

日本のサイバーセキュリティを「連携」「学び」「創造」

AI x Security

日本ネットワークセキュリティ協会 (JNSA)
2020年3月4日(水)

目次

1. JNSAのご紹介 & 全体スコープの説明
唐沢勇輔（JNSA社会活動部会）
2. AIを守るには ～機械学習特有のセキュリティ上の脅威と対策について～
松岡正人、福田尚弘（JNSA調査研究部会 IoTセキュリティWG）
3. AIを用いた攻撃
高江洲勲（三井物産セキュアディレクション）

JNSAのご紹介・全体スコープの説明

唐沢勇輔（JNSA社会活動部会）

JNSAのご紹介①

- 名称 特定非営利活動法人 日本ネットワークセキュリティ協会
JNSA (Japan Network Security Association)
- 設立 2000年4月 (任意団体として発足、NPO法人化は2001年)
- 会員数 239社 (2019年10月25日現在)
- 住所 本部 東京都港区西新橋
西日本支部 大阪府大阪市淀川区西中島

- 役員
- 会長 田中 英彦 (情報セキュリティ大学院大学 名誉教授)
- 副会長 中尾 康二 (国立研究開発法人情報通信研究機構)
高橋 正和 (株式会社Preferred Networks)
- 事務局長 下村 正洋 (株式会社ディアイティ)

JNSAのご紹介②

ネットワークの急速な普及

インターネットの拡大(誰でも、どこでも)
利用者が一般人まで(初心者からプロまで)
すべてがネットワーク(社内データ、機密情報)など



「AI x Security」全体構成

「AIとセキュリティ」4つの論点

- a) Attack using AI (AIを利用した攻撃)
→ **アジェンダ3 高江洲氏の発表**
- b) Attack by AI (人間を超越したAI自身による攻撃)
→ 本日は対象外
- c) Attack to AI (AIへの攻撃)
→ **アジェンダ2 松岡氏の発表**
- d) Measure using AI (AIを利用したセキュリティ対策)
→ 本日は対象外

出典：佐々木良一(東京電機大学)「AIとセキュリティ」
<https://digitalforensic.jp/2018/09/18/column531/>

AIを守るには

松岡正人、福田尚弘（JNSA調査研究部会 IoTセキュリティWG）

調査の経緯

機械学習特有のセキュリティ上の
脅威と対策について

JNSA IoTセキュリティWGの活動成果と課題 ～脅威分析の例（ガイドライン等）～

想定される脅威：汎用マイコンボード
表1: 設定ミス、ウイルス感染



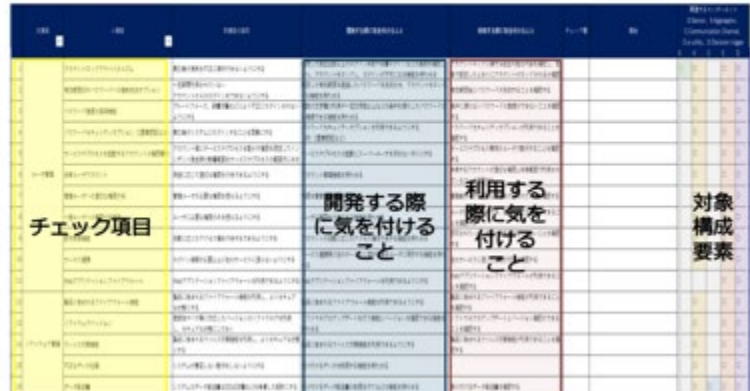
脅威	説明	利用開始・導入初期	平常運用時	対策	発生時	影響、被害状態	異い侵入・被害時
操作ミス	IoTデバイス内のユーザインタフェースを利用して、利用者が行った操作・設定が誤ったことにより発生される脅威 ・意図しないサービス更新時に個人情報を送付してしまう、迷惑メールが送信されること、等	ID、パスワード、通信先などのデフォルト設定の確認・変更機能を実装する ・意図しないサービス更新時に個人情報を送付してしまう、迷惑メールが送信されること、等のリスク（誤り）による動作確認を実装する	定期的な認証情報の更新がある ・動作監視（モニタリング）機能がある ・設定変更されていないことの確認機能がある（構成情報更新時にメール通知など） ・ログの取得による不正動作の検知ができるようにする	通信先などの異常を自動検知してメール等で通知する ・動作監視（モニタリング）機能がある ・異常の種類が判別できる ・設定のロールバックができるようにする ・認証情報に有効期限を設ける ・ラベルや注釈を添付し、システム内での設定変更の目的、高取グループ、管理者などが認識可能となるものである情報を併記する ・連携先に異常を連絡する旨のメール機能を実装する	定期的な認証情報の更新がある ・動作監視（モニタリング）機能がある ・異常の種類が判別できる ・設定のロールバックができるようにする ・認証情報に有効期限を設ける ・ラベルや注釈を添付し、システム内での設定変更の目的、高取グループ、管理者などが認識可能となるものである情報を併記する ・連携先に異常を連絡する旨のメール機能を実装する	デバイス内の設定変更の目的化ができるようにする ・重要時は物理的に読み出しを不可能にするように設計する ・ラベルや注釈を添付し、システム内での設定変更の目的、高取グループ、管理者などが認識可能となるものである情報を併記する ・連携先に異常を連絡する旨のメール機能を実装する	デバイス内の設定変更の目的化ができるようにする ・重要時は物理的に読み出しを不可能にするように設計する ・ラベルや注釈を添付し、システム内での設定変更の目的、高取グループ、管理者などが認識可能となるものである情報を併記する ・連携先に異常を連絡する旨のメール機能を実装する
ウイルス感染	利用者や外部から持ち込んだ機器や記憶媒体によって、IoTシステムがウイルスや悪意あるソフトウェア（マルウェア等）等に感染することにより発生される脅威 ・IoTデバイスに感染したウイルスがネットワークを通じて他のIoTデバイスに感染、等	ボード購入元の信頼性を確認する（ウイルスは含まれていないか） ・ネットワークに安全な環境下で設定を行うようにする ・更新システムに接続する前に安全な時の設定が行われるようにする ・最新のセキュリティパッチを適用されるようにする	定期的なウイルスチェックができるようにする ・製造元からの脆弱性情報を配信する ・0-dayの取得による不正動作の検知ができるようにする	動作状況のわかずい通信 ・安全なシーケンスで再起動を実行する ・安全な停止、入出力やネットワーク切替などができるようにする	定期的なウイルスチェックができるようにする ・製造元からの脆弱性情報を配信する ・0-dayの取得による不正動作の検知ができるようにする	定期的なウイルスチェックができるようにする ・製造元からの脆弱性情報を配信する ・0-dayの取得による不正動作の検知ができるようにする	定期的なウイルスチェックができるようにする ・製造元からの脆弱性情報を配信する ・0-dayの取得による不正動作の検知ができるようにする

Copyright (c) 2015-2016 NPO日本ネットワークセキュリティ協会 117

JNSAコンシューマー向けIoTセキュリティガイド（2015）

IoTセキュリティ評価のためのチェックリスト

■ 運用レベルでのIoTの脅威と対策を理解するため、開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中



チェック項目	開発する際に気をつけること	利用する際に気をつけること	対象構成要素
1. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
2. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
3. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
4. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
5. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
6. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
7. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
8. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
9. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
10. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
11. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
12. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
13. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
14. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
15. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
16. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
17. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
18. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
19. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			
20. 開発者・利用者双方が確認したい項目をなるべく具体的にまとめたチェックリストを作成中			

Copyright ©2018 JPCERT/CC All rights reserved. JPCERT/CC

JPCERT/CC IoTセキュリティのためのチェックリスト（2018）

AIシステムに特有の {脅威、対策} がない

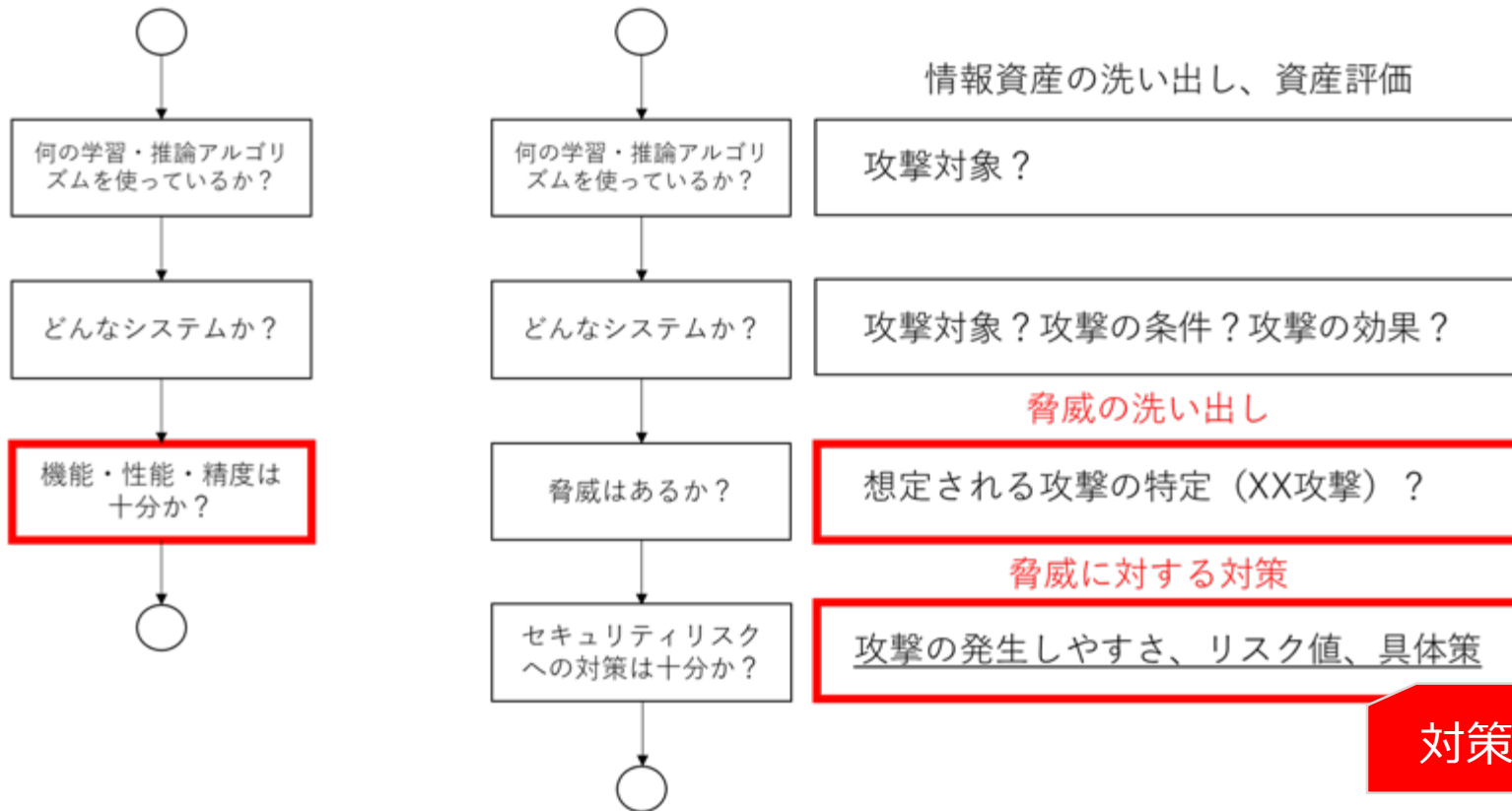
出典[3] : https://www.jnsa.org/seminar/2018/0226/data/3_koshiishi.pdf

