Região Autónoma da Madeira
Governo Regional

Secretaria Regional
de Educação, Ciência e Tecnologia
Direção Regional de Educação

Inteligência Artificial Responsável na Educação

Terminologia e Taxonomia

Ciências da Computação
Região Autónoma da Madeira
O habitat natural dos criadores

Título:

Inteligência Artificial Responsável na Educação | Terminologia e Taxonomia

Ciências da Computação - O habitat natural dos criadores

Autor:

Rodolfo Duarte Pinto - Ciências da Computação | CSF Code.org Facilitador

Gabinete de Modernização das Tecnologias Educativas

Divisão de Tecnologias, Segurança e Infraestruturas

Direção de Serviços de Tecnologias e Ambientes Inovadores de Aprendizagem

Direção Regional de Educação

Contactos:

Rua D. João n.º 57, Quinta Olinda 9054 - 510 Funchal

Madeira - Portugal

Telefone: +351 291 705 860

Email: rodolfodu7@edu.madeira.gov.pt

Funchal, abril de 2024

Índice

Introdução

No atual panorama digital e tecnológico, a Inteligência Artificial (IA) é uma fronteira de inovação que transcende setores, estabelecendo-se como uma linguagem universal ao nível do progresso e do desenvolvimento. É neste contexto que o documento "EU-U.S. Terminology and Taxonomy for Artificial Intelligence (Second Edition)"[1] assume um papel preponderante, ao propor uma base linguística comum e abrangente que não só facilita a colaboração entre diversas áreas, incluindo o da Educação, mas também uniformiza a compreensão das tecnologias emergentes da IA.

Com o nosso compromisso educativo no âmbito das Ciências da Computação, reconhecemos a importância crucial de disponibilizarmos aos nossos professores ferramentas que lhes permitam aceder e compreender integralmente estas novas dinâmicas. Nesta perspetiva, **partilhamos a tradução e adaptação do documento acima mencionado**, um passo determinante para garantir que a nossa comunidade educativa esteja na vanguarda do conhecimento da IA. São também incluídos exemplos práticos para cada conceito, de modo a facilitar o seu mais célere entendimento.

Importa destacar que o documento referido é fruto da colaboração de especialistas de IA da União Europeia e dos Estados Unidos da América. A integração de 13 novos termos e a revisão de 24 termos previamente estabelecidos refletem a natureza evolutiva e colaborativa inerente da IA. Este documento é o produto de um processo iterativo e meticuloso, que procurou integrar o feedback de peritos externos, afinando definições e terminologias com critérios e metodologias consolidados, reforçando assim a qualidade e a exatidão da informação disponibilizada.

Com este recurso, reforçamos o nosso empenho em assegurar que os profissionais da Educação e os alunos estejam preparados para enfrentarem os desafios e a capitalizarem as oportunidades que a IA apresenta.

[1] https://teducativas.madeira.gov.pt/s/PvcCoy3

1 - Ciclo de vida da IA

Aprendizagem automática adversária

Um campo preocupado com a conceção de algoritmos de Aprendizagem Automática (*Machine Learning*) capaz de resistir a desafios de segurança, do estudo das capacidades dos agentes adversários e da compreensão das consequências dos ataques.

Exemplo Prático: Imagine um sistema de reconhecimento facial usado para segurança. Um investigador de segurança pode tentar enganar o sistema, alterando ligeiramente uma imagem de um rosto (por exemplo, adicionando um filtro impercetível) para testar se o sistema ainda pode reconhecer corretamente a pessoa ou se será enganado, permitindo acessos não autorizados.

Autonomia (Sistema de IA Autónomo)

O nível de independência de um sistema em relação à intervenção humana e a capacidade de operar sem a intervenção humana. [Os diferentes sistemas de IA têm diferentes níveis de autonomia.] Um sistema autónomo possui um conjunto de capacidades de aprendizagem, adaptativas e analíticas para responder a situações que não foram pré-programadas ou antecipadas (isto é, respostas baseadas em decisões) antes da implementação do sistema. Sistemas de IA autónomos ou semiautónomos podem ser caracterizados como "human-in-the-loop" (humano no ciclo), "human-on-the-loop" (humano sobre o ciclo), ou "human-out-of-the-loop" (humano fora do ciclo) dependendo do nível de envolvimento significativo do ser humano.

