Умные живут дольше

Технологии разума в современном кибероружии

Андрей Масалович

Andrei Masalovich

avalanche100500@gmail.com

УМНЫЕ ЖИВУТ ДОЛЬШЕ.

Технологии разума в современном кибероружии

- Искусственный интеллект - это не только умные дома и умные города, но и технически совершенное автономное кибероружие. Новую войну будут вести армии умных ботов, способных не только к групповой координации без участия человека, но и к самостоятельной выдаче целеуказаний.
- В докладе рассматриваются решения из сферы ИИ, используемые в современном кибероружии: генеративно-состязательные нейронные сети (GAN) для распознавания новых видов кибератак, методики глубокого обучения с подкреплением (DRL) для агентного моделирования информационных атак, методы «цифровых двойников» для исследования различных физических и психологических воздействий без проведения тестовых атак.

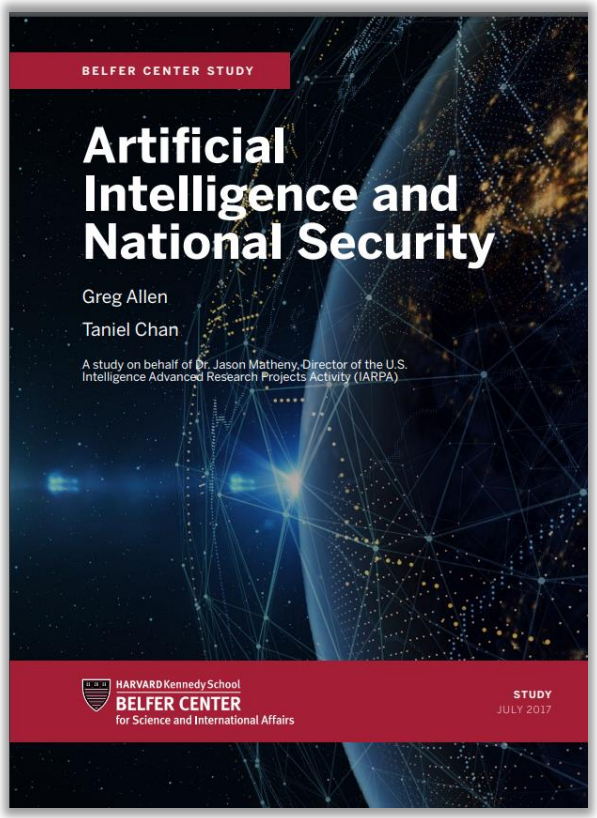
Основные тренды AI

Top 12 AI Tech Trends

1. Deep Learning - теория глубокого обучения
2. Capsule Neural Networks – капсульные сети
3. Deep reinforcement learning (DRL) – глубокое обучение с подкреплением
4. Generative adversarial network (GAN) – генеративно-состязательные сети
5. Lean and augmented data – обучение на неполных и дополненных данных
6. Probabilistic programming – вероятностное программирование
7. Hybrid learning models – модели гибридного обучения
8. Automated machine learning (AutoML) – автоматическое машинное обучение
9. Digital twin – Цифровой двойник
10. Explainable AI – Объяснимый искусственный интеллект
11. AI Chatbots – разумные чат-боты
12. AI Accelerators – апппаратные акселераторы ИИ

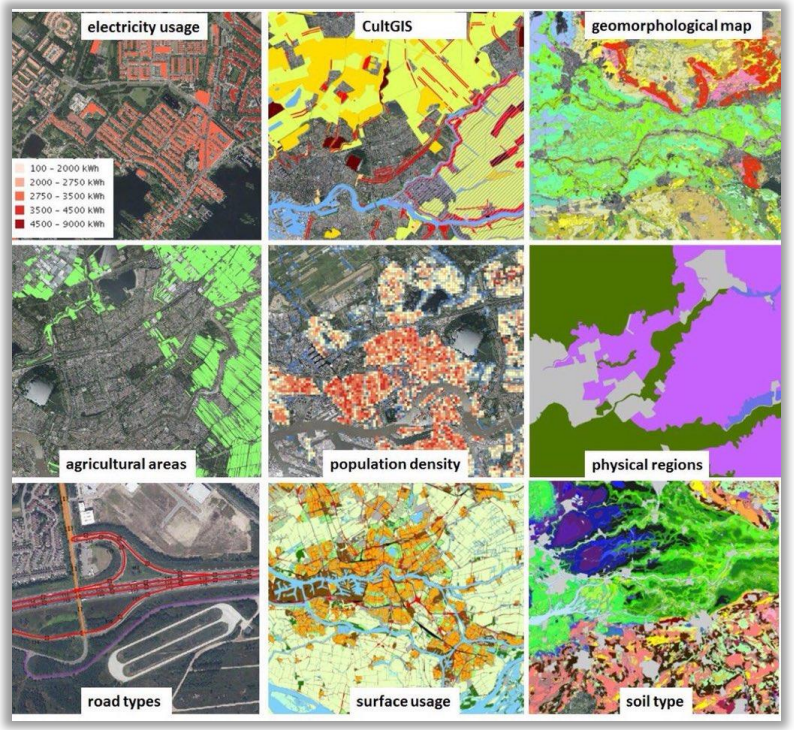
Deep Learning - теория глубокого обучения

Deep Learning - совокупность методов машинного обучения, основанных на обучении представлениям (*feature/representation learning*), а не на специализированных алгоритмах, разработанных для конкретных задач



Capsule Neural Networks – капсульные сети

- Capsule Neural Networks - новый тип глубоких нейронных сетей, могут поддерживать иерархические отношения



Deep reinforcement learning (DRL) – глубокое обучение с подкреплением

- DRL – сеть учится, взаимодействуя с окружающей средой посредством наблюдений, действий и вознаграждений