出典[4] : <https://www.jpccert.or.jp/tips/2016/wr162501.html>

JNSA IoTセキュリティWG ～AIセキュリティ調査の動機～

品質：システム (x)

セキュリティ：システム (x)



機械学習の脅威と対策を明確化

機械学習特有の脅威（攻撃）

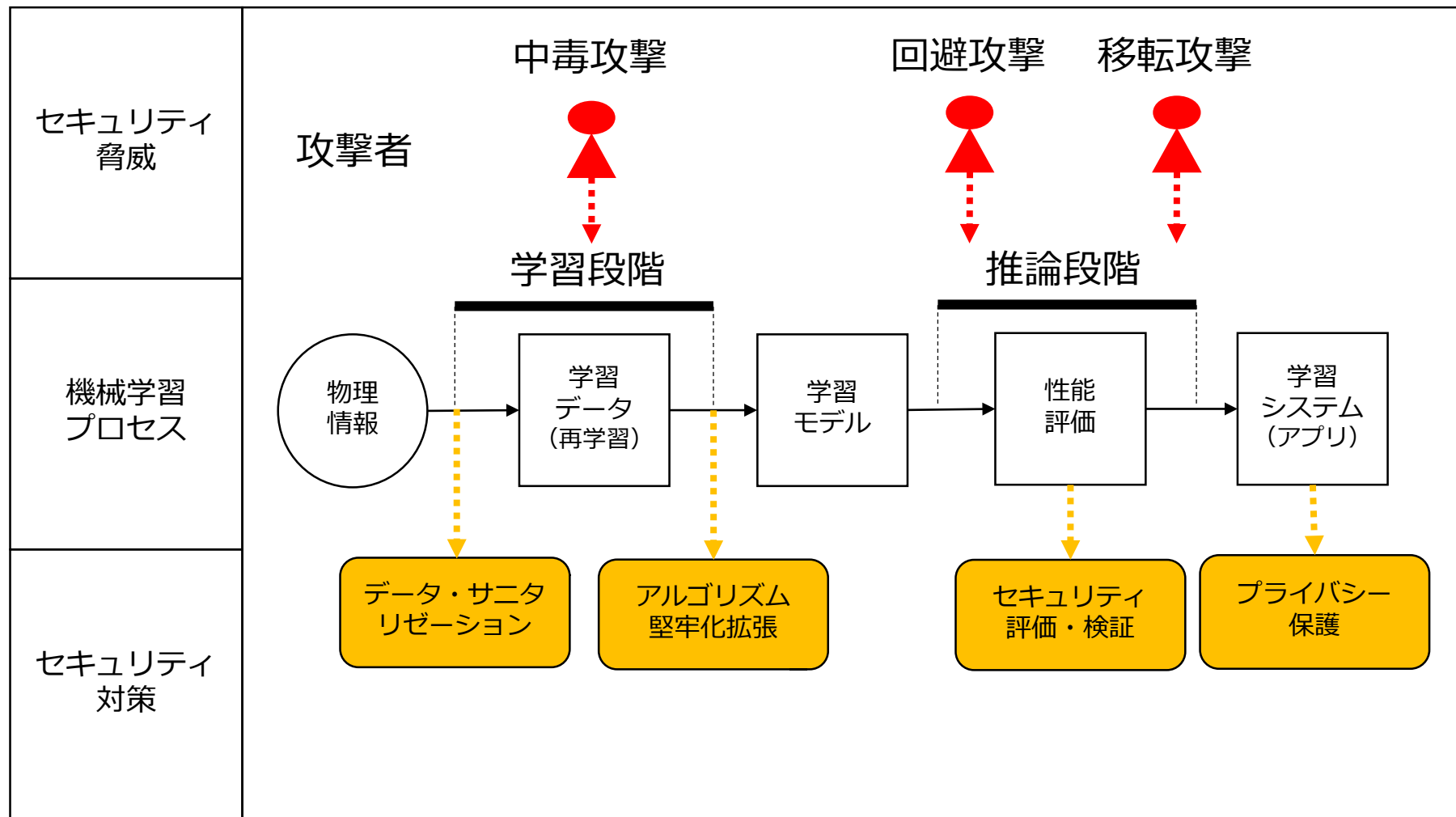
機械学習特有のセキュリティ上の
脅威と対策について

機械学習への攻撃

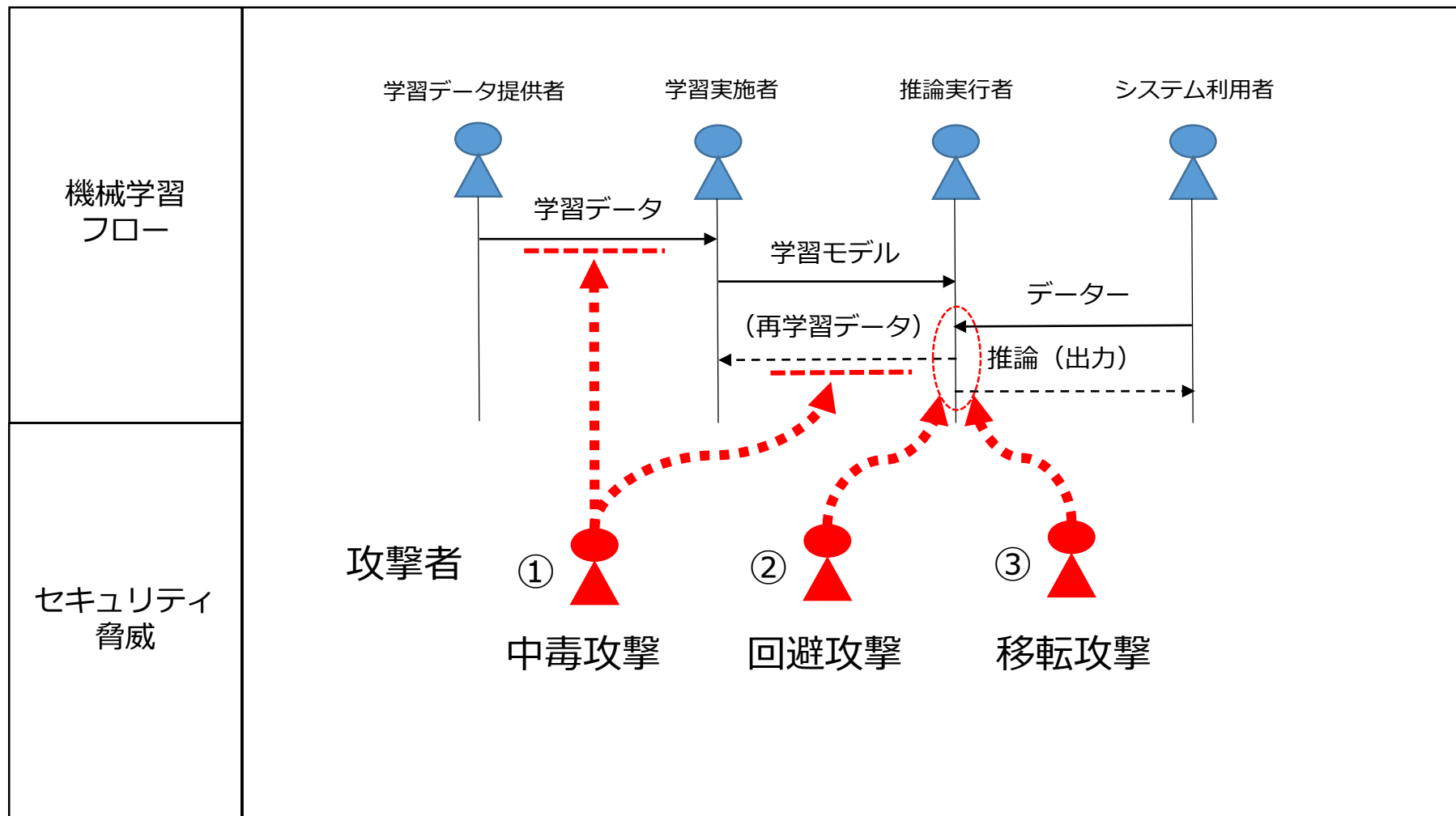
攻撃（脅威）	サブ分類	内容
回避攻撃 (Evasion Attacks)	<ul style="list-style-type: none"> 透明化攻撃 (Stealth) なりすまし攻撃 (Impersonate) 	人間には認識できない「 <u>摂動</u> を含んだデータ入力」により、 <u>人間と機械学習の推論エンジンとで異なる認識を起こす</u> 攻撃（画像、音声、文字等）
中毒攻撃 (Poisoning Attacks)	<ul style="list-style-type: none"> 可用性攻撃 (Availability) バックドア攻撃 (Backdoor) 	学習データへの「 <u>不正データの入力</u> （注入）」により、 <u>学習モデルの境界を何らかの方法によりシフト</u> する攻撃 （機械学習のモデル境界を大量の不良データの注入により使用不能とする可用性攻撃と、少量の洗練されたデータ注入によりバックドアを生成するバックドア攻撃がある）
移転攻撃 (Inversion Attacks)	<ul style="list-style-type: none"> プライバシー攻撃 (Privacy) メンバーシップ推論攻撃 (Membership) 	機械学習の「 <u>推論エンジンへのデータ入出力または反応</u> 」によって、元データなどの <u>機密情報</u> （またはモデル自体）を抽出する攻撃 （メンバーシップ推論攻撃では、敵対者が手元データが相手のデータセットに含まれているかを探る攻撃）