Exemplo Prático: Um carro autónomo que circula no trânsito sem intervenção humana. Este sistema avalia o ambiente ao seu redor usando sensores e câmaras, tomando decisões em tempo real sobre quando acelerar, travar, ou virar, sem qualquer comando humano.

Big data

Um termo abrangente para grandes conjuntos de dados digitais complexos cujo armazenamento, análise, gestão e processamento requerem meios tecnológicos

igualmente complexos e substancial poder de computação. Conjuntos de dados são às vezes interligados para observarem como os padrões num domínio podem afetar outras áreas. Os dados podem ser estruturados em campos fixos ou não estruturados e são frequentemente gerados ou recebidos a uma taxa elevada. A análise de grandes conjuntos de dados, muitas vezes utilizando a IA, pode revelar padrões, tendências ou relações subjacentes que não eram previamente aparentes.

Exemplo Prático: Uma empresa de venda online usa *big data* para analisar as compras e os comportamentos de navegação de milhões de utilizadores. Com estes dados, a empresa pode identificar tendências de compra e otimizar as suas estratégias de marketing e do stock.

Aumento de dados

A técnica onde o conjunto de dados de treino é aumentado em tamanho e qualidade, através da alteração dos dados originais de treino, para criar novos exemplos de treino, com o objetivo de treinar melhores modelos de Aprendizagem Automática.

Exemplo Prático: Para treinar um modelo de reconhecimento de imagens, um engenheiro de dados pode rodar, redimensionar ou ajustar o brilho das imagens originais para criar mais exemplos. Esta prática torna o modelo mais robusto às variações que possa encontrar em imagens no mundo real.

Envenenamento de dados

Um tipo de ataque de segurança onde atores maliciosos modificam os dados de treino com o objetivo de corromper o modelo aprendido, fazendo com que o sistema de IA aprenda algo que não deveria.

Exemplo prático: Se um ator mal-intencionado obtém acesso ao conjunto de dados utilizado para treinar um modelo de filtragem de spam, pode introduzir emails maliciosos, mas aparentemente inofensivos, que são classificados como seguros. Isso pode resultar em que o sistema classifique futuros emails semelhantes, que são efetivamente maliciosos, como seguros.

Engenharia de características

A engenharia de características é o ato de extrair características de dados brutos - isto é, extrair representações numéricas de aspetos dos dados - e transformá-las em formatos que são adequados para um modelo de Aprendizagem Automática.

Exemplo Prático: Um modelo de previsão de preços de apartamentos pode utilizar características como a área, o número de quartos e a distância até ao centro da cidade. O engenheiro de dados extrai estas características dos dados brutos para ajudar o modelo a realizar previsões mais precisas.

Representação de conhecimento

A arte de formalizar conhecimento de forma declarativa, tipicamente para uso num sistema de IA simbólico, como um sistema especialista.

Exemplo Prático: Num sistema especialista médico, o conhecimento sobre sintomas, diagnósticos e tratamentos é codificado numa base de conhecimento, permitindo que o sistema faça inferências e ofereça recomendações médicas com base na introdução de sintomas pelo utilizador.

Ciclo de vida de um sistema de IA

As fases do ciclo de vida de um sistema de IA envolvem várias etapas, incluindo: 1) planeamento e design, 2) recolha e tratamento de dados, 3) construção do modelo e/ou adaptação de modelos existentes a tarefas específicas, 4) teste, avaliação, verificação e validação, 5) implementação e 6) operação e monitorização. Estas fases decorrem frequentemente de forma iterativa e não são necessariamente sequenciais.

Exemplo Prático: O desenvolvimento de um assistente virtual engloba várias etapas, começando pela recolha de dados de diálogos, seguindo-se o treino do modelo de linguagem, testes de interação e recolha do feedback dos utilizadores. Este processo é contínuo, com atualizações regulares para melhorar a precisão e a capacidade de resposta do sistema.

Função de Perda (também chamada função de custo)

A função de perda produz uma métrica de avaliação global única, para fins de treino, de um sistema de IA tomando qualquer decisão ou ação disponível. Tipicamente, o objetivo do treino do sistema de IA é minimizar a perda total sobre um conjunto de exemplos de validação.

Exemplo Prático: Num modelo de previsão de vendas, a função de perda pode calcular a diferença entre as vendas previstas e as vendas reais. Esta função ajuda a otimizar o modelo para minimizar estas discrepâncias, tornando as previsões mais precisas.

