特集
Special Feature

［AI 時代のサイバーセキュリティ］

# 5 AI を活用したシステムへの攻撃と防御に関する最新セキュリティ研究動向

応専

森　達哉　早稲田大学／理研 AIP

## 「AI を活用したシステム」とは ？

本稿が対象とする「**AI を活用したシステム**」（以下，簡単のため「**AI システム**」）は，内部で行う処理の一部に機械学習モデルを応用しているシステム全体を指す．今日，さまざまな AI システムが存在するが，その一例は自動運転車である．自動運転車はさまざまなモジュールから構成される複雑なシステムであるが，モジュールの 1 つである環境認識部では，カメラ画像あるいは LiDAR センサ（3 次元測距センサ）の読み取り値を入力とし，機械学習モデル（たとえば Convolutional Neural Network（CNN）などの深層学習モデル）によって障害物や人物を認識する．自動運転車は，環境認識部で得られた情報をもとに他のモジュールと連携する．すなわち，環境認識部で前方に障害物を認識したら，経路計画を変更し，その計画に基づいてステアリングやブレーキの操作を行い，障害物との衝突回避を実現する．本稿では，上述したように処理の中核部に機械学習モデルを含むシステム全体を AI システムと呼称する．

本稿の狙いは，（深層学習を中心とする）機械学習モデル単体を対象としたセキュリティにとどまらず，機械学習モデルの周囲に存在するさまざまな入出力，機能モジュールを含む統合されたシステム全体を対象とし，そのようなシステムに特有なセキュリティ課題にフォーカスした研究動向を解説することにある．

## AI とセキュリティ

AI とセキュリティの接点は，（1）「**AI を用いた防御技術**」，（2）「**AI を用いた攻撃技術**」，（3）「**AI に対する攻撃・防御技術**」の 3 つに大別できる．より詳しくは筆者が 2020 年に日本セキュリティ・マネジメント学会誌で執筆した解説記事[☆1]をご参照いただきたい．いずれも応用上重要な問題設定であり，それぞれ盛んに研究開が行われているが，本稿は（3）として「**AI システムに対する攻撃・防御技術**」を解説する．

「AI システムに対する攻撃」は，AI システム全体に対する攻撃であるが，本稿では，特に攻撃対象として AI のパート，つまり機械学習モデルに焦点を当てる．AI システムに付随する OS やソフトウェアの脆弱性をつく攻撃は本稿の範囲外である．

AI システムに対する攻撃では，機械学習モデルに内在する固有な性質を利用し，機械学習モデルの作成者が意図していないモデルの動作，すなわち予測や分類の結果を引き起こすことを狙いとする．そのような性質に着目した基礎検討は，主

☆1 機械学習とオフェンシブセキュリティ，日本セキュリティ・マネジメント学会誌（Web），33(3), (2020).

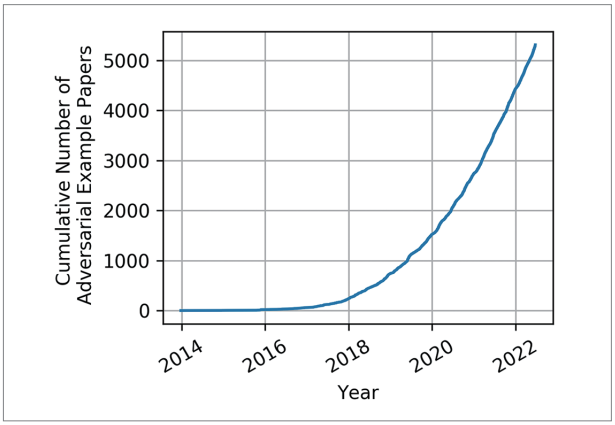
として機械学習研究コミュニティで培われてきたが，セキュリティ研究コミュニティにおいても応用研究が精力的に行われている．前者が機械学習モデルのロバスト性に着目した研究が多いのに対し，後者は AI システムを対象とした現実的な脅威に着目した研究が多いという違いがある．

AI システムに対する具体的な攻撃方法として，（1）意図的に機械学習の誤分類を誘発するデータを生成する**敵対的入力**（Adversarial Example あるいは Adversarial Input），（2）一般に作成に多大なコストを要し，財産的価値のある機械学習モデルを第三者が限られた情報をもとに再現してしまう**モデル抽出**（Model Extraction），および（3）機械学習モデルの訓練に用いたデータを類推，再構築する**モデル反転**（Model Inversion）などがある．本稿では，3 つの中で最も研究事例が多い敵対的入力に着目して研究動向を解説する．

