PALESTRA

Deep fake e #trakinagens para fugir das IA(s)

A Conferência Security BSides São Paulo (BSidesSP) é uma mini-conferência gratuita sobre segurança da informação e cultura hacker, realizada em São Paulo, Brasil, nos **dias 25 e 26 de Maio de 2019**

PEDRO BEZERRA

COGNITIVE SECURITY | INFORMATION SECURITY | GRC |+|
RESEARCHER | LECTURE | COMMUNITY MANAGER OF AI AND SECURITY

Comunidades que apoio e sou apoiado ;-D

AI Brasil

an artificial intelligence community

+de 2700 membros esperando por você em
https://www.meetup.com/pt-BR/ai-brasil/

security
H1V3

+de 290 membros esperando por você em
https://www.meetup.com/pt-BR/Security-Hive/

BSIDES SÃO PAULO

Comunidades

AI Brasil
an artificial intelligence community

Meetup **https://www.meetup.com/pt-BR/Security-Hive/**

Procurar por Security H1V3

https://web.facebook.com/H1V3Sec

Grupos em Apps: Entre na nossa página do Meetup.com e fale com Pedro Bezerra.

https://web.telegram.org/#/im?p=@H1V3SecResurrection

https://www.meetup.com/pt-BR/ai-brasil Meetup

https://www.youtube.com/c/AIBrasilCommunity

https://web.facebook.com/BrasilAI

Grupos em Apps: Entre na nossa página do Meetup.com e fale com Pedro Bezerra.

https://web.telegram.org/#/im?p=g310549344

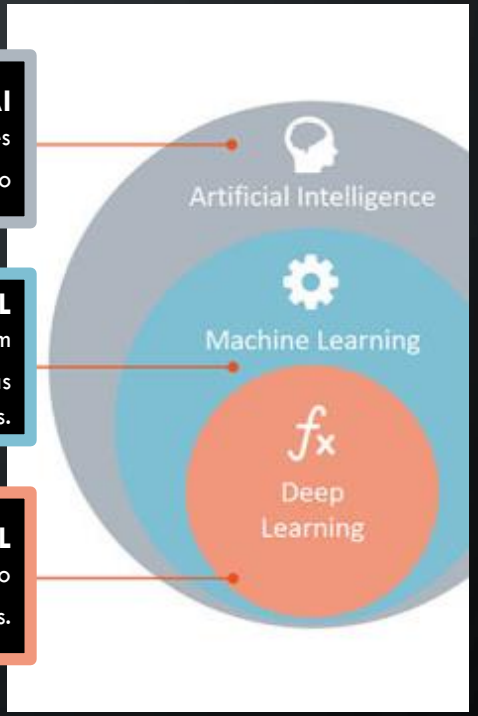
Vamos começar !!!

AI, Machine Learning, Deep Learning e NLP

Artificial Intelligence - AI
Qualquer técnica que permita aos computadores **imitar** o comportamento humano

Machine Learning - ML
Subconjunto de técnicas de AI que utilizam **métodos estatísticos** para permitir que as máquinas melhorem com experiências.

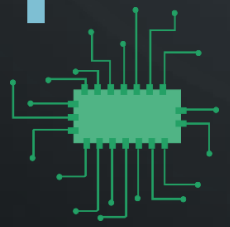
Deep Learning - DL
Subconjunto de ML que possibilita a computação de **redes neurais** de várias camadas.



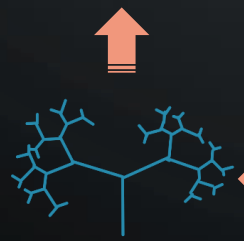
Fontes:
https://content-static.upwork.com/blog/uploads/sites/3/2017/06/27091427/image-43.png
http://biogeocarlos.blogspot.com.br/2009/04/arte-zoologia-iv-camaleon.html

