionais pode constituir uma fonte de autoridade para responsabilizar os criadores e/ou implantadores de tecnologias de

IA, a sua aplicação ao nível da governação da IA tem-se revelado pouco eficaz e incapaz de promover mudanças estruturais (Su, 2022). De facto, as abordagens académicas que procuram mapear e compreender os princípios éticos e os valores sociais incorporados nas tentativas de regular, a uma escala global, o desenvolvimento de tecnologias de IA, apontam para uma lacuna assinalável entre a enunciação de princípios abstratos e a concretização de práticas de operacionalização que assegurem o efetivo desenvolvimento do potencial da IA para uma distribuição de benefícios sustentável, solidária e orientada para o bem-estar de toda a sociedade (Hagendorff, 2020; Jobin et al., 2019; Newman, 2020; Resseguier e Rodrigues, 2021).

Ainda que o GPAN IA proclame a tentativa de cruzar a enunciação de princípios éticos abstratos universais com indicações sobre a forma de operacionalizar tais princípios em sistemas sociotécnicos, diversas vozes críticas mostram como a produção destas diretrizes se orientou por uma abordagem principialista (Resseguier e Rodrigues, 2021), excluindo propostas e visões alternativas (Heilinger, 2022; Roche et al., 2022). Isto significa olhar para a ética como uma réplica suavizada do direito e, como tal, traduzida em princípios gerais, abstratos e prescritivos. Ora, uma abordagem principialista torna a ética mal equipada para lidar com práticas concretas. Mais, esta abordagem potencia uma desconexão entre a ética e os impactos sociais, políticos e materiais da IA, nomeadamente o recrudescimento de desigualdades sociais e raciais, injustiças e danos ambientais (Munn, 2022).

2.2. Contornos da participação e envolvimento dos públicos

As referências à importância da participação e do envolvimento dos públicos predominam nas estratégias políticas de IA. Porém, diversos estudos mostram que a menção que os documentos de natureza política fazem à participação e ao envolvimento de diversos públicos tende a ser abstrata e é frequentemente ofuscada, quer por outros papéis que são atribuídos ao público, quer por outros valores e preocupações políticas. Wilson (2022), por exemplo, mostra como nas estratégias nacionais de IA levadas a cabo por instituições governamentais de 16 países[24], o público é frequentemente perspetivado em papéis distintos: Utilizadores de serviços e produtos associados a IA; destinatários de benefícios abstratos da IA; força de trabalho que precisa de qualificação e formação; ou um elemento importante na sustentação de uma sociedade democrática próspera que desbloqueia as potencialidades da IA. O autor alerta, ainda, para a forma como a participação pública na governação da IA é enunciada nas estratégias que a contemplam: Trata-se mais de um gesto retórico ou de uma reflexão tardia do que um compromisso claro com o envolvimento de diversos públicos no desenho e implementação da IA (Wilson, 2022, pp. 7-8). Na verdade, apenas três das 16 estratégias nacionais de IA analisadas por Wilson (2022, pp. 4-5)

24 Wilson (2022) analisou documentos escritos e publicados por instituições governamentais dos seguintes países: Alemanha; China; Coreia do Sul; Dinamarca; Estados Unidos da América; Estónia; Finlândia; Holanda; Hungria; Luxemburgo; Noruega; Portugal; Reino Unido; República Checa; Suécia; Uruguai.

foram produzidas com a colaboração formal de ONGs ou de grupos constituídos por múltiplas partes interessadas (*multi-stakeholders groups*).

Também Ulnicane e colegas (2021b) exploraram os enquadramentos que sustentam o apelo à participação pública em 49 documentos políticos dedicados à IA. Os autores revelam a existência de elevadas expectativas quanto à possibilidade de a participação pública poder representar uma solução para resolver preocupações relacionadas com a concentração de poder, o recrudescimento das desigualdades, a falta de diversidade, e enviesamentos. Até hoje, porém, são escassas as considerações sobre como lidar, na prática, com desafios bem conhecidos por quem tem estado envolvido em iniciativas de participação pública em ciência e tecnologia, designadamente a dificuldade em alcançar um consenso entre diversas visões societais, os elevados recursos que a concretização de exercícios de participação pública exige, e os riscos de captação por interesses instalados (Ulnicane et al., 2021b, pp. 170-171). Permanece, assim, polémica a ideia de uma configuração visível e palpável de estratégias políticas de IA produzidas com o envolvimento de diversos públicos.

Mas para compreendermos de forma abrangente o lugar da participação e do envolvimento dos públicos no contexto de políticas de ciência e tecnologia, como é o caso da IA, é preciso articular, como sugerem Macq e colegas (2020), a análise da participação em processos de tomada de decisão política (*participation in decision-making processes*) com a participação em processos de produção de conhecimento e inovação (*participation in knowledge and innovation-making processes*). Isto significa que também é necessário explorar como é que as práticas concretas de investigação incorporam o envolvimento e a participação de diversos públicos e traduzem as perspetivas públicas sobre as implicações éticas e sociais da IA.

Numa revisão sistemática que mapeia o cenário dos estudos empíricos realizados acerca das visões dos públicos sobre os desafios éticos da IA, Machado e colegas (2023) mostram que as principais motivações subjacentes ao envolvimento dos públicos nestes estudos prendem-se com a promoção da inovação e da legitimação, visando, por um lado, a coprodução de conhecimento sobre as implicações éticas e sociais da IA ao incluir saberes localizados dentro e fora da esfera da ética "formal" e, por outro lado, a promoção da confiança pública e da aceitabilidade da IA e das políticas que a apoiam. São pouco frequentes as motivações relacionadas com a educação dos públicos ou com o seu empoderamento para participar na IA, a disseminação do interesse pela IA ou a politização, isto é, a abordagem de injustiças e exclusões históricas. As autoras identificam, ainda, os públicos que são convidados a pronunciar-se sobre as implicações éticas e sociais da IA, e concluem que o envolvimento de audiências não científicas, em particular utilizadores (reais ou potenciais) de produtos e serviços relacionados com IA, coexiste com o envolvimento de grupos profissionais (designadamente profissionais de saúde) e de responsáveis pelo desenvolvimento de tecnologias de IA (criadores e/ou implantadores). Esta coexistência evidencia a necessidade de inclusão de diversos públicos, enquanto se salvaguarda o conhecimento dos especialistas em IA.

Machado e colegas (2023) revelam, por fim, disparidades na atenção que é dedicada aos diferentes desafios éticos da IA. Observa-se um menor enfoque na diversidade, não discriminação, equidade, e bem-estar societal e ambiental, por comparação com tópicos relacionados com o desenvolvimento de uma IA centrada no ser humano, a privacidade e a governação dos dados alcançadas por via de métodos técnicos. Este balanço ilustra uma mudança nos sistemas periciais que enquadram a IA, mais afastados dos tradicionais especialistas em regulação e cada vez mais próximos dos engenheiros de privacidade e assessores de riscos. Ainda assim, o reconhecimento da solidez técnica e segurança, da transparência, e da responsabilização como requisitos éticos fundamentais da IA mostra como os públicos são sensíveis a limitações associadas aos sistemas periciais, abrindo espaço para políticas de otimização de algoritmos, num contexto em que as tecnologias de IA são perspetivadas como corrigíveis e em constante evolução.

Os escassos estudos que avaliam as perspetivas públicas sobre as implicações éticas e sociais da IA com base em inquéritos representativos da população tendem a apontar no mesmo sentido, mostrando a complexidade e a variabilidade das visões públicas em função de diferentes utilizações da IA (ver, por exemplo, Ada Lovelace Institute[25] and The Alan Turing Institute[26], 2023; Awad et al., 2020; Dupont et al., 2023; Kieslich et al., 2022; Ploug et al., 2021; Willems et al., 2022, 2023). Prevalecem visões positivas sobre a maioria das tecnologias de IA, com benefícios esperados, em particular, nas áreas da saúde, da ciência e da segurança, desde que subordinadas à ação e supervisão humanas. Já as preocupações públicas concentram-se nas aplicações associadas à robótica avançada (nomeadamente a existência de veículos sem condutor e de armas autónomas) e à educação (como o uso de IA para realizar exames e trabalhos em casa), traduzindo inquietações mais abrangentes quanto à falta de transparência e à responsabilização pelos sistemas de IA e os seus resultados, clamando pela necessidade de assegurar aos utilizadores a possibilidade de avaliação e/ou contestação através de vias de recurso acessíveis. A regulamentação é considerada necessária, sobretudo para proteger direitos fundamentais como a privacidade.

2.3. Para uma ética de cuidado

Converter em práticas concretas os princípios gerais, abstratos e prescritivos associados a uma abordagem principialista é fundamental para conectar a ética com os impactos sociais, políticos e materiais da IA. Com propósitos semelhantes, ainda que aplicáveis a inovações tecnológicas distintas, diversos académicos associados aos Estudos Sociais da Ciência e Tecnologia têm sugerido o acionamento de uma ética de cuidado (ver, por exemplo, de la Bellacasa, 2011; Gill et al., 2017; Kerr et al., 2018;

25 O *Ada Lovelace Institute* é um instituto de investigação independente sediado no Reino Unido cuja missão declarada consiste em assegurar que a IA produz resultados para as pessoas e para a sociedade, ou seja, que as oportunidades, benefícios e privilégios gerados pela IA são distribuídos e experienciados de forma justa e equitativa.

26 O *Alan Turing Institute* é o instituto nacional de ciência de dados e IA do Reino Unido. Fundado em 2015, tem como propósito dar passos significativos e fazer grandes progressos no desenvolvimento e uso da ciência de dados e da IA para mudar o mundo para melhor.

Lindén e Lydahl, 2021; Martin et al., 2015). Na esteira destas propostas, refletimos de seguida sobre os contributos que uma ética de cuidado pode proporcionar para reconfigurar o debate em torno das implicações éticas e sociais da IA. Esta reflexão serve de mote para contextualizar a discussão em torno da IA na educação, na saúde e na justiça, os três campos concretos da vida social abordados na segunda parte deste livro.

Assente em quatro pilares fundamentais – diversidade, inclusão, solidariedade e bem-estar social –, a ética de cuidado é sensível à complexidade dos problemas coletivos. Essa sensibilidade resulta da atenção que confere aos contextos particulares, às relações sociais concretas e às configurações morais individuais e coletivas envolvidas nas tecnologias de IA (Resseguier e Rodrigues, 2021; Villegas-Galaviz e Martin, 2023). Este enquadramento introduz duas mudanças importantes nas abordagens que têm dominado a produção de orientações éticas para a IA. Primeiro, promove o envolvimento, em particular, das comunidades mais afetadas e vulneráveis, cujas vozes raramente têm expressão na governação da ciência e tecnologia. Segundo, entrecruza as tecnologias de IA, as pessoas e o ambiente na avaliação dos riscos, em alinhamento com uma perspetiva "mais-do-que-humana" (Gill et al., 2017; Latimer e Gomez, 2019; Martin et al., 2015).

Orientar a ética para o cuidado significa assumir que cuidar é, em simultâneo, um compromisso ético-político e uma prática material situada num contexto (Lindén e Lydahl, 2021). Por outras palavras, a ética de cuidado dá visibilidade à materialidade dos impactos sociais, políticos e ambientais mais problemáticos das tecnologias de IA, desde o seu desenvolvimento (na produção de algoritmos e das bases de dados usadas para os treinar) até aos contextos de aplicação e potenciais utilizações. Estes impactos são muitas vezes negligenciados pelas narrativas dominantes (de la Bellacasa, 2011), que tendem a apresentar as tecnologias de IA como imateriais e intangíveis. Referimo-nos, entre outros, aos seguintes impactos: Resultados híbridos e controversos das tecnologias de IA; danos e inversões nas promessas e benefícios esperados das utilizações da IA na prática clínica; implementação prematura de algoritmos cuja reprodutibilidade e generalização raramente estão asseguradas; construção de modelos que poderão originar previsões enviesadas; eficácia, segurança, interpretabilidade e opacidade dos modelos; reconfigurações na confiança institucional e pessoal; e a redistribuição de responsabilidades nos processos de tomada de decisão.

Cabe então perguntar: Como é que a ética de cuidado dá visibilidade a desafios societais mais problemáticos? Fá-lo através de duas formas: Situa os impactos da IA em realidades práticas; e contextualiza os impactos da IA nas relações de poder institucionais, organizacionais, grupais e interpessoais, salientando o seu caráter profundamente desigual e assimétrico (Crawford, 2024 [2021]; Villegas-Galaviz e Martin, 2023). Esta abordagem às implicações éticas e sociais da IA é particularmente relevante quando consideramos os enormes interesses económicos e comerciais envolvidos e a forma como estes têm moldado a agenda de diversas iniciativas no

âmbito da regulação e legislação do desenvolvimento e implementação de tecnologias de IA.

Afastando-se da produção de orientações e regulamentos cuja finalidade radica na determinação de normas prescritivas, institucionalizadas e estandardizadas, a ética de cuidado promove uma revisitação de questões existenciais e humanistas fundamentais (Lagerkvist et al., 2022). O objetivo é providenciar instrumentos e recursos que garantam que todos os atores sociais, nas diversas posições situadas que ocupam, possam fazer escolhas que promovam uma IA de "confiança" e "responsável", cujo desenvolvimento efetivo assente numa distribuição de benefícios sustentável, solidária e orientada para o bem-estar de toda a sociedade.

3. Uma abordagem sociotécnica da Inteligência Artificial

Introdução

A partir de uma abordagem da Sociologia, neste capítulo perspetivamos a Inteligência Artificial (IA) enquanto fenómeno sociotécnico, que resulta de imbricações complexas entre tecnologia e sociedade e da interação de processos históricos, sociais, culturais, políticos, económicos e técnicos. Assumindo o pressuposto central da construção social da ciência e tecnologia, elencamos questões e dimensões de análise que se configuram úteis para o debate em torno dos desafios sociais e éticos da IA no século XXI. Procedemos à sua contextualização no âmbito de transformações sociais e políticas ocorridas nas últimas duas décadas, que possibilitaram um investimento maior (privado e público) em tecnologias de IA. Referimo-nos, designadamente, à aceleração e expansão planetária do capitalismo alicerçado em extração e circulação massiva de dados digitais, assim como à vertente mitológica da inevitabilidade da IA e à retórica visionária subjacente aos discursos tecno-otimistas veiculados por empresas tecnológicas, alguns cientistas na "vanguarda" e atores governamentais e políticos.

Ao realçar o papel das tecnologias de IA como agentes da ontologia do social, ou seja, como entidades sociais que integram interações sociais, numa relação simétrica e dinâmica entre humanos e máquinas, expomos os desafios teórico-metodológicos que esta abordagem relacional da IA convoca. Mencionamos, em particular, a inclusão de todos os atores (humanos e não humanos; visíveis, silenciados e invisibilizados) que estão implicados no mundo social da IA em contextos diversos, e o envolvimento crítico com os futuros da IA imaginados pelos seus "empreendedores", desde empresários tecnológicos a políticos e organizações internacionais.

Mapeamos ainda os principais contributos de uma abordagem sociotécnica para a desconstrução da caixa negra da IA e dos princípios da universalidade, neutralidade e racionalização que imperam nas narrativas dominantes sobre a IA, mostrando a existência de flexibilidade interpretativa na apreciação das implicações sociais e éticas da IA à luz de expectativas diversificadas e localmente interpretadas por atores sociais com posicionamentos específicos. Por fim, elencamos as principais tendências da abordagem distintiva da Sociologia em relação à IA e apresentamos as características mais relevantes de algumas metodologias adequadas ao estudo da IA enquanto fenómeno sociotécnico.

Em síntese, neste capítulo pretendemos:

- Explicar o que significa abordar a IA enquanto fenómeno sociotécnico.
- Identificar as questões de investigação e as dimensões de análise convocadas por uma abordagem sociotécnica da IA.
- Mapear os contributos da Sociologia para compreender os contextos, os discursos e as interações envolvidos na expansão da IA no século XXI.
- Dar a conhecer algumas metodologias e técnicas de investigação usadas na abordagem sociotécnica da IA.

3.1. A Inteligência Artificial como um fenómeno sociotécnico

Adotando a perspetiva da Sociologia, podemos falar da IA como sendo um "fenómeno sociotécnico" (Søraa, 2023, pp. 12-13). Isto implica reconhecer que não se trata de um fenómeno que está apenas circunscrito a aspetos técnicos e científicos, mas é sim o resultado de uma interação complexa entre ciência, tecnologia e as interações sociais, estruturas de poder e desigualdades presentes nas sociedades.

Imaginemos a seguinte situação: Um determinado sistema de IA é desenvolvido para apoiar os serviços estatais de segurança social na tomada de decisão sobre quais serão as famílias que devem beneficiar de subsídios escolares. Muito provavelmente, os programadores desse sistema de IA vão definir, no desenho do código e algoritmo, quais os pressupostos subjacentes para ser alcançada determinada decisão. Ou seja, quais são os critérios para atribuir ou recusar esse subsídio estatal perante determinadas características que definirão o "perfil" da família (entre outras, a composição familiar, os rendimentos, o número de filhos, e o tipo de habitação). Um exemplo concreto de como a IA pode ser abordada como um fenómeno sociotécnico consiste em questionar quais são os valores e as normas sociais subjacentes a esses pressupostos de categorização social das famílias usados na programação do sistema de IA. Em suma, nas palavras de Joyce e colegas:

> Profundamente interligados com a sociedade, esses sistemas [de IA] são aquilo a que os estudiosos de ciência e tecnologia chamam *sociotécnico*, um termo que chama a atenção para a forma como os valores, as práticas institucionais e as desigualdades estão incorporados no código, na conceção e na utilização da IA. (Joyce et al., 2021, p. 1)

Outro meio para compreendermos melhor a ideia da IA como um fenómeno sociotécnico consiste em analisar como e porquê se assistiu, nos últimos anos, a um crescimento expressivo deste tipo de tecnologias. Conforme relatamos no primeiro capítulo, os primórdios da IA remontam aos anos de 1950. Ou seja, durante décadas os avanços do campo da IA foram modestos. No entanto, ao longo dos últimos anos foram-se reunindo condições sociais, económicas e políticas que possibilitaram um investimento cada vez maior em tecnologias de IA. Por exemplo, a crescente produção, armazenamento e circulação de dados digitais (processos esses em boa medida reforçados durante os anos da pandemia COVID-19) fez com que o investimento privado e público em computação avançada crescesse exponencialmente, facilitando, com isso, o desenvolvimento e a utilização da IA.

Foi a partir do momento em que a IA começou a ser objeto de interesse da parte de grandes empresas, que este campo deixou de ser apenas do interesse de alguns cientistas para se alargar expressivamente e se tornar um fenómeno de amplas repercussões sociais, culturais e políticas (Liu, 2021). Esta constatação evidencia que o desenvolvimento de determinada tecnologia não depende apenas da vontade e motivação da comunidade científica que a desenvolve, mas também da conjugação

de circunstâncias favoráveis ao investimento e interesse nessa mesma tecnologia. A este respeito, um fator geralmente muito importante é o potencial de determinada tecnologia poder ser comercializada em grande escala e gerar lucros.

No entanto, não são apenas as circunstâncias sociais e económicas que têm implicações no desenvolvimento da IA. O inverso também é válido: Ou seja, a IA tem efeitos sobre a sociedade e a economia, podendo afetar relações de poder e as condições de existência e bem-estar das populações. Pensemos, por exemplo, na automatização de tarefas e na robótica introduzidas em processos de manufatura e produção industrial: A IA tanto pode desencadear desemprego (pela substituição de pessoas por máquinas), como pode libertar os trabalhadores de tarefas repetitivas para poderem exercer funções mais criativas e complexas, ou mesmo contribuir para criar novos empregos.

Outro aspeto crítico tem que ver com o amplo potencial da IA para provocar profundas transformações sociais e culturais, podendo gerar impactos complexos e ambíguos: Por exemplo, a IA pode reforçar desigualdades sociais pré-existentes e criar novas formas de desigualdade, opressão e discriminação; mas também pode oferecer o potencial de contribuir para uma maior igualdade e justiça social (Joyce et al., 2021; Zajko, 2022). Nas secções seguintes iremos debater em profundidade os contributos da Sociologia para o debate e investigação científica das relações complexas entre IA e sociedade.

3.2. Questões de investigação e dimensões de análise

Adotar uma abordagem sociotécnica convoca a exploração de várias dimensões de análise para responder a questões distintas, mas entrelaçadas. Elaboramos a tabela 5, onde sistematizamos algumas das principais questões que cada dimensão de análise procura responder no âmbito de uma abordagem sociotécnica da IA.

Tabela 5
Uma abordagem sociotécnica da Inteligência Artificial: Questões de investigação e dimensões de análise.

Dimensões de análise	Questões principais
Normas e valores	Como é que as normas, valores e crenças de uma sociedade influenciam o desenvolvimento da IA e surgem incorporados em dados, algoritmos e sistemas de IA?
Relações de poder	Como é que podemos escrutinar ou desafiar relações de poder existentes e mapear o modo como a IA incorpora, reproduz e consolida desigualdades estruturais e sistémicas?
Implicações éticas	Quais as implicações éticas da implementação da IA em diferentes contextos de utilização, incluindo questões de privacidade, discriminação e justiça?
Dinâmicas organizacionais	Como é que a IA afeta a cultura e estrutura organizacionais?
Aceitação e/ou resistência social	Como é que as atitudes públicas em relação à IA influenciam a aceitação e/ou resistência à implementação da IA em diferentes contextos?
Fronteiras humanos/máquinas	Como é que as fronteiras entre humanos e máquinas se (re)configuram em determinados contextos?
Antecipar e mitigar riscos sociais	Que riscos sociais poderão emergir no desenho e implementação da IA e que estratégias podem ser desenvolvidas para mitigar esses riscos?

Para responder a estas questões complexas, uma abordagem sociotécnica convoca três níveis principais de análise, que estão interligados (Lindgren e Holmström, 2020). Em primeiro lugar, os contextos históricos, sociais, culturais, económicos e políticos mais amplos que enquadram o desenvolvimento, utilização e perspetivas sociais sobre a IA, sobretudo nos últimos anos, destacando-se reflexões em torno da expansão do chamado capitalismo de dados digitais. Em segundo lugar, os discursos sobre os modernos sistemas de IA que enformam a construção social de mitos, retóricas, expectativas e controvérsias associados a este fenómeno. Por fim, em terceiro lugar, as interações e mediações suscitadas pela IA e as suas implicações na formação de identidades, conhecimentos e relações sociais. Nas próximas secções, sumariamos os contributos de uma abordagem sociotécnica para compreender os contextos, os discursos e as interações envolvidos na expansão da IA no século XXI.

3.3. Os contextos

Pensando no contexto histórico, social, cultural, económico e político que enquadra o desenvolvimento atual de tecnologias de IA, um primeiro conceito que se configura útil é o de capitalismo de dados, definido por Sarah West da seguinte forma: "O capitalismo de dados é, na sua essência, um sistema em que a mercantilização dos nossos dados permite uma redistribuição do poder na era da informação" (West, 2019, p. 23). A autora argumenta que o sistema capitalista produz e reforça relações de poder que favorecem os atores sociais e as organizações que têm acesso e capacidade para dar sentido aos dados digitais.

Na sua abordagem do capitalismo de dados, Sarah West afirma a proximidade conceptual com a ideia de capitalismo de vigilância, proposta por Shoshanna Zuboff (2015). O capitalismo de vigilância postula a emergência de uma nova forma de capitalismo assente na acumulação de vestígios digitais, e em que os lucros derivam da vigilância unilateral e da modificação do comportamento humano baseadas na mediação informática generalizada, produzindo as suas próprias relações sociais e, com isso, as suas conceções específicas de poder (Zuboff, 2015, p. 77). No entanto, West entende que o capitalismo de dados não se esgota apenas em questões de vigilância unilateral; trata-se, sobretudo, da forma como o mercado confere aos dados novos tipos de poder informativo e capitaliza esse poder, tornando-o invisível em nome da transparência e da eficácia (West, 2019, p. 22).

Ainda que nem West nem Zuboff tenham tratado especificamente do fenómeno da IA, ambas as autoras abordaram temas conexos, como as novas formas de poder e de autoridade suscitadas pela massificação dos processos de extração, armazenamento, transformação e circulação de dados digitais possibilitados por técnicas de *Big Data* ("grandes dados"[27]). Estas mudanças sociais têm produzido aquilo que, de uma forma geral, se pode designar como dataficação, ou seja, a transformação da ação social em dados digitais quantitativos que permitam o acompanhamento em tempo real

27 Consultar o glossário para mais informações.

e a análise preditiva do comportamento humano, sendo essa informação percebida como uma nova forma de valor[28] (van Dijck, 2014).

As tecnologias de IA baseiam-se em quantidades massivas de dados digitais. O tipo de conhecimento que produzem baseia-se num treino extensivo, e computacionalmente intensivo, com grandes conjuntos de dados acompanhados de regras e recompensas predefinidas (ver capítulo 1). Como tal, a IA depende de estruturas históricas, sociais, culturais, económicas e políticas associadas ao capitalismo de dados e à dataficação.

O livro de Kate Crawford (2024 [2021]), sugestivamente intitulado *Atlas da IA – Poder, política e os custos planetários da Inteligência Artificial*, ilustra este tipo de abordagem da IA como fenómeno expressivo de formais atuais de capitalismo de dados e informacional. A autora explica como a IA depende inteiramente de um conjunto muito vasto de estruturas políticas e sociais, e devido ao capital necessário para construir a IA à escala e às formas de otimização desejáveis, os sistemas de IA são, em última análise, concebidos para servir os interesses dominantes existentes e reproduzindo relações sociais e compreensões do mundo. Neste sentido, Crawford afirma que a IA é um registo de poder (2024 [2021], p. 17) que exige conectar as questões de poder e justiça: Da epistemologia aos direitos laborais, da extração de recursos à proteção de dados, da desigualdade racial às alterações climáticas (2024 [2021], p. 29). Nas suas palavras:

> A inteligência artificial é uma ideia, uma infraestrutura, uma indústria, um modo de exercer poder e uma forma de ver; é igualmente a manifestação de um capital altamente organizado, apoiado por vastos sistemas de extração e logística, com cadeias de fornecimento que abrangem todo o planeta. Tudo isto faz parte do que é a inteligência artificial – uma expressão de duas palavras sobre a qual se cartografa um complexo conjunto de expectativas, ideologias, desejos e medos. A IA pode parecer uma força espectral – enquanto computação incorpórea – mas estes sistemas são tudo menos abstratos. São infraestruturas físicas que estão a remodelar a Terra, ao mesmo tempo que alteram a forma como o mundo é visto e compreendido. (Crawford, 2024 [2021], p. 27)

Realçamos, ainda, o livro intitulado *A razão algorítmica: O novo governo do eu e do outro*, da autoria de Cláudia Aradau e Tobias Blanke (2022). Esta obra debruça-se especificamente sobre o papel dos algoritmos e da IA naquilo que os autores chamam de capitalismo digital. Os autores propõem-se abordar as condições sociais e políticas que tornaram os algoritmos – e tecnologias conexas como o *Big Data* e a IA – uma espécie de resposta para problemas globais, diversos e dispersos. Fizeram-no

28 O termo dataficação foi generalizado por Mayer-Schönberger e Cukier (2013). Estes autores falam de um novo paradigma pelo qual governos e empresas apostam cada vez mais na extração de grandes quantidades de dados de redes sociais e outras plataformas de comunicação digitais. As técnicas de *Big Data* possibilitam não só a extração de quantidades massivas de dados em tempo real como transformam esses dados, aparentemente desconexos, em índices numéricos que permitem produzir informações incontáveis sobre o comportamento humano e sobre interações sociais.

transformando relações de poder, ao mesmo tempo que criaram "as condições de possibilidade de implementação de algoritmos para governar a conduta de indivíduos e populações, de amigos e inimigos, de normalidade e anormalidade em distintos mundos sociais e fronteiras políticas" (Aradau e Blanke, 2022, p. 3).

3.4. Os discursos

3.4.1. Mitos, metáforas e expectativas

Vários autores que adotam uma perspetiva sociotécnica têm vindo a focar o papel fundamental dos mitos, das narrativas e da retórica[29] na projeção de discursos[30] sobre a IA no espaço público, influenciando fortemente o modo como a maioria das pessoas pensam e falam a respeito da IA. Por exemplo, metáforas como "inteligência" artificial ou "aprendizagem" das máquinas intervêm de forma duradoura no discurso social, alimentando mitos e expectativas futuras tanto junto do público em geral como no seio de comunidades de especialistas (Bareis e Katzenbach, 2022; Campolo e Crawford, 2020; Natale e Ballatore, 2017). Outra fonte discursiva que marca o discurso popular sobre IA são histórias sobre máquinas com semelhanças humanas, muito presentes não só na ficção científica contemporânea como em narrativas míticas que perduram há séculos[31] (Mayor, 2018; Sheikh et al., 2023).

Os discursos sobre IA surgem associados a expectativas altamente otimistas. Este é, segundo Kornelia Konrad (2006), um processo comum quando estamos a falar de

29 De modo muito sintético e necessariamente simplista, é importante esclarecer que enquanto o discurso se refere à expressão de ideias em qualquer forma de comunicação, a narrativa é uma forma específica de discurso que conta uma história. Por sua vez, a retórica refere-se a modos de persuadir, enquanto os mitos são narrativas tradicionais que explicam crenças, práticas ou fenómenos culturais fundamentais dentro de uma sociedade.

30 Adotamos uma noção muito ampla de discurso, nos termos propostos por Adele Clarke (2005, pp. 148-149). Referimo-nos a qualquer forma de comunicação, cultural e historicamente situada, que tanto pode incluir linguagem verbal como não verbal, elementos visuais, símbolos, coisas não humanas, objetos, etc. Estes podem influenciar perceções, criar objetos de conhecimento, e incluir formas de representação e de veiculação de significados que permitem, de acordo com alguns autores, realizar análises de poder (Foucault, 1972, 1973). Não cabe no escopo deste capítulo explorar a diversidade teórica a propósito de discursos, mas podemos apontar, a título de exemplo, a diferença fundamental entre as abordagens do interacionismo simbólico (Mead, 1934/1962; Strauss, 1978), que defendem que os indivíduos e as coletividades são produzidos por via da sua participação nos mundos sociais, incluindo por via dos seus discursos; e a abordagem pós-estruturalista proposta por Foucault (1973), quando este argumenta que os indivíduos e as coletividades são constituídos por via de discursos e disciplinas.

31 Várias culturas, e em diferentes regiões do mundo, têm histórias sobre personagens que podem ser caracterizadas como formas artificiais de inteligência. Dando como exemplo a mitologia grega, Dédalo, arquiteto e inventor do mundo antigo, terá criado Talos, um super-soldado mecânico (um robô) para proteger a ilha de Creta. Hephaistos, o ferreiro dos deuses, tinha ajudantes mecânicos na sua oficina. Para castigar a humanidade, Zeus criou a mulher mecânica Pandora, que derramava todo o tipo de sofrimento sobre os humanos quando abria o seu frasco (a "caixa de Pandora"). A história de cientistas que criam uma forma de vida artificial que acaba por se voltar contra o seu criador tornou-se um arquétipo dos riscos da tecnologia moderna representado na ficção científica. Este tema está presente em inúmeros filmes, incluindo clássicos como *Blade Runner* (1982), *O Exterminador do Futuro* (1984) e *Matrix* (1999).

novas tecnologias. Para a autora, as expectativas desempenham um papel fundamental no ritmo e no desenrolar de um ciclo de inovação:

> As novas tecnologias estão frequentemente sujeitas a expectativas muito elevadas. Normalmente, as expectativas podem ser amplamente aceites durante um período de tempo, tanto pelos apoiantes de uma nova tecnologia como por vozes críticas resignadas à inevitabilidade do desenvolvimento de uma determinada tecnologia. (Konrad, 2006, p. 429)

A autora alerta, porém, para potenciais mudanças nas dinâmicas das expectativas. Numa fase inicial de implementação de uma tecnologia, as expectativas partilhadas desempenham um papel central na criação da dinâmica necessária para os processos de inovação e na coordenação de atores sociais heterogéneos e diferentemente posicionados numa rede de inovação. Mas estas expectativas otimistas podem vir a ser altamente problemáticas quando os ciclos de entusiasmo se transformam em fases de desilusão. Neste caso, observa-se frequentemente um efeito prejudicial na credibilidade de atores específicos ou de um campo de inovação. Ou então, quando as expectativas permanecem amplamente aceites mesmo depois da fase inicial de entusiasmo com tecnologias novas, estas chegam a um ponto em que deixam de estar sujeitas a um exame crítico (Konrad, 2006, p. 430). Isto significa que a dinâmica das expectativas em torno de uma determinada tecnologia é eminentemente social e coletiva, dependendo da coordenação e articulação entre atores sociais dispersos e pautados por objetivos diferenciados.

A recente conjunção de expectativas sociais marcadamente otimistas em torno das tecnologias de IA, acompanhadas de alocações massivas de recursos tecnológicos e financeiros, a par com o agudizar de controvérsias e discursos sobre receios e danos, conduzem a perspetivar a IA do século XXI como um fenómeno paradigmaticamente novo. Fala-se numa revolução (Sejnowski, 2018), num *tsunami* (Manning, 2015), num trauma epistémico (Pasquinelli, 2015) ou ainda – numa abordagem mais crítica – em mitos tecnológicos (Bareis e Katzenbach, 2022; Roberge et al., 2020) ou crenças mágicas (Elish e Boyd, 2018).

