IoT 時代における AI とセキュリティに関する統合的研究の構想 Beyond Attackers を目指して

佐々木良一[1]　金子朋子[2]　　吉岡信和[2]

概要：セキュリティの研究は残念ながら常に攻撃者の後追いであった．近年，AI（Artificial Intelligence）を用いたセキュリティ対策の研究も増えているが大部分が後追いであるという傾向は変わらない．その問題を解決し，Beyond attackers をより多く実現するため本研究ではまず AI とセキュリティに関する研究を分析し，（ａ）Attack using AI（AI を利用した攻撃），（ｂ）Attack by AI（AI 自身による攻撃），（ｃ）Attack to AI（AI への攻撃），（ｄ）Measure using AI（AI を利用したセキュリティ対策）に分類できることを明確にし，それぞれの研究の現状と課題を整理した．次に（ｄ）Measure using AI と他の観点からのアプローチを組み合わせることにより，事前予測を含む種々のサイバー攻撃に対応でき，AI への攻撃や AI による攻撃に強いセキュリティ対策用 AI システムの構築が容易になり， Beyond Attackers の可能性が高まる見通しを得た．併せて，（ｂ）Attack by AI（AI 自身による攻撃）と（ｄ）Measure using AI（AI を利用したセキュリティ対策）の組み合わせに限定して，将来出現しうる攻撃のリストアップ法と対策抽出方法を示すとともに AI-embedded attack に例を取り適用を開始した．

キーワード：セキュリティ，AI，機械学習，統合的研究，Beyond attackers

Consideration on an Integrated Research for AI and Security in the IoT Era　- Aiming “Beyond Attackers”-

RYOICHI SASAKI[1]　TOMOKO KANEKO[2]　NOBUKAZU YOSHIOKA[2]

Abstract: Unfortunately, security research has always followed attackers. In recent years, research on security measures using AI (Artificial Intelligence) has increased, but the tendency that most of them are behind attackers remains unchanged. In order to solve the problem and to realize more "Beyond attackers", in this research, we first analyze the research on AI and security, and classified them into four categories:　(a) Attack using AI , (b) Attack by AI, (c) Attack to AI, and (d) Measure using AI. In addition, the status and issues of each research were summarized. Next, by combining (d) Measure using AI and approaches from other perspectives, we make it possible to deal with various cyber attacks including pre-prediction, and construct an AI system for security measures that is strong against attacks to AI and attacks by AI. At the same time, we examined the combination of (b) Attack by AI and (d) Measure using AI, showed the method of listing the attacks that may appear in the future and the selecting method of measures, and started application research on the AI-embedded attack.

Keywords: Security, AI, Machine Learning, Integrated Research, Beyond Attackers

1. はじめに

IoT（Internet of Things）時代を迎え，機械学習（Machine Learning）を中心とする人工知能（Artificial Intelligence：以下 AI）の研究が各分野で再度注目を浴びている．一方，IoT 時代にはセキュリティの確保は従来以上に重要な課題となっている．残念ながらセキュリティの研究は，攻撃が実現してから，それをどう解決するかを検討する，常に，「後追い研究」であった．そしてセキュリティ対策に AI を導入してもその傾向は変わらなかった．

そこで，AI とセキュリティの関係を整理し，それらをうまく組み合わせることにより，「後追い研究」の問題を解決できないかと考えた．そのため，最初に AI とセキュリティの両方に関連する研究には次の４種類があることを明確にした．

（ａ）Attack using AI（AI を利用した攻撃）
（ｂ）Attack by AI（AI 自身による攻撃）
（ｃ）Attack to AI（AI への攻撃）
（ｄ）Measure using AI（AI を利用したセキュリティ対策）

次に，それぞれについて研究の概況と課題を明確にした．そして，これらをうまく組み合わせる統合的な研究も必要であり，（ｄ）Measure using AI と他の 3 種の研究を組み合わせることにより攻撃者を先回りして対策を行う「Beyond attackers」がより多く実現できる見通しを得た．併せてそのための具体的方法を整理するとともに，新しい攻撃方法の類推と対策方法の例を示した．