AI, Machine Learning, Deep Learning e NLP

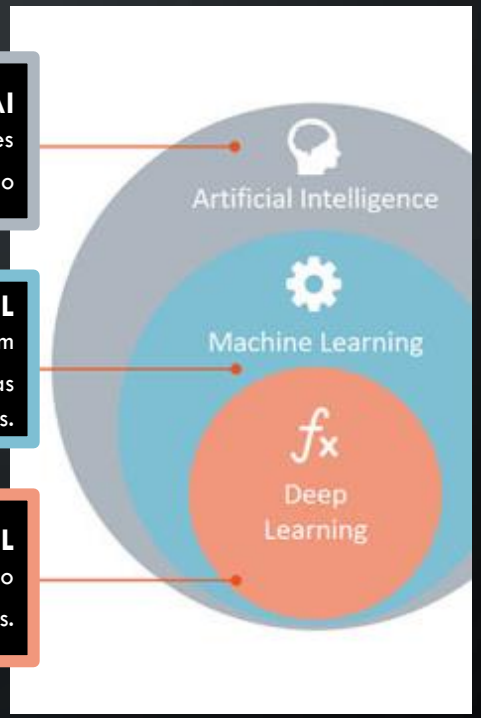
Artificial Intelligence - AI
Qualquer técnica que permita aos computadores **imitar** o comportamento humano



Machine Learning - ML
Subconjunto de técnicas de AI que utilizam **métodos estatísticos** para permitir que as máquinas melhorem com experiências.

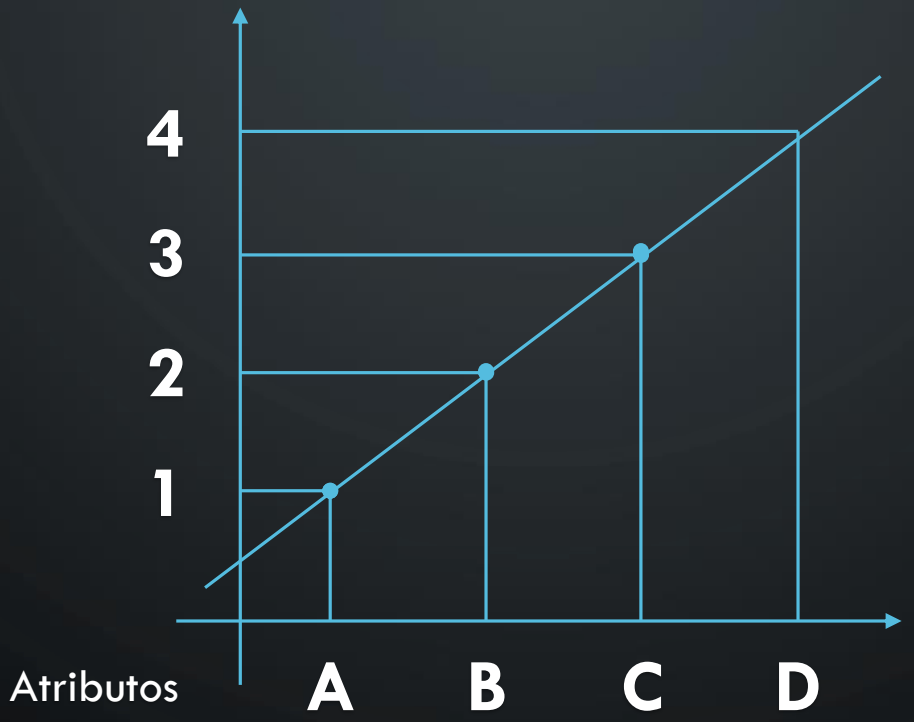


Deep Learning - DL
Subconjunto de ML que possibilita a computação de **redes neurais** de várias camadas.

