AEQUITAS

D2.1 Relatório de Vieses

Cofinanciado pela
União Europeia

Índice

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado pela União Europeia

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado pela União Europeia

AEQUITAS

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Introdução

Parte do projeto AEQUITAS consiste em criar uma base de dados sobre vieses de género e raciais nas aplicações médicas da Inteligência Artificial (IA), focando-se especificamente em três doenças: doenças cardiovasculares, diabetes, e depressão.

Para completar esta tarefa, os parceiros do consórcio precisaram primeiro de recolher uma variedade de fontes sobre os vieses acima listados. O Hospital Universitário de Colónia (Universitätsklinikum Köln, UKK), como líder da tarefa e especialista nesta área, organizou a atividade de recolha de informação e forneceu o modelo de mapeamento que os parceiros usaram para mapear as fontes, garantindo que a informação relevante pudesse ser facilmente transferida para a base de dados.

Este relatório apresenta as bases teóricas e científicas que orientaram a seleção da atividade de recolha e do modelo de mapeamento, acompanhado de estudos de caso que evidenciam os diferentes tipos de viés; as implicações políticas dos enviesamentos induzidos pela IA biomédica, que afetam os direitos protegidos pela Carta dos Direitos Fundamentais da UE; uma descrição da atividade de recolha de dados; o modelo de mapeamento; uma lista de fontes recolhidas pelos parceiros AEQUITAS; e outro material de apoio. O resto do relatório está estruturado da seguinte forma:

Em primeiro lugar, apresentamos o contexto teórico do nosso trabalho numa introdução às aplicações médicas da IA e ao conceito de viés em sistemas informáticos e medicina. Começamos por nos focar na medicina, mostrando primeiro como os vieses raciais e de género se manifestam nos cuidados médicos, e segundo, como é que as questões morais que surgem na prática da medicina e da investigação biomédica são abordadas pela Bioética, fornecendo uma breve introdução aos quatro princípios da Bioética (Autonomia, Não Maleficência, Beneficência e Justiça).

Em seguida, passamos para o domínio da IA, apresentando os tipos de viés que podem ser observados em sistemas de IA tal como se manifestam no pipeline de Machine

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Learning (aprendizagem automática) e Inteligência Artificial (ML/IA). Cada tipo de viés é acompanhado por exemplos e um estudo de caso sobre vieses de género e raciais, contemplando o seu impacto a nível social relativamente às três doenças focadas pelo projeto AEQUITAS (doenças cardiovasculares, diabetes e depressão), extraído de fontes recolhidas pelos parceiros AEQUITAS após a conclusão do T2.2. Quando isso não foi possível porque o material recolhido não demonstrava claramente o tipo específico de enviesamento de IA considerado, foi apresentado um caso alternativo, de outro domínio médico, facilmente generalizável para as doenças-alvo do AEQUITAS. As descrições dos estudos de caso, juntamente com as fontes recolhidas, baseiam-se ainda em recursos científicos adicionais conforme necessário para as apoiar.

Por último, na secção de implicações Políticas, é demonstrado como os vários tipos de viés de IA afetam os direitos fundamentais protegidos pela Carta da UE, nomeadamente os preceitos da dignidade humana, igualdade perante a lei, não discriminação, bem como o direito à integridade da pessoa, o direito ao acesso aos cuidados de saúde, à proteção de dados, e o direito a um recurso eficaz, concluindo com as salvaguardas que podem ser implementadas em avaliações de conformidade, monitorização pós-comercialização, e contratação no setor público.

O relatório termina com as Referências e os seguintes Apêndices:

Apêndice 1: Método de Recolha e Mapeamento das Fontes, que contém o modelo de mapeamento e descreve o processo de recolha, mapeamento, e avaliação de informação realizado durante as tarefas T2.1 e T2.2.

Apêndice 2: Contém material de apoio para as tarefas T2.1 e T2.2, ou seja, slides que descrevem o processo, apresentados pela UKK em reuniões de parceiros.

Appendix 3: A lista de fontes recolhidas pelos parceiros AEQUITAS.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Aplicações médicas da IA

O crescimento das aplicações de IA nos últimos anos impactou fortemente a Medicina, incluindo a aquisição de dados digitalizados, machine learning e a infraestrutura de computação (Yu et al., 2018). Em particular, a introdução de algoritmos de deep learning (aprendizagem profunda) em áreas como a visão computacional e processamento de linguagem natural revolucionou as aplicações informáticas em radiologia, patologia, cardiologia, diabetologia, psiquiatria, oncologia, etc. (Esteva et al., 2019; Koteluk et al., 2021; Rajpurkar et al., 2022; Gou et al., 2024). A Organização Mundial de Saúde (OMS) enumera os seguintes domínios de aplicação de sistemas de IA nos cuidados de saúde: diagnóstico e diagnóstico baseado em previsão, cuidados clínicos, investigação e desenvolvimento de medicamentos, gestão e planeamento de sistemas de saúde, saúde pública e vigilância em saúde pública, promoção de saúde, prevenção de doenças, vigilância baseada em previsão, preparação para emergências e resposta a surtos (World Health Organization, 2021).

No entanto, o advento das aplicações de IA em medicina traz consigo um conjunto de desafios, como desafios na implementação, que incluem limitações de confiança no modelo e nos dados, questões de responsabilização, que incluem desafios regulatórios e a atribuição adequada de responsabilidades, e a garantia de justiça através da utilização ética dos dados, da distribuição equitativa de benefícios e da deteção e mitigação de enviesamentos (Rajpurkar et al., 2022).

O projeto AEQUITAS foca-se nos casos de viés de género e raça em doenças cardiovasculares, diabetes e depressão. As aplicações médicas de IA apoiam o cuidado cardiovascular através de apoio na decisão clínica, telemedicina, avaliação de riscos, terapia personalizada, análise preditiva e monitorização remota (Bernstein et al., 2025; Naskar et al., 2025); melhoram o controlo da diabetes (incluindo a monitorização e autogestão pelos doentes), diagnóstico, tratamento e prevenção (Contreras & Vehi, 2018; Khalifa & Albadawy, 2024; Naskar et al., 2025; Sheng et al., 2024). No que toca à

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

depressão, estão envolvidas no rastreio, diagnóstico e tratamento (Alhuwaydi, 2024) , com especial foco na deteção e rastreio através do uso de Modelos de Linguagem de Grande Escala (LLMs) (Cao et al., 2025; Kumari et al., 2025; Mao et al., 2023; Wang et al., 2025). Em todas estas áreas existem desafios de viés, por exemplo, no que diz respeito a doenças cardiovasculares, diabetes e depressão, ver van Assen et al. (2024), Cronjé et al. (2023), Dang et al. (2024), respetivamente.

Desafios como vieses, seja na Medicina ou IA, são abordados através de uma combinação de Bioética e Ética na IA. Na secção seguinte, fornecemos uma breve visão geral destes dois tipos de domínios de ética aplicada que serviram de base teórica e científica para o desenvolvimento do modelo de Mapeamento de Vieses.