A este respeito é interessante considerar o termo "sublime tecnológico", invocado por Leo Marx (2000) para descrever o modo como durante o século XIX, com as primeiras obras-primas da engenharia, como o caminho de ferro, o sublime, anteriormente dirigido aos fenómenos naturais e aos enigmas da física, é cada vez mais "dirigido para a tecnologia ou, melhor, para a conquista tecnológica da matéria" (Marx, 2000, p. 197). A evocação deste sublime tecnológico encarna a celebração do progresso tecnológico e esconde os seus problemas e contradições (Marx, 2000, p. 207), e ajuda a compreender como a agência pode ser afastada dos humanos e projetada para a IA (Bareis e Katzenbach, 2022, p. 860).

Os Estudos Sociais da Ciência e Tecnologia têm dedicado ampla atenção ao modo como os discursos sobre os avanços tecnológicos e a retórica visionária veiculados por empresas, alguns cientistas na "vanguarda" e atores governamentais e políticos projetam expectativas e histórias sobre o futuro (van Lente, 2016). As novas

descobertas tecnológicas e científicas – e a IA não é exceção – estão regularmente ligadas a narrativas modernistas de progresso e à ideia da tecnologia como meio de inovação de mercado e engenharia social. Estes discursos projetam futuros desejados, assim como anseios e aspirações – embora caiba perguntar *de quem*, e, nessa medida: Quem é favorecido, desfavorecido ou silenciado/invisibilizado por determinados discursos públicos e de que forma esses futuros podem ser contestados (Brown et al., 2017; Oomen et al., 2022)? Mais: Quem protagoniza a construção e disseminação de expectativas e de mitos sobre a IA? Quais as características principais da retórica em torno da IA? Que cenários futuros são projetados?

Por exemplo, ao analisar os discursos de diferentes governos nacionais em torno da IA, geralmente marcados por uma perspetiva de futuro, podemos compreender como funciona o poder do Estado em termos de perspetivação de como deve ser feita a seleção de prioridades de desenvolvimento, alocação de recursos e investimento em infraestruturas. Uma análise das políticas nacionais que projetam o futuro da IA permite igualmente analisar qual é o papel que o Estado atribui ao envolvimento de diferentes organizações, empresas, setores (como a educação) e cidadãos no desenvolvimento da IA. A este respeito, um estudo conduzido por Christopher Wilson sobre estratégias nacionais de 16 países em relação à IA, já referido no segundo capítulo, concluiu que os cidadãos são essencialmente enquadrados pelos atores governativos nas seguintes categorias: Destinatários dos benefícios abstratos da IA; utilizadores de serviços e produtos orientados para a IA; ou força de trabalho que necessita de formação e de melhoria de competências para lidar com a IA, de modo a contribuir para libertar o potencial da IA, tido como essencial para uma sociedade próspera (Wilson, 2022, pp. 7-8).

Os estudos existentes sobre o modo como os diferentes governos nacionais elaboram uma retórica em torno da IA mostram uma tendência para todas as políticas nacionais projetarem discursos que enquadram a IA como um desenvolvimento tecnológico adquirido e massivamente disruptivo que irá mudar fundamentalmente a sociedade. Em consequência, a necessidade de adotar a IA em todos os sectores-chave da sociedade é retratada retoricamente como inevitável, independentemente do país em causa (Bareis e Katzenbach, 2022). Ao mesmo tempo, as implicações sociais e éticas do desenvolvimento da IA surgem secundarizadas em relação aos esperados efeitos benéficos na economia e inovação.

Neste contexto, a ideia de inevitabilidade do desenvolvimento da IA surge como um mito cultural. A este propósito, é importante reter o que diz Vincent Mosco na sua abordagem dos mitos como dispositivos de estruturação para a ordenação sociotécnica, em particular a relevância da análise do poder dos mitos, o qual não decorre do seu grau de veracidade:

> Os mitos não são verdadeiros nem falsos (...). Compreender um mito é mais do que provar que ele é falso. Significa perceber porque é que o mito existe, porque é tão importante para as pessoas, o que significa e o que nos diz sobre as esperanças e os sonhos das pessoas. (Mosco, 2005, p. 29)

Uma análise de mitos bem-sucedidos, como é o caso do mito da inevitabilidade da IA, permite elucidar sobre a hierarquia de valores sociais e as estruturas de poder subjacentes à predominância de determinados valores sociais em detrimento de outros (por exemplo, dar prioridade à competitividade económica e eficiência em detrimento da igualdade e acesso a um trabalho digno). Mitos bem-sucedidos implicam também um processo de despolitização ao reduzirem maciçamente a complexidade e dissociando os desenvolvimentos tecnológicos dos seus contextos sociais e políticos. Esta função de despolitização será, porventura, o elemento retórico mais importante quando falamos de IA. Nas palavras de Bareis e Katzenbach:

> Este é o paradoxo dos imaginários da IA: Os mitos sobre IA soam fantásticos e desencadeiam as nossas fantasias, embora simultaneamente minem a imaginação política e a prática política ao criarem expectativas de uma solução tecnológica reconfortante para problemas estruturais da sociedade. Embora grande parte destes debates seja ainda bastante controversa, parece que estamos já a assistir a um processo de encerramento de um conjunto de questões fundamentais. (2022, p. 876)

3.4.2. Caixas negras e flexibilidade interpretativa

Tal como acontece com a retórica política em torno da IA, os discursos de cientistas e empresas que desenvolvem e promovem IA parecem associar-se a reivindicações de conhecimentos em formato de resultados categóricos irrefutáveis, ainda que os seus significados sejam maioritariamente descontextualizados. Quando factos ou artefactos produzidos pela ciência e tecnologia são bem-sucedidos, tendem a ser tomados como garantidos e desta forma adquirem um carácter de inevitabilidade, como se fossem a melhor ou a única solução possível para um determinado conjunto de problemas (Sismondo, 2012, p. 133).

Por outras palavras, quando a IA é apresentada como produzindo resultados categóricos transforma-se necessariamente em algo prescritivo, o que significa que "cada norma e cada categoria valoriza um ponto de vista e silencia outro" (Bowker e Star, 1999, p. 5). Neste contexto de produção de conhecimento irrefutável é adequada a utilização do termo caixa negra, que descreve um dispositivo previsível de entrada e saída, cujo funcionamento interno não precisa de ser conhecido para ser utilizado. Atendendo a que os discursos dominantes em torno da IA a apresentam como inevitável, poderemos falar de uma caixa negra? Se sim: Antes da IA se ter tornado uma caixa negra, com factos e artefactos bem-sucedidos, que controvérsias e disputas existiam? O que é que acontece a essa caixa negra quando surgem controvérsias em torno da IA? Quem é que protagoniza essas controvérsias e que tipo de disputas surgem? Como é que as controvérsias são desencadeadas (abertas) e encerradas (fechadas)?

Os ciclos de inovação e a receção pública de novas tecnologias apontam para a existência de flexibilidade interpretativa (Sismondo, 2012, p. 120), para se referirem ao processo pelo qual a apreciação das implicações de uma determinada tecnologia é feita à luz de expectativas localmente interpretadas por atores sociais com

posicionamentos específicos. Ou seja, apesar de existirem expectativas partilhadas – a convicção que a IA continuar-se-á a desenvolver e que existem benefícios, mas também riscos – a interpretação de diferentes grupos sociais será distinta. Por exemplo, enquanto os líderes tecnológicos irão propor soluções técnicas para a resolução dos problemas potencialmente suscitados pela IA (por exemplo, mais treino de máquinas, o que implicará alargar a recolha de dados digitais e a produção de mais protocolos de classificação de conteúdos), os investigadores da Sociologia poderão propor mecanismos regulatórios da utilização e desenvolvimento da IA e de responsabilização por danos e proteção e reparação das vítimas. Conforme aponta James Steinhoff (2023), todos estes atores, ainda que heterogéneos e movidos por interesses e agendas muito distintas, fazem parte de uma mesma rede de inovação, na qual os atores que reivindicam a reflexão em torno das implicações sociais e éticas da IA tendem a ocupar uma posição subordinada.

Na nossa perspetiva, a rede de inovação que sustenta a IA está a evoluir dinamicamente para um maior hibridismo e fusão de fronteiras, sendo importante uma compreensão mais matizada e holística do modo como esta rede funciona em contextos diversos (Machado et al., 2023). Por exemplo, é ainda muito escasso o conhecimento sobre o papel do erro e das falhas ao nível da investigação e desenvolvimento da IA. Do mesmo modo, importa envolver os diversos públicos na criação de precauções éticas e regulatórias em matéria de IA (Sieber et al., 2024). A este propósito, as abordagens centradas nas questões humanistas estão a ganhar destaque (Lagerkvist et al., 2002), sublinhando-se a necessidade de incorporar direitos fundamentais nos sistemas tecnológicos (ver capítulo 2) e de abordar a crescente hibridez entre entidades humanas e não humanas (ver secção seguinte).

3.5. As interações

Em 1985, o sociólogo Steve Woolgar publicou um artigo com o seguinte título sugestivo e provocatório: *Por que não uma sociologia das máquinas? O caso da sociologia e da inteligência artificial*. Defendendo a necessidade de ir além da análise dos impactos da IA e atender à sua génese e construção social, Woolgar critica o papel restrito atribuído à Sociologia na discussão da investigação em IA. A tendencial circunscrição à avaliação do impacto das tecnologias de IA na sociedade deu azo, na perspetiva de Woolgar, a abordagens focadas em três tópicos principais: 1) Caracterização dos contextos em que a IA é aplicada e como entrou na esfera do quotidiano; 2) as atitudes sociais em relação à IA; e 3) as potenciais implicações da introdução da IA em diferentes organizações, como o sistema educativo.

O autor reconhece as virtudes destas abordagens centradas nos impactos, mas sugere uma expansão da perspetiva sociológica em direção ao estudo da génese da IA e dos modos de produção de conhecimento científico e de agendas de investigação que lhe estão associadas, entre outros aspetos. Woolgar propõe, concretamente, que se aborde o tipo de dicotomias que estão na base de discursos e práticas no campo particular de produção e desenvolvimento da IA (por exemplo, humano/máquina; cognitivo/social; racionalidade/inteligência), de modo a destacar os significados e

atribuições de sentidos que são mobilizados para legitimar certas ações e agendas de investigação. Desta forma, argumenta Woolgar, a Sociologia pode resguardar-se dos riscos de a sua abordagem ser cooptada e estabelecida por via do próprio discurso de quem desenvolve e comercializa IA – os chamados "peritos ou especialistas" da IA. Nas suas palavras:

> Os perigos em adotar acriticamente o discurso [dos especialistas em IA] são particularmente evidentes no caso da distinção entre homem e máquina que é usada com efeitos consideráveis pelos especialistas (...) para que estes possam reivindicar a sua perícia (humana) específica para falarem de sistemas periciais (máquinas). Com isto, definem a natureza e o carácter do objeto de estudo, estabelecem quais são os tópicos adequados para investigar e argumentam que são os únicos com competência para falarem sobre esses objetos. (Woolgar, 1985, pp. 565-566)

Steve Woolgar acaba por argumentar que o fenómeno da IA tem particular importância para a Sociologia por levar a questionar o que se entende por social e comportamento humano:

> A IA oferece a oportunidade de reavaliar os nossos pressupostos sobre o comportamento, a ação, as suas origens e agência e, mais importante ainda, as nossas tentativas de compreensão desses aspetos. Considero que é instrutivo reavaliar cuidadosamente o argumento que há algo de especial no comportamento humano. Ou, dito de forma mais cuidadosa (uma vez que não é minha intenção pronunciar-me sobre o "carácter real" do comportamento humano), é importante analisar como a Sociologia presume que o comportamento humano é diferente do desempenho de uma máquina. Como é que conceções prevalecentes relacionadas com a atividade da máquina e do comportamento social moldam a explicação sociológica? (Woolgar, 1985, p. 568)

Mais recentemente, Ceyda Yolgörmez (2021) reforçou o apelo a uma mudança metodológica no estudo da IA numa perspetiva em que se pense a IA como parte integrante de uma interação social, numa relação simétrica entre humanos e máquinas. Fazem parte do "dinamismo" dessa relação a indeterminação e a aleatoriedade, aspetos que estão subjacentes tanto a humanos como à IA. Será, por isso, impossível prever totalmente o espaço de interações potenciais e reais entre humanos e máquinas. Yolgörmez sugere a adoção do conceito de agência distribuída para pensarmos na nova sociabilidade suscitada pela crescente hibridez de fronteiras entre humano e máquina. A autora perspetiva a agência como uma noção coletiva, constituída por vários processos, em vez de ser uma capacidade de um indivíduo; e é neste sentido que os não-humanos em geral, e a IA em particular, se tornam relevantes para a dimensão sociológica (Yolgörmez, 2021, p. 158). Pergunta a autora:

> O que é que chamamos "social"? O que é uma sociedade e quais são as implicações das nossas conceções sobre o que são as sociedades? Como é que os estudos sobre o "social" organizam esta compreensão e constroem os limites da nossa imaginação sociológica? (Yolgörmez, 2021, p. 144)

Na esteira das questões enunciadas por Ceyda Yolgörmez, acrescentamos outras: Como desenvolver uma Sociologia da IA por via do questionamento de fronteiras convencionais entre máquina e humano? O que significa uma agência partilhada e quais as suas implicações na abordagem de novos modos de sociabilidade (re)criados pela IA? O que distingue uma Sociologia relacional de uma abordagem centrada em impactos sociais?

As questões relativas às tecnologias têm sido tradicionalmente deixadas para os campos da engenharia, e as ciências sociais foram pensadas como estando apenas equipadas para lidar com os fenómenos sociais que emergem em torno das tecnologias. No entanto, Yolgörmez propõe uma outra abordagem, tomando as relações com e das máquinas como pertinentes para as relações sociais, argumentando sobre a necessidade de uma Sociologia relacional. Nas suas palavras:

> A IA é um objeto de estudo instável, uma vez que não se enquadra nos limites tradicionais e puros do humano *versus* não-humano. Pelo contrário, a IA emerge de emaranhados de relações sócio-materiais, e o seu papel na emergência da agência permite-nos classificá-la como um ser que reside e se encontra no domínio social. No entanto, não pretendo enquadrar a IA como tal – não se trata de uma operação de definição rígida –, mas sim de que esta pode ser uma outra forma de pensar a IA e que, nesta forma de pensar, o social não é uma área exclusivamente humana. Em vez disso, o social tem a ver com um encontro, com relacionalidade, e pode contribuir para uma expansão do pensamento sociológico ao permitir-lhe olhar simetricamente para as entidades que entram em relações. (Yolgörmez, 2021, p. 157)

A autora considera ainda que a IA é um caso-limite em que a Sociologia pode experimentar um campo de investigação não tradicional e descobrir até que ponto as fronteiras da disciplina podem ser reformuladas. Neste sentido, este esforço é uma resposta à chamada crise das ciências sociais (Yolgörmez, 2021, p. 159).

Outro aspeto importante desta abordagem relacional reside na visão alternativa às propostas que perspetivam a IA como patamar de crítica à lógica do capitalismo. Uma abordagem relacional da IA reconhece a importância desta crítica, e refere a urgência de pensar criticamente o modo como as tecnologias de IA reproduzem as forças históricas do capitalismo, do colonialismo, do patriarcado e do racismo e disseminam e consolidam essas lógicas nas sociedades, influenciando assimetricamente os grupos sociais. Porém, Yolgörmez entende que estas análises excluem outras formas de ver e interpretar a IA:

> Nesta linha de estudos, a IA surge como um instrumento do tecnocapitalismo e não tem uma verdadeira agência por si só; a IA só pode promover a agenda dos sistemas em que está inserida. (...) Embora tudo isto seja verdade (...) há outras formas, talvez mais consequentes, de pensar a IA sociologicamente. (Yolgörmez, 2021, p. 145)

A autora defende que a exploração das relacionalidades pode conduzir a resultados imprevisíveis e, assim, "escapar a ser totalizadas sob a lógica do capitalismo tardio. Este foco na relacionalidade demonstrará novas formas de imaginar as diferenças entre humanos e máquinas, mantendo a sua relevância para o olhar sociológico" (Yolgörmez, 2021, p. 159).

3.6. O olhar distintivo da Sociologia

Na Figura 2, sumariamos a nossa perspetiva sobre o olhar distintivo que a abordagem sociológica proporciona para compreender os contextos, os discursos e as interações envolvidos na expansão da IA no século XXI.

Figura 2
Uma abordagem sociológica da Inteligência Artificial: Níveis de análise.

O papel da Sociologia no estudo do desenvolvimento e utilização da IA configura-se cada vez mais urgente e necessário, à medida que o avolumar do investimento privado e público em sistemas sociotécnicos de IA traz a constatação que muitos mais sistemas de IA serão desenvolvidos e integrados nas organizações e sociedades nos próximos anos. O investimento em IA tem a pretensão de trazer as tecnologias de IA para a educação, os cuidados de saúde, a justiça penal, os serviços sociais, o planeamento urbano e outras áreas cruciais da vida social. Entre outros aspetos, uma abordagem sociológica implica identificar os múltiplos significados da IA enquanto fenómeno sociotécnico e propor abordagens teóricas e metodológicas para compreender as implicações vastas deste fenómeno. Zheng Liu (2021) define a especificidade da abordagem sociológica da IA como aquela que 1) encara o desenvolvimento da IA como um fenómeno social; 2) toma este fenómeno como seu objeto de estudo; 3) analisa as interações com as condições sociais, culturais, económicas e políticas mais amplas em que o desenvolvimento e utilização da IA ocorrem e pelas quais são afetados.

Liu (2021) propõe uma tipologia de abordagens da IA, distinguindo a abordagem da Sociologia das chamadas abordagens científicas da IA (realizadas a nível das "invenções" académicas em torno da IA) e das abordagens técnicas (correspondentes à comercialização de tecnologias de IA). Para a autora, na abordagem da Sociologia

é possível distinguir duas perspetivas teóricas principais: Em primeiro lugar, uma abordagem da IA de inspiração marxista, que foca as relações de exploração laboral dos humanos presentes no desenvolvimento destas tecnologias no contexto do sistema capitalista. Em segundo lugar, uma abordagem de pendor culturalista, que foca as distintas interpretações sobre os sentidos e significações a atribuir à IA que são construídas e mobilizadas por diferentes grupos sociais, socorrendo-se de diferentes recursos culturais e tradições para desenvolver narrativas de IA que ajudem a promover as suas agendas específicas (Liu, 2021, pp. 8-9).

Independentemente da abordagem teórica específica da Sociologia, consideramos crucial conferir atenção aos processos sociais de criação de significados ao mesmo tempo que se revela que as interações entre atores humanos e não humanos são sempre socialmente situadas, múltiplas e relacionais. No que diz respeito à IA, destacam-se os contributos destas perspetivas para o estudo das desigualdades por via de algoritmos, no código e dados, e a atenção conferida à forma como os contextos moldam a conceção e a utilização da IA. Conforme foi explicitado nas secções anteriores, várias abordagens dos sistemas sociotécnicos da IA permitem esclarecer como é que as desigualdades operam a nível individual, organizacional e em sistemas globais. Uma abordagem sociológica intersecional (Joyce et al., 2021, p. 6) contribui para a crescente discussão sobre o impacto desigual da IA, tendo em conta as histórias globais do capitalismo, e as formas como as variações históricas e contemporâneas das desigualdades criadas pelos sistemas capitalistas são reproduzidas e exacerbadas nos e pelos sistemas digitais.

Por fim, algumas abordagens sociológicas da IA defendem a necessidade de uma postura que contribua para a transformação social e para ajudar a afirmar a agência sobre estas tecnologias através de três tipos de ações: Crítica e política de recusa; combate às desigualdades através da própria tecnologia; e consolidação de contributos para colaborar na governação dos algoritmos (Zajko, 2022).

Em relação ao primeiro tipo de ação – crítica e política de recusa–, o autor remete para estudos que reivindicam que certas decisões não devem ser automatizadas por tecnologias de IA (por exemplo, Benjamin, 2019). Zajko defende a ideia que os estudos da crítica e da recusa das tecnologias de IA contêm "um argumento positivo sobre o que certas decisões devem implicar, em termos de envolvimento humano. Por isso, argumentar que não devemos construir ou implementar um sistema é valioso e muitas vezes apropriado como forma de alcançar futuros desejáveis" (Zajko, 2022, pp. 6-7).

No que diz respeito ao segundo tipo de ação – o combate às desigualdades através do envolvimento de sociólogos na produção da própria tecnologia de IA –, Zajko alerta para o perigo de cooptação, mas não deixa de reconhecer as virtualidades da abordagem pela qual a Sociologia pretende transformar as tecnologias de IA em tecnologias mais justas e igualitárias ao envolver-se na produção técnica das tecnologias (por exemplo, tentando eliminar viés e discriminação nos dados usados para o treino dos algoritmos), por via de colaborações interdisciplinares.

O terceiro tipo de ação – contribuir para a governação dos algoritmos – diz respeito à forma como a investigação sociológica pode estar envolvida na construção normativa do mundo, participando na governação da tecnologia e abordando os perigos muito reais da governação algorítmica (Zajko, 2022, pp. 10-11). Nos últimos anos, tem-se assistido a uma quantidade considerável de estudos que documentam, comparam e criticam diferentes políticas ou regimes de regulação de algoritmos e IA, incluindo estratégias governamentais, declarações de princípios corporativos, normas e regulamentos. O autor chama a atenção para o facto de que os modos como a Sociologia pode contribuir para a política pública sobre governação de IA depende das oportunidades políticas disponíveis, tais como a abertura dos processos formais à contribuição académica, ou até que ponto as consultas públicas são realmente utilizadas para informar a política ou ao invés desempenharem um papel legitimador ou performativo (ver também Machado et al., 2023).

3.7. Metodologias de investigação

3.7.1. Metodologias quantitativas e qualitativas

A abordagem sociotécnica abre a porta à mobilização de diversas metodologias, ou seja, é possível utilizar múltiplos métodos e técnicas para desenhar um estudo e para recolher, analisar e interpretar dados. Quando pretendemos abordar um tópico relacionado com os desafios sociais e éticos da IA no século XXI, importa escolher as metodologias que são mais apropriadas para responder às questões que nos preocupam e ao tipo de dados empíricos que precisamos de obter.

As categorizações convencionais distinguem as metodologias quantitativas (por exemplo, questionários) das metodologias qualitativas (como as entrevistas, análise documental e observação/etnografia). Em termos gerais, as metodologias quantitativas visam medir atitudes ou práticas relacionadas com a IA com base em abordagens estruturadas assentes na definição de hipóteses. Estas permitem generalizar os resultados para a população, investigar associações e estabelecer relações de causa-efeito. Porém, as metodologias quantitativas revelam limitações quanto à compreensão das perspetivas dos participantes e dos contextos históricos, sociais, culturais, económicos e políticos mais amplos onde estes se inserem, e nem todas as dimensões da IA são passíveis de quantificação. Identificamos, de seguida, alguns estudos empíricos concretos que exploraram as implicações sociais e éticas do desenvolvimento, implementação e utilização de tecnologias de IA com base em metodologias quantitativas.

Awad e colegas (2018), por exemplo, usaram uma plataforma experimental *online* com o objetivo de quantificar as expectativas públicas em torno dos princípios éticos que deveriam guiar o comportamento de máquinas, e como estas expectativas variavam entre indivíduos e países, reunindo respostas de 2,3 milhões de utilizadores da Internet em quase todo o mundo. Já Ploug e colegas (2021) aplicaram um inquérito baseado em escolhas a uma amostra representativa da população adulta na

Dinamarca (num total de 1027 participantes) com o objetivo de captar as preferências públicas quanto ao desempenho e explicabilidade de decisões tomadas através da IA nos cuidados de saúde, e determinar em que medida essas preferências dependiam das características dos participantes, incluindo a confiança depositada na tecnologia e no sistema de saúde, assim como os receios e benefícios esperados da IA. O uso de vinhetas num estudo experimental *online*, com uma amostra representativa de 1048 participantes, permitiu a Willems e colegas (2023) testar de que forma os cidadãos austríacos alcançavam um compromisso entre a utilidade percebida de aplicativos baseados em IA e as preocupações com a privacidade, em situações que envolviam a utilização de serviços públicos. Referimos, por fim, o relatório do *Ada Lovelace Institute* e do *The Alan Turing Institute* (2023), que deu a conhecer os resultados de um inquérito nacional representativo com cerca de 4000 adultos na Grã-Bretanha. Este estudo explorou três dimensões principais das experiências e atitudes públicas em relação a 17 aplicações da IA na vida pública e pessoal, desde o comum reconhecimento facial para desbloquear telemóveis a aplicações menos visíveis e futuras, como a elegibilidade para benefícios estatais e a existência de carros sem condutor, designadamente: 1) O nível de conhecimento sobre cada uma das aplicações da IA; 2) as perceções sobre os benefícios, preocupações e riscos associados a cada tecnologia; e 3) as preferências quanto à regulação e governação da IA.

As metodologias qualitativas, por outro lado, pretendem descrever e compreender as situações ou práticas sociais e os significados atribuídos à IA com base em abordagens flexíveis, adaptáveis a eventuais imprevisibilidades que possam acontecer ao longo do trabalho de investigação. Estas permitem captar as vozes e as visões detalhadas dos participantes e compreender em profundidade os contextos que enquadram o desenvolvimento, utilização e perspetivas sociais sobre a IA, assim como os discursos sobre os modernos sistemas de IA e as interações suscitadas pela expansão da IA no século XXI.

É frequente o recurso a entrevistas individuais, semiestruturadas ou em profundidade, e mais raramente a grupos focais. McCradden e colegas, por exemplo, conduziram seis grupos focais com 41 representantes do público em geral no Canadá para compreender as suas visões quanto ao uso de dados de saúde na investigação em IA (McCradden et al., 2020a), e 30 entrevistas individuais com doentes, cuidadores e profissionais de saúde para explorar as suas visões sobre os desafios éticos implicados na utilização de IA nos cuidados de saúde (McCradden et al., 2020b). Já Bastian e colegas (2021) entrevistaram 17 profissionais que trabalhavam em dois jornais renomados na Holanda e na Suíça para analisar as perceções dos profissionais dos média quanto ao impacto de sistemas algorítmicos de recomendação de notícias nas suas normas profissionais e nas missões dos órgãos de comunicação social, e como estas normas e missões poderiam ser integradas no desenho desses mesmos sistemas algorítmicos de recomendação de notícias. Destacamos, ainda, o trabalho de Aquino e colegas (2023), que conduziram 72 entrevistas semiestruturadas com especialistas em IA e/ou em clínica para explorar as estratégias acionadas para tentar mitigar o viés algorítmico e perspetivar a questão ética da responsabilização pela existência de vieses algorítmicos.

Existem, ainda, os estudos qualitativos de natureza etnográfica. A etnografia permite observar e apreender discursos e interações em contextos reais, incluindo os espaços digitais, possibilitando a obtenção de informação rica, aprofundada e detalhada normalmente associada ao estudo de pequenos grupos ou comunidades. A flexibilidade que proporciona é desafiante, e torna possível o cruzamento de diversas técnicas de recolha de dados. Henriksen e Blond (2023), por exemplo, articularam a realização de entrevistas semiestruturadas, a análise documental, e a observação participante numa etnografia conduzida numa empresa de IA situada na Escandinávia (região que abrange a Dinamarca, a Suécia e a Noruega). Isso permitiu-lhes perceber como é que a IA foi performada e executada à medida que os desenvolvedores acionavam dois sistemas preditivos em conjunto com as partes interessadas nos setores públicos das finanças e da saúde, respondendo a três questões principais: 1) Que indivíduos eram priorizados e destacados? 2) Que tipo de competências e de agência eram promovidas? 3) Quem beneficiou, de facto, da implementação dos sistemas de IA? Já Lee e colegas (2022) realizaram uma investigação etnográfica e participativa para compreender como é que jovens (14-24 anos) sub-representados nas áreas de ensino associadas à ciência, tecnologia, engenharia e matemática (conhecidas por *STEM*) atribuíam sentido à IA nas suas vidas e na sociedade, e como é que as suas relações com a tecnologia evoluíram quando criaram as suas próprias ferramentas de IA.

Outros estudos qualitativos recorrem à análise documental de relatórios institucionais, *websites*, blogues, artigos, e notícias publicadas em jornais, entre outros documentos. Rogers e colegas (2021), por exemplo, usaram diversos materiais públicos (oito artigos científicos, dois documentos regulatórios, três *websites* e dois comunicados de imprensa) que reportavam as fases de desenvolvimento, produção de evidência e implementação de duas aplicações de IA na saúde com o objetivo de analisar, em detalhe, as implicações éticas envolvidas no uso de sistemas de IA para auxiliar processos de tomada de decisão clínica. Já Alfrink e colegas (2022) combinaram a análise de 10 documentos relacionados com um projeto de implementação de sistemas de IA na gestão urbana com a realização de nove entrevistas a utilizadores dos mesmos, no contexto de uma investigação-ação participativa que pretendia explorar visões e experiências em torno do que significa um sistema de IA transparente. Neste estudo específico, os autores analisaram como é que os especialistas que desenham, desenvolvem e governam sistemas de IA urbanos compreendem a transparência, e como é que os seus utilizadores experienciam um sistema de IA transparente.

Também as novas práticas artísticas e imagens que têm emergido à luz da cultura visual potenciada pelas tecnologias de IA constituem um objeto de estudo privilegiado. Na esteira da abordagem semiótica de Barad (2003), pensa-se o social e o científico/tecnológico em conjunto na construção de futuros alternativos, procurando compreender como as práticas tecnocientíficas material-discursivas delimitam as fronteiras entre humanos e máquinas para explorar as exclusões e os espaços de contestação. De Vries e Schinkel (2019), por exemplo, mostraram como é que os artistas criticam ou contrariam as normatividades dos algoritmos e os efeitos de vigilância materializados nas tecnologias de reconhecimento facial quando desenham as suas obras de arte, inspirando a criação de diversos imaginários sobre os

usos de algoritmos. Já Borgdorf e colegas (2020) mostraram como os diálogos e os encontros entre a criatividade e invenção artísticas e os Estudos Sociais da Ciência e Tecnologia podem gerar novos conhecimentos e ações ao reconfigurar as conceções tradicionais sobre o que são dados empíricos e ao incitar abordagens intervencionistas inovadoras com base em modalidades distintas de participação e envolvimento dos públicos.

Uma abordagem metodológica alternativa para estudar a IA é-nos proposta por Cláudia Aradau e Tobias Blanke (2022), no livro intitulado *A razão algorítmica: O novo governo do eu e do outro*. Com o propósito de compreender como é que operações algorítmicas dispersas e desalinhadas, a nível global, são conectadas através de uma razão algorítmica ascendente, os autores utilizaram métodos qualitativos e digitais para investigar cenas (*scenes*) e controvérsias, designadamente: A vigilância em massa e o escândalo *Cambridge Analytica* no Reino Unido (que envolveu a exploração de dados pessoais de cerca de 87 milhões de utilizadores do *Facebook*); o policiamento preditivo nos Estados Unidos da América; o uso de tecnologias de reconhecimento facial na China; os ataques com *drones* direcionados no Paquistão; ou a regulação dos discursos de ódio na Alemanha. A originalidade desta lente analítica reside em olhar para as profundas transformações associadas às tecnologias de IA de uma forma mais mundana, material e envolvente. Tornou-se assim possível para os autores compreender como é que a racionalidade algorítmica (decomposição, recomposição e partição) molda o governo do Eu e do Outro ao redesenhar fronteiras e reconfigurar diferenças que se concretizam na construção do Outro como perigo ou ameaça, no poder das plataformas e na produção de valor económico. Esta abordagem possibilita, ainda, explorar as fricções, as recusas e as resistências com que as intervenções políticas se deparam quando tentam tornar os algoritmos governáveis.

3.7.2. Metodologias mistas

Nos últimos anos, as metodologias qualitativas e quantitativas têm vindo a ser combinadas nas designadas metodologias mistas. Estas envolvem a recolha, análise, integração e interpretação de dados qualitativos e quantitativos para responder a questões de investigação. A utilização de metodologias mistas está geralmente associada a um dos seguintes objetivos: Triangulação, ou seja, comparação de dados quantitativos e qualitativos com o propósito de corroborar os resultados; complementaridade, isto é, aprimoramento ou clarificação dos resultados, onde as componentes qualitativa e quantitativa representam peças diferentes de um *puzzle*; ou expansão, quando os dados são apresentados "lado a lado", mantendo-se intactos pois respondem a diferentes questões (Creswell, 2015). Um exemplo é o projeto *Shaping AI*[32], que contempla uma análise crítica e comparativa das trajetórias globais dos discursos públicos em torno da IA na Alemanha, Reino Unido, Canadá e França, entre 2012 e 2021, com base no estudo de controvérsias registadas nos média, na política, na investigação, e na participação e envolvimento dos públicos.

32 O *Shaping AI* é um projeto de investigação social multidisciplinar financiado pela iniciativa *European Open Research Area* (Fevereiro 2021 – Fevereiro 2024), cuja descrição sumária está disponível em https://www.shapingai.org/ [Acesso a 24 de julho de 2024].

