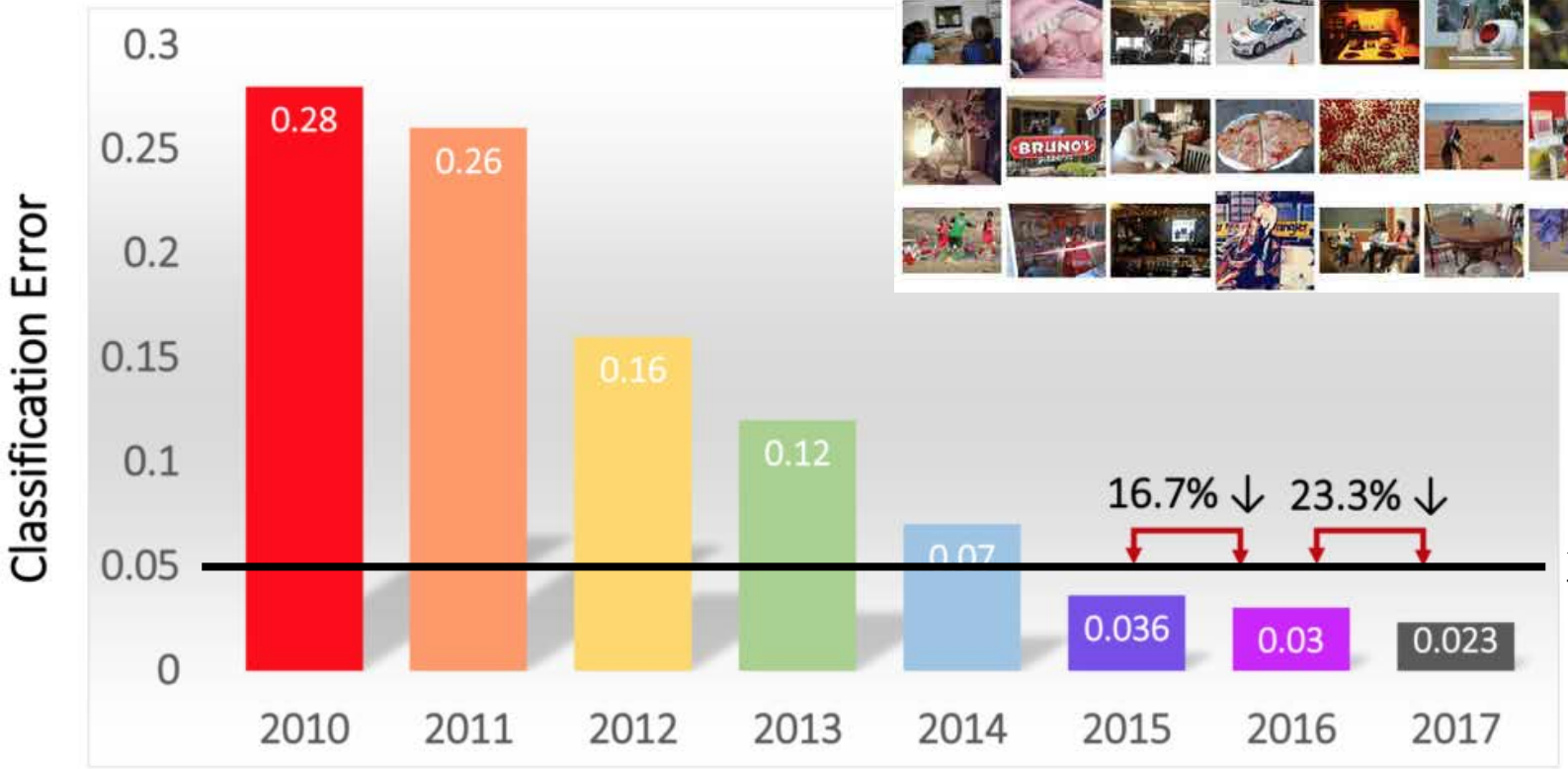
第9回 成長戦略ワーキング・グループにおけるヒアリング
「デジタル時代における刑事法の在り方について」