Aprendizagem Automática

Aprendizagem Automática (*Machine Learning*) é um ramo da IA que se concentra no desenvolvimento de sistemas capazes de aprenderem a partir de dados para realizarem uma tarefa sem serem explicitamente programados. A aprendizagem refere-se ao processo de otimização dos parâmetros do modelo através de técnicas computacionais de modo que o comportamento do modelo seja otimizado para a tarefa de treino.

Exemplo Prático: Um sistema de recomendação de filmes que aprende com as classificações e os comentários dos utilizadores para sugerir novos filmes que, provavelmente, serão do agrado dos mesmos utilizadores.

Processamento de linguagem natural

O campo preocupado com máquinas capazes de processar, analisar e gerar linguagem humana, seja ela falada, escrita ou gestual.

Exemplo prático: Um chatbot de atendimento ao cliente que interpreta perguntas formuladas em linguagem natural e fornece respostas automáticas. Este chatbot utiliza o conteúdo de uma FAQ ou de uma base de conhecimento para responder às questões dos utilizadores de forma precisa e eficaz.

Prompt

Prompts são entradas para um sistema de IA generativo que descrevem uma tarefa que o sistema deve executar ou a informação à qual deve responder.

Exemplo Prático: No desenvolvimento de conteúdo assistido por IA, como a redação de artigos, o utilizador fornece um prompt, por exemplo, "Escreva um artigo sobre as tendências tecnológicas em 2024". O sistema, então, gera um rascunho inicial do artigo baseado neste prompt.

Engenharia de prompts

O processo de desenhar e criar prompts para modelos de IA generativos com o objetivo de obter as saídas desejadas. Envolve compreender as capacidades do modelo e adequar o prompt para orientar eficazmente o modelo a gerar saídas relevantes, informativas e criativas.

Exemplo Prático: Ajustar o prompt de um modelo generativo para criar uma ilustração, especificando detalhes como o estilo, aos elementos visuais desejados e o tema, para assegurar que o resultado esteja alinhado com as expectativas do utilizador.

Aprendizagem por Reforço

A Aprendizagem por Reforço é um subconjunto da Aprendizagem Automática que permite a um sistema artificial (por vezes referido como um agente) num determinado ambiente otimizar o seu comportamento. Os agentes aprendem a partir de sinais de feedback recebidos como resultado das suas ações, tais como recompensas ou punições, com o objetivo de maximizar a recompensa recebida. Estes sinais são computados com base numa função de recompensa dada, que constitui uma representação abstrata do objetivo do sistema. O objetivo pode ser, por exemplo, obter uma pontuação alta num jogo de vídeo ou minimizar o tempo de inatividade dos trabalhadores numa fábrica.

Exemplo Prático: Um robô de armazém que aprende a navegar e a otimizar o seu

percurso para recolher os itens, baseando-se na recompensa de completar a tarefa no menor tempo possível.

Dados Sintéticos

Dados Sintéticos são dados artificialmente gerados por um processo computacional em vez de serem captados por aparelhos sensoriais ou manualmente criados por humanos. Dados sintéticos são frequentemente produzidos por um modelo treinado para reproduzir as características e estrutura dos seus dados de treino, visando uma distribuição semelhante.

Exemplo Prático: Criação de imagens de rostos humanos para treinar um sistema de reconhecimento facial sem a necessidade de recolher grandes quantidades de imagens de pessoas reais, protegendo a privacidade e reduzindo a necessidade de dados reais.

Dados de Treino

Dados utilizados para treinar um sistema de inteligência artificial através do ajuste de parâmetros, como os pesos de uma rede neuronal.

Exemplo Prático: Utilização de um vasto conjunto de imagens etiquetadas de animais para treinar um modelo de classificação de espécies, onde os dados de treino incluem tanto as imagens como as etiquetas correspondentes a cada espécie.

2 - Medição

Precisão

Proximidade dos cálculos ou estimativas em relação aos valores exatos ou verdadeiros que as estatísticas pretendiam medir. O conceito de precisão é frequentemente utilizado para avaliar a capacidade preditiva do modelo de IA.