1 東京電機大学
Tokyo Denki University
2　国立情報学研究所
National Institute of Informatics

2 節で 4 種の研究の分類法とそれぞれの研究の概要と課題を示す．3 節でこれらを組み合わせた統合アプローチ案の構想を記述したうえで，4 節で具体的な統合的アプローチ法の一例を示す．

AI に関する研究もセキュリティに関する研究も非常に多く，これらを組み合わせた研究も増えてきている．Google Scholar で「Cyber Security AND Artificial Intelligence」をキーワードとして検索すると 2007 年には 1870 件だったものが 2019 年には 17,000 件と約 9 倍に増加している．しかし，本稿で扱うような統合的な研究にターゲットを合わせた研究は見当たらない．

2. AI とセキュリティに関する 4 つの観点と研究課題

2.1 Attack using AI

今後，不正者による AI を利用したサイバー攻撃は増加してくると考えられる．基本的なものは，人間がやっていた攻撃を，AI を用いて自動化するというものだろう．その 1 例は次のとおりである（図 1 参照）．

① 最近，ボットを利用して，コンサートなどのチケットの買い占めが試みられている．

② ボットによるアクセスの防止のために，コンピュータが判読を苦手とする画像認証などを利用している．

③ しかし，AI を利用して画像認証の解読確率を向上させたようであり，2018 年 8 月のイープラスサイトでのケースではチケット購入のアクセスのうち 9 割超が bot だったという．

特に，AI 機能付きのマルウェアは近い将来，確実に誕生するだろう．今後は小さな種々の AI 機能付きマルウェアが侵入し，協力しながら環境に最も適した攻撃をするようになっていくのではないかと考えている．少なくとも研究レベルでは，このような動きを考慮して今後の対策を考えておくことが大切となるだろう．

最近では単純な自動化ではなく，AI を使用し，セキュリティ対策を回避して標的をピンポイントで攻撃する方法も提案されている．その 1 例が，2018 年 8 月に開催された Black Hat USA において，IBM Research の研究者らが開発した，DeepLocker であろう[1][2]．DeepLocker は「AI の仕組みそのものに攻撃を埋め込む「AI-embedded attack」というアプローチを採用しており，「処理過程がブラックボックス化されているために動作解析が困難」という特性を利用することで，セキュリティ対策製品による検知を回避する高い隠匿性能を獲得できる[2]．これを発展させたいろいろな攻撃法が最近検討されている．

図1 画像認証画面の一例

2.2 Attack by AI

AI が人間に及ぼす悪影響で最も大きな問題は人間を上回る能力を有する AI が誕生し将来的に人間が絶滅させられるではないかということである．

Google の研究者のレイ・カールワイツは 2045 年には AI の能力が，人間を超越するシンギュラリティが生じ，反乱すら起きるかもしれないとしている[3]．また，スティーブン・ホーキングは「人工知能の発明は人類史上最大の出来事だった．だが同時に『最後』の出来事になってしまう可能性もある」といっている[3]．また，「2001 年宇宙の旅」や，「ターミネータ」など映画の世界では，AI の反乱が数多く描かれており，人々の関心が高いことがわかる．

一方，日本の研究者は次のような理由で AI の反乱といったようなことは起こらないという意見が強い．

（1）「強い AI（汎用 AI）」ではなく「弱い AI（専用 AI）」の研究が中心であり，弱い AI が汎用的な能力を発揮し，さらに高度な AI を自動的に作ることは困難である．