3.7.3. Metodologias participativas

Quando a utilização de metodologias quantitativas, qualitativas ou mistas é acionada com objetivos democráticos e participativos, valorizando-se a colaboração direta e significativa de participantes que representam os interesses ou façam parte dos grupos que são afetados pelo tópico em estudo numa investigação orientada para promover ações ou mudanças ou para resolver problemas complexos, estamos perante uma investigação participativa. Não é nosso propósito esmiuçar a diversidade de abordagens e orientações que enquadram as metodologias participativas (para esse efeito, ver a proposta de síntese elaborada por Vaughn e Jacquez, 2020), antes realçar o que as une: Fazer investigação com as pessoas e não acerca delas, produzindo conhecimento com impacto no mundo real. Isto significa incorporar o envolvimento e a participação de diversos públicos (investigadores, partes interessadas, membros de comunidades ou de grupos afetados pelos usos de tecnologias de IA) na investigação sobre as implicações sociais e éticas da IA, recorrendo a instrumentos, tarefas ou atividades que facilitem a participação e promovam a tomada de decisão compartilhada e a aprendizagem mútua. No caso da participação em processos de tomada de decisão política, alguns destes exercícios incluem os júris de cidadãos e a realização de *workshops* (Reeve et al., 2023).

Ora, como vimos no capítulo 2, algumas iniciativas que visam o envolvimento dos públicos alegadamente com base em metodologias participativas têm sido instrumentalizadas ao serem usadas com o propósito de legitimar a inovação e de promover a ideia de uma IA responsável segundo os cânones dos seus empreendedores. A abordagem sociológica à IA requer, por isso, metodologias de investigação inovadoras e imaginação para poder desconstruir contextos, discursos e interações no campo da IA no século XXI.

4. A Inteligência Artificial na educação

Introdução

Os potenciais impactos da introdução de ferramentas de Inteligência Artificial (IA) na educação tem suscitado amplo debate nas duas últimas décadas, existindo um vasto conjunto de literatura académica, relatórios e estudos internacionais sobre o tema. Em 2021, a OCDE publicou uma coletânea com textos de vários especialistas versando sobre distintos aspetos do panorama da educação digital, incluindo considerações sobre o uso de IA e robôs nesse setor (OCDE, 2021). Este estudo reconhece os contributos positivos da IA, designadamente em termos de "eficácia, equidade e relação custos-eficiência dos sistemas educativos", mas deixou o seguinte apelo:

> As tecnologias inteligentes (*smart technologies*) são sistemas híbridos humano-IA. Envolver os utilizadores finais na sua conceção, dar controlo aos seres humanos nas decisões importantes e negociar a sua utilização com a sociedade de forma transparente é fundamental para as tornar úteis e socialmente aceitáveis. (OCDE, 2019, sumário executivo)

Não obstante o reconhecimento das potencialidades da IA na educação, é importante notar que a sua implementação responsável e confiável neste setor requer considerações éticas (Holmes et al., 2022a; Vincent-Lancrin e van der Vlies, 2020) e ponderação de implicações sociais. Alguns exemplos dos desafios sociais e éticos a considerar no que diz respeito ao uso de IA no setor da educação (detalhados nas secções 4.4. e 4.5. deste capítulo) vão desde a garantia da privacidade dos dados de estudantes e professores, a medidas que previnam e combatam a exclusão digital, e estratégias de regulação que tenham como tónica manter sempre a supervisão humana sobre as decisões propostas por sistemas de IA. Além disso, várias organizações internacionais e especialistas em IA na educação defendem que esta deve ser encarada como uma ferramenta complementar ao ensino, e não como um substituto completo para a interação humana (Comissão Europeia, 2022; Conselho Europeu, 2019; OCDE, 2021; Schiff, 2020).

Outro aspeto importante a considerar é o facto que a aplicação de uma determinada tecnologia nunca é linear. Por outras palavras, conforme é referido nos três primeiros capítulos deste livro, os usos potenciais de uma determinada tecnologia nunca são apenas condicionados pelas características técnicas da mesma, mas também pelos contextos históricos, sociais, culturais, políticos e económicos em que são utilizadas e pelas características e comportamento do utilizador. Como faz notar Schiff:

> Os tecnólogos da educação que acreditam que os seus produtos serão utilizados de uma determinada forma estão condenados a ficar desapontados ou, pelo menos, surpreendidos, uma vez que os professores e os alunos modificam e inovam através de um processo de adaptação estratégica. Uma ferramenta imaginada para um determinado fim acaba por ser reutilizada para outra finalidade. (2020, p. 334)

Ainda que a IA seja frequentemente encarada como uma solução para muitos dos principais problemas da educação (por exemplo, a falta de professores, o insucesso escolar dos alunos, e o fosso crescente entre alunos ricos e pobres) (OCDE, 2021), uma compreensão abrangente das implicações sociais e éticas da IA na educação suscita a necessidade de considerar múltiplas questões, nomeadamente: Quais os objetivos da utilização da IA na educação? Onde é que a IA é utilizada, e por quem? Como é que a IA funciona e é operacionalizada a diversos níveis (desde o aluno individual a turmas inteiras, ou redes de colaboração a nível nacional e transnacional)?

Procurando responder a estas questões, nas próximas secções identificamos os principais temas que têm sido trabalhados no campo da IA no setor da educação, em particular no que respeita a aprendizagem com IA, apontando para a importância de articular o atual enfoque na dimensão tecnológica com aspetos sociais e éticos na abordagem à literacia em matéria de IA. Prosseguimos com o mapeamento das aplicações de IA na educação, desde os sistemas de tutoria inteligentes, jogos digitais, simulações, sistemas de realidade virtual e robôs educativos, até aos mais recentes sistemas de aumento da inteligência e sistemas de aprendizagem personalizada. Por fim, apresentamos, de forma crítica, algumas recomendações e propostas de atividades para que educadores e diretores de instituições de ensino possam operacionalizar a aplicação de princípios éticos ao usar a IA e dados digitais no ensino e na aprendizagem, incorporando reflexões sobre a influência de fatores políticos e económicos, a reprodução de desigualdades sociais e discriminação, e o ideário de modelo pedagógico que subjaz às tecnologias de IA na educação.

Em síntese, neste capítulo pretendemos:

- Mapear os temas da IA na educação.
- Discutir como é que a IA tem sido utilizada em processos de aprendizagem.
- Dar a conhecer algumas aplicações de IA em contexto de ensino e aprendizagem.
- Explorar, de forma crítica, formas de operacionalizar a aplicação de princípios éticos no setor da educação que considerem aspetos sociais.

4.1. Temas da Inteligência Artificial na educação

Na esteira da proposta de Holmes e colegas (2019, 2022b), subdividimos em quatro grandes áreas as temáticas da IA na educação, embora sem fronteiras rígidas entre elas: "Aprendizagem com a IA"; "utilização da IA para aprender sobre a aprendizagem"; "aprender sobre a IA"; e "preparação para a IA". Descrevemos, de seguida, cada uma destas áreas de forma sumária.

A *aprendizagem com a IA* envolve a utilização de ferramentas orientadas para a IA no ensino e na aprendizagem e a utilização da IA para apoiar diretamente os estudantes, incluindo sistemas tutoriais inteligentes, sistemas de tutoria baseados no diálogo, ambientes de aprendizagem exploratórios, avaliação automática da escrita, redes de aprendizagem, *chatbots* e IA para apoiar os alunos com deficiências.

Contempla, ainda, a utilização da IA para apoiar os sistemas administrativos associados ao ensino, desde a definição de horários à gestão da aprendizagem.

A *utilização da IA para aprender sobre a aprendizagem* envolve a análise de dados resultantes da utilização de IA na aprendizagem, para saber como os alunos aprendem, como progridem na aprendizagem ou quais as conceções de aprendizagem que são eficazes. O objetivo é facultar dados que permitam tomar decisões, por exemplo, ao nível do planeamento de programas de ensino e aprendizagem.

Aprender sobre a IA envolve aumentar o conhecimento e as competências de IA dos estudantes de todas as idades (ou seja, desde o ensino básico e secundário até ao ensino superior) e dos seus professores. Em suma, significa promover a literacia em IA na sua dimensão tecnológica (Holmes et al., 2022b, p. 19).

Por fim, a *preparação para a IA* implica alertar e formar os cidadãos para os possíveis impactos da IA nas suas vidas, ajudando-os a ir além do mero entusiasmo ou receio em relação a tecnologias de IA, e capacitando-os para compreender questões como, por exemplo, o modo como a IA apoia tomadas de decisão e os possíveis enviesamentos daí decorrentes podem fragilizar a privacidade e liberdades individuais, ou podem alterar significativamente os empregos. Neste contexto, estaremos a falar de literacia em IA na sua dimensão humana (Holmes et al., 2022b, p. 19).

É importante notar que, até há pouco tempo, prevalecia uma abordagem da literacia em IA do ponto de vista técnico e, nesse sentido, o ensino sobre IA tem sido principalmente da responsabilidade dos cientistas da computação. O enfoque na dimensão tecnológica da IA tende a desviar a atenção dos aspetos sociais e éticos. Uma maneira de começar a abordar essas dimensões passa por incentivar todos os professores, desde os que lecionam disciplinas das ciências até às humanidades e artes, e não apenas os professores na área das ciências da computação, a explorar com os seus estudantes os usos potenciais, benefícios, impactos, desafios e riscos da IA. Por exemplo, dado que a IA pode ser usada para gerar automaticamente imagens digitais e escrever poemas, os professores de arte e de literatura podem perguntar aos seus estudantes: Se uma máquina pode ser capaz de atos criativos, o que significa ser humano (Bringula, 2023)?

Em 2019, o Comissário Europeu para os Direitos Humanos, num documento de recomendações sugestivamente intitulado "Descompactando a Inteligência Artificial: 10 passos para proteger os Direitos Humanos" (*Unboxing Artificial Intelligence: 10 steps to protect Human Rights*), proferiu a seguinte recomendação:

> Os Estados-Membros devem investir no nível de literacia em matéria de IA junto do público em geral, através de esforços sólidos de sensibilização, formação e educação, incluindo (em especial) nas escolas. Este esforço não se deve limitar à educação sobre o funcionamento da IA, mas também sobre o seu potencial impacto – positivo e negativo – nos direitos humanos. Devem ser feitos esforços especiais para chegar aos grupos marginalizados e aos que

> estão em desvantagem no que respeita à literacia informática em geral. (Conselho Europeu, 2019, pp. 14-15)

Outras questões que devem ser consideradas incluem os princípios de "confiança" e de "responsabilidade" explorados no capítulo 2, conforme identificado por vários organismos internacionais como a OCDE (2023), a UNESCO (2021), o Fórum Económico Mundial (2023) e a Comissão Europeia (2020). Neste contexto, um importante debate recai sobre como garantir que as tecnologias de IA estejam alinhadas com valores sociais e orientadas para a criação de um futuro mais justo, equitativo e inclusivo, enraizado em princípios éticos como o bem-estar, respeito da autonomia, proteção da privacidade, solidariedade, participação democrática, equidade, diversidade, prudência, responsabilidade e desenvolvimento sustentável (Ulnicane et al., 2021a, 2021b). Atendendo a que estes princípios são gerais e abstratos, o maior desafio será definir como operacionalizar a sua aplicação no âmbito do setor específico da educação. Este aspeto será desenvolvido na secção 4.4. do presente capítulo.

4.2. A aprendizagem com Inteligência Artificial

Em termos gerais, quando se pensa na utilização de IA em processos de aprendizagem são apontados os seguintes aspetos: Personalizar a aprendizagem; apoiar professores e educadores; e melhorar o acesso à educação (Crompton e Burke, 2023; Holmes et al., 2022b). Mas de que forma é que a IA pode ser mobilizada para atingir estes propósitos?

No caso da *personalização da aprendizagem*, a IA pode adaptar o conteúdo de ensino de acordo com as necessidades e competências individuais dos estudantes. Isso permite que cada um progrida ao seu próprio ritmo. Os sistemas de IA podem também fornecer *feedback* imediato sobre o desempenho em tarefas e avaliações. Isso ajuda os estudantes a identificar áreas de melhoria em tempo real. Do mesmo modo, assistentes virtuais educacionais podem oferecer apoio permanente aos estudantes, respondendo a perguntas, fornecendo orientações e ajudando na resolução de problemas.

No âmbito do *apoio a professores e educadores*, estes podem usar ferramentas de IA para criar recursos de aprendizagem mais envolventes. A IA pode, por exemplo: Recomendar materiais de estudo e cursos *online* com base no estilo de aprendizagem e nas metas educacionais de cada aluno; detetar padrões nos dados de desempenho dos estudantes e identificar áreas em que possam estar a enfrentar dificuldades; e tornar os processos de avaliação mais eficientes e confiáveis, reduzindo a carga de trabalho dos professores na correção de tarefas e exames. É ainda destacado o potencial efeito da IA em economia de tempo, na medida em que se considera que a IA pode assumir tarefas administrativas, como avaliações e organização de materiais.

Quanto a *melhorar o acesso à educação*, refere-se a forma como as plataformas de educação *online* com IA podem disponibilizar conteúdo educacional para um público global, independentemente de limitações geográficas. Outra vantagem invocada em relação à IA é poder ajudar na adaptação de materiais de ensino para estudantes

com necessidades especiais, tornando a educação mais acessível e inclusiva. Alguns exemplos deste último aspeto são os seguintes: Aplicações como a conversão da oralidade em texto, a conversão do texto em voz, a legendagem automática, etc., permitem que estudantes cegos, com deficiência visual, surdos ou com dificuldades auditivas participem em ambientes e práticas educativas "tradicionais". Algumas tecnologias de IA facilitam o diagnóstico e a correção de algumas necessidades especiais (por exemplo, disgrafia) e apoiam a aprendizagem socio-emocional de estudantes com autismo, para que possam participar mais facilmente no ensino regular.

4.3. Panorama das aplicações de Inteligência Artificial na educação

A presença de computadores nas instituições de ensino e a digitalização do setor da educação não são fenómenos novos. Os seguintes exemplos apresentam tecnologias digitais com funções educativas, algumas das quais em aplicação já desde a década de 1980 (ver Baker, 2021; Belpaeme e Tanaka, 2021).

Os *tutores informáticos ou sistemas de tutoria inteligentes* proporcionam aos alunos uma experiência de aprendizagem em que o sistema adapta a apresentação de conteúdos com base num modelo pré-definido de como ensinar e geralmente incorporando uma avaliação contínua do estudante.

Os *jogos digitais* de aprendizagem incorporam a aprendizagem numa atividade divertida que se assemelha a um jogo. O grau de gamificação pode variar, desde atividades que integram a aprendizagem no jogo e que podem nem sequer parecer uma atividade de aprendizagem (por exemplo, os jogos *SimCity* e *Civilisation*[33]) até atividades de aprendizagem mais óbvias, em que o aluno recebe recompensas pelo seu desempenho (por exemplo, atirar uma banana a um macaco depois de responder corretamente a um problema de matemática no jogo educativo *MathBlaster*).

As *simulações* são imitações computorizadas de um processo ou de uma atividade que seria difícil ou dispendiosa de realizar no mundo real como atividade educativa, usando-se, para esse efeito, laboratórios virtuais.

Os *sistemas de realidade virtual* integram os estudantes em representações *3D* (a três dimensões) de atividades do mundo real. Tal como as simulações, estes sistemas de realidade virtual tornam viável a participação em atividades a partir de casa ou de um laboratório informático que seriam dispendiosas, perigosas ou simplesmente impossíveis de realizar de outra forma. Os sistemas de realidade aumentada integram informações e experiências adicionais em atividades do mundo real, desde pormenores que aparecem em *pop-up* e ecrãs de ambiente (informações que estão disponíveis no ambiente sem que seja necessário focá-las) até à sobreposição de um

33 *SimCity* é uma série de jogos de simulação, na qual o jogador constrói e administra uma cidade; enquanto que no *Civilisation* o objetivo é construir um grande império, começando na era antiga, mas desenvolvendo o império através das eras e concorrendo com diversas outras civilizações que podem tornar-se aliadas ou inimigas.

mundo diferente sobre o atual. Tanto a realidade aumentada como a realidade virtual recorrem frequentemente a auscultadores para apresentar informações visuais aos alunos.

• Os *robôs educativos* têm uma presença física e interagem com os alunos para apoiar a sua aprendizagem. Embora os robôs neste contexto estejam disponíveis desde os anos 1980, desenvolvimentos recentes estimam que os robôs possam vir a assumir o papel de tutores.

Até agora, a IA em educação focou-se principalmente em sistemas tutoriais inteligentes. Porém, mais recentemente, com a expansão de dados digitais e o desenvolvimento da área de Análise de Aprendizagem (*learning analytics*), também designada como extração de dados educativos (*educational data mining*), a IA tem como objetivo utilizar as quantidades crescentes de dados digitais provenientes da educação para compreender melhor os processos de aprendizagem e fazer inferências sobre os estudantes e os contextos em que estes aprendem (Baker, 2019). A Análise de Aprendizagem aplica técnicas de IA de aprendizagem automática (*machine learning*) – processo que explora o estudo e a construção de algoritmos (instruções para resolver problemas específicos e efetuar cálculos) – à educação, ou seja, usando problemas específicos da educação. Desafios como a inferência de conhecimentos dos estudantes em tempo real, a previsão do futuro abandono escolar e a compreensão dos fatores que levam à desmotivação dos alunos têm suscitado um interesse particular neste domínio.

Os modelos derivados da Analítica de Aprendizagem são frequentemente utilizados em dois tipos de tecnologia: Sistemas de aumento da inteligência (*intelligence augmentation systems*) e sistemas de aprendizagem personalizada (*personalised learning systems*) (Baker, 2016).

Os *sistemas de aumento da inteligência*, também designados por sistemas de apoio à decisão, comunicam informações às partes interessadas, como os professores ou diretores de turma e de escolas, de forma a apoiar a tomada de decisões. Embora possam simplesmente fornecer dados brutos, muitas vezes fornecem informações, ou seja, dados brutos tratados através de modelos de aprendizagem automática, que se convertem em previsões ou recomendações. Neste contexto, os sistemas de aumento da inteligência são frequentemente utilizados para fazer previsões sobre os potenciais resultados futuros dos alunos e também fornecendo razões para essas previsões. Por exemplo, os sistemas de análise preditiva podem ser usados para tentar compreender quais os estudantes que correm o risco de abandonar o ensino secundário ou de não concluir o ensino superior, com o objetivo de sugerir intervenções para prevenir esse insucesso ou abandono escolar.

Nos *sistemas de aprendizagem personalizada*, a tecnologia de IA pode identificar o grau de domínio de cada aluno e proporcionar-lhe atividades de aprendizagem de acordo com o seu grau de conhecimento ou estado de aprendizagem, dentro das orientações e objetivos especificados pelo professor (Molenaar, 2021). Este sistema está geralmente vinculado a uma aprendizagem autorregulada dos alunos – a sua

capacidade de fazer escolhas "adequadas" durante a aprendizagem que melhorem os seus resultados e a sua eficiência formativa. Este tipo de tecnologia tem também a capacidade de reconhecer quando os alunos estão a utilizar estratégias ineficazes ou ineficientes e de lhes fornecer recomendações ou sugestões para voltarem a uma trajetória mais eficaz (D'Mello, 2021; DeFalco et al., 2017).

Com o intuito de clarificar os conceitos técnicos associados à IA e explicitar as suas aplicações em contexto de ensino e aprendizagem, apresentamos a seguinte síntese, que reproduzimos a partir de um glossário sobre IA e Dados Digitais proposto pela Comissão Europeia (2022) para o contexto da educação (Tabela 6).

Tabela 6
Aplicações de Inteligência Artificial na educação.

Termo técnico (IA)	Significado	Aplicações na educação
Algoritmo	Um processo ou conjunto de regras a serem seguidas em cálculos ou outras operações de resolução de problemas, neste caso, por um computador.	Os algoritmos de IA podem revelar padrões no desempenho dos alunos e ajudar os professores a otimizar as suas estratégias/ metodologias de ensino para personalizar a aprendizagem e melhorar os resultados.
Análise de aprendizagem	Envolve a medição, a recolha, a análise e a comunicação de dados sobre os alunos e os seus contextos, com o objetivo de compreender e otimizar a aprendizagem e os ambientes em que esta ocorre.	Os sistemas de gestão da aprendizagem registam dados sobre a interação dos alunos com os materiais do curso, a sua interação com os professores e outros colegas e o seu desempenho nas avaliações digitais. As escolas podem utilizar a análise destes dados para monitorizar o desempenho dos alunos, prever o desempenho global e facilitar a prestação de apoio através de *feedback* personalizado a cada aluno.
Análise preditiva	Utilização de algoritmos estatísticos e de técnicas de aprendizagem automática para fazer previsões sobre o futuro utilizando dados atuais e históricos.	A análise preditiva pode fornecer informações sobre os alunos que necessitam de apoio adicional, não só com base no seu desempenho atual e histórico, mas também no seu desempenho futuro previsto.
Automação	Um sistema informático que pode executar tarefas sem necessitar de supervisão humana contínua é descrito como autónomo.	As escolas e os professores podem utilizar sistemas informáticos para executar muitas tarefas repetitivas e morosas, como o registo de horários, a assiduidade e inscrições.
Big Data	Conjuntos de dados extremamente grandes e complexos que desafiam as capacidades tradicionais de processamento de dados.	Através da análise de grandes volumes de dados, os educadores podem potencialmente identificar áreas em que os alunos têm dificuldades ou prosperam, compreender as necessidades individuais dos alunos e desenvolver estratégias para uma aprendizagem personalizada.

Chatbot	Um programa que comunica através de texto ou voz imitando as interações entre humanos.	Os *chatbots* podem ser conselheiros virtuais para os alunos e, nesse processo, adaptar-se ao seu ritmo de aprendizagem, ajudando assim a personalizar a sua aprendizagem. As suas interações com os alunos podem também ajudar a identificar as matérias em que estes precisam de ajuda.
Processamento de linguagem natural	É uma forma de IA que coloca os computadores a ler e a responder, simulando a capacidade humana de compreender a linguagem quotidiana.	O processamento de linguagem natural pode ser aplicado a sistemas de tutoria virtual. Por exemplo, utilizando o reconhecimento de voz para identificar problemas na capacidade de leitura de um aluno.
Realidade Aumentada (RA)	É uma experiência interativa em que os ambientes e objetos do mundo real são complementados por modelos 3D, gerados por computador, e sequências animadas que são apresentadas como se estivessem num ambiente do mundo real.	A RA cria oportunidades para os professores ajudarem os alunos a apreender conceitos abstratos através da interação e da experimentação com materiais virtuais. Este ambiente de aprendizagem interativo oferece oportunidades para implementar abordagens de aprendizagem práticas e melhorar a experiência de aprendizagem.

Fonte. Adaptado de Comissão Europeia, 2022, pp. 32-36.

4.4. Desafios éticos

Em 2022, a Comissão Europeia publicou um conjunto de recomendações sobre o uso de IA e dados digitais no ensino e na aprendizagem (Comissão Europeia, 2022). Essas recomendações, especificamente dirigidas a professores do ensino básico e secundário, abordam considerações e os requisitos éticos, sendo apresentados conselhos práticos aos educadores e aos diretores de instituições de ensino. Este documento discute ainda as competências emergentes para a utilização ética da IA e dos dados digitais, sugerindo formas de sensibilização e de envolvimento da comunidade. Reproduzimos, de seguida, algumas dessas recomendações (Tabela 7).

Tabela 7
Principais recomendações sobre o uso de Inteligência Artificial na educação.

Requisitos éticos	**Algumas recomendações**
Ação e supervisão humanas	O papel do professor está claramente definido enquanto o sistema de IA está a ser utilizado?
	Como é que o sistema de IA afeta o papel didático do professor?
	Os professores e os diretores das escolas têm toda a formação e informação necessárias para utilizar eficazmente o sistema e garantir que é seguro e não causa danos ou viola os direitos dos alunos?

Transparência	São claros quais os aspetos que a IA pode assumir e quais os que não pode assumir no sistema?
	Os professores e os diretores das escolas compreendem como funcionam os algoritmos específicos de avaliação ou de personalização no sistema de IA?
	As instruções e informações são acessíveis e apresentadas de forma clara, tanto para os professores como para os estudantes?
	Até que ponto as previsões, avaliações e classificações do sistema de IA são fiáveis para explicar e avaliar a pertinência da sua utilização?
Diversidade, não discriminação e equidade	O sistema é acessível a todos da mesma forma, sem quaisquer barreiras?
	O sistema oferece modos de interação adequados aos alunos com deficiências ou necessidades educativas especiais?
	Existem procedimentos para garantir que a utilização da IA não conduzirá a discriminação ou comportamento injusto para os utilizadores?
	Existem procedimentos para detetar e lidar com a parcialidade ou desigualdades que possam surgir?
Bem-estar societal e ambiental	Como é que o sistema de IA pode afetar o bem-estar social e emocional dos estudantes e dos professores? Como é que isso é monitorizado?
	O sistema de IA indica claramente que a sua interação social é simulada e que não tem capacidade real de expressar sentimentos ou de empatia?
	Os estudantes ou os seus pais estão envolvidos na decisão de utilizar o sistema de IA?
	A utilização do sistema cria algum dano ou receio para os indivíduos ou para a sociedade?
Privacidade e governação dos dados	Existem mecanismos para garantir que os dados sensíveis são mantidos anónimos?
	Existem procedimentos para limitar o acesso aos dados apenas a quem deles necessita?
	O acesso aos dados dos estudantes está protegido e armazenado num local seguro e é utilizado apenas para os fins para os quais os dados foram recolhidos?
	Existe um mecanismo que permite aos professores e dirigentes escolares assinalar questões relacionadas com a privacidade ou a proteção de dados?
	Os estudantes e os professores são informados sobre o que acontece com os seus dados, como são utilizados e para que fins?
	É possível personalizar as definições de privacidade e de dados?
	O sistema de IA está em conformidade com o Regulamento Geral sobre a Proteção de Dados?
Solidez técnica e segurança	Existe uma estratégia para monitorizar e testar se o sistema de IA está a cumprir os objetivos, finalidades e aplicações pretendidas?
	Existem mecanismos de supervisão adequados para a recolha, armazenamento, processamento, minimização e utilização de dados?
Responsabilização	Quem é responsável pela monitorização contínua dos resultados produzidos pelo sistema de IA e pela forma como os resultados estão a ser utilizados para melhorar o ensino, a aprendizagem e a avaliação?
	Como está a ser avaliada a eficácia e o impacto do sistema de IA e como é que esta avaliação tem em conta os valores fundamentais da educação?
	Quem é responsável pelas decisões finais tomadas relativamente à aquisição e implementação do sistema de IA?
	Existe um protocolo que defina claramente os serviços de apoio e manutenção e as medidas a tomar para resolver os problemas comunicados?

Fonte. Adaptado de Comissão Europeia, 2022, pp. 19-21.

A abordagem apresentada dos desafios éticos suscitados pelo uso de tecnologias de IA na educação tem limitações importantes, desde logo porque não considera uma dimensão essencial: A ética da/na educação (Holmes et al., 2022b, p. 520). É necessário considerar também a ética das expectativas de professores e de estudantes, do investimento feito em sistemas de IA no conjunto das decisões de como usar os recursos existentes, a ponderação desse investimento no contexto em que é aplicado, daquilo que pode ser considerado uma forma de conhecimento útil, dos papéis dos professores e dos estudantes e das relações de poder subjacentes, e das abordagens pedagógicas particulares que serão afetadas ou alteradas. Todos estes aspetos podem assumir uma relevância mais acutilante quando se trata de introduzir a IA na educação de grupos vulneráveis, como crianças, pessoas com deficiências ou populações sujeitas a adversidades e vulnerabilidades de diferente natureza (por exemplo, minorias étnicas, refugiados, pessoas não binárias e outros grupos ou indivíduos sujeitos a exclusão social, discriminação e violação de direitos humanos) (Holmes et al., 2022b).

4.5. Desafios sociais

O uso de tecnologias de IA na educação suscita a necessidade de considerar aspetos sociais mais gerais, que extravasam as considerações éticas anteriormente expostas. Isto acontece não só porque uma análise dos impactos de uma tecnologia – e a IA não é exceção – deve considerar sempre o contexto mais lato em que esta é utilizada; como também o próprio objetivo de introduzir a IA na educação vai ter implicações profundas nos cidadãos e nas sociedades. Por outras palavras, as escolhas feitas pela sociedade em relação ao tipo de pedagogia a implementar, o que e como ensinar, o que são os conhecimentos válidos a adquirir pelos estudantes e qual é o papel a conferir aos professores, traduz aquilo que é valorizado e o modo como se concebe o que é ser cidadão. De seguida, apresentamos algumas das problemáticas sociais subjacentes à IA na educação.

Fatores políticos e económicos

A introdução de tecnologias de IA na educação tem subjacente um determinado modelo de ensino e de aprendizagem e essa escolha é sempre um ato político. O modelo pedagógico subjacente às tecnologias de IA está orientado para a aquisição de competências práticas e de resolução de problemas, que pode ser adequado em determinados contextos, mas não em outros. Este tipo de ensino é sobretudo utilitarista e orientado para competências mensuráveis de modo quantitativo, não sendo adequado para um ensino sustentado em valores humanistas, que afirmem a dignidade humana, a compaixão, a moral, a ética e a democracia (Holmes et al., 2022b).

Do mesmo modo, a introdução da IA na educação pode ser considerada à luz dos objetivos comerciais que movem as empresas que desenvolvem estas tecnologias. Cabe assim perguntar até que ponto as escolas realmente necessitam destas tecnologias, pois o desenvolvimento das mesmas pode não ser baseado em necessidades

pedagógicas reais, mas sim artificialmente criadas por empresas que procuram o lucro (Knox, 2020). Em última instância, um investimento forte na introdução de IA nas escolas pode estar a transferir para as empresas tecnológicas a definição de modelos "adequados" de ensino e aprendizagem.

Discriminação e desigualdades

Muitas vezes é invocado que um dos grandes benefícios da IA na educação é robustecer o acesso universal à educação, contribuindo, assim, para diminuir desigualdades. Mas o inverso também pode acontecer: Por exemplo, a maioria das ferramentas de IA para utilização na educação exige um certo nível de competência técnica e de conhecimentos linguísticos. Por conseguinte, a IA pode exacerbar, em vez de atenuar, as desigualdades na educação, afetando particularmente as comunidades marginalizadas e cavando um fosso ainda maior entre as escolas que têm acesso a infraestrutura tecnológica e digital e aquelas que não têm.

É importante fazer notar que as tecnologias de IA para a educação tendem a ser treinadas e desenvolvidas em inglês americano padrão, pelo que a diversidade linguística e cultural continua a ser um desafio. A sua utilização dificilmente poderá acontecer de modo sensível e respeitador de especificidades nacionais e locais, e de modo a proteger a cultura e identidade nacional ou étnica, cultural, religiosa e linguística (Holmes et al., 2022b, p. 46).

Por fim, é importante salientar que as tecnologias de IA são concebidas, desenvolvidas e implementadas com base em condicionamentos sociais e culturais. Desta forma, a própria tecnologia incorpora valores sociais que podem projetar e reproduzir preconceitos contra ou a favor de uma pessoa, objeto ou posição. Os preconceitos podem surgir de muitas formas nas tecnologias de IA. Por exemplo, nos sistemas de IA baseados em dados, como os produzidos através da aprendizagem automática, a parcialidade na recolha de dados e no treino do programa informático pode fazer com que um sistema de IA reproduza essa parcialidade. Nos sistemas de IA baseados na lógica, como os sistemas baseados em regras, a parcialidade pode surgir devido à forma como os desenhadores podem definir as regras que se aplicam num determinado contexto. Em ambas as circunstâncias, as suposições feitas pelos algoritmos de IA podem amplificar os preconceitos existentes, relacionados com o género, a raça ou estatuto de deficiência.

Na maioria dos casos, as tecnologias de IA aplicadas em contexto de educação são desenvolvidas através de dados de treino que são extraídos de bases de utilizadores iniciais. Assim, estas tecnologias aprendem sobre o envolvimento, a aprendizagem e os estilos cognitivos dos estudantes a partir de um subconjunto de utilizadores. É plausível que estes utilizadores venham de sistemas escolares com mais recursos e que possam comprar e experimentar estas novas ferramentas. Estes utilizadores podem também vir de países mais ricos e, especialmente, de países onde o desenvolvimento da IA na educação é predominante. Neste contexto, a questão da utilização intercultural e do preconceito emerge como uma preocupação real, o que vem abalar

o argumento que seriam os países de baixo rendimento aqueles que poderiam ser mais positivamente afetados pela IA (Schiff, 2020).

Modelo pedagógico

Além das considerações já feitas relativamente ao modelo pedagógico subjacente às tecnologias de IA na educação (natureza utilitarista, predomínio de determinados valores em detrimento de outros), é importante atender a que este tipo de ensino se adequa mais facilmente às áreas da ciência, tecnologia, engenharia e matemática (conhecidas por *STEM*), sendo que a produção de conteúdos para as áreas das artes, humanidades e ciências sociais pode ser mais desafiadora e complexa para as empresas tecnológicas que desenvolvem IA. Num contexto de hipervalorização e entusiasmo em torno da IA, há áreas de ensino e aprendizagem que apresentam especificidades que não são compatíveis com as técnicas de IA dominantes, o que pode fazer com que se acentue o processo de desvalorização política e económica destes ramos do saber.

Outro aspeto importante a considerar é a transformação do papel dos professores. Os defensores da introdução da IA na educação argumentam que estas tecnologias se destinam a empoderar os professores e não a substitui-los. Por exemplo, é corrente a ideia que as tecnologias de IA podem libertar os professores de tarefas rotineiras, dando-lhes, em contrapartida, mais tempo e condições para se dedicarem a vertentes mais criativas e empáticas do processo de ensinar. No entanto, vários fatores podem vir a fragilizar o papel dos professores num contexto de ampliação da introdução da IA na educação. Em primeiro lugar, continua o desenvolvimento de investigação na área da criação de ambientes de ensino e aprendizagem totalmente virtuais, sem a presença de um professor. Em segundo lugar, a agência do professor em termos de lhe ser possível escolher se usa ou não IA nas suas aulas é geralmente limitada (estando essa decisão mais do lado de líderes políticos, de diretores de escolas e de outros atores com poder para decidir). O que parece consensual é a constatação que um dos objetivos principais de introduzir a IA na educação é diminuir o rácio professor/estudante, com base na convicção que a IA introduz maior eficiência no sistema educativo (Schiff, 2020, p. 340).

