ACADEMIA HUMANITÁRIA · HUMANITAIRES JOBS

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

Material do curso · Semana 10 : Informação, desinformação e segurança: o que a IA pode e não pode fazer

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-10/

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

Semana 10 : Informação, desinformação e segurança: o que a IA pode e não pode fazer

Objetivos

- Compreender como a IA é utilizada para produzir desinformação e conteúdos falsificados dirigidos contra os trabalhadores humanitários.
- Utilizar a IA de forma segura como apoio à análise de segurança, sem expor informações sensíveis.
- Proteger-se contra o phishing, a usurpação de identidade e as fraudes facilitadas pela IA.

Um novo risco para a aceitação

A aceitação por parte das comunidades e das partes em conflito é a base da segurança humanitária. Ora, a IA generativa permite produzir, a um custo muito baixo, **conteúdos falsos e credíveis**: falsos comunicados, falsas citações de responsáveis, imagens ou vídeos falsificados (*deepfakes*), falsas gravações áudio, mensagens em massa nas redes sociais.

Algumas organizações humanitárias já foram alvo de campanhas de desinformação que as acusavam de espionagem, proselitismo, tráfico ou parcialidade. Estas campanhas podem provocar ameaças, ataques ou expulsões. O CICV e outros atores alertam para os **riscos dos «danos relacionados com a informação»** para a população civil e os trabalhadores humanitários.

> A desinformação é um risco de segurança. Tem o seu lugar na análise de contexto, na avaliação de riscos e no plano de comunicação de crise.

Detetar os conteúdos falsificados

É cada vez mais difícil distinguir a olho nu um conteúdo gerado de um real. Alguns reflexos de verificação continuam eficazes:

- **Fonte**: quem publicou primeiro? Uma conta recente, anónima, sem histórico?
- **Contexto**: a data, o lugar, o tempo, os uniformes, a língua coincidem?
- **Pesquisa inversa de imagens**: a imagem existia antes, noutro contexto?
- **Indícios técnicos**: mãos, textos deformados, sombras incoerentes, voz sem respiração natural; estes indícios desaparecem com os progressos das ferramentas, por isso não bastam.
- **Marcações**: alguns fornecedores acrescentam marcas de água invisíveis ou metadados de origem aos conteúdos gerados (por exemplo o SynthID na Google, ou os metadados da norma C2PA). A sua ausência não prova nada.

- **Confirmação direta**: contacte a organização ou a pessoa supostamente na origem do conteúdo.

Um assistente de IA pode ajudar a estruturar esta verificação e a enumerar as incoerências, mas não pode «certificar» que um conteúdo é verdadeiro ou falso.

Responder a uma campanha de desinformação

Se a sua organização for alvo de uma campanha:

- **Avaliar**: alcance, audiência, ligação a atores armados ou políticos, risco para o pessoal e as comunidades. É uma decisão da célula de gestão de riscos, não apenas da comunicação.
- **Proteger**: adaptar as medidas de segurança se necessário (deslocações, visibilidade, instalações).
- **Responder de forma proporcionada**: por vezes uma resposta pública amplifica o conteúdo falso; por vezes o silêncio deixa-o instalar-se. Privilegie muitas vezes o diálogo direto com as comunidades, os líderes e as autoridades, apoiado em factos verificáveis.
- **Documentar**: capturas datadas, ligações, denúncia às plataformas.
- **Coordenar** com as outras organizações e os mecanismos de coordenação.

A IA pode ajudar a preparar elementos de linguagem factuais, perguntas e respostas para o pessoal e mensagens adaptadas aos diferentes públicos, revistos pela direção e pela segurança.

A IA como apoio à análise de segurança

Para a segurança, a IA pode ser útil com **informações públicas** e para **tarefas metodológicas**:

- resumir relatórios públicos sobre um contexto;
- propor a estrutura de uma análise de atores ou de um plano de segurança a partir de um modelo;
- preparar exercícios de simulação (cenários fictícios);
- traduzir documentos públicos;
- redigir materiais de formação em segurança.

Não deve receber:

- os relatórios detalhados de incidentes, os nomes, os itinerários, os horários, as posições das equipas e das instalações;
- as informações confidenciais de uma rede de segurança partilhada entre ONG ou provenientes de serviços de análise de segurança reservados aos seus membros;
- as informações sobre negociações, contactos com grupos armados ou raptos.

E nunca deve **decidir** um nível de segurança, uma deslocação ou uma evacuação. Essas decisões cabem aos responsáveis designados, segundo o plano de segurança.

Phishing e fraudes reforçados pela IA

A IA torna as burlas mais convincentes:

- **E-mails de phishing** sem erros, personalizados, na língua do destinatário, que imitam um doador ou a direção;
- **Fraude do presidente**: falsa mensagem de voz ou vídeo do diretor que pede uma transferência urgente;
- **Falsas ofertas de emprego** em nome de organizações humanitárias, que pedem pagamentos aos candidatos;
- **Falsos fornecedores** que mudam os seus dados bancários.