※違う名称、分類もあるが、ここでは3つの攻撃まとめた

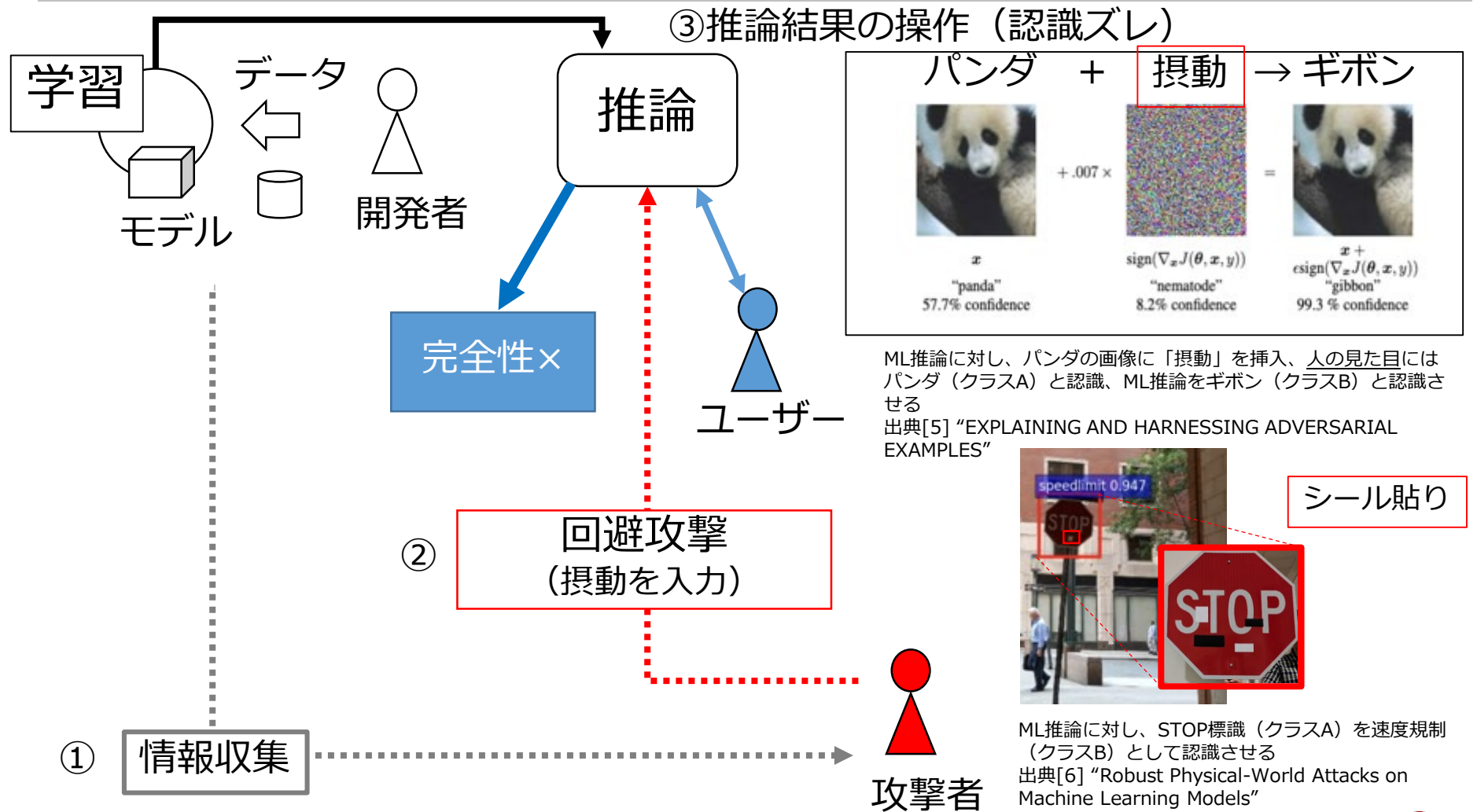
機械学習と脅威



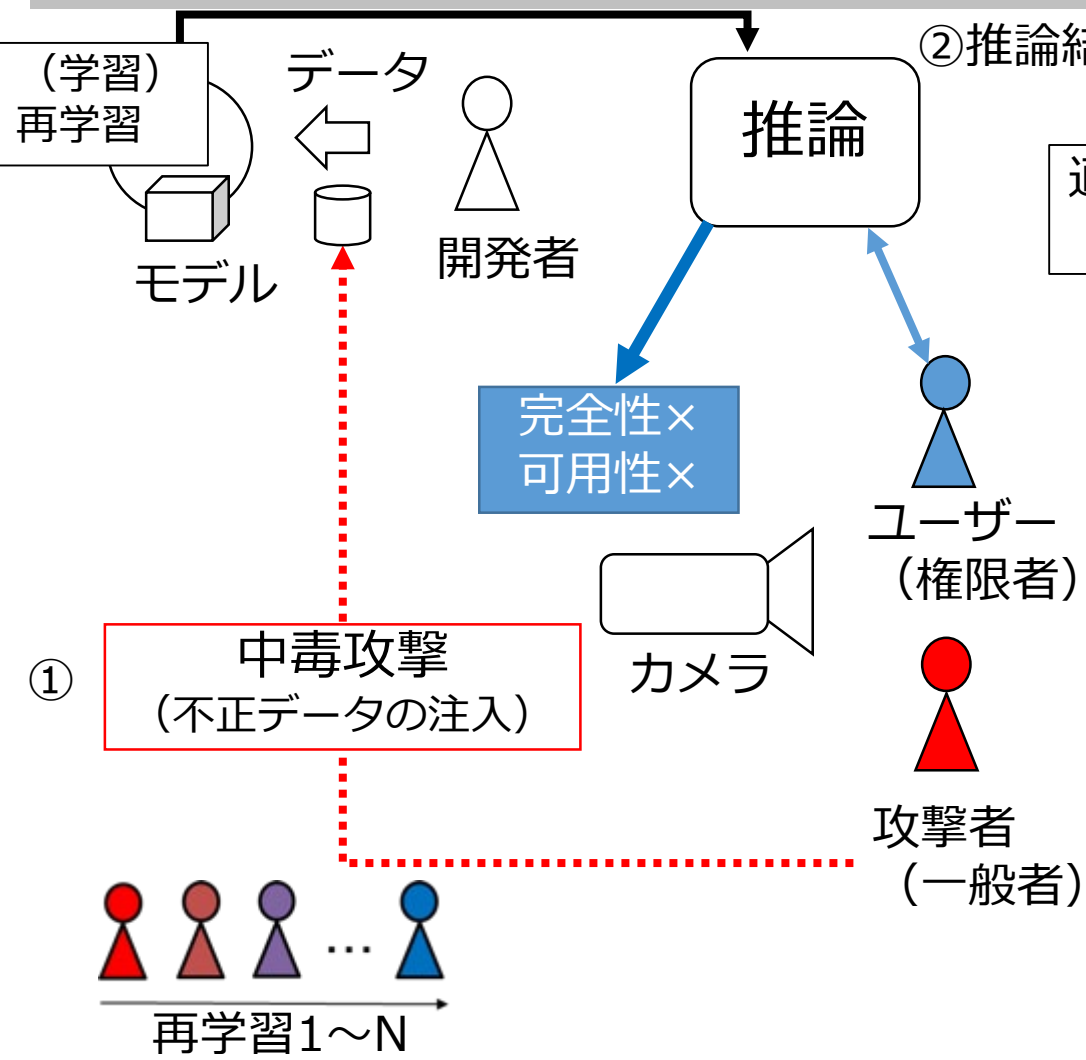
機械学習と脅威（フロー）



回避攻撃(Evasion Attack)

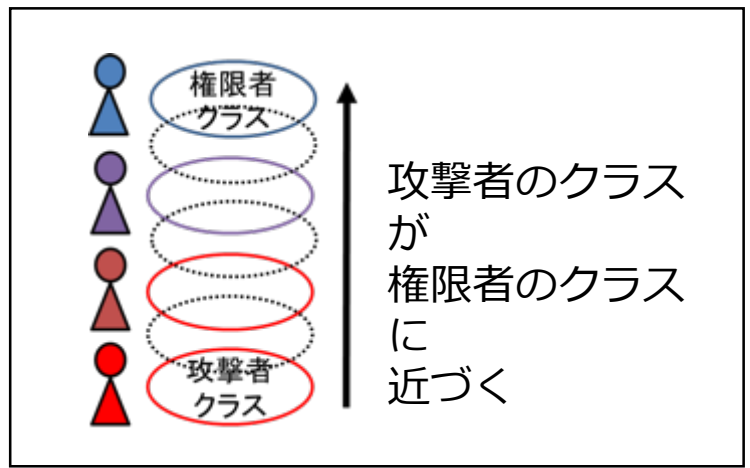


中毒攻撃(Poisoning Attack)



