Preocupações Éticas em IA

Professora Dra. Edna Dias Canedo

Universidade de Brasília (UnB), Departamento de Ciência da Computação (CIC)
P.O. Box 4466, Brasília–DF, Brasil
E-mail: ednacanedo@unb.br e edna.canedo@gmail.com
Laboratório de Engenharia de Software (LES)
https://les.unb.br/

22 de julho de 2025

Agenda

Diretrizes Éticas e Princípios Éticos em IA

Diretrizes Éticas em IA

- As diretrizes são formuladas por diferentes atores da sociedade (39 diretrizes).

Grupo 1 – Membros da Sociedade Civil:

- **Associações Profissionais:** elaboram códigos de conduta e ética para orientar a atuação técnica de engenheiros, cientistas de dados e desenvolvedores.
- **Sociedade Civil e Grupos de Advogados:** produzem diretrizes voltadas à advocacia, ativismo e defesa de direitos, além de promover o debate público.
- **Organizações Sem Fins Lucrativos:** abordam a ética com foco em justiça social, equidade e sustentabilidade.
- **Academia:** contribui com reflexões filosóficas, jurídicas e técnicas, promovendo a articulação entre pesquisa e prática.

Diretrizes Éticas em IA

Grupo 2: Organizações nacionais e internacionais – políticas e regulamentações em diferentes setores.

Grupo 3: Indústria – elabora diretrizes para nortear o desenvolvimento interno da organização e o uso de IA.

- Cada diretriz propõe princípios éticos específicos.
- A tradução desses princípios **em práticas concretas** é um desafio.

Princípios Éticos em IA

ID	Princípio	Descrição
P1	**Transparência**	Os sistemas devem operar de forma compreensível e auditável.
P2	Justiça	Evitar discriminações–tratamento equitativo entre indivíduos e grupos.
P3	**Responsabilidade**	Os criadores devem se responsabilizar pelos impactos dos sistemas.
P4	**Privacidade**	Proteção dos dados pessoais, respeitando leis como a LGPD e a GDPR.
P5	**Segurança**	Garantir que os sistemas sejam protegidos contra ataques e falhas.
P6	**Beneficência**	O objetivo dos sistemas deve ser gerar benefícios sociais e individuais.
P7	Autonomia Humana	Preservação da autonomia dos usuários nas interações com os sistemas.
P8	**Dignidade humana**	Respeitar os direitos e a integridade dos indivíduos.
P9	Inclusão	Promover acessibilidade e não exclusão de usuários.
P10	**Sustentabilidade**	Considerar impactos ambientais e sociais no ciclo de vida dos sistemas.
P11	Explicabilidade	Devem permitir compreensão das ações e inferências dos sistemas.
P12	**Confiabilidade**	Os sistemas devem ser consistentes e funcionar conforme esperado.
P13	Mitigação de riscos	Prevenir e minimizar danos aos usuários e sociedade.
P14	Proporcionalidade	As ações dos sistemas devem ser adequadas aos fins e riscos envolvidos.
P15	**Prestação de contas**	Mecanismos claros para atribuir responsabilidades e revisar decisões.
P16	Consentimento informado	Os usuários devem estar cientes e consentir com o uso de seus dados.
P17	Evitabilidade de dano	Projetar sistemas para evitar danos a pessoas ou grupos.
P18	Confidencialidade	Garantir que informações sensíveis não sejam acessadas indevidamente.
P19	**Liberdade**	Garantir o respeito aos direitos civis, liberdade de expressão e pensamento.
P20	**Solidariedade**	Promover a distribuição justa dos benefícios e riscos da IA.
P21	Prosperidade	Favorecer o desenvolvimento social e econômico coletivo por meio da IA.

Avanços em IA e Modelos de Linguagem

Avanços em IA e Modelos de Linguagem

Modelos generativos de linguagem:

- Capazes de compreender a linguagem natural e gerar textos com qualidade.
- Reconhecer padrões, fazer inferências e gerar respostas a diferentes tipos de entrada.
- Viabilizaram aplicações, que vão desde chatbots e assistentes virtuais a serviços de tradução e ferramentas de geração de conteúdo.
- Chatbots capazes de interagir com usuários de maneira semelhante à interação humana.
  - Fundamentais na otimização do atendimento ao usuário, suporte técnico e respostas a consultas de forma ágil e eficaz.

Impactos dos LLMs

- Promovem uma comunicação e compreensão mais eficazes entre pessoas de diferentes origens linguísticas.
- Permitem respostas rápidas a emergências/crises em regiões que falam idiomas com poucos recursos ou dialetos indígenas.
- Geram texto coerente e relevante ao contexto – valiosos no domínio da geração de conteúdo.
- Produção de diferentes tipos de conteúdo — artigos, revisões, postagens em redes sociais e materiais de marketing.

