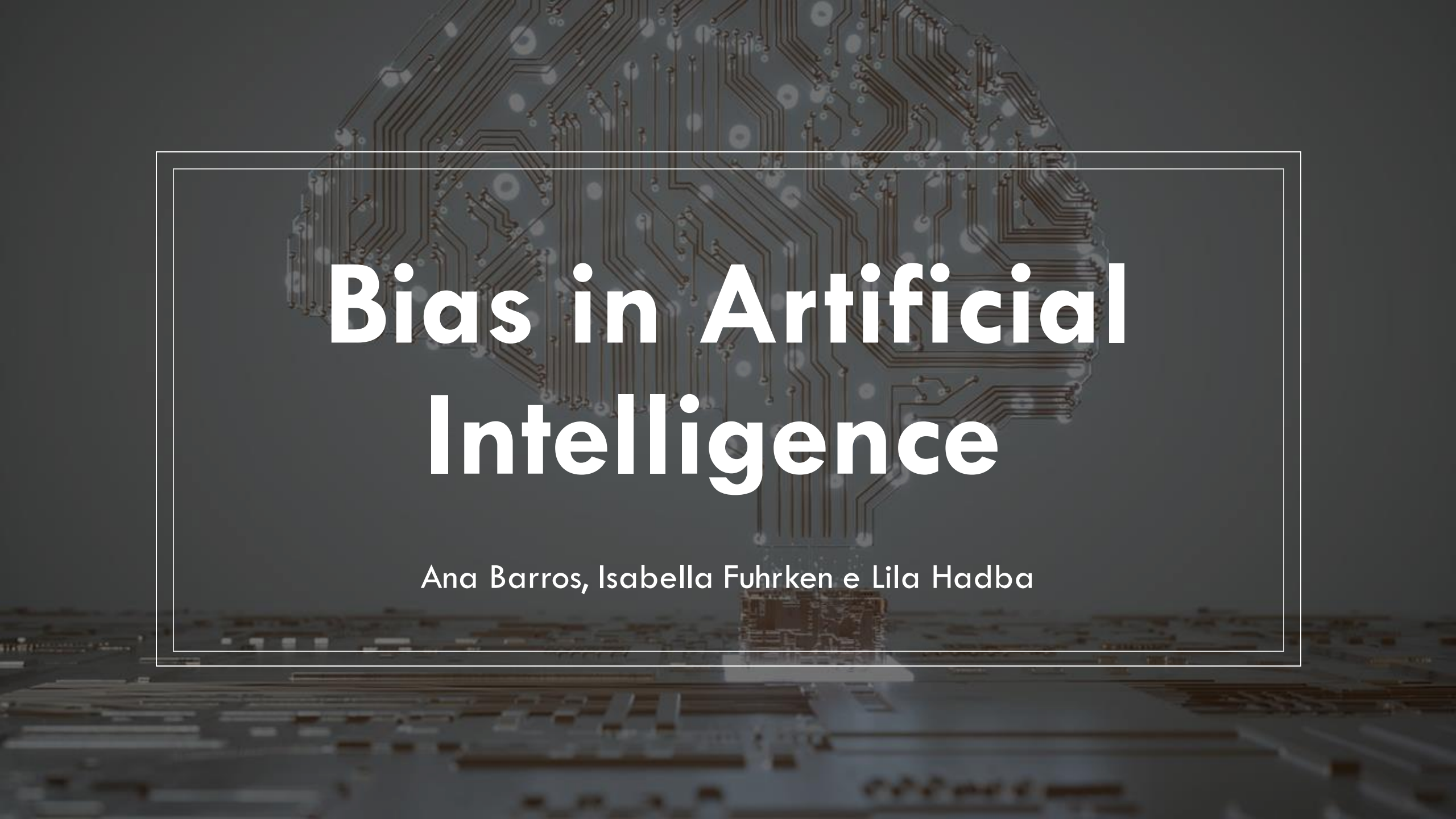