（2）そのようなことが起こる可能性があるとしても，AI システムに，例えば＜ロボット 3 原則＞のような制約を与えることにより反乱を抑えることができる．

西垣通氏は「西洋で人工知能の反乱を恐れるのは，一神教の影響で神に代わって創造主になることに対する恐れでないか．」といっている[4]．

私自身は次のように考えている．

（1） AI が反乱を起こす可能性は極めて低い．

（2） しかし，原子力プラントへの津波来襲に伴う事故にみられるように，人間のリスクに対する知覚能力は極めて低い．

（3） また，反乱がおきると取り返しのつかないことになっている可能性が強い．

（4） したがって，動きを慎重に見守っていくことが大切である．

例えば＜ロボット 3 原則＞のような反乱防止のメタ制約を与える一方，AI が自由に進化をするとした場合，本当に

反乱を抑えることができるかどうかをシミュレーションなどで確認する研究なども進めておく必要があると考えている．

2.3 Attack to AI

AI システムへの攻撃には次のようなものがあると考えられる（図２参照：[5]を参考に作成）．

① 機械学習システムの停止やファイル情報，通信路情報の盗み出しなどの攻撃：これは従来のシステムへの攻撃と基本的に同じなので，ここでは対象外とする．

② 訓練済みモデルの誤分類を誘発するノイズ付加攻撃：判定・予測用データにノイズ等が加えられると，判定・予測の精度が低下し，誤判定等が誘発されうる．例えば，動物名を判定するシステムに対し，パンダの画像に微細 なノイズを加えることにより，人間が見ればパンダだが，テナガザルと誤判断させるような攻撃が知られている．

③ 機械学習に対する偏った訓練データを意図的に与えるなどが原因で，不適切な判断をさせてしまう攻撃：米マイクロソフトのチャットボット「Tay」は，クラウドソーシングを利用して学習させた．ところが悪意を持ったユーザが協力して差別的な意見を繰り返し入力，Tay は差別発言を繰り返すようになってしまった．

④ 学習モデルへデータを入出力することにより情報を漏洩させる攻撃：機械学習システムでは，判定・予測エンジンの 入出力から，訓練データにかかる情報が漏洩する可能性がある．例えば，氏名等の個人を識別する情報と顔画像を訓練データとして使用した 顔画像認識システムにおいて，判定・予測エンジ ンの入出力から，訓練データとして用いられた特定の個人の顔画像を高い確率で推定する研究事例が知られている．

これらは，重要な課題であり，今いろいろな研究がおこなわれている分野であるが，②③はさらに研究を加速すべきであると考えられる．

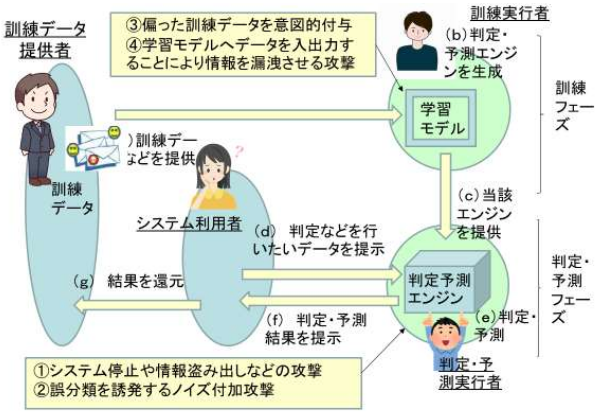


図2 機械学習の利用形態と攻撃方法の概要

2.4 Measure using AI

2.4.1 方式の概要

セキュリティ対策に AI を用いるアプローチである．Google Scholar を用いて調査すると非常に多くの論文が出てくる．

これらの論文の調査やＷＥＢ上の製品紹介から，次のようなセキュリティ対策のためにすでに機械学習を中心とする AI が使われていることが分かった．

・「マルウェアの検出」
・「ログの監視・解析」
・「継続的な認証」
・「トラフィックの監視・解析」
・「セキュリティ診断」
・「スパムの検知」
・「情報流出」など，

AI を使ったというセキュリティ対策ツールは各社から発売されており，そのメリットがＷＥＢ上で述べられている．ただ実際のフィールドでどの程度有効かについてはよくわからないものも多い．

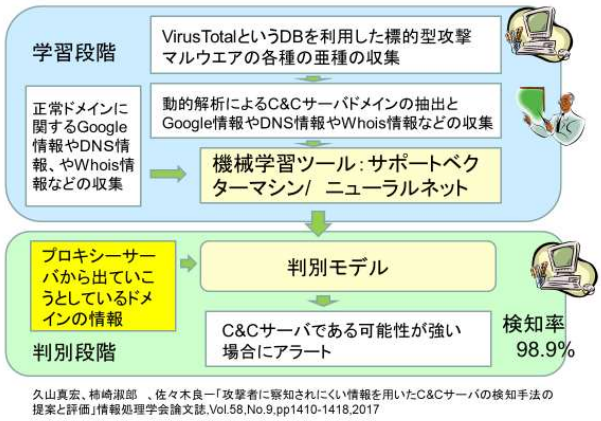
いずれにしても，AI を利用したセキュリティ対策の研究は今後ますます重要になっていくと考えられる．