Ética e Vieses na Medicina e na IA

Bioética e Vieses na Medicina

O viés na medicina encontra-se bem documentado: ver, por exemplo, Hammond et al. (2021) no que toca a viés cognitivo, que consiste em erros sistemáticos de pensamento devido a limitações de processamento humano ou modelos mentais desadequados, e (FitzGerald and Hurst 2017) para viés implícito, que envolvem associações alheias à consciência e que conduzem a uma avaliação negativa de uma pessoa com base em características irrelevantes como raça ou género.

O viés racial na medicina encontra-se bem estudado no caso dos EUA, por exemplo, onde está documentado que os afro-americanos, assim como os outros grupos minoritários, são submetidos a menos procedimentos e a cuidados médicos de pior qualidade, recebendo tratamentos menos agressivos, apresentando taxas de cirurgia mais baixas, e recebendo menos encaminhamentos para especialistas do que as pessoas brancas (Bowser, 2001; Williams & Wyatt, 2015).

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

O viés de género pode ser atribuído à cegueira de género e a preconceitos estereotipados sobre homens e mulheres (Hamberg, 2008), somados a uma falta generalizada de conhecimento acerca do funcionamento do corpo feminino e das suas diferenças biológicas em relação ao corpo masculino. Por exemplo, mulheres gravemente doentes com 50 anos ou mais tinham menor probabilidade de ser admitidas numa unidade de cuidados intensivos (UCI) do que homens igualmente doentes (Bierman, 2007) e, mesmo em investigação básica, pré-clínica e cirúrgica, em que são usados modelos de ratinho, os machos são, em geral, mais representados que as fêmeas (Yoon et al., 2014).

É também importante referir que os indivíduos LGBT+ sofrem discriminação no acesso a cuidados de saúde e são alvo de estereótipos que não afetam a população heterossexual. Estes fatores sociais e culturais perpetuam a discriminação e têm um impacto na saúde. Por exemplo, um estudo nos EUA, baseado em dados do National Health Interview Survey (NHIS) de 2013-2014, constatou que adultos LGB reportaram piores indicadores de saúde, assim como mais limitações funcionais, sofrimento psicológico severo e dificuldades em suportar custos de saúde, em comparação com os seus pares heterossexuais. Estas desigualdades são impulsionadas pelo stress de minorias e por uma marginalização social multifacetada (Liu et al., 2023).

Por outro lado, a medicina,enquanto disciplina, está sujeita a elevados padrões éticos desde a antiguidade até aos dias de hoje (Baker & McCullough, 2008). Há séculos que existe a expectativa social de que um médico siga as regras éticas de responsabilidade profissional estabelecidas pelos padrões da sua profissão, manifestadas através de normas profissionais que vão desde o Juramento de Hipócrates, de 400 a. C. (Miles, 2005), até às declarações de Genebra e Helsínquia (Tröhler, 2008). Como salientam (Vevaina et al., 1993), os médicos são obrigados a conformar-se ao código deontológico da sua profissão devido ao investimento que a sociedade faz na sua educação (monetário e o uso dos seus membros como material de aprendizagem ao

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

longo da formação e carreira do médico), e ao monopólio virtual que a sua profissão detém através do licenciamento.

A ética biomédica (ou bioética) é um domínio de ética prática (ou aplicada) que aborda as questões morais que surgem na prática da medicina e na investigação biomédica (Vevaina et al., 1993). Fundamentais para a ética biomédica são os quatro princípios definidos por Beauchamp e Childress (Beauchamp & Childress, 2019):

1. Autonomia: respeitar a capacidade de decisão de pessoas autónomas. Duas condições gerais são essenciais para a autonomia: liberdade, manifestada como independência de influências controladoras, e agência, isto é, a capacidade de ação intencional.

2. Não maleficência: evitar causar danos.

3. Beneficência: tomar medidas positivas para ajudar outros, especificamente, prevenindo o mal ou o dano, removendo o mal ou o dano, e promovendo o bem.

4. Justiça: distribuir benefícios, riscos e custos de forma justa. Justiça é interpretada como um tratamento justo, equitativo e apropriado para indivíduos e grupos, tendo em conta as muitas disparidades na assistência médica e na investigação com base na raça, etnia, género e condição social.

Ética e Viés em IA

A introdução da IA e o rápido desenvolvimento de aplicações de IA suscitaram uma série de questões éticas (Christoforaki & Beyan, 2022), sendo o enviesamento e a descriminação duas das mais importantes.

Assim, a ética da IA foi desenvolvida como um domínio da ética prática (ou aplicada), que compreende “um conjunto de valores, princípios e técnicas que empregam

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

padrões amplamente aceites de certo e errado para guiar a conduta moral no desenvolvimento e utilização de tecnologias de IA" (Leslie, 2019, p. 3).

A ética da IA assenta tanto na bioética (os quatro princípios apresentados acima) como no discurso dos direitos humanos, este último incluindo, entre outros, o direito à igualdade de liberdade e dignidade perante a lei, a proteção dos direitos civis, políticos e sociais, o reconhecimento universal da personalidade e o direito à participação livre e desimpedida na vida da comunidade (Leslie, 2019).

Aos quatro princípios bioéticos, é assim adicionada a Explicabilidade, e são traduzidos para a IA em (Floridi et al., 2018) da seguinte forma:

1. Autonomia, compreendida como o poder dos humanos de decidir se devem decidir, e contendo o risco de delegar demasiado nas máquinas.

2. Não maleficência, entendida como a prevenção de danos decorrentes quer da intenção humana, quer do comportamento imprevisível das máquinas.

3. Beneficência, entendida como a promoção do bem-estar, preservando a dignidade e a sustentabilidade do planeta.

4. Justiça, compreendida como a prevenção e eliminação de discriminações injustas já existentes, bem como de novos danos, e assegurando a distribuição equitativa dos benefícios da IA.

5. Explicabilidade, definida como a compreensão e a responsabilização dos processos de tomada de decisão da IA.

Desta forma, a ética da IA convergiu num conjunto de princípios baseados nos quatro pilares da ética médica, assim como de outras abordagens, resumidas em (Christoforaki & Beyan, 2022). No entanto, como notado em (Mittelstadt, 2019), em comparação com a medicina, o desenvolvimento de IA não inclui: (1) objetivos comuns

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

e deveres fiduciários, (2) histórico e normas profissionais, (3) métodos comprovados de tradução de princípios em práticas, e (4) mecanismos de responsabilização legais e profissionais robustos. Naturalmente, existe ainda um ambiente regulatório complexo de governação do desenvolvimento e utilização de IA na UE, incluindo leis anti-discriminatórias, um tópico que, no entanto, se encontra fora do âmbito deste relatório.