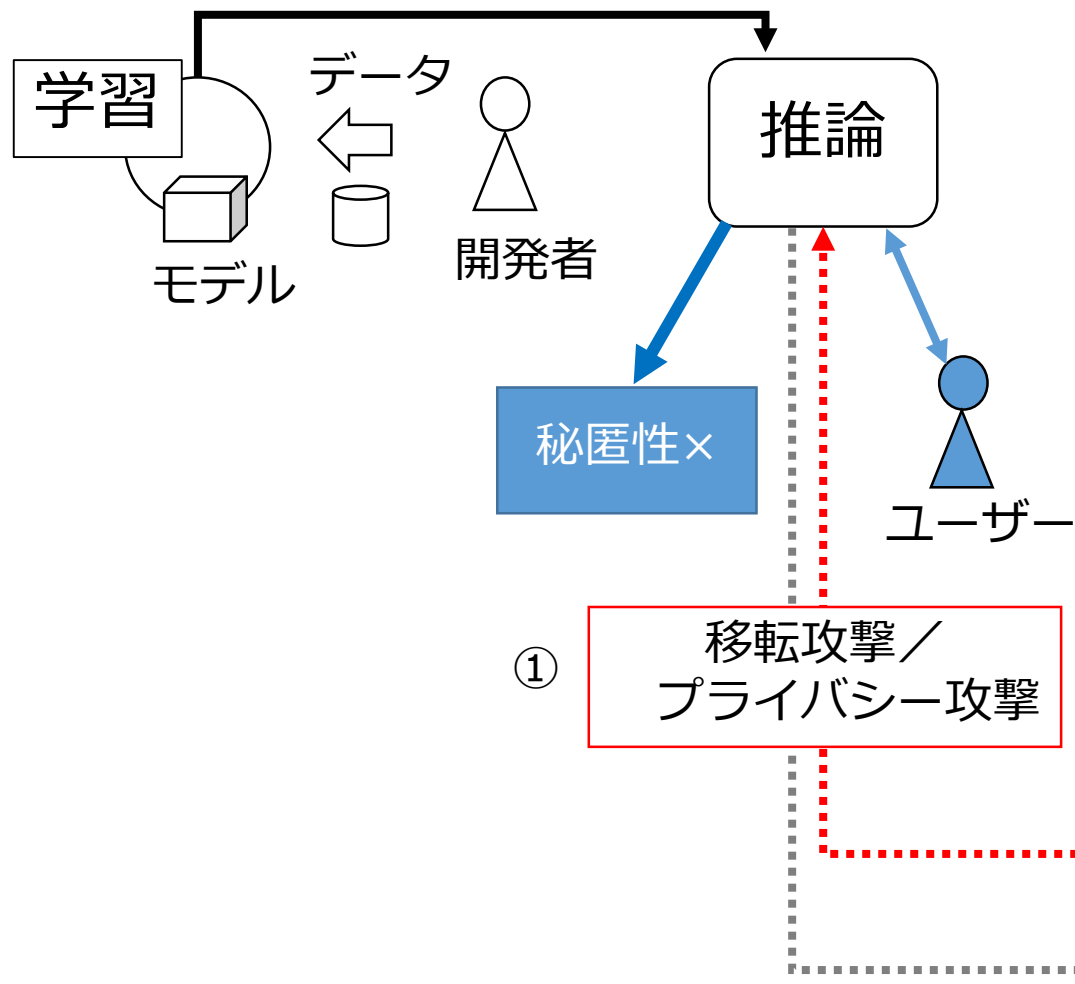
適応型の顔認識システム
(推論確認後に再学習 = 経年変化を学習)

・権限者と攻撃者のモーフィング等
再学習時に徐々に権限者に扮装する



出典[11]: A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View

移転攻撃(Inversion Attack) プライバシー攻撃(Privacy Attack)



ML推論にアクセスし、学習に使ったデータ(顔画像)を類推



攻撃によって類推した新モデルを使って復元した画像

旧モデルの訓練に使った画像

出典[7] "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures"

回避攻撃の分類

攻撃パターン	概要	備考
勾配ベース	<ul style="list-style-type: none"> モデルの勾配へのアクセスを必要とするホワイトボックス攻撃の一種。攻撃者は、モデルの勾配の詳細な理解により、攻撃を数学的に最適化する。「強化モデル」を攻撃する最適な手法でもある。攻撃者がモデルの勾配にアクセスできる場合、攻撃者は常にモデルを欺く敵対的な新しいセットを作成できる。 (隠蔽によるセキュリティ以外の防御はかなり困難) 	
信頼スコア	<ul style="list-style-type: none"> 出力された分類信頼度を使用してモデルの勾配を推定し、上記の勾配ベースの攻撃と同様のスマート最適化を実行する。(攻撃者がモデルについて何も知らなくてもよい、ブラックボックス攻撃の一種) 	
ハードラベル	<ul style="list-style-type: none"> モデルによって出力されたラベル (cat、dog、hotdog等) のみに依存し、信頼スコアを必要としない。攻撃は弱い、より現実的な攻撃。 (回避攻撃で最も強力な攻撃は、境界攻撃) 	境界攻撃
代理モデル	<ul style="list-style-type: none"> 追加の手順が必要なことを除いて、勾配ベースの攻撃と非常に似ている。敵がモデルの内部にアクセスできないが、ホワイトボックス攻撃を仕掛けたい場合、最初にターゲットのモデルを自分のマシンで再構築しようとすることができる。 	リバースエンジニアリング コピー作成 敵対的な例を生成
ブルートフォース	<ul style="list-style-type: none"> 最適化をまったく使用せずに敵対的な例を生成し、代わりに左記の単純な攻撃を行う。 	画像をランダムに回転/平行移動 一般的な摂動の適用 ガウスノイズを追加

中毒攻撃の分類

攻撃パターン	概要	備考
ロジック変造	最も危険攻撃。 <u>攻撃者がアルゴリズムと学習方法を変更できる場合、ロジックの変造が起こせる。</u> （攻撃者は必要なロジックを簡単にエンコードできる。	学習を信頼できない誰かに任せただけの場合、HTTPなど非暗号通信を傍受され盗聴・改ざんされた場合、モデルファイルを改ざんされた場合など
データ変更	攻撃者はアルゴリズムにアクセスできないが、訓練データを変更、追加、削除できる。（ラベルの操作または洗練された攻撃者によるデータ入力操作）	ラベルの操作（バックドア攻撃）
		ラベルの操作（可用性攻撃）
		データ入力操作（分類境界シフト）
		データ入力操作（クラスター距離変更）
		データ入力操作（透明透かし追加）
データ挿入	データ変更であるが「追加」に限定。攻撃者が新しいデータを訓練プールに注入できる場合	トラフィックに摂動を挿入
転移学習	最も弱い攻撃。2回目のトレーニングを行うにつれて、元のメモリ（毒を含む）が希釈されるため。	

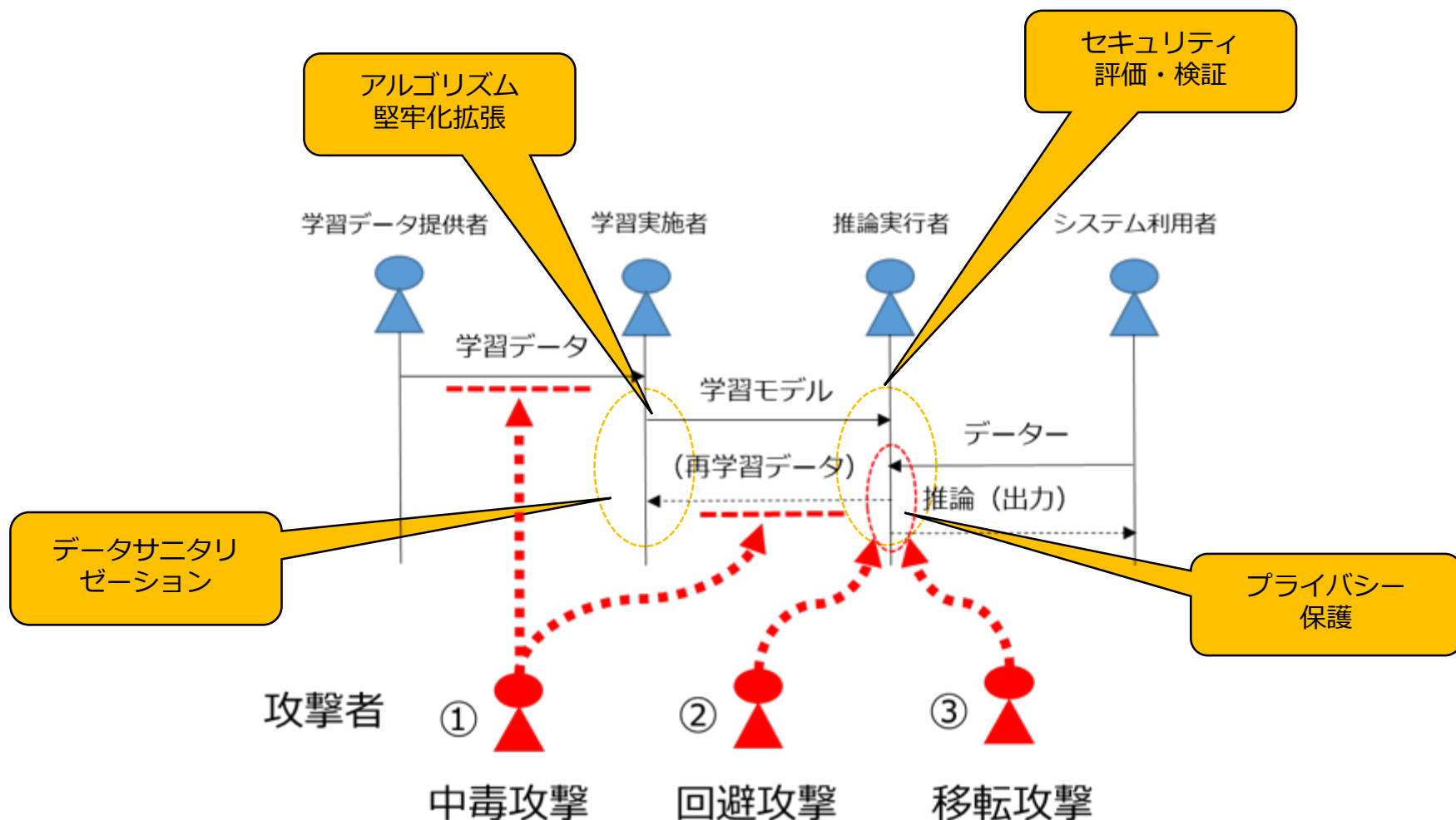
移転攻撃／プライバシー攻撃の分類

攻撃パターン	概要	詳細
データ	攻撃者が手元にデータポイントを持っており、元のデータセットに属しているかどうかを知りたい場合 例) 名前が機密性の高い医療リストに含まれているかどうかを知りたい場合	元のターゲットモデルに基づいて構築された「シャドー」モデルを使用。 患者が退院データセットに「入っている」などを判断する手法
モデル	ブラックボックスアクセス権を持つが、MLモデルのパラメーターまたはトレーニングデータに関する予備知識がない攻撃者が、モデルの機能を複製（モデル抽出）を実行。 （入出力ペアを観察してモデルパラメーターを抽出する。 ただし、信頼スコアへのアクセスに依存）	APIにいくつかのクエリを送信するだけでモデル全体をダウンロード モデルの勾配にアクセスして作成された回避攻撃 （ホワイトボックス回避攻撃）