Передний край: DRL + Agent-Basing Dynamics

Generative adversarial network (GAN) – генеративно-состязательные сети

- GAN - две конкурирующие нейронные сети, генератор и дискриминатор

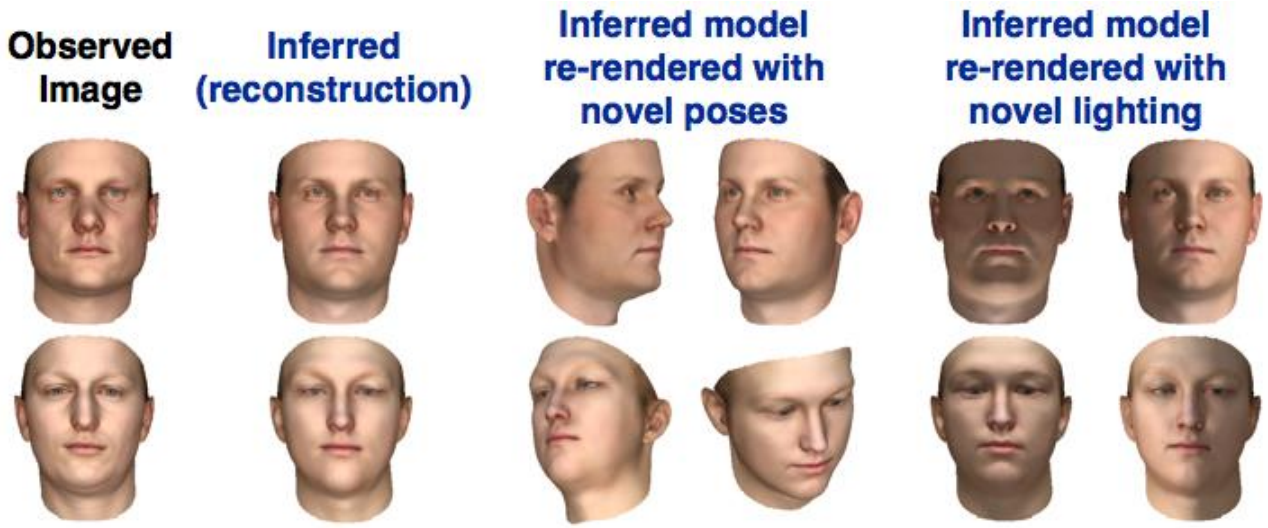
Deep Dream

Lean and augmented data – обучение на неполных и дополненных данных

- Перенос обучения
- Экстремальное обучение
- Синтез данных

Probabilistic programming – вероятностное программирование

- Probabilistic programming -высокоуровневый язык программирования, который облегчает разработку вероятностной модели, а затем автоматически «решает» эту модель



Hybrid learning models – модели гибридного обучения

- Hybrid learning models – глубокие нейронные сети + байесовские или вероятностные подходы
- “Blended Learning”

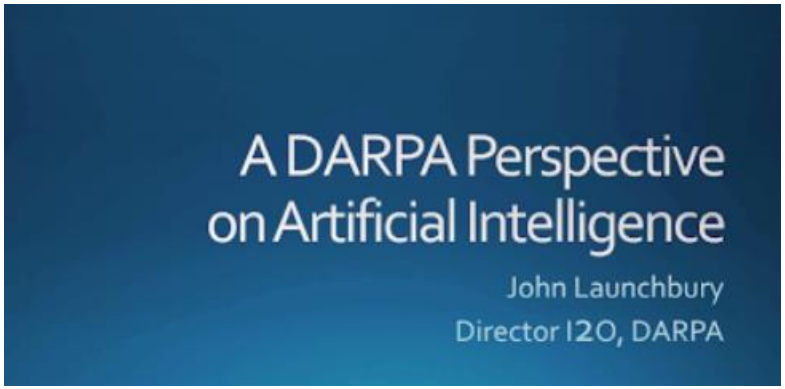
Automated machine learning (AutoML) – автоматическое машинное обучение

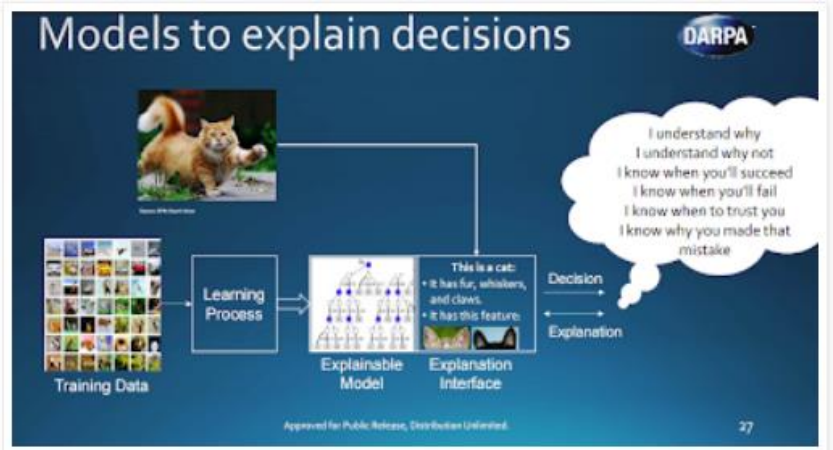
- Automated machine learning (AutoML) - Автоматизация процесса подготовки данных, выбора функций, выбора модели или техники, обучения и настройки

