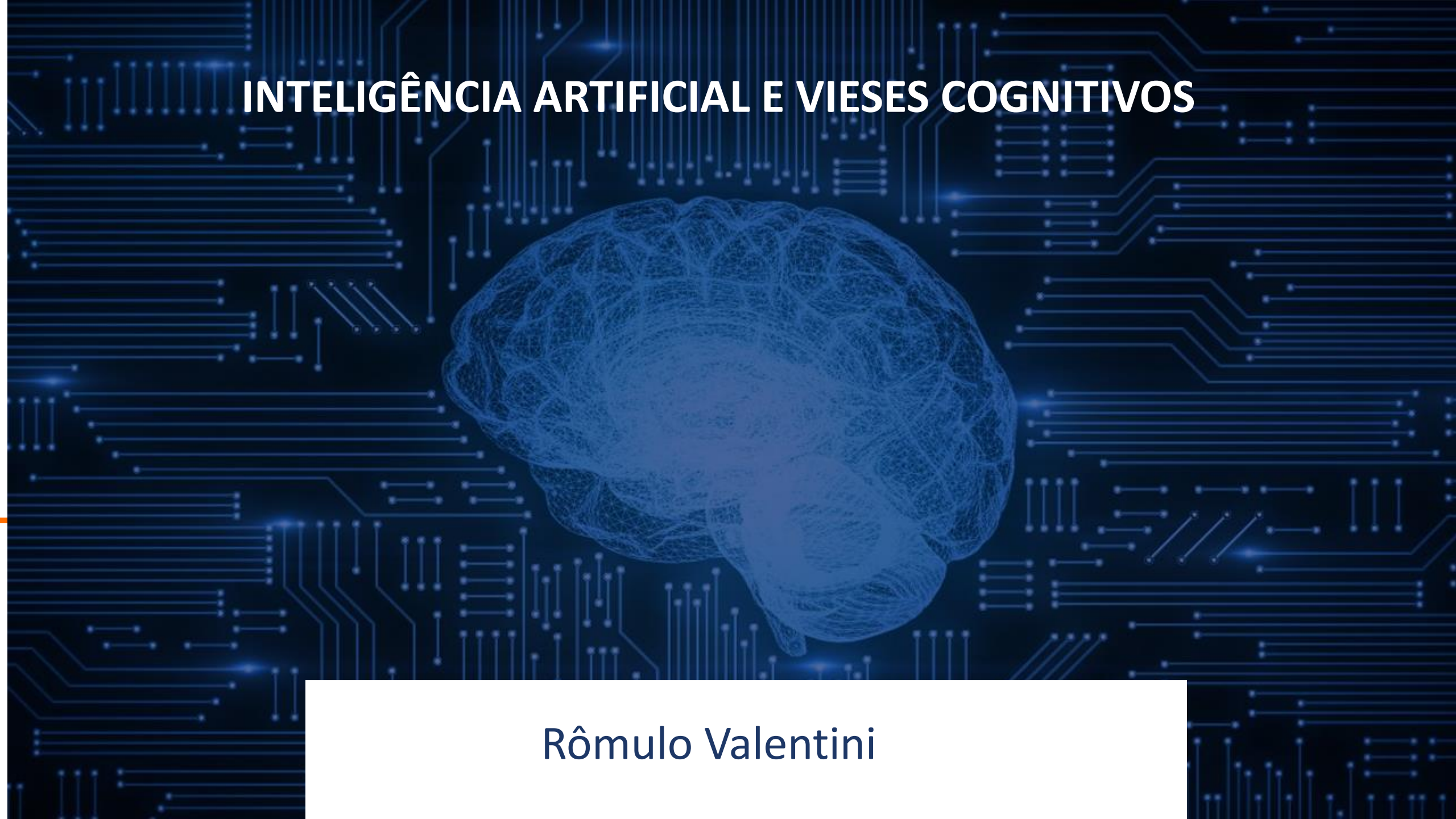
# INTELIGÊNCIA ARTIFICIAL E VIESES COGNITIVOS



Rômulo Valentini

# A Indústria 4.0

*"A revolução que se presencia agora teve início em 2011 quando o governo alemão apresentou na Feira de Hannover uma série de estratégias voltadas à tecnologia capazes de transformar a organização dos sistemas de produção por meio do surgimento de "fábricas inteligentes", capazes de produzir de forma mais eficiente com a utilização "**sistemas ciber-físicos**" para comunicação e integração entre máquinas, pessoas e recursos"*

**SCHWAB, Klaus. The Fourth Industrial Revolution.**
**Genebra: World Economic Forum, 2016.**

# Codificação do trabalho

O chamado **“trabalho intelectual”** nada mais é do que a aplicação da inteligência humana para a realização de tarefas concretas. O “input” (problema) normalmente é variado, mas existe um “output” (resultado) esperado, consistente na entrega do trabalho pelo profissional, **materializada em um serviço ou produto que pode ser mensurado**.