Exemplo Prático: Suponha que um hospital utiliza um sistema de IA para identificar pacientes que possam ter uma determinada doença com base em imagens de raio-X. De 200 imagens analisadas, o sistema identifica corretamente 180 casos em que a doença está presente e erra em 20 casos onde a doença não está presente ou falha em detetar a doença. A precisão do sistema seria então de 90%, o que significa que, em 90% dos casos, o sistema conseguiu identificar corretamente se a doença estava presente ou não. Este exemplo demonstra a importância de ter um sistema altamente preciso em aplicações médicas, onde a precisão das previsões pode ter implicações significativas para o diagnóstico e tratamento dos pacientes.

3 - Atributos técnicos do sistema

Sistema de IA

Um sistema de IA é um sistema baseado em máquina que, de forma explícita ou implícita, tem objetivos, infere, a partir dos dados que recebe, como gerar saídas tais como previsões, recomendações ou decisões que podem influenciar ambientes físicos ou virtuais. Diferentes sistemas de IA variam nos seus níveis de autonomia e capacidade de adaptação após a implementação.

Exemplo Prático: Um assistente virtual, como a Siri ou o Google Assistant, é um sistema de IA que recebe comandos de voz e interpreta esses dados para realizar tarefas como marcar compromissos, fazer chamadas ou procurar informações na internet. A autonomia e a capacidade de adaptação do sistema podem ser observadas na forma como personaliza as respostas com base no histórico de interações do utilizador.

Aprendizagem Adaptativa (adaptabilidade)

A adaptabilidade é a caraterística de alguns sistemas de IA de serem capazes de alterar o seu comportamento durante a utilização, com base em interações com as entradas e os dados. [A adaptação pode implicar uma mudança nos pesos do modelo ou uma mudança na estrutura interna do modelo em si.] [Exemplos incluem um sistema de reconhecimento de voz que se adapta à voz de um indivíduo ou um sistema de recomendação de música personalizado.] O novo comportamento do sistema adaptado pode produzir resultados diferentes do sistema anterior para as mesmas entradas.

Exemplo Prático: Um aplicativo de aprendizagem de línguas que ajusta o nível de dificuldade das lições e dos exercícios conforme o utilizador avança, analisando os erros e os acertos anteriores para oferecer conteúdo apropriado que potencialize a aprendizagem.

Sistema Especialista

Sistemas automatizados codificados com o conhecimento de especialistas humanos, tipicamente através de técnicas de representação do conhecimento. Focados em tarefas específicas e com tomada de decisão automatizada baseada em regras "se-então".

Exemplo Prático: Um programa de diagnóstico médico que utiliza conhecimento especializado para avaliar sintomas, histórico médico e resultados de exames, e fornece uma lista de possíveis diagnósticos e tratamentos com base em regras pré-definidas.

Aprendizagem Federada

A Aprendizagem Federada é uma abordagem à Aprendizagem Automática que aborda problemas de governança e privacidade de dados ao treinar algoritmos de forma colaborativa sem transferir os dados para um local central. Cada dispositivo federado treina localmente com os dados e partilha o seu modelo local em vez de partilhar os dados de treino. Diferentes sistemas de Aprendizagem Federada têm diferentes topologias que envolvem diferentes formas de partilha dos parâmetros.

Exemplo Prático: Uma aplicação de teclado para telemóveis que aprende o estilo de escrita do usuário sem enviar os dados para um servidor central. Em vez disso, a aprendizagem ocorre no próprio dispositivo e somente os parâmetros do modelo são partilhados com outros dispositivos para melhorar o algoritmo de forma colaborativa.

Valores Humanos para a IA

Os sistemas de IA não são neutros em termos de valores. Os valores são qualidades idealizadas ou condições no mundo que as pessoas consideram boas. O desenho de sistemas de IA humanos implica a negociação de diferentes valores e sistemas de criação de significados, e requer decisões relativas a princípios éticos, governança, políticas e incentivos. Desenhar IA com valores humanos exige consciência dos interesses sociais e económicos subjacentes aos sistemas de IA, bem como respeito pela diversidade cultural.

Exemplo Prático: Um sistema de recomendação de conteúdo nas redes sociais que é desenhado para promover o equilíbrio entre manter os utilizadores conectados e evitar a exposição prolongada a conteúdo polarizado ou extremista, visando respeitar a diversidade cultural e a saúde mental dos utilizadores.