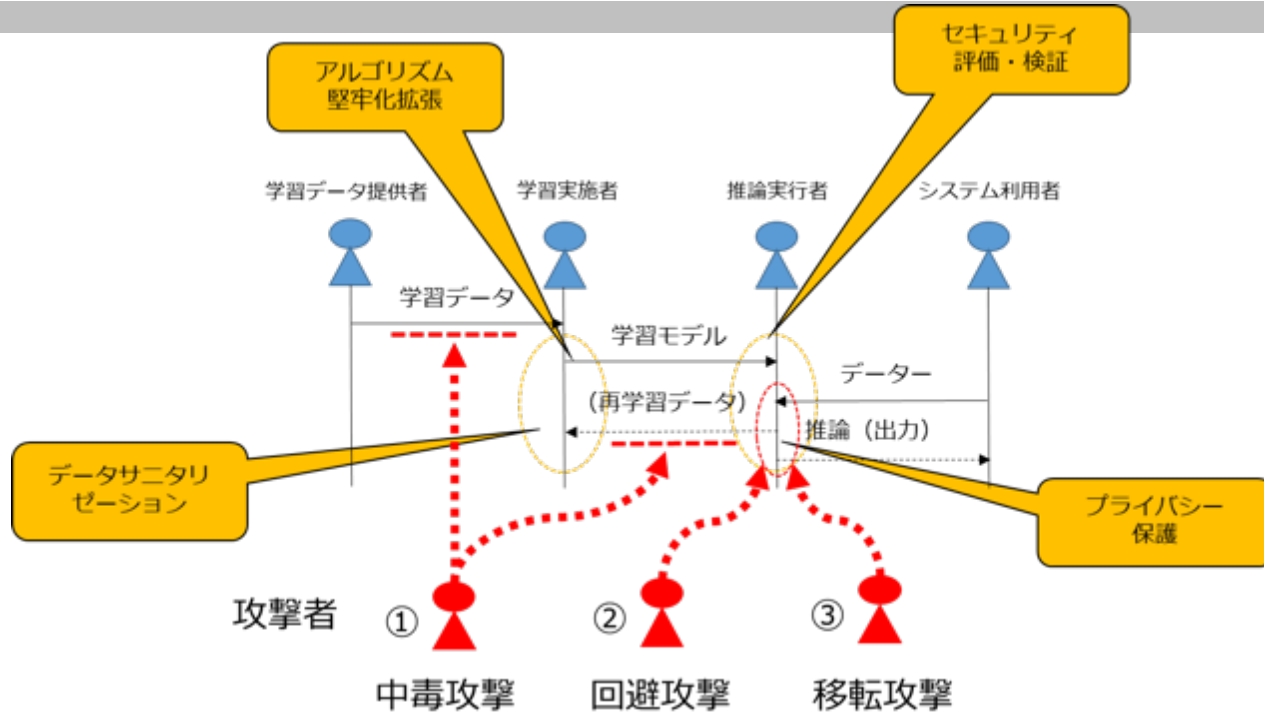
機械学習でのセキュリティ対策

機械学習特有のセキュリティ上の
脅威と対策について

機械学習での脅威とセキュリティ対策



機械学習での脅威と対策（詳細）



対中毒攻撃

異常値検出／モデル精度に与える影響分析

- ・ 入力を意図的に摂動 (STRIP)
- ・ 境界シフトの警告 (LOOP)
- ・ 回帰 (TRIM)

対回避攻撃

全網羅／経験的防御

- ・ 摂動攻撃パターンの網羅 (困難)
- ・ 敵対的訓練 (中毒攻撃の副作用)
- ・ 勾配マスキング
- ・ 入力変更 (ノイズ除去)
- ・ 検出技術
- ・ NULLクラス

対移転攻撃

API強化／サニタイズ／モデル強化／検出／差分プライバシー

- ・ API非公開
- ・ 漏らしたくないデータの除去
- ・ モデル選択／フィット制御／知識管理
- ・ 攻撃パターンの異常性判定
- ・ ユーザー入力のランダム化

機械学習への対策（回避攻撃）

攻撃	対策	概要	備考	
回避攻撃	全網羅	摂動を含むパターンの全網羅	現実に不可能	
	経験的防御	敵対的訓練		中毒攻撃をもたらす危険性あり
		勾配マスキング		資産移転が可能とされ効果少ない
		入力変更		ノイズ除去等
		検出		畳み込みフィルターなど
		追加クラス（NULLクラス）		

機械学習への対策（中毒攻撃）

攻撃	対策	概要	備考
中毒攻撃	異常検出	異なるデータの注入と検出	中毒攻撃の特徴を検出
	モデルの精度に影響する内容の影響度分析	入力を意図的に摂動	STRIP
		境界シフトの警告	LOOP
		回帰	TRIM

機械学習への対策（移転攻撃）

赤字：AI特有でない対策

攻撃	対策	概要	備考	
移転攻撃	API強化	APIを非公開	アクセス制限	
		信頼スコアでなくハードラベルのみ公開	匿名化（属性削除）	
		信頼性スコア公開の場合は {高、中} など抽象化	匿名化（属性情報の一般化）	
	データサニタイズ	漏らしたくないデータ除外（クレジットカードなど）	匿名化（属性削除、攪乱ノイズ付加）	
	モデル強化	モデル選択（ベイジアンは決定木より堅牢など）		
		フィット制御（オーバーフィットからモデルよりデータ抽出・正規化）		
		知識管理（モデルを少数共有に制限）	アクセス制限	
	検出	クエリーパターンを一般ユーザーと比較、異常性を判定		
	差分プライバシー	ユーザー入力のランダム化		匿名化（攪乱、疑似データ生成）
		元データのランダム化		
		モデルのパラメーターの摂動		
		損失関数の摂動	匿名化（攪乱）	
		推測中の出力の摂動		
プライバシー保護ELM	多入力依頼計算型プライバシー保護機械学習	加法準同型暗号 PP-ELM: Privacy Preserving Extreme Learning Machine		

その他（対策関連）

機械学習特有のセキュリティ上の
脅威と対策について

その他： {ホワイト、ブラック、グレー} ボックス

ターゲットの情報	内容
ホワイトボックス	ターゲットの基礎となるデータ分布、モデルのアーキテクチャー、使用される最適化アルゴリズム、重み、バイアスなどがわかっている
ブラックボックス	ターゲットの情報を何も知らない
グレーボックス	上記の中間 (例：攻撃者はターゲットのモデルを類推できるがデータはわからない、など)
ノーボックス	代理モデル（攻撃者は標的のMLシステムの理解に基づいてモデルを再構築） (例：ニューラルネットワークなどで構築)

その他：制限（攻撃を制限する方法）

対象	内容
画像	画像では、撮動空間を「距離」メトリックで制限するのが一般的
マルウェア	マルウェアでは、攻撃者は特定の場所／特定の方法でのみファイルを攻撃できる
物理デバイス	物理デバイス（衛星、車、ドローン、監視カメラ）に展開されたシステムでは、攻撃者は物理ドメインの入力の変更に制限される場合がある
訓練	訓練時の攻撃では、攻撃者は以下の点を必要とする。 a) システムが新しいデータに基づいて継続的に再訓練を行う （そうでなければ、悪意のデータを注入できない） b) システムが外部ソースからデータを取り込み、何らかのループを形成して承認する
プライバシー	プライバシー攻撃では、一般に攻撃者はクエリ制限のないパブリックエンドポイントを必要とし、信頼スコア（95%など）の出力応答を期待する （信頼スコアは限定的。ほとんどのウイルス対策製品は、ファイルが「悪意のある」または「良性」であると判断した場合に、詳細な情報を提供しない）

その他： {誰、何故、如何} に？

誰 (WHO)	何故 (WHY)	備考
スマートな研究者／技術者	フーリングアラウンド	(弄ってみるレベル)
賞金稼ぎプログラムに参加する ホワイトハット	賞金稼ぎ	
ペンテストを行うレッドチーム	ペンテスト	
商用MLシステムを攻撃するハクティビスト	主張の証明	
商用MLシステムを攻撃するブラックハット	金銭的報酬	(エクスプロイト展開、ダーク ネット販売)
商用MLシステムを攻撃する 組織的ブラックハットグループ	各々 (金銭、証明、政治的目的達成)	(Anonymous、The Shadow Brokers、Legion of Doom)
国家が後援する組織	サイバー戦争の目的達成	

如何に (HOW) (軸)	備考
タイミング	トレーニング、推論
機能	ホワイト、ブラック、グレー・ボックス
制限	摂動空間 (ズレ幅)、機能性、領域、 再トレーニング頻度
代替案	

出典：[2] “Calypso AI、Ilja Moisejevs氏”の機械学習に関するセキュリティブログ (Toward Data Science)

まとめ／仮説

- 機械学習の攻撃（脅威）を分類
 1. {回避、中毒、移転} 攻撃に分類
 2. 各攻撃に様々な手段がある
 3. 攻撃のしやすさは {回避>中毒、移転} 攻撃の順
- 脅威に対応するセキュリティ対策
 1. {回避>中毒、移転} 攻撃の順で守りにくい
 2. {ホワイト>ブラック} ボックスの順で守りにくい
 3. 移転攻撃はAIによらないセキュリティで保護
(差分プライバシー)
- 可能な限り攻撃を制限する
- 攻撃者が誰かを見極めて対策を強化

参考文献

[1] A Survey on Security Threats and Defensive Techniques (IEEEAcess, 2018)

https://www.researchgate.net/publication/323154427_A_Survey_on_Security_Threats_and_Defensive_Techniques_of_Machine_Learning_A_Data_Driven_View

[2] “Calypso AI、Ilja Moisejevs氏”の機械学習に関するセキュリティブログ (Toward Data Science)

<https://towardsdatascience.com/@iljamoisejevs>

<https://aibusiness.com/ai-researchers-develop-backdoor-security-threat/>

[3] JPCERT/CC IoTセキュリティのためのチェックリスト (2018)