Portanto, é teoricamente possível o desenvolvimento de um processo de **“codificação do trabalho”** por meio do qual engenheiros de software tentam **“algoritmizar” as tarefas exercidas pelos trabalhadores** e, com isso, conseguir obter os **mesmos resultados** (serviços e produtos) **com menor necessidade de trabalho humano qualificado.**

# Um robô pode fazer meu trabalho?

Na lógica do sistema de produção capitalista, uma máquina **não precisa ter uma performance perfeita ou superior à humana** para ser utilizada em larga escala para suprimir postos de trabalho. Basta que apresente um **melhor "custo-benefício" em termos de produção.**

Uma máquina pode, inclusive, gerenciar outros empregados humanos mediante instruções pré-programadas no algoritmo (*subordinação algorítmica*). **Um robô não apenas pode fazer o seu trabalho, mas pode vir a ser o seu chefe!**

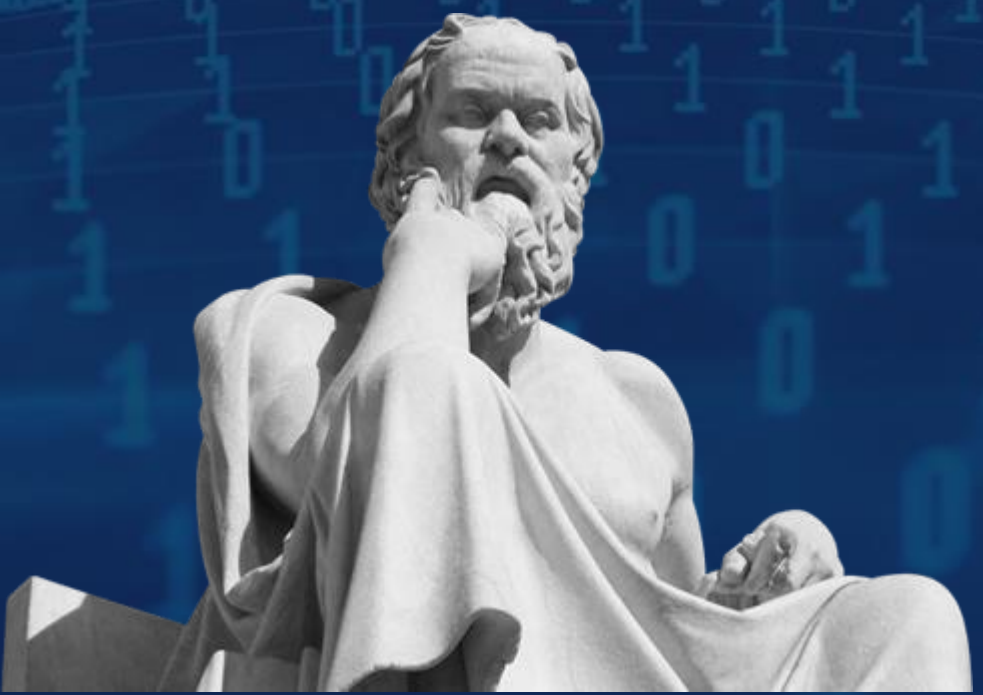
# Vieses

Um **viés** é uma "falha" cognitiva no processo de formação do raciocínio, que faz com que o agente, de maneira inconsciente, adote determinadas tendências e maneiras de pensar e agir.


(PLOUS, Scott. **The psychology of judgment and decision making.** New York : McGraw-Hill,1993, p.233.)