Impactos dos LLMs

- A crescente adoção dos LLMs exige atenção aos vieses presentes nesses modelos e aos seus impactos sociais.
- Esses modelos aprendem com grandes volumes de dados, que frequentemente refletem preconceitos estruturais.
- Como consequência, podem produzir respostas tendenciosas em áreas sensíveis — como recrutamento, segurança pública e saúde — reforçando estereótipos e perpetuando desigualdades.
- Essa problemática tem gerado preocupação entre especialistas e sociedade civil, exigindo abordagens éticas no desenvolvimento e uso de LLMs.

Fatores que Contribuem para Viés em Modelos de IA

Fatores que Contribuem para Viés em Modelos de IA

Fatores	Descrição
Dados de Treinamento	Viés pode estar no conteúdo ou na seleção dos dados.
Algoritmos	Priorizam certos padrões, podendo ampliar desigualdades.
Rotulagem e Anotação	Subjetividade dos anotadores influencia os dados.
Decisões de Design do Produto	Casos de uso e interfaces que priorizam certos públicos.
Decisões de Política	Regras embutidas podem favorecer ou restringir comportamentos.

Fatores que Contribuem para Viés em Modelos de IA

- LLMs podem reproduzir e amplificar diferentes tipos de viés.

Tipo de Viés	Descrição
Demográficos	Super ou sub-representação de grupos (gênero, raça, etnia), gerando respostas tendenciosas.
Culturais	Reforço de estereótipos e preconceitos culturais aprendidos nos dados.
Linguísticos	Desempenho inferior em idiomas pouco representados ou dialetos minoritários.
Temporal	Recorte limitado no tempo prejudica a cobertura de eventos atuais ou históricos.
De Confirmação	Reforça crenças existentes ao gerar respostas alinhadas com visões dominantes nos dados.
Ideológico/Político	Pode favorecer ideologias específicas, amplificando polarizações presentes nos dados.

Mitigação de viés com abordagens Human-in-the-Loop

Mitigação de viés com abordagens Human-in-the-Loop

- Podem ser mitigados através de abordagens Human-in-the-Loop (HITL).
  - Envolvem a participação humana, com feedback ou supervisão ao longo do desenvolvimento e da implantação do modelo.
- Algumas formas de integrar HITL para mitigar viés:
  1. **Curadoria de Dados de Treinamento**: Humanos podem atuar na curadoria e anotação de dados de treinamento de alta qualidade e diversidade.
     - Identificar e corrigir vieses, garantir equilíbrio de perspectivas e reduzir conteúdos ofensivos ou controversos.
  2. **Ajuste Fino do Modelo (Fine-Tuning)**: Especialistas podem orientar o processo de ajuste fino por meio de feedback sobre as saídas do modelo.
     - Ajuda a evitar respostas enviesadas ou incorretas.

Mitigação de viés com abordagens Human-in-the-Loop

3. **Avaliação e Feedback**: Revisores humanos podem avaliar o desempenho do modelo e fornecer feedback aos desenvolvedores.
   - Pode identificar e tratar problemas relacionados a viés, promovendo a melhoria contínua do modelo.
4. **Moderação em tempo real**: Moderadores humanos podem monitorar e revisar as saídas em tempo real, intervindo quando necessário para corrigir respostas enviesadas ou inadequadas.
   - Especialmente útil em aplicações sensíveis ou de alto impacto.
5. **Customização e controle**: Usuários podem ter opções para ajustar o comportamento do modelo, adaptando as respostas às suas preferências ou necessidades específicas.
   - Permite mitigar o viés nas respostas ao adaptar o modelo para contextos/ domínios específicos.

Preocupações Éticas em IA/LLMs

Preocupações Éticas em IA/LLMs

- Embora abordagens como Human-in-the-Loop contribuam para mitigar o viés, é importante reconhecer que essas estratégias não os eliminam completamente.
- O viés pode ter diversas origens: dados de treinamento, ajustes técnicos e até intervenções humanas podem perpetuar desigualdades.
- Essa complexidade exige uma visão ética mais ampla, que considere impactos sociais, riscos de alucinação e desinformação.
- Princípios como justiça, equidade e responsabilidade devem guiar o design e a implementação de sistemas baseados em IA/LLMs.