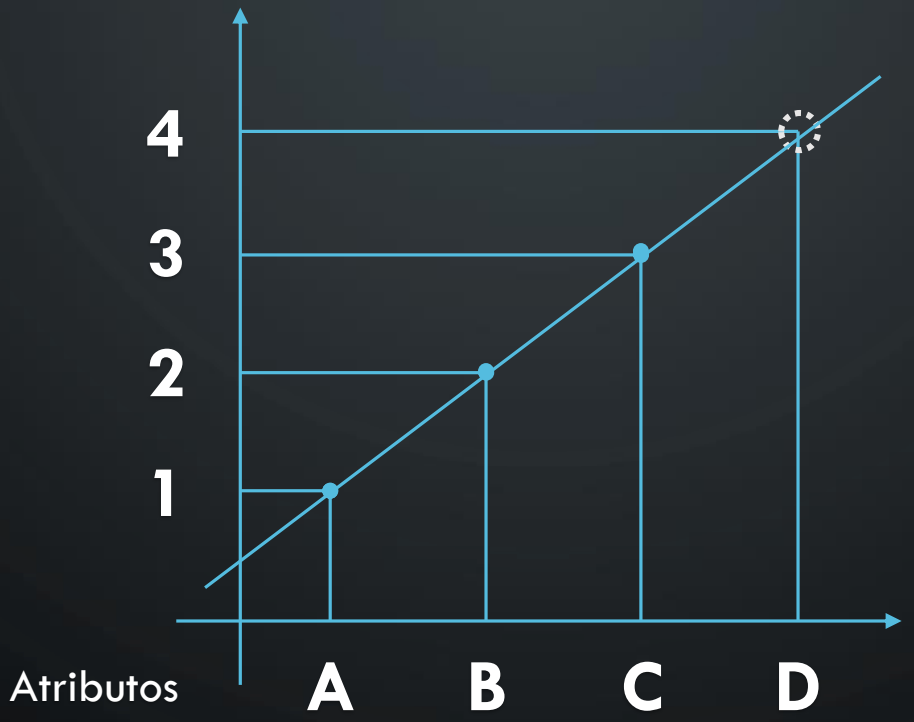


Fontes:
https://content-static.upwork.com/blog/uploads/sites/3/2017/06/27091427/image-43.png
http://biogeocarlos.blogspot.com.br/2009/04/arte-zoologia-iv-camaleon.html

