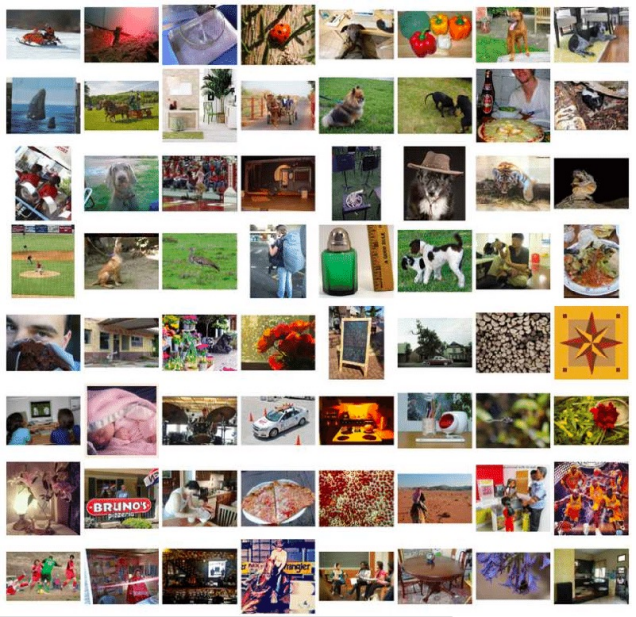
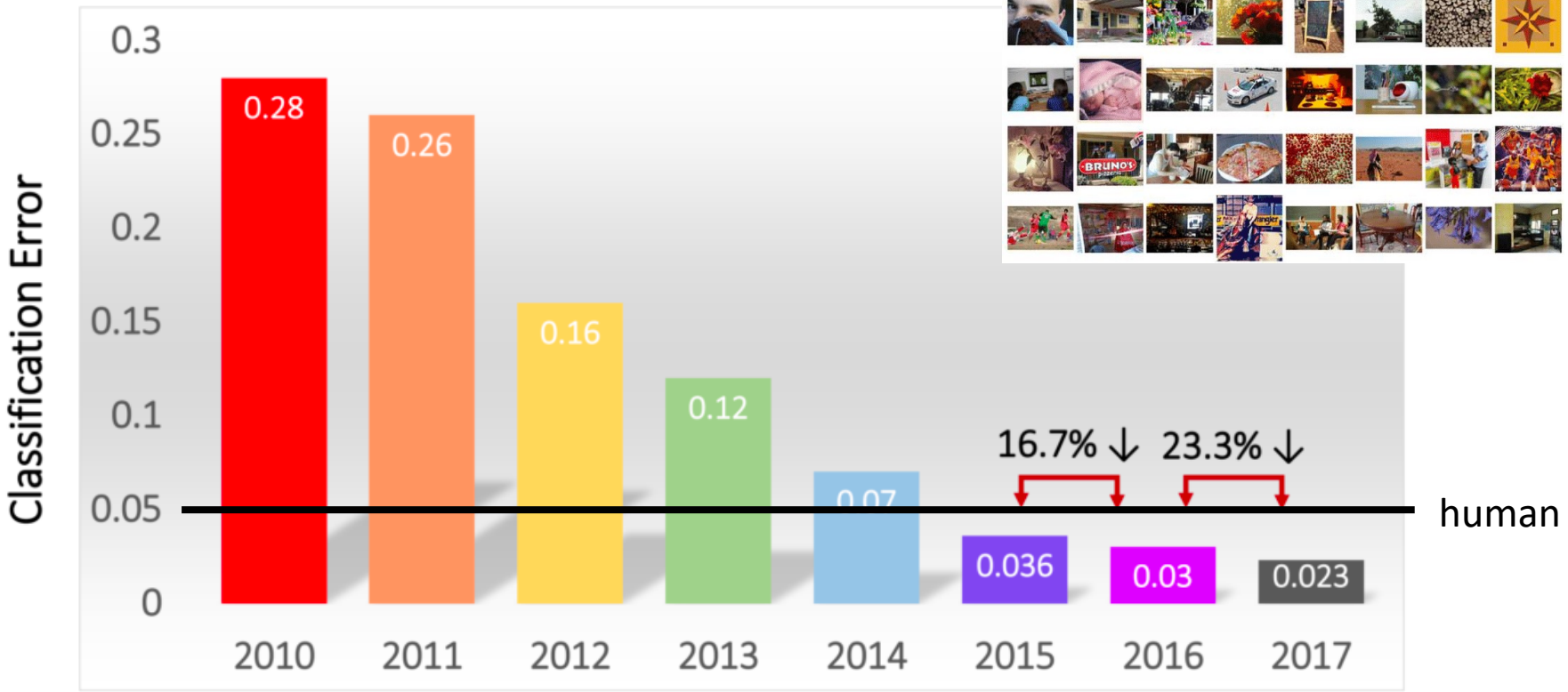
人間から信頼されるAIを目指して: AIセキュリティの観点から

