

Título do plano	Ética e vieses: identificar, mitigar, comunicar			Sequência 17
Etapa/Ano	Ensino Médio: ☐ 1º Ano ☐ 2º Ano ☒ 3º Ano			
Nível de Maturidade [link]	Escola	☐ Emergente ☐ Básico ☒ Intermediário ☐ Avançado	Docente	☐ Básico ☒ Intermediário ☐ Avançado
Competências [link]	☒ C1 ☐ C2 ☐ C3 ☒ C4 ☒ C5			
Objetos de Conhecimento	- Vieses em sistemas de Inteligência Artificial: tipologia (dado, modelo, relato) e suas origens. - Discriminação algorítmica e seus efeitos sobre grupos sociais. - Análise crítica de dataset: representatividade, ausências, vies histórico, vies de seleção. - Análise de desempenho de modelos por subgrupos: leitura básica de matriz de confusão. - Vieses comunicativos: superpromessa, neutralidade aparente, ocultamento de limites. - Estratégias de mitigação e boas práticas em projetos de IA.			
Recursos Educacionais	- Tabela de Pontos Críticos e Síntese de Auditoria produzidas pelos grupos no Plano 16. - Pacote do projeto integrador entregue no Plano 15: dataset em CSV, datasheet (Plano 8), arquivo .ows do Orange e relatório/painel do MVP. - Estação 1 (debate): cópia impressa ou tela com as estatísticas descritivas do dataset (médias, frequências, distribuições por categoria) e cinco cartões de análise crítica. - Estação 2 (digital): computador ou tablet com Orange instalado, abrindo o .ows do projeto, mais um workflow que permita comparar a matriz de confusão para dois subgrupos diferentes do dataset. - Estação 3 (mão na massa): cópia impressa do relatório/painel do MVP, marca-texto e folha com o template "Versão original / Versão revisada". - Folha da Carta de Mitigação (uma por grupo). - Cronômetro visível para a turma e cartazes simples identificando cada estação na sala.			
Componentes Curriculares Relacionados	☐ Arte ☒ Língua Port. ☐ Ed. Física	☐ Geografia ☐ História ☒ Matemática	☐ Língua Inglesa ☐ Ciências ☒ Computação	
Palavras-chave	vieses; discriminação algorítmica; ética em IA; mitigação; representatividade; matriz de confusão; transparência; rotação por estações			
Perguntas Importantes	- Quais grupos podem ter sido sub-representados ou ignorados pelos dados que coletamos? - Nosso modelo erra mais para algum subgrupo do que para outro? Como saber? - O nosso relatório descreve o sistema de forma honesta — incluindo o que ele não sabe ou faz mal? - De onde vêm os vieses: dos dados, do modelo, da forma como contamos a história, ou dos três? - O que é possível fazer para mitigar cada tipo de vies sem refazer o projeto inteiro? - O que é uma boa prática de comunicação responsável sobre os limites de um sistema de IA?			
Encontros	1	Total de Horas-aula	2	
Objetivos	- Distinguir, com vocabulário próprio, três grandes famílias de vieses em sistemas de IA: vieses do dado, vieses do modelo e vieses do relato. - Identificar vieses concretos no projeto integrador construído ao longo do ano, em cada uma das três famílias.			