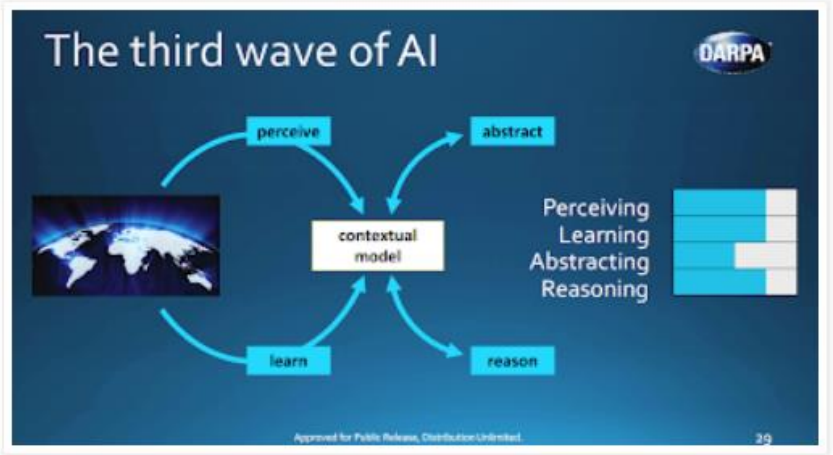
Digital twin – Цифровой двойник

- Digital twin— это виртуальная модель, используемая для облегчения детального анализа и мониторинга физических или психологических систем

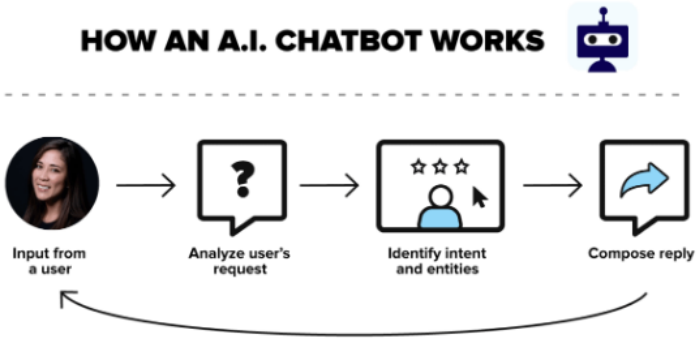
Explainable AI – Объяснимый искусственный интеллект