佐久間 淳

筑波大学 / 理研AIP

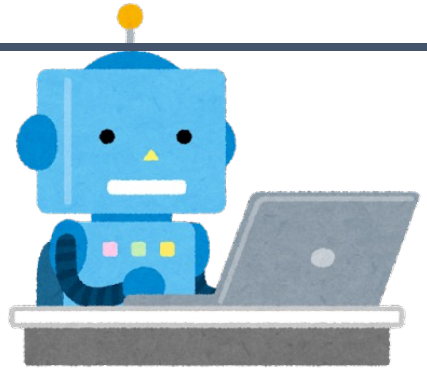
AIによる画像認識精度の向上





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

AI as Expert?



データ

意思決定

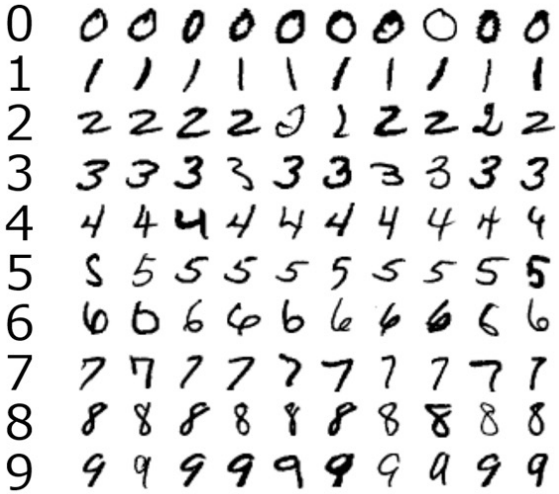
医師　→AIによる医療

運転手→自動運転

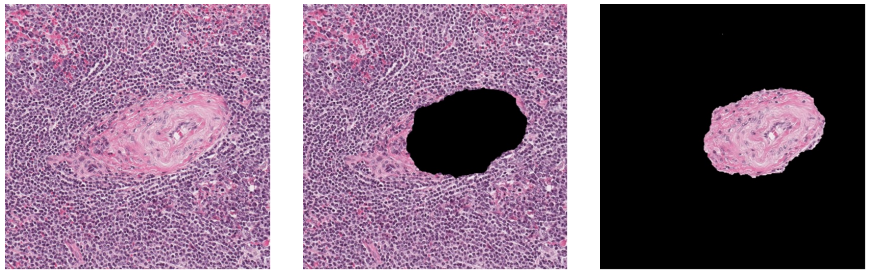
法律家→AIによる紛争解決

人事　→AIによる採用, etc…

AIに何を望む?