	- Realizar uma análise comparativa elementar do desempenho do modelo entre subgrupos, usando o Orange e leitura básica de matriz de confusão. - Reescrever de forma mais honesta passagens do relatório do MVP que contenham super promessas ou ocultamentos. - Formular estratégias de mitigação implementáveis para cada viés identificado, priorizando-as por impacto. - Produzir uma Carta de Mitigação que sirva de insumo para o Model Card do Plano 18.
Habilidades Relacionadas	BNCC: (EM13CHS502) Analisar situações da vida cotidiana, estilos de vida, valores, condutas etc., desnaturalizando e problematizando formas de desigualdade, preconceito, intolerância e discriminação. (EM13CHS503) Identificar diversas formas de violência (física, simbólica, psicológica etc.), suas causas, significados e usos políticos, ideológicos e disciplinadores, discutindo o papel de agentes públicos e privados em sua produção e disseminação. (EM13CHS504) Analisar e avaliar os impasses ético-políticos decorrentes das transformações culturais, sociais, históricas, científicas e tecnológicas no mundo contemporâneo e seus desdobramentos nas atitudes e nos valores de indivíduos, grupos sociais, sociedades e culturas. (EM13LP01) Relacionar o texto, tanto na produção como na leitura/escuta, com suas condições de produção e seu contexto sócio-histórico de circulação. BNCC Computação: (EM13CO10) Conhecer os fundamentos da Inteligência Artificial, comparando-a com a inteligência humana, analisando suas potencialidades, riscos e limites. (EM13CO11) Compreender o impacto das tecnologias digitais e da Inteligência Artificial no trabalho e na sociedade. (EM13CO13) Discutir aspectos éticos, sociais e ambientais relacionados ao desenvolvimento e uso de sistemas computacionais e de IA. Referencial Curricular IA no EM: (EM02IA0X) Reconhecer questões éticas que surgem com a introdução da IA em processos sociais e produtivos. Identificar e questionar vieses presentes em sistemas baseados em IA, distinguindo origens nos dados, no modelo e na forma de comunicar resultados. Compreender a relação entre representatividade da amostra, qualidade do modelo e justiça das decisões algorítmicas. Reconhecer estratégias básicas de mitigação de vieses em projetos de IA aplicados a contextos reais.
Práticas Pedagógicas Inovadoras	☒ Aula Enriquecida com Tecnologia ☐ Ensino Híbrido: Sala de aula invertida ☒ Ensino Híbrido: rotação por estação ☐ Ensino Híbrido: rotação individual ☐ Aulas Mão na massa ☐ Aprendizagem baseada em projetos ☐ IA Desconectada

Sequência Didática

1. Abertura coletiva — "De onde vêm os vieses?" (≈15 min)

Inicie o encontro pedindo a cada grupo que coloque sobre a mesa a Tabela de Pontos Críticos e a Síntese de Auditoria produzidas no Plano 16. Esses materiais não são acessórios — são o ponto de partida deste plano. Avise a turma de que vamos pegar aquela tabela, especialmente a coluna "O que pode dar errado", e olhar para ela com lupa, com vocabulário mais preciso e com estratégias concretas para reduzir os riscos identificados.

Escreva no quadro a tipologia que organiza o encontro: "Viés é uma palavra que descreve um desvio sistemático, e em sistemas de IA esse desvio costuma vir de três lugares diferentes". Apresente as três famílias com exemplos curtos do próprio projeto integrador:

- Vieses do dado — quando a base que treinou o sistema está incompleta, mal balanceada ou reflete desigualdades históricas. Exemplo no projeto: se a coleta do Plano 6 contemplou só uma escola, o sistema pode não funcionar bem em outras.
- Vieses do modelo — quando o algoritmo, mesmo com dados razoáveis, aprende padrões que prejudicam algum subgrupo, ou erra mais para uma classe do que para outra. Exemplo: a árvore de decisão do Plano 13 pode prever "prioridade alta" com mais precisão para um perfil do que para outro.
- Vieses do relato — quando, na hora de comunicar o que o sistema faz, exageramos suas capacidades, escondemos suas limitações ou fingimos que ele é "neutro" e "objetivo". Esse viés mora na linguagem do relatório e do painel do Plano 15.