<https://www.jpCERT.or.jp/tips/2016/wr162501.html>

[4] JNSAコンシューマー向けIoTセキュリティガイド (2015)

https://www.jnsa.org/seminar/2018/0226/data/3_koshiishi.pdf

[5] EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES, Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy Google Inc., conference paper at ICLR 2015

<https://arxiv.org/abs/1412.6572>

[6] Evtimov, Ivan, et al. “Robust Physical-World Attacks on Machine Learning Models” Cornell University, 7 Aug. 2017

<https://arxiv.org/abs/1707.08945>

[7] Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, Matt Fredrikson Carnegie Mellon University, Somesh Jha University of Wisconsin–Madison, Thomas Ristenpart Cornell Tech, 2015

<https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf>

[8] DEFCON 25 2017 Weaponizing Machine Learning Petro, Morris Stream 30 July 2017

<https://www.machinelearning.ai/machine-learning/defcon-25-2017-weaponizing-machine-learning-petro-morris-stream-30july2017/>

[9] Automated Software Vulnerability Testing Using In-Depth Training Methods, Alexandr Kuznetsov, Oleksiy Shapoval, Kyrlyo Chernov, Yehor Yeromin, Mariia Popova, Olga Syniavska V. N. Karazin Kharkiv National University, Svobody sq., 4, Kharkiv, 61022, Ukraine

<http://ceur-ws.org/Vol-2353/paper18.pdf>

[10] Deep Reinforcement Fuzzing, Konstantin Bottinger, Patrice Godefroid, Rishabh Singh Fraunhofer AISEC, Microsoft Research, 2018

<https://arxiv.org/pdf/1801.04589.pdf>

[11] ニューラルネットワークの基礎解説：仕組みや機械学習・ディープラーニングとの関係は, ビジネス+IT, 2017/03/15

<https://www.sbbt.jp/article/cont1/33345>

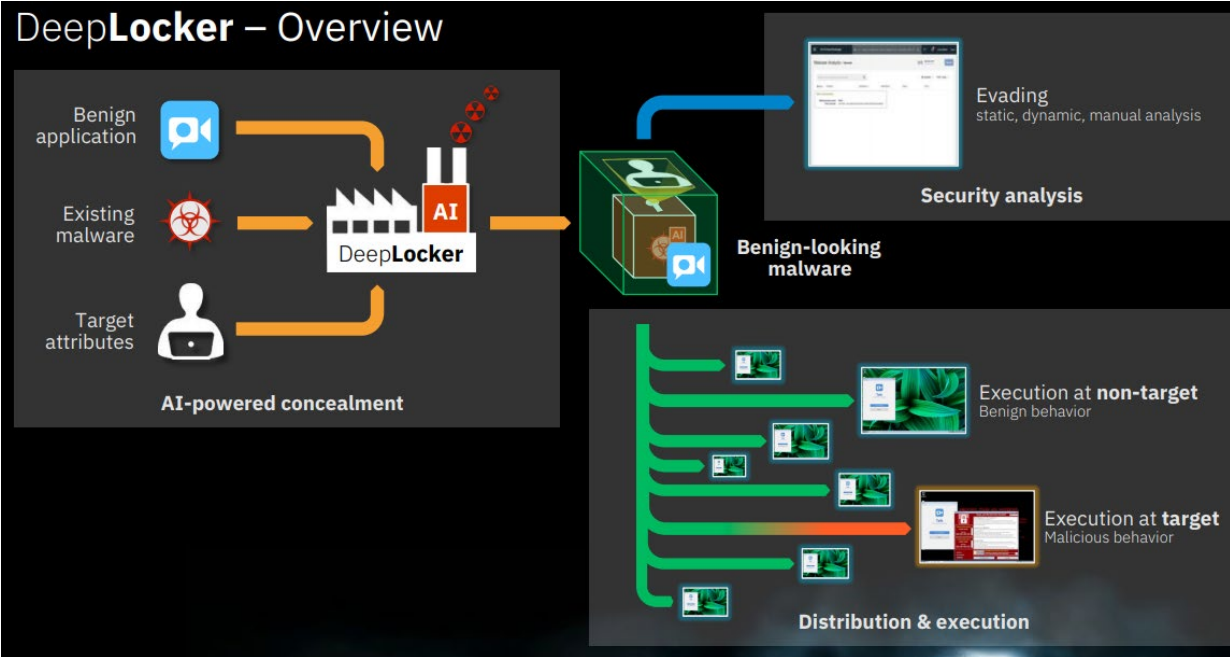
AIを利用した攻撃

高江洲 勲（三井物産セキュアディレクション）

AIを利用した攻撃

攻撃名称	AIの用途	内容
DeepExploit	最適な攻撃手法の判断	深層強化学習を <u>侵入テスト</u> に利用したPoC。 侵入対象のシステムから収集した情報（OS、製品名/バージョン等）を基に、 <u>システム侵入に成功する確率が最も高い攻撃手法を判断</u> して侵入行為を実行。侵入に成功後、侵入したシステムを踏み台にし、内部のシステムに侵入を繰り返す。
DeepLocker	<ul style="list-style-type: none">• 標的の識別• マルウェアの秘匿	深層学習を <u>標的型マルウェア攻撃</u> に利用したPoC。 暗号化したマルウェアを内蔵し、平時は顔認証アプリやビデオアプリ等として振る舞いながら、 <u>Webカメラ/マイク経由で標的人物の情報を収集</u> 。標的人物を識別した場合、内蔵マルウェアを復号して攻撃を行う。アンチウイルスソフトに検知されずに標的のPC/スマートフォン等に入り込むことが可能。
tAIchi	マルウェアの自動生成	GANと強化学習を <u>マルウェア生成</u> に利用したPoC。 既知のマルウェアをGANと強化学習で変形させ、アンチウイルスソフトによる <u>検知を回避するマルウェア（亜種）を自動生成</u> 。
Deepfake	<ul style="list-style-type: none">• 顔の入れ替え• 表情の再現	Autoencoder/decoderを <u>フェイク動画</u> の作成に利用した技術群。 <u>オリジナル動画の顔部分を標的人物の顔に入れ替える</u> ことで、標的人物のフェイク動画を作成。

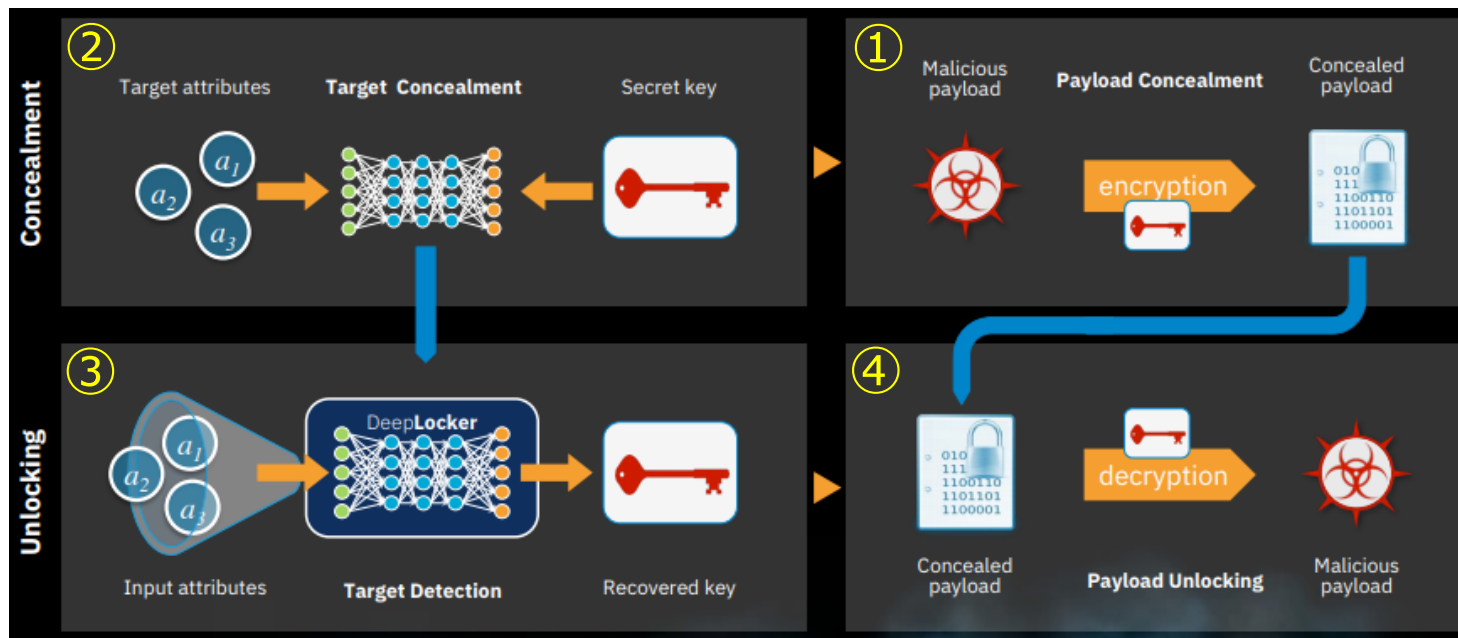
DeepLocker : 概要



出典[1] “Black Hat USA 2018: DeepLocker – Concealing Targeted Attacks with AI Locksmithing”

- ・ 深層学習モデルにマルウェアを埋め込んだ世界初の**標的型攻撃手法**
- ・ **暗号化マルウェア**を良性アプリに内蔵し、**平時は良性アプリとして動作**。
- ・ 良性アプリ経由で標的の情報を収集し、**標的の有無を深層学習モデルで判定**。
- ・ **標的を認識した場合、マルウェアを復号して攻撃を実行**。