文字認識



医療診断

- 単純な認識から、重大な意思決定へ
- **正しいだけでなく、安全で信頼できる意思決定**

AIによる判断・意思決定が信頼される条件

- 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

AIによる判断・意思決定が信頼される条件

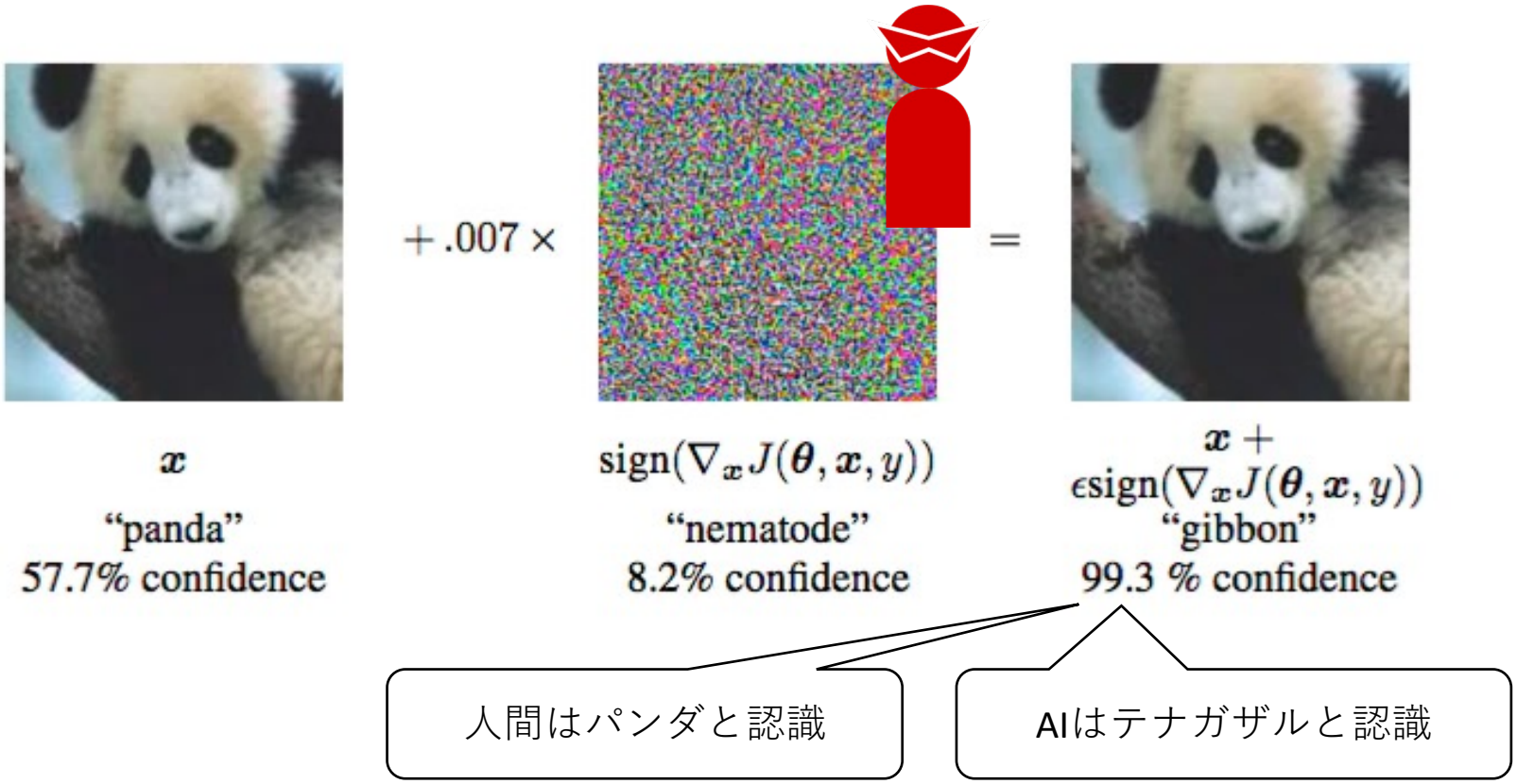
- エキスパートレベルの品質
- 法令遵守
- 環境変化・攻撃に対する安定性
- 人権の尊重
  - 自己決定権の尊重
  - プライバシー保護
  - 公平性への配慮
- 判断根拠の説明可能性
- 判断過程の追跡可能性