2.4.2 著者らの研究

セキュリティ対策に対し AI を用いる研究として著者らは次のような研究を実施している．

（１）機械学習を利用した標的型攻撃用 C&C サーバの自動判別システム（図３参照）

近年流行している標的型攻撃では，マルウェアに感染した後に C&C サーバとの間で様々な通信を行う．そのため，出口対策として C&C サーバのブラックリストを用い，ブラックリストに載ったサーバとの通信を制限することにより，被害の発生を防止することが出来る．しかし，ブラックリストは常に古い C&C 情報しか持たず，新しく C&C にされてしまった C&C サーバにアクセスしてしまう可能性がある．そこで，C&C サーバと通常のサーバの DNS 情報や WHOIS 情報などを調べ，ニューラルネットワークなどのツールを使い判別モデルを作成した．そして，この手法に実データを適用し C&C サーバの判別を行った結果，約 99.3%と高い検知率を得ることができ，有効性の見通しを得ることができた．また，攻撃者に察知されにくい情報だけを用いても 98.9%の検知率を上げられることが明らかになった．ここで C&C サーバに関する情報は Virous Total を用いて求めた．時間経過が具体的にどのように影響を与えるかについては今後の課題である．詳しくは，文献[５][６]などを参照願いたい．



久山真宏、柿崎淑郎 、佐々木良一「攻撃者に察知されにくい情報を用いたC&Cサーバの検知手法の提案と評価」情報処理学会論文誌,Vol.58,No.9,pp1410-1418,2017

図3 標的型攻撃用C&Cサーバの自動判別システム

（2）ルールベースシステムやベイジアンネットワークを利用した知的ネットワークフォレンジックシステム

近年，企業や政府機関を対象としたサイバー攻撃の数が増加しているのはご存じのとおりである．このような組織では，標的とされた攻撃に対する対策を準備する必要があるが，支援システムの助けなしに攻撃中にこれらの手段を実行することは非常に困難である．そこで，執筆者らは，ルールベースシステムやベイジアンネットワークなどの人工知能技術を用いて攻撃事象を推定し，攻撃対策を導くための LIFT（Live and Intelligent Network Forensic Technologies）システムを開発した（図４参照）．このシステムは，収集されたログを分析し，攻撃の手がかり（兆候）を検出し，次にベイジアンネットワークを使用して，検出された手がかりから各攻撃事象の可能性を推定する．確信度が十分に大きい場合，その攻撃事象が発生していると考える．確信度が小さい場合は，LIFT システムはログからの追加の手がかりの収集を行ったうえで同様の処理を確信度が十分大きくなるまで繰り返す．さらに，LIFT システムは，攻撃事象からルールベースシステムにより対策項目を選定し，対策の実施をガイドしたり，自動操作を実行したりする．著者らは，LIFT システムのプロトタイプを開発し，このプロトタイプを過去に発生した攻撃シーケンスに適用した．その結果，LIFT は，事象を推定したり対策を正しく指示したりという目的とする機能を実現することが確認できた．ここで，機械学習ではなくルールベースシステムやベイジアンネットワークを利用したのは，異常事象に関するデータが十分得られないからである．この LIFT システムについて詳しくは，文献[8]を参照願いたい．

なお，AI を応用して事象や応急対策を明確にした後，侵入元や，侵入範囲を推定する必要があるが，ここでは機械学習などの AI を使わず，各機器のプロセスログとパケットログを Onmitsu[9]という執筆者らが開発したツールを用いて求め，それに機器間の接続情報を加えたうえで，検知アルゴリズムに基づき侵入元や，侵入範囲を推定する方法を開発している．アルゴリズムを用いる方法にしたのはこの方が，アルゴリズムが正しいならば見落としがなく，推定した理由の説明が容易であり，そのあと具体的対策を実施するのに適していると考えたからである．このように AI を用いる方法と，アルゴリズムを用いる方法を組み合わせて利用するというアプローチも有効であると考えられる．