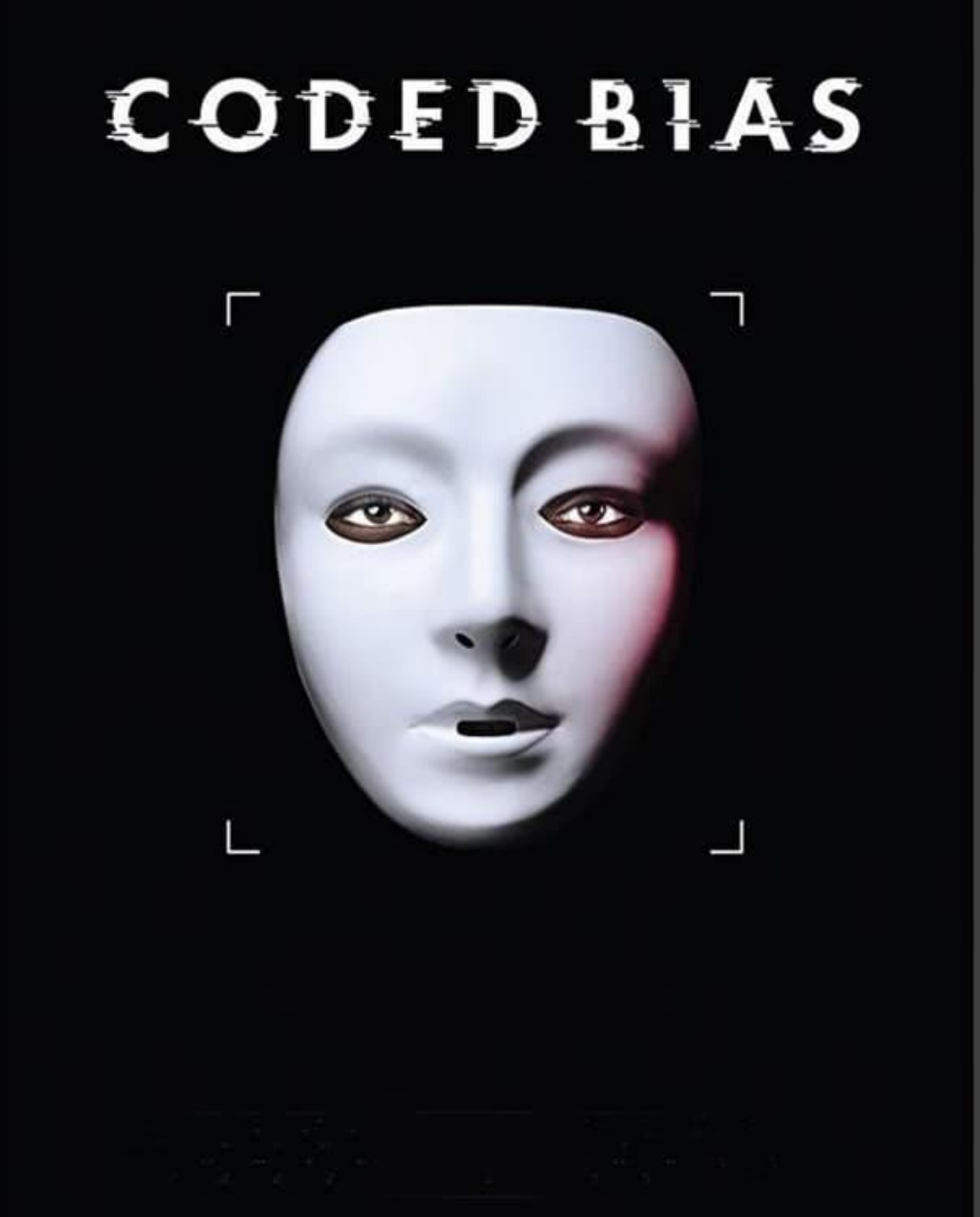