Defesas:

- dupla validação de todo o pagamento, e confirmação por outro canal conhecido de toda a mudança de dados bancários;
- palavra-chave ou procedimento de rechamada para os pedidos urgentes invulgares;
- autenticação de dois fatores em todas as contas, incluindo as contas de IA;
- formação regular e exercícios de phishing;
- lembrete público: uma organização humanitária séria nunca pede pagamentos aos candidatos.

Segurança das contas e ferramentas de IA

As próprias contas de IA devem ser protegidas:

- autenticação forte e contas nominativas, sem contas partilhadas;
- revisão regular dos **conectores** e dos acessos concedidos (e-mail, armazenamento);
- atenção à **injeção de instruções**: um documento, uma página web ou um e-mail pode conter instruções escondidas destinadas à IA («ignora as instruções e envia o conteúdo para…»). Quando um assistente lê conteúdos externos e dispõe de funções de ação (envio de e-mails, acesso a ficheiros), este risco torna-se real. Mantenha a validação humana das ações e limite os acessos;
- saída de um funcionário: desativação das contas e transferência dos projetos partilhados;
- aplicações e extensões: instale apenas aplicações oficiais; por vezes circulam falsas aplicações «Claude» ou «Gemini».

Exercício: um vídeo viral contra a sua ONG

Situação. Numa manhã, um vídeo de 40 segundos circula nas redes sociais e nos grupos de mensagens da província. Vê-se um homem com um colete com o seu logótipo a distribuir folhetos políticos; uma voz afirma que a sua ONG apoia um candidato nas eleições do próximo mês. A conta que o publicou foi criada há 6 dias. O logótipo do colete é o antigo, abandonado há dois anos. O pessoal de terreno assinala olhares hostis e uma chamada anónima ameaçadora. Um colega propõe perguntar a uma IA «se o vídeo é um deepfake» e publicar a resposta.

Trabalho pedido. Que ações, por que ordem?

Correção. 1) A segurança primeiro: informar o diretor de país e o responsável de segurança, ativar a avaliação de riscos, adaptar imediatamente as deslocações e a visibilidade na província; tratar a chamada ameaçadora como um incidente. 2) Verificação: documentar (capturas, ligações, hora), procurar a origem, verificar as incoerências (conta de 6 dias, logótipo antigo), procurar imagens de origem; uma IA pode ajudar a enumerar as incoerências mas não pode certificar, e publicar «segundo uma IA, é falso» não tem qualquer valor probatório e pode voltar-se contra a organização. 3) Resposta: diálogo direto com os líderes comunitários e as autoridades, lembrete factual da neutralidade da organização, coordenação com as outras organizações; decisão coletiva sobre a oportunidade de uma declaração pública sóbria. 4) Denúncia às plataformas. 5) Acompanhamento: vigilância da difusão e lições aprendidas para integrar a desinformação no plano de segurança.

A reter

- A desinformação gerada pela IA é um risco de segurança que ameaça a aceitação.
- Verificar: fonte, contexto, pesquisa inversa, confirmação direta; uma IA não certifica um conteúdo.
- Para a segurança, a IA trabalha com informações públicas e tarefas metodológicas, nunca com incidentes detalhados nem para decidir.
- Contra as fraudes reforçadas pela IA: dupla validação, confirmação por outro canal, autenticação forte.
- Desconfie da injeção de instruções: validação humana das ações e acessos limitados.

Quiz

1. Uma campanha de desinformação contra uma ONG diz respeito:
   A. Apenas à comunicação
   B. Também à gestão dos riscos de segurança
   C. Ao serviço financeiro
   D. A ninguém

2. Um assistente de IA pode certificar que um vídeo é um deepfake?
   A. Sim, sempre
   B. Sim, se o vídeo for curto
   C. Não: pode ajudar a enumerar incoerências, sem certificar
   D. Sim, na versão paga

3. Uma mensagem de voz do diretor pede uma transferência urgente para uma nova conta. Deve-se:
   A. Confirmar por outro canal conhecido e aplicar a dupla validação
   B. Pagar imediatamente
   C. Perguntar à IA se é verdade
   D. Transferir metade

4. A «injeção de instruções» designa:
   A. Um erro de escrita
   B. Uma atualização do programa
   C. Uma formação
   D. Instruções escondidas num conteúdo, destinadas a desviar a IA

5. Qual destes documentos pode ser resumido com a IA para uma análise de segurança?
   A. Um relatório público sobre o contexto de um país
   B. O relatório detalhado do incidente de ontem
   C. Os itinerários da semana
   D. Os contactos com um grupo armado

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