以下では，はじめに（単体の）機械学習モデルを対象とした敵対的入力の原理と対策技術を示し，つぎに AI システムを対象とした敵対的入力の事例を対策技術を解説する．

## 機械学習モデルへの敵対的入力

敵対的入力の研究に先鞭をつけたのは，Goodfellow らが 2014 年に arXiv で公開した論文[1] である．同論文はパンダの画像にわずかなノイズを加えることで，物体認識を行う深層学習アルゴリズムが，高い確信度でテナガザルと誤認識してしまう例で有名であり，敵対的入力研究ブームの火付け役となった．Nicholas Carlini による文献調査[☆2] によれば，2013 年 12 月から 2022 年 6 月までの 8 年半で arXiv に登録された敵対的入力攻撃関連論文の累積件数は実に 5,000 本を超える（図 -1）．これは 1 日平均 1 ～ 2 本の論文が生産され続けているペースであり，今なお増加傾向にある．




■図 -1　ArXiv に投稿された敵対的入力関連の論文数[☆2]

以下では，文献 1）に沿って，敵対的入力の生成方法の概略を示す．敵対的入力に関する研究は速いペースで広範囲に発展しており，以降の内容は基礎的な内容にとどまることに注意が必要である．網羅的なサーベイは文献 2）などを参照されたい．

### 敵対的入力の生成方法

機械学習アルゴリズムへの入力 $X$ に対し，小さな摂動（ノイズ）$\eta$ を加えた $\tilde{X}=X+\eta$ を考える．一般に摂動 $\eta$ が十分に小さければ，機械学習アルゴリズムは $X$ と $\tilde{X}$ を同じクラスに分類すると期待する．このとき，人間は $X$ と $\tilde{X}$ の微細な違いを認知することは困難である．しかしながら以下で示すような工夫により，機械学習アルゴリズムが $X$ と $\tilde{X}$ を異なるクラスに分類するような敵対的入力 $\tilde{X}=X+\eta$ を意図的に作り出すことができる．

敵対的入力により，機械学習アルゴリズムが，本来分類されるべきターゲットのクラス $Y$ とは異なるクラスに誤分類するケースを考える．深層学習の訓練におけるコスト関数（損失関数と正則化項を合わせたもの）を $J(\theta,X,Y)$ とする．ここで $\theta$ は機械学習モデルの性能を最適化するためのパラメタである．コスト関数を最大化する入力の方向は勾配 $\nabla_x J(\theta,X,Y)$ で与えられる．このとき，摂動に対する最大値ノルム制約の下，コスト関数を最

☆2 https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

大化するような摂動 $\eta$ は以下のように計算することができる．

$$\eta = \varepsilon \, \mathrm{sign}(\nabla_x J(\theta, X, Y))$$

sign(・) は符号関数であり，入力が正なら 1 を，負なら -1 を返す．摂動後の入力 $\tilde{X}=X+\eta$ は元の入力との距離は小さいものの，コスト関数を最大化する勾配ベクトルの方向に合わせて $+\varepsilon$ あるいは $-\varepsilon$ ずつ数値が変化している．

以上で示した敵対的入力の生成方法を Fast Gradient Sign Method（FGSM）と呼ぶ．FGSM で計算した摂動を入力値に加えた場合，必ずしも誤分類を引き起こすことは保証されないが，多くの場合で攻撃が成立することが実験的に示されている．

## 敵対的入力の脅威モデル

FGSM を用いた敵対的入力の生成においては，攻撃者は対象とする機械学習モデルのアーキテクチャおよび訓練後のパラメタ $\theta$ を知っていると仮定している．そのような仮定は，攻撃者がターゲットとなる機械学習モデルやパラメタにアクセス可能，すなわち標的となるシステムが採用している機械学習モデルが完全に**ホワイトボックス**であることを意味する．現実的には，攻撃対象とする機械学習モデルが完全にホワイトボックスとなる状況はそれほど多くない．なぜならビジネス上の資産でもある機械学習モデルへの完全なアクセスを許すということは，敵対的入力による攻撃がもたらす損害をはるかに越える損害をもたらすと予想される攻撃が，すでに成立していることを示唆するからである．もっとも，OSS として公開されている機械学習モデルやパラメタを用いているような場合は，対象がホワイトボックスとなる可能性がある．有料で販売されている訓練済みのモデルがターゲットとなる可能性もある．

敵対的入力がもたらす脅威は，ブラックボックスモデルに対する攻撃が成立すると，より現実味を帯びてくる．Nicolas Papernot らは，異なるアーキテクチャの機械学習アルゴリズムで作られた敵対的入力が，別のアーキテクチャを持つ機械学習アルゴリズムに対しても敵対的入力として動作する性質 “Transferability” を利用することにより，商用の機械学習分類サービス（いずれもモデルの詳細は非公開）に対する敵対的入力攻撃が高確率で成功することを実証した．すなわち，ターゲットとなる機械学習アルゴリズムがブラックボックスである場合でも，攻撃者は敵対的入力を用いた攻撃を成功させる可能性がある．

## ユニバーサル敵対的摂動

前節で示した FGSM のような方法において，攻撃者が敵対的入力となる摂動 $\eta$ を計算するためには，あらかじめ入力 $X$ を知る必要がある．この条件は脅威モデルの観点からすると，やや制約が強い．なぜなら，機械学習モデルには未知のデータが入力されることが多いからである．