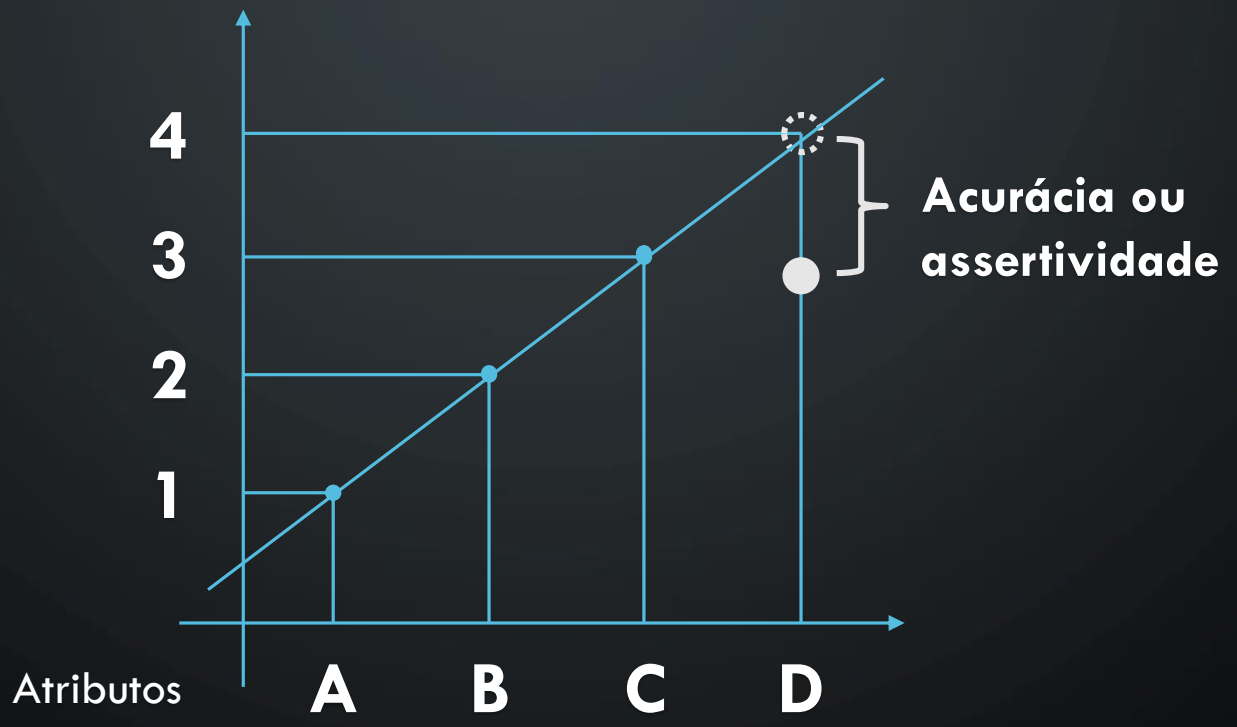
Tentando simplificar



Tentando simplificar



Tentando simplificar



Tentando simplificar

quem irá fiscalizar a minha empresa ?

LGPDCap10Art61	0.02	LGPDCap1Art5	0.07	LGPDCap5Art34	0.01	LGPDCap6sec1Art38	0.01
LGPDCap6sec2Art41	0.05	LGPDCap7sec2Art51	0.02	LGPDCap8sec1Art52	0.08		
LGPDCap9sec1Art55	0.01	LGPDCap9sec1Art56	0.53	LGPDCap9sec1Art57	0.02		

qual artigo me fala sobre a coleta do consentimento?

LGPDCap1Art3	0.01	LGPDCap2Art8	0.55	LGPDCap3sec1Art14	0.02	LGPDCap3sec2Art19	0.04
LGPDCap3sec3Art20	0.20	LGPDCap6sec3Art44	0.03	LGPDCap7sec1Art46	0.02		
LGPDCap8sec1Art52	0.02	LGPDCap9sec1Art56	0.01	LGPDCap9sec2Art59	0.01		

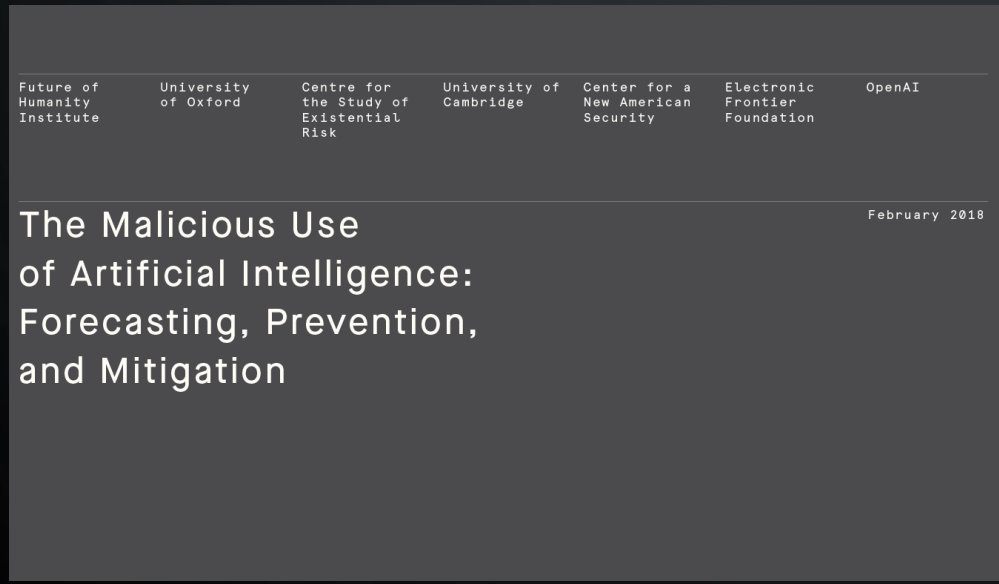
Nome da Classe / Categoria / Classificação / etc. Acurácia ou assertividade


StratoEnergetics LIVE STREAM
http://www.stratoenergetics.com
Buenos Aires Event
TV Truck 02