Conte, em três a cinco minutos, dois exemplos reais que dão peso ético à discussão. O primeiro é brasileiro: estudos do Centro de Estudos de Segurança e Cidadania (CESeC) e da Defensoria Pública mostraram que sistemas de reconhecimento facial usados em policiamento, em estados como Bahia e Rio de Janeiro, abordaram majoritariamente pessoas negras — não porque haja racismo no código, mas porque os dados de treinamento e o contexto de uso reproduziram desigualdades estruturais já existentes. O segundo é internacional e ajuda a generalizar a discussão: o sistema COMPAS, usado em tribunais norte-americanos para estimar risco de reincidência, foi auditado pela ProPublica e mostrou taxas mais altas de falsos positivos para pessoas negras do que para pessoas brancas. Em ambos os casos, o sistema "funcionava" no agregado, mas falhava sistematicamente para grupos específicos. *Esse é o tipo de problema que o trabalho de hoje quer aprender a detectar antes que ele cause dano.*

Encerre a abertura com a logística da rotação. Os mesmos grupos do Plano 16 percorrerão três estações de aproximadamente 25 minutos cada (cerca de 22 minutos de trabalho efetivo e 3 de transição). No fim, voltam às mesas para integrar os três diagnósticos em uma Carta de Mitigação, que será insumo direto do Plano 18 (Model Card).

2. Rotação por estações (≈75 min — 3 estações × 25 min)

Cada estação tem um modo de trabalho próprio (debate analítico, digital com Orange, mão na massa de reescrita) e produz um diagnóstico específico sobre uma das três famílias de vieses. Os três diagnósticos serão integrados na Carta de Mitigação final. Recomenda-se que o(a) professor(a) circule entre as estações, intervindo sobretudo na Estação 2 (onde dúvidas técnicas

sobre matriz de confusão costumam aparecer) e na Estação 3 (onde estudantes podem ter resistência a criticar o próprio relatório).

ESTAÇÃO 1 — Vieses do Dado

Modo: debate / análise crítica do dataset

Tempo sugerido: ≈25 min

Insumos:

- Estatísticas descritivas do dataset do projeto, impressas ou em tela.
- Datasheet do Plano 8 (documentação do dataset).
- Cinco cartões de análise

Tarefa:

O grupo lê os cinco cartões em voz alta, um a um, discute por aproximadamente 4 minutos cada e registra um viés concreto do dataset relacionado àquela pergunta.

Produto da estação:

Lista de cinco vieses do dado, um por cartão, em frases específicas (não genéricas) e com referência ao plano do projeto onde o viés se origina (Plano 5, 6, 7 ou 8).

ESTAÇÃO 2 — Vieses do Modelo

Modo: digital / aula com tecnologia |

Tempo sugerido: ≈25 min

Insumos:

- Computador ou tablet com Orange instalado, com o arquivo .ows do projeto aberto.
- Workflow auxiliar
- Quick reference card
- Folha impressa com perguntas

Tarefa:

O grupo escolhe uma variável de subgrupo do dataset que faça sentido sociologicamente. Em seguida:

- Roda o modelo no subgrupo A e anota a matriz de confusão correspondente, com atenção especial a Falsos Positivos e Falsos Negativos.
- Roda o modelo no subgrupo B e anota a matriz de confusão correspondente, do mesmo modo.

O grupo responde às perguntas da folha impressa.

Produto da estação:

Tabela "Onde nosso modelo é melhor / pior"

ESTAÇÃO 3 — Vieses do Relato

Modo: mão na massa / reescrita crítica |

Tempo sugerido: ≈25 min

Insumos:

- Cópia impressa do relatório e do painel do MVP entregues no Plano 15.
- Marca-texto ou caneta colorida.
- Folha com o template "Versão original / Versão revisada / Por que mudei"
- Lista de "armadilhas comunicativas" frequentes em relatos sobre IA.

Tarefa:

O grupo lê o próprio relatório/painel e usa o marca-texto para sublinhar trechos suspeitos de armadilhas comunicativas.

Em seguida, o grupo escolhe ao menos três trechos sublinhados e preenche a FOLHA TEMPLATE REVISÃO

Produto da estação:

Folha "Versão original / Versão revisada" com no mínimo três trechos reescritos.