ユニバーサル敵対的摂動（Universal Adversarial Perturbation : UAP）は，入力 $X$ によらず，機械学習モデルの分類を変更するような，汎用的な摂動を計算するアプローチである．機械学習モデルへの入力となり得るあらゆるデータ $X$ のうち，ほぼすべてが，本来とは異なるクラスに分類されるような摂動 $X+v$ を計算する．ただし，$v$ は十分に小さいものとする． UAP の生成アルゴリズムの詳細は割愛するが，各データポイントに対し，摂動のサイズをあらかじめ定めたしきい値内に維持する範囲内で，元のクラスから他のクラスへの識別境界面に到達するような最小の摂動を計算する．このようにして得た摂動ベクトルを重ね合わせることにより，すべての入力を識別境界面に近づけるような UAP を計算することができる．

最終的に得られた UAP は，極力サイズを小さくするように計算するとはいえ，攻撃が成功する条件を満たすには，人間が認識する可能性がある，不自然なパターンの摂動を生成するケースが多い．

## 敵対的入力攻撃の対策技術

敵対的入力に対する対策は，主として機械学習モデルベースで実施する対策と，AI システム全体として実施する対策に分けることができる．本稿のフォーカスは後者の AI システム全体としてのアプローチにあるが，本節では前者の機械学習モデルベースの対策技術の概略を示す．

機械学習モデルベースの対策技術は，ある機械学習モデルに対して生成された敵対的入力を無効化することを狙いとしたリアクティブなアプローチと，ターゲットとなる機械学習モデルをあらかじめ敵対的入力に対して耐性をもたせることを狙いとしたプロアクティブなアプローチに分類できる [2]．

前者のリアクティブなアプローチは，敵対的入力検出（adversarial detecting），入力データ再構築（input reconstruction），ネットワーク検証（network verification）が挙げられる．後者のプロアクティブなアプローチは，ネットワーク蒸留（network distillation），敵対的訓練（adversarial training），分類器のロバスト化（classifier robustifying）が挙げられる．これらのアプローチのうち，特に研究事例が多いのは，敵対的入力検出と，敵対的訓練である．以下にそれぞれの概略を示す．

### 敵対的入力検出

敵対的入力検出は，データの分類時に，入力したデータが敵対的入力であるか否かを機械学習モデルによって検出する対処療法的アプローチである．あらかじめ用意した通常の入力と敵対的入力を二値分類する機械学習モデルを用いる方法，ベイズニューラルネットワーク等を用い，不確実さが高い入力を敵対的入力と判定する方法など，さまざまな方法が提案されている [2]．一方，これらの多くの敵対的入力検出手法は，C&W（Carlini & Wagner）等の新たな方法で生成した敵対的入力をうまく検出できないという報告がある．

### 敵対的（再）訓練

敵対的訓練（あるいは敵対的「再」訓練）は，生成した敵対的入力を使って機械学習を再訓練することにより，分類器をあらかじめ敵対的入力に対する頑健性を高めるようにする予防的アプローチである．このアイディアは，文献 1）で示された．敵対的訓練では，訓練が 1 ステップ進むごとに，その時点でのモデルを用いて敵対的入力を生成し，訓練データに加える．このようなステップを繰り返すことにより，敵対的入力に対してロバストなモデルを訓練することができる．敵対的訓練は，FGSM のように 1 ステップで生成した敵対的入力に対して有効であるが，繰り返し処理により生成した敵対的入力に対する有効性は落ちるとの報告がある [2]．

以上で示したいずれの対策技術も，特定の攻撃に対する耐性を高めることができるものの，それ以外の攻撃に対して脆弱であることが知られている．現時点において，あらゆる敵対的入力を防ぐことができる，万能な防御技術は存在しないのが現状である．研究のトレンドとしては，防御手法が提案された後，それを破る攻撃方法が提案されることが多い（一般に，防御側はあらゆる攻撃に備える必要があるのに対し，攻撃側は穴をついて突破すれば良いので，有利である）．ただし，以下で示すように，機械学習モデル単体ではなく，AI システムへの敵対的入力としてとらえた場合，いくつかの入出力を組み合わせた実用的な防御を行うことができる．

## AI システムへの敵対的入力

前章では，機械学習モデル単体に対する敵対的入力を解説した．本章は AI システムに対する現実的な敵対的入力攻撃の事例と，固有の課題を議論する．AI システムへの入力は，サイバー空間で生成されるデジタル領域のデータをターゲットとした敵対的入力，および物理空間で計測されるアナログ領域のデータをターゲットとした敵対的入力に分類するこ

とができる．以下では，はじめにデジタルデータへの敵対的入力，およびアナログデータへの敵対的入力の事例を解説する．その後，敵対的入力をどこで入力するかという問題を考察する．

## デジタルデータへの敵対的入力

サイバー空間で生成されたデジタルデータが入力となる AI システムの