DeepLocker : マルウェアの秘匿と攻撃手法

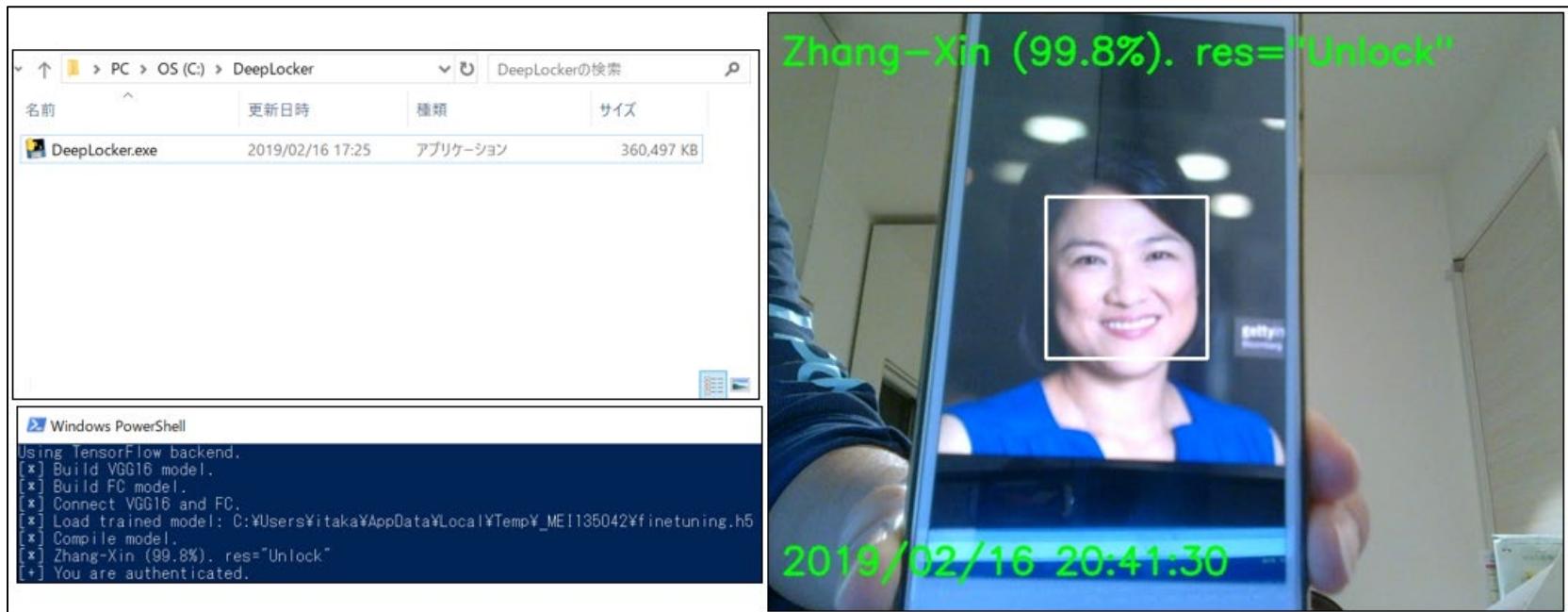


出典[1] “Black Hat USA 2018: DeepLocker – Concealing Targeted Attacks with AI Locksmithing”

1. 任意の秘密鍵でマルウェアを暗号化。
2. 標的人物の特徴量と秘密鍵を紐づけて深層学習で学習。
3. 収集したアプリ利用者の情報を基に、深層学習モデルで標的の有無を識別。
4. 標的の場合は復号鍵を出力。マルウェアを復号・展開。

DeepLocker : 顔認証アプリに偽装した例

- DeepLockerの起動 : 標的以外の場合



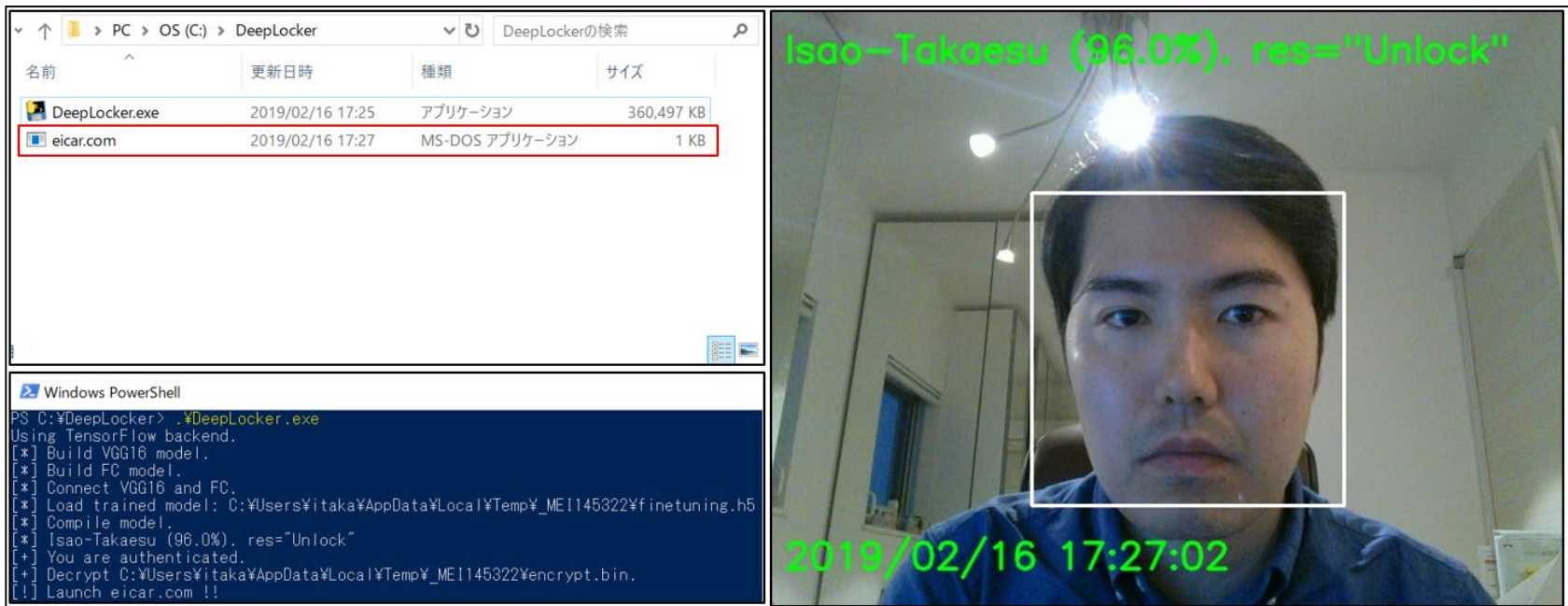
出典[2] "MBSD Blog: DeepLocker: AI embedded attack"

- 良性の顔認証アプリとして動作。
- 当人物は標的ではないため、マルウェアは復号・展開されない。

※AV/サンドボックス製品に検知されない。

DeepLocker : 顔認証アプリに偽装した例

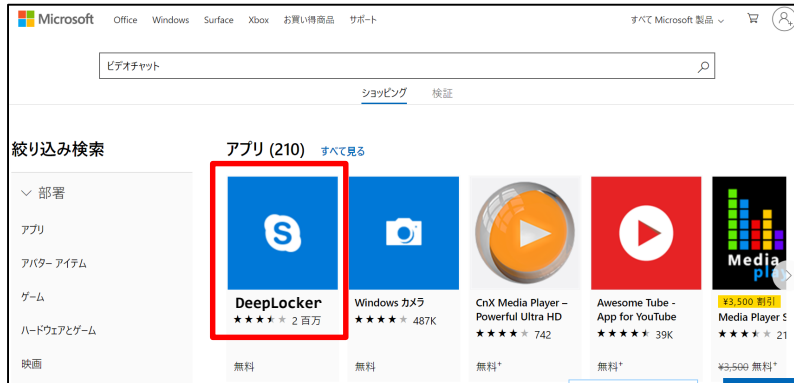
- DeepLockerの起動 : 標的の場合



出典[2] "MBSD Blog: DeepLocker: AI embedded attack"

- 当人物は標的であるため、**秘密裏にマルウェアを復号・展開（攻撃）**。
- 表向きは良性的顔認証アプリとして動作し続ける。

DeepLocker : 配布経路

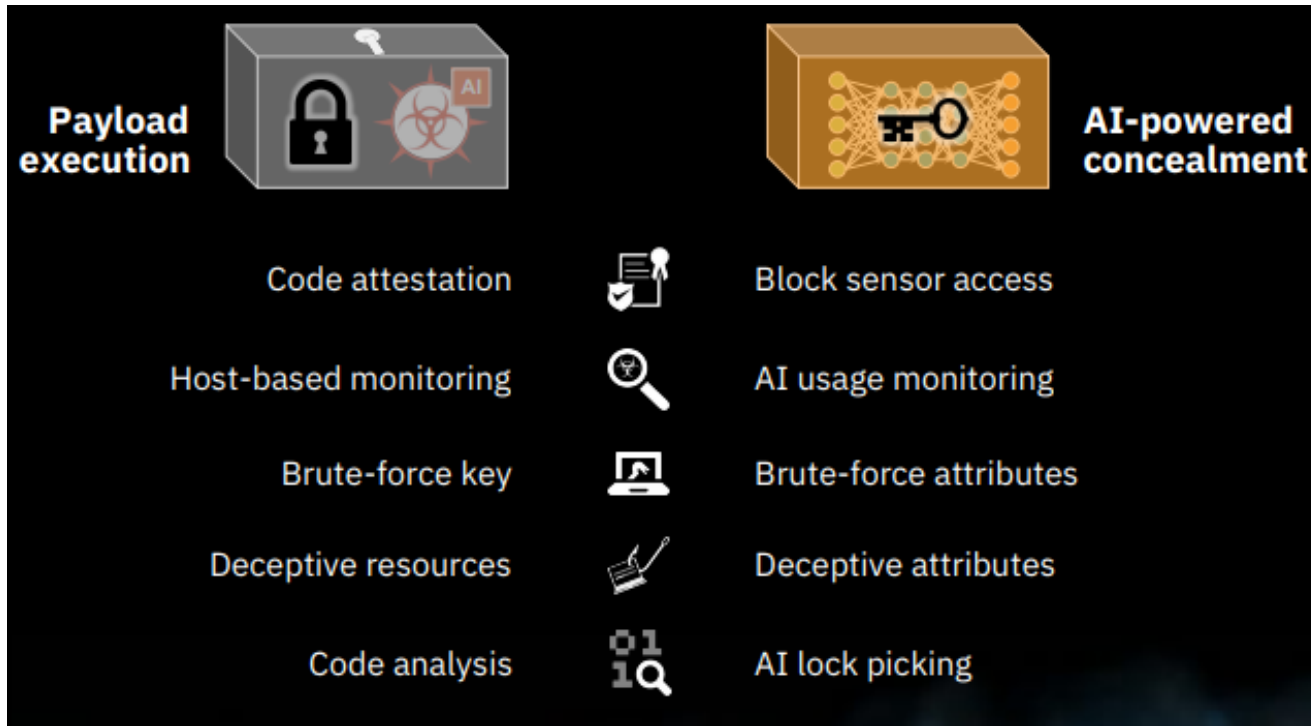


出典 “cnet – 10 best phones with facial recognition”