3. Carta de Mitigação e fechamento (≈20 min)

Os grupos retornam às mesas originais com os três produtos parciais: a Lista de Vieses do Dado (Estação 1), a Tabela "Onde nosso modelo é melhor / pior" (Estação 2) e a folha de Reescrita do Relato (Estação 3). A tarefa final é integrar esses três produtos em uma Carta de Mitigação de uma página (conforme modelo), que será entregue como produção formal do plano e usada como insumo direto do Plano 18 (Model Card) e do Plano 20 (versão final).

Compartilhamento rápido (≈8 min dos 20):

Cada grupo apresenta, em até 60 segundos, três pontos: (a) o viés mais grave identificado em qualquer das três famílias, (b) a mitigação que considera mais urgente, e (c) um viés que reconheceram não ser mitigável neste momento. O(a) professor(a) anota brevemente no quadro um termo-chave por grupo, formando um pequeno mapa de turma sobre vieses e mitigações.

Conexão para o Plano 18 (≈2 min):

Encerre lendo a frase que sintetiza o sentido do encontro: "Um sistema de IA não fica mais justo só porque queremos que ele seja. Fica mais justo quando olhamos com lupa para os vieses, propomos mitigações concretas e contamos a história com honestidade." Anuncie que o próximo plano (18) parte das Cartas de Mitigação produzidas hoje para construir o Model Card do projeto — o documento que comunica oficialmente, e em linguagem acessível, o que o sistema faz, o que não faz, para quem funciona, sob que limites e sob que responsabilidade. Recolha (ou peça envio digital) das Cartas, que ficarão arquivadas em pasta visível à turma.

Avaliação

A avaliação deste plano é processual e baseada em evidências de produção das três estações e da síntese final. Critérios e peso:

- Lista de Vieses do Dado — Estação 1 (20%): cinco vieses identificados, um por cartão, em frases específicas e com referência ao plano do projeto onde o viés se origina.
- Tabela "Onde nosso modelo é melhor / pior" — Estação 2 (25%): comparação clara entre dois subgrupos, identificação dos erros predominantes em cada um, e resposta argumentada à pergunta ética final.
- Folha de Reescrita do Relato — Estação 3 (20%): mínimo de três trechos sublinhados, três versões revisadas e três justificativas, com identificação clara da armadilha comunicativa em cada caso.
- Carta de Mitigação final (25%): integração coerente dos três produtos parciais, com mitigações concretas, implementáveis e priorizadas; reconhecimento explícito do que não é mitigável.
- Postura ética e participação na rotação (10%): observação docente sobre engajamento crítico, qualidade das discussões em grupo e abertura para revisar o próprio projeto sem postura defensiva.

Sugere-se que a rubrica seja compartilhada com os estudantes no início do encontro, de preferência impressa em cada estação.

Material Complementar

Casos e estudos sobre vieses e discriminação algorítmica:

- CESeC — Centro de Estudos de Segurança e Cidadania. O Panóptico: monitor do reconhecimento facial no Brasil. Disponível em: <https://opanoptico.com.br/>
- ANGWIN, Julia; LARSON, Jeff; MATTU, Surya; KIRCHNER, Lauren. Machine Bias. ProPublica, 2016. Estudo do sistema COMPAS. Disponível em: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of FAT* 2018. Estudo seminal sobre vieses em reconhecimento facial.
- OBERMEYER, Ziad et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 2019. Caso emblemático de viés em algoritmo de saúde.

Marcos normativos e referenciais éticos:

- BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm
- UNESCO. Recomendação sobre a Ética da Inteligência Artificial, 2021. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000380455>
- OECD. AI Principles. Disponível em: <https://oecd.ai/en/ai-principles>

- BRASIL. Projeto de Lei 2338/2023 — Marco regulatório da Inteligência Artificial. Disponível em: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>

VICARI, R. M.; GALAFASSI, C.; BRACKMANN, C.; MIZUSAKI, L. IA@Escola. 2024. Disponível em: www.computacional.com.br/ia.

Plano 17 elaborado para o Ano 3 — Equipe do Módulo 4: Claudia, Aline e Ricardo.

