Análise Crítica de um Dilema Ético em IA

Caso: Fraudes com deepfake de voz e vídeo

Em 2019, criminosos usaram IA de clonagem de voz para enganar o CEO de uma empresa de energia no Reino Unido, que acreditou estar falando com o chefe da matriz. Ele autorizou uma transferência de US$ 243.000 para uma conta fraudulenta. Casos semelhantes cresceram, inclusive em golpes contra pessoas físicas e campanhas de desinformação.

Framework usado

- NIST AI Risk Management Framework (AI RMF 1.0): governar, mapear riscos, medir e mitigar. - Princípios da UNESCO e do ACM Code of Ethics: evitar dano, proteger direitos humanos, combater desinformação.

Aplicando o NIST AI RMF ao caso

1) Governança

Falta de governança clara sobre quem pode treinar, distribuir e usar ferramentas de clonagem de voz/vídeo.

2) Mapear riscos

Uso indevido para fraude bancária, manipulação política e pornografia de vingança.

3) Medir

Necessidade de métricas como taxa de falsos positivos em detectores de deepfake.

4) Mitigar

Opções: watermarking digital, regulação, detectores acessíveis, educação digital.

Análise ética

Desenvolvedores têm obrigação de evitar danos previsíveis e restringir usos nocivos. Uso fraudulento de IA viola direitos humanos.

Posição profissional fundamentada

O uso de IA generativa não é antiético por si só, mas sem controles técnicos, jurídicos e sociais, os riscos superam os benefícios.

Recomendações práticas

- Desenvolvedores: restrições de uso, watermarking, termos claros. - Empresas: detectores automáticos de deepfakes. - Governos: leis específicas para fraude com IA. - Sociedade: alfabetização midiática.

Conclusão

Fraudes com deepfake mostram como a IA pode ser uma arma de manipulação. É necessário aplicar governança, regulação e responsabilidade ética.