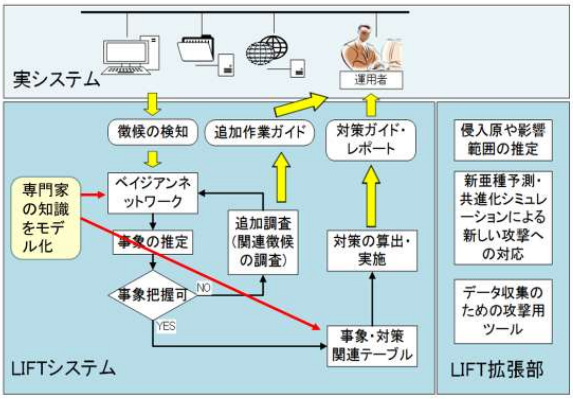


図4 LIFTシステムの概要

2.4.3 セキュリティ対策への AI 応用の特徴

セキュリティ対策のために AI を実際に利用するにあたっては，以下のような問題を解決することが望ましい．

（1） AI システム特に機械学習システムは適切に分類された大規模なデータセットを得ることが望ましいが，この分野でこのようなデータセットを入手するのは一般に困難である．特に，サイバー攻撃は時間とともに特性が変化することが多く，それぞれの期間における大量のデータの入手が必要となるが困難なことが多い．

（2） 機械学習システムはいくつかのケースの誤検出を代償にして精度を高めることができる．しかし，ソフトウェアの世界では，善良なアプリケーションをいくつか誤ってブロックするアンチウイルスを許そうとせず，1%より低い誤検出率を要求することが多い．したがって，誤検知率が十分小さくならない場合は，誤検知があっても影響を十分小さくできるような仕組みと組み合わせる必要がある．

（3） セキュリティ分野では結果が「説明可能」であることが望ましいが，ニューラルネットワークや深層学習などの高度な機械学習アルゴリズムは，人間が読むことのできる言葉で説明するのが困難な場合が多い．

これらをどう解決するかがセキュリティ対策に AI を適用する上で重要な課題となる．

３．統合的研究アプローチに関する考察

以上，4 つの観点からそれぞれの研究状況と残された課題などを示してきた．これらの研究の深化はそれぞれ重要であるが，Beyond attackers を可能とするような大きな成果に結びつけるのは困難である．図５に示すように他の観点からのアプローチを，（d）Measure using AI（AI を利用したセキュリティ対策）とうまく結合するアプローチが必要

になると考えている．

（ａ）Attack using AI（AI を利用した攻撃）とうまく結合することにより，①AI を使った攻撃方法を検討し防御方法を事前実装することにより新しい攻撃が出てきても事前に対応できる可能性がある．

また，（ｂ）Attack by AI（AI 自身による攻撃）と組み合わせることにより，②　AI の反乱に備えた検知方法の実装することにより，事故や不正などにより AI が間違えた行動をとり始めた時，AI を用いたセキュリティ対策システムが事前に異常を検知できるようになる可能性がある．

（ｃ）Attack to AI（AI への攻撃）と組み合わせることにより，③　AI への攻撃に強い対策方法を実装したセキュリティ対策システムの開発が可能となる可能性がある．

これらを組み合わせることにより，事前予測を含む種々のサイバー攻撃に対応でき，AI への攻撃や AI による攻撃に強いセキュリティ対策用 AI システムの構築が容易になると考えられる．事前予測を含んでおり，完全な実現は見果てぬ夢であるが，部分的には Beyond Attackers の可能性が高まると考えられる．

これらの効果や，具体的方法を掘り下げて研究することが重要であり，今後，大切な研究課題であると考えている．なお，良い対策であるかどうかを判断するためにはサイバー攻撃・防御実験用共通基盤やリスク評価手法が必要になると考えられる．