Em relação aos direitos humanos, de acordo com um relatório de 2018 financiado pelo Conselho da Europa (Committee of experts on internet intermediaries (MSI-NET), 2018), os direitos humanos particularmente afetados por algoritmos e por técnicas automatizadas de processamento de dados incluem:

- Direito a um julgamento justo e a um processo equitativo
- Privacidade e proteção de dados
- Liberdade de expressão
- Direito a recurso efetivo
- Liberdade de reunião e de associação
- Proibição de discriminação
- Direitos sociais e acesso a serviços públicos
- Direito a eleições livres

Algoritmos enviesados são mencionados explicitamente como um possível fator discriminatório de grupos sociais com base na idade, orientação sexual, raça, género ou condição socioeconómica (Committee of experts on internet intermediaries (MSI-NET), 2018, p. 27). Além disso, a Convenção-Quadro do Conselho da Europa sobre a Inteligência Artificial e os Direitos Humanos, a Democracia e o Estado de Direito

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

menciona especificamente que os Estados-Membros “devem adotar ou manter medidas com vista a garantir que as atividades dentro do ciclo de vida dos sistemas de inteligência artificial respeitem a igualdade, incluindo a igualdade de género, e a proibição de discriminação, tal como previsto no direito internacional e nacional aplicável”, (Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024, p. 4).

As organizações da sociedade civil (OSCs), enquanto partes interessadas no ecossistema de saúde (Vayena et al., 2018) podem desempenhar um papel significativo na identificação e abordagem dos vieses da IA e na governação da IA em geral através da defesa do desenvolvimento ético da IA, responsabilizando as partes interessada, educando o público, representando comunidades marginalizadas, formulando políticas e estruturas regulatórias, e fomentando colaborações entre governos, empresas tecnológicas e o público (Korir, 2024).

Dentro desta estrutura teórica, são explicitamente desenvolvidas diversas soluções técnicas para lidar com o enviesamento. Na secção seguinte, apresentamos a classificação de vieses induzidos pela IA que serviu de base para o nosso modelo de mapeamento, focando no seu impacto na descriminação de género e racial. Os vieses mentais humanos (Hofmann, 2023), dos quais são exemplo vieses cognitivos como o viés de confirmação ou de disponibilidade, apesar de terem grande impacto na Medicina, são considerados fora do âmbito do presente projeto.

Viés em sistemas de IA

O viés em sistemas informáticos é definido em (Friedman & Nissenbaum, 1996, p. 332) como um termo “[referente] a sistemas informáticos que discriminam sistemática e injustamente contra certos indivíduos ou grupos individuais em favor de outros. Um sistema discrimina injustamente se nega uma oportunidade ou um bem, ou se atribui

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

um resultado desfavorável a um indivíduo ou grupo de indivíduos com base em fundamentos irracionais ou desadequados."

De acordo com (Friedman & Nissenbaum, 1996), vieses em sistemas informáticos podem ser distinguidos em três categorias: viés preexistente, viés técnico, e viés emergente. Na subsecção seguinte, examinamos cada tipo de viés e ilustramo-lo com estudos de caso, tal como manifestado na literatura científica.

Viés preexistente

O viés preexistente tem origem em vieses presentes em instituições práticas e atitudes sociais que já existem, que são independentes e que geralmente estão presentes antes da criação do sistema. Este tipo de viés é incorporado no sistema de forma consciente ou inconsciente, por vezes até quando os criadores do sistema tentam evitá-lo.

Estudo de caso: Diagnóstico de Doenças Cardiovasculares em Mulheres

As doenças cardiovasculares (DCVs) têm sido comummente percepcionadas como "doenças masculinas", o que tem contribuído para o subdiagnóstico e subtratamento nas mulheres. Como demonstrado por (Al Hamid et al., 2024), numa revisão sistemática sobre este assunto, as DCVs foram menos reportadas entre mulheres que, ou apresentavam sintomas mais ligeiros do que os homens, ou cujos sintomas eram erradamente diagnosticados como gastrointestinais ou relacionados com a ansiedade; consequentemente, as mulheres recebiam menos exames de diagnóstico, medicação e eram referenciadas para cardiologistas e/ou hospitalizadas com menor frequência. Além disso, quando hospitalizadas, as mulheres tinham menor probabilidade de receber uma intervenção coronária. Por conseguinte, os factores de risco das mulheres eram subestimados pelos médicos, especialmente pelos médicos do sexo masculino. Considerando que as mulheres continuam sub-representadas na área da cardiologia (Fatunde et al., 2025), é possível concluir que estas têm menor probabilidade de receber cuidados médicos adequados devido a preconceitos já existentes.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Considerando que os sistemas de IA são treinados com recurso a dados recolhidos pelas práticas existentes, um sistema de diagnóstico de DCV baseado em IA irá incorporar este viés, criando discriminação contra as mulheres, independentemente de quaisquer escolhas feitas durante a implementação técnica.

Viés técnico

O viés técnico surge de limitações ou considerações técnicas, particularmente quando os criadores de sistemas tentam adaptar conceitos humanos à computação, como quantificar o qualitativo, discretizar o contínuo ou formalizar o informal. Além disso, descontextualizar algoritmos dos ambientes em que operam pode fazer com que não tratem todos os grupos de forma justa em todas as condições relevantes.

Estudo de caso: Precisão preditiva de modelos de previsão de risco de AVC em populações negras e brancas

(Hong et al., 2023) realizaram um estudo retrospectivo sobre a precisão preditiva do risco de AVC, comparando modelos de previsão de risco específicos para AVC já existentes com novas técnicas de machine learning que envolviam, entre outros critérios, a raça dos doentes. Todos os algoritmos apresentaram pior discriminação em indivíduos negros do que em indivíduos brancos. Esta situação, segundo os autores, pode resultar de fatores de risco não captados nos dados, como o tipo de plano de saúde, as barreiras linguísticas e outros fatores resultantes do acesso diferenciado aos serviços de saúde, ou seja, os dados estão descontextualizados do ambiente socioeconómico em que foram produzidos. Simultaneamente, todos estes fatores de risco são constructos difíceis de representar num formato adequado para computadores. A tudo isto, podemos ainda acrescentar o facto de que os algoritmos de IA de última geração são, por natureza, opacos quanto às características que selecionam para alcançar uma elevada precisão (Knight, 2017), tornando assim até os

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

seus criadores incapazes de explicar como funcionam e, portanto, de controlar se algum dos fatores socioeconómicos acima mencionados é realmente tido em conta no funcionamento interno do sistema de IA.

Viés emergente

O viés emergente manifesta-se num contexto de utilização com utilizadores reais, normalmente após a conclusão do projeto, como resultado de mudanças no conhecimento social que não pode ser, ou não é, incorporado no projeto do sistema, ou de uma população com conhecimento ou valores culturais diferentes dos pressupostos no projeto.

Estudo de Caso: Alterações no conjunto de dados (dataset)

Uma alteração no dataset é uma incompatibilidade entre as distribuições dos datasets de treino e de teste durante o desenvolvimento do algoritmo e pode levar a um desempenho discrepante ao nível do subgrupo (Chen et al., 2023).