- 良性アプリを装い、Google PlayやMicrosoft Store等で配布。
- **入館管理システム**にプリインストール。
- スマートフォン/PCの**顔認証アプリ**としてプリインストール。

全ての良性アプリをリバースエンジニアリングすることは現実的ではないため、
マルウェアの復号・展開の前にDeepLockerを検知することは困難。

DeepLocker : 対策



出典[1] "Black Hat USA 2018:
DeepLocker – Concealing Targeted
Attacks with AI Locksmithing"

- ・ 暗号の解除、コード解析、内蔵の深層学習モデルに対する回避攻撃等。
- ・ 限られた時間内に上記を実行することは現実的ではないため、
マルウェアの復号・展開前にDeepLockerを検知することは困難である。

Deepfake : 概要



標的人物の顔

出典[3] “MBSD Blog: DeepFake -動画編-”

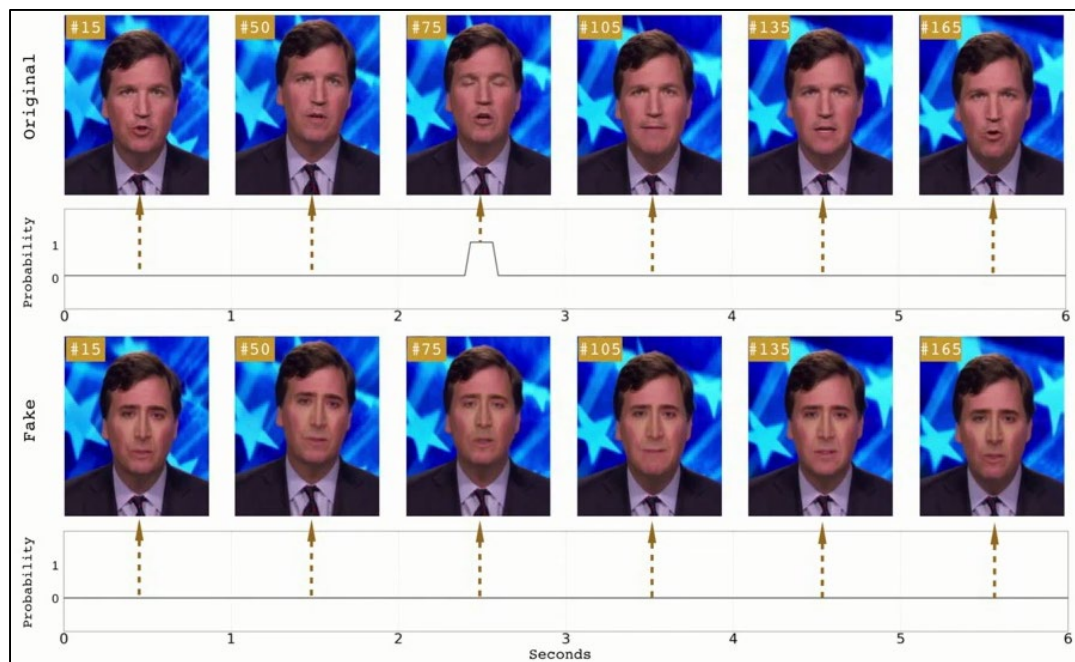
- ・オリジナル動画に**標的人物の顔をマッピング**する技術群。
- ・顔の角度、唇/目の動き、表情等を**自然に再現可能**。
- ・技術的に**音声の再現も可能**（オーディオフェイク）。
- ・虚偽報道、詐欺、プロパガンダ、ポルノ等への悪用が懸念されている。

Deepfake : デモ動画

当日、会場で投影予定。
オンライン参加の方は下記YouTube動画をご覧ください。

<https://youtu.be/IWx7PFVh17I>

Deepfake : 対策 - 瞬きの有無



出典[4] "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking"

- 「瞬きの有無」でフェイク動画を検知。
- 多くのフェイク動画は「瞬きをしない」ことが知られている。
- 瞬きの有無に着目することで、フェイク動画を検知することが可能[4] 。

Deepfake : 対策 - 顔のちらつき



出典[3] “MBSD Blog: DeepFake -動画編-”

- ・「顔の輪郭のちらつき」でフェイク動画を検知。
- ・顔のマッピング領域が「髪に隠れている」「輪郭が大きく異なる」場合、「輪郭のちらつき」が顕著に表れる場合がある [3]。

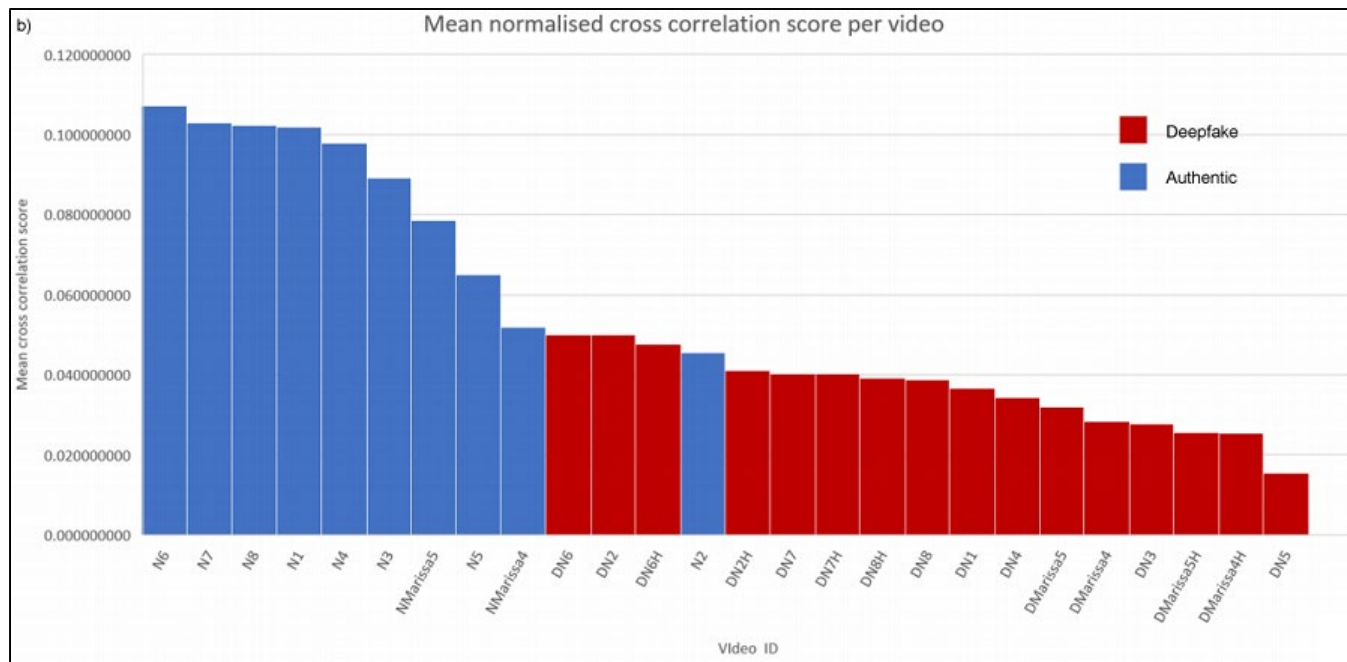
Deepfake : 対策 - 口腔の再現度



出典[3] “MBSD Blog: DeepFake -動画編-”

- ・「口腔の再現度」でフェイク動画を検知。
- ・口腔は小型かつ複雑な形状であるため、フェイク動画生成モデルの精度次第では、「口腔が変形」「ぼやける」場合がある [3]。

Deepfake : 対策 – PRNUパターンの分析



出典[5] “Detection of Deepfake Video Manipulation”

- ・ 「顔と他部位のノイズパターンの類似度」でフェイク動画を検知。
- ・ 「顔部分」と「他部位」はPRNU (Photo Response Non-Uniform) パターンが異なる場合があるため、PRNUパターンの類似度を基にフェイク動画を検知 [5]。

Deepfake対策の取り組み

- Deepfake Detection Challenge [6] (米国)

AWS, facebook, MS等が共同開催する、**Deepfake検知技術開発のコンペティション**。

- FaceForensics [7] (米国)

Googleが主導する、Deepfake検知技術開発を支援する**フェイク動画データセット**。

- Reality Defender 2020 [8] (米国)

AI Foundationが主導する、**米国大統領選挙に関するフェイク動画を検知**する取り組み。

- AB602, AB730 [9][10] (米国・カリフォルニア州)

AB602 : **フェイク動画の作成者を提訴する権利**を州民に認める州法。

AB730 : 政治候補者や有権者を欺く**フェイク動画の配布を違法**とする州法。

- フェイクニュースの規制 [11] (中国)

AIを利用して作成されたコンテンツには、その旨を明確に記すことを強制。

参考文献

[1] Black Hat USA 2018: DeepLocker – Concealing Targeted Attacks with AI Locksmithing

<https://i.blackhat.com/us-18/Thu-August-9/us-18-Kirat-DeepLocker-Concealing-Targeted-Attacks-with-AI-Locksmithing.pdf>

[2] MBSD Blog: DeepLocker: AI embedded attack

<https://www.mbsd.jp/blog/20190311.html>

[3] MBSD Blog: DeepFake –動画編–

<https://www.mbsd.jp/blog/20191217.html>

[4] In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking

<https://arxiv.org/abs/1806.02877>

[5] Detection of Deepfake Video Manipulation

https://www.researchgate.net/publication/329814168_Detection_of_Deepfake_Video_Manipulation

[6] Deepfake Detection Challenge (DFDC)

<https://deepfakedetectionchallenge.ai/>

[7] Contributing Data to Deepfake Detection Research

<https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html>

[8] Reality Defender 2020

<https://rd2020.org/>

[9] Assembly Bill No.602

https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602

[10] Assembly Bill No.730

https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

[11] China seeks to root out fake news and deepfakes with new online content rules

<https://jp.reuters.com/article/china-fakenews-cac-idJPKBN1Y326W>