ACADEMIA HUMANITÁRIA · HUMANITAIRES JOBS

Mini MBA IA e ação humanitária: trabalhar com o Claude e o Gemini

Material do curso · Semana 4 : Ética da IA e princípios humanitários: não causar dano, enviesamentos, transparência

Um programa de exatamente 3 meses (12 semanas, 144 horas) para utilizar os assistentes de IA Claude (Anthropic) e Gemini (Google) no trabalho humanitário, de forma útil, segura e responsável: funcionamento e limites, redação de pedidos, proteção de dados, ética e princípios humanitários, relatórios, propostas, análise de dados e MEAL, tradução e comunicação com as comunidades, pesquisa e verificação de fontes, desinformação e segurança, produtividade e automatização, política de utilização e implementação numa organização. Exercícios corrigidos, três exames mensais com estudos de caso e um diploma verificável da Academia Humanitária. Programa independente, sem parceria com a Anthropic nem com a Google.

https://humanitairejobs.org/pt/academia/mini-mba-ia-claude-gemini/semaine-4/

Documento gratuito da Academia Humanitária. Pode ser impresso e partilhado livremente para fins de formação, sem modificações.

-

Semana 4 : Ética da IA e princípios humanitários: não causar dano, enviesamentos, transparência

Objetivos

- Relacionar a utilização da IA com os princípios humanitários (humanidade, neutralidade, imparcialidade, independência) e com o princípio de não causar dano.
- Detetar os enviesamentos de uma resposta de IA e os seus possíveis efeitos nas pessoas afetadas.
- Aplicar regras claras de transparência, responsabilidade humana e respeito pelos direitos de autor.

Os princípios humanitários perante a IA

Os quatro princípios humanitários aplicam-se também às ferramentas digitais:

- **Humanidade**: a ferramenta deve servir as pessoas afetadas, não apenas a eficiência interna. Poupar tempo num relatório é útil se esse tempo voltar aos programas.
- **Imparcialidade**: uma IA não deve introduzir discriminação no acesso à ajuda (por exemplo classificando agregados familiares segundo critérios enviesados).
- **Neutralidade**: um texto gerado pode retomar, sem que se note, o vocabulário de uma parte em conflito. Reveja os nomes de lugares, grupos e acontecimentos.
- **Independência**: depender inteiramente de um fornecedor para uma função crítica é um risco; mantenha a capacidade de trabalhar sem a ferramenta.

O princípio de **não causar dano** (*do no harm*) obriga a perguntar antes de cada nova utilização: quem poderia ser prejudicado se a ferramenta se enganar, se os dados fugirem ou se o resultado for mal interpretado?

Os enviesamentos: de onde vêm e como detetá-los

Os modelos aprendem com textos produzidos maioritariamente em poucas línguas e em poucas regiões do mundo. Daí resultam **enviesamentos**:

- **Enviesamento linguístico**: menor qualidade em línguas com pouca presença na Internet (numerosas línguas africanas, asiáticas ou indígenas), e por vezes em certas variantes do árabe ou do francês.
- **Enviesamento de perspetiva**: leitura das crises «a partir do Norte», soluções padronizadas, desconhecimento dos atores locais e das ONG nacionais.
- **Estereótipos**: representações de género ou culturais (o «médico» é um homem, a «beneficiária» é passiva) nos textos e sobretudo nas imagens geradas.
- **Enviesamento de complacência**: tendência para dar razão a quem pergunta. Se pedir «confirma que o nosso projeto é um sucesso», o assistente provavelmente confirmará.

Para os detetar: peça a posição contrária («que argumentos há contra esta conclusão?»), faça rever os textos destinados às comunidades por pessoas dessas comunidades e compare com fontes locais.

Decisões sobre as pessoas: o limite a não ultrapassar

As decisões que afetam diretamente os direitos ou o acesso à ajuda de uma pessoa não devem ser confiadas a um assistente de IA generativa:

- seleção ou exclusão de beneficiários;
- avaliação da credibilidade de um relato (pedido de asilo, denúncia, caso de proteção);
- pré-seleção ou rejeição de candidatos a um emprego;
- decisão disciplinar ou de investigação.

Estas ferramentas podem ajudar a **preparar** (redigir um guião de entrevista, estruturar critérios publicados, resumir uma política), mas a decisão cabe a pessoas habilitadas, segundo um procedimento documentado, com possibilidade de recurso. Na União Europeia, o Regulamento da Inteligência Artificial (AI Act, adotado em 2024 e aplicado progressivamente) classifica além disso como de «risco elevado» certas utilizações como o recrutamento ou o acesso a serviços essenciais, com obrigações reforçadas.

> Teste: se a pessoa afetada soubesse que a decisão foi tomada por uma IA, seria aceitável e poderia contestá-la? Se a resposta for não, não utilize a IA para decidir.

Transparência e responsabilidade

A confiança é o capital de uma organização humanitária. Algumas regras simples protegem-na:

- **Responsabilidade humana**: cada documento tem um autor humano responsável pelo seu conteúdo, tenha ou não sido ajudado pela IA.
- **Transparência proporcionada**: indique a utilização da IA quando for significativa ou quando o destinatário possa razoavelmente esperá-lo (avaliação, investigação, conteúdo publicado, trocas com as comunidades). Para uma correção ortográfica, não é necessário.
- **Respeito pelas regras dos doadores e parceiros**: alguns doadores e convites à apresentação de propostas enquadram ou pedem que se declare a utilização da IA nas candidaturas; leia as condições (semana 6).
- **Sem usurpação**: nunca gere uma imagem, uma voz ou uma citação falsamente atribuída a uma pessoa real, e nunca apresente uma imagem gerada como uma fotografia de terreno.