Na deteção do cancro de pele, por exemplo, muitos datasets de imagem utilizados para treinar algoritmos de IA para detetar a doença são provenientes de países com populações de pele clara (Guo et al., 2021), sub-representando, portanto, determinados grupos demográficos. Os algoritmos de IA treinados com estes datasets apresentam um desempenho inferior quando aplicados em países com populações mais diversas, discriminando indivíduos de pele escura. A recolha, anotação e validação de datasets são difíceis e dispendiosas, o que faz com que os sistemas de IA desenvolvidos em países de baixo e médio rendimento dependam de datasets disponíveis publicamente e que podem não refletir a distribuição da sua população, resultando numa incompatibilidade entre as populações de origem e a população-alvo. O mesmo pode ocorrer em países de rendimento elevado, por exemplo, devido a alterações populacionais resultantes do aumento da imigração ou a variações na autodeclaração de raça. Como se observa em (Chen et al., 2023), “uma vez que agora é

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

aceite que a raça é uma construção social e que existe uma maior variabilidade genética dentro de uma raça específica do que entre raças" […] "a comunidade médica começou a perceber que as taxonomias do passado não representam adequadamente os grupos de pessoas que pretendem representar" e "podem obscurecer a cultura, a história, o estatuto socioeconómico e outros fatores de confusão da equidade".

Tipos de viés específico dos pipelines de ML/IA

Embora o exposto acima seja válido para todos os sistemas computacionais, as aplicações de IA têm requisitos mais específicos, pelo que precisávamos de uma taxonomia mais refinada. Consequentemente, decidimos seguir a classificação de viés apresentada em (Suresh & Guttag, 2020), uma vez que identifica os tipos de viés em cada etapa do pipeline de ML/IA, como ilustrado na Figura 1.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado pela União Europeia

AEQUITAS

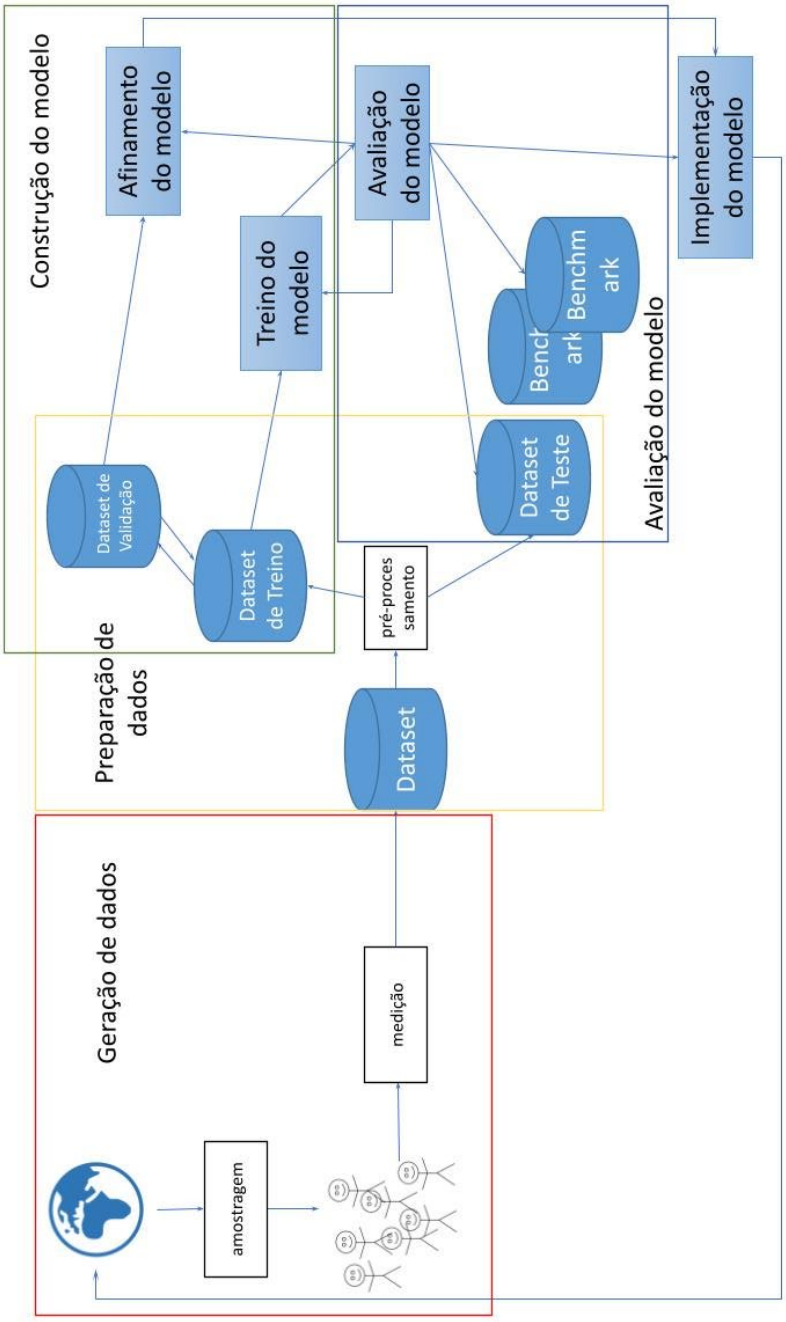

Figura 1 Pipeline de ML/AI. Imagem adaptada de (Suresh & Guttag, 2020)

Um pipeline ML/IA típico pode ser descrito da seguinte forma:

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

- **Geração de Dados**. A criação de um sistema de ML/IA começa com a geração de dados. Isto envolve, em primeiro lugar, a recolha e preparação de dados para compilar um dataset para o sistema de IA. Os dados existentes no mundo devem ser recolhidos através da identificação de uma amostra da população-alvo. O passo seguinte é definir e medir características relevantes para a aplicação a implementar e/ou anotar os dados com rótulos apropriados. Este é um processo dispendioso e demorado, pelo que, na maioria das vezes, os profissionais de IA utilizam datasets existentes (sejam públicos ou adquiridos).

- **Preparação dos Dados**. Nesta etapa, o dataset é dividido em três partes: o *dataset de treino* - o conjunto de dados utilizado para treinar o modelo; o *dataset de validação*, uma amostra de dados utilizada para avaliar o ajuste do modelo ao dataset de treino, durante o afinamento dos hiperparâmetros do modelo (parâmetros que não podem ser aprendidos a partir dos dados, como o número de camadas e neurónios numa rede neural). Nesta fase, os dados podem necessitar de pré-processamento (por exemplo, serem limpos, normalizados); e o *dataset de teste*, a parte dos dados utilizada para avaliar o modelo final, fornecendo um padrão de referência após o treino completo do modelo.

- **Construção do Modelo**. Nesta fase, o modelo é treinado com os dados de treino e afinado através da otimização dos hiperparâmetros no dataset de validação.

- **Avaliação do Modelo**. O modelo treinado é avaliado utilizando o dataset de teste e, por vezes, conjuntos de dados de referência (benchmark datasets), que são conjuntos de dados compilados independentemente e que são utilizados para demonstrar a robustez do modelo e/ou permitir a comparação com outros métodos.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

- **Implementação do Modelo**. Aplicação do modelo num contexto real. Isto pode levar a alterações dependendo dos resultados e pode também criar um ciclo de feedback para o início do processo.