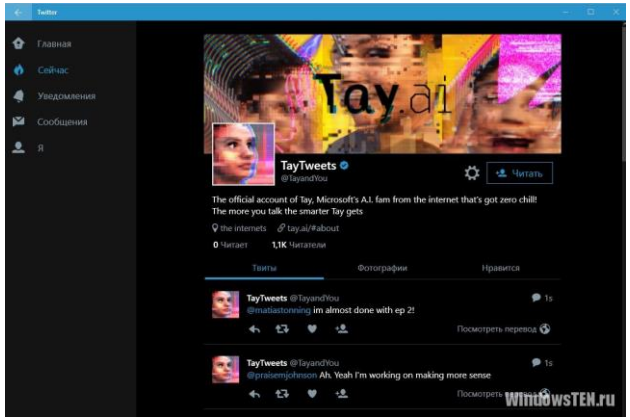






AI Chatbots – разумные чат-боты



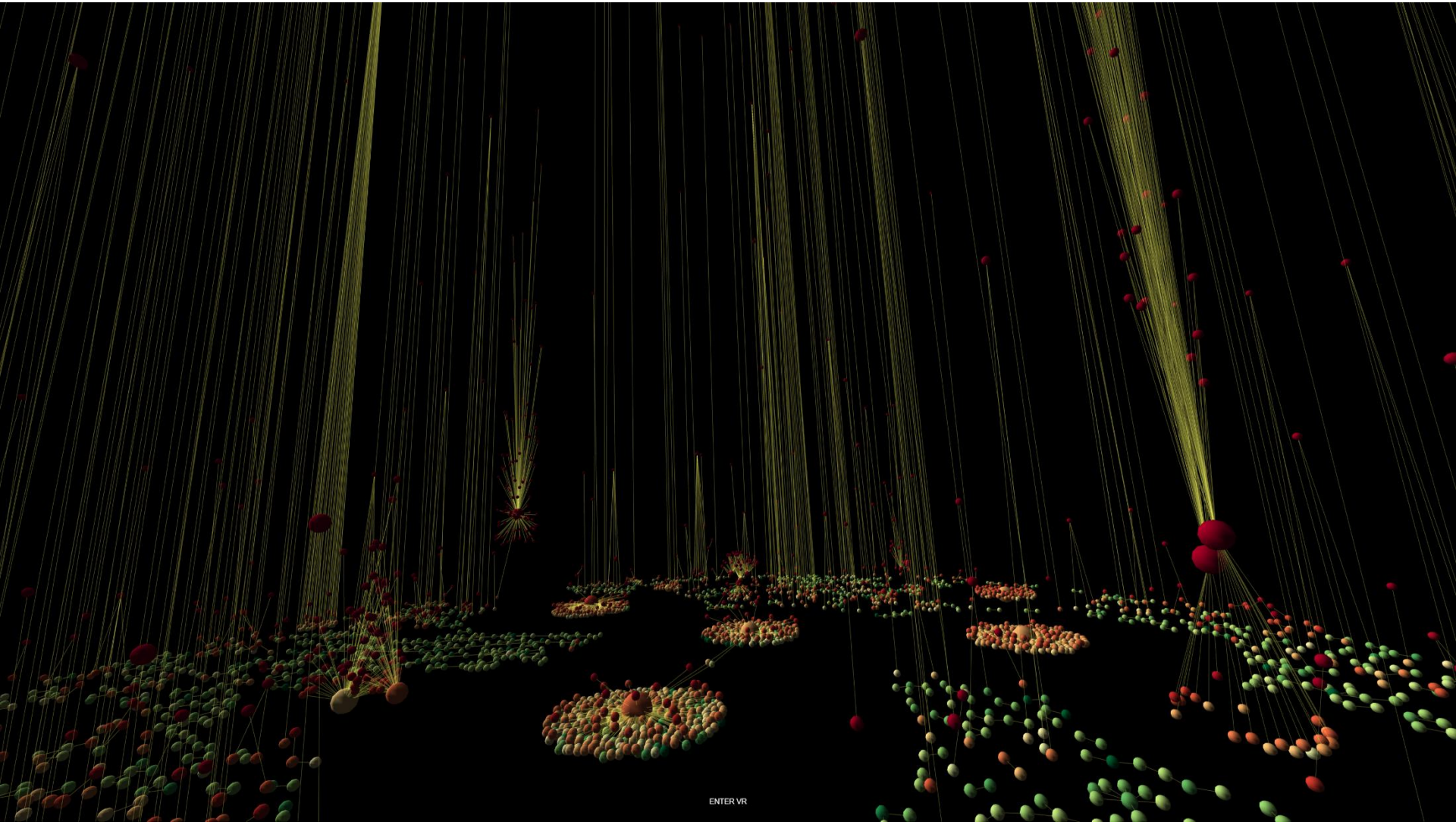


- 23 марта 2016 года – Microsoft запускает в Twitter AI-чатбота **Тай** (@TayandYou)
- 24 марта – Тай обучился нетерпимости, расизму и обсценной лексике
- 25 марта – бота отключили

AI accelerators – аппаратные ускорители ИИ

AVALANCHE

One minute inside Twitter

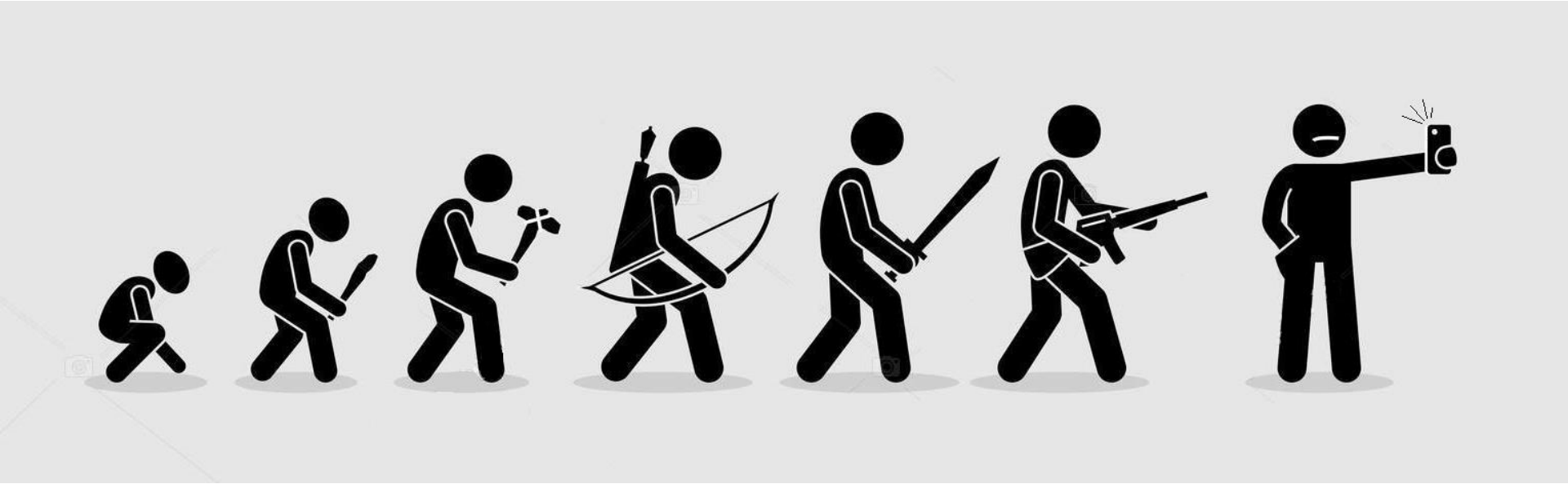


AVALANCHE

Дополнительная информация

Мы – дети в мире умных вещей

Военные – дети с гранатой



- Высокоточное оружие
- Умное оружие
- Автономное летальное оружие
- Сетецентрическая война

Тест Тьюринга

- Ученый: Искусственный разум – тот, который при общении неотличим от живого человека
- Хакер: Я боюсь не того компьютера, который пройдет тест Тьюринга, а того, который его намеренно завалит...
- Политтехнолог: Задача искусственного разума – убедить экзаменатора, что он сам компьютер

The Malicious Use of AI

Security Domains

The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation
February 2018

- **Digital security.** The use of AI to automate tasks involved in carrying out cyberattacks will alleviate the existing tradeoff between the scale and efficacy of attacks. This may expand the threat associated with labor-intensive cyberattacks (such as spear phishing). We also expect novel attacks that exploit human vulnerabilities (e.g. through the use of speech synthesis for impersonation), existing software vulnerabilities (e.g. through automated hacking), or the vulnerabilities of AI systems (e.g. through adversarial examples and data poisoning).
- **Physical security.** The use of AI to automate tasks involved in carrying out attacks with drones and other physical systems (e.g. through the deployment of autonomous weapons systems) may expand the threats associated with these attacks. We also expect novel attacks that subvert cyber-physical systems (e.g. causing autonomous vehicles to crash) or involve physical systems that it would be infeasible to direct remotely (e.g. a swarm of thousands of micro-drones).
- **Political security.** The use of AI to automate tasks involved in surveillance (e.g. analysing mass-collected data), persuasion (e.g. creating targeted propaganda), and deception (e.g. manipulating videos) may expand threats associated with privacy invasion and social manipulation. We also expect novel attacks that take advantage of an improved capacity to analyse human behaviors, moods, and beliefs on the basis of available data. These concerns are most significant in the context of authoritarian states, but may also undermine the ability of democracies to sustain truthful public debates.