Por fim, ainda no que toca ao modelo pedagógico subjacente às tecnologias de IA, uma preocupação constante dos especialistas na área da IA na educação são os riscos de manipulação. À medida que se desenvolvem tecnologias de IA com capacidade para prever, medir e responder a dados emocionais, levantam-se preocupações sérias sobre as implicações destas respostas (Borenstein e Arkin, 2017; Schiff, 2020, p. 342). Idealmente, estes sistemas emocionais seriam utilizados para apoiar e melhorar os estudantes do ponto de vista educativo e socio-emocional. Mas quem controla o que são respostas emocionais "adequadas"? Como proteger os estudantes mais vulneráveis (devido à idade, estatuto socioeconómico ou outras características pessoais)? Quais as implicações sociais e cognitivas de expor estudantes (por exemplo, crianças) a estímulos emocionais produzidos por uma máquina e muitas vezes

em formatos que podem causar adição (pelo estímulo visual, pela repetição e por conteúdos pensados mais por lógicas comerciais do que educacionais)?

4.6. Atividades para debate

Apresentam-se exemplos de casos para debate à luz das questões sociais e éticas analisadas neste capítulo e desenvolvidas em termos gerais e abstratos com mais detalhe no capítulo 2. Considere ainda as diferentes dimensões e os diversos níveis de análise apresentados no capítulo 3, selecionando aqueles que lhe pareçam mais adequados à análise de cada caso e justificando porquê.

Caso 1

Uma escola está a ponderar introduzir um sistema de tutoria inteligente, que adapta automaticamente a apresentação de conteúdos combinando um modelo pré-definido de programa de ensino com mecanismos de adaptação às necessidades do estudante. Este sistema de IA baseia-se em dados históricos do estudante para adaptar as experiências de aprendizagem aos seus níveis previstos de conhecimento. Além disso, fornece *feedback* constante ao estudante e informações em tempo real sobre o seu progresso, que são apresentadas num painel de controlo alocado ao professor.

Proceda a uma análise desta situação, abordando questões como:

- *Ação e supervisão humanas*: Como é que o sistema de IA afeta o papel didático do professor?
- *Transparência*: Os professores e os diretores das escolas compreendem como funcionam os algoritmos específicos de avaliação ou de personalização neste sistema de IA?
- *Diversidade, não-discriminação e equidade:* O sistema oferece modos de interação adequados aos alunos com deficiências ou necessidades educativas especiais?
- *Privacidade e governação de dados*: O acesso aos dados dos estudantes está protegido e armazenado num local seguro e os dados são utilizados apenas para os fins para os quais foram recolhidos?
- *Solidez técnica e segurança:* Existe uma estratégia para monitorizar e testar se o sistema de IA está a cumprir os objetivos, finalidades e aplicações pretendidas?
- *Responsabilização:* Quem é responsável pela monitorização contínua dos resultados produzidos pelo sistema de IA e pela forma como os resultados estão a ser utilizados para melhorar o ensino, a aprendizagem e a avaliação?

Caso 2 (real)

No verão de 2020, as medidas de distanciamento social impostas pela pandemia do COVID-19 levaram a que os exames de acesso à universidade (designados "níveis A"), no Reino Unido, não pudessem ser realizados. O governo decidiu proceder a uma estimativa de classificações utilizando um algoritmo de IA. Para esse efeito, foram utilizados três tipos de dados: 1) A distribuição histórica das notas das escolas nos três anos anteriores (2017-2019); 2) A classificação de cada aluno dentro da sua própria escola numa determinada disciplina com base na avaliação de um professor sobre a sua classificação provável se os "níveis A" tivessem decorrido como planeado; 3) Os resultados dos exames anteriores de cada aluno por disciplina.

Deste processo resultou que o algoritmo analisou a distribuição histórica das notas de uma escola e, em seguida, decidiu a nota de um aluno com base nessa classificação. Por exemplo, se um aluno estivesse a meio do *ranking* de classificações, então a sua nota seria aproximadamente igual à que a pessoa na mesma classificação obteve em anos anteriores. Ou então, por exemplo, se ninguém de uma determinada escola tivesse obtido a nota mais alta nos últimos três anos, seria impossível atribuir a alguém dessa escola essa nota. Além disso, o algoritmo deu mais peso ao fator 2 (avaliação projetada pelo professor) se houvesse menos de 15 alunos na turma na qual o estudante estava inserido.

A divulgação dos resultados deu origem a um protesto público. Foi particularmente criticado o efeito díspar que o algoritmo de classificação teve ao fazer com que os estudantes de turmas mais pequenas tivessem mais probabilidades de beneficiar da inflação de notas do que os de turmas maiores. Ou seja, o comportamento do algoritmo em relação a turmas pequenas fez com que as escolas privadas registassem um aumento na proporção de alunos que obtiveram classificações mais elevadas. A crítica pública apontou para a discriminação algorítmica, neste caso refletida numa desvalorização dos resultados dos estudantes que frequentaram escolas públicas e numa inflação de notas para estudantes de escolas privadas, prejudicando assim os alunos de um meio socioeconómico mais baixo.

Caso 3

Empresas que desenvolvem tecnologias de IA frequentemente apresentam as tecnologias de reconhecimento facial – uma tecnologia que identifica e verifica rostos em imagens ou vídeos, analisando características únicas da face humana, como formato dos olhos, nariz e boca – como uma ferramenta útil para as escolas. Apontam como benefícios do uso de reconhecimento facial em escolas os seguintes aspetos:

- *Segurança:* O reconhecimento facial pode ser utilizado para controlar o acesso às instalações escolares, garantindo que apenas pessoas autorizadas entrem; e proporcionar alertas de intrusão, se detetarem a presença de pessoas não autorizadas em áreas restritas.

- *Controlo dos estudantes*: Desde o registo automático de presenças; à monitorização de comportamentos para identificar atividades incomuns ou desviantes na sala de aula, recreio e outros espaços.
- *Identificação de emergências médicas*: Deteção precoce de sintomas médicos, como febre, contribuindo para a saúde geral da comunidade escolar.
- *Controlo de acesso a serviços*: Pode ser utilizado para gerir o acesso a bibliotecas, cantinas e outras instalações, garantindo que apenas alunos autorizados utilizem esses serviços; ou ainda, dispensando a utilização de dinheiro ou de cartões.

Os críticos da utilização de reconhecimento facial em escolas apontam os seguintes riscos (Galligan et al., 2020):

- *Reprodução e ampliação do racismo*: Tal como acontece com outras tecnologias de IA, o reconhecimento facial é apresentado como um sistema objetivo e neutro, mas na prática reflete os preconceitos estruturais e sistémicos das sociedades. À semelhança do que acontece em outros contextos, como atividades de policiamento e aeroportos, é provável que o uso de reconhecimento facial em escolas atinja de modo desproporcional e injusto estudantes de grupos étnicos mais vulneráveis à exclusão e estigmatização sociais. Além disso, vários estudos apontam que o reconhecimento facial é mais sujeito a erros e imprecisão quando se trata de identificar pessoas que não correspondem ao padrão físico do "homem branco".
- *"Naturalização" da vigilância*: A implementação de reconhecimento facial em escolas pode conduzir à incorporação da ideia de que é "normal" e aceitável ser constantemente vigiado, o que pode trazer consequências psicológicas e sociais negativas para os estudantes. Além disso, existem perigos reais de ameaças à privacidade e proteção de dados; além do facto deste tipo de vigilância poder extravasar finalidades de segurança e tornar-se um mecanismo de controlo desproporcionado do comportamento dos estudantes.
- *Imposição de comportamentos*: O reconhecimento facial nas escolas também é suscetível de disciplinar os estudantes de formas inesperadas, estreitando a definição de "estudante aceitável" e punindo aqueles que não se enquadram nessa definição. Assim, a presença desta tecnologia pode reprimir expressões de individualidade, por exemplo, conduzindo os estudantes a alterar o seu comportamento no que diz respeito ao estilo de vestuário e de penteado, para evitarem serem indisciplinados. É ainda provável que o reconhecimento facial apresente taxas de sinalização de "não conformidade" junto de estudantes com deficiência, não binários e não brancos.
- *Uso inadequado dos dados*: Na ausência de legislação rigorosa nesta matéria, existe o risco dos dados recolhidos pelas tecnologias de reconhecimento facial poderem ser usados para treinar algoritmos ou para finalidades comerciais da parte das empresas que recolhem, armazenam e fazem a gestão destes dados.

5. A Inteligência Artificial na saúde humana

Introdução

A disseminação da Inteligência Artificial (IA) no setor da saúde humana tem-se expandido de modo assinalável em diversas especialidades profissionais e campos de atuação, desde a prestação de cuidados e serviços até atividades de gestão e administração de hospitais, clínicas e centros de saúde. As aplicações da IA na saúde podem proporcionar novas oportunidades para melhorar os processos de diagnóstico e tratamento, o desenvolvimento de medicamentos e a vigilância em saúde pública. Podem, ainda, contribuir para a resiliência e a sustentabilidade dos sistemas de saúde e potenciar o acesso universal à saúde (Laï et al., 2020; Organização Mundial de Saúde, 2021).

Porém, a expansão da IA no âmbito da saúde humana requer uma reflexão cuidadosa em torno das suas implicações sociais e éticas, cuja abrangência e complexidade têm convocado abordagens prudentes e cautelosas ao otimismo tecnológico que tende a rodear os discursos de quem desenvolve IA e de decisores políticos (ver capítulo introdutório). Vários autores apelam ao desenvolvimento de estudos prévios à implementação da IA para assegurar que a ética e os direitos humanos estão no coração do desenho, da implementação e dos usos destas tecnologias no campo da saúde (Braun et al., 2021; Fiske et al., 2019; Morley et al., 2020; Murphy et al., 2021). Cabe perguntar, entre outras questões: Em que áreas da saúde é utilizada a IA, e com que objetivos? Que preocupações e desafios têm suscitado as utilizações da IA na saúde? Que valores sociais e princípios éticos têm guiado o debate sobre a governação da IA na saúde, e quem tem liderado essa discussão? Por fim, como conciliar a necessidade de recolha, uso e partilha de dados de saúde das pessoas, essenciais para o desenvolvimento de aplicações de IA na saúde, com princípios de equidade, transparência e justiça social?

Procurando responder a estas questões, neste capítulo exploramos os benefícios esperados das utilizações de IA sobretudo no âmbito dos cuidados e serviços de saúde, na investigação em saúde, e no desenvolvimento de fármacos, mas também as controvérsias e preocupações que estas têm gerado. Terminamos este capítulo com uma reflexão sobre os desafios sociais e éticos suscitados pelas aplicações de tecnologias de IA na saúde, essencial para que a pesquisa, o desenvolvimento e a implementação da IA na saúde possam beneficiar a sociedade como um todo, de forma responsável.

Em síntese, neste capítulo pretendemos:

- Traçar o panorama de algumas aplicações de tecnologias de IA no âmbito dos cuidados e serviços de saúde, da investigação em saúde, e do desenvolvimento de fármacos.
- Compreender os desafios sociais e éticos que a utilização da IA na saúde humana pode convocar.
- Refletir sobre como é que o desenvolvimento e a implementação da IA na saúde podem promover a equidade e a justiça social.

5.1. Panorama de aplicações da Inteligência Artificial no setor da saúde humana

Os usos da IA na esfera da saúde humana podem assumir, em termos gerais, duas ramificações: Virtual e física (Hamet e Tremblay, 2017). A componente virtual inclui um conjunto diversificado de aplicações, desde sistemas de registos de saúde eletrónicos constituídos por dados clínicos e de saúde recolhidos eletronicamente para cada pessoa e produzidos por diversas entidades que prestam cuidados de saúde, ao uso de redes neuronais[34], modelos computacionais inspirados pelo sistema nervoso central de um animal, para apoiar processos de decisão no âmbito de diagnósticos e/ou tratamentos. Já a componente física é bem ilustrada com o caso de robôs que auxiliam na realização de cirurgias e na prestação de cuidados a pessoas com mobilidade limitada ou declínio cognitivo, ou que suportam próteses inteligentes[35] para pessoas portadoras de alguma deficiência, além dos robôs sociais usados com propósitos terapêuticos em situações de demência ou autismo (Fiske et al., 2019; Valles-Peris et al., 2021). Acrescem, ainda, os nanorobôs, dispositivos extremamente pequenos projetados para realizar tarefas a nível molecular ou celular[36], que se têm tornado ferramentas importantes para injetar medicamentos em áreas específicas no corpo humano, como por exemplo a administração de medicamentos diretamente em células cancerígenas.

Reproduzimos, de seguida, algumas utilizações da IA na saúde humana a partir do mapeamento realizado pela Organização Mundial de Saúde em 2021 (Organização Mundial de Saúde, 2021)[37]. Esta proposta resulta de um trabalho de dois anos que envolveu um grupo de especialistas em ética, tecnologia digital e direito, assim como peritos associados a Ministérios da Saúde com o objetivo de produzir um conjunto de recomendações harmonizadas sobre ética e governação da IA na saúde dirigidas a desenvolvedores de IA; ministérios da saúde; e instituições e profissionais de saúde.

Cuidados de saúde

Diagnóstico e previsão de diagnóstico: A IA pode apoiar o diagnóstico de diversas formas, designadamente na radiologia e na imagiologia médica. Alguns exemplos incluem o diagnóstico radiológico em oncologia (entre outros, colonoscopia, mamografia e otimização de doses de radiação), o diagnóstico de retinopatia diabética, e a deteção de tuberculose. A IA pode, ainda, tornar o diagnóstico mais rápido e preciso em casos como o AVC e a pneumonia. A IA também pode ser usada para prever

34 Consultar o glossário para mais informações.

35 A computação aplicada a próteses pretende desenvolver membros artificiais com tecnologias computacionais para pessoas com algum tipo de deficiência motora ou com ausência de algum membro.

36 Consultar o glossário para mais informações.

37 A Organização Mundial de Saúde identificou as tecnologias de IA desenvolvidas e usadas em países de rendimento elevado. Algumas destas tecnologias também eram usadas em países de rendimento médio e baixo, designadamente nas seguintes áreas: Diagnóstico; avaliação do risco de morbilidade e de mortalidade; vigilância e controlo de surtos; e políticas e planeamento em saúde (Organização Mundial da Saúde, 2021, p. 6).

doenças ou eventos importantes de saúde antes que os mesmos aconteçam (por exemplo, prevenir doenças cardiovasculares e a diabetes, entre outras).

Atendimento clínico: A IA pode ser usada para cruzar diversos registos de saúde eletrónicos durante uma consulta, intersetando dados clínicos e de saúde recolhidos eletronicamente para cada pessoa por diversas entidades que prestam cuidados de saúde, contribuindo para identificar pessoas em risco de doença e para auxiliar os profissionais de saúde a tomar decisões sobre o tratamento. Ao potenciar a automatização de algumas tarefas, é esperado que possa reduzir a sobrecarga de trabalho e acarretar benefícios como libertar tempo para que os profissionais de saúde possam ouvir empaticamente as pessoas (Blease et al., 2020) e os médicos se foquem na resolução dos casos mais complexos.

Reconfiguração do papel do paciente/utente: A IA pode transformar a forma como as pessoas gerem as suas condições médicas, em particular no caso de doenças crónicas ou problemas de saúde mental, na medida em que podem auxiliar no autocuidado. Referimos, por exemplo, o eventual apoio que pode resultar da possibilidade de conversar com *chatbots* ou o recurso a instrumentos e tecnologias de monitorização do estado de saúde especificamente concebidos para pessoas com incapacidades.

Mudança dos cuidados hospitalares para os cuidados domiciliários: A IA pode facilitar a transição de cuidados do hospital para o domicílio, já iniciada pela telemedicina. Os sistemas de monitorização remota ilustram este potencial, de que são exemplo os assistentes virtuais de saúde que apoiam e interagem com os pacientes ou os tratamentos observados diretamente por chamada de vídeo. Também o vestuário com sensores incorporados que monitorizam o estado de saúde (por exemplo, ao medir a frequência cardíaca ou a pressão arterial) e informam os utilizadores e profissionais de saúde podem ajudar a otimizar eventuais intervenções imediatas para apoiar a gestão da saúde das pessoas (doentes ou "saudáveis") fora do contexto clínico.

Extensão dos cuidados "clínicos" para além do sistema formal de saúde: Algumas tecnologias de IA podem ser adquiridas e usadas pelas pessoas para monitorizar a sua saúde, acedendo assim a serviços de saúde não mais confinados aos limites dos próprios sistemas de saúde.

Alocação de recursos e priorização: A IA pode auxiliar nos processos de tomada de decisão sobre a melhor forma de estabelecer prioridades e de disponibilizar recursos que são escassos. Os algoritmos podem ser treinados e usados para racionar a disponibilidade de serviços de saúde, por exemplo, identificando as pessoas que devem receber cuidados intensivos ou quando determinadas intervenções devem ser suspensas.

Investigação em saúde e desenvolvimento de medicamentos

Usar dados provenientes de registos de saúde eletrónicos: Desde que devidamente desenhada e treinada com dados de qualidade e apropriados, a IA pode contribuir para

identificar as melhores práticas clínicas e desenvolver orientações para novos modelos de prestação de cuidados.

Genómica: Espera-se que a IA possa auxiliar a medicina genómica e de precisão na complexa análise de grandes conjuntos de dados que reúnem a totalidade do material genético de uma pessoa, contribuindo para melhorar a compreensão da doença ou para identificar novos biomarcadores (isto é, características biológicas mensuráveis) que ajudem a prever, diagnosticar e tratar doenças, bem como a desenvolver fármacos.

Simplificar e acelerar o desenvolvimento de fármacos: A IA pode tornar este processo mais curto e efetivo e menos caro, ao transformar um trabalho intensivo (ou seja, um processo que requer uma grande quantidade de mão-de-obra) num processo intensivo em capital e dados através do uso da robótica e da genómica. Estima-se uma evolução significativa desta área nas próximas décadas, com testes virtuais de medicamentos (sem animais nem humanos) e prescrições personalizadas para cada pessoa.

Gestão e planeamento dos sistemas de saúde

Automatização de tarefas logísticas e repetitivas, como o agendamento de consultas e a gestão dos produtos sanguíneos disponíveis em *stock*, pode contribuir para melhorar a eficiência dos processos e fluxos de trabalho, incluindo atividades de enfermagem e de gestão.

Apoiar processos de decisão e planeamentos complexos. A IA pode contribuir para otimizar a alocação de recursos de saúde por regiões, de acordo com os respetivos desafios de saúde sazonais, ou prever o tempo que os profissionais de saúde devem permanecer junto de comunidades com maiores carências.

Saúde pública e vigilância em saúde pública

Promoção da saúde: A IA pode contribuir para identificar populações ou locais onde predominam comportamentos de "elevado risco" ou a quem deverão ser dirigidas mensagens com conteúdos específicos de saúde. Trata-se de micro-segmentação, ou seja, da divisão da população em grupos, diferenciando-os de acordo com determinados atributos, como os seus estilos de vida.

Prevenção da doença: A IA tem sido usada para auscultar as causas subjacentes a maus resultados de saúde pública, incluindo na saúde ambiental e na saúde ocupacional. Referimos, por exemplo, a identificação de riscos ambientais através de sensores que analisam os níveis de poluição em determinados espaços.

Vigilância e preparação para emergências: A IA adiciona "vestígios" digitais de atividades humanas (como blogues, vídeos e pesquisas na Internet) ao tipo de dados recolhidos para efeitos de vigilância em saúde pública e usa essa evidência para criar modelos matemáticos para a tomada de decisões. Um dos exemplos é a *Flue*

Google Trends, um serviço deste motor de busca que disponibiliza dados em tempo real sobre a propagação da gripe em diversos países com base num modelo estatístico assente nas pesquisas realizadas sobre complicações, remédios, sintomas e medicamentos antivirais para a gripe. Já a recolha de dados em tempo real acerca da movimentação e da localização das pessoas foi usada para construir modelos de IA que previam as dinâmicas de transmissão regional da pandemia da COVID-19 e orientaram o controlo e a vigilância de fronteiras (Whitelaw et al., 2020).

Resposta a surtos: A IA pode ser usada para estudar a transmissão de vírus e para desenvolver potenciais vacinas e tratamentos, contribuindo para melhorar a identificação e gestão de surtos.

Vemos, assim, que a IA pode transformar a prestação de cuidados e os serviços de saúde, a investigação biomédica e a saúde pública, e tornar mais eficientes, sustentáveis e acessíveis os sistemas de saúde. No entanto, será importante equacionar, por exemplo, se a reprodutibilidade e generalização dos algoritmos estão asseguradas, ou se, pelo contrário, só se adequam a determinados grupos sociais. Imaginemos um hospital que implementa um sistema de IA para auxiliar radiologistas na interpretação de exames de mamografia para deteção de cancro de mama. Este sistema é treinado num grande conjunto de dados de pacientes, e alcança uma alta precisão na deteção de tumores em mulheres de uma determinada faixa etária e etnia, que representam a maioria dos pacientes no conjunto de dados de treino. No entanto, ao ser implementado na prática, descobre-se que a precisão do algoritmo é significativamente menor em mulheres mais jovens ou mais velhas, ou em mulheres de etnias minoritárias. Isso ocorre porque o algoritmo foi treinado predominantemente em dados de pacientes de uma faixa etária específica e de uma única etnia, e não pode ser generalizado para grupos demográficos diferentes. Neste caso, a falta de reprodutibilidade e generalização do algoritmo pode resultar em diagnósticos imprecisos e, consequentemente, em tratamentos inadequados ou atrasados para certos grupos de pacientes.

Outra questão central refere-se à necessidade de garantir a qualidade e precisão dos dados usados, de modo a que os modelos de apoio à decisão não originem previsões enviesadas. Imaginemos um sistema de IA desenvolvido para prever quais pacientes têm maior probabilidade de serem readmitidos no hospital após obterem alta clínica. Este sistema é treinado em dados históricos de pacientes, incluindo informações como idade, sexo, condições médicas pré-existentes, readmissões anteriores e detalhes do tratamento recebido. No entanto, se os dados utilizados para treinar o modelo forem enviesados ou incompletos, isso pode levar a previsões igualmente enviesadas. Por exemplo, se o sistema for treinado principalmente em dados de pacientes de uma determinada faixa etária ou de uma única instituição de saúde, as previsões podem não ser precisas para pacientes de outras faixas etárias ou de diferentes instituições. Além disso, se os dados históricos refletirem disparidades no acesso aos cuidados de saúde ou tratamentos diferentes para grupos demográficos específicos, o modelo pode reproduzir e até amplificar essas desigualdades, resultando em previsões enviesadas e injustas.

Neste contexto complexo, é nosso propósito explorar nas próximas secções os principais desafios sociais e éticos que as aplicações da IA na saúde humana têm suscitado.

5.2. Princípios éticos "consensuais"

O Grupo de Peritos designados pela Organização Mundial de Saúde identificou seis princípios fundamentais na prossecução do uso ético da IA na saúde (Organização Mundial de Saúde, 2021, pp. 23-30). Amplamente disseminados como os primeiros princípios "consensuais" que orientam a conceção, desenvolvimento e uso da IA na saúde, são eles:

Proteger a autonomia, de modo a que sejam as pessoas, profissionais ou pacientes, quem controla os sistemas de saúde e as decisões médicas, desenhando e implementando os sistemas de IA para auxiliar as pessoas a tomar decisões informadas, com respeito pela privacidade e confidencialidade dos dados e garantia de consentimento informado e válido.

Promover o bem-estar e a segurança humanas, e o interesse público, sem causar danos (físicos ou mentais, incluindo a estigmatização ou discriminação) a pessoas ou grupos, assegurando a eficácia, segurança, precisão e qualidade das tecnologias de IA.

Garantir a transparência, explicabilidade e inteligibilidade, de forma a que as tecnologias de IA sejam compreendidas por todas as pessoas (desenvolvedores, utilizadores e reguladores), disponibilizando informação regular e atempada quer para consulta e debate públicos sobre o desenho e usos das tecnologias de IA, quer para efeitos de auditoria e de prestação de contas, mesmo antes da implementação de uma tecnologia de IA.

Promover a responsabilização e a prestação de contas, cabendo às partes interessadas assegurar que as tecnologias de IA podem desempenhar determinadas tarefas e são usadas nas circunstâncias apropriadas e por pessoas devidamente treinadas, com supervisão ativa de profissionais de saúde, pacientes e desenvolvedores de sistemas de IA, assim como do público em geral e das autoridades e agências regulatórias, assegurando vias de recurso acessíveis perante a ocorrência de impactos adversos.

Assegurar a inclusão e a equidade, encorajando o acesso amplo às tecnologias de IA e a respetiva utilização, independentemente de características como a idade, género, rendimento, raça, etnia, local de residência ou língua, e a participação ativa de todas as pessoas que poderão ser afetadas pelas tecnologias de IA na respetiva conceção e avaliação, de modo a evitar vieses e discriminação.

Promover uma IA responsiva e sustentável, onde projetistas, desenvolvedores e utilizadores avaliam (de forma contínua, sistemática e transparente) se as tecnologias de IA respondem adequadamente às expectativas e aos requisitos exigidos nos contextos onde são usadas, existindo respostas institucionais face a situações de ineficácia ou insatisfação com o propósito de resolver os problemas encontrados. Importa, ainda,

reunir esforços para promover a sustentabilidade dos sistemas de saúde (ao educar e treinar os profissionais de saúde, por exemplo) e do ambiente (ao minimizar a pegada ecológica e aumentar a eficiência energética).

A noção de princípios éticos "consensuais", como os identificados pelo Grupo de Peritos designados pela Organização Mundial de Saúde para o uso ético da IA na saúde, pode ser vista como uma tentativa de estabelecer diretrizes universais para a aplicação ética destas tecnologias. No entanto, é importante reconhecer que esses princípios podem refletir predominantemente a visão e interesses das elites, que muitas vezes estão envolvidas na formulação de políticas e de diretrizes (ver capítulos 2 e 3).

Ao adotar esses princípios como padrões consensuais, há o risco de que as preocupações e perspetivas de grupos marginalizados ou sub-representados sejam negligenciadas ou minimizadas. Além disso, a natureza abstrata e genérica desses princípios pode não considerar adequadamente as complexidades éticas e contextuais envolvidas na implementação da IA na saúde em diferentes comunidades e contextos socioculturais. Portanto, uma reflexão crítica sobre a noção de princípios éticos consensuais deve questionar até que ponto esses princípios refletem verdadeiramente valores éticos universais ou se são influenciados por interesses particulares e visões hegemónicas. Essa crítica pode abrir espaço para um debate mais amplo e inclusivo sobre ética na IA na saúde humana, garantindo que as vozes de todos os envolvidos sejam consideradas na formulação de diretrizes éticas e políticas.

5.3. Desafios sociais e éticos

Diversas revisões sistemáticas de literatura académica focalizadas na identificação dos desafios sociais e éticos implicados nas aplicações de tecnologias de IA na saúde evidenciam a importância que tende a ser atribuída a princípios relacionados com o respeito da autonomia humana, equidade, explicabilidade, privacidade e prestação de contas, mas alertam para o facto de ser menos frequente a discussão em torno da prevenção de danos (por exemplo, Karimian et al., 2022; Murphy et al., 2021). À semelhança do que acontece noutras áreas (ver capítulo 2), também na saúde se verifica a dominância de um debate geral e abstrato, que revela limitações quanto à consideração de princípios éticos no que diz respeito à conceção e implementação efetiva de tecnologias de IA na saúde, destacando-se a rara consideração de instrumentos práticos que permitam testar e atualizar exigências éticas ao longo do ciclo de vida das tecnologias de IA (Karimian et al., 2022; Morley et al., 2020).

Murphy e colegas (2021) mostram como o debate gira sobretudo em torno das aplicações da IA nos cuidados de saúde, em particular os robôs cuidadores e os processos de diagnóstico, e na medicina de precisão, silenciando em larga medida o debate em torno da saúde pública e da saúde global, designadamente no contexto dos países de baixo e médio rendimento. De facto, as concetualizações de valor público na saúde digital estão maioritariamente relacionadas com aspetos económicos, ou seja, os benefícios e as contribuições do uso de tecnologias de IA na saúde tendem a ser

medidos em termos de mercados, criação de empregos e ganhos financeiros, com uma tendência para marginalizar outras interpretações orientadas para a perspetiva de acrescentar valor para a sociedade, designadamente a saúde pública, a sustentabilidade a longo-prazo ou o bem comum (Gross e Geiger, 2023).

O debate em torno da IA na saúde humana ramifica-se ainda em dois aspetos importantes, interligados, mas ainda assim distintos: Por um lado, questões éticas relacionadas com o uso, desenvolvimento e implementação de tecnologias de IA na saúde ("ética na IA"). Isso inclui preocupações sobre justiça, transparência, responsabilidade e impactos sociais, económicos e políticos das tecnologias de IA. Por outro lado, a "IA ética" quando falamos da ideia de desenvolver IA que seja intrinsecamente ética, ou seja, que incorpore princípios éticos na própria conceção e funcionamento de algoritmos e sistemas de IA, como garantir a privacidade dos dados, evitar discriminação algorítmica e maximizar o benefício para a sociedade (Arbelaez et al., 2024). A relevância destes aspetos torna-se ainda mais crucial perante a necessidade de auscultar a perspetiva de diversos grupos quanto aos desafios sociais e éticos da IA na saúde. Os públicos convidados a pronunciarem-se são, sobretudo, profissionais de saúde e excecionalmente cuidadores ou doentes, estando largamente sub-representados quer grupos vulneráveis, como pessoas com deficiências, quer os líderes em informática da saúde, criadores e/ou implantadores de IA, ou gestores de organizações de saúde (Karimian et al., 2022). Quem tem liderado a discussão em torno de uma IA ética na saúde raramente menciona o envolvimento com utilizadores finais e beneficiários (Murphy et al., 2021).

Sumariamos, de seguida, os principais desafios sociais e éticos dos usos da IA na saúde humana (Tabela 8), categorizados de acordo com os sete requisitos para uma IA de confiança propostos pelo Grupo de Peritos de Alto Nível sobre a IA designados pela Comissão Europeia (desenvolvidos em detalhe no capítulo 2). Exploraremos com maior detalhe alguns destes desafios nas próximas secções.

Tabela 8
Alguns desafios sociais e éticos do uso da Inteligência Artificial na saúde humana.

Requisitos	Desafios sociais e éticos
Ação e supervisão humanas	Falta de centralidade das pessoas (profissionais ou doentes/utentes).
	Potencial para ignorar necessidades ou preferências individuais.
	Ausência de processos de decisão partilhada.
	Efeitos disruptivos na relação entre profissionais de saúde e doentes/utentes.
Solidez técnica e segurança	Ocorrência de erros e funcionamento desadequado das tecnologias de IA.
	Em que medida o desempenho da IA pode ser generalizado e reprodutível em diferentes contextos (socioculturais, económicos, geográficos)?
	Como avaliar os efeitos, diretos e colaterais, das tecnologias de IA?
	Ausência de ensaios clínicos prospetivos para avaliar sistemas de IA na saúde.
	Rápida obsolescência das tecnologias de IA.

Privacidade e governação dos dados	Falta de consentimento para partilhar dados de saúde e ausência de controlo individual dos dados.
	Uso indevido de dados, desconfianças relacionadas com a existência de interesses comerciais nos dados de saúde, e preocupações quanto ao risco dessa informação poder ser usada por entidades bancárias, empregadores, companhias de seguro ou governos.
	Como proteger a privacidade e a confidencialidade dos dados pessoais de saúde, designadamente no âmbito da procura transfronteiriça de cuidados?
	Em que circunstâncias a perda de privacidade é aceitável em prol de um bem maior?
Transparência	Opacidade e escassa interpretabilidade dos algoritmos poderão colocar entraves à comunicação entre profissionais de saúde e doentes/utentes quanto aos benefícios e riscos envolvidos numa decisão baseada em tecnologias de IA.
	Falta de explicações transparentes e contextualizadas.
Diversidade, não discriminação e equidade	Uso de dados limitados, de baixa qualidade e não representativos podem refletir, perpetuar e agravar iniquidades em saúde.
	Modelos poderão originar previsões enviesadas ao basearem-se em fatores como a raça/etnia, idade, género, e tipo de seguro de saúde individual, beneficiando determinados grupos sociais em detrimento de grupos vulneráveis. Como mitigar o viés algorítmico?
	Falta de equidade, diversidade e justiça na disponibilidade de tecnologias de IA.
	Necessidade de considerações éticas específicas para países com escassas infraestruturas ou recursos para informar as pessoas, comunicar incertezas, obter consentimento e gerar dados robustos.
	Excesso de confiança nas tecnologias de IA pode exacerbar desigualdades de acesso a tecnologias médicas entre grupos sociais e entre países.
Bem-estar societal e ambiental	Como planear a implementação de tecnologias de IA tendo em consideração as mudanças socioculturais e clínicas exigidas?
	Como articular formas de governação global perante os enormes interesses económicos?
	Como medir eventuais mudanças no estado de saúde das populações?
	Crescente vigilância massiva das populações, pela recolha desproporcionada de dados, eventualmente usados com finalidades médicas e não médicas (por exemplo no sistema de justiça criminal, como veremos no capítulo 6 deste livro) sem o consentimento explícito das pessoas.
Responsabilização	Dificuldades logísticas na implementação de tecnologias de IA.
	Como (re)distribuir as responsabilidades nos processos de tomada de decisão e na prestação de contas?
	Como devem ser reguladas as tecnologias e aplicações digitais destinadas à autogestão da saúde?
	As responsabilidades que recaem sobre as pessoas quanto ao uso de IA para promover o autocuidado podem ser percecionadas como causadoras de stress adicional e podem limitar o acesso a serviços de saúde formais.