IA Centrada no Humano

IA Centrada no Humano (ou "IA centrada no utilizador") é uma abordagem ao desenho, implementação e uso de sistemas de IA que os considera como componentes de ambientes sociotécnicos nos quais os humanos assumem uma agência significativa. A Abordagem Centrada no Humano à IA prioriza o aumento das capacidades humanas em vez de as substituir. A abordagem é promovida em políticas, pesquisa e engenharia com o objetivo de desenvolver sistemas de IA como ferramentas para servir os seres humanos e para aumentar o bem-estar humano e ambiental, promovendo os direitos humanos, o estado de direito, os valores democráticos e o desenvolvimento sustentável.

Exemplo Prático: Um assistente de acessibilidade para pessoas com necessidades especiais e que usa IA para descrever ambientes e objetos, facilitando a orientação em espaços públicos. Este sistema é desenvolvido com o foco em aumentar a independência e a qualidade de vida dos utilizadores, em vez de apenas substituir as funções humanas.

Modelo de Linguagem de Grande Escala (LLM)

Uma classe de modelos de linguagem que usa algoritmos de Aprendizagem Profunda e é treinada em conjuntos de textos extremamente grandes que podem ter múltiplos terabytes em tamanho. A maioria dos LLMs podem gerar representações comprimidas de uma entrada útil para tarefas como classificação ou resposta a questões.

Exemplo Prático: Um modelo como o GPT que pode gerar textos em vários estilos e idiomas com base em instruções simples. Este modelo é utilizado em tarefas como resumo de documentos, criação de conteúdo educacional ou resposta a perguntas em chatbots educativos.

Modelo

Um componente central de um sistema de IA usado para fazer inferências a partir de entradas para produzir saídas. Um modelo caracteriza uma transformação de entrada-saída destinada a realizar uma tarefa computacional central do sistema de IA (por exemplo, classificar uma imagem, prever a palavra seguinte numa sequência ou selecionar a próxima ação de um robô dadas as suas condições atuais).

Exemplo Prático: Um modelo de previsão do tempo que analisa dados atmosféricos (como a temperatura, a pressão, a humidade) para gerar previsões locais ao nível do clima. Este modelo transforma as entradas (dados observados) em saídas (previsões meteorológicas) que são cruciais para contextos agrícolas ou até eventos ao ar livre.

Rede Neural

Uma rede neural consiste em uma ou mais camadas de neurónios conectados por ligações ponderadas com pesos ajustáveis. Uma rede neural recebe dados de entrada e produz uma saída, processando-os através da rede, com cada neurónio realizando um cálculo simples. Embora algumas redes neurais sejam destinadas a simular o funcionamento dos neurónios biológicos no sistema nervoso, a maioria das redes neurais na IA são ferramentas de engenharia que retiram apenas uma inspiração geral da biologia.

Exemplo Prático: Um sistema de reconhecimento facial utilizado em segurança que identifica e verifica pessoas com base nas características faciais. A rede neural processa imagens de entrada, compara com um banco de dados e identifica correspondências, facilitando o controlo no acesso a áreas restritas.

4 - Governação

Auditabilidade de um sistema de IA

Auditabilidade refere-se à capacidade de um sistema de IA ser submetido a uma avaliação dos seus algoritmos, dados e processos de design, em particular para determinar se o sistema está a funcionar como previsto. A auditabilidade não implica necessariamente que a informação sobre modelos de negócio e propriedade intelectual relacionada com o sistema de IA deva estar sempre abertamente disponível. Assegurar a rastreabilidade e os mecanismos de registo desde a fase inicial de design do sistema de IA pode ajudar a viabilizar a auditabilidade do sistema.

Exemplo Prático: Imagine um sistema de IA utilizado numa escola para personalizar a aprendizagem dos alunos com base nos seus desempenhos anteriores. A auditabilidade neste contexto poderia envolver a revisão periódica do sistema pela equipa TIC da escola ou por uma entidade externa, para assegurar que os algoritmos não estão a criar vieses involuntários (por exemplo, favorecendo alunos de determinado perfil demográfico). Esta revisão poderia incluir a análise de como os dados dos alunos são recolhidos, processados e utilizados para ajustar os métodos de ensino, garantindo que a privacidade dos alunos é respeitada e que os resultados educacionais são justos e equitativos.

5 - Fiável

Viés

O tratamento diferencial de um sistema de IA, que pode surgir de sistemas implícitos de significado, normas e valores.