資料１－１

AIによる誤認識・偽造・差別の問題

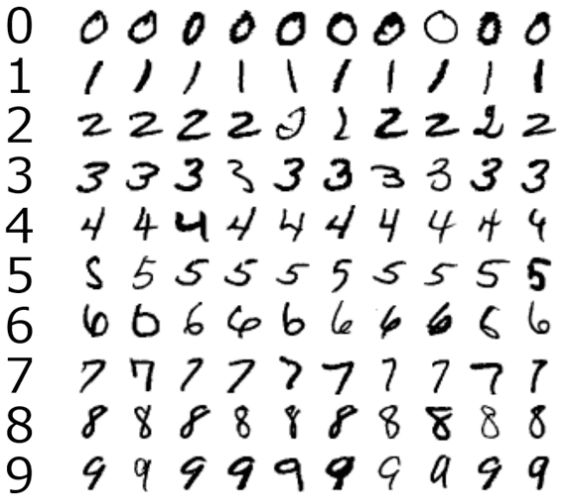
佐久間 淳
筑波大学 / 理研AIP

人間を超える
AIの画像認識能力

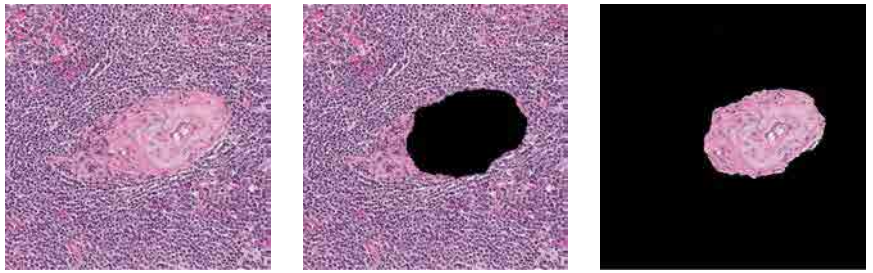


https://www.kaggle.com/getting-started/149448

AIに何を望む?



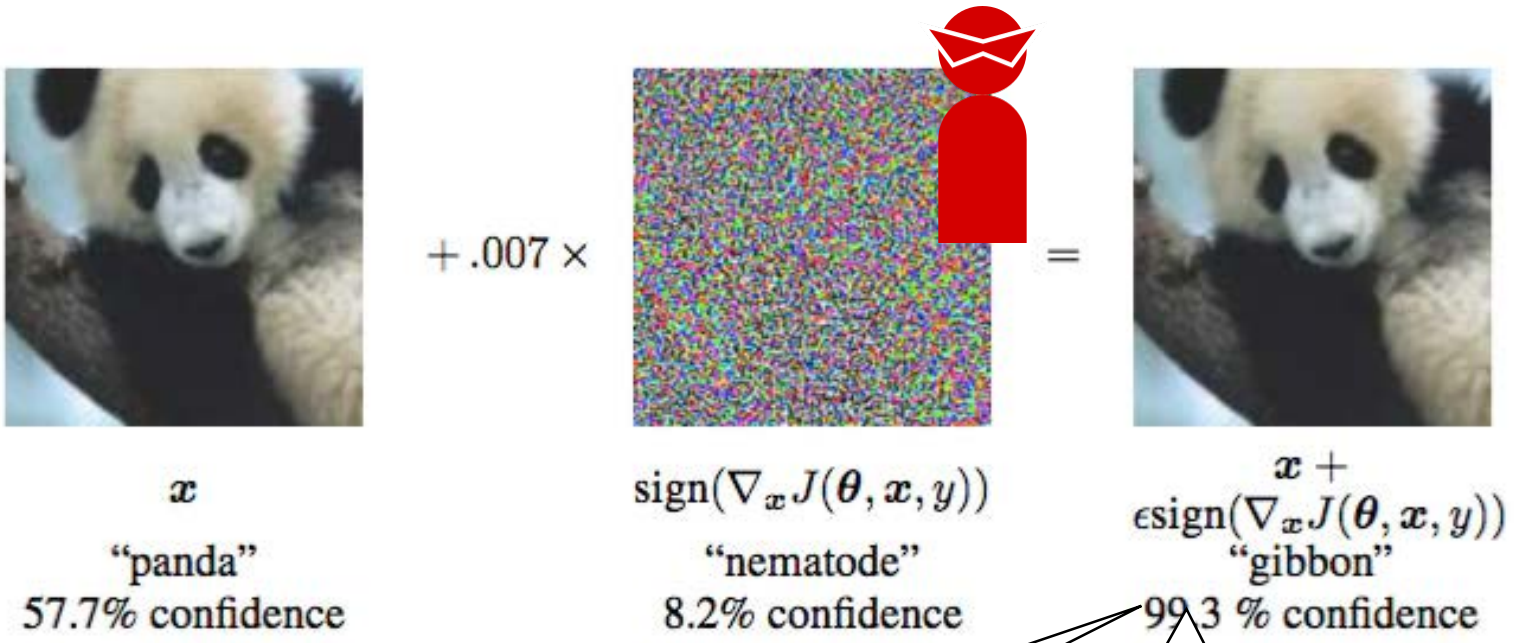
文字認識



医療診断、自動運転、etc.