Fonte. Karimian et al., 2022; Murphy et al., 2021; Organização Mundial de Saúde, 2021, pp. 31-64.

5.3.1. Ação e supervisão humanas

Alguns desafios sociais e éticos a considerar prendem-se com os receios quanto à desqualificação e à substituição dos profissionais de saúde por tecnologias de IA e à erosão do julgamento humano em questões de saúde. Acrescem preocupações relacionadas com a perda de autonomia das pessoas: Por exemplo, em sistemas de diagnóstico automatizados, pacientes e profissionais de saúde podem ser menos incentivados a questionar ou entender os processos por trás de um diagnóstico, confiando total e acriticamente na tecnologia.

De realçar, ainda, inquietações suscitadas por uma eventual dependência excessiva das tecnologias de IA, que poderá simplificar demasiado os processos de decisão em saúde (Blease et al., 2020), potencialmente comprometendo a qualidade do cuidado. Por exemplo, sistemas de IA que prescrevem medicamentos com base em algoritmos podem não levar em consideração todos os aspetos do histórico médico do paciente, interações medicamentosas ou preferências individuais. Esta dependência excessiva das tecnologias de IA pode ainda significar a ausência de decisões partilhadas entre médicos e pacientes. Por exemplo, um sistema de IA que recomende um determinado tratamento sem existir uma mediação humana para explicar o enquadramento, os riscos e os benefícios, pode resultar numa desconexão entre o paciente e o processo de tratamento, levando a uma menor adesão ao mesmo e a resultados menos satisfatórios.

Como assegurar que caberá às pessoas a última palavra, e que as suas necessidades e preferências individuais de saúde não são ignoradas? Algumas propostas alertam para a importância de garantir que sejam as pessoas a decidir quando, onde e se devem usar determinadas tecnologias de IA (Valles-Peris et al., 2021) e em que circunstâncias máquinas e pessoas poderão complementar-se e enriquecer-se mutuamente (Blease et al., 2020; Cresswell et al., 2018). Algumas das circunstâncias de complementaridade e enriquecimento mútuo incluem o uso de *chatbots* e assistentes virtuais alimentados por IA para fornecer informações básicas sobre condições médicas, responder a perguntas comuns dos pacientes e ajudar na triagem inicial de sintomas. Os profissionais de saúde podem então intervir quando necessário para fornecer orientação mais personalizada, esclarecer dúvidas específicas e oferecer suporte emocional aos pacientes. Outra situação seria ao nível do diagnóstico médico: As máquinas podem analisar grandes conjuntos de dados médicos, incluindo o histórico do paciente, exames de imagem e resultados de testes laboratoriais, para identificar padrões e sugerir diagnósticos. Os médicos e profissionais de saúde podem usar a sua experiência clínica e intuição para interpretar os resultados fornecidos pela IA, considerar o contexto do paciente e tomar decisões informadas sobre o diagnóstico e o tratamento.

5.3.2. Solidez técnica e segurança

Uma preocupação central na área da saúde prende-se com a possível ocorrência de erros ou o funcionamento desadequado das tecnologias de IA e a sua rápida obsolescência. Cabe então perguntar: Como garantir a usabilidade e a confiabilidade nas tecnologias de IA na saúde? As tecnologias de IA na saúde devem ser sujeitas a auditorias e monitorização contínuas para garantir a sua confiabilidade ao longo do tempo, devem ser regularmente avaliadas quanto à sua precisão e desempenho, com ajustes feitos conforme necessário. Uma dimensão a considerar será a avaliação adequada dos riscos, não só em termos técnicos, mas também sociais e éticos. Um exemplo concreto seria realizar uma análise de impacto social e ético antes de implementar um sistema de IA para triagem de pacientes, considerando questões como equidade no acesso ao cuidado e considerar que algoritmos de triagem podem inadvertidamente introduzir preconceitos e discriminação.

Outra necessidade premente é, por exemplo, a avaliação dos efeitos, diretos e colaterais, que as aplicações de tecnologias de IA podem ter na saúde humana. Referimos, em particular, as implicações sociais e éticas da utilização de tecnologias de IA cujo desempenho não pode ser generalizado a diferentes contextos socioculturais, económicos e geográficos (aspeto que será trabalho na secção 5.3.5.). Um outro problema social e ético suscitado pelas tecnologias de IA na saúde prende-se com a disponibilização de diagnósticos quando a disponibilidade de opções de tratamento é escassa (Organização Mundial de Saúde, 2021). Ou seja, se a IA é capaz de diagnosticar uma condição médica para a qual não há opções de tratamento acessíveis ou eficazes, isso pode agravar desigualdades no acesso à saúde. Pacientes diagnosticados com uma condição para a qual não podem receber tratamento adequado podem sentir-se injustiçados e os profissionais de saúde e os sistemas de saúde podem enfrentar dilemas éticos em relação à divulgação de diagnósticos quando as opções de tratamento são escassas.

5.3.3. Privacidade e governação dos dados

A IA na saúde humana depende da disponibilidade de grandes conjuntos de dados de saúde para treinar algoritmos e modelos de análise de dados. Esses dados podem incluir informações médicas sensíveis, como histórico de doenças, resultados de exames, procedimentos cirúrgicos e registos de medicamentos. Neste contexto, a partilha e/ou o uso indevido de dados relativos à saúde são as principais preocupações no que respeita a privacidade e governação dos dados. Por exemplo, diversos estudos têm mostrado que as pessoas desconfiam da existência de interesses comerciais nos dados de saúde usados no âmbito de tecnologias de IA (Hallowell et al., 2022; McCradden et al., 2020b; Rogers et al., 2021) e revelam inquietações quanto ao risco dessa informação poder ser usada por entidades bancárias, empregadores, companhias de seguro ou governos (Amann et al., 2023). Os impactos da violação de dados e da perda de privacidade e confidencialidade podem afetar mais os grupos sociais vulneráveis afetados pela IA, como as pessoas idosas (Wang et al., 2019) ou com problemas de saúde mental (Blease et al., 2020).

No contexto de uma análise das implicações sociais e éticas da IA em saúde, reforça-se o apelo a métodos seguros de armazenamento e proteção de dados pessoais sensíveis (McCradden et al., 2020b). A IA suscita questões novas e controversas em relação à governação de dados de saúde: Por exemplo, as circunstâncias em que os robôs podem armazenar e processar os dados de saúde que recolhem (De Graaf et al., 2022) ou serem menos intrusivos na informação que enviam a profissionais de saúde sobre os níveis de adesão aos tratamentos dos seus utilizadores (Jenkins e Draper, 2015). À medida que os robôs se tornam mais comuns em contextos de assistência médica, como em terapia assistida por robôs ou cuidados domiciliários, surge a questão de como esses dispositivos devem lidar com os dados sensíveis dos pacientes. Por um lado, os robôs podem ser projetados para armazenar e processar dados de saúde como parte da sua função assistencial. Por exemplo, um robô de assistência domiciliária pode recolher informações sobre os padrões de sono de um paciente, a sua atividade física e outros indicadores de saúde. No entanto, isso levanta preocupações sobre a segurança e privacidade desses dados, bem como sobre quem tem acesso a eles e como são utilizados. Além disso, há debates sobre como os robôs devem comunicar informações de saúde aos profissionais de saúde. Por exemplo, um robô pode monitorizar a adesão de um paciente ao tratamento e enviar relatórios aos médicos sobre os níveis de adesão. No entanto, há questões éticas sobre a quantidade de informação que os robôs devem compartilhar, especialmente quando se trata de informações sensíveis sobre o comportamento do paciente. Essas questões destacam a necessidade de desenvolver diretrizes claras e políticas robustas para governar a recolha, armazenamento e uso de dados de saúde por parte dos robôs e outras tecnologias de IA na saúde.

A procura de um equilíbrio entre estes valores fundamentais (confidencialidade, privacidade e saúde) afigura-se um importante desafio social e ético no contexto da utilização da IA em saúde (Laï et al., 2020). A procura deste equilíbrio tem sido expressada através do debate sobre dois aspetos complementares: Primeiro, a ideia de que os titulares dos dados devem ser completamente informados acerca de como os seus dados serão usados e ter a opção de prestar o seu consentimento informado para esse efeito, exercendo controlo sobre como é que a sua própria informação pessoal é usada (Isbanner e O'Shaughnessy, 2022). Segundo, o apelo ao desenvolvimento de regulamentação, cuja importância é particularmente relevante num contexto em que os desenvolvedores de IA tendem a protelar esse processo, alegando a complexidade, a lentidão e os obstáculos criados pela regulação ao avanço esperado do desenvolvimento tecnológico (Duke, 2022).

5.3.4. Transparência

A opacidade e a escassa interpretabilidade dos algoritmos de tecnologias de IA podem colocar entraves importantes à comunicação entre profissionais de saúde e utentes quanto aos benefícios e riscos envolvidos numa decisão baseada em IA (Choung et al., 2023; Duke, 2022; Wang et al., 2019). Quando os algoritmos de IA são percebidos como "caixas negras" (ver capítulo 3), ou seja, quando os processos pelos

quais chegam a uma decisão não são claros ou interpretáveis, isso pode prejudicar a confiança e a compreensão dos pacientes e dos próprios profissionais de saúde.

Para superar esse obstáculo na área da saúde, é necessário desmistificar e tornar os processos de IA mais transparentes e interpretáveis. A desmistificação da IA poderá passar, no caso da saúde, pela explicação clara e compreensível dos seguintes aspetos: Primeiro, o que é e o que não é IA. Quando a IA estiver a ser usada num contexto de saúde, todas as pessoas envolvidas nesse processo deverão estar conscientes de que os seus dados estão a ser recolhidos e usados e como é que isso está a ser feito. Segundo, como é que cada aplicação da IA funciona (implicações, riscos e benefícios). Esta explicação poderá incluir informação sobre a composição do conjunto de dados usados para treino, como operam os algoritmos, e como é que a IA toma uma decisão.

5.3.5. Diversidade, não discriminação e equidade

A conceção, desenvolvimento e implementação de tecnologias de IA na saúde que espelham e reificam desigualdades sociais através do uso de dados limitados e não representativos e de vieses algorítmicos que podem beneficiar mais uns grupos sociais do que outros constitui um dos principais desafios sociais e éticos. Referimos as múltiplas consequências negativas que daí podem advir, como a perpetuação e o agravamento de iniquidades em saúde com base na etnicidade, idade, género, estatuto socioeconómico ou condição de saúde. Isto significa a existência de diferenças evitáveis e injustas no estado de saúde ou na distribuição dos recursos de saúde entre diferentes grupos sociais e países, que resultam das condições nas quais as pessoas nascem, crescem, vivem, trabalham e envelhecem.

Um aspeto central a considerar é a transferibilidade e eficácia das tecnologias de IA em todos os grupos sociais (Rogers et al., 2021), o que nem sempre acontece. Referimos um exemplo frequentemente citado na literatura (Obermeyer et al., 2019): A utilização de um sistema de IA para alocar cuidados em diversos serviços de saúde nos Estados Unidos, que disponibilizou mais cuidados a pacientes caucasianos do que a pacientes negros quando as necessidades eram mais elevadas entre os pacientes negros. Isto terá acontecido porque a IA foi desenvolvida a partir de dados relativos à subutilização histórica de serviços de saúde por parte de pacientes negros, assumindo que isso significaria menos necessidades de cuidados de saúde. Para assegurar a transferibilidade e eficácia das tecnologias de IA em todos os grupos sociais, é fundamental incorporar a diversidade, a inclusão e a pluralidade cultural como valores centrais nas aplicações da IA na saúde. Só assim é que a IA poderá contribuir para reduzir iniquidades em saúde e para compensar preconceitos sobre determinadas pessoas ou grupos sociais.

Realçamos, ainda, a importância de apoiar o acesso de todas as pessoas, em todos os países, às tecnologias de IA na saúde, o que convoca um redireccionamento no debate atual em direção à consideração de questões sociais e éticas contextualizadas e específicas para casos onde as infraestruturas tecnológicas e de saúde e/ou os recursos para informar as pessoas, comunicar incertezas e obter consentimento são

escassos. Por exemplo, em alguns países de baixo e médio rendimento, a utilização de tecnologias de IA na saúde poderá exigir investimentos significativos que desencorajarão a sua implementação, desde o investimento em infraestruturas relacionadas com tecnologias de informação e comunicação à necessidade de recolher dados para treino.

Cabe, porém, perguntar: Mas o que fazer para assegurar a inclusão e a equidade, e mitigar o viés algorítmico? Aquino e colegas (2023), por exemplo, procuraram conhecer algumas das estratégias acionadas por especialistas em IA e/ou clínicos para abordar estes desafios: Desde a divulgação pública das limitações ao envolvimento de utentes, representação de grupos marginalizados e considerar a equidade nos métodos de amostragem. O debate permanece em aberto: Quem deve ser responsável por lidar com os vieses da IA na saúde (desenvolvedores, profissionais de saúde, produtores e vendedores, políticos, reguladores, cientistas em IA, e utentes) e como fazê-lo?

5.3.6. Bem-estar societal e ambiental

Os impactos que a utilização de tecnologias de IA no setor da saúde pode ter no trabalho e emprego, assim como nas relações socioprofissionais, emergem como as preocupações principais no que respeita o bem-estar societal e ambiental. Referimos, por exemplo, os seguintes receios: Como lidar com eventuais mudanças suscitadas pela IA nas práticas socioprofissionais, com implicações muitas vezes ainda desconhecidas e incertas? Como enfrentar riscos de potencial desqualificação profissional pelo facto da IA poder ser desenvolvida para realizar tarefas específicas que anteriormente eram realizadas por profissionais de saúde, como por exemplo interpretar exames ou diagnosticar certas condições médicas? Como saber se as tecnologias de IA são eficientes na prestação de cuidados de saúde? Um aspeto importante a considerar são as discrepâncias entre os discursos altamente otimistas sobre as tecnologias de IA na saúde e a sua utilidade efetiva nos contextos reais e nas práticas em que são usadas (Laï et al., 2020; Organização Mundial de Saúde, 2021).

Acrescem inquietações quanto à possibilidade de que haja perda de empatia, humanidade e/ou sensibilidade nas relações socioprofissionais, o que poderá dificultar os processos de comunicação entre pacientes, cuidadores e profissionais de saúde. Por exemplo, a entrada de um robô cuidador na vida de uma pessoa idosa para auxiliar na prestação de cuidados de saúde domiciliários pode criar tensões entre a própria pessoa, os seus cuidadores formais (profissionais de saúde) e os seus cuidadores informais (familiares ou amigos, por exemplo), que podem sentir-se vigiados ou monitorizados e sem espaço para explicar eventuais decisões sobre os tratamentos (Jenkins e Draper, 2015), lamentando a eventual perda de contacto humano.

5.3.7. Responsabilização

A questão da responsabilização tende a ser percecionada como uma questão social e ética importante na utilização de tecnologias de IA na saúde por parte de diversos

públicos. Um desafio a considerar prende-se com a diluição de responsabilidades nos processos de tomada de decisão e na prestação de contas. Imaginemos, por exemplo, que um sistema de IA recomenda um determinado fármaco e uma certa dose para o paciente A, e da sua administração resulta um efeito adverso muito grave que obriga à hospitalização do paciente. A quem caberá a responsabilidade por esta situação: A quem desenvolveu o sistema de IA, potencialmente com falhas, erros ou vieses; ao médico que seguiu a recomendação da IA; a ambos; ou a nenhum deles? Esta questão é particularmente relevante quando diversos estudos alertam para a tendência de deslocar as responsabilidades relacionadas com o desempenho das tecnologias de IA para fora da esfera da competência das indústrias tecnológicas que desenvolvem e comercializam as tecnologias de IA, recaindo as responsabilidades especialmente sobre os utilizadores finais (Laï et al., 2020; Nichol et al., 2023).

Estes desafios abrem espaço para a individualização das responsabilidades, o que significa passar para as mãos de cada utilizador a responsabilidade do que possa acontecer no decurso de uma eventual utilização de tecnologias de IA. Neste enquadramento, alega-se que os pacientes têm a possibilidade de escolher e de tomar decisões sobre as tecnologias de IA, o que os motivará a usar os seus próprios recursos (por exemplo, o tempo e o conhecimento) para promover o autocuidado e a autogestão da sua saúde através do uso de IA. Pensemos, por exemplo, nas pessoas a quem é recomendado o uso de camisolas com sensores incorporados para medir a frequência cardíaca ou a pressão arterial, que informam de imediato os profissionais de saúde para que estes possam intervir quando necessário. Ora, há determinadas circunstâncias que podem limitar ou mesmo impedir que algumas pessoas usem estas camisolas, como a obrigatoriedade de usar determinado vestuário no exercício da sua atividade profissional. Também o facto de alguns pacientes poderem sentir-se responsáveis, individualmente, por usar estas camisolas ininterruptamente pode causar-lhes fadiga e *stress* adicional perante algum esquecimento ou quando, por algum motivo, precisam de tirar a camisola. Por fim, a sensação de monitorização contínua provocada pelo uso das camisolas com sensores pode fazer com que os utilizadores se sintam mais confiantes com o seu estado de saúde e, portanto, menos propensos a visitar um médico regularmente para exames de rotina. Isso pode resultar numa diminuição da vigilância sobre possíveis problemas de saúde que não são detetáveis apenas pelo sistema de monitorização. Como consequência, as políticas de saúde podem ser ajustadas de forma a reduzir a oferta de serviços, alegando uma diminuição na procura por cuidados médicos, o que, por sua vez, pode ter o impacto negativo de limitar o acesso da população aos serviços de saúde.

5.4. Atividades para debate

Apresentam-se exemplos de casos para debate à luz das questões sociais e éticas analisadas neste capítulo e desenvolvidas em termos gerais e abstratos com mais detalhe no capítulo 2. Considere ainda as diferentes dimensões e os diversos níveis de análise apresentados no capítulo 3, selecionando aqueles que lhe pareçam mais adequados à análise de cada caso e justificando porquê.

Caso 1

Imagine que a IA está a ser usada num hospital para produzir recomendações quanto à medicação e à dosagem adequadas aos pacientes. A IA recomenda um determinado fármaco e uma certa dose para o paciente A. Porém, o médico não compreende como é que IA chegou a esta recomendação. O algoritmo usado pela IA é muito sofisticado e impossível de entender pelo médico. O médico deve seguir a recomendação da IA? Se o paciente descobrir que a prescrição foi recomendada pela IA, mas ninguém o tinha informado sobre isso, o que sentirá o paciente? O médico terá o dever (moral e/ou legal) de informar o paciente que recorreu a uma tecnologia de IA? (adaptado de Organização Mundial de Saúde, 2021, p. 48)

Reflita sobre estas questões, ponderando os desafios sociais e éticos que poderão surgir no âmbito dos seguintes requisitos: Ação e supervisão humanas; transparência; e responsabilização.

Caso 2

Devido a constrangimentos financeiros, um hospital público pretende facultar a uma empresa privada acesso a dados dos pacientes (exames, comportamentos e historial médico), em troca da implementação de um sistema de IA que melhore substancialmente a capacidade dos médicos diagnosticarem doenças graves, com rapidez e segurança. O algoritmo só será bem-sucedido se os dados forem abundantes e transferíveis. Esta exigência dificulta o conhecimento antecipado sobre a forma como os dados serão usados. Além disso, torna-se difícil garantir a privacidade e assegurar o consentimento dos pacientes. (Adaptado de Whittlestone et al., 2019, p. 22)

Proceda a uma análise desta situação, abordando questões como:

- *Solidez técnica e segurança:* Em que medida o desempenho da IA pode ser generalizado e reprodutível em diferentes contextos (socioculturais, económicos, geográficos)?
- *Privacidade e governação dos dados:* Em que circunstâncias a perda de privacidade é aceitável?
- *Diversidade, não discriminação e equidade:* Que abordagens podem mitigar um eventual viés algorítmico?

6. A Inteligência Artificial no sistema de justiça

Introdução

O sistema de justiça visa garantir a ordem social, proteger os direitos individuais e resolver disputas de maneira justa e equitativa. Em que medida a introdução da Inteligência Artificial (IA) no sistema de justiça pode ajudar – ou, pelo contrário, limitar ou constranger– as finalidades do sistema de justiça?

O presente capítulo procurará responder a esta questão. Se, por um lado, se argumenta que a IA pode melhorar o acesso à justiça (por exemplo, por via de plataformas digitais ou com a ajuda de assistentes virtuais) e contribuir para a diminuição dos custos processuais (por exemplo, automatizando algumas tarefas administrativas); por outro lado, tem-se refletido sobre como o uso da IA pode reproduzir ou aumentar a discriminação institucional e a injustiça estrutural. Nas palavras de Rafanelli (2022), é imperativo ter em consideração que os usos da IA correspondem a formas de poder, o que suscita questões de (in)justiça:

> Delegar tarefas na IA é por vezes descrita como um modo de retirar poder às mãos humanas. Esta ideia é um erro sério (…) seríamos negligentes se não víssemos a utilização da IA para aplicação da lei, vigilância ou armas autónomas como uma forma de algumas pessoas exercerem poder sobre outras. Nestes casos, o poder humano opera através de programas de computador, mas são programas escritos por humanos, treinados com dados criados por humanos e postos a funcionar por alguns humanos para monitorizar, regular, controlar e exterminar outros. (…) A IA é uma ferramenta com a qual os humanos exercem poder, e não um substituto do poder humano, pelo que a sua utilização levanta questões de justiça. É nossa responsabilidade, enquanto consumidores, programadores e investigadores, garantir que estas questões não ficam sem resposta. (Rafanelli, 2022, pp. 5-6)

Começaremos, na próxima secção, por facultar exemplos de algumas maneiras pelas quais a IA pode ser aplicada no sistema de justiça, em particular no setor judicial (tribunais e processo judicial), nas forças policiais, nas prisões e na segurança pública, controlo de fronteiras e cooperação internacional na justiça criminal. Nas secções seguintes abordamos os desafios sociais e éticos específicos suscitados pela presença da IA em cada um destes diferentes setores do sistema de justiça, convidando os leitores a refletir sobre os modos pelos quais as tecnologias de IA podem ser utilizadas de forma justa, responsável e transparente e para o bem coletivo. Após uma análise crítica em torno da aplicação dos princípios éticos propostos pela Comissão Europeia para a Eficiência na Justiça (CEPEJ, 2018) – respeito pelos direitos fundamentais; não discriminação; qualidade e segurança; transparência, imparcialidade e equidade; e "controlo do utilizador" – finalizamos este capítulo com propostas de atividades e debates a partir de casos concretos.

Em síntese, neste capítulo pretendemos:

- Dar a conhecer algumas aplicações da IA no sistema de justiça.

- Discutir os desafios sociais e éticos que a utilização de tecnologias de IA suscita em diferentes setores do sistema de justiça.
- Refletir sobre como utilizar as tecnologias de IA de forma justa, responsável e transparente e para o bem coletivo em cada um dos setores do sistema de justiça.

6.1. Panorama de aplicações da Inteligência Artificial no sistema de justiça

A IA tem o potencial de transformar diversos aspetos do sistema de justiça, englobando o setor judicial (tribunais e processo judicial), as forças policiais, as prisões e o contexto de atuação de agências de segurança pública, controlo de fronteiras e cooperação internacional na justiça criminal, entre outras instituições que lidam com a administração da justiça numa sociedade, conforme descrevemos de seguida.

Setor judicial

No sistema judicial, ou seja, na atividade dos tribunais e em processos judiciais, algumas das principais aplicações de IA são as seguintes (de Oliveira et al., 2022; Nowotko, 2021):

Análise de documentos legais: A IA pode ser utilizada para analisar grandes volumes de documentos legais e de jurisprudência, identificando padrões, tendências e informações relevantes.

Assistência jurídica virtual: Assistentes virtuais podem oferecer informações legais básicas, orientação inicial e responder a perguntas comuns.

Triagem de casos: Sistemas baseados em IA podem ajudar na triagem de casos, avaliando a relevância e a complexidade de cada processo, podendo facultar informação pertinente para a tomada de decisão sobre os recursos a alocar a cada caso.

Decisão judicial: Algoritmos podem analisar dados históricos para prever possíveis decisões judiciais.

Mediação online: Plataformas de resolução de disputas *online*, com a ajuda da IA, podem facilitar a mediação entre partes, oferecendo soluções alternativas para litígios.

Automatização de tarefas repetitivas, como classificação e organização de documentos.

Forças policiais

A IA também pode ser usada nas atividades policiais. Referimos, por exemplo, as seguintes utilizações (Berk, 2021; Neiva et al., 2023):

Previsão de crimes: A IA pode analisar grandes conjuntos de dados para identificar padrões e tendências, ajudando as forças policiais a prever e prevenir crimes.

Análise de vídeos de vigilância: Algoritmos de visão computacional podem analisar vídeos de vigilância para identificar atividades suspeitas.

Reconhecimento facial: Algoritmos e técnicas computacionais que identificam e autenticam características faciais humanas (padrões específicos, como a distância e tamanho dos olhos, contornos e proporções do rosto) podem ajudar na identificação rápida de suspeitos em multidões ou ajudar a localizar pessoas desaparecidas.

Prevenção de crimes cibernéticos: A IA pode analisar padrões de tráfego na Internet para identificar possíveis ataques cibernéticos ou atividades suspeitas, ou outro tipo de atividades criminosas *online*.

Gestão de recursos: Algoritmos podem otimizar as rondas de veículos policiais para responder rapidamente a incidentes ou para prever áreas de alta probabilidade de crimes, podendo ainda analisar dados operacionais para melhorar a eficiência e a alocação de recursos.

Simulações virtuais: A IA pode ser usada em simulações virtuais para treinar profissionais de polícia para situações de alto risco.

Prisões

A IA pode, ainda, ser aplicada de várias maneiras em prisões para melhorar a eficiência operacional, a segurança e a gestão de recursos. Algumas possíveis aplicações são as seguintes (Puolakla e Van De Steene, 2021; Završnik, 2020):

Sistemas de vigilância inteligente: Câmaras equipadas com tecnologia de visão computacional podem monitorizar e analisar padrões de movimento para identificar atividades suspeitas e comportamentos de risco e alertar os funcionários sobre potenciais problemas.

Avaliação de riscos: Algoritmos podem ser usados para avaliar o risco de comportamento violento ou tentativas de fuga por parte dos reclusos, por exemplo, analisando dados históricos para identificar padrões que indiquem a possibilidade de incidentes.

Monitorização de saúde: Dispositivos em rede podem ser usados para monitorizar o estado de saúde de reclusos, detetando sinais precoces de emergências médicas.

Previsão de necessidades: Algoritmos de análise preditiva podem prever as necessidades operacionais, ajudando na alocação eficiente de recursos, como pessoal, alimentos e suplementos médicos.

Manutenção preditiva: A IA pode ser usada para prever falhas em infraestruturas e equipamentos.

Reabilitação e educação: Sistemas baseados em IA podem adaptar programas de reabilitação e educação de acordo com as necessidades individuais dos reclusos. Podem

ainda monitorizar o progresso dos reclusos em programas de reabilitação, adaptando as abordagens conforme necessário.

Simulações virtuais: A IA pode ser usada em simulações para treinar funcionários para situações de emergência, melhorando a prontidão e a eficácia.

Segurança pública, controlo de fronteiras e cooperação internacional na justiça criminal

A IA tem sido cada vez mais utilizada em diversas situações relacionadas com a segurança pública, o controlo de fronteiras e a cooperação internacional na justiça criminal. Aqui estão alguns exemplos (eu-LISA e EUROJUST, 2022):

Análise de vídeo em tempo real: Sistemas de câmaras equipadas com IA podem analisar vídeos em tempo real para detetar atividades suspeitas ou atividades fora do padrão na movimentação de multidões.

Reconhecimento facial: Sistemas computacionais de análise da face humana podem ser usados para identificar suspeitos no seio de multidões, aeroportos ou outros pontos de acesso.

Monitorização de redes sociais em busca de atividades suspeitas ou ameaças à segurança, com o objetivo de antecipar eventos e responder a potenciais crises.

Controlo de fronteiras: Verificação automática de documentos, deteção de falsificações e identificação de pessoas suspeitas.

Análise de dados de satélite: A análise de imagens de satélite com técnicas de IA pode ser usada para monitorizar áreas de fronteira, identificando padrões de movimentação que podem indiciar atividades ilegais. Podem ainda servir para analisar zonas de guerra.

Gestão de tráfego e segurança viária: A IA pode ser usada para monitorizar o tráfego em tempo real, identificando padrões que podem indicar acidentes ou atividades suspeitas.

Sistemas de alerta antecipado: Utilizando análise de dados em larga escala, a IA pode fornecer sistemas de alerta antecipado para eventos como desastres naturais, permitindo uma evacuação mais eficiente e uma resposta mais rápida.

Vemos, assim, que a IA pode tornar mais eficiente e acessível o sistema de justiça. Porém, é também necessário atender a que a utilização deste tipo de tecnologia pode criar riscos adicionais para os direitos humanos e contribuir para reforçar processos discriminatórios e desigualdades sociais, conforme é referido nos três primeiros capítulos deste livro. Por isso, é crucial refletir sobre os modos pelos quais as tecnologias de IA podem ser utilizadas de forma justa, responsável e transparente e para o bem coletivo em cada um dos diferentes setores do sistema de justiça. É este o propósito que guia a escrita das próximas secções.

6.2. A Inteligência Artificial no sistema judicial

Tem-se verificado uma atenção crescente à presença da utilização de IA no sistema judicial. É possível distinguir duas perspetivas principais a nível do debate académico sobre este fenómeno (de Oliveira et al., 2022): Por um lado, os estudos que visam compreender as possíveis aplicações da IA na resolução de litígios e no trabalho jurídico e explorar, concretamente, as potencialidades para tornar os tribunais mais eficientes e céleres. Por outro lado, os estudos que desenvolvem uma reflexão crítica, ponderando como é que o uso de IA nos tribunais pode implicar situações que levam à discriminação e à injustiça.

A Comissão Europeia para a Eficiência na Justiça (CEPEJ), criada pelo Comité de Ministros do Conselho Europeu em 2002, com o intuito de estabelecer um organismo inovador para melhorar a qualidade e eficiência dos sistemas judiciais europeus e reforçar a confiança dos utilizadores nesses sistemas, é um dos organismos que tem promovido o debate sobre os benefícios e riscos da introdução de IA nos sistemas judiciais. Em 2018, a CEPEJ publicou um documento intitulado "Carta Ética Europeia sobre a Utilização da Inteligência Artificial nos Sistemas Judiciais e no seu Ambiente", com o seguinte objetivo geral:

> A utilização de tais instrumentos e serviços nos sistemas judiciais tem por objetivo melhorar a eficácia e a qualidade da justiça e deve ser incentivada. Deve, no entanto, ser levada a cabo de forma responsável, com o devido respeito pelos direitos fundamentais das pessoas, tal como estabelecidos na Convenção Europeia dos Direitos do Homem e na Convenção sobre a Proteção dos Dados Pessoais, e em conformidade com outros princípios fundamentais (…), que deverão orientar a definição das políticas públicas de justiça neste domínio. (CEPEJ, 2018, p. 5)

Este documento começa por diferenciar entre o uso de IA no domínio da decisão judicial no campo do direito civil, comercial ou administrativo e o seu uso no domínio de matérias criminais (este último considerado mais sensível):

> O tratamento das decisões judiciais pela inteligência artificial, segundo os seus criadores, é suscetível, em matéria civil, comercial e administrativa, de contribuir para melhorar a previsibilidade da aplicação do direito e a coerência das decisões judiciais, desde que sejam respeitados os princípios a seguir enunciados. Em matéria penal, a sua utilização deve ser considerada com as maiores reservas, a fim de evitar discriminações baseadas em dados sensíveis, em conformidade com as garantias de um processo equitativo. (CEPEJ, 2018, p. 5)

Os estudos académicos sobre a IA no setor dos tribunais têm-se debruçado sobre várias temáticas sobretudo orientadas para a exploração das potencialidades destas tecnologias. Neste contexto, os temas mais comuns são as implicações da IA ao nível do trabalho jurídico, podendo trazer maior celeridade (Alarie et al., 2018; de Sousa

et al., 2022); diminuição de custos e diminuição de erros humanos em processos judiciais (Chalkidis et al., 2020; Xiao et al., 2021).

Os estudos que apelam a uma abordagem mais reflexiva e cautelosa sobre as implicações da IA nos tribunais salientam que a expansão de aplicações de IA neste contexto requer uma profunda ponderação sobre um possível reforço de desigualdades, vieses e injustiças. Infelizmente, verifica-se que são ainda escassos os estudos académicos nesta vertente crítica. Assim, consideramos imprescindível convocar para esta matéria alguns estudos das ciências sociais sobre tribunais que, embora não tenham abordado especificamente a IA neste setor, apresentam resultados sólidos ao nível das implicações da ação dos tribunais no plano da cidadania e da democracia.