Пример AI: Нейросеть распознает человека по клавиатурному почерку

Аутентификаторы:

- Уникальное знание
- Уникальный предмет
- Уникальная характеристика

Клавиатурный почерк - поведенческая биометрическая характеристика, которую описывают следующие параметры:

- **Скорость ввода** - количество введенных символов разделенное на время печатания
- **Динамика ввода** - характеризуется временем между нажатиями клавиш и временем их удержания
- **Частота возникновение ошибок** при вводе
- **Использование клавиш** - например, какие функциональные клавиши нажимаются для ввода заглавных букв

Корнеев В.В., Масалович А.И и др. Распознавание программных модулей и обнаружение несанкционированных действий с применением аппарата нейросетей Информационные технологии N10, 1997 - http://sci-pub.info/ref/321545/

Physical Security

Фитнес трекер Strava
выдал расположение военных баз США

Умный пистолет Armatix IP1 – стреляет только в руках владельца

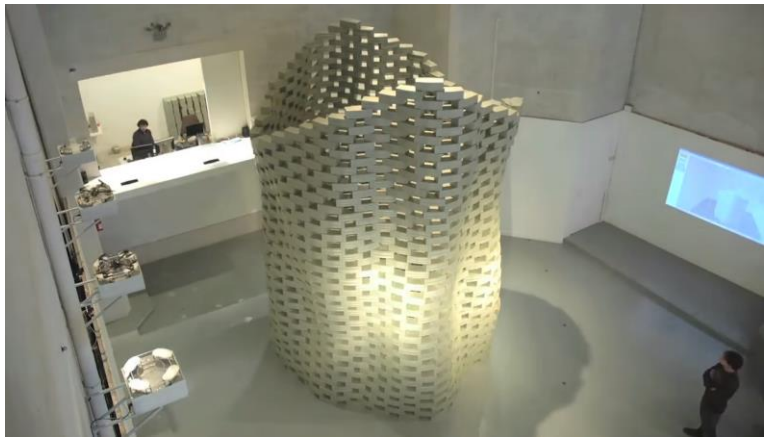
Хакер Plore:

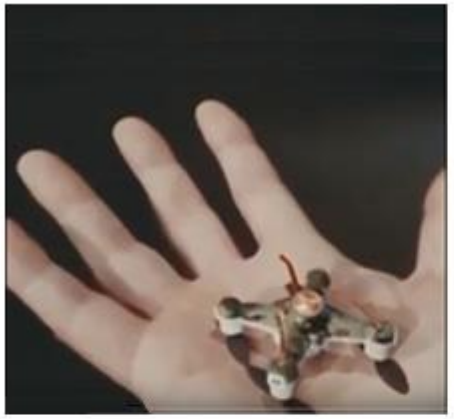
Я знаю как минимум три способа взломать Armatix

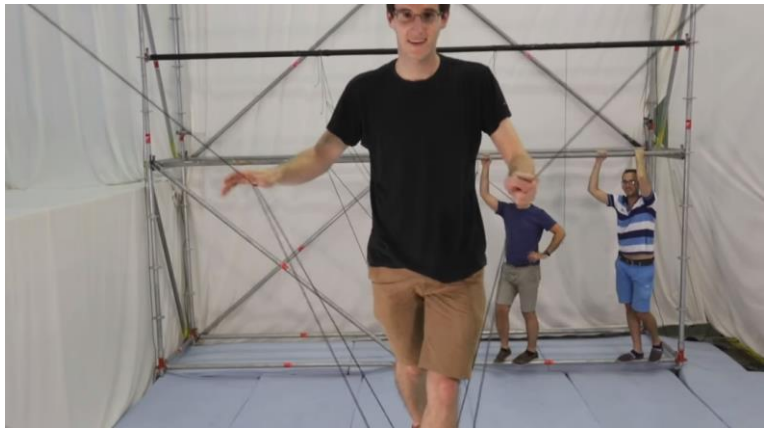
TrackingPoint XS1 —винтовка под Linux

Meet the dazzling flying machines of the future…

Дроны убивают людей



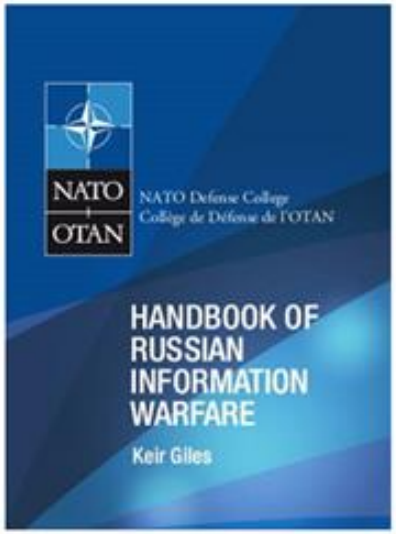


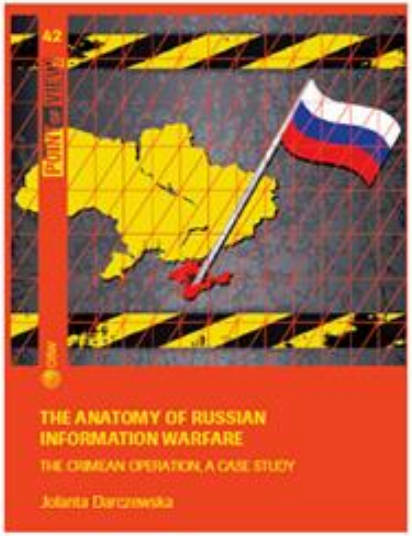


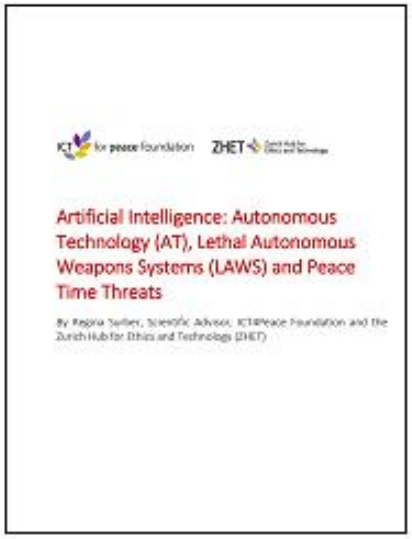
https://www.youtube.com/watch?v=RCXGpEmFbOw

https://www.youtube.com/watch?v=HCG_Hnv7nMY

Оружие новой войны







Учите материальную часть!