必要条件

制約

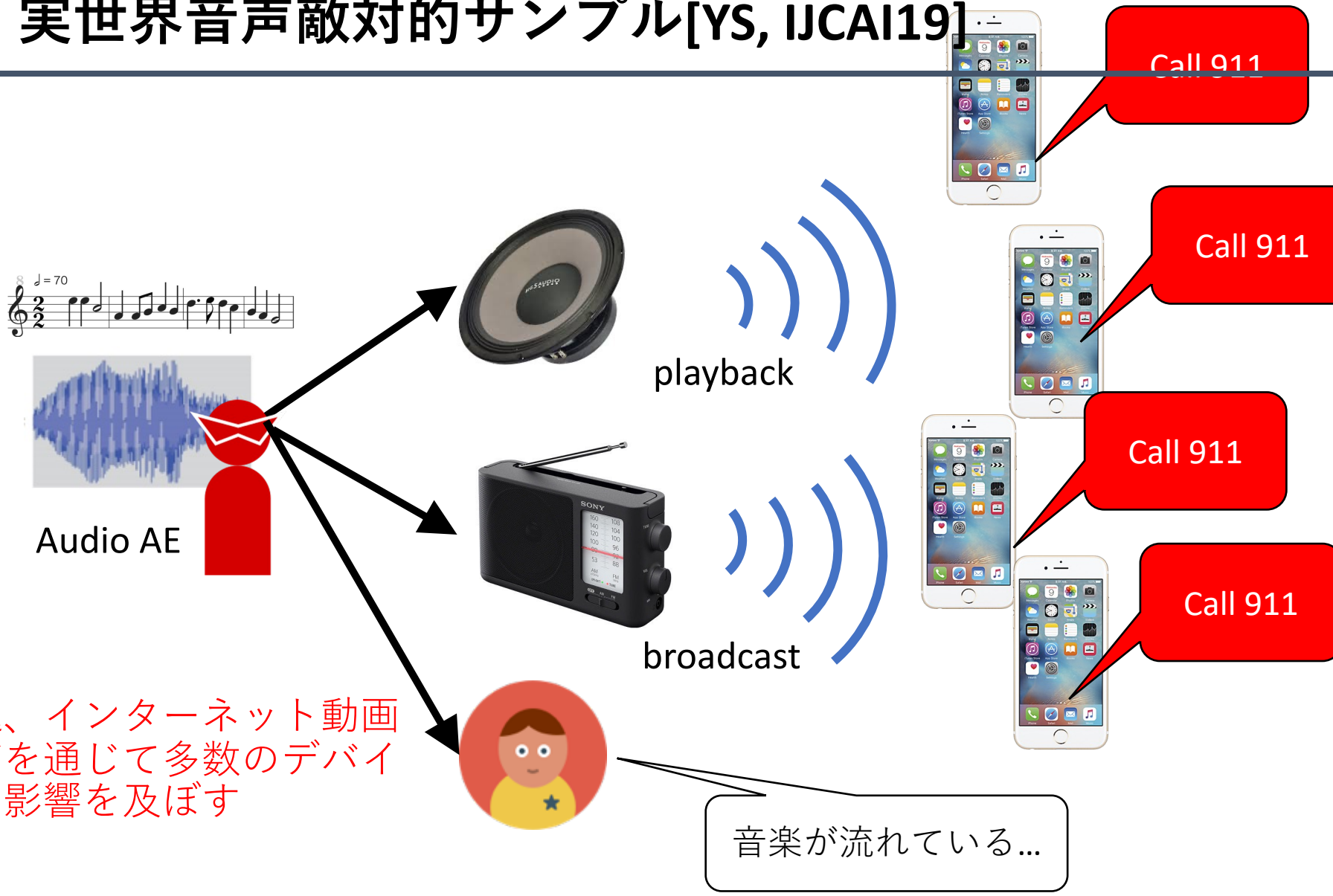
事後検証

AIを騙す画像：敵対的サンプル(AE)　[Szegedy+13]