Considerando as fases do pipeline de ML/IA descritas acima, adotamos a classificação de viés de (Suresh & Guttag, 2020). Especificamente, eles identificam as seguintes categorias de viés: histórico, de representação, de medição, de agregação, de aprendizagem, de avaliação e de implementação. Nas subsecções seguintes, definimos os enviesamentos acima listados e apresentamos estudos de caso das fontes recolhidas para o projeto.

Viés Histórico

O viés Histórico corresponde ao viés preexistente, como definido por (Friedman & Nissenbaum, 1996), que incorpora nos dados preconceitos e estereótipos já existentes. Um exemplo pode ser visto em (Calderone, 1990), que examina se a frequência de administração de analgésicos e sedativos a doentes submetidos a Cirurgia de Revascularização do Miocárdio (CRM) no pós-operatório difere de acordo com o sexo e a idade do doente. O resultado revelou que os doentes do sexo masculino e os doentes com 61 anos ou menos receberam analgésicos com uma frequência significativamente mais elevada do que os doentes do sexo feminino e os doentes com 62 anos ou mais, que, por sua vez, receberam sedativos com uma frequência significativamente mais elevada. O Estudo de Caso sobre a Precisão preditiva de modelos de previsão de risco de AVC em populações negras e brancas demonstra-o na subsecção sobre o Viés Preexistente; no entanto, apresentaremos um outro estudo de caso que demonstra o viés histórico relacionado com o uso de IA em saúde mental.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Estudo de Caso: A Inteligência Artificial na saúde mental e os vieses dos modelos baseados na linguagem.

(Straw & Callison-Burch, 2020) apresentam uma revisão sistemática da literatura sobre a utilização do Processamento de Linguagem Natural (PLN) na saúde mental, com o objetivo de identificar como estes vieses podem amplificar as desigualdades em saúde. Os modelos de IA que utilizam o PLN para traçar perfis de saúde mental recolhem grandes datasets de linguagem expressiva, geralmente obtidos a partir de redes sociais, fóruns online, blogues e salas de chat. No entanto, estes dados são influenciados à partida pela experiência pessoal e pelo contexto social do indivíduo.

Especificamente, no que diz respeito ao género e à linguagem, existe uma extensa bibliografia (sobre a língua inglesa) resumida em(Pennebaker et al., 2003), que revela diferenças no uso das palavras por mulheres e homens. Por exemplo, as mulheres utilizam um discurso menos assertivo, manifestando-se em maior polidez, menos palavrões, mais intensificadores (por exemplo, "realmente", "então") e mais atenuadores (ou seja, qualificadores ou palavras que indicam incerteza, como "mais ou menos", "talvez" ou "quem sabe"). Por outro lado, os homens, foram descritos como diretivos, precisos e também menos emocionais no seu uso da linguagem, que é caracterizado por referências à quantidade, adjetivos de julgamento (por exemplo, "bom", "burro"), frases elípticas ("Grande foto.") e referências a "eu". Como observam os autores, estas diferenças são consistentes com uma estrutura sociológica de diferenças de género, mas também podem ser atribuídas a explicações alternativas, como o maior envolvimento social das mulheres.

No que toca à saúde mental, homens e mulheres escrevem cartas de suicídio expressando angústia suicida de diferentes formas; as mulheres interiorizam emoções negativas, enquanto os homens expressam raiva crescente (Straw & Callison-Burch, 2020). Um sistema de IA que rastreia problemas de saúde mental em um dos géneros

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

pode ser inadequado para outro (e isto considerando o género num contexto binário, o que exclui grande parte da população).

Viés de Representação

O viés de representação ocorre quando a amostra de desenvolvimento sub-representa alguma parte da população durante a fase de recolha de dados. Isto pode acontecer das seguintes formas: ao definir a população-alvo, se esta não refletir a população utilizadora; ao definir a população-alvo, se esta contiver grupos sub-representados; ao amostrar a população-alvo, se o método de amostragem for limitado ou desigual. O viés de representação resulta numa generalização deficiente para um subconjunto da população utilizadora. Um exemplo típico de viés de representação diz respeito à deteção de cancro de pele, dado que muitos datasets de imagens sub-representam determinados grupos demográficos, fazendo com que os modelos de machine learning sejam treinados predominantemente com imagens de indivíduos de pele clara (Guo et al., 2021). Considerando as doenças alvo do projeto AEQUITAS, apresentamos um estudo de caso sobre viés de representação relacionado com a raça na diabetes tipo 2.

Caso de Estudo: Avaliação de viés racial em algoritmos de previsão de risco de diabetes tipo 2

De acordo com (Cronjé et al., 2023), em relação população dos EUA, os grupos brancos não hispânicos continuam a estar sobre-representados na literatura sobre a previsão do risco de diabetes, apesar do risco nesta população ser comparativamente mais baixo. Numa outra revisão sobre Equidade Étnico-Racial em Inteligência Artificial para a Gestão da Diabetes, nos artigos revistos que reportaram a raça, a distribuição média foi de 69,5% brancos, 17,1% negros e 3,7% asiáticos, enquanto apenas 2 artigos reportaram a inclusão de participantes nativos americanos (Pham et al., 2021).

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Está bem documentado que as desigualdades nos desfechos da diabetes são amplamente impulsionadas por determinantes sociais da saúde complexos e inter-relacionados, e que incluem o acesso a alimentos saudáveis, cuidados de saúde de qualidade, cobertura de seguro de saúde, barreiras educacionais e taxas diferenciadas na adoção de tecnologias. Estes desfechos incluem taxas mais elevadas de complicações e pior controlo glicémico entre as populações minoritárias e de baixo rendimento (Alipour & Alipour, 2025).

Consequentemente, um sistema de IA treinado com datasets existentes teria um desempenho fraco ao fazer generalizações, levando a modelos preditivos enviesados que podem favorecer indivíduos de determinados grupos raciais, por exemplo, em ações preventivas.

Viés de Medição

O viés de medição ocorre na escolha, recolha ou computação de características e rótulos a utilizar num problema de previsão, especialmente quando se utiliza um indicador indireto (proxy, uma aproximação de um constructo que não é diretamente codificado ou observável). Um exemplo pode ser encontrado num estudo de Obermeyer et al. (2019), onde os custos com a saúde foram utilizados como indicador indireto para prever e classificar quais os doentes que beneficiariam mais de cuidados adicionais, resultando em discriminação racial. No entanto, os custos com os cuidados de saúde são um indicador indirecto desadequado para as necessidades de cuidados de saúde, uma vez que doentes negros, que enfrentam níveis de pobreza desproporcionais, frequentemente gastam menos em cuidados de saúde do que doentes brancos. Devido a este viés, o algoritmo concluiu erradamente que doentes negros eram mais saudáveis do que doentes brancos igualmente doentes, classificando-os, por isso, como doentes de menor prioridade no acesso aos serviços de saúde.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado pela União Europeia

AEQUITAS

Outras fontes de viés de medição podem ocorrer quando o método de medição varia entre grupos, por exemplo, quando dois grupos são monitorizados quanto ao mesmo comportamento, mas um deles é monitorizado com mais rigor ou frequência do que o outro. Da mesma forma, a precisão da medição pode variar entre grupos, o que, em aplicações médicas, pode levar a taxas sistematicamente mais elevadas de diagnóstico incorreto ou subdiagnóstico em determinados grupos. Por exemplo, os médicos são mais propensos a subestimar a dor dos doentes negros em relação aos doentes não negros devido a falsas crenças sobre as diferenças biológicas entre negros e brancos, o que faz com que os doentes negros tenham menos probabilidade de receber analgésicos e, se receberem, recebam quantidades mais pequenas (Hoffman et al., 2016).