# Vieses dos seres humanos

**Os seres humanos estão sujeitos a uma série de vieses cognitivos, sendo que alguns deles são explorados e potencializados pelas redes:**

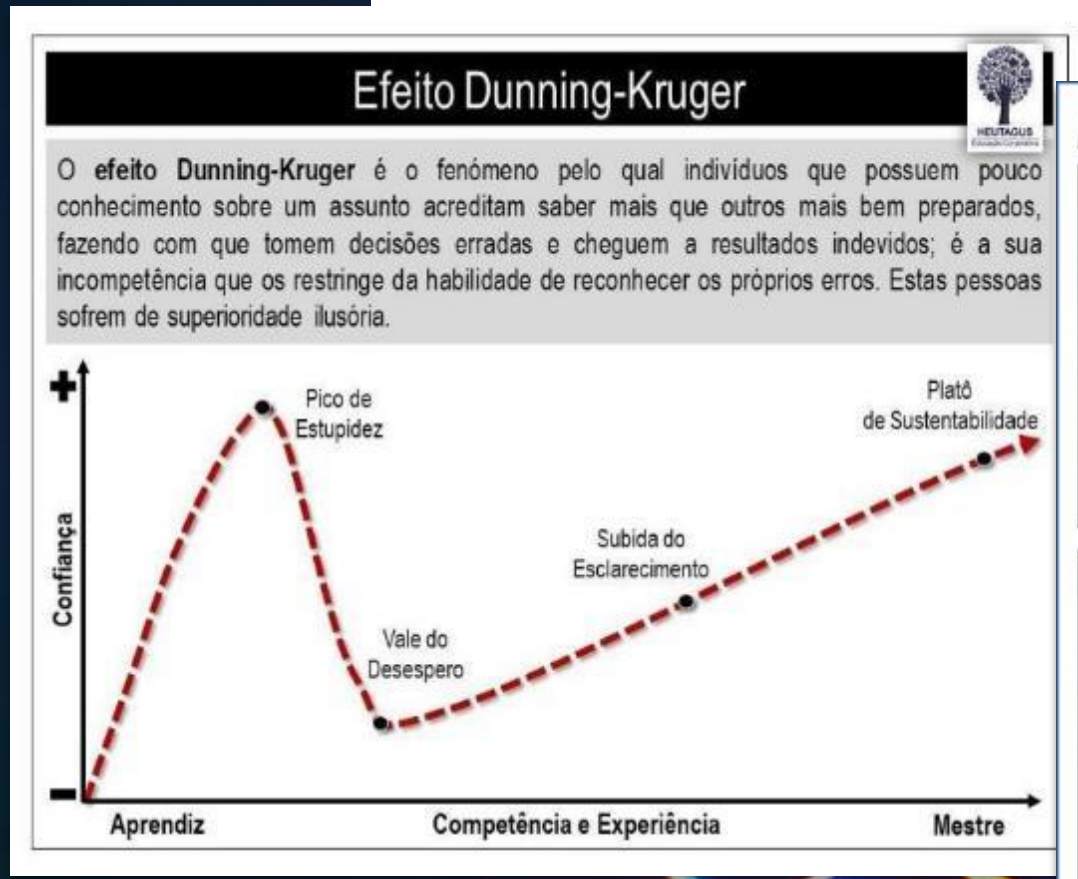


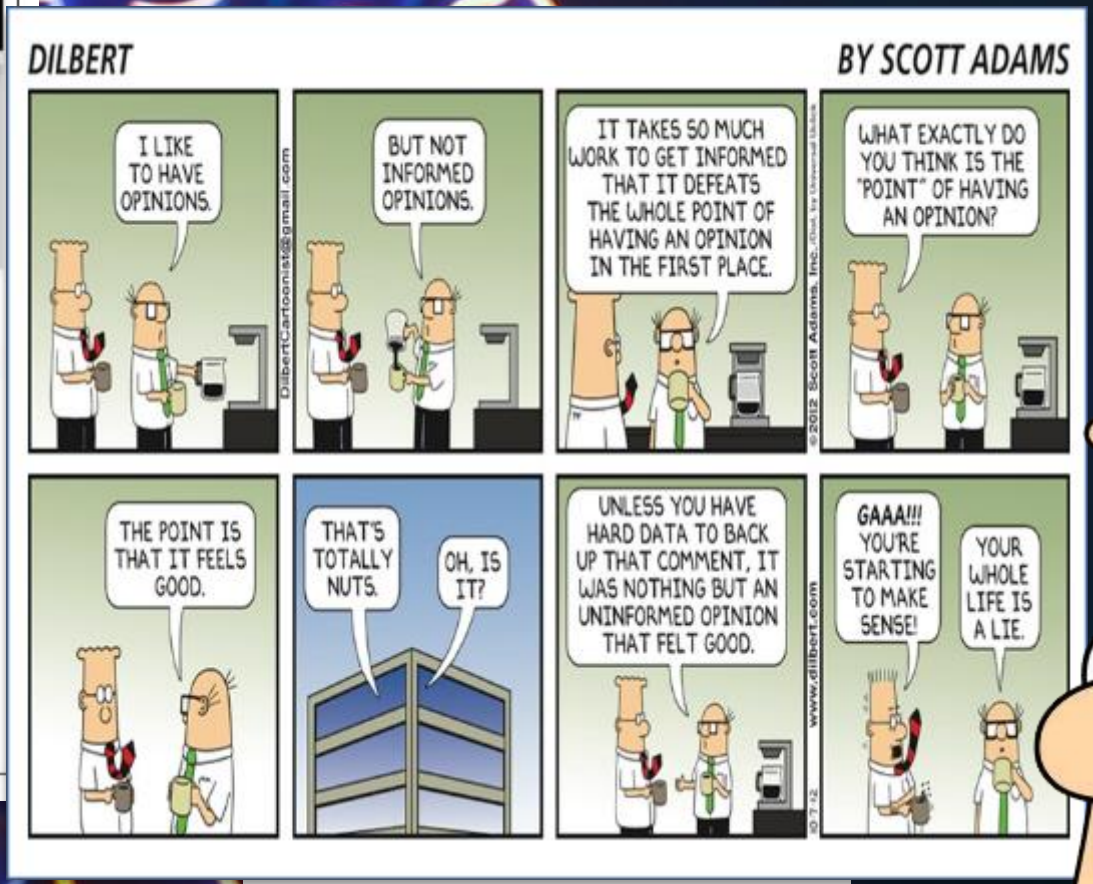
**Viés de confirmação**, a tendência de procurar informações que validem suas crenças ou hipóteses, independentemente de serem ou não verdadeiras, bem como interpretá-las de modo que elas confirmem as preconcepções próprias.

**VIÉS ATENCIONAL** a tendência de prestar atenção a estímulos sensoriais ou emocionais dominantes em sua ambiente e negligenciar dados relevantes ao fazer julgamentos de correlação ou associação.

**Outros vieses comuns e relevantes:** Ancoragem, Cascata de disponibilidade, Efeito adesão, Viés da expectativa, Escalada irracional de compromisso, Heurística de disponibilidade.

# Por que somos enviesados?




https://dilbert.com/strip/2012-10-07

# Vieses da Máquina

As técnicas de inteligência artificial em geral são aplicadas a partir do treinamento de algoritmos especializados que "aprendem" a executar tarefas a partir de análise de um conjunto de dados (treinamento), normalmente coletados e/ou produzidos por seres humanos, motivo pelo qual um sistema informatizado pode se enviesar de diversos modos.