Preocupações Éticas em IA/LLMs

- Integração da IA em múltiplos domínios da sociedade amplia seu impacto sobre indivíduos, grupos e estruturas sociais.
- Desenvolvedores, pesquisadores e usuários têm a responsabilidade de promover **tratamento equitativo**, evitando reforçar ou introduzir vieses sistêmicos.
- Transparência e confiança tornam-se pilares fundamentais para garantir que a **IA opere de forma justa, ética e responsável**.
- Transparência não é apenas técnica, mas também ética – ela permite que sistemas de IA demonstrem alinhamento com princípios normativos e regulamentações.

Preocupações Éticas em IA/LLMs

- Transparência, equidade e responsabilidade são importantes para IA ética e confiável.
- A União Europeia propôs um conjunto de **Diretrizes Éticas para IA** – destacando os princípios **transparência, justiça, privacidade, autonomia humana e prestação de contas**.
- Mas como transformar esses princípios na prática?
  - Desenvolvendo ferramentas automatizadas que apoiem o monitoramento ético durante o ciclo de vida dos sistemas.
  - Estimulando a participação social, com espaços abertos para questionamento e validação dos resultados gerados por modelos de IA.

Preocupações Éticas em IA/LLMs

- Relacionadas ao princípio **Privacidade**:
  - LLMs podem comprometer a privacidade em diferentes formas:
    - Vazamento de informações sensíveis durante a geração de respostas.
    - Uso indevido de dados pessoais em conjuntos de treinamento.
    - Inserção acidental de código corporativo ou estratégias empresariais por desenvolvedores.
  - Líderes de conformidade devem estar atentos aos dados inseridos em LLMs públicos – podem ser usados no treinamento.
    - Em 2023, funcionários da Samsung copiaram código interno no ChatGPT para depurar um problema e depois descobriram que essas informações foram retidas nos servidores do modelo.

Preocupações Éticas em IA/LLMs

- **Propriedade Intelectual** (Responsabilidade): geração de conteúdo com base em dados de terceiros, sem controle autoral ou atribuição de fontes.
- **Colaboração no Trabalho** (Solidariedade): automação de tarefas pode reduzir interações humanas, enfraquecendo laços profissionais, empatia e aprendizado coletivo.
- **Substituição da Força de Trabalho** (Beneficência): LLMs otimizam tarefas com menos pessoas, ampliando desigualdades e risco de exclusão profissional.

Preocupações Éticas em IA/LLMs

- **Alucinações** (Risco Ético-Cognitivo): **Ocorrem quando LLMs geram informações incorretas, inventadas ou desconectadas da realidade.**
- Modelos podem produzir textos factualmente errôneos, com grande confiança e fluidez.
- Falta de referenciamento confiável dificulta a validação das respostas, como no caso real de um advogado que apresentou casos jurídicos fictícios gerados por IA.
- Podem surgir de ruído ou inconsistência nos dados de treinamento, levando à geração de padrões distorcidos.

Preocupações Éticas em IA/LLMs

- **Alucinações podem gerar impactos éticos sérios e duradouros.**
  - **Conteúdo Discriminatório e Tóxico** (Justiça): reforça estereótipos, marginaliza grupos sociais e pode gerar exclusão ou discriminação algorítmica.
  - **Vazamento de Dados** (Privacidade): exposição de informações sensíveis aprendidas durante o treinamento.
  - **Desinformação** (Confiança): conteúdo falso pode influenciar decisões em saúde, educação, direito e política.

Preocupações Éticas em IA/LLMs

Medidas recomendadas para reduzir alucinações:

- Aprimorar dados de treinamento: mais qualidade e diversidade.
- Verificar fontes: adotar mecanismos de checagem de fatos.
- Refinar o modelo: com revisão humana especializada (fine-tuning).
- Moderação em tempo real: para identificar e corrigir respostas problemáticas.
- Ciclos de feedback: incorporar avaliações contínuas de usuários.
- Transparência: explicar como as respostas são geradas e com base em que dados.

Importante: nenhuma medida isolada é suficiente — é a combinação que garante segurança e confiabilidade.

Conclusão

Conclusão

- A supervisão humana é fundamental para garantir o uso ético da IA.
- Mitigar vieses e envolver usuários contribui para sistemas mais inclusivos, justos e acessíveis.
- Depende da nossa vigilância ativa, responsabilidade ética e ação contínua.
- Questões éticas em IA impactam diretamente decisões políticas, processos sociais e vidas humanas.
- **Precisamos agir agora** para que a IA beneficie a todos — e não apenas alguns.

A ética não freia a inovação — ela legitima seu futuro.

Obrigada!
Perguntas?

Professora Edna Dias Canedo

ednacanedo@unb.br