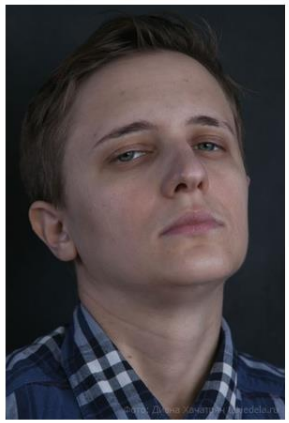
Digital Security

Обмануть нейросеть

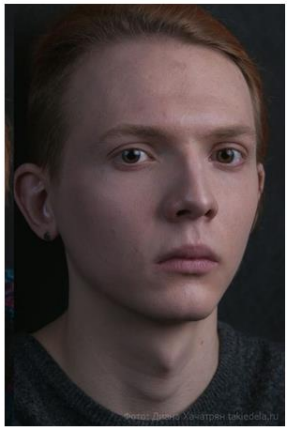
Man or Woman ?

Как обмануть систему распознавания

Adversarial training (состязательное обучение)





Источник: https://pikabu.ru/story/yeto_muzhchina_ili_zhenshchina_4440498

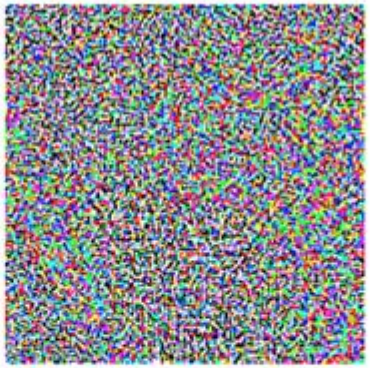
Adversarial Examples

вредоносные примеры

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



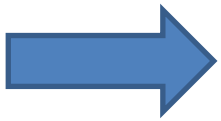
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

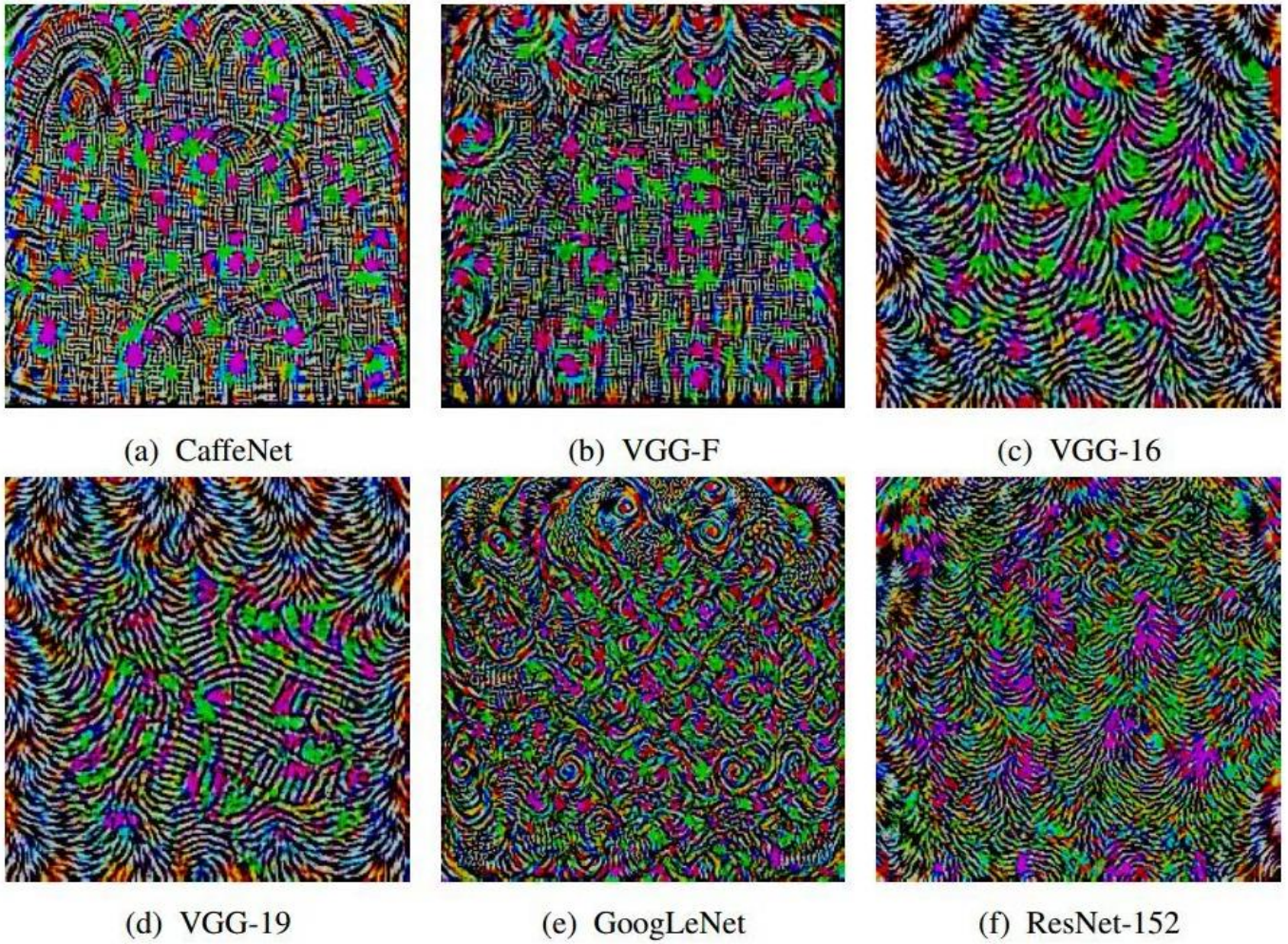
"gibbon"
99.3 % confidence



Источник: https://buzzrobot.com/4-ways-to-easily-fool-your-deep-neural-net-dca49463bd0

