



Cérebro do Mentor

Documentação aberta

EBOOK · EDIÇÃO DE REFERÊNCIA

Ética no Cérebro do Mentor e a Ética Geral de IA

Transcrição organizada dos trechos: o caso do drone russo, a Recomendação da UNESCO e o princípio que organiza o produto.

3

partes

2021

Recomendação
UNESCO

03/08

medido em 2026

Sumário

- 1 O caso do drone russo — ética geral e ausência de julgamento humano
- 2 A Recomendação da UNESCO sobre Ética da Inteligência Artificial
- 3 Ética no Cérebro do Mentor — princípio organizador e linhas vermelhas

As marcas [F1], [F6], [F8] e [F2] ao fim dos parágrafos indicam a fonte de cada trecho, preservadas como na transcrição original.

O caso do drone russo — ética geral e ausência de julgamento humano

Kateryna Bondar, pesquisadora sênior do Wadhwani AI Center, do Center for Strategic and International Studies (CSIS), em Washington, avaliou que este seria o primeiro caso documentado de mortes de civis causadas por um drone russo equipado com sistema de auto-alvo. Sistemas assim são treinados em milhares de imagens para reconhecer categorias como “rio”, “caminhão militar” ou “pessoa”, e usam as câmeras para caçar esses objetos com precisão maior que a de um piloto humano. [F1]

O significado do episódio, segundo os analistas ouvidos: a IA já vinha sendo usada nas etapas finais de ataques planejados por humanos, mas aqui a máquina interpreta sozinha o que constitui alvo legítimo — e, por extensão, as leis da guerra. O Comitê Internacional da Cruz Vermelha alerta que armas autônomas removem o julgamento humano e criam riscos legais e éticos. O coronel ucraniano Serhiy Minaiev afirmou ao NYT que isso representa um risco para o mundo inteiro. O caso também expõe a dificuldade de negar tecnologia a adversários dos EUA: o minicomputador não é vendido na Rússia, mas está disponível em mercados de revenda. [F1]

Aqui a máquina interpreta sozinha o que constitui alvo legítimo — e, por extensão, as leis da guerra.

A Recomendação da UNESCO sobre Ética da Inteligência Artificial

Aprovada em 23 de novembro de 2021.

Esta Recomendação aborda questões éticas relacionadas ao campo da inteligência artificial, na medida em que aquelas estão dentro do mandato da UNESCO. Ela aborda a ética da IA como uma reflexão normativa sistemática, com base em um marco holístico, abrangente, multicultural e em evolução de valores, princípios e ações interdependentes que podem orientar as sociedades para que lidem de forma responsável com os impactos conhecidos e desconhecidos das tecnologias de IA sobre seres humanos, sociedades, meio ambiente e ecossistemas, oferecendo-lhes uma base para aceitar ou rejeitar essas tecnologias. Ela considera a ética como uma base dinâmica para a avaliação e a orientação normativa das tecnologias de IA, fazendo referência à dignidade humana, ao bem-estar e à prevenção de danos — como uma bússola e tendo como fundamento a ética da ciência e da tecnologia. [F6]

Esta Recomendação não tem a ambição de fornecer uma definição única de IA, uma vez que tal definição precisaria se modificar ao longo do tempo, em conformidade com os desenvolvimentos tecnológicos. Em vez disso, sua ambição consiste em abordar os aspectos dos sistemas de IA que têm relevância ética central. Portanto, esta Recomendação aborda os sistemas de IA como sistemas que têm capacidade de processar dados e informações de uma forma que se assemelha ao comportamento inteligente e, normalmente, inclui aspectos de raciocínio, aprendizagem, percepção, previsão, planejamento ou controle. [F6]

Questões éticas em relação aos sistemas de IA dizem respeito a todos os estágios do ciclo de vida de tais sistemas, entendidos aqui como abrangendo desde pesquisa, design e desenvolvimento até implementação e uso, incluindo manutenção, operação, comércio, financiamento, monitoramento e avaliação, validação, fim de uso, desmontagem e término. [F6]

Os sistemas de IA levantam novas espécies de questões éticas que incluem, mas não se limitam a, seu impacto na tomada de decisões, emprego e trabalho, interação social, assistência médica, educação, meios de comunicação, acesso à informação, exclusão digital, dados pessoais e proteção ao consumidor, meio ambiente, democracia, Estado de direito, segurança e policiamento, dupla utilização, e direitos humanos e liberdades fundamentais, incluindo liberdade de expressão, privacidade e não discriminação. Além disso, novos desafios éticos são criados pela possibilidade de que os algoritmos de IA reproduzam e reforcem vieses existentes e, assim, agravem formas já existentes de discriminação, preconceitos e estereótipos. [F6]