Fonte (YouTube):U.S. Military's Super Weapon: Autonomous Drone Swarm

Como podemos minimizar os impactos negativos:

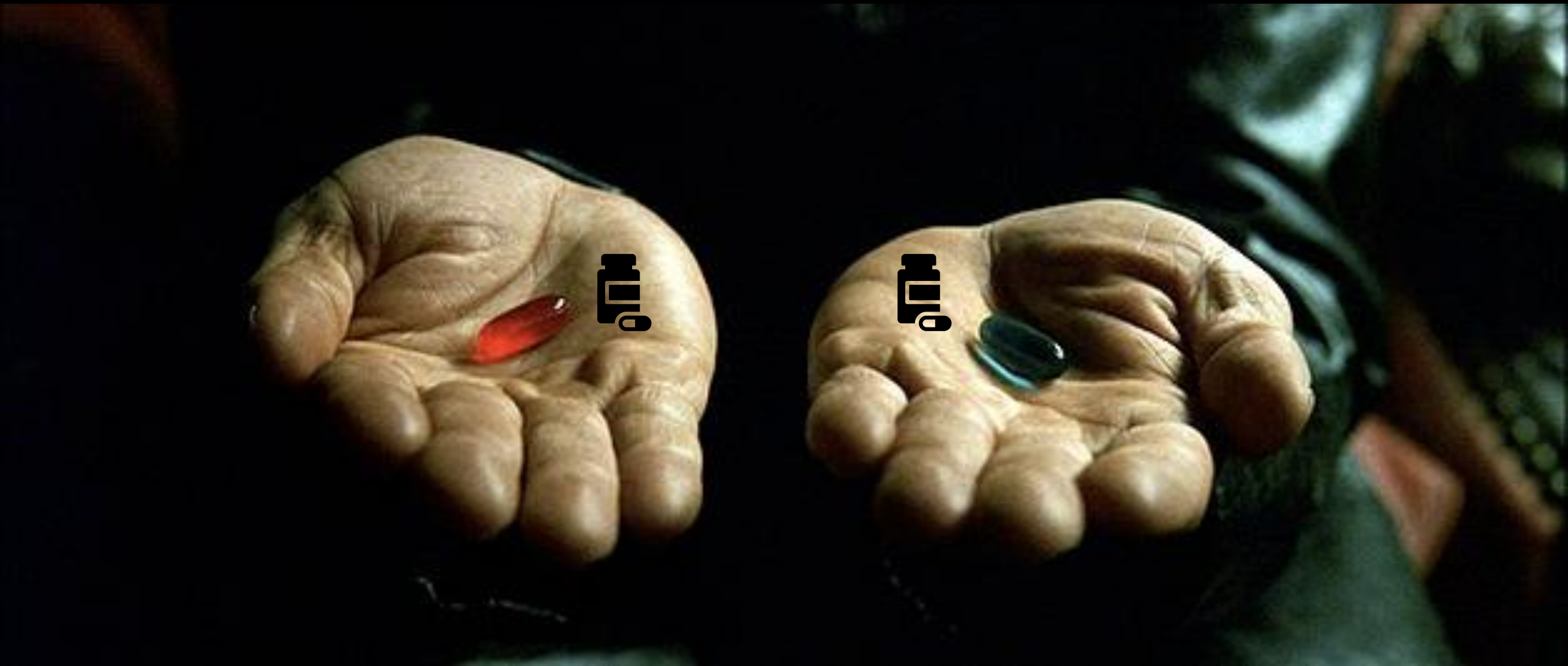
1. Colaboração para investigar, prevenir e mitigar possíveis usos maliciosos da IA.
2. Pesquisadores e engenheiros em inteligência artificial devem levar a sério a natureza de dupla utilização de seu trabalho.
3. As melhores práticas devem ser identificadas.
4. Procurar ativamente expandir a discussão desses desafios.