Há já várias décadas que estudos das ciências sociais apontam que a ação dos tribunais contribui para perpetuar e reforçar desigualdades sociais (Gomes et al., 2013; Santos, 1995; Tantikul, 2024). Algumas das formas pelas quais isso pode acontecer incluem vieses raciais e étnicos que resultam, por exemplo, de disparidades significativas nas sentenças que afetam os grupos historicamente mais criminalizados (minoritas étnicas, estrangeiros, migrantes); acesso desigual à justiça, pelo qual os grupos socioeconómicos mais vulneráveis que não podem pagar por representação legal adequada podem enfrentar desvantagens substanciais no sistema judicial; discriminação de género, que se pode manifestar nas decisões judiciais, como sentenças mais severas para mulheres em comparação com homens por crimes semelhantes; o recurso excessivo à prisão preventiva, especialmente em casos relacionados com indivíduos em situação de precariedade económica, que pode levar ao aprofundamento de situação de vulnerabilidade económica e familiar; leis que têm impactos desproporcionais em certos grupos podem também contribuir para a desigualdade (por exemplo, leis de drogas que resultam em sentenças mais longas para determinadas comunidades desfavorecidas). Cabe então perguntar: Será que a introdução de tecnologias de IA nos tribunais vai contribuir para reforçar as desigualdades sociais ou pode ajudar a prevenir e mitigar esse tipo de efeito?

Para ajudar a responder a esta questão iremos analisar um dos temas mais debatidos a propósito da IA nos tribunais: Os potenciais benefícios e riscos da implementação de modelos de "justiça preditiva".

6.2.1. A justiça preditiva

A justiça preditiva consiste na análise de grandes quantidades de decisões judiciais por tecnologias de IA, a fim de fazer previsões sobre o resultado de certos tipos de litígios especializados (por exemplo, indemnizações por despedimento ou pensões alimentares). As correlações consideradas relevantes permitem criar modelos que, quando aplicados a novos casos, produzem uma previsão da decisão (por exemplo, o montante da indemnização a pagar a uma vítima) (CEPEJ, 2018, p. 75).

A chamada justiça preditiva também se propõe estabelecer as probabilidades de sucesso (ou insucesso) de um processo num tribunal. Estas probabilidades são calculadas através da modelização estatística de decisões anteriores, utilizando métodos

de dois grandes domínios da informática: O processamento da linguagem natural e a aprendizagem automática.

As principais etapas na operacionalização da justiça preditiva são as seguintes:

- *Recolha massiva de dados de processos judiciais*. Esta etapa pode incluir documentos judiciais, decisões anteriores, estatísticas de casos, leis e regulamentos, entre outros.
- É aplicada a técnica de *processamento de linguagem natural* (PLN), que converte ocorrências de linguagem humana em representações formais passíveis de inserção em programas de computador.
- Com base nos dados processados (por PLN), são criadas *características relevantes para alimentar os modelos de aprendizagem automática*. Essas características podem incluir palavras-chave, entidades mencionadas, padrões gramaticais e outros elementos relevantes.
- Os *algoritmos de aprendizagem da máquina são treinados* com conjuntos de dados históricos para identificar padrões e correlações entre as características extraídas e os resultados judiciais.
- *Validação do modelo*: Os modelos são validados usando conjuntos de dados de teste separados. Isso ajuda a garantir que o modelo seja capaz de fazer previsões precisas em novos casos.
- *Previsões e recomendações*: Uma vez treinado e validado, o modelo pode ser usado para fazer previsões em casos novos, fornecendo a probabilidade de certos resultados judiciais e podendo, deste modo, facilitar a tomada de decisões.

A utilização de modelos de justiça preditiva tem conhecido alguma expansão em diferentes países, mas oferece muitas reservas no campo da justiça criminal. Como alerta a Comissão Europeia para a Eficiência na Justiça (CEPEJ), a utilização de modelos de justiça preditiva nos processos penais significa riscos acrescidos, por comparação com outras áreas da aplicação do direito. Por exemplo, pode provocar o ressurgimento de doutrinas deterministas (a crença pela qual se considera que há causas pré-definidas que explicam o comportamento criminal, como por exemplo causas biológicas) e pode afetar de modo desproporcionado as comunidades à partida mais vulneráveis à criminalização e discriminação (CEPEJ, 2019, pp. 53-56).

Por outras palavras: A justiça preditiva usando IA pode reproduzir processos sociais de viés e discriminação e que surgem camuflados por via de crenças na objetividade e neutralidade da máquina. Uma forma pela qual isto pode acontecer é se os dados de treino dos algoritmos de IA usados para desenvolver modelos preditivos contiverem vieses, uma vez que o modelo preditivo vai refletir esses preconceitos. Outro modo de reprodução de vieses acontece quando as decisões judiciais preditivas – isto é, propostas pela tecnologia de IA com base nos dados com que foi treinada – ignoram variáveis socioeconómicas importantes que explicam o contexto de um caso. Daqui podem resultar decisões que não consideram adequadamente as circunstâncias subjacentes. Por fim, e não menos importante, a recolha e o uso de

dados para treinar modelos de justiça preditiva podem levantar questões de privacidade, especialmente se envolver informações sensíveis.

6.3. A Inteligência Artificial no policiamento

6.3.1. O policiamento preditivo

Uma das facetas mais debatidas em relação ao uso de tecnologias de IA nas atividades policiais diz respeito ao chamado policiamento preditivo (Meijer e Wessels, 2019). Na literatura, não existe uma definição unânime de policiamento preditivo, mas há algum consenso sobre as suas principais características. Muitos autores indicam que o policiamento preditivo implica a aplicação de técnicas quantitativas para prever onde as atividades criminosas podem ocorrer num futuro (próximo). Por outras palavras, o policiamento preditivo é um conceito que se baseia na premissa de que é possível prever os crimes que voltarão a ocorrer no futuro com base em análise informática de informações sobre crimes cometidos anteriormente (McCue, 2014; Norton, 2013; Williams et al., 2017).

Não obstante as origens do policiamento preditivo com base em computadores se possam situar na década de 1970, foi a partir dos anos 2010 que a previsão da ocorrência de crimes, tipicamente associada à prevenção, se tornou uma tendência mais acentuada por via de uma maior digitalização de dados de interesse para a investigação policial e de um maior desenvolvimento de *softwares* específicos para a área do policiamento e segurança (Wilson, 2020). Outros fatores terão contribuído para a expansão do policiamento preditivo, tais como medidas de austeridade que limitam os recursos humanos e conduzem à procura de novas soluções alegadamente mais eficientes; uma perceção crescente de que a polícia deve adotar uma postura preventiva, colocando a ênfase na antecipação de riscos e na prevenção; e um aumento do volume e da complexidade dos dados disponíveis passíveis de serem computorizados, tornando-se necessário criar ferramentas de processamento de dados e análise cada vez mais sofisticadas (Babuta e Oswald, 2021).

Os métodos de policiamento preditivo podem ser divididos em quatro grandes categorias: Métodos que visam a previsão de crimes, ou a previsão de locais e horários com um risco acrescido de crime; métodos que visam a previsão de infratores, ou a identificação de indivíduos em risco de cometer crimes (ou reincidir) no futuro; métodos que visam a previsão de infratores, ou a criação de perfis semelhantes aos de infratores passados; e métodos que visam a previsão de vítimas de crimes, utilizados para identificar grupos ou indivíduos que são suscetíveis de se tornarem vítimas de crimes (Perry, 2013).

As previsões baseadas nestas ferramentas analíticas podem ser usadas a vários níveis: A um nível macro, para o planeamento estratégico e a definição de prioridades de alocação de recursos; a um nível operacional, para identificar alvos prováveis para a intervenção policial e prevenção do crime, podendo orientar a tomada de decisões

no que respeita à afetação de forças policiais a determinados locais; e a tomada de decisões ou avaliações de risco relacionadas com indivíduos (Moses e Chan, 2018).

Para efeitos de policiamento preditivo são recolhidos e cruzados dados de diversos tipos de fontes: Desde bases de dados comerciais a informação biográfica, dados biométricos, informação financeira, dados de georreferenciação, dados associados a redes de interação, dados sobre emprego, viagens, migração e registo criminal (Leese, 2022). O policiamento preditivo suscita diversas questões, designadamente as seguintes: Em primeiro lugar, a eventual rotulagem de certas pessoas como suspeitas durante o processo de policiamento preditivo pode suscitar questões de proteção de dados, mas também potencialmente afetar o direito à presunção de inocência.

Em segundo lugar, os dados recolhidos para efeito de policiamento preditivo dependem frequentemente, pelo menos em parte, do tratamento de dados que não estão originalmente relacionados com o crime, mas que são inicialmente recolhidos por empresas privadas no contexto da sua atividade comercial normal (por exemplo, bancos, telecomunicações, viagens). Os esquemas de policiamento preditivo também se baseiam tipicamente em *software* preditivo produzido por empresas privadas, quer se trate de fornecedores especializados neste domínio ou de grandes empresas tecnológicas. Assim, surgem riscos acentuados de vigilância e recolha desproporcionada de dados sobre os cidadãos, deixando lacunas em termos de transparência, prestação de contas e responsabilidade.

Por fim, os modelos de policiamento preditivo baseiam-se em técnicas de associação de riscos a determinados indivíduos, comunidades, nacionalidades e locais, o que contribui para processos de discriminação e preconceito. Importa considerar que os modelos preditivos de policiamento assentam em noções prévias sobre quem são (e/ou qual a sua aparência física, modo de vestir, modo de falar, etc.) as comunidades "suspeitas". Segundo a proposta de Christina Pantazis e Simon Pemberton, uma comunidade suspeita consiste em:

> Um subgrupo da população que é destacado para a atenção do Estado como sendo "problemático". Especificamente, em termos de policiamento, os indivíduos podem ser visados, não necessariamente como resultado de suspeitas de infração, mas simplesmente devido à sua presumível pertença a esse subgrupo. (Pantazis e Pemberton, 2009, p. 649)

Ou seja, as atividades de policiamento envolvem sempre processos de categorização social que operam com base em classificações socias que distinguem entre suspeitos e não suspeitos. De acordo com os sociólogos Simon Cole e Michael Lynch (2006), o que impera é uma visão "objetivista" da suspeição, mas que esta deve ser substituída por uma abordagem "construtivista". Nas palavras destes autores:

> Uma visão objetivista [convencional] sustenta que os suspeitos existem e são identificados por características específicas (...) em contraste, os construtivistas sustentam que os suspeitos são construídos através da interação social com os agentes, agências e processos do sistema de justiça criminal (...)

> [através do] estigma demográfico, socioeconómico e cultural que os agentes da justiça criminal associam ao estatuto de suspeito. (Cole e Lynch, 2006, p. 40)

Um processo "objetivista" de construção da suspeição acontece com os modelos preditivos de policiamento e a utilização de informação biométrica (por exemplo, reconhecimento facial) baseados em sistemas de IA. Quando são selecionados e recolhidos determinados dados e que são abstraídos dos seus contextos, de modo a serem traduzidos para códigos numéricos adaptados aos sistemas de IA, os dados acabam por ser reagrupados de acordo com determinados critérios tidos como objetivos. Este processo de recolha e reorganização dos dados veicula uma certa representação do mundo numa forma numérica e computável, centrando-se naquilo que considera ser a veracidade desses mesmos dados. Esta representação do mundo, potenciada pela crença na objetividade e maior eficiência da IA, tende a ignorar os efeitos subjetivos das pré-noções de suspeição que recaem sobre determinadas comunidades, sendo este um processo conjuntamente elaborado por várias instâncias do sistema de justiça (Machado et al., 2020; Matzner, 2016).

6.3.2. Reconhecimento facial

Além do *software* preditivo, outra tecnologia de IA usada em atividades policiais é o reconhecimento facial. O reconhecimento facial associado à IA é uma tecnologia que combina técnicas de visão computacional e aprendizagem automática para identificar e autenticar indivíduos com base em características faciais únicas. As etapas de utilização desta tecnologia são as seguintes:

- *Captura de imagem*: O processo começa com a captura de uma imagem ou vídeo que contenha rostos humanos. Isso pode ser feito por câmaras tradicionais ou câmaras especiais projetadas para a recolha de informações específicas do rosto.
- *Deteção de rosto*: São aplicados algoritmos para localizar e isolar as regiões faciais nas imagens. Esses algoritmos podem identificar características como olhos, nariz, boca e contornos do rosto.
- *Extração de características*: Após a deteção do rosto, a tecnologia de reconhecimento facial utiliza algoritmos para extrair características distintivas do rosto, como a distância entre os olhos, o formato do nariz e a disposição das características faciais.
- *Criação de vetor de características*: As características extraídas são convertidas em um vetor ou conjunto de números que representam as características faciais únicas da pessoa.
- *Aprendizagem automática*: A IA, por meio de técnicas de aprendizagem automática, treina em grandes conjuntos de dados para reconhecer padrões e variações nas características faciais. Isso permite que o sistema melhore a sua capacidade de identificar rostos com base em experiências anteriores.
- *Base de dados e comparação*: Os vetores de características dos rostos são armazenados em bases de dados. Quando uma nova imagem é recolhida, o sistema

compara as características extraídas com as armazenadas na base de dados para identificar ou verificar determinada pessoa.
- *Tomada de decisão*: Com base na comparação, o sistema de IA toma uma decisão, como autenticar a identidade da pessoa ou alertar sobre uma possível correspondência.

O reconhecimento facial associado à IA é utilizado numa variedade de aplicações, desde situações do quotidiano (por exemplo, desbloquear *smartphones* ou computadores, aceder a edifícios por via de autenticação de identidade), a contextos clínicos (por exemplo, diagnóstico precoce de doenças associadas a transformações faciais impercetíveis ao olho humano ou deteção de emoções em casos de doença mental), mas também em atividades de policiamento e segurança pública.

A utilização de tecnologias de reconhecimento facial em contexto de policiamento tem gerado amplas controvérsias: Desde queixas relacionadas com a elevada taxa de erros desta tecnologia, que podem reforçar discriminação de género e de raça (por exemplo, esta tecnologia tem revelado várias falhas na deteção de rostos que não se alinhem com o "padrão facial" associado a homens brancos – ou seja, apresenta uma taxa de erro mais elevada com mulheres, pessoas negras, pessoas não binárias e pessoas com deficiências), a receios que a tecnologia de reconhecimento facial instigue uma vigilância desproporcionada sobre os cidadãos, que colide com vários direitos humanos. Uma das críticas é o facto de os dados biométricos (neste caso, os rostos) de milhares de pessoas serem recolhidos sem o seu consentimento e usados em bases de dados de suspeitos de crimes, podendo ser cruzados com informações com interesse para identificação de suspeitos e investigação criminal.

6.4. A Inteligência Artificial em prisões

De acordo com alguns estudos dedicados à aplicação de tecnologias de IA em prisões, a utilização deste tipo de tecnologia pode abarcar essencialmente três áreas: A segurança e vigilância, a gestão administrativa de recursos e serviços (Puolakka e Van De Steene, 2021), e o uso de técnicas para avaliar o risco de reincidência dos reclusos (Zivani e Mahlangu, 2022).

Um estudo conduzido por Puolakka e Van De Steene (2021), englobando prisões de 20 países em diferentes regiões do mundo, concluiu que a utilização de IA é ainda residual ou está ainda em fase de experimentação. No entanto, foram identificadas algumas jurisdições que utilizam IA em contexto prisional. Referindo-se ao caso das prisões asiáticas, os autores identificaram os casos da China continental, Hong-Kong, Singapura e India, onde várias prisões utilizam IA no âmbito da segurança e vigilância. Em muitas prisões destes países, os reclusos são vigiados e controlados de modo permanente, inclusive no interior das celas, através de uma rede de câmaras e sensores que usam reconhecimento facial e análise de padrões de movimentação humana, permitindo uma análise em tempo real destinada a detetar qualquer comportamento considerado anómalo.

O recurso à IA para propósitos de vigilância e segurança nas prisões foi identificado em outros países além dos asiáticos. Por exemplo, prisões nos Estados Unidos da América usam sistemas de IA aplicados a chamadas telefónicas realizadas por e com reclusos que conjugam reconhecimento de discurso, análise semântica e aprendizagem automática. O objetivo destes sistemas de IA é construir grandes bases de dados com conversas de reclusos e desenvolver técnicas de análise que venham a permitir detetar padrões suspeitos de conversação que permitam identificar atividades ilícitas como, por exemplo, contrabando, tráfico de droga ou planeamento de crimes (Cassen-Weiss, 2019). No Reino Unido, algumas prisões utilizam IA para processamento de linguagem natural e técnicas de aprendizagem automática a partir de textos retirados dos processos judiciais dos reclusos para incorporação dessa informação em estimativas de risco de violência (Puolakka e Van De Steene, 2021, p. 8).

Uma das áreas em que se tem debatido o uso de IA em contexto prisional diz respeito à introdução de robôs, *chatbots* e assistentes virtuais, que possam, entre outras tarefas, servir como companheiros de confinamento a reclusos que enfrentam castigos severos, como a "cela solitária" (forma especial de punição pela qual o recluso é encarcerado numa cela individual e privado de qualquer contacto humano). No entanto, este tipo de sugestão tem sido criticada por se recear que a introdução de "companheiros digitais" ou robôs para combater a solidão sirva para legitimar a continuidade de políticas que favorecem punições extremas (Berry, 2023). É também questionável a tendência para a substituição de humanos por máquinas em processos que exigem a gestão de emoções, correndo-se o risco de perder um aspeto-chave da reabilitação: Relações humanas de qualidade. Por fim, a utilização de dados extraídos de ambientes prisionais para construir modelos de avaliação de risco e de previsão do comportamento futuro dos reclusos suscita questões éticas muito sensíveis, na medida em que o ambiente de onde são extraídos esses dados é bastante particular, representando um contexto em que as pessoas acumulam a privação de liberdade com várias vulnerabilidades (psicológicas, identitárias e sociais). É muito complexo avaliar riscos e necessidades específicas da população reclusa a partir de dados gerados pelo próprio sistema prisional e a partir do ambiente prisional. Como chamam a atenção Puolakka e Van De Steene:

> Ainda estamos muito longe de ter sistemas penitenciários que possam prevenir a reincidência criminal, daí que devamos ser extremamente cautelosos em confiar em sistemas de IA que são treinados com dados gerados em ambientes prisionais. Se houver viés cognitivo no modo como avaliamos as necessidades e os riscos dos reclusos, esse viés será repetido pelos algoritmos e, no pior cenário, estes podem justificar o processo de enviesamento. (Puolakka e Van De Steene, 2021, p. 134)

6.5. A Inteligência Artificial no controlo de fronteiras e cooperação transnacional em justiça criminal

Ao longo dos últimos anos, agências de investigação criminal e segurança têm intensificado esforços para expandir e consolidar a cooperação internacional no combate ao terrorismo e criminalidade organizada (Amelung et al., 2020). Na Europa, destacam-se três agências: Em primeiro lugar, a Europol – a Agência da União Europeia para a Cooperação Policial – formada em 1998 para lidar com a inteligência criminal e combater o grave crime organizado internacional e o terrorismo através da cooperação entre as autoridades competentes dos Estados-Membros da União Europeia. Em segundo lugar, a Eurojust, uma agência da União Europeia criada em 2002, constituída por juízes, procuradores e policiais dos diferentes Estados-Membros, pela necessidade de reduzir o crime organizado dentro da UE, assim como todas as questões relacionadas com o controlo de fronteiras. Em terceiro lugar, a agência da União Europeia para a Gestão Operacional de Sistemas Informáticos de Grande Escala no Espaço de Liberdade, Segurança e Justiça (eu-LISA), fundada em 2011 para assegurar o funcionamento ininterrupto de sistemas informáticos de grande escala na gestão de políticas de asilo, fronteiras e migração da União Europeia.

Em anos recentes, a Eurojust e a eu-LISA têm realizado esforços significativos para acelerar a transformação digital e a adoção de soluções informáticas baseadas na IA no domínio da Justiça e dos Assuntos Internos (eu-LISA e Eurojust, 2022). Em 2020, a eu-LISA publicou um relatório sobre a IA na Gestão Operacional de Sistemas Informáticos de Grande Escala, que descreve a gama de oportunidades para a IA no âmbito da atividade principal da eu-LISA (eu-LISA, 2020). Especificamente, existem vários projetos em que a eu-LISA já implementou ou está a ponderar a implementação de IA, designadamente nas seguintes áreas: Intercâmbio de informações sobre condenações penais; registos criminais de nacionais de países terceiros; desenvolvimento e gestão operacional de plataformas digitais de colaboração de equipas de investigação de diferentes Estados-Membros, que apoiará a colaboração judicial transfronteiriça.

A eu-LISA e a Eurojust, com o apoio da Comissão Europeia, estão a investir em duas tecnologias de IA principais aplicadas às atividades de investigação criminal, controlo de fronteiras e cooperação transnacional nesses domínios: A visão computacional (que está na base, por exemplo, das tecnologias de reconhecimento facial) e o processamento de linguagem natural. Ambas as tecnologias são tidas como particularmente relevantes em situações em que é necessário o tratamento de dados não estruturados em grande escala. Esses dados não estruturados tanto podem estar em formato de texto (por exemplo, mensagens de correio eletrónico, documentos escritos diversos) como em formato de imagem estática ou dinâmica (neste último caso, um exemplo é a transmissão de vídeo em direto).

As tecnologias de IA baseadas em processamento de linguagem natural podem apoiar a compreensão semântica no domínio da justiça, que é imprescindível não só para entender o significado subjacente da linguagem jurídica, mas também para

incorporar em assistentes de voz e *chatbots*, para a tradução automática em investigações transfronteiriças, e para o resumo de textos ou para a anonimização de documentos. Esta tecnologia, em combinação com outras abordagens como os gráficos de conhecimento[38], pode também ser eficazmente utilizada na investigação jurídica, facilitando a identificação de jurisprudência relevante, a atualização de nova informação jurídica, a extração de relações entre entidades jurídicas e, sobretudo, a compreensão do significado real por trás das palavras utilizadas no sistema de justiça.

As tecnologias associadas à visão computacional, e em particular as tecnologias de reconhecimento biométrico (como é o caso do reconhecimento facial), podem ser utilizadas no contexto de investigações criminais, em particular quando a análise de grandes volumes de dados de imagem ou vídeo é necessária para a identificação de pessoas (por exemplo, potenciais vítimas ou autores de crimes). É ainda possível a geração de biometria sintética: Ou seja, os fluxos de vídeo podem ser anonimizados, substituindo rostos reais por imagens sintéticas, protegendo assim as identidades das pessoas cujos rostos são capturados em vídeo em espaços públicos.

Na União Europeia, os sistemas de informação de grande escala relacionados com as fronteiras que integram o processamento algorítmico estão normalmente ligados, através de uma variedade de fluxos de dados, às autoridades nacionais, à Europol ou, frequentemente, a ambas. Os dados provenientes das autoridades responsáveis pela aplicação da lei podem ser introduzidos nos sistemas relacionados com as fronteiras e os dados destes sistemas podem eventualmente ser fornecidos ou tornados acessíveis às autoridades responsáveis pela aplicação da lei. A segurança e a gestão das fronteiras têm sido os principais motores do desenvolvimento de sistemas de informação centralizados e descentralizados no chamado "Espaço de Liberdade, Segurança e Justiça" (ELSJ) [39] da União Europeia.

A área de controlo de fronteiras e de cooperação transnacional no combate ao terrorismo e criminalidade organizada tem sido o alvo de maior investimento em termos de digitalização e de utilização de tecnologias de IA no sistema de justiça. Isto acontece por vários fatores, designadamente os seguintes: É considerada uma área prioritária para a segurança pública, daí que os governos e os organismos internacionais tenham um interesse adicional em investir recursos para o desenvolvimento de tecnologia neste setor; é um contexto em que circulam muitos dados passíveis de digitalização e conversão para dados computacionais (por exemplo, dados de

38 Um gráfico de conhecimento é uma representação estruturada de conhecimento que representa diferentes entidades e as suas relações. É uma forma de organizar informações de maneira interconectada A combinação de gráficos de conhecimento e tecnologia de processamento de linguagem natural (PLN) pode ser especialmente valiosa na área da pesquisa jurídica. Por exemplo, permitindo a extração de informação jurídica e a análise de relações complexas entre entidades jurídicas, podendo neste caso ajudar a identificar precedentes importantes, influências jurisprudenciais e conexões entre casos jurídicos.

39 O Espaço de Liberdade, Segurança e Justiça (ELSJ) é um conjunto de políticas de assuntos internos e de justiça destinadas a garantir a segurança, os direitos e a livre circulação na União Europeia (UE). As áreas abrangidas incluem a harmonização do direito internacional privado, acordos de extradição entre os estados-membros, políticas de controlo das fronteiras internas e externas, vistos de viagem comuns, políticas de imigração e de asilo e cooperação policial e judiciária.

passageiros, registos criminais, dados biométricos); é uma das arenas da vida social onde são mais visíveis os mecanismos de vigilância e de controlo de comunidades suspeitas (na medida em que grupos mais vulneráveis à suspeição, como refugiados, pessoas que pedem asilo e pessoas de várias nacionalidades circulam de modo particularmente visível em espaços como aeroportos e fronteiras). Na área de controlo de fronteiras e de cooperação internacional para combate ao terrorismo e criminalidade vigoram práticas de avaliação de risco que são consideradas passíveis de introduzir níveis elevados de discriminação, pois surgem associadas à construção social da suspeição com base em determinados dados demográficos. Pessoas de determinado género, grupo etário, etnia, nacionalidade, religião e/ou profissão estão mais sujeitas ao escrutínio de agentes de autoridade, circunstâncias que se agudizam em contextos de mobilidade internacional (Amelung e Machado, 2019; Machado e Granja, 2020; Machado et al., 2020). Como vários estudos apontam, a criminalização de determinadas populações não acontece apenas com um enfoque territorial (como é comum, por exemplo, ao nível das atividades de policiamento, em que determinados bairros ou zonas de cidade são mais objeto de vigilância policial do que outros), sendo crescente a expansão do fenómeno de vigilância das autoridades para o contexto dos emigrantes, refugiados e requerentes de asilo e outros grupos envolvidos em processos de mobilidade internacional (Aas, 2011; Ajana, 2013; Amoore, 2013).

Ao longo dos últimos anos, o interesse das autoridades públicas por dados relacionados com migração e asilo parece estar a aumentar em muitos Estados-Membros da União Europeia. Alguns países introduziram alterações jurídicas nos seus procedimentos em relação a requerentes de asilo, nomeadamente para permitir que as autoridades apreendam e analisem os dados contidos em dispositivos pessoais (por exemplo, *smartphone*s) destas pessoas, com o intuito de determinar a sua identidade e os seus itinerários de viagem. As tecnologias usadas em fronteiras e aeroportos são múltiplas e podem incluir sistemas de vigilância automatizados com diferentes capacidades de deteção, desde reconhecimento facial a deteção de batimentos cardíacos e câmaras térmicas.

A crescente criminalização de comunidades migrantes conheceu desenvolvimentos acentuados pela ação de sistemas de informação computadorizados que operam em rede e que se associam à recolha, armazenamento e análise de um volume e diversidade de dados digitais com uma escala sem precedentes, que se conjuga com o desejo das agências de segurança da União Europeia de tornar os sistemas de vigilância transnacional mais "interoperáveis" (Leese, 2022). Muitos destes sistemas informáticos são aplicados em ferramentas de vigilância de aeroportos e outros espaços há décadas. Contudo, com os avanços da IA esta recolha e análise de dados tem-se tornado mais automatizada, de modo a que os computadores – e não os seres humanos – façam determinações preliminares sobre possíveis ameaças e sobre a forma como as autoridades devem reagir. A IA promete aumentar esta vigilância, por via da automatização e interoperabilidade, tornando as ferramentas mais poderosas e capazes de processar e interpretar mais dados do que no passado.

A implementação destas tecnologias, que muitas vezes foi mais rápida do que os quadros legislativos e recomendações para regular a sua utilização, tem suscitado preocupações relacionadas com a fragilização da privacidade e a crescente vigilância não só de migrantes e viajantes, mas de populações inteiras.

6.6. Desafios sociais e éticos

A "Carta Ética Europeia sobre a Utilização da Inteligência Artificial nos Sistemas Judiciais e no seu Ambiente", elaborada pela Comissão Europeia para a Eficiência na Justiça (CEPEJ, 2018), estipula cinco princípios fundamentais a salvaguardar para o sistema judicial. Na nossa perspetiva, esses princípios devem ser aplicados não apenas ao sistema judicial, mas a todos os setores do sistema de justiça. Os cinco princípios fundamentais são os seguintes:

- O princípio do *respeito pelos direitos fundamentais*: Garantir que a conceção e a aplicação de ferramentas e serviços de IA sejam compatíveis com os direitos fundamentais.
- O princípio da *não discriminação*: Impedir especificamente o desenvolvimento ou a intensificação de qualquer discriminação dirigida a indivíduos ou grupos de indivíduos.
- O princípio da *qualidade e segurança*: No que diz respeito ao tratamento das decisões e dos dados judiciários, utilizar fontes certificadas e dados com modelos elaborados de forma multidisciplinar, num ambiente tecnológico seguro.
- O princípio da *transparência, imparcialidade e equidade*: Tornar acessíveis e compreensíveis os métodos de tratamento dos dados e autorizar auditorias externas.
- O princípio do *"controlo do utilizador"*: Garantir que os utilizadores sejam intervenientes informados e controlem as escolhas efetuadas.

A aplicação destes cinco princípios não é isenta de críticas. Comecemos por abordar a reflexão promovida pela própria CEPEJ (2018), num documento intitulado "Estudo aprofundado sobre a utilização da IA nos sistemas judiciais, nomeadamente as aplicações de IA que processam decisões e dados judiciais" (Ronsin et al., 2018). Segundo os autores, há que questionar de modo reflexivo a aplicação desses cinco princípios e a esse respeito começam por destacar que uma questão premente tem que ver com o facto da utilização de IA poder não ser compatível com os direitos individuais consagrados na Convenção Europeia dos Direitos do Homem (CEDH). Estes incluem o direito a um julgamento justo (em especial o direito a um juiz natural estabelecido por lei, o direito a um tribunal independente e imparcial, e a igualdade de armas nos processos judiciais) e, nos casos em que não tenham sido tomadas medidas suficientes para proteger os dados comunicados em dados abertos, o direito ao respeito pela vida privada e familiar (Ronsin et al., 2018, p. 15). Importa atender a que a iniciativa para o desenvolvimento destes instrumentos tecnológicos provém em grande parte do sector privado, cuja clientela até agora tem sido maioritariamente constituída por companhias de seguros, advogados e serviços jurídicos que pretendem reduzir

a insegurança jurídica e a imprevisibilidade das decisões judiciais. No entanto, os decisores públicos começam a ser cada vez mais solicitados por um sector privado que deseja ver estas ferramentas integradas nas políticas públicas, o que, como salientam Rosin e colegas (2018), requer uma abordagem cautelosa. Nas suas palavras:

> É essencial que qualquer debate público envolva todas as partes interessadas, sejam profissionais do direito, empresas de tecnologia jurídica ou cientistas, a fim de lhes permitir transmitir todo o alcance e possível impacto da introdução de aplicações de inteligência artificial nos sistemas judiciais e definir o quadro ético em que devem operar. Subsequentemente, este debate poderia ir além de um quadro puramente "empresarial", envolvendo os próprios cidadãos. (Ronsin et al., 2018, p. 16)

Uma outra dimensão a considerar é que os sistemas de IA geram riscos que interagem com desafios sociais e éticos não diretamente relacionados com a IA. Por exemplo, a IA reproduz problemas sistémicos de viés e discriminação que estão inseridos em estruturas económicas e sociais que produzem efeitos negativos cumulativos; e é imprescindível considerar formas complexas de responsabilidade, na medida em que estas tecnologias envolvem empresas e agências governamentais em situações difusas de agência moral e jurídica (Bakiner, 2023).

O sistema de justiça pode servir como um espelho das desigualdades e relações de poder presentes numa sociedade, tendo suscitado vários estudos que envolvem uma análise crítica das estruturas sociais e económicas que perpetuam a discriminação. Estes estudos abrangem várias temáticas:

- *A aplicação desigual das leis*: A maneira como as leis são aplicadas muitas vezes reflete os preconceitos e as desigualdades presentes na sociedade. Isso pode incluir penas mais severas para crimes semelhantes e a criminalização de certos comportamentos que afetam desproporcionalmente grupos específicos. Em alguns casos, as leis e práticas judiciais podem ter um impacto desproporcional sobre comunidades em situação de vulnerabilidade social e económica e minorias étnicas.
- *Policiamento discriminatório*: O policiamento muitas vezes reflete e amplifica as desigualdades sociais existentes. Práticas como a discriminação racial ou étnica em abordagens policiais podem resultar em prisões e condenações injustas.
- *Ciclo de pobreza e suspeição*: Existe um ciclo interconectado de pobreza e suspeição, visível em taxas mais elevadas de encarceramento de pobres ou uma vigilância mais invasiva em migrantes em situação de vulnerabilidade social e económica.

Perante este cenário, a IA pode conduzir a uma reprodução e consolidação de desigualdades e de injustiça. De facto, se os dados usados para treinar modelos de IA contêm vieses e refletem discriminação, a IA pode reproduzir e amplificar desigualdades existentes. Do mesmo modo, se os conjuntos de dados utilizados para treino de sistemas de IA não representam adequadamente todas as comunidades e grupos,

o modelo pode ter uma visão distorcida e limitada da realidade, resultando em decisões injustas e discriminatórias.

6.7. Atividades para debate

Apresentam-se exemplos de casos para debate à luz das questões sociais e éticas analisadas neste capítulo e desenvolvidas em termos gerais e abstratos com mais detalhe no capítulo 2. Considere ainda as diferentes dimensões e os diversos níveis de análise apresentados no capítulo 3, selecionando aqueles que lhe pareçam mais adequados à análise de cada caso e justificando porquê.