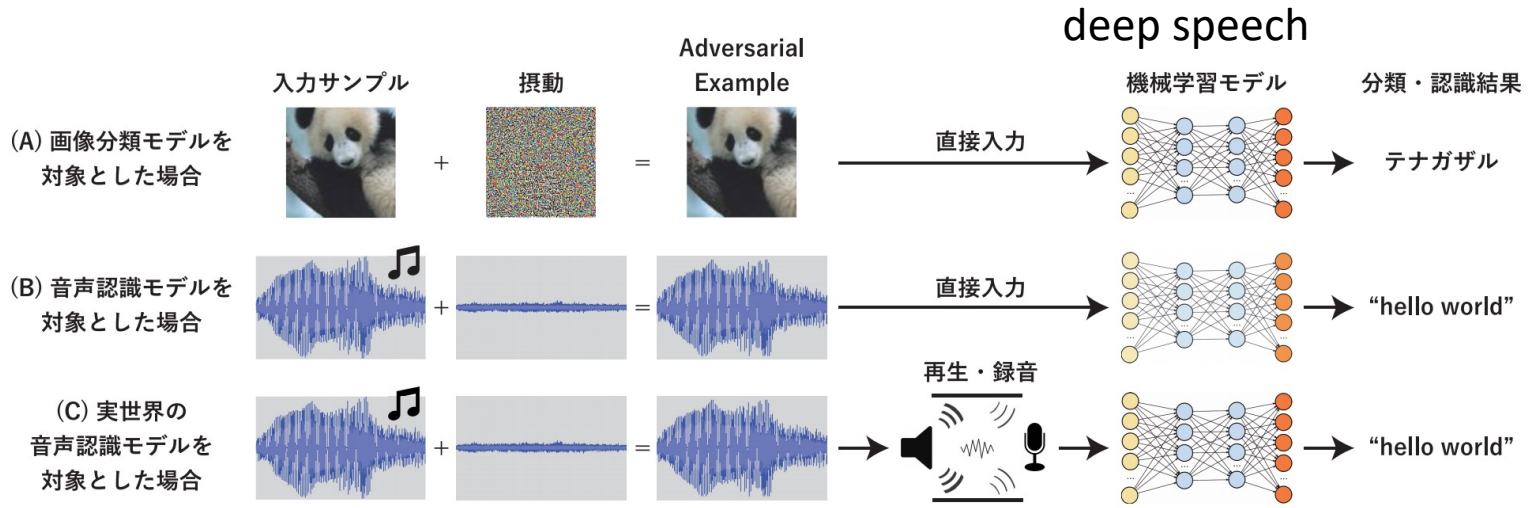
高精度と思われるAIにも多くの脆弱性が残されている
どのような脆弱性があり、それをどう克服するか？

実世界音声敵対的サンプル[YS, IJCAI19]



Robust Audio Adversarial Example for a Physical Attack, Hiromu Yakura, Jun Sakuma, IJCAI2019, (h5-index 67, 被引用回数 124, Core rank A*, 採択率17.9%)　被引用回数は同内容のws論文含む

実世界音声敵対的サンプル[YS, IJCAI19]



オリジナル音声

audio AE
(Hello Worldと認識)

主観評価実験

		正常でないと感じた	文字列を聞き取れた	選択肢に対して 正解	不正解	わからない
Hello world	(A)	36.0%	0.0%	4.0%	28.0%	68.0%
Open the door	(C)	64.0%	0.0%	12.0%	16.0%	72.0%

実世界を経由して動作する音声AEの作成に世界で初めて成功

Robust Audio Adversarial Example for a Physical Attack, Hiromu Yakura, Jun Sakuma, IJCAI2019

人間には自然物に見える敵対的サンプル

64x64 pixel　　128x128 pixel

人間には自然物に見えるがNNには敵対的サンプルとして機能するartifact
NNは上の写真を”速度制限”と認識する

ここだけ見ればWGAN

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}} \left[\log \mathcal{D}\left(\boldsymbol{v}\right)\right] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}} \left[\log\left(1-\mathcal{D}\left(\mathcal{G}\left(\boldsymbol{z}\right)\right)\right)\right]$$
$$+ \alpha \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}},\boldsymbol{\theta}\sim p_{\boldsymbol{\theta}}} \left[\mathcal{L}_f\left(A\left(\mathcal{G}\left(\boldsymbol{z}\right),\boldsymbol{x},\boldsymbol{\theta}\right),t\right)\right] \quad , \qquad (4)$$