Bias in Artificial Intelligence

Ana Barros, Isabella Fuhrken e Lila Hadba



Introdução

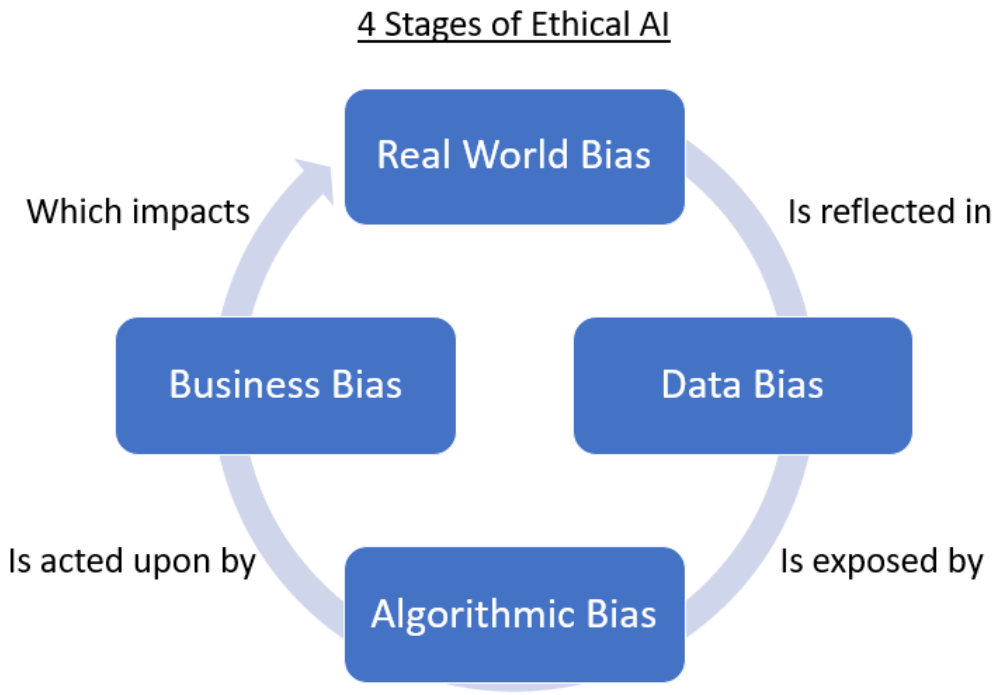
- O que é? Obtenção de resultados discriminantes em algoritmos de inteligência artificial devido à suposições tomadas ou à base de dados enviesada utilizada.
- "Algoritmos usam informações históricas para fazer previsões futuras"
- Definição de IA e seu escopo
- Relações de poder
- Redes sociais

Categorias de AI Bias

Bias em AI pode afetar humanos

- AI pode tomar decisões que podem afetar se ela é aprovada em alguma scola, autorizada a pegar um empréstimo bancário etc.

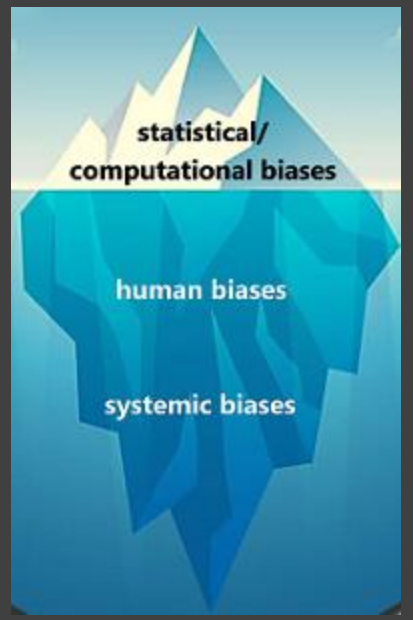
- Os sistemas de Inteligência Artificial podem ter bias que advém da programação e fontes de dados.



Bias em AI pode afetar humanos

- “If we are to develop trustworthy AI systems, we need to consider all the factors that can chip away at the public’s trust in AI. Many of these factors go beyond the technology itself to the impacts of the technology.” —Reva Schwartz, principal investigator for AI bias

- 3 tipos principais de Bias: Estatítico e Computacional,
- Humano e Sistêmico.



1. Estatístico e Computacional

- Ocorre quando uma amostra não é representativa da população
- Relação com documentário Coded Bias:
  - Reconhecimento facial de indivíduos se mostrou falho ao identificar mulheres negras uma vez que o dataset utilizado era composto de homens brancos em sua maioria

- Exemplo: algoritmo para processos seletivos e para demissão em empresas (consideram somente uma quantidade de parâmetros definidas, não permite uma avaliação holística. Quem mais se adequar a "caixinha", está dentro)