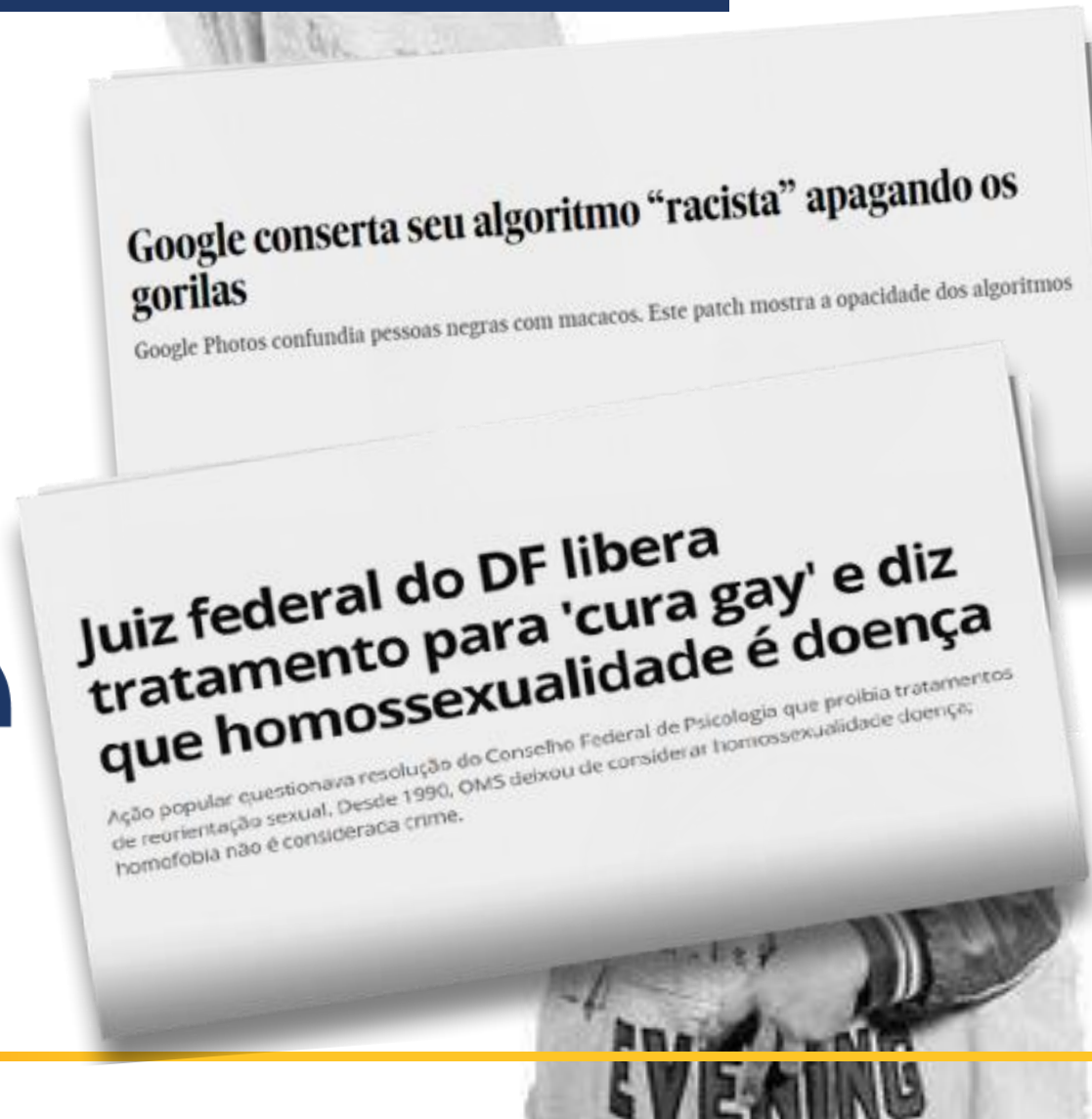
Normalmente os vieses das máquinas surgem nos sistemas em três aspectos: erros na definição do problema (especificação dos requisitos), erros na coleta de dados (input) e erro no tratamento de dados (estruturação).

Mesmo que o viés seja identificado através da saída (output) é muito difícil identificar quando ele foi introduzido e como retirá-lo do modelo (problema da *caixa preta*).

# Vieses humanos X Vieses das máquinas
# Cada computador, uma sentença?

> Algoritmos de aprendizagem automática podem ser muito bons na detecção de relações difíceis de observar entre os dados, pode ser possível detectar associações obscuras entre certas variáveis de casos concretos e os resultados de julgamentos específicos. Seria um resultado impactante se a aprendizagem automática trouxesse evidências sugerindo que os juízes geralmente baseavam suas decisões por fundamentos diferentes das suas justificativas declaradas. A análise dinâmica dos dados pode trazer o debate se determinados julgamentos foram proferidos por fatores diferentes dos que foram expressos na fundamentação da sentença.

**SURDEN, Harry. Machine Learning and Law. Washington Law Review, Vol. 89, No. 1, 2014. Disponível em SSRN: https://ssrn.com/abstract=2417415. Acesso em 02/11/2017. p.108-109. Tradução nossa.**



# Como corrigir o viés das máquinas?

**1) Definir o conceito de "decisão justa".** Na computação o conceito de "justo" é aquele for tido como tal em termos matemáticos. (ex: "Moral Machine do MIT"). Esse conceito deve ser conjugado com os conceitos jurídicos e éticos de justiça.

**2) Eliminar a restrição de contexto social, ou "viés do programador".** A forma como cientistas da computação são ensinados a definir problemas sociais muitas vezes é inadequada, por ausência de preparo do profissional. (ex: usar somente fotos de pessoas brancas para treinar uma IA). Equipes de desenvolvimento multidisciplinares e diversificadas ajudam a mitigar o viés.

**3) Evitar e descartar processos de aprendizagem imperfeitos.** Muitas vezes os dados utilizados para testar a performance do modelo tem o mesmo viés que os dados utilizados para treinar o modelo: (ex: sistema identifica o que "não é negro" a partir da definição de "negro"). É preciso observar diretrizes de desenviesamento em todas as etapas do processo.

# Obrigado!

Rômulo Valentini

*rsvalentini@gmail.com*