Caso 1 (real)

> Por iniciativa do Ministério da Justiça Francês, dois tribunais de recurso aceitaram testar um *software* de justiça preditiva em vários processos de decisões civis, com base em dados jurisprudenciais internos. A empresa que desenvolveu o *software* propôs-se realizar uma análise quantificada dos montantes de indeminização atribuídos pelos tribunais, bem como uma classificação geográfica das discrepâncias observadas em pedidos e julgamentos semelhantes. O objetivo declarado do *software* era criar uma ferramenta de decisão para reduzir, se necessário, a variabilidade excessiva das decisões judiciais, em nome do princípio da igualdade dos cidadãos perante a lei.
>
> O resultado da experiência, contraditoriamente debatido entre os dois tribunais de recurso, revelou infelizmente a ausência de valor acrescentado da versão testada do *software* para o trabalho de reflexão e de tomada de decisão dos magistrados; mais significativamente, foram revelados enviesamentos de raciocínio do *software* que conduziram a resultados aberrantes ou inadequados, devido à confusão entre as meras ocorrências lexicais da fundamentação judicial e as causalidades que tinham sido decisivas no raciocínio dos juízes. (CEPEJ, 2018, p. 42)

Comente este caso, referindo-se aos riscos suscitados pelas propostas de decisão judicial baseadas em modelos de "justiça preditiva". Pode articular o mapeamento desses riscos com uma reflexão crítica que contenha uma apreciação sobre os mitos em torno da crença na infalibilidade da IA e os desafios sociais e éticos da automatização de tarefas.

Caso 2 (real)

Na Finlândia é muito comum que os reclusos possam trabalhar e estudar, muitas vezes em regime aberto (ou seja, saindo para o exterior). Aproveitando este contexto, uma empresa finlandesa que desenvolve IA – a *Metroc* – protocolou com prisões a utilização da força laboral de reclusos em tarefas de treino de algoritmos. No caso concreto, os reclusos classificam dados para treinar um modelo linguístico associado a um motor de busca destinado a ajudar as empresas de construção civil a encontrar projetos de construção recentemente aprovados. Por exemplo, é solicitado aos reclusos que façam a distinção, por via de um simples *click*, entre uma janela e um edifício. Por este trabalho os reclusos recebem uma pequena compensação.

Em todo o mundo, milhões dos chamados *clickworkers* treinam modelos de IA, ensinando às máquinas a diferença entre peões e palmeiras, ou que combinação de palavras descreve violência ou abuso sexual. Normalmente, estes trabalhadores vivem em países pobres, onde os salários são baixos.

Nesta situação concreta, a empresa *Metroc* obtém trabalhadores baratos, que falam finlandês, enquanto o sistema prisional pode oferecer aos reclusos um emprego que, segundo a empresa, os prepara para o mundo do trabalho digital após a sua libertação. A utilização de reclusos para treinar a IA cria paralelos incómodos com o tipo de trabalho mal pago e por vezes explorador que tem existido frequentemente a jusante no desenvolvimento de tecnologia de IA. Mas na Finlândia, o projeto tem recebido um apoio generalizado.

Comente este caso à luz do que aprendeu sobre as desigualdades perpetuadas pela IA.

Caso 3

Aeroportos em diferentes regiões do mundo têm investido na digitalização do controlo de passageiros, automatizando a identificação de pessoas com bases em dados biométricos que variam da recolha e armazenamento de impressões digitais a imagens de rostos. Os procedimentos de recolha, armazenamento, circulação e partilha destes dados surgem rodeados por falta de transparência e prestação de contas, desconhecendo-se se é ou não cumprida legislação em vigor ou em que medida estes procedimentos são ou não dotados de excecionalidade. Vários ativistas e académicos têm clamado por maior transparência e prestação de contas, receando que estas tecnologias não só veiculem práticas injustas de discriminação como facilmente possam ser aplicadas em vigilância da população em outros espaços públicos.

Discuta a problemática da utilização de IA para o controlo e monitorização de passageiros, com base no debate sobre as respetivas implicações sociais e éticas.

Caso 4 (real)

Em 2022, o governo do Reino Unido anunciou os seus planos de adoção de *smartwaches* para monitorizar migrantes com registo criminal. A intenção seria uma vigilância permanente destes indivíduos, obrigando-os a tirar fotografias deles próprios várias vezes ao dia. Enquanto que o governo alega que esta é uma abordagem alternativa à custódia ou prisão preventiva (menos dispendiosa e "mais humana"), especialistas em direitos humanos e criminologia apresentam outros argumentos (retirados de Blount, 2024):

Sabe-se que o reconhecimento facial é uma tecnologia imperfeita e perigosa que tende a discriminar as pessoas de cor e as comunidades marginalizadas. Estas inovações no policiamento e na vigilância são frequentemente impulsionadas por empresas privadas, que lucram com a corrida dos governos à vigilância total e ao controlo das populações. (…) Através de tecnologias e algoritmos opacos, facilitam a discriminação governamental e as violações dos direitos humanos sem qualquer responsabilidade. Nenhum outro país da Europa utilizou esta tecnologia desumanizante e invasiva contra os migrantes. *Lucie Audibert, advogada e responsável jurídica da Privacy International*

A monitorização eletrónica é uma tecnologia de controlo intrusiva. Alguns indivíduos desenvolvem sintomas de ansiedade, depressão, ideação suicida e deterioração geral da saúde mental. O Ministério do Interior ainda não sabe ao certo quanto tempo as pessoas permanecerão sob vigilância. Não apresentou quaisquer provas da necessidade da monitorização eletrónica nem demonstrou que [faz] com que as pessoas cumpram melhor as regras de imigração. O que precisamos é de soluções humanas, não degradantes e baseadas na comunidade. *Monish Bhatia, professor de criminologia na Birkbeck, Universidade de Londres*

Comente este caso, à luz dos desafios sociais e éticos da IA, em particular no sistema de justiça.

Conclusão

Ao finalizar este livro, reforçamos a importância de reconhecer que a compreensão em profundidade e a reflexão sobre os desafios sociais e éticas da Inteligência Artificial (IA) no século XXI requerem uma análise cuidadosa arredada de visões deterministas da tecnologia. A Sociologia emerge como uma área de conhecimento particularmente adequada para essa missão, ao desvendar o desenvolvimento da IA como um fenómeno sociotécnico, examinando as suas interações com contextos históricos, sociais, culturais, políticos e económicos mais amplos. Esse enfoque analítico é possível dentro de uma variedade de perspetivas no campo das teorias do social, desde aquelas que enfatizam aspetos culturais e simbólicos, explorando diferentes interpretações e significados atribuídos à IA por diversos grupos sociais, até abordagens sociológicas interseccionais que consideram o impacto desigual da IA, levando em conta as narrativas globais do capitalismo e como as desigualdades são perpetuadas nos e pelos sistemas digitais e tecnológicos.

Além de explorar os desafios sociais da IA, este livro também abordou as suas implicações éticas, adotando uma perspetiva distinta dos enquadramentos tradicionais. A necessidade de traduzir os princípios éticos em práticas concretas para lidar com os impactos da IA nos domínios sociais, políticos e materiais é, nessa perspetiva, crucial. O presente livro junta-se aos apelos de uma abordagem ética baseada no cuidado para enfrentar esses desafios, destacando a importância da sensibilidade às complexidades dos problemas coletivos e da inclusão das vozes das comunidades afetadas pela IA. Essa ética enfatiza a interconexão entre as tecnologias de IA, as pessoas e o ambiente, adotando uma perspetiva mais ampla que reconhece os impactos sociais, políticos e ambientais das tecnologias desde a sua conceção até à sua implementação. Além disso, a ética do cuidado reconhece que cuidar é tanto um compromisso ético-político quanto uma prática material situada. Ao destacar os desafios sociais relacionados com a IA, a ética do cuidado ressalta a necessidade de situar os impactos da IA em contextos práticos e de os contextualizar nas relações de poder desiguais e assimétricas. Essa abordagem é especialmente relevante dada a influência dos interesses económicos e comerciais na agenda de regulamentação da IA. Diferentemente das abordagens prescritivas convencionais, a ética do cuidado promove uma reavaliação de questões humanistas e existenciais, visando fornecer ferramentas que garantam escolhas responsáveis e uma IA orientada para o bem-estar social.

Portanto, a abordagem da IA como fenómeno sociotécnico em articulação com a ética do cuidado oferecem uma estrutura sólida para uma análise crítica e abrangente da vida em sociedade. Os campos da educação, saúde e justiça foram selecionados neste livro como exemplos de aplicabilidade da nossa reflexão.

Considerando a esfera da educação, este livro revela como uma abordagem da IA enquanto fenómeno sociotécnico convoca a necessidade de considerar como essa tecnologia pode afetar o ambiente educacional, desde a sala de aula até aos sistemas de ensino mais amplos. Isso implica examinar não apenas o potencial educacional da IA, mas também as suas implicações culturais, sociais e éticas. Por exemplo, como a IA pode influenciar a equidade no acesso à educação ou perpetuar preconceitos e

desigualdades existentes? Uma ética do cuidado destaca a importância de garantir que as decisões relacionadas com a implementação da IA na educação sejam guiadas pelo bem-estar das comunidades educativas e pela equidade educacional.

No campo da saúde, uma abordagem da IA como fenómeno sociotécnico considera não apenas os avanços tecnológicos, mas também as implicações sociais e éticas da sua utilização. Isso inclui questões como o acesso igualitário aos cuidados de saúde, a privacidade dos pacientes, e o viés algorítmico em diagnósticos e tratamentos. A ética do cuidado enfatiza a importância de colocar o bem-estar do paciente no centro das decisões relacionadas com a IA na saúde, garantindo que os sistemas de saúde sejam sensíveis às necessidades individuais e não perpetuem desigualdades existentes.

A análise detalhada do campo da justiça também beneficia destes conceitos. Ao considerar a IA como um fenómeno sociotécnico, é essencial examinar como essas tecnologias influenciam o acesso à justiça, os direitos individuais e a equidade no sistema legal. A ética do cuidado destaca a importância de garantir que a aplicação da IA no sistema judicial seja guiada pela justiça e pelo respeito pelos direitos humanos, evitando a perpetuação de preconceitos e desigualdades.

Ao longo deste livro, exploramos as complexidades das interações entre as tecnologias de IA e as sociedades, enfatizando a necessidade de envolver os públicos nas decisões sobre o desenvolvimento e uso destas tecnologias disruptivas. Reconhecemos que o debate em torno da IA é muitas vezes dominado por interesses e poderes específicos, negligenciando as vozes das comunidades marginalizadas e limitando o papel dos cidadãos a funções que servem principalmente às dinâmicas económicas e de mercado. Ao desafiar estes pressupostos, procuramos ampliar o diálogo público sobre a IA, levantando questões sobre quem define o que é "bom" para a sociedade, quais são os valores sociais a prevalecer e como a IA pode ser projetada e utilizada de forma a beneficiar a sociedade como um todo.

Além disso, exploramos maneiras de desenvolver uma atitude prudente e responsável em relação aos riscos da IA e examinamos como os princípios éticos podem ser aplicados em práticas concretas para promover uma IA mais ética e compatível com o bem-estar da humanidade. Ao responder a estes desafios, os sociólogos podem contribuir para a transformação social em relação à IA por meio de ações críticas e reflexivas, comprometidas com o combate às desigualdades perpetuadas e (re) criadas por meio da tecnologia, procurando tornar a IA mais justa e igualitária, e evitando a cooptação por parte de interesses económicos e comerciais.

Esperamos que este livro possa contribuir para democratizar o debate sobre a IA e criar oportunidades para que todas as vozes sejam ouvidas e consideradas no desenvolvimento destas tecnologias. Um dos nossos objetivos foi mostrar a importância da construção de um espaço mais democrático e participativo no qual as decisões relacionadas com as tecnologias de IA sejam informadas por uma compreensão mais ampla e contextualizada das questões sociais e éticas envolvidas.

Pensando nas mensagens centrais deste livro e olhando para o futuro, identificamos quatro áreas que merecem atenção e investigação académica adicionais.

Em primeiro lugar, o empoderamento de comunidades marginalizadas: Há uma necessidade de investigar estratégias mais eficazes para amplificar as vozes das comunidades marginalizadas no desenvolvimento e governação da IA, garantindo uma representação mais equitativa e inclusiva nos processos de tomada de decisão.

Em segundo lugar, a conscientização sobre a necessidade de aumentar a sensibilidade e o reconhecimento da diversidade cultural: Futuras pesquisas devem concentrar-se em explorar como incorporar de forma mais eficaz as diversas perspetivas culturais na conceção e implementação de tecnologias de IA, visando mitigar preconceitos culturais e estereótipos e promover soluções mais inclusivas e culturalmente sensíveis.

Em terceiro lugar, promover ações de envolvimento público na tomada de decisões relacionadas com a IA: É necessário investigar e desenvolver abordagens inovadoras para facilitar o envolvimento público na governação da IA, promovendo uma cultura de inovação responsável e alinhando os desenvolvimentos tecnológicos com valores e aspirações sociais mais amplos.

Em quarto lugar, é crucial consolidar a cidadania crítica digital (e não apenas a literacia digital, como proposto por governos e líderes da indústria de IA): Pesquisas futuras devem explorar como capacitar os cidadãos a tomar decisões informadas e participar ativamente dos debates sobre a implementação e regulamentação da IA.

Em suma, esperamos que este livro estimule uma reflexão mais ampla e profunda sobre os desafios da IA na sociedade, incentivando ações que promovam uma implementação ética e responsável dessa tecnologia em benefício de toda a humanidade.

Referências bibliográficas

Aas, K.F. (2011). 'Crimmigrant' bodies and bona fide travelers: Surveillance, citizenship and global governance. *Theoretical Criminology*, *15*(3), pp. 331-346. https://doi.org/10.1177/1362480610396643

Ada Lovelace Institute and The Alan Turing Institute (2023). *How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain*. Disponível em https://adalovelaceinstitute.org/report/public-attitudes-ai [Acesso a 24 de julho de 2024].

Ajana, B. (2013). *Governing through biometrics: The biopolitics of identity*. Palgrave Macmillan.

Alarie, B., Niblett, A., & Yoon, A. H. (2018). How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, *68*, pp. 106-124. https://doi.org/10.3138/utlj.2017-0052

Alfrink, K., Keller, I., Doorn, N., & Kortuem, G. (2022). Tensions in transparent urban AI: Designing a smart electric vehicle charge point. *AI & Society*, *38*(3), pp. 1049-1065. https://doi.org/10.1007/s00146-022-01436-9

Amann, J., Vayena, E., Ormond, K. E., Frey, D., Madai, V. I., & Blasimme, A. (2023). Expectations and attitudes towards medical artificial intelligence: A qualitative study in the field of stroke. *PLoS One*, *18*(1), e0279088. https://doi.org/10.1371/journal.pone.0279088

Amelung, N., & Machado, H. (2019). 'Bio-bordering' processes in the EU: De-bordering and re-bordering along transnational systems of biometric database technologies. *International Journal of Migration and Border Studies, 5*(4), pp. 392-408. https://doi.org/10.1504/IJMBS.2019.105813

Amelung, N., Granja, R., & Machado, H. (2020). *Modes of bio-bordering: The hidden (dis)integration of Europe*. Palgrave Pivot.

Amoore, L. (2013). *The politics of possibility: Risk and society beyond probability*. Duke University Press.

Aquino, Y. S. J., Carter, S. M., Houssami, N., Braunack-Mayer, A., Win, K. T., Degeling, C., Wang, L., & Rogers, W. A. (2023). Practical, epistemic and normative implications of algorithmic bias in healthcare artificial intelligence: A qualitative study of multidisciplinary expert perspectives. *Journal of Medical Ethics*, jme-2022-108850. https://doi.org/10.1136/jme-2022-108850

Aradau, C., & Blanke, T. (2022). *Algorithmic reason. The new government of self and other*. Oxford University Press. https://library.oapen.org/handle/20.500.12657/58142

Arbelaez, O. L., Lorenzini, G., Milford, S.R., Shaw, D., Elger, B.S., & Rost, M. (2024). Integrating ethics in AI development: A qualitative study. *BMC Medical Ethics*, *25*(1), 10. https://doi.org/10.1186/s12910-023-01000-0

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), pp. 59-64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J.B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*(2), pp. 134-143. https://doi.org/10.1038/s41562-019-0762-8

Babuta, A., & Oswald, M. (2021). Machine learning predictive algorithms and the policing of future crimes: Governance and oversight. In J.L.M. McDaniel, & K. Pease (Eds.). *Predictive policing and artificial intelligence* (pp. 214-236). Taylor & Francis. https://doi.org/10.4324/9780429265365-11

Baker, R. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), pp. 600-614. https://doi.org/10.1007/s40593-016-0105-0

Baker, R. (2019). Challenges for the future of educational data mining: The Baker Learning analytics prizes. *Journal of Educational Data Mining*, *11*(1), pp. 1-17. https://doi.org/10.5281/zenodo.3554745

Baker, R. (2021). Artificial intelligence in education: Bringing it all together. In *OECD digital education outlook 2021: Pushing the frontiers with artificial intelligence, blockchain, and robots* (pp. 43-54). OECD Publishing. https://doi.org/10.1787/f54ea644-en

Bakiner, O. (2023). The promises and challenges of addressing artificial intelligence with human rights. *Big Data & Society*, *10*(2). https://doi.org/10.1177/20539517231205476

Barad, K. (2003). Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs*, *28*(3), pp. 801-831. https://www.journals.uchicago.edu/doi/10.1086/345321

Bareis, J., & Katzenbach, C. (2022). Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values*, *47*(5), pp. 855-881. https://doi.org/10.1177/01622439211030007

Bastian, M., Helberger, N., & Makhortykh, M. (2021). Safeguarding the journalistic DNA: Attitudes towards the role of professional values in algorithmic news recommender designs. *Digit Journal*, *9*(6), pp. 835-863. https://doi.org/10.1080/21670811.2021.1912622

Beck, U. (1992). *Risk society: towards a new modernity*. Sage Publications.

Belpaeme, T., & Tanaka, F. (2021). Social robots as educators. In *OECD digital education outlook 2021: Pushing the frontiers with artificial intelligence, blockchain, and robots* (pp. 143-158). OECD Publishing. https://doi.org/10.1787/1c3b1d56-en

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity Press.

Benkler, Y. (2019). Don't let industry write the rules for AI. Nature, 569(7755), p. 161. https://doi.org/10.1038/ d41586-019-01413-1

Berger, P. (2001[1963]). *Perspectivas sociológicas. Uma visão humanística*. Petrópolis.

Berk, R. (2021). Artificial intelligence, predictive policing, and risk assessment for law enforcement. *Annual Review of Criminology*, *4*, pp. 209-237. https://doi.org/10.1146/annurev-criminol-051520-012342

Berry, C. R. (2023). Decrypting the gaze of electronic monitoring (EM): A comparative book review of Daems' Electronic monitoring and Gacek's portable prisons. Tagging offenders in a culture of surveillance. *Journal of Criminology*, *56*(2-3), pp. 359-367. https://doi.org/10.1177/26338076231173150

Bijker, W. E., Hughes, T. P., & Pinch, T. J. (Eds.). (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. MIT Press.

Blease, C., Locher, C., Leon-Carlyle, M., & Doraiswamy, M. (2020). Artificial intelligence and the future of psychiatry: Qualitative findings from a global physician survey. *Digital Health*, *6*, 2055207620968355. https://doi.org/10.1177/2055207620968355

Blount, K. (2024). Using artificial intelligence to prevent crime: Implications for due process and criminal justice. *AI & Society, 39*, pp. 359-368. https://doi.org/10.1007/s00146-022-01513-z

Borenstein J. & Arkin, R. (2017). Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior. *AI & Society 32(4)*, pp. 499-507. https://doi.org/10.1007/s00146-016-0684-1

Borgdorf, H., Peters, P., & Pinch, T. (2020). *Dialogues between artistic research and science and technology Studies*. Routledge.

Bowker, G., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics*, *47*(12), e3. https://doi.org/10.1136/medethics-2019-105860

Bringula, R. (2023). What do academics have to say about ChatGPT? A text mining analytics on the discussions regarding ChatGPT on research writing. *AI Ethics*. https://doi.org/10.1007/s43681-023-00354-w

Brown, N., Rappert, B., & Webster, A. (2017). Introducing contested futures: From looking into the future to looking at the future. In N. Brown, B. Rappert, & A. Webster (Eds.). *Contested futures: A sociology of prospective techno-science* (pp. 3-20). Routledge.

Campolo, A., & Crawford, K. (2020). Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society*, *6*, pp. 1-19. https://doi.org/10.17351/ests2020.277

Cassen-Weiss, D. (2019). Prisons and jails use artificial intelligence to monitor inmate phone calls. *ABA Journal*. Disponível em https://www.abajournal.com/news/article/prisons-and-jails-use-artificial-intelligence-to-monitor-inmate-phone-calls [Acesso a 24 de julho de 2024].

CEPEJ [Comissão Europeia para a Eficiência na Justiça] (2018). *European ethical charter on the use of artificial intelligence (AI) in judicial systems and their environment*. Disponível em https://rm.coe.int/ethical--charter-en-for-publication-4-december-2018/16808f699c [Acesso a 24 de julho de 2024].

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2020). Neural legal judgment prediction in English. In *Proceedings of the Conference ACL 2019- 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4317-4323). https://doi.org/10.18653/v1/p19-1424

Choung, H., David, P., & Ross, A. (2023). Trust and ethics in AI. *AI & Society*, *38*(2), pp. 733-745. https://doi.org/10.1007/s00146-022-01473-4

Clarke, A. (2005). *Situational analysis: Grounded theory after the postmodern turn*. Sage Publications.

Cole, S., & Lynch, M. (2006). The social and legal construction of suspects. *Annual Review of Law and Social Science*, *2*(1), pp. 39-60. https://doi.org/10.1146/annurev.lawsocsci.2.081805.110001

Comissão Europeia (2020). *LIVRO BRANCO sobre a inteligência artificial. Uma abordagem europeia virada para a excelência e a confiança*. Disponível em https://op.europa.eu/pt/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a1 [Acesso a 24 de julho de 2024].

Comissão Europeia (2021). *Proposta de Regulamento do Parlamento Europeu e do Conselho que estabelece regras harmonizadas em matéria de inteligência artificial (regulamento inteligência artificial) e altera determinados atos legislativos da União*. Disponível em https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A52021PC0206 [Acesso a 24 de julho de 2024].

Comissão Europeia [European Commission] (2022). *Directorate-General for Education, Youth, Sport and Culture Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*. Publications Office of the European Union. https://data.europa.eu/doi/10.2766/153756 [Acesso a 24 de julho de 2024].

Conselho Europeu [Council of Europe Commissioner for Human Rights] (2019). *Unboxing artificial intelligence: 10 steps to protect human rights*. Disponível em https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64 [Acesso a 24 de julho de 2024].

Crawford, K. (2024 [2021]). *Atlas da IA*. *Poder, política e custos planetários da inteligência artificial*. Relógio D'Água.

Creswell, J, W. (2015). *A concise introduction to mixed methods research*. Sage Publications.

Cresswell, K., Cunningham-Burley, S., & Sheikh, A. (2018). Health care robotics: Qualitative exploration of key challenges and future directions. *Journal of Medical Internet Research*, *20*(7), e10410. https://doi.org/10.2196/10410

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*, 22. https://doi.org/10.1186/s41239-023-00392-8

D'Mello, S. (2021). Improving student engagement in and with digital learning technologies. In *OECD digital education outlook 2021: Pushing the frontiers with Artificial Intelligence, Blockchain, and Robots* (pp. 79-104). OECD Publishing. https://doi.org/10.1787/8a451974-en

De Graaf, M. M. A., Hindriks, F. A., & Hindriks, K. V. (2022). Who wants to grant robots rights? *Frontiers in robotics and AI*, *8*, 781985. https://doi.org/10.3389/frobt.2021.781985

de la Bellacasa, M. P. (2011). Matters of care in technoscience: Assembling neglected things. *Social Studies of Science*, *41*(1), pp. 85-106. https://doi.org/10.1177/0306312710380301

de Oliveira, L.F., da Silva Gomes, A., Enes, Y., Branco, T.V.C., Pires, R.P., Bolzon, A., & Demo, G. (2022). Path and future of artificial intelligence in the field of justice: A systematic literature review and a research agenda. *SN Social Sciences*, *2*(180). https://doi.org/10.1007/s43545-022-00482-w

de Sousa, W.G., Fidelis, R.A., de Souza Bermejo, P.H., da Silva Gonçalo, A.G., & de Souza Melo, B. (2022). Artificial intelligence and speedy trial in the judiciary: Myth, reality or need? A case study in the Brazilian Supreme Court (STF). *Government Information Quarterly*, *39*(1), 101660. https://doi.org/10.1016/j.giq.2021.101660

de Vries, P., & Schinkel, W. (2019). Algorithmic anxiety: Masks and camouflage in artistic imaginaries of facial recognition algorithms. *Big Data & Society*, *6*(1). https://doi.org/10.1177/2053951719851532

DeFalco, J., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S., & Lester, J.C. (2017). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, *28*(2), pp. 152-193. https://doi.org/10.1007/s40593-017-0152-1

Duke, S. A. (2022). Deny, dismiss and downplay: Developers' attitudes towards risk and their role in risk creation in the field of healthcare-AI. *Ethics and Information Technology*, *24*(1). https://doi.org/10.1007/s10676-022-09627-0

Dupont, J., Wirde, S., & Ali, V. (2023). *What does the public think about AI?* Disponível em https://publicfirst.co.uk/ai/ [Acesso a 24 de julho de 2024].

Elish, M. C., & Boyd, D. (2018). Situating methods in the magic of Big Data and AI. *Communication Monographs*, *85*(1), pp. 57-80. https://doi.org/10.1080/03637751.2017.1375130

eu-LISA (2020). *Artificial intelligence in the operational management of large-scale IT systems – Research and technology monitoring report – Perspectives for eu-LISA*. Publications Office. Disponível em https://data.europa.eu/doi/10.2857/58386 [Acesso a 24 de julho de 2024].

eu-LISA e EUROJUST (2022). *Artificial intelligence supporting cross-border crime cooperation in criminal justice*. Disponível em https://www.eurojust.europa.eu/publication/artificial-intelligence-supporting-cross-border-cooperation-criminal-justice [Acesso a 2 de julho de 2024].

Felt, U., & Wynne, B. (2007). *Taking European knowledge society seriously. Report of the expert group on science and governance*. European Commission Directorate-General Research. Disponível em https://op.europa.eu/en/publication-detail/-/publication/5d0e77c7-2948-4ef5-aec7-bd18efe3c442 [Acesso a 24 de julho de 2024].

Ferretti, T. (2022). An institutionalist approach to AI ethics: Justifying the priority of government regulation over self-regulation. *Moral Philosophy and Politics, 9*(2), pp. 239-265. https://doi.org/10.1515/mopp-2020-0056

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, *21*(5), e13216. https://doi.org/10.2196/13216

Fórum Económico Mundial [World Economic Forum] (2023). *Adopting AI responsibly: Guidelines for procurement of AI solutions by the private sector*. Disponível em https://www.weforum.org/reports/adopting-ai-responsibly-guidelines-for-procurement-of-ai-solutions-by-the-private-sector/ [Acesso a 24 de julho de 2024].

Foucault, M. (1972). *The archeology of knowledge and the discourse on language*. Harper.

Foucault, M. (1973). *The order of things: An archeology of the human sciences*. Vintage/Random House.

Galligan, C., Rosenfeld, H., Kleinman, M., & Parthasarathy, S. (2020). *Cameras in the classroom. Facial recognition technology in schools (Technology Assessment Report)*. Gerald R. Ford School of Public Policy. Science, Technology, and Public Policy, University of Michigan. Disponível em https://stpp.fordschool.umich.edu/sites/stpp/files/uploads/file-assets/cameras_in_the_classroom_full_report.pdf [Acesso a 24 de julho de 2024].

GatesNotes (2023). *HISTORY HELPS. The risks of AI are real but manageable*. Disponível em https://www.gatesnotes.com/The-risks-of-AI-are-real-but-manageable [Acesso a 24 de julho de 2024].

Giddens, A. (1990). *The consequences of modernity*. Polity Press.

Gill, N., Singleton, V., & Waterton, C. (2017). The politics of policy practices. *Sociological Review*, *65*, pp. 3-19. https://doi.org/10.1177/0081176917710429

Gomes, C., Duarte, M., Fernando, P., Ferreira, A., Dias, J., & Campos, A. (2013). *Contextos e desafios da transformação das magistraturas: Contributos dos estudos sociojurídicos*. Vida Económica. Disponível em https://hdl.handle.net/10316/79911 [Acesso a 24 de julho de 2024].

Gross, N., & Geiger, S. (2023). Choreographing for public value in digital health? *Big Data & Society*, *10*(2). https://doi.org/10.1177/20539517231220622

GPAN IA [AI HLEG] (2019a). *A definition of AI: Main capabilities and disciplines*, European Commission. Disponível em: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341 [Acesso a 24 de julho de 2024].

GPAN IA [AI HLEG] (2019b). *Ethics guidelines for trustworthy AI*. European Commission. Disponível em https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Acesso a 24 de julho de 2024].

GPAN IA [AI HLEG] (2019c). *Policy and investment recommendations for trustworthy AI*. European Comission. Disponível em https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence [Acesso a 24 de julho de 2024].

GPAN IA [AI HLEG] (2020). *The assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment*. European Comission. Disponível em https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment [Acesso a 24 de julho de 2024].

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds & Machines, 30*, pp. 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hallowell, N., Badger, S., Sauerbrei, A., Nellaker, C., & Kerasidou, A. (2022). "I don't think people are ready to trust these algorithms at face value": Trust and the use of machine learning algorithms in the diagnosis of rare disease. *BMC Medical Ethics*, *23*(112). https://doi.org/10.1186/s12910-022-00842-4

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism: Clinical and Experimental*, *69S*, pp. S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

Héder, M. (2021). AI and the resurrection of technological determinism. *Információs Társadalom XXI*, *2*, pp. 119-130. https://dx.doi.org/10.22503/inftars.XXI.2021.2.8

Heilinger, J. C. (2022). The ethics of AI ethics. A constructive critique. *Philosophy & Technology*, *35*(61). https://doi.org/10.1007/s13347-022-00557-9

Henriksen, A., & Blond, L. (2023). Executive-centered AI? Designing predictive systems for the public sector. *Social Studies of Science*, *53*(5), pp. 738-760. https://doi.org/10.1177/03063127231163756

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S., Santos, O., Rodrigo, M., Cukurova, M., Bittencourt, I., & Koedinger, K. (2022a). Ethics of AI in education: towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *32*, pp. 504-526. https://doi.org/10.1007/s40593-021-00239-1

Holmes, W., Persson, J., Chounta, I.A., Wasson, B., & Dimitrova, V. (2022b). *Artificial intelligence and education. A critical view through the lens of human rights, democracy and the rule of law*. Education Department Council of Europe: Council of Europe Publishing. Disponível em https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd [Acesso a 24 de julho de 2024].

Isbanner, S., & O'Shaughnessy, P. (2022). The adoption of artificial intelligence in health care and social services in Australia: Findings from a methodologically innovative national survey of values and attitudes (the AVA-AI Study). *Journal of Medical Internet Research*, *24*(8), e37611. https://doi.org/10.2196/37611

Jenkins, S., & Draper, H. (2015). Care, monitoring, and companionship: Views on care robots from older people and their carers. *International Journal of Social Robotics*, *7*(5), pp. 673-683. https://doi.org/10.1007/s12369-015-0322-y

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, pp. 389-399. https://doi.org/10.1038/s42256-019-0088-2

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., & Shestakofsky, B. (2021). Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius*, *7*. https://doi.org/10.1177/2378023121999581

Karimian, G., Petelos, E., & Evers, S.M.A.A. (2022). The ethical issues of the application of artificial intelligence in healthcare: A systematic scoping review. *AI and Ethics*, *2*, pp. 539-551. https://doi.org/10.1007/s43681-021-00131-7

Kerr, A., Hill, R. L., & Till, C. (2018). The limits of responsible innovation: Exploring care, vulnerability and precision medicine. *Technology in Society*, *52*, pp. 24-31. https://doi.org/10.1016/j.techsoc.2017.03.004

Kieslich, K., Keller, B., & Starke, C. (2022). Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society*, *9*(1), pp. 1-15. https://doi.org/10.1177/20539517221092956

Knox, J. (2020). Artificial intelligence and education in China. *Learning, Media and Technology*, *45*(3), pp. 298-311. https://doi.org/10.1080/17439884.2020.1754236

Konrad, K. (2006). The social dynamics of expectations: The interaction of collective and actor-specific expectations on electronic commerce and interactive television. *Technology Analysis & Strategic Management, 18*(3-4), pp. 429-444. https://doi.org/10.1080/09537320600777192

Lagerkvist, A., Tudor, M., Smolicki, J., Ess, C.M., Lundström, J.E., & Rogg, M. (2022). Body stakes: An existential ethics of care in living with biometrics and AI. *AI & Society*. https://doi.org/10.1007/s00146-022-01550-8

Laï, M. C., Brian, M., & Mamzer, M. F. (2020). Perceptions of artificial intelligence in healthcare: Findings from a qualitative survey study among actors in France. *Journal of Translational Medicine*, 18(14). https://doi.org/10.1186/S12967-019-02204-Y

Latimer, J., & Gomez, D. L. (2019). Intimate entanglements: Affects, more-than-human intimacies and the politics of relations in science and technology. *Sociological Review*, *67*(2), pp. 247-263. https://doi.org.10.1177/0038026119831623

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Harvard University Press.