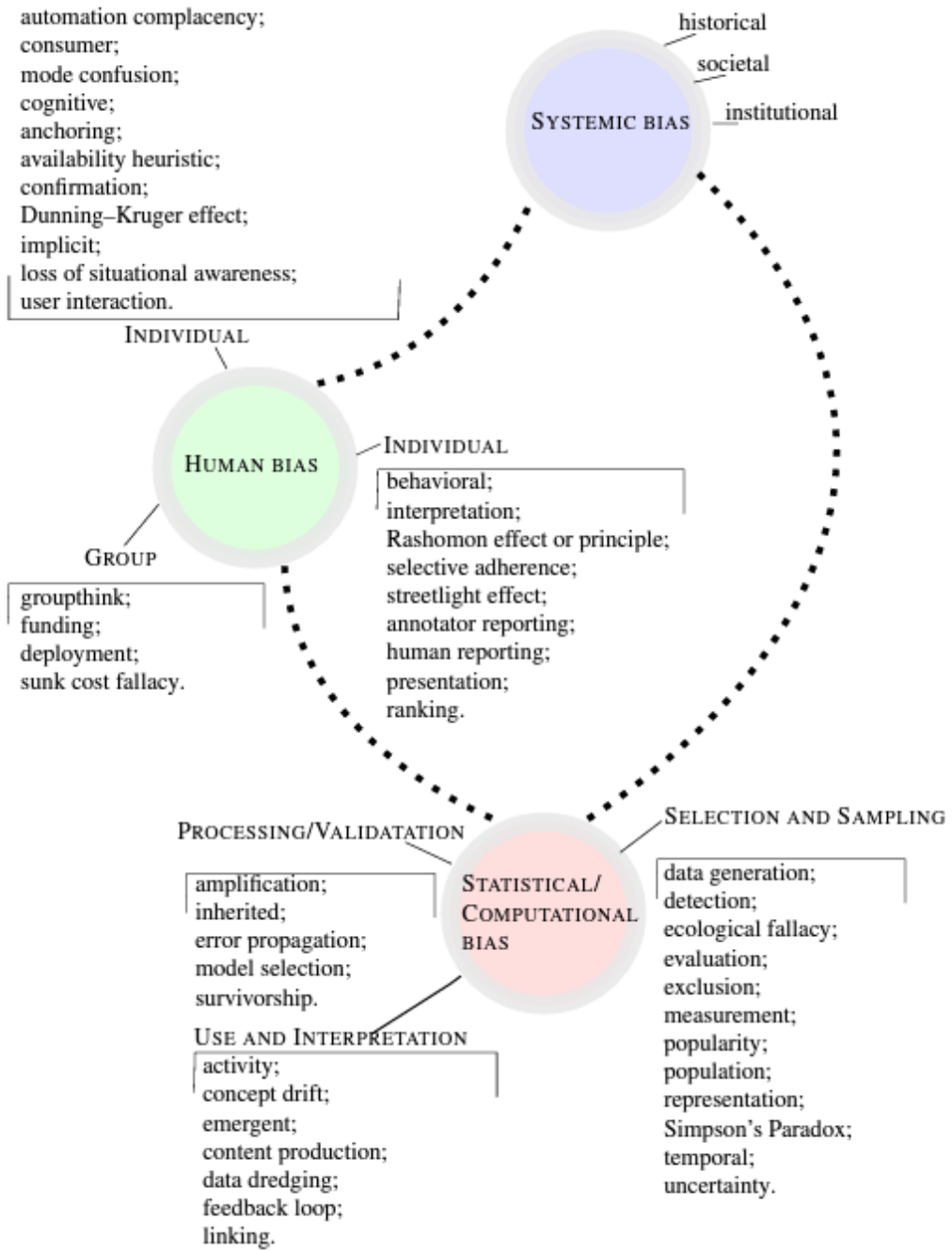
2. Humano

- Refletem erros sistemáticos no pensamento humano com base em um número limitado de princípios heurísticos e predição de valores para operações de julgamento mais simples
- Erros inconscientes que afetam os julgamentos e decisões individuais
- Exemplo: identidade visual no marketing

3. Sistêmico

- Procedimentos e práticas de instituições que resultam na valorização ou desvalorização de determinado grupo social
- Pode ser resultado de preconceitos e/ou do seguimento de normas existentes
- Exemplos: institucional, racismo e sexismo. Ex. Ao identificar o rosto de uma pessoa negra, atribuir a ela uma maior tendência à criminalidade.

4. Resumo das Categorias



Conclusões

- Para retirar todos os AI Bias seria necessário limpar o dataset que será utilizado;
- Se não for possível ter uma mente humana totalmente sem Bias, é impossível ter um sistema inteligente sem Bias;
- O que é possível? Minimizar o AI Bias por meio de testes e desenvolvendo sistemas com princípios responsáveis em mente.

Minimizing bias will be critical if artificial intelligence is to reach its potential and increase people's trust in the systems.

Six potential ways forward for artificial-intelligence (AI) practitioners and business and policy leaders to consider

McKinsey & Company

Referências

https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

Documentário Netflix – Coded Bias (2020)

https://research.aimultiple.com/ai-bias/

https://blogs.gartner.com/anthony_bradley/2020/01/15/4-stages-ethical-ai-algorithmic-bias-not-problem-part-solution/

Q&A

Obrigada!