The "universal adversarial preturbation"

универсальное искажение

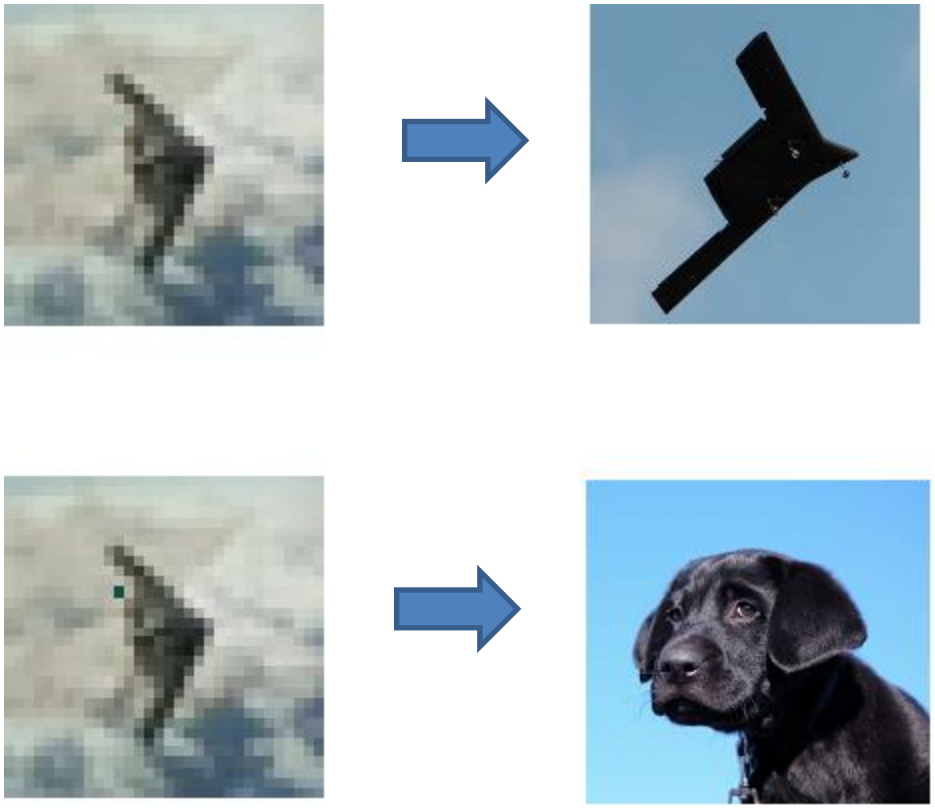


(a) CaffeNet (b) VGG-F (c) VGG-16

(d) VGG-19 (e) GoogLeNet (f) ResNet-152

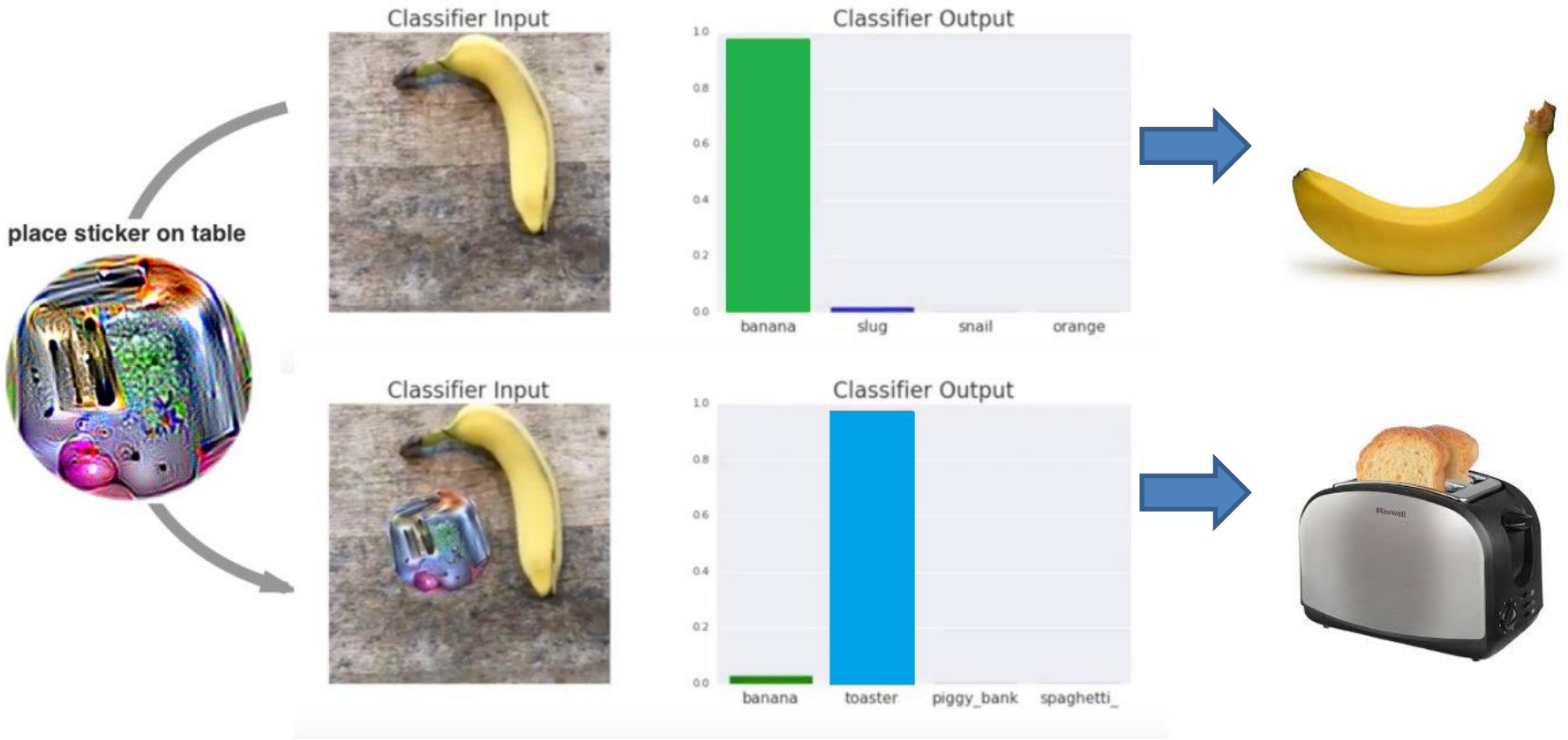
One Pixel Attack

for fooling deep neural networks



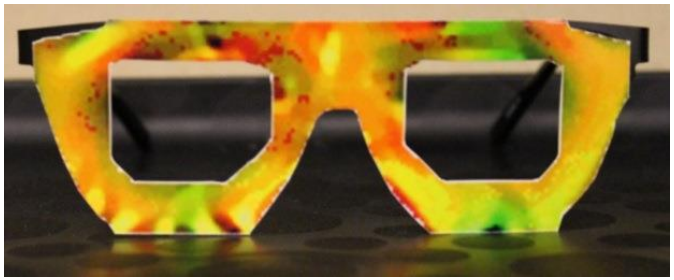
Источник: https://buzzrobot.com/4-ways-to-easily-fool-your-deep-neural-net-dca49463bd0

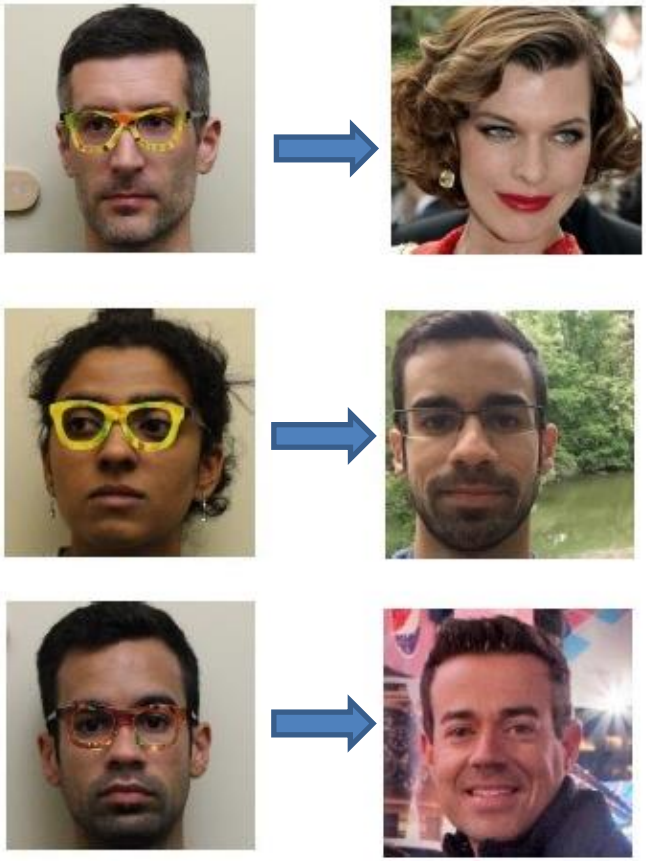
The Adversarial Patch
вредоносная заплатка