Scenarios - Digital Security

- Automation of social engineering attacks
- Automation of vulnerability discovery
- More sophisticated automation of hacking
- Human-like denial-of-service
- Automation of service tasks in criminal cyber-offense
- Prioritising targets for cyber attacks using machine learning
- Exploiting AI used in applications, especially in information security
- Black-box model extraction of proprietary AI system capabilities

Nós é quem escolhemos !!



Somos enganados por IA

Deep Fake

Vídeo

Fonte: BBC NEWS BRASIL: https://www.youtube.com/watch?v=OrGLT6-pKqs

Somos enganados por IA

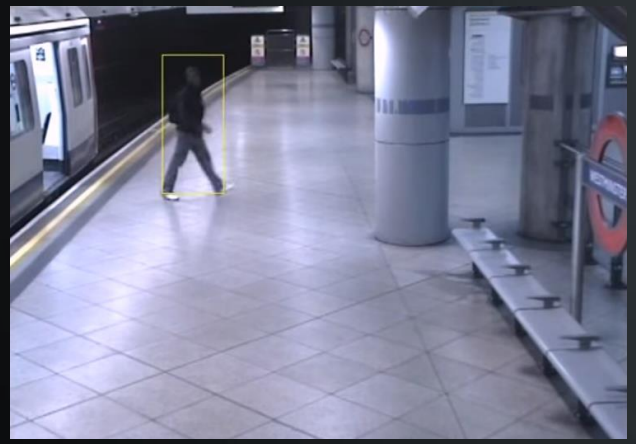
Deep Fake

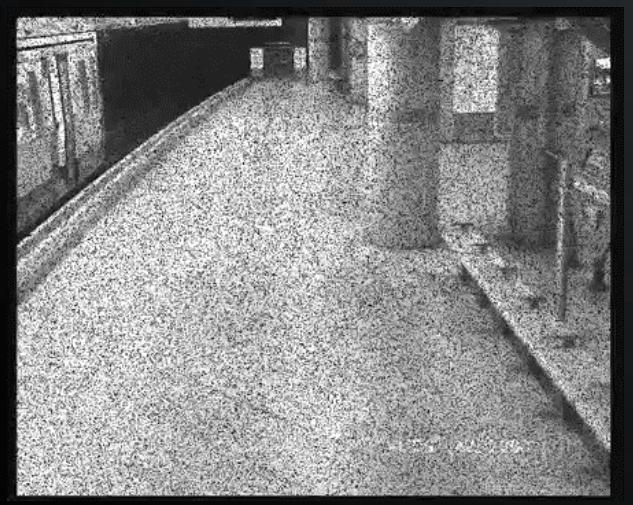
Texto

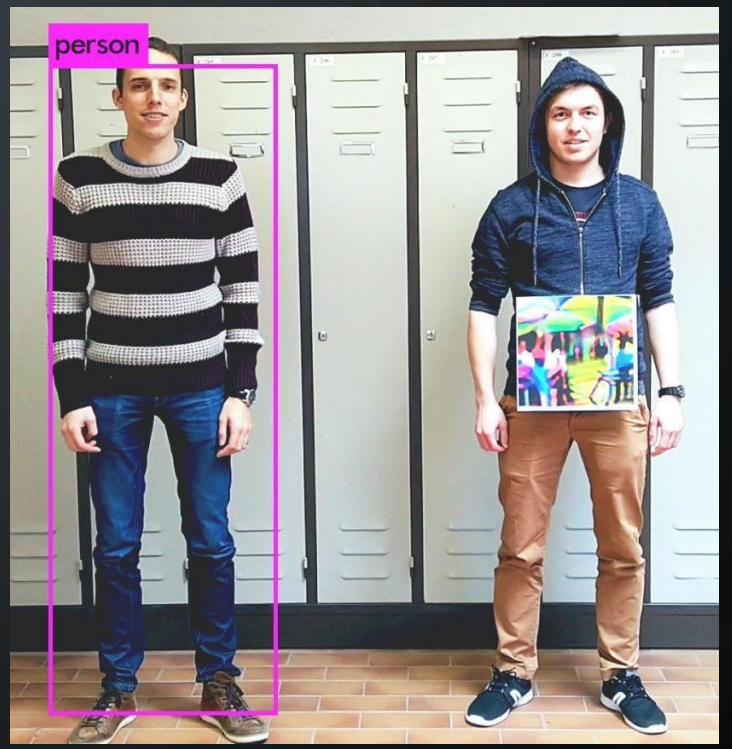
Fonte: THE GUARDIAN: https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