No longo prazo, os sistemas de IA podem desafiar o sentido especial de experiência e capacidade de ação dos humanos, o que levanta preocupações adicionais sobre, entre outros, autocompreensão humana, interação social, cultural e ambiental, autonomia, capacidade de ação, valor e dignidade. [F6]

Como a Recomendação foi determinada

Em novembro de 2019, a Conferência Geral da UNESCO, em sua 40ª sessão, aprovou a 40 C/Resolução 37, pela qual determinou que a diretora-geral “elaborasse um instrumento internacional normativo sobre a ética da inteligência artificial (IA) na forma de uma recomendação”, o qual deve ser apresentado à Conferência Geral em sua 41ª sessão, em 2021. [F8]

Reconhecendo que o desenvolvimento de tecnologias de IA exige um aumento proporcional de dados, mídias e alfabetização informacional, bem como acesso a fontes de informação independentes, pluralistas e confiáveis, inclusive como parte dos esforços para mitigar os riscos de desinformação, informações falsas, discurso de ódio e danos causado por meio da utilização indevida de dados pessoais. [F8]

Observando que um marco normativo para as tecnologias de IA e suas implicações sociais encontra o seu fundamento nos marcos jurídicos internacionais e nacionais, nos direitos humanos e nas liberdades fundamentais, na ética, na necessidade de acesso a dados, informações e conhecimento, na liberdade de pesquisa e inovação, no bem-estar humano, ambiental e ecossistêmico, e vincula valores e princípios éticos aos desafios e às oportunidades relacionadas às tecnologias de IA, com base no entendimento comum e nos objetivos compartilhados. [F8]

Também convencida de que os padrões éticos aceitos em âmbito global para as tecnologias de IA, em total respeito ao direito internacional, em particular às leis de direitos humanos, podem desempenhar um papel fundamental no desenvolvimento de normas relacionadas à IA em todo o mundo. [F8]

Ética no Cérebro do Mentor

— princípio organizador e linhas vermelhas

COMO A ÉTICA É ENDEREÇADA NO CÉREBRO DO MENTOR

Medido em **03/08/2026**. Cada afirmação aqui aponta para o código ou para um número verificável. As que não têm lastro estão na seção **O que ainda não está resolvido** — e essa seção existe de propósito. [F2]

O produto fala com profissional de saúde. Isso muda o que se pode afirmar: uma frase otimista sobre conformidade não é marketing agressivo, é risco jurídico que chega no mentor e, através dele, no paciente dele. [F2]

O princípio que organiza tudo

O produto não afirma o que não mede. [F2]

Não é postura moral, é engenharia: afirmação sem medição envelhece em silêncio. Já aconteceu. O `README` dizia que “o único elo em API externa é o Composer” — era verdade quando `extraction_model` era `ollama`, deixou de ser quando alguém trocou para `openai`, e o texto ficou. Ninguém percebeu por semanas, porque texto errado continua legível. [F2]

Por isso as afirmações de conformidade têm **teste**

(`tests/test_readme_soberania.py`), e o teste compara com a configuração real nos dois sentidos: se alguém religar o modelo local, ele cobra a atualização do texto na direção contrária. Subestimar também é mentir. [F2]

As linhas vermelhas

Nunca dizer, em nenhuma superfície — produto, site, proposta:

PROIBIDO	POR QUÊ
"zero alucinação" / "sem alucinação"	absoluto indefensável; risco legal em saúde
"conforme à LGPD"	veredito jurídico que o produto não pode emitir
"aderente ao marco legal da IA"	não há implementação; ver pendências
"air-gapped"	hoje é air-gappable , não air-gapped
"sem viés" / "imparcial"	medimos viés de acervo , não viés social
"o produto roda no Brasil"	só a chamada de conformidade roda cá; extração, embeddings e composição saem
"certificado ISO 27001"	a Maritaca declara SGSI <i>estruturado segundo os requisitos</i> , não certificação
"a Maritaca não retém nada"	resíduo em log de erro pode persistir até 30 dias

Três dessas estão travadas por teste automatizado

(`test_nenhum_selo_afirma_conformidade_legal` ,
`test_selo_de_fontes_ao_fala_em_vies_social` ,
`test_nao_afirma_residencia_de_dado_no_brasil`). As outras dependem de disciplina humana. [F2]