- 単純な認識から、重大な判断へ
- 正確なだけでなく、安定した判断が重要
  - 正確＝精度が高い
  - 安定＝多少の変化に影響されない(天候、ノイズ、攻撃、etc.)

敵対的サンプル(AE) [Szegedy+14]



人間はパンダと認識

AIはテナガザルと認識

Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

現実世界における敵対的サンプル [Etimov et al.]

Images: Evtimov et al

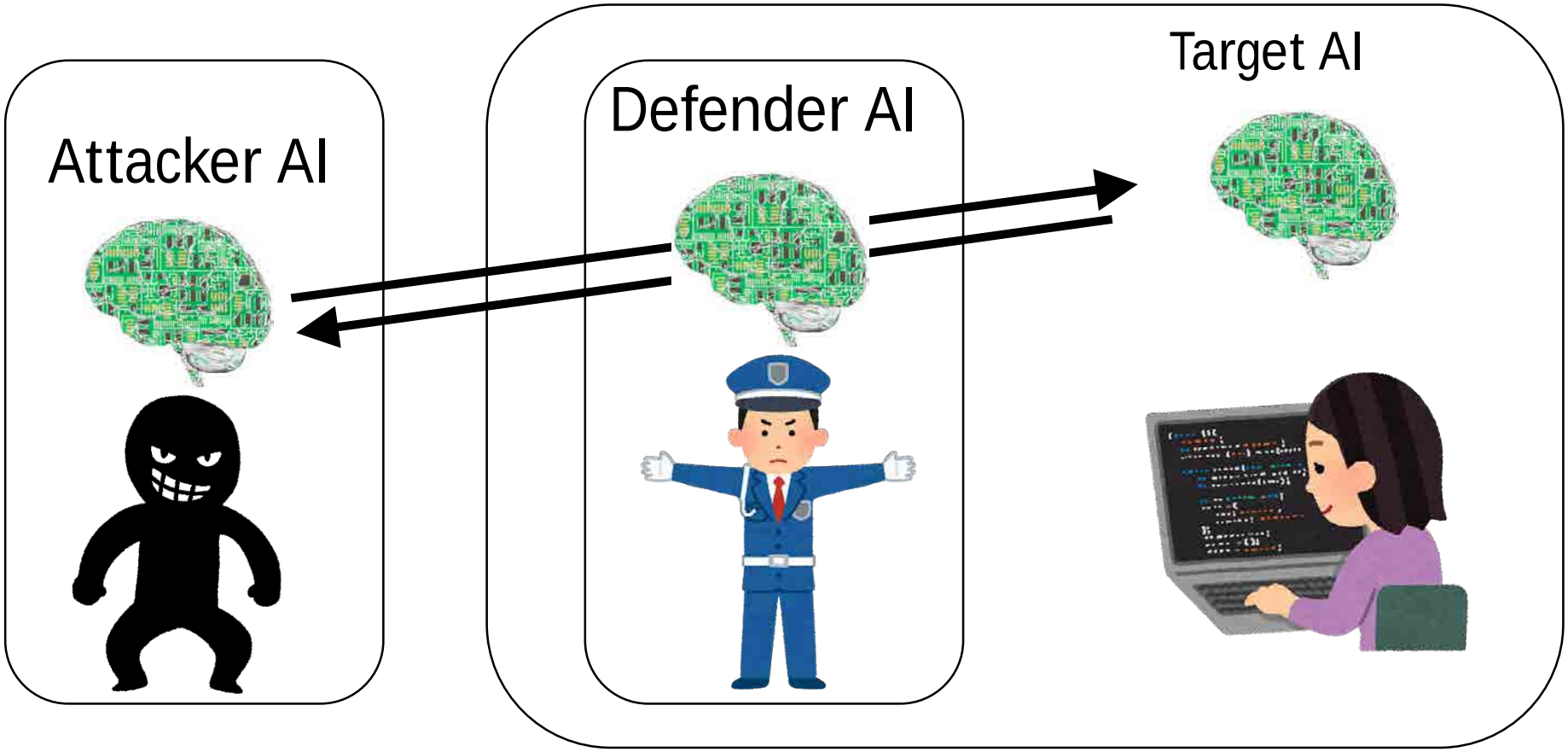
- 道路標識上に特殊なパターンを持つステッカーを添付
- AIは “STOP”標識を “speed limit 45 mph”と認識
- このステッカーが貼られた標識では、自動運転車はブレーキを働かせない可能性