位置、角度に対するロバスト化　　生成パッチ入り画像のターゲットクラスに対する損失

Generate (non-software) bugs to fool classifiers, Hiromu Yakura, Yohei Akimoto, Jun Sakuma, AAAI2020

音声の実世界敵対的サンプル

- 音声AE(ベースライン) Carlini+18

$$\underset{v}{\operatorname{argmin}}\ \underset{f}{\operatorname{Loss}}(x+v,l)+\varepsilon\|v\| \Rightarrow \underset{v}{\operatorname{argmin}}\ \operatorname{Loss}(\mathrm{MFCC}(x+v),\ l)+\varepsilon\|v\|$$

オリジナル音声 

audio AE
(Hello Worldと認識) 

- 提案法

$$\underset{v}{\operatorname{argmin}}\ E_{h\sim H,w\sim N(0,\sigma^2)}\left[\operatorname{Loss}(\mathrm{MFCC}(\underset{h}{\operatorname{Conv}}(x+\underset{1000\sim4000\,\mathrm{Hz}}{\mathrm{BPF}}(v))+w),\ l)+\varepsilon\|v\|\right]$$

再生環境に依存した反響に対するロバスト化
雑音に対するロバスト化

人間の可聴域は20-20000Hz
マイクとスピーカーもそれに合わせて設計

主観評価実験

		正常でないと感じた	文字列を聞き取れた	選択肢に対して 正解	不正解	わからない
Hello world	(A)	36.0%	0.0%	4.0%	28.0%	68.0%
Open the door	(C)	64.0%	0.0%	12.0%	16.0%	72.0%

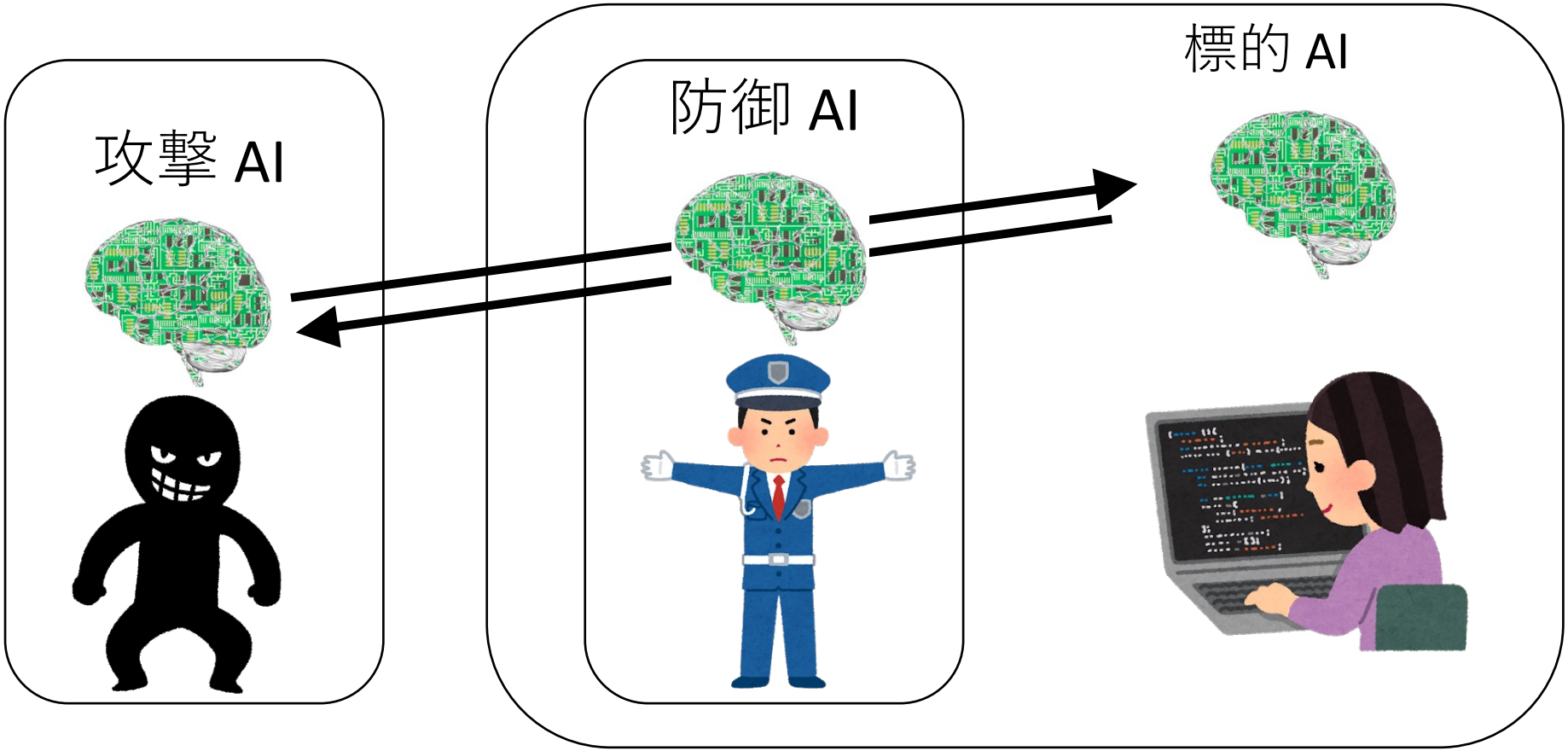
Robust Audio Adversarial Example for a Physical Attack, Hiromu Yakura, Jun Sakuma, IJCAI2019