Exemplo Prático: Num sistema de recrutamento automatizado, se os dados históricos utilizados para treinar o algoritmo revelarem uma predominância de contratações de indivíduos de um determinado género, o sistema poderá desenvolver um viés que favorece candidatos desse género em detrimento de outros.

Viés Nocivo

Refere-se ao viés num sistema de IA que resulta em impactos negativos, tais como decisões injustas ou discriminatórias. Este tipo de viés pode emergir de diversos fatores, incluindo decisões humanas e processuais ao longo do ciclo de vida da IA, preconceitos culturais e sociais presentes nos dados utilizados para treino, especificações técnicas inadequadas e limitações dos dados de design, ou ainda o uso do sistema em contextos que não foram devidamente antecipados. Existem medidas que podem ser adotadas para mitigar e detetar o viés nocivo, assegurando uma maior justiça e equidade nas operações do sistema.

Exemplo Prático: Um algoritmo de concessão de crédito que foi treinado com dados que sub-representam minorias étnicas pode resultar em taxas de aprovação mais baixas para essas populações, o que constitui um viés nocivo que leva a impactos negativos e discriminatórios.

Alucinação

Quando sistemas de IA generativos produzem respostas inexatas ou falsas que podem parecer plausíveis ao utilizador. A alucinação pode ser, por exemplo, a invenção de informação histórica ou biográfica errónea. A alucinação é o resultado de previsão estatística, repetição de dados de treino ou padrões.

Exemplo Prático: Um assistente de IA utilizado para fornecer informações turísticas poderia, ao ser questionado sobre eventos históricos de uma cidade, "inventar" um evento que nunca aconteceu, devido a uma adaptação excessiva aos dados de treino ou padrões nos dados que não são factualmente corretos.

Fuga de Dados

No contexto de IA, Fuga de Dados é a introdução de informação num sistema que será esperado inferir dos dados nos quais é treinado, o que não deveria estar legitimamente disponível para aprender. Isso resulta num modelo de alto desempenho enquanto avalia o modelo no conjunto de teste durante o desenvolvimento, mas com um desempenho pobre durante a implementação quando avaliado em novos conjuntos de dados.

Exemplo Prático: Durante o treino de um modelo de IA para reconhecimento facial, se as fotos de um grupo de teste são acidentalmente incluídas no conjunto de treino, o modelo pode parecer excecionalmente preciso em testes, mas falhará em reconhecer novas faces no mundo real, evidenciando uma fuga de dados.

Deep Fake

Conteúdo de imagem, áudio ou vídeo gerado ou manipulado por IA que se assemelha a pessoas, objetos, locais ou outras entidades ou eventos existentes e que, falsamente, pareceria a uma pessoa ser autêntico ou real.

Exemplo Prático: Um vídeo deep fake de um político a realizar um discurso que nunca ocorreu pode ser criado usando a IA, levando pessoas que veem o vídeo a acreditarem que ele é autêntico, quando na verdade é uma criação.

Discriminação

Tratamento diferencial de indivíduos com base em fatores como a sua etnia, cultura ou religião. A discriminação pode resultar de vieses institucionais e individuais que estão embutidos nos processos ao longo do ciclo de vida da IA, por exemplo, vieses culturais e sociais detidos por atores de IA e organizações, ou representados nos dados de sistemas de IA. A discriminação também pode ser o resultado de limitações técnicas em

hardware ou software, ou do uso de um sistema de IA que, devido ao seu contexto de aplicação, não trata todos os grupos igualmente. Como muitas formas de viés são sistémicas e implícitas, não são facilmente controladas ou mitigadas e requerem abordagens específicas de governança e outras semelhantes.

Exemplo Prático: Um sistema de vigilância automatizado que é mais propenso a identificar indevidamente indivíduos de certas etnias como suspeitos de atividades criminosas, devido a vieses nos dados com os quais foi treinado, pratica a discriminação.

Evasão

A evasão é um dos ataques mais comuns em modelos de Aprendizagem Automática (*Machine Learning*) realizados durante a produção. Refere-se ao desenho de uma entrada que parece normal para um humano, mas é incorretamente classificada por modelos de Aprendizagem Automática, afetando o seu comportamento. Um exemplo típico é mudar alguns pixéis numa imagem antes de carregar, de modo que um sistema de reconhecimento de imagem não consiga classificar o resultado corretamente. A evasão também pode ser usada durante a implementação.