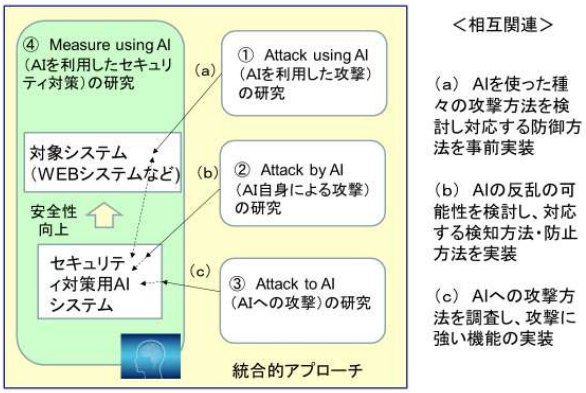


図5　AIとセキュリティに関する統合的アプローチ

4．Attack by AI をベースにした Measure using AI アプローチ

ここでは，3 つあるアプローチのうち，まず，（ａ）Attack using AI（AI を利用した攻撃）とうまく結合する方法の検討を行う．そして，①AI を使ったさまざまな攻撃方法を検討し防御方法を事前実装することにより新しい攻撃が出てきても事前に対応できるようにするための技術の研究について検討を行う．

4.1　既存のアプローチ方法

次のようなアプローチが考えられる．

（１）新しい攻撃方法のアイディアをセキュリティの専門家に対するアンケートで示してもらい，可能性の強いものを抽出したうえで，効率的対策を考案し，システムに組み込む．きちんと答えてもうための体制つくりや，不正に利用できなくする仕組みなどの検討が同時に必要となる．

（２）AI の手法の 1 つであるマルチエージェントを用いる方法である．先に述べた LIFT の研究・開発者の一人である八槇博史らは，図６に示すようなマルチエージェントを用いた共進化モデルを開発している[10]．ここでは，攻撃用のマルウェアと守備側のシステムが，それぞれ進化しつつ，お互いを淘汰していくモデルとなっている．

（３）新しいタイプの攻撃方法を取り上げ，その要素を変更することにより，類似攻撃を類推する（詳しくは 4.2 参照）．

これらを採用することにより，まだ誕生していないが今後誕生する可能性のあるマルウェアに対する対策も事前に取り込め，Beyond attackers をより多く実現することができる．いろいろな進化が考えられ，これらをさらに詳細化していくことが必要となる．

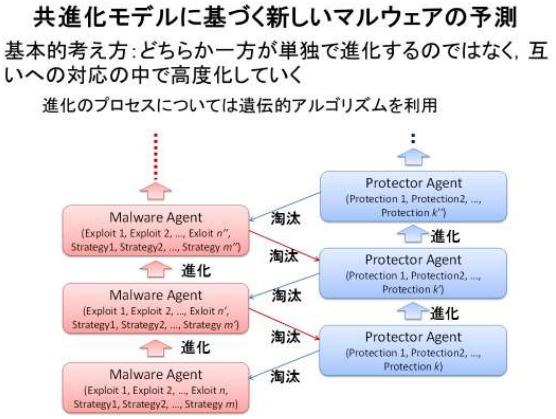


図6　新機能を持つ類似マルウエアの予測法

4.2　現在構想している対策法

DeepLocker のような AI-embedded attack 型の新しいタイプの攻撃方法が出てきたら，その基本コンセプトをうまく利用することによって，攻撃方法を拡張していき，それに対する解決方法を今後普及していくと考えられる WEB 会議システムなどに適用していくアプローチが考えられる（図 7 参照）．例えば，基本システムや基本機能は DeepLocker と同じで，不正方法だけが異なる DeepURL という方法はすでに提案されている[11]．

著者らは，DeepTrans と名付けた次のような攻撃方法を検討中である．ここでは，音声認識や自動翻訳などの AI 機能を持つ WEB 会議システム（Zoom など）において，通常状態は翻訳を行い画面上に文字を表示しているが，特別な単語（例えば中国であれば「天安門事件」）などを検知した

場合には，その会議を強制的に中止したり，ブラックリストに記録するなどの攻撃が考えられる．この際，特殊な単語の検知や，会議の強制中止の部分は暗号化などで，わからなくしておく．

これに対し，対応側としては，暗号化している部分を簡易な乱数チェックによって，利用する際にチェックしておき，暗号化部分のモジュールを処理したのち，起こっては困る処理に移ろうとする場合は，アラートを上げるようにしておくなどの対応が考えられる．