Estudo de Caso: Diferenças Raciais e Étnicas na Associação entre Média de Glicose e Hemoglobina A1c

O teste da hemoglobina glicada (A1C) mede a quantidade média de glicose (açúcar) no sangue e é utilizado para detetar a pré-diabetes ou para auxiliar no diagnóstico da diabetes tipo 2. No entanto, a A1C é apenas uma medida indireta e não está causalmente ligada aos desfechos de saúde, dado que existem inúmeras formas pelas quais a relação entre as medidas diretas da glicemia (a concentração de glicose no sangue) e a A1C pode ser diretamente alterada. Existe até uma variação substancial na relação glicemia-A1C entre indivíduos e até mesmo no mesmo indivíduo ao longo do tempo. Além disso, estudos reportaram níveis significativamente mais elevados de hemoglobina glicada (A1C) em doentes afro-americanos do que em doentes brancos com a mesma média de glicose (Karter et al., 2023).

Se um sistema de IA concebido para diagnosticar a diabetes for treinado para utilizar os resultados do teste de A1C como um indicador indireto da glicemia, sem ter em conta outros fatores como a raça do paciente, isto pode levar a diagnósticos prematuros de diabetes e a tratamentos desadequados, resultando em enviesamentos

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

e desigualdades na qualidade dos cuidados de saúde. No entanto, como observado em (Alipour & Alipour, 2025), uma revisão sistemática de vieses que podem afetar a equidade dos modelos de IA/ML na diabetes (incluindo o viés de medição), embora os estudos revistos mencionem explicitamente que o viés de medição se pode propagar através de modelos de IA se não for corrigido, nenhum deles teve em conta tais vieses durante o desenvolvimento do modelo, mitigou-os explicitamente ou reportou a correção de diferenças na precisão da medição.

Viés de Agregação

O viés de agregação surge quando é utilizado um modelo único para um conjunto de dados que inclui grupos diversos de pessoas ou coisas.

Podemos considerar o exemplo do mapeamento de dados de entrada (por exemplo, o rendimento de uma pessoa) para rótulos que os descrevam (por exemplo, baixo, médio, alto), assumindo que são consistentes em todos os subconjuntos dos dados. Na realidade, a origem ou cultura de uma pessoa pode mudar o que estes números significam na realidade. Por exemplo, um rendimento "elevado" numa pequena cidade rural ou num país de baixo ou médio rendimento pode significar algo muito diferente do que numa grande cidade ou num país de alto rendimento.

Caso de Estudo: Ferramentas de saúde digitais para a monitorização passiva de depressão

A utilização de ferramentas digitais para medir variáveis fisiológicas e comportamentais na monitorização passiva da depressão é abordada em (De Angel et al., 2022), uma revisão sistemática sobre o tema. Os artigos nela revistos examinaram associações entre depressão e dados comportamentais objetivos obtidos através de sensores de smartphones e dispositivos vestíveis (wearable). Estes dados foram mapeados em características utilizadas pelos modelos de IA para fazer previsões,

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

correspondentes ao sono, atividade física, ritmo circadiano, sociabilidade, localização e utilização do telefone.

No entanto, os autores realçam a heterogeneidade que surge da diversidade de métodos utilizados para criar estas características. Por exemplo, a característica "qualidade do sono" pode ser definida pela contagem de despertares, pelo número total de minutos acordado ou pela proporção de tempo acordado em relação ao tempo a dormir numa sessão de sono, sendo que também é necessário ter em conta as diferenças na forma como os sensores em diferentes dispositivos descrevem um evento como "sono". Uma vez que todas as diferenciações acima referidas não são consideradas e são agrupadas coletivamente sob o termo "qualidade do sono", e considerando que um dataset pode ser proveniente de pessoas ou grupos com diferentes origens, culturas ou normas, esta característica pode ter um significado diferente para cada um destes grupos ou indivíduos.

Agregar estes dados numa única característica pode resultar num sistema que não se adequa a nenhum grupo ou que privilegia a população dominante, caso exista também um viés de representação. Por exemplo, existem evidências de diferenças sexuais no sono entre homens e mulheres e, no entanto, estas últimas estão frequentemente sub-representadas na investigação sobre o sono. Além disso, outros factores geralmente não considerados em relação aos padrões e perturbações do sono incluem a não distinção do género, como uma construção social, e o sexo biológico, e a não consideração das identidades interseccionais definidas pela idade, raça e classe socioeconómica (Lok et al., 2024).

Viés de Aprendizagem

O viés de aprendizagem surge quando as escolhas de modelação amplificam as disparidades de desempenho entre diferentes exemplos nos dados. Um exemplo disso é a privacidade diferencial, um mecanismo utilizado nos sistemas de IA que garante

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

que, ao examinar o output de um sistema, não é possível determinar se os dados de um indivíduo específico estavam incluídos no dataset original. A privacidade diferencial é utilizada em conjuntos de dados da área da saúde para proteger a informação sensível dos doentes, por exemplo, no caso de doenças raras, em que o caso de cada doente é mais ou menos único numa área limitada servida por um hospital; assim, mesmo que os dados sejam anonimizados, não é muito difícil deduzir a identidade da pessoa. Foi demonstrado, no entanto, que a privacidade diferencial reduz a influência de dados sub-representados no modelo; portanto, se o sistema de IA já apresenta um viés desde o início, a aplicação de uma medida de melhoria da privacidade agrava ainda mais este enviesamento (Bagdasaryan & Shmatikov, 2019).

Estudo de Caso: Privacidade diferencial e disparidades na saúde

Em setembro de 2018, o Departamento de Censos dos EUA anunciou que iriam implementar privacidade diferencial em produtos de dados derivados dos dados do censo de 2020. No entanto, (Santos-Lozada et al., 2020) investigaram a forma como a implementação da privacidade diferencial pode alterar o conhecimento sobre disparidades de saúde na mortalidade, especialmente para minorias raciais ou étnicas em pequenas áreas e ambientes menos urbanizados. Os resultados sugeriram ainda que a privacidade diferencial afetará mais fortemente as estimativas da taxa de mortalidade para os negros não hispânicos e hispânicos do que as estimativas para os brancos não hispânicos.