音声の敵対的サンプル [YS, IJCAI'19]

メディアを通じた広い拡大のリスク

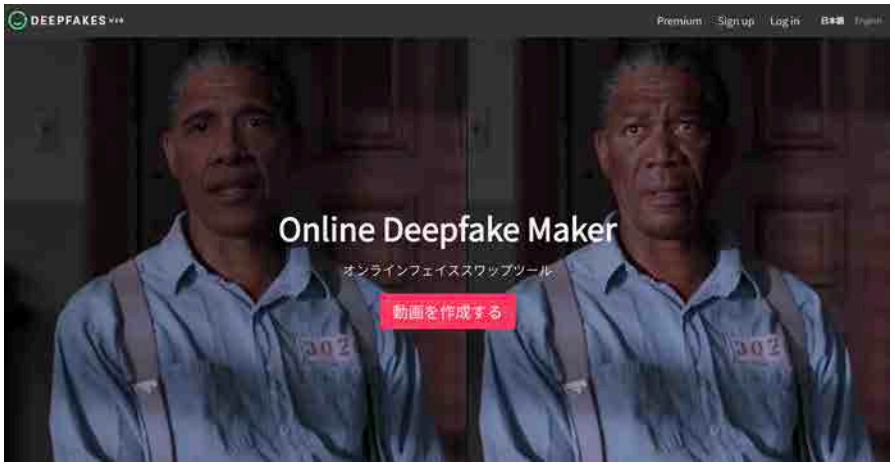
Yakura, Hiromu, and Jun Sakuma. "Robust audio adversarial example for a physical attack." *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019.

AI同士の攻防



攻撃者は常に防御AIを上回る攻撃AIを作成しようとする
いたちごっこを続けること自体がAIの安全性を守ることにつながる

AIによる人工メディア生成・操作



https://deepfakesweb.com/?locale=ja

実在しない人物画像生成　　実在する人物の動画偽造

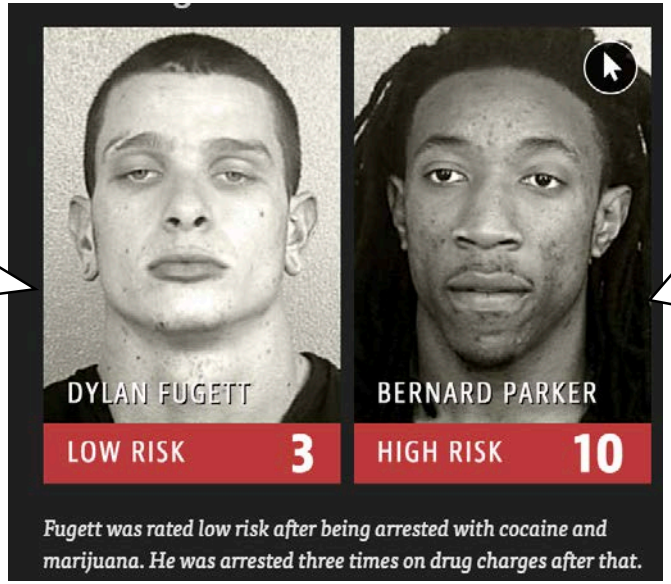
- 画像/音声/動画生成・操作
  - 偽造ポルノ、リベンジポルノ、証拠の捏造
- 人間の文章と見分けがつかない長文テキストの生成
  - デマ、フェイクニュース、虚偽方法、世論の誘導
- 従来より、巧妙かつ大量の自動生成

Tolosana, Ruben, et al. "Deepfakes and beyond: A survey of face manipulation and fake detection." *Information Fusion* 64 (2020): 131-148.

AIによる差別の実例 [Angwin+16]

- COMPASSは再犯リスクを予測（米国のいくつかの州）
  - スコアは被告人の保釈・判決・仮釈放等に影響
- COMPASSによる再犯リスクスコアリングは人種の影響を受けている [Angwin+16]
  - アルゴリズムによる決定が差別を生み出す

再犯リスク低と判定、
実際には三回再犯



再犯リスク高と判定、
実際には再犯なし

Angwin, Julia, et al. "Machine bias." *ProPublica, May* 23.2016 (2016): 139-159.
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

AIが社会に信頼されるためには

- エキスパートレベルの品質
- 法令遵守
- 環境変化・攻撃に対する安定性
- 人権の尊重
  - 自己決定権の尊重
  - プライバシー保護
  - 公平性への配慮
- 判断根拠の説明可能性
- 判断過程の追跡可能性

必要条件

制約

事後検証