今後，このような攻撃を詳細化するとともに，コスト効果の大きな対応策の組み合わせを求めていきたい．

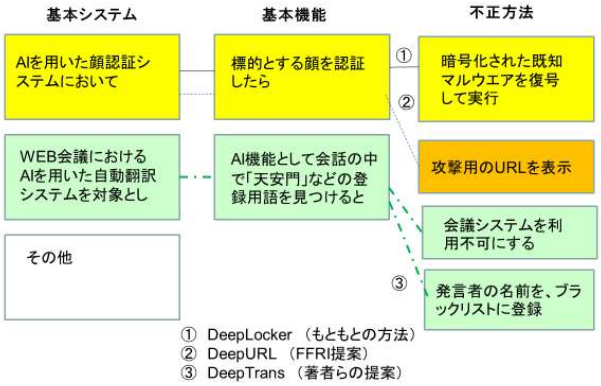


図7　類似攻撃のリストアップ法

５．おわりに

セキュリティの研究は残念ながら常に攻撃者の後追いであったという問題を解決し，Beyond attackers を一部あっても実現するため，本研究ではまず AI とセキュリティに関する研究を分析し，

（ａ）Attack using AI（AI を利用した攻撃），

（ｂ）Attack by AI（AI 自身による攻撃），

（ｃ）Attack to AI（AI への攻撃），

（ｄ）Measure using AI（AI を利用したセキュリティ対策）

の 4 種類があることを明確にし，それぞれの研究の現状と課題を整理した．

次に（ｄ）Measure using AI と他の観点からのアプローチを組み合わせることにより，従来出現していなかった攻撃方法を把握し，対策方法を考案しうる可能性を見出した．併せて，（ｂ）Attack by AI（AI 自身による攻撃）と（ｄ）Measure using AI（AI を利用したセキュリティ対策）の組み合わせに限定して，Beyond attackers を可能とする安全性の高いシステムを実現するため，従来の方法を整理するとともに，WEB 会議システムへの類似の新しい攻撃方法と，その防止策の基本アイディアを示した．

今後，このような攻撃を詳細化するとともに，コスト効果の大きな対応策の組み合わせを求めていきたい．

参考文献

[1] “DeepLocker - Concealing Targeted Attacks with AI Locksmithing”, https://www.blackhat.com/us-18/briefings/schedule/#deeplocker---concealing-targeted-attacks-with-ai-locksmithing-11549

[2] 高江洲　勲　” DeepLocker : AI-embedded attack “https://www.mbsd.jp/blog/20190311.html

[3] https://ja.wikipedia.org/wiki/人工知能

[4] [西垣通「ビッグデータと人工知能」中公新書，2016

[5] 井上紫織，宇根正志 「金融分野における機械学習システムの活用とセキュリティ対策」 2019 https://www.boj.or.jp/research/wps_rev/rev_2019/data/rev19j02.pdf

[6] MASAHIRO KUYAMA, YOSHIO KAKIZAKI, RYOICHI SASAKI “Method for detecting a malicious domain by using only well-known information” International Journal of Cyber-Security and Digital Forensics (IJCSDF) 5(4): 166-174

[7] 久山真宏，柿崎淑郎 ，佐々木良一「攻撃者に察知されにくい情報を用いた C&C サーバの検知手法の提案と評価」情報処理学会論文誌,Vol.58,No.9,pp1410-1418,2017

[8] Ryoichi Sasaki et al.“Development and Evaluation of Intelligent Network Forensic System LIFT Using Bayesian Network for Targeted Attack Detection and Prevention” International Journal of Cyber-Security and Digital Forensics (IJCSDF) 7(4): pp344-353, 2018

[9] 三村聡志，佐々木良一「プロセス情報と関連づけた通信情報保全手法の提案」情報処理学会論文誌, Vol.57,No.9,pp1944-1953

[10] 石川博也，八槇博史「サイバー空間における攻撃と防御の共進かシミュレーション」 情報処理学会 CSS2016，2016

[11] FFRI エンジニアブログ「AI-Embedded Attack についての考察」2019-09-24 https://engineers.ffri.jp/entry/2019/09/24/000000