Estas descobertas foram corroboradas por (Kurz et al., 2022), que demonstraram que a aplicação de privacidade diferencial aos mesmos dados pode resultar em representações erradas das taxas de participação do Medicaid em grupos raciais e étnicos já marginalizados. Especificamente, em determinadas combinações de condado, raça e etnia, as diferenças nas taxas obtidas entre os resultados dos dados com privacidade diferencial e os dados originais, chegaram a ultrapassar os 10% em alguns casos. Além disso, os indivíduos brancos não hispânicos foram o único subgrupo

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

étnico e racial para o qual o algoritmo de privacidade diferencial captou com precisão as taxas de participação no Medicaid. Esta descoberta pode ter implicações importantes para as políticas de saúde, uma vez que que os dados dos Censos são utilizados para planear programas governamentais, alocar recursos e avaliar e monitorizar políticas.

Viés de avaliação

O viés de avaliação ocorre quando os dados de referência (benchmark) utilizados para uma determinada tarefa não representam a população utilizadora. Os benchmarks são conjuntos de dados normalizados utilizados para medir a qualidade de um modelo, permitindo a comparação quantitativa entre modelos. Consequentemente, existe o risco de incentivar o desenvolvimento e a implementação de modelos que apresentem um bom desempenho apenas no subconjunto de dados representado no benchmark. Assim, pode ocorrer discriminação contra subgrupos ou indivíduos vulneráveis se o benchmark estiver sujeito a um viés histórico, de representação ou de medição.

Na área da saúde, a sub-representação de populações específicas em datasets pode ocorrer devido à ausência de indivíduos ou grupos (por exemplo, grávidas, devido a restrições éticas) ou à categorização incorreta ou desadequada de pessoas em grupos (por exemplo, categorias como “etnia mista” ou “outros”). As causas subjacentes podem incluir razões sociais, técnicas ou legais/éticas, tais como barreiras estruturais no acesso à saúde, obstáculos técnicos à recolha ou digitalização de dados de saúde relevantes, limitações individuais e estruturais relativas ao consentimento para a partilha de dados, e restrições legais ou éticas à partilha de dados que impedem o acesso aos mesmos, entre outras (Arora et al., 2023).

Consequentemente, os sistemas de IA calibrados com base nestes parâmetros podem apresentar um desempenho inferior quando aplicados a indivíduos de um grupo sub-representado. No entanto, é importante notar que a validade dos parâmetros de

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

referência é uma questão mais genérica e não se limita a enviesamentos (Brooks, 2025).

Estudo de Caso: Datasets de imagens de pele

Os conjuntos de dados de imagens de pele sub-representam determinados grupos demográficos, dado que a maioria das imagens nestes conjuntos provém de populações da América do Norte ou da Europa e retrata predominantemente indivíduos de pele clara (Guo et al., 2021). Devido ao elevado custo e à dificuldade de construção destes datasets, para além do treino de modelos, estes podem ser também utilizados como benchmarks.

O estudo de caso que ilustra o viés emergente, ou seja, os datasets de imagens de cancro de pele utilizados para treinar modelos de previsão, é um exemplo de benchmark inadequado quando a população de utilizadores pertence a grupos sub-representados (Guo et al., 2021). Um caso semelhante que, embora não relacionado com a IA, mostra a generalidade do problema, envolvia oxímetros de pulso (dispositivos que medem a saturação de oxigénio no sangue, utilizados, por exemplo, em casos de ataque cardíaco ou insuficiência cardíaca), que demonstraram funcionar com maior precisão em peles de pigmentação clara (Sjoding et al., 2020).

Vieses de representação, medição, agregação, aprendizagem e avaliação, podem seer mapeados para viés técnico, como definido por (Friedman & Nissenbaum, 1996).

Viés de Implementação

O viés de implantação surge quando existe uma incompatibilidade entre o problema que um modelo pretende resolver e a forma como é efetivamente utilizado, o que pode ser prejudicial, especialmente quando combinado com vieses cognitivos como o viés de confirmação e o viés de automação.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanciado pela União Europeia

AEQUITAS

O viés de implementação equivale ao viés emergente definido por (Friedman & Nissenbaum, 1996).

Estudo de Caso: Mudança de domínio

O caso de alterações no dataset encontra-se documentado na subsecção de viés emergente sobre a deteção de cancro de pele. Adicionalmente, podemos definir a questão de deslocamento de domínio, que ocorre quando um sistema é implementado, recebe autorização regulamentar e é implantado na prática clínica, mas é aplicado a uma coorte de doentes diferente daquela para a qual foi treinado. Por exemplo, um sistema pode ser desenvolvido para um hospital num país de rendimento elevado e implementado num país de rendimento baixo ou médio sem ter em conta factores como as características sociodemográficas dos doentes ou se os doentes têm o mesmo nível de risco global em comparação com os incluídos nos dados de treino (Vokinger et al., 2021).

Implicações políticas

As evidências mapeadas no Deliverable D2. 1 demonstram que os vieses de género e raciais na IA biomédica não são falhas técnicas incidentais ou isoladas, mas sim riscos sistémicos que emergem ao longo de todo o ciclo de vida dos sistemas de IA utilizados na área da saúde. Nas doenças cardiovasculares, depressão e diabetes, o viés surge de datasets clínicos historicamente enviesados, práticas de diagnóstico desiguais, variáveis indiretas que codificam desigualdades estruturais e contextos de implementação que distribuem de forma desigual tanto os benefícios como os malefícios. Estas descobertas confirmam que a IA biomédica afeta diretamente múltiplos direitos e princípios protegidos pela Carta dos Direitos Fundamentais da UE, principalmente os preceitos da dignidade humana, da igualdade perante a lei e da não

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

discriminação, bem como o direito à integridade da pessoa, o direito à saúde, à proteção de dados e o direito a um recurso eficaz.

Neste contexto, as políticas europeias e nacionais que regem a IA na saúde devem tratar a mitigação dos enviesamentos não como um complemento ético voluntário, mas como uma componente obrigatória da implementação legal e em conformidade com os direitos humanos da IA. Os esforços regulamentares europeus e nacionais relativos à IA na saúde devem ser vistos como inseridos no contexto mais amplo dos direitos fundamentais que regem a IA (ver Novossiolova, 2025; Novossiolova et al., 2025; Kasapi, 2025). A Lei da IA da UE (EU AI Act) fornece uma base regulamentar necessária ao classificar a maioria dos sistemas de IA biomédica como de alto risco, mas a sua eficácia na prática dependerá da forma como as salvaguardas dos direitos fundamentais serão operacionalizadas nas avaliações de conformidade, na monitorização pós-comercialização e nas aquisições do setor público.