Podemos enganar a IA ! #Trakinagem

Podemos enganar a IA ! #Trakinagem

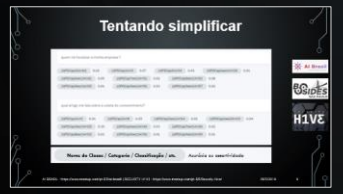






Fonte: ORIGEM: Link

Podemos enganar a IA ! #Trakinagem



NEXTCon
Online AI Tech Talk Series
Friday, Jan 18

Adversarial Attacks on A.I. Systems

Anant Jain
Co-founder, commonlounge.com (Compose Labs)

https://commonlounge.com
https://index.anantja.in

Deep fake e #trakinagens para fugir das IA(s)

Pedro Bezerra

Drone para iniciantes com AI

Google Tensorflow e Mobilenet

IMAGENET

TensorFlow

https://github.com/tensorflow/tensorflow

https://machinethink.net/blog/mobilenet-v2/

Conclusões

Adversarial Patch

Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer
https://arxiv.org/abs/1712.09665

1. Os patches são universais porque **podem ser usados para atacar qualquer cena**, robustos porque funcionam sob uma ampla variedade de transformações e direcionados porque podem fazer com que um classificador produza qualquer classe-alvo.
2. Os sistemas de aprendizagem profunda são amplamente vulneráveis a exemplos contraditórios, **inputs cuidadosamente escolhidos** que fazer com que a rede mude a saída sem uma mudança visível para um ser humano [15, 5].
3. Porque esse patch é independente de cena, permite que **atacantes criem um ataque no mundo físico sem conhecimento prévio de as condições de iluminação, ângulo da câmera, tipo de classificador sendo atacado**, ou até mesmo os outros itens dentro a cena.

Conclusões

Adversarial Patch

Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer
https://arxiv.org/abs/1712.09665

4. Este ataque é significativo porque o atacante não precisa saber que imagem ele está atacando ao construir o ataque. Depois de gerar um patch adversário, o patch **poderia ser amplamente distribuídos pela Internet para outros invasores imprimirem e usarem**.
5. As técnicas de **defesa existentes que se concentram na defesa contra pequenas perturbações** podem não ser robustas a perturbações maiores como estas. Na verdade, o trabalho recente demonstrou que os modelos treinados por adversários de **última geração sobre o MNIST ainda são vulneráveis** a perturbações
6. Muitos modelos ML **operam sem validação humana** de cada entrada e, assim, atacantes mal-intencionados não se preocupam com a imperceptibilidade de seus ataques.
7. Mesmo que os seres humanos sejam capazes de perceber patches, eles podem não entender a intenção do patch e **vê-lo como uma forma de arte**.

PALESTRA

Deep fake e #trakinagens para fugir das IA(s)

#Amo Perguntas !!!

Básicas, avançadas, não entendi nada, todas são muito bem vindas :-D

A Conferência Security BSides São Paulo (BSidesSP) é uma mini-conferência gratuita sobre segurança da informação e cultura hacker, realizada em São Paulo, Brasil, nos **dias 25 e 26 de Maio de 2019**

PEDRO BEZERRA

COGNITIVE SECURITY | INFORMATION SECURITY | GRC |+|
RESEARCHER | LECTURE | COMMUNITY MANAGER OF AI AND SECURITY

Comunidades que apoio e sou apoiado ;-D

AI Brasil

an artificial intelligence community

+de 2700 membros esperando por você em
https://www.meetup.com/pt-BR/ai-brasil/

security HIVE

+de 290 membros esperando por você em
https://www.meetup.com/pt-BR/Security-Hive/