Lee, C. H., Gobir, N., Gurn, A., & Soep, E. (2022). In the black mirror: Youth investigations into artificial intelligence. *ACM Transactions on Computing Education*, *22*(3), pp. 1-25. https://doi.org/10.1145/3484495

Leese, M. (2022). Fixing state vision: Interoperability, biometrics, and identity management in the EU. *Geopolitics*, *27*(1), pp. 113-133. https://doi.org/10.1080/14650045.2020.1830764

Lindén, L., & Lydahl, D. (2021). Editorial: Care in STS. *Nordic Journal of Science and Technology Studies*, *9*(1), pp. 3-12. https://doi.org/10.5324/njsts.v9i1.4000

Lindgren, S., & Holmström, J. (2020). A social science perspective on artificial intelligence: Building blocks for a research agenda. *Journal of Digital Social Research*, *2*(3), pp. 1-15. https://doi.org/10.33621/jdsr.v2i3.65

Liu, Z. (2021). Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass*, *15*(3), pp. 1-13. https://doi.org/10.1111/soc4.12851

Machado, H., & Granja, R. (2020). *Forensic genetics in the governance of crime*. Palgrave Pivot. https://doi.org/10.1007/978-981-15-2429-5

Machado, H., Granja, R., & Amelung, N. (2020). Constructing suspicion through forensic DNA databases in the EU. The views of the Prüm professionals. *The British Journal of Criminology*, *60*(1), pp. 141-159. https://doi.org/10.1093/bjc/azz057

Machado, H., Silva, S., & Neiva, L. (2023). Publics' views on ethical challenges of artificial intelligence: A scoping review. *AI & Ethics*. https://doi.org/10.1007/s43681-023-00387-1

Macq, H., Tancoigne, E., & Strasser, B. J. (2020). From deliberation to production: public participation in science and technology policies of the European Commission (1998–2019). *Minerva*, *58*(4), pp. 489-512. https://doi.org/10.1007/s11024-020-09405-6

Manning, C. D. (2015). Computational linguistics and deep learning. *Computational Linguistics*, *41*(4), pp. 701-707.

Martin, A., Myers, N., & Viseu, A. (2015). The politics of care in technoscience. *Social Studies of Science*, *45*(5), pp. 625-641. https://doi.org/10.1177/0306312715602073

Marx, L. (2000). *The machine in the garden: Technology and the pastoral ideal in America*. Oxford University Press.

Matzner, T. (2016). Beyond data as representation: The performativity of Big Data in surveillance. *Surveillance and Society*, *14*(2), pp. 197-210. https://doi.org/10.24908/ss.v14i2.5831

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mayor, A. (2018). *Gods and robots: Myths, machines, and ancient dreams of technology*. Princeton University Press.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *Proposal for the Dartmouth summer research project on artificial intelligence*. Disponível em http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf [Acesso a 24 de julho de 2024].

McCradden, M. D., Baba, A., Saha, A., Ahmad, S., Boparai, K., Fadaiefard, P., & McCradden, M. D., Sarker, T., & Paprica, P. A. (2020a). Conditionally positive: A qualitative study of public perceptions about using health data for artificial intelligence research. *BMJ Open*, *10(*10), e039798. https://doi.org/10.1136/bmjopen-2020-039798

McCradden, M. D., Baba, A., Saha, A., Ahmad, S., Boparai, K., Fadaiefard, P., & Cusimano, M. D. (2020b). Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: A qualitative study. *CMAJ Open*, *8*(1), pp. E90-E95. https://doi.org/10.9778/cmajo.20190151

McCue, C. (2014). *Data mining and predictive analysis: Intelligence gathering and crime analysis*. Butterworth-Heinemann.

Mead, G. (1934/1962). *Mind, self and society*. Chicago University Press.

Meijer, A., & Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration*, *42*(12), pp. 1031-1039. https://doi.org/10.1080/01900692.2019.1575664

Merriem-Webster. (2023). *Artificial intelligence*. Disponível em https://www.merriam-webster.com/dictionary/artificial%20intelligence#dictionary-entry-1 [Acesso a 30 de dexembro de 2023].

Metzinger, T. (2019). Ethics washing made in Europe. *Tagesspiegel*, 8 April. Disponível em https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html [Acesso a 24 de julho de 2024].

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), pp. 501-507. https://doi.org/10.1038/s42256-019-0114-4

Molenaar, I. (2021). Personalisation of learning: Towards hybrid human-AI learning technologies. In *OECD digital education outlook 2021: Pushing the frontiers with Artificial Intelligence, Blockchain, and Robots* (pp. 57-77). OECD Publishing. https://doi.org/10.1787/2cc25e37-en

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine*, *260*, 113172. https://doi.org/10.1016/j.socscimed.2020.113172

Mosco, V. (2005). *The digital sublime: Myth, power, and cyberspace*. MIT Press.

Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: assumptions, evaluation, and accountability. *Policing and Society*, *28*(7), pp. 806-822. https://doi.org/10.1080/10439463.2016.1253695

Munn, L. (2022). The uselessness of AI ethics. *AI Ethics*, *3*, pp. 869-877. https://doi.org/10.1007/s43681-022-00209-w

Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics*, *22*(1), 14. https://doi.org/10.1186/s12910-021-00577-8

Natale, S., & Ballatore, A. (2017). Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence: The International Journal of Research into New Media Technologies*, *26*(1), pp. 3-18. https://doi.org/10.1177/1354856517715164

Neiva, L., Machado, H., & Silva, S. (2023). The views about Big Data among professionals of police forces: A scoping review of empirical studies. *International Journal of Police Science & Management*, *25*(2), pp. 208--220. https://doi.org/10.1177/14613557231166225

Newman, J. C. (2020). *Decision points in AI governance: Three case studies explore efforts to operationalize AI principles*. UC Berkeley, Center for Long-Term Cybersecurity.

Nichol, A.A., Halley, M.C., Federico, C.A., Cho, M.K., & Sankar, P.L. (2023). Not in my AI: Moral engagement and disengagement in health care AI development. *Pacific Symposium on Biocomputing*, *28*, pp. 496-506.

Norton, A. A. (2013). Predictive policing: The future of law enforcement in the Trinidad and Tobago police service (TTPS). *International Journal of Computer Applications*, *62*(4), pp. 32-36. https://doi.org/10.5120/10070-4680

Nowotko, P. (2021). AI in judicial application of law and the right to a court. *Procedia Computer Science*, *192*, pp. 2220-2228. https://doi.org/10.1016/j.procs.2021.08.235

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), pp. 447-453. https://doi.org/10.1126/science.aax2342

Ochigame, R. (2019). The invention of "ethical AI": How big tech manipulates academia to avoid regulation. *Intercept*. Disponível em https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/ [Acesso a 24 de julho de 2024].

OCDE [OECD] (2019). *Recommendation of the Council on artificial intelligence (OECD/LEGAL/0449)*. Disponível em https://oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf [Acesso a 24 de julho de 2024].

OCDE [OECD] (2021). *OECD Digital Education Outlook 2021: Pushing the frontiers with artificial intelligence, blockchain, and robots*. OECD Publishing. Disponível em https://read.oecd-ilibrary.org/education/oecd-digital-education-outlook-2021_589b283f-en#page1 [Acesso a 24 de julho de 2024].

OCDE [OECD] (2023). *Recommendation of the Council on artificial intelligence*. Disponível em https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 [Acesso a 24 de julho de 2024].

Oomen, J., Hoffman, J., & Hajer, M. A. (2022). Techniques of futuring: On how imagined futures become socially performative. *European Journal of Social Theory*, *25*(2), pp. 252-270. https://doi.org/10.1177/1368431020988826

Organização Mundial de Saúde [World Health Organisation] (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. Disponível em https://www.who.int/publications/i/item/9789240029200 [Acesso a 24 de julho de 2024].

Pantazis, C., & Pemberton, S. (2009). From the "old" to the "new" suspect community: Examining the impacts of recent UK counter-terrorist legislation. *British Journal of Criminology*, *49*(5), pp. 646-666. https://doi.org/10.1093/bjc/azp031

Pasquinelli, M. (Ed.) (2015). *Alleys of your mind: Augmented intelligence and its traumas*. Meson Press.

Perry, W. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Phan, T., Goldenfein, J., Mann, M., & Kuch, D. (2022). Economies of virtue: The circulation of 'ethics' in big tech. *Science as Culture*, *31*(1), pp. 121-135. https://doi.org/10.1080/09505431.2021.1990875

Ploug, T., Sundby, A., Moeslund, T. B., & Holm, S. (2021). Population preferences for performance and explainability of artificial intelligence in health care: Choice-based conjoint survey. *Journal of Medical Internet Research*, *23*(12), e26611. https://doi.org/10.2196/26611

Puolakka, P., & Van De Steene, S. (2021). Artificial intelligence in prisons in 2030: An exploration on the future of AI in prisons. *Advancing Corrections Journal*, *11*, pp. 128-138. Disponível em https://rm.coe.int/ai-in-prisons-2030-acjournal/1680a40b83 [Acesso a 24 de julho de 2024].

Rafanelli, L. M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517221080676

Reeve, O., Colom, A., & Modhvadia, R. (2023). *What do the public think about AI? Understanding public attitudes and how to involve the public in decision-making about AI*. Ada Lovelace Institute. Disponível em https://www.adalovelaceinstitute.org/evidence-review/what-do-the-public-think-about-ai/#methodology-27 [Acesso a 24 de julho de 2024].

Resseguier, A., & Rodrigues, R. (2021). Ethics as attention to context: Recommendations for the ethics of artificial intelligence. *Open Research Europe*, *1*, 27. https://doi.org/10.12688/openreseurope.13260.2

Roberge, J., Senneville, M., & Morin, K. (2020). How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720919968

Roche, C., Wall, P. J., & Lewis, D. (2022). Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI Ethics*. https://doi.org/10.1007/s43681-022-00218-9

Rogers, W. A., Draper, H., & Carter, S. M. (2021). Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues. *Bioethics, 36*(4), pp. 624-633. https://doi.org/10.1111/bioe.12885

Ronsin, X., Lampos, V., & Maîtrepierre, A. (2018). Annex I – In-depth study on the use of AI in judicial systems, notably AI applications processing judicial decisions and data. In *European ethical charter on the use of artificial intelligence (AI) in judicial systems and their environment*. Disponível em https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c [Acesso a 24 de julho de 2024].

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Santos, B. S. (1995). *Os tribunais nas sociedades contemporâneas: O caso Português*. Afrontamento.

Sarangi, S. & Sharma, P. (2019). *Artificial intelligence. Evolution, ethics and public policy*. Routledge.

Schiff, D. (2020). Out of the laboratory and into the classroom: The future of artificial intelligence in education. *AI & Society*, *36*, pp. 331–348. https://doi.org/10.1007/s00146-020-01033-8

Sejnowski, T. J. (2018). *The deep learning revolution*. MIT Press.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: Definition and background. In *Mission AI: The new system technology*. Springer. https://doi.org/10.1007/978-3-031-21448-6_2

Sieber, R., Brandusescu, A., Adu-Daako, A., & Sangiambut, S. (2024). Who are the publics engaging in AI? *Public Understanding of Science*. https://doi.org/10.1177/09636625231219853

Sigfrids, A., Leikas, J., Salo-Pöntinen, H., & Koskimies, E. (2023). Human-centricity in AI governance: a systemic approach. *Frontiers in Artificial Intelligence, 6*, p. 976887. https://doi.org/10.3389/frai.2023.976887

Sismondo, S. (2012). *An introduction to science and technology studies* (2nd ed.). Wiley-Blackwell.

Søraa, R. (2023). *AI for diversity*. Routledge.

Steinhoff, J. (2023). AI ethics as subordinated innovation network. *AI & Society*. https://doi.org/10.1007/s00146-023-01658-5

Strauss, A. (1978). A social worlds perspective. *Studies in Symbolic Interaction*, 1, pp. 119-128.

Su, A. (2022). The promise and perils of international human rights law for AI governance. *Law, Technology and Humans*, *4*(2), pp. 166-182. https://doi.org/10.5204/lthj.2332

Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2). https://doi.org/10.1177/20539517231206794

Tantikul, T. (2024). Judicial indifference in criminal sentencing: Explaining inequality of the Thai Fines. *The British Journal of Criminology*, *64*(2), pp. 343-360. https://doi.org/10.1093/bjc/azad033

Ulnicane, I. (2022). Emerging technology for economic competitiveness or societal challenges? Framing purpose in artificial intelligence policy. *Global Public Policy and Governance*, *2*, pp. 326-345. https://doi.org/10.1007/s43508-022-00049-8

Ulnicane, I., Okaibedi, E.D., Knight, W., Ogoh, G., & Stahl, B. (2021a). Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews*, *46*(1-2), pp. 71-93. https://doi.org/10.1080/03080188.2020.1840220

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C. & Wanjiku, W. (2021b). Framing governance for a contested emerging technology: Insights from AI policy. *Policy and Society*, *40*(2), pp. 158-177. https://doi.org/10.1080/14494035.2020.1855800

UNESCO (2021). *Recomendação sobre a ética da inteligência artificial*. Disponível em https://unesdoc.unesco.org/ark:/48223/pf0000381137_por [Acesso a 24 de julho de 2024].

União Europeia (2024). *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 (Regulamento da Inteligência Artificial)*. Disponível em https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=OJ:L_202401689 [Acesso a 24 de julho de 2024].

Valles-Peris, N., Barat-Auleda, O., & Domenech, M. (2021). Robots in healthcare? What patients say. *International Journal of Environmental Research and Public Health*, *18*(9933). https://doi.org/10.3390/ijerph18189933

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), pp. 197-208. https://doi.org/10.24908/ss.v12i2.4776

van Lente, H. (2016). Forceful futures: From promise to requirement. In N. Brown, B. Rappert, & A. Webster (Eds.). *Contested futures: A sociology of prospective techno-science* (pp. 43-64). Routledge.

van Maanen, G. (2022). AI ethics, ethics washing, and the need to politicize data ethics. *Digital Society*, *1*(9). https://doi.org/10.1007/s44206-022-00013-3

Vaughn, L. M., & Jacquez, F. (2020). Participatory research methods. Choice points in the research process. *Journal of Participatory Research Methods*, *1*(1). https://doi.org/10.35844/001c.13244

Villegas-Galaviz, C., & Martin, K. (2023). Moral distance, AI, and the ethics of care. *AI & Society*. https://doi.10.1007/s00146-023-01642-z

Vincent-Lancrin, S., & van der Vlies, R. (2020). Trustworthy artificial intelligence (AI) in education: Promises and challenges. *OECD Education Working Papers*, No. 218. https://doi.org/10.1787/a6c90fa9-en

Wang, S., Bolling, K., Mao, W., Reichstadt, J., Jeste, D., Kim, H. C., & Nebeker, C. (2019). Technology to support aging in place: Older adults' perspectives. *Healthcare (Basel)*, *7*(2), 60. https://doi.org/10.3390/healthcare7020060

Weingart, P., Joubert, M., & Connoway, K. (2021). Public engagement with science. Origins, motives and impact in academic literature and science policy. *PloS One*, *16*(7), e0254201. https://doi.org/10.1371/journal.pone.0254201

West, S. M. (2019). Data capitalism: Redefining the logics of surveillance and privacy. *Business & Society*, *58*(1), pp. 20-41. https://doi.org/10.1177/0007650317718185

Whitelaw, S., Mamas, M. A., Topol, E., & Van Spall, H. G. C. (2020). Applications of digital technology in COVID-19 pandemic planning and response. *The Lancet. Digital health*, *2*(8), pp. e435-e440. https://doi.org/10.1016/S2589-7500(20)30142-4

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. Nuffield Foundation. Disponível em https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data--and-AI-report-Nuffield-Foundat.pdf [Acesso a 24 de julho de 2024].

Willems, J., Schmidthuber, L., Vogel, D., Ebinger, F., & Vanderelst, D. (2022). Ethics of robotized public services: The role of robot design and its actions. *Government Information Quarterly*, *39*(2), 101683. https://doi.org/10.1016/j.giq.2022.101683

Willems, J., Schmid, M. J., Vanderelst, D., Vogel, D., & Ebinger, F. (2023). AI-driven public services and the privacy paradox: Do citizens really care about their privacy? *Public Management Review*, *25*(11), pp. 2116-2134. https://doi.org/10.1080/14719037.2022.2063934

Williams, M.L., Burnap, P., & Sloan, L. (2017). Crime sensing with Big Data: The affordances and limitations of using open source communications to estimate crime patterns. *The British Journal of Criminology*, *57*(2), pp. 320–340. https://doi.org/10.1093/bjc/azw031

Wilson, C. (2022). Public engagement and AI: A values analysis of national strategies. *Government Information Quartely*, *39*(1), 101652. https://doi.org/10.1016/j.giq.2021.101652

Wilson, D. (2020). Predictive policing management: A brief history of patrol automation. *New Formations: A Journal of Culture/Theory/Politics*, 98, pp. 139-155. https://www.muse.jhu.edu/article/747025

Wooldrige, M. (2021). *A brief history of artificial intelligence: What it is, where we are, and where we are going*. Flatiron Books.

Woolgar, S. (1985). Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology*, *19*(4), pp. 557-572. https://doi.org/10.1177/0038038585019004005

Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021). Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, *2*, pp. 79-84. https://doi.org/10.1016/j.aiopen.2021.06.003

Yolgörmez, C. (2021). Machinic encounters: A relational approach to the sociology of AI. In J. Roberge, & M. Castelle (Eds.). *The cultural life of machine learning: An incursion into AI critical studies* (pp. 143-166). Palgrave Macmillan.

Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, *16*(3), pp. 1-16. https://doi.org/10.1111/soc4.12962

Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. *ERA Forum*, *20*, pp. 567-583. https://doi.org/10.1007/s12027-020-00602-0

Zivani, E., & Mahlangu, G. (2022). Digital prison rehabilitation and successful re-entry into a digital society: A systematic literature review on the new reality on prison rehabilitation. *Cogent Social Sciences*, 8(1). https://doi.org/10.1080/23311886.2022.2116809

Zuboff, S. (2015). *The age of surveillance capitalism. The fight for a human future at the new frontier of power*. Profile Books.

Glossário

Algoritmo: Um processo ou conjunto de regras a serem seguidas em cálculos ou outras operações de resolução de problemas, neste caso, por um computador.

Análise de aprendizagem: Envolve a medição, a recolha, a análise e a comunicação de dados sobre os alunos e os seus contextos, com o objetivo de compreender e otimizar a aprendizagem e os ambientes em que esta ocorre.

Análise preditiva: Utilização de algoritmos estatísticos e de técnicas de aprendizagem automática para fazer previsões sobre o futuro utilizando dados atuais e históricos.

Aprendizagem da máquina (ou aprendizagem automática) (*machine learning*): Refere-se a um subcampo da IA que se concentra no desenvolvimento de algoritmos e modelos que permitem a um sistema aprender padrões a partir de dados, sem ser explicitamente programado. O objetivo principal é capacitar as máquinas a tomar decisões ou realizar tarefas sem intervenção humana constante, com base na análise de dados e na identificação de padrões. Existem vários tipos de algoritmos de aprendizagem da máquina, sendo os mais comuns a aprendizagem supervisionada, não supervisionada e por reforço. Este tipo de desenvolvimento de IA é aplicado numa variedade de áreas, como reconhecimento de padrões, processamento de linguagem natural, visão computacional, diagnósticos médicos, recomendações de produtos, entre outros.

Aprendizagem não supervisionada (*unsupervised learning*): O algoritmo é alimentado com um conjunto de dados não rotulados. O objetivo é descobrir padrões e estruturas nos dados sem a orientação explícita dos rótulos de saída. Os algoritmos de aprendizagem não supervisionada podem ser usados para tarefas como *clustering* (agrupamento de dados semelhantes) ou redução de dimensionalidade (processo de redução do número de variáveis ou características de um conjunto de dados, mantendo ao mesmo tempo o máximo de informações relevantes possível, tornando-o mais fácil entender, visualizar e processar).

Aprendizagem por reforço (*reinforcement learning*): O agente de aprendizagem (a máquina) interage com um ambiente dinâmico, tomando decisões sequenciais em função do *feedback* que recebe na forma de recompensas ou penalidades. O objetivo é aprender uma estratégia que maximize a soma esperada de recompensas ao longo do tempo. Diferente da aprendizagem supervisionada, não são facultados pares de entrada-saída e a máquina precisa de explorar o ambiente e aprender com a experiência.

Aprendizagem profunda (*deep learning*): Refere-se ao treino de modelos computacionais chamados de redes neuronais artificiais para realizar tarefas diretamente a partir dos dados. Este tipo de aprendizagem é chamada de "profunda" porque envolve o uso de múltiplas camadas de unidades de processamento, chamadas de "neurónios", para aprender representações complexas dos dados. É sobretudo aplicada quando há grandes quantidades de dados disponíveis e a complexidade das relações nos dados é alta. Exemplos de aplicação de aprendizagem profunda são os seguintes: Reconhecimento de imagens (classificando imagens em

categorias ou identificando e delimitando objetos); reconhecimento e processamento de linguagem natural (converter áudios em texto, traduzir textos de uma língua para outra, determinar os sentimentos expressos em textos); jogos de estratégia (têm sido usados para treinar máquinas capazes de superar humanos em jogos complexos como o *Go* ou *StarCraft*); medicina e diagnóstico (identificar anomalias em exames de imagem e auxiliar na interpretação de registos médicos); geração de conteúdo (música, texto, representações visuais); previsão climática e análise temporal financeira, entre muitas outras.

Aprendizagem supervisionada (*supervised learning*): O algoritmo é treinado com um conjunto de dados rotulados, onde cada exemplo de treino consiste num par de entrada e correspondente saída desejada. O objetivo é aprender uma função que mapeia as entradas para as saídas. Durante o treino, o modelo faz previsões com base nas entradas e compara com as saídas reais, ajustando os seus parâmetros iterativamente para minimizar a diferença entre as previsões e os rótulos reais. Uma vez treinado, o modelo pode fazer previsões precisas para novos dados.

Atuadores (*actuators*): Dispositivos ou componentes que permitem um sistema interagir com o ambiente. São responsáveis por converter informação digital em ações físicas. Os atuadores não são exclusivos da IA, mas desempenham um papel fundamental em campos como a robótica e automação. Por exemplo, na robótica, os atuadores são responsáveis por realizar movimentos com base nas instruções fornecidas pelo sistema de IA. Esses movimentos podem envolver qualquer coisa, desde o controlo das articulações de um braço robótico até ao controlo das rodas de um robô móvel.

Automação: Um sistema informático que pode executar tarefas sem necessitar de supervisão humana contínua é descrito como autónomo.

Big Data: Conjuntos de dados extremamente grandes e complexos que desafiam as capacidades tradicionais de processamento de dados. Geralmente, esses conjuntos de dados são caracterizados por três principais aspetos (3 V's): Volume (grandes quantidades de dados gerados), velocidade (com o fluxo de dados em tempo real, os dados podem ser gerados a velocidades incrivelmente elevadas) e variedade (dados estruturados tradicionais como bases de dados a dados semi-estruturados e não estruturados como vídeos, textos, etc.).

Chatbot (originalmente *chatterbot*): Aplicação de *software* ou *interface* da Internet que pretende imitar a conversação humana através de interações de texto ou voz. Utilizam sistemas de IA capazes de manter uma conversa com um utilizador em linguagem natural e de simular a forma como um ser humano se comportaria como parceiro de conversação.

ChatGPT: Modelo de linguagem de IA desenvolvido pela OpenAI, uma organização de pesquisa em IA com sede nos Estados Unidos, fundada em dezembro de 2015, baseado na arquitetura GPT-3.5, que é a terceira geração do modelo GPT (*Generative Pre-trained Transformer*). GPT-3.5 é treinado com uma enorme quantidade de texto recolhido da Internet, o que lhe confere a capacidade de gerar texto em uma

ampla variedade de estilos e contextos. Foi projetado para gerar respostas com base em padrões estatísticos, que são coerentes e relevantes para uma ampla gama de perguntas e estímulos fornecidos pelos utilizadores.

Internet das Coisas (IoT, do inglês *Internet of Things*): Refere-se a uma rede de objetos físicos, dispositivos, veículos, prédios e outros itens incorporados com sensores, *software* e conectividade de rede, permitindo a recolha e troca de dados. Esses objetos podem interagir entre si, muitas vezes sem a necessidade de intervenção humana. O objetivo da IoT é criar um ambiente onde os objetos do mundo real possam estar interconectados, recolhendo e compartilhando informações para melhorar a eficiência, a automação, a segurança e a conveniência numa variedade de setores. Por exemplo, numa casa inteligente, pode-se ter eletrodomésticos, lâmpadas, fechaduras e outros dispositivos conectados à Internet, podendo ser controlados remotamente através de um aplicativo em *smartphone* e, além disso, podem comunicar entre si para otimizar o uso de energia, a segurança e o conforto. A IoT traz desafios em termos de segurança, privacidade e gestão de grandes volumes de dados gerados por esses dispositivos interconectados.

Nanorobôs: Os nanorobôs são dispositivos extremamente pequenos com escala nanométrica (10^{-9} metros) projetados para realizar tarefas específicas a nível molecular ou celular. Na área da medicina, podem ser projetados para entregar medicamentos diretamente às células ou tecidos doentes e para realizar procedimentos cirúrgicos dentro do corpo humano.

Processamento de linguagem natural (PLN): É uma área da IA que se concentra na interação entre computadores e linguagem humana, capacitando os computadores a ler, a responder, a analisar e a gerar linguagem em diferentes contextos e aplicações, simulando a capacidade humana de compreender a linguagem quotidiana. Por meio de técnicas como análise sintática, semântica e pragmática, juntamente com algoritmos de aprendizagem da máquina e processamento de dados, o PLN permite que os computadores processem grandes volumes de texto, extraiam informações relevantes e forneçam respostas "significativas" (relevantes para o utilizador) em tempo real. Isso torna possível a aplicação do PLN em diversas áreas, como tradução automática, análise de sentimentos em redes sociais, resumo automático de textos, extração de informações de documentos, entre outras.

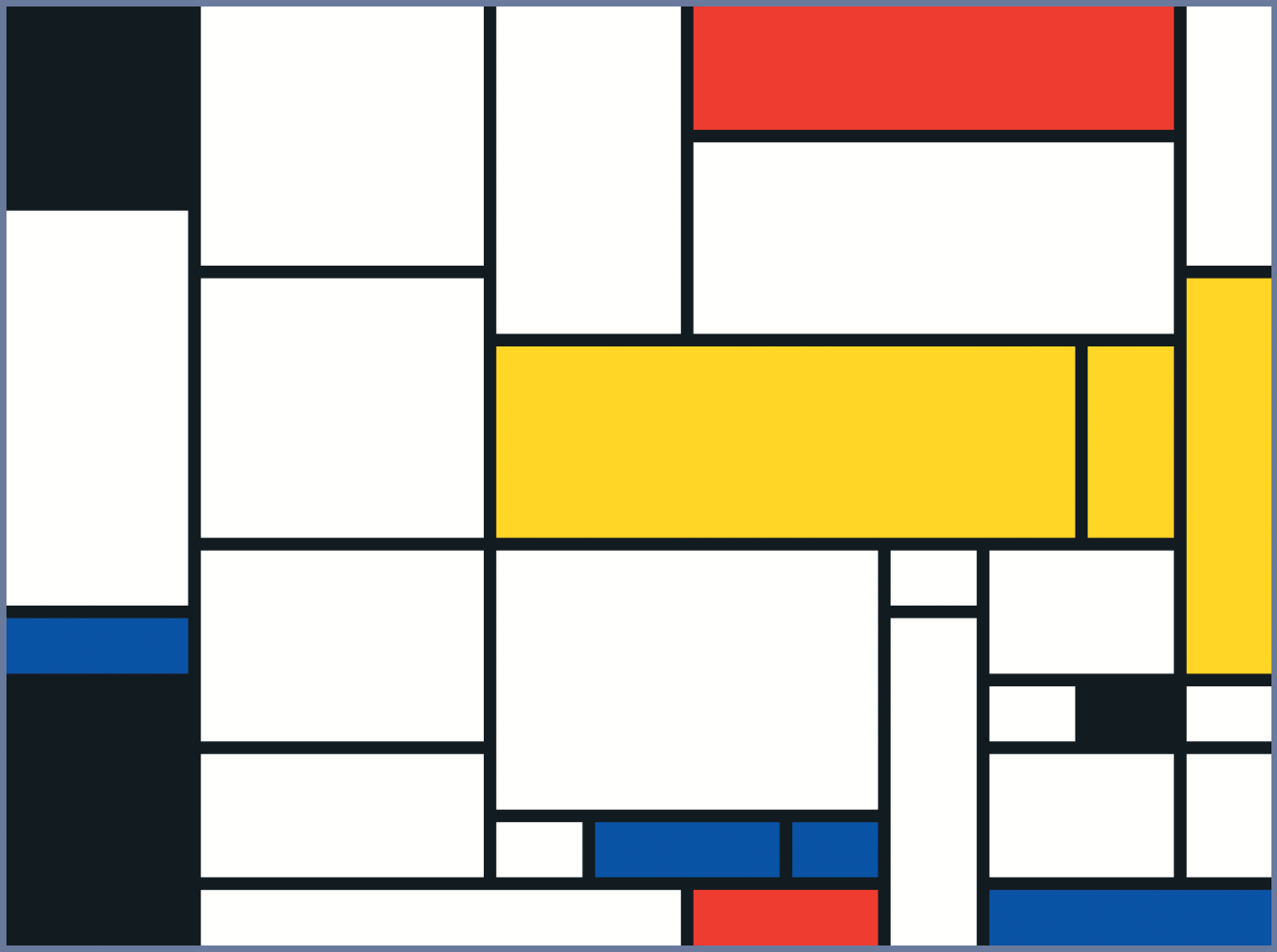
Realidade aumentada: É uma experiência interativa em que os ambientes e objetos do mundo real são complementados por modelos *3D* (três dimensões), gerados por computador, e sequências animadas que são apresentadas como se estivessem num ambiente do mundo real.

Redes neuronais: As redes neuronais são modelos computacionais inspirados pelo sistema nervoso central de um animal. Utilizam algoritmos estatísticos e técnicas de aprendizagem automática para reconhecer padrões escondidos a partir de dados atuais e históricos e correlações em dados não processados, classificando-os através da colocação em grupos, o que lhes permite fazer previsões sobre o futuro.

Notas biográficas

Helena Machado é Professora Catedrática de Sociologia no Instituto de Ciências Sociais da Universidade do Minho e investigadora no CRIA-UMinho/IN2PAST. A sua investigação situa-se na intersecção entre a sociologia, os estudos de ciência e tecnologia, os estudos críticos de vigilância e a criminologia. O seu trabalho tem abordado as implicações sociais e éticas de tecnologias emergentes, com um foco particular na genética forense e, mais recentemente, na Inteligência Artificial. Atualmente, investiga o uso de tecnologias de reconhecimento facial e deteção de emoções, explorando as suas implicações para a governamentalidade e os desafios éticos, culturais e políticos. Helena Machado tem sido agraciada com prestigiados financiamentos, incluindo do Conselho Europeu de Investigação. É membro fundadora e coordenadora da Rede de Investigação em Ciências Sociais dedicada ao estudo da Inteligência Artificial, dados digitais e algoritmos (AIDA – Artificial Intelligence, Data & Algorithms).

Susana Silva é Professora Associada de Sociologia no Instituto de Ciências Sociais da Universidade do Minho e investigadora no CRIA-UMinho/IN2PAST. Tem-se envolvido em estudos interdisciplinares que intersetam metodologias quantitativas e qualitativas para analisar políticas e cuidados de saúde integrados e centrados nos cidadãos e para discutir formas contemporâneas de governação e regulamentação da investigação no campo da saúde e forense. O seu trabalho tem explorado as implicações sociais e éticas decorrentes do uso de tecnologias emergentes e controversas, com um foco particular em tecnologias reprodutivas e genéticas. Atualmente, desenvolve investigação sobre ética pragmática, ética de cuidado, e participação cidadã na regulação de aplicações da Inteligência Artificial em contextos clínicos e de reprodução humana. É membro da Rede de Investigação em Ciências Sociais dedicada ao estudo da Inteligência Artificial, dados digitais e algoritmos (AIDA – Artificial Intelligence, Data & Algorithms).



Este livro nasce da necessidade de disponibilizar, numa linguagem acessível, uma análise científica e pedagógica sobre Inteligência Artificial (IA) que fomente a capacidade crítica e reflexiva. Com um olhar atento aos principais desafios sociais e éticos da IA no século XXI — sobretudo na educação, saúde e justiça —, esta obra explora questões de poder e elementos políticos e culturais que subjazem às narrativas dominantes sobre a IA.

Ao desvendar estas dimensões e projetar cenários futuros, queremos abrir espaço para as visões, necessidades e expectativas de diferentes grupos sociais, promovendo uma IA que beneficie toda a sociedade. Este livro convida-nos a refletir sobre como a IA pode gerar vantagens para alguns, mas também impactos negativos para os mais vulneráveis, exigindo uma reflexão coletiva e antecipatória.

Procuramos alargar o diálogo público sobre a IA, questionando: Quem define o que é "bom" para a sociedade? Que valores sociais devem prevalecer? Como projetar e aplicar a IA de forma a garantir o bem-estar comum? Um convite para imaginar uma IA mais inclusiva e responsável.

UMinho Editora

Universidade do Minho

ISBN 978-989-9074-51-4

9 789899 074514 >