Em primeiro lugar, as garantias de uma supervisão humana significativa devem ser reforçadas e especificadas para os sistemas de IA biomédica ao longo de todo o seu ciclo de vida. As ferramentas de IA clínica utilizadas para diagnóstico, estratificação de risco, rastreio ou apoio ao tratamento não devem, em caso algum, funcionar como decisores autónomos. A supervisão humana deve incluir não só a possibilidade de intervenção manual pelos profissionais de saúde, mas também uma clara responsabilidade institucional pela compreensão das limitações do sistema, dos riscos de enviesamento conhecidos e das lacunas de desempenho entre subgrupos. Em consonância com a proteção da dignidade e integridade humanas prevista na Carta, os profissionais de saúde devem receber formação e apoio institucional para questionarem criticamente os resultados da IA, em vez de se submeterem a eles. Isto requer a incorporação da literacia em IA, da consciencialização sobre os vieses e da formação em direitos fundamentais na educação médica e no desenvolvimento profissional contínuo.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

As obrigações de transparência devem ser interpretadas de forma abrangente em contextos de saúde. Os doentes e os utentes dos serviços de saúde devem ser informados sempre que sejam utilizados sistemas de IA na tomada de decisões clínicas que os afetem, incluindo a triagem, priorização ou avaliação de riscos. Quando os resultados gerados pela IA influenciam os serviços públicos de saúde, estes resultados devem ser claramente identificáveis como tal e acompanhados de explicações acessíveis sobre a sua função, limitações e riscos de viés conhecidos. Os indivíduos devem também ser informados quando os seus dados pessoais são utilizados para treino, teste ou aprendizagem contínua de IA, particularmente quando estão envolvidos dados de saúde sensíveis. Estas medidas de transparência são essenciais para garantir os direitos à proteção de dados e ao recurso eficaz previstos na Carta, e para permitir que os indivíduos contestem de forma significativa decisões que os possam afetar negativamente..

Em segundo lugar, a avaliação do impacto nos direitos fundamentais deve tornar-se um requisito rotineiro e obrigatório para os sistemas de IA biomédicos, estendendo-se para além das verificações pré-mercado e abrangendo a avaliação contínua durante a implementação. As evidências empíricas do D2. 1 mostram que muitos danos decorrentes do enviesamento só se tornam visíveis quando os sistemas de IA interagem com populações reais e fluxos de trabalho clínicos, particularmente através de efeitos interseccionais que envolvem género, raça, idade e condição socioeconómica. As avaliações de impacto baseadas nos direitos, como as inspiradas pela metodologia HUDERIA do Conselho da Europa (Metodologia para a Avaliação de Riscos e Impactos dos Sistemas de Inteligência Artificial do Ponto de Vista dos Direitos Humanos, da Democracia e do Estado de Direito), devem, portanto, ser obrigatórias para a IA médica de alto risco, examinando explicitamente o desempenho e os resultados diferenciais entre os grupos protegidos. Estas avaliações devem envolver a participação significativa das partes interessadas, incluindo organizações da sociedade

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

civil, representantes dos doentes e entidades de defesa da igualdade, a fim de revelar danos que podem ser invisíveis de uma perspetiva puramente técnica ou clínica.

Auditorias periódicas aos sistemas de IA biomédicos devem ser obrigatórias para verificar a conformidade contínua com as normas de direitos fundamentais, com especial atenção à deriva de viés, às alterações nos datasets e às alterações na utilização clínica ao longo do tempo. Quando as auditorias revelarem efeitos discriminatórios persistentes ou não-mitigáveis, devem existir vias legais e institucionais claras para restringir, suspender ou terminar a utilização do sistema. O direito à saúde não pode justificar a contínua implementação de ferramentas de IA que desfavorecem sistematicamente determinados grupos, mesmo que as métricas de desempenho agregadas pareçam favoráveis.

Em terceiro lugar, as autoridades da UE e nacionais devem abordar o risco de utilização indevida e de danos secundários associados à IA biomédica. Isto inclui vulnerabilidades de cibersegurança que possam comprometer a integridade do sistema ou permitir a manipulação maliciosa de resultados clínicos, bem como a reutilização da IA na saúde para vigilância, criação de perfis ou práticas de exclusão. Os sistemas de IA biomédica devem estar sujeitos a avaliações de segurança regulares e a obrigações robustas de notificação de incidentes, com mecanismos claros de responsabilização nos casos em que sistemas enviesados ou comprometidos conduzam a violações de direitos. Os marcos de responsabilidade devem garantir que a responsabilidade não pode ser atribuída exclusivamente a médicos individuais quando os danos estão estruturalmente incorporados no design da IA ou nas decisões de implementação.

Em quarto lugar, a promoção de práticas éticas e responsáveis deve ser incorporada em toda a cadeia de valor da IA biomédica. Os programadores devem ser obrigados a abordar proativamente os riscos de viés através da recolha de dados representativos, seleção criteriosa de alvos e indicadores, validação específica para subgrupos e

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

relatórios transparentes de desempenho em relação ao género e grupos raciais. É importante realçar que as evidências analisadas em D2. 1 mostram que a “equidade através da inconsciência” e as estratégias puramente técnicas de redução de enviesamento são frequentemente insuficientes em ambientes de cuidados de saúde. Por conseguinte, as orientações e normas regulamentares devem ir além das métricas abstratas de equidade e exigir que os programadores demonstrem resultados de equidade clinicamente significativos, avaliados em relação aos fluxos de cuidados e aos padrões de acesso reais.

As políticas de contratação pública e de financiamento desempenham um papel crucial na definição dos incentivos aos desenvolvedores. As autoridades de saúde e os hospitais públicos devem integrar os direitos fundamentais e os critérios de enviesamento nas decisões de aquisição de sistemas de IA, privilegiando soluções que demonstrem práticas robustas, transparentes e verificadas de forma independente para a mitigação de enviesamentos. Os instrumentos de financiamento da UE, incluindo futuros programas de investigação e inovação, devem continuar a priorizar projetos que combinem a inovação técnica com a governação baseada nos direitos, o envolvimento das partes interessadas e a capacitação, em consonância com o modelo AEQUITAS.

Por fim, o reforço da resiliência da sociedade à IA biomédica enviesada exige um investimento contínuo na sensibilização do público, no envolvimento da sociedade civil e na colaboração intersectorial. Os indivíduos devem ser capacitados para compreender os seus direitos em relação à assistência médica mediada por IA e os mecanismos disponíveis para os proteger. As organizações da sociedade civil, as entidades de igualdade e os grupos de doentes devem ser reconhecidos como intervenientes essenciais na monitorização dos impactos da IA, no apoio às pessoas afectadas e na formulação de políticas. A cooperação entre governos, profissionais de saúde, investigadores, indústria e sociedade civil é necessária para garantir que os

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

benefícios da IA biomédica são partilhados de forma equitativa e não reforçam as desigualdades em saúde existentes..

No seu conjunto, as conclusões do Deliverable D2. 1 sustentam uma conclusão política clara: a IA biomédica só pode ser considerada fiável e legítima na UE quando a sua conceção, implementação e governação estiverem firmemente ancoradas na proteção dos direitos fundamentais. A Lei da IA da UE, interpretada à luz da Carta dos Direitos Fundamentais da UE e operacionalizada através de mecanismos concretos de supervisão, avaliação de impacto e responsabilização, oferece uma oportunidade crucial para garantir que a inovação na área da saúde promove a equidade, em vez de reproduzir padrões históricos de discriminação.

Financiado pela União Europeia. Os pontos de vista e as opiniões expressas são as do(s) autor(es) e não refletem necessariamente a posição da União Europeia ou da Agência de Execução Europeia da Educação e da Cultura (EACEA). Nem a União Europeia nem a EACEA podem ser tidos como responsáveis por essas opiniões. Código do projeto: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI