資料2-4

AIセキュリティについて

佐久間淳
筑波大学/理研AIP

情報セキュリティとAIセキュリティ

- 情報システムの役割
  - 蓄積 (e.g. ストレージ)
  - 管理 (e.g. DB)
  - 変換 (e.g. 情報処理)
  - 伝達 (e.g. インターネット)
- 情報
  - デジタル前提
- 挙動
  - 決定的
  - ルールベース
- セキュリティ
  - 機密性
  - 完全性
  - 可用性

- AIシステムの役割
  - 生成 (e.g. 画像生成)
  - 表現・認識 (e.g. 音声認識)
  - 分析 (e.g. 画像診断)
  - 意思決定 (e.g. 自動運転)
- 情報
  - アナログ前提
- 挙動
  - 確率的
  - 統計ベース
- セキュリティ
  - 機密性?
  - 完全性?
  - 可用性?

＋α？

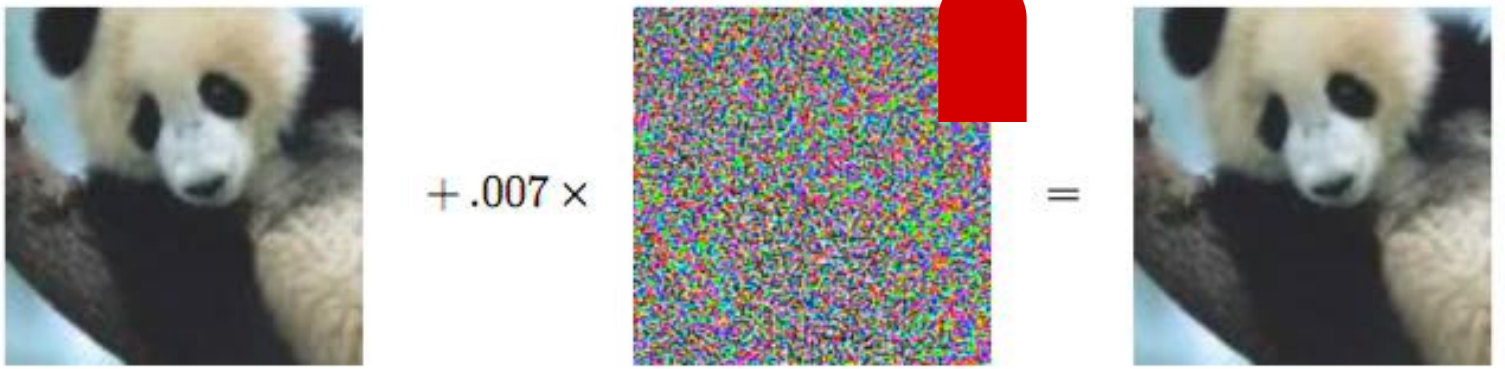
AIシステムの完全性

- 情報システムの完全性
  =システムがあるべき挙動から逸脱しない
  - 「あるべき挙動」は論理的に記述可能
  - 「逸脱」も検出可能

- AIシステム(e.g. 画像診断)の完全性
  =画像診断システムがあるべき挙動から逸脱しない
  - 「あるべき挙動」=専門医師と同じ思考プロセスによる診断
  - 「あるべき挙動」がルールベースで記述できない
  - 情報システムと同じ意味での「完全性」は定義できない

敵対的サンプル [Szegedy+14]

巧妙に設計
されたノイズ



$\boldsymbol{x}$

"panda"
57.7% confidence

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

"nematode"
8.2% confidence

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

"gibbon"
99.3 % confidence

人間はパンダと認識

AIは手長猿と認識

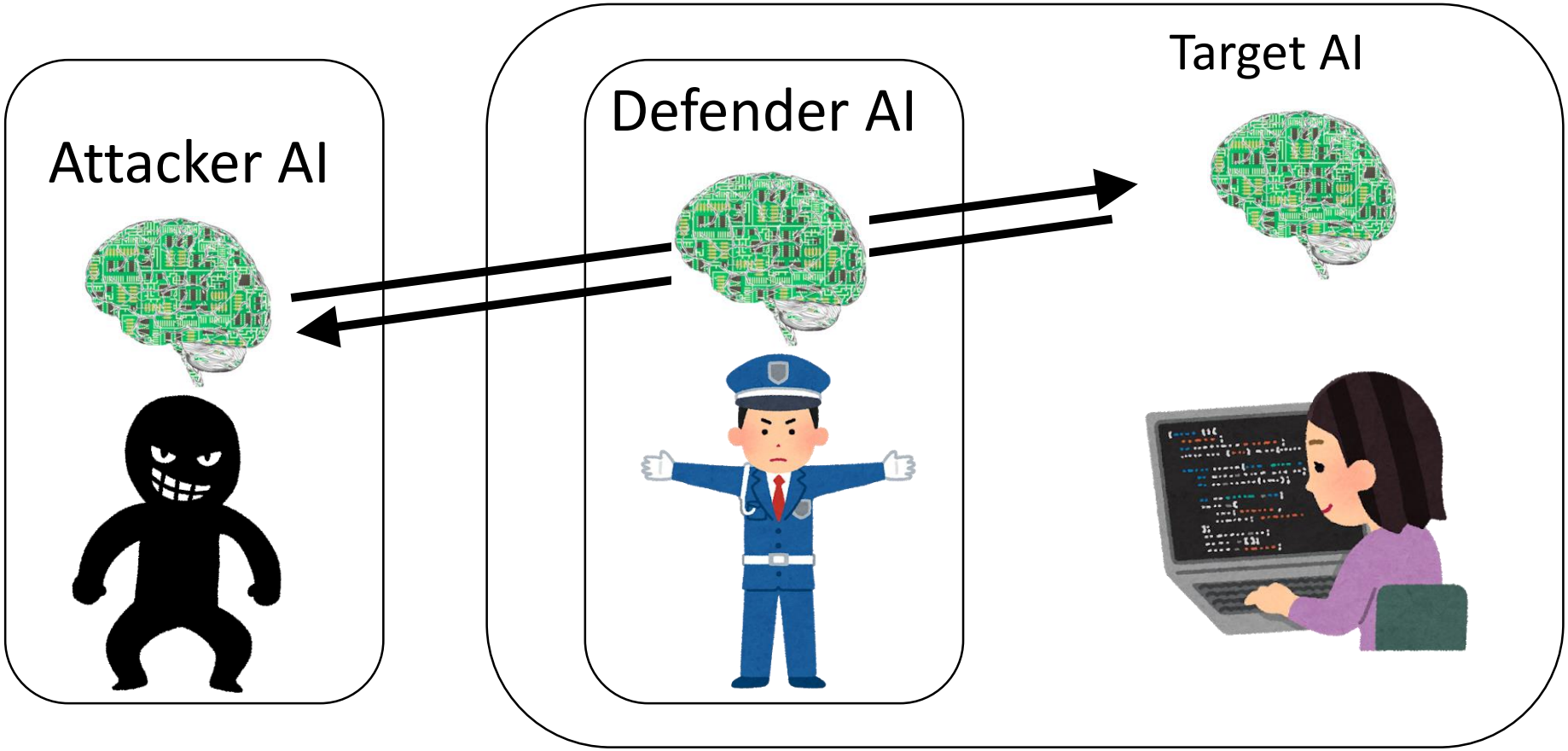
Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

敵対的サンプル・攻防の歴史

Defense	Attack
	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

Arms race of AIs



AI/深層学習のセキュリティ

- AI/深層学習
  - 複雑で大規模
  - ルールベースでなく統計ベース
  - データに依存した確率的な振る舞い
  - 性能は良いが中身はブラックボックス
  - 攻撃者が有利、防御は不利

- 「情報通信」のセキュリティは千年経っても解決していない

→ 継続的な攻撃と防御の研究によるセキュリティの維持

- 「認識」「意思決定」のセキュリティも、おそらく同様

信頼できるAIの実現に向けて

- 専門家レベルの意思決定
- 法令遵守
- 攻撃耐性
- 人権尊重
  - 自己決定権の尊重
  - プライバシーの保護
  - 公平性保証 (差別的決定の排除)
- 説明性
- 透過性

AIによる意思決定の必要条件

AIによる意思決定の制約条件

AIによる意思決定の検証