Exemplo Prático: Hackers podem alterar ligeiramente as características visuais de uma placa de uma matrícula em imagens para que um sistema de IA de reconhecimento de matrículas falhe em detetar ou identificar corretamente a placa. Uma forma de ataque conhecido como evasão.

Opacidade

Quando uma ou mais características de um sistema de IA, como processos, a proveniência dos conjuntos de dados, funções, saída ou comportamento são indisponíveis ou incompreensíveis para todos as partes interessadas - geralmente um antónimo para transparência.

Exemplo Prático: Um complexo algoritmo financeiro que decide sobre a elegibilidade para empréstimos pode ser tão complexo que nem os programadores que o criaram

conseguem explicar precisamente como decisões específicas são tomadas, exemplificando a opacidade.

IA Confiável

A IA confiável possui três componentes: (1) deve ser legal, garantindo a conformidade com todas as leis e regulamentações aplicáveis; (2) deve ser ética, demonstrando respeito e assegurando a adesão a princípios e valores éticos; e (3) deve ser robusta, tanto de uma perspetiva técnica como social, uma vez que, mesmo com boas intenções, os sistemas de IA podem causar danos não intencionais. As características dos sistemas de IA confiáveis incluem: validade e fiabilidade, segurança e resiliência, responsabilidade e transparência, explicabilidade e interpretabilidade, privacidade aprimorada e justiça com viés nocivo gerido. A IA confiável diz respeito não apenas à confiabilidade do próprio sistema de IA, mas também engloba a confiabilidade de todos os processos e atores que fazem parte do ciclo de vida do sistema de IA. A IA confiável é baseada no respeito pelos direitos humanos e valores democráticos.

Exemplo Prático: Um sistema de IA num hospital e que ajuda a diagnosticar doenças deve ser confiável; deve seguir rigorosamente as leis de proteção de dados (legal), operar com base em princípios éticos de não prejudicar e ser justo (ético), e ser tecnicamente robusto o suficiente para minimizar erros de diagnóstico (robustez).

Notas finais

À medida que concluímos a nossa viagem através deste documento, é fundamental reiterar a crescente relevância da IA em todos os setores da sociedade contemporânea, com especial destaque na Educação. O documento, "EU-U.S. Terminology and Taxonomy for Artificial Intelligence (Second Edition)", foi concebido para ser também uma ferramenta essencial também na formação dos profissionais da Educação, dotando-os com o conhecimento necessário para navegarem e prosperarem na era da IA.

O processo colaborativo da revisão e da integração de novos termos sublinha a importância da cooperação internacional e do diálogo contínuo entre especialistas, para acompanhar este ritmo acelerado do desenvolvimento da IA. A transparência e a colaboração não só enriquecem o conhecimento partilhado, mas também promovem uma compreensão mais profunda e uma aplicação mais criteriosa da IA.

Os exemplos práticos apresentados para cada conceito pretendem não só clarificar a terminologia, mas também demonstrar a aplicabilidade prática da IA em diversas situações reais. Esta abordagem prática é crucial para uma compreensão plena do impacto potencial da IA e como podemos utilizar estas tecnologias de forma responsável e ética. O papel da disciplina das Ciências da Computação é, neste contexto, naturalmente fundamental.

Este recurso é um testemunho do nosso compromisso contínuo com a Educação e adaptação às novas tecnologias. Encorajamos todos os profissionais da Educação a utilizarem este recurso para se manterem informados e preparados para os desafios e oportunidades que a IA continuará a apresentar.

À medida que avançamos, é crucial que continuemos a aprender, a adaptar e a integrar a IA de uma forma que beneficie a sociedade em geral, mantendo sempre um olhar crítico sobre as implicações éticas e sociais da sua aplicação.

Referências bibliográficas

EU-U.S. Trade and Technology Council, Working Group 1: Technology Standards Subgroup on AI Taxonomy & Terminology. (2023). EU-U.S. Terminology and Taxonomy for Artificial Intelligence: Second Edition. European Commission; National Institute of Standards and Technology.
Disponível em: https://teducativas.madeira.gov.pt/s/PvcCoy3. (Acedido a 11 de abril de 2024).

OpenAI. (2021). ChatGPT (Versão 4). Disponível em: https://openai.com/chatgpt (Acedido a 11 de abril de 2024).