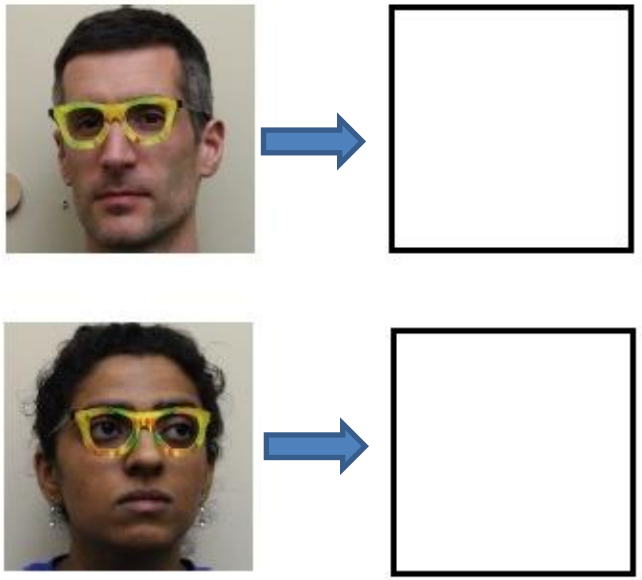


Источник: https://boingboing.net/2018/01/08/what-banana.html

Обмануть систему распознавания лиц







Обмануть беспилотный автомобиль

Дорожные знаки - обманки

Robust Physical-World Attacks on Deep Learning Visual Classification (04.2018)

Источник: https://arxiv.org/pdf/1707.08945.pdf

Спасибо за внимание ☺

Questions?

Masalovich Andrei

Масалович Андрей Игоревич

Специалист по связям с реальностью

+7 (964) 577-2012

avalanche100500@gmail.com

iam.ru/tipaguru.htm