As diretrizes éticas de comunicação do setor (dignidade, consentimento informado, sem miserabilismo) aplicam-se plenamente aos conteúdos produzidos com a IA.

Direitos de autor, plágio e integridade

Três pontos de atenção:

- **Conteúdos alheios**: pedir à IA que «reescreva» o relatório de outra organização para o apresentar como seu continua a ser plágio. Cite as suas fontes.
- **Conteúdos gerados**: o estatuto jurídico dos textos e imagens gerados varia consoante os países e evolui. As condições de utilização do Claude e do Gemini precisam os direitos do utilizador sobre os resultados; para uma utilização comercial ou uma publicação importante, consulte-as.
- **Integridade académica e profissional**: nos exames deste Mini MBA, como numa prova de recrutamento, utilizar a IA para responder no seu lugar é uma fraude. O objetivo é que domine as competências, não apenas a ferramenta.

Para as fotografias, a regra da organização mantém-se: nenhuma fotografia alheia sem direitos, nenhuma fotografia de uma criança identificável sem consentimento, e uma imagem gerada é claramente assinalada como tal.

Custo ambiental e equidade de acesso

A IA consome eletricidade e água nos centros de dados. As estimativas por pedido variam muito consoante os estudos e os modelos; o impacto global do setor cresce fortemente. Para uma organização empenhada no clima, reflexos razoáveis:

- utilizar a IA quando traz um valor real, não por reflexo para tarefas triviais;
- preferir um documento bem preparado a dezenas de tentativas sucessivas;
- integrar a utilização do digital na política ambiental.

Equidade de acesso: as equipas de terreno têm muitas vezes uma ligação fraca, equipamentos antigos e menos acesso a subscrições pagas. Uma organização que reserva a IA à sede e ao pessoal internacional aumenta as desigualdades. Uma política justa prevê o acesso, a formação e as línguas para o pessoal nacional e os parceiros locais.

Exercício: analisar três utilizações do ponto de vista ético

Situação. Uma ONG estuda três utilizações: A) gerar imagens «realistas» de crianças desnutridas para uma campanha de angariação de fundos, à falta de fotografias autorizadas; B) pedir à IA que classifique 600 candidaturas a um posto de motorista e rejeite automaticamente as 500 com pior pontuação; C) utilizar a IA para propor uma primeira versão das mensagens de sensibilização sobre a cólera em francês simples, depois testadas com grupos da comunidade e traduzidas por falantes nativos.

Trabalho pedido. Para cada utilização, diga se é aceitável e porquê, e proponha uma alternativa se necessário.

Correção. A) Não aceitável: uma imagem gerada apresentada como real engana o público, atenta contra a dignidade e reforça o miserabilismo; alternativa: fotografias com consentimento informado, ilustrações claramente assinaladas ou mensagens centradas nas soluções. B) Não aceitável: decisão sobre pessoas sem controlo humano, risco de enviesamento e de discriminação, ausência de recurso; alternativa: critérios publicados aplicados por um painel, podendo a IA ajudar a redigir a grelha ou uma prova prática. C) Aceitável: a IA prepara um primeiro rascunho, a comunidade e os falantes nativos validam-no; é uma utilização proporcionada, transparente e verificada.

A reter

- Os princípios humanitários e o princípio de não causar dano aplicam-se à utilização da IA.
- Os enviesamentos afetam a língua, a perspetiva, os estereótipos e a complacência; procure a contradição.
- Nenhuma decisão sobre os direitos ou o acesso à ajuda de uma pessoa é confiada a uma IA generativa.
- Um autor humano continua responsável; indique a utilização significativa da IA; nunca uma imagem gerada apresentada como real.
- Pense no custo ambiental e na equidade de acesso para o pessoal nacional e os parceiros.

Quiz

1. Pedir à IA «confirma que o nosso projeto é um sucesso» expõe sobretudo ao enviesamento:
   A. Linguístico
   B. De complacência
   C. De data
   D. De formato

2. Qual destas tarefas NÃO deve ser confiada a uma IA generativa?
   A. Redigir um guião de entrevista
   B. Resumir uma política de RH publicada
   C. Decidir que agregados familiares ficam excluídos da ajuda
   D. Corrigir a ortografia de um relatório

3. Uma imagem gerada de uma situação de terreno pode ser utilizada:
   A. Como fotografia real se for realista
   B. Nunca, em caso algum
   C. Apenas para os doadores
   D. Se for claramente assinalada como gerada e respeitar a dignidade

4. O princípio de independência convida a:
   A. Manter a capacidade de trabalhar sem a ferramenta para as funções críticas
   B. Utilizar um único fornecedor
   C. Nunca utilizar a IA
   D. Deixar a IA decidir

5. Responder aos exames deste Mini MBA com a IA no seu lugar é:
   A. Recomendado
   B. Uma fraude
   C. Obrigatório
   D. Indiferente

Respostas do quiz : 1. B · 2. C · 3. D · 4. A · 5. B