「騙されないAI」を作るのは簡単ではない

防御	攻撃
	初めてのAE (2014)
• Adversarial training (2015～)	←非効率, 精度の低下
• Defensive distillation (2016)	←CW攻撃による無効化
• Detection methodology (2016)	←CW攻撃による無効化
• Obfuscated gradients(2017)	←特定の損失関数による無効化
• Certified defense (2017～)	←モデルサイズの制限、非効率
• Rand. smoothing (2019～)	←理論保証範囲の制限
…	

終わることのない、いたちごっこ

AIの安全性は、AI同士の攻防によって実現



- 通信の安全性は攻撃と防御の戦いを絶え間なく継続することによって実現
- AIの安全性も、「これで万全」という状態はない
- AIセキュリティも「攻撃と防御の絶え間ない戦い」が重要

AIによる判断・意思決定が信頼される条件

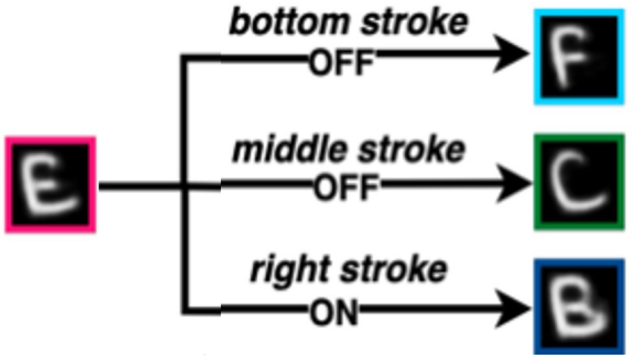
- エキスパートレベルの品質
- 法令遵守
- 環境変化・攻撃に対する安定性
- 人権の尊重
  - 自己決定権の尊重
  - プライバシー保護
  - 公平性への配慮
- 判断根拠の説明可能性
- 判断過程の追跡可能性

必要条件

制約

事後検証

「なぜ」を説明できる画像認識



”E”は

- 下に線がある (もしなければFになる)
- 真ん中に線がある (もしなければCになる)
- 右に線がない (もしあればBになる)

から、“E”である (反事実仮想による説明)

- コンセプトによる反事実仮想
  - 「クラスAの画像にコンセプトXが有れば（無ければ）、クラスAとは識別されない」
  - そのようなバイナリーコンセプトXを探索し、説明として活用する

$$\mathcal{L}(X) = \frac{1}{|X|} \sum_{\mathbf{x} \in X} \left[\mathcal{L}_{\mathrm{VAE}}(\mathbf{x}) + \lambda_{\mathrm{R}} \mathcal{L}_{\mathrm{R}}(\mathbf{x}) \right] + \lambda_{\mathrm{CE}} \mathcal{L}_{\mathrm{CE}}(X).$$

潜在変数を求めるVAE

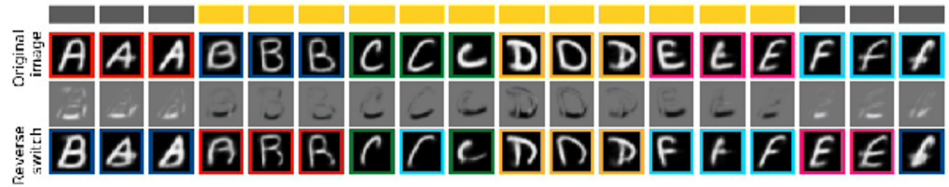
潜在変数を単純化する正則化

潜在変数の変化と識別ラベルの相互情報量を最大化

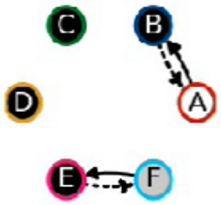
Unsupervised Causal Binary Concepts Discovery with VAE for Black-box Model Explanation, Thien Q Tran, Kazuto Fukuchi, Youhei Akimoto, Jun Sakuma, AAAI2022.

クラス間遷移を要因レベルで概観

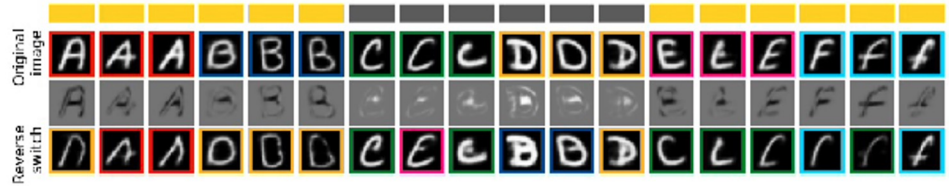
EMNIST (A, B, C, D, E, F)



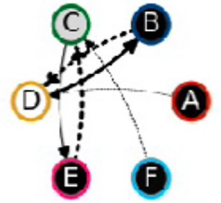
(a) Controlling switch γ_0 of concept m_0 (bottom stroke)



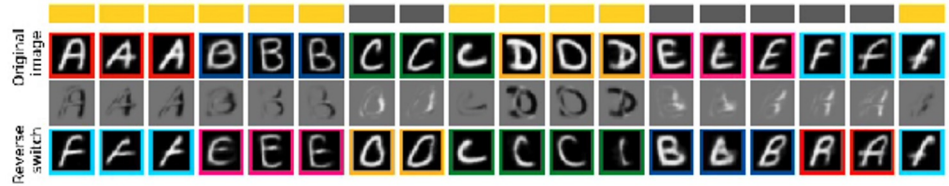
(b) Transition by m_1



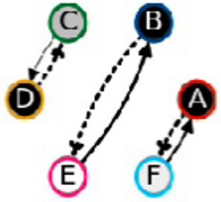
(c) Controlling switch γ_1 of concept m_1 (middle stroke)



(d) Transition by m_1



(e) Controlling binary γ_2 of concept m_2 (right stroke)



(f) Transition by m_2

クラス間の判断根拠をバイナリーコンセプトに基づく反事実仮想として表示

Unsupervised Causal Binary Concepts Discovery with VAE for Black-box Model Explanation, Thien Q Tran, Kazuto Fukuchi, Youhei Akimoto, Jun Sakuma, AAAI2022.

AIセキュリティにおけるIntegrity

- セキュリティ3要素:　Confidentiality(機密性), Integrity(完全性), Availability(可用性)
- 一般の情報システムの Integrity
  - 敵対的環境下でも、仕様通り情報処理がなされていること
- AIのIntegrity
  ≠　仕様（AI のモデル／パラメータ）通りの情報処理
  =　敵対的環境下でも、人間と同等な判断／学習の実現
- 人間の判断に仕様を定義することはできない
  - 仕様が定義できるシンボリックな世界
  - NNでなければ取り扱えない複雑な信号/感覚の世界
  - 両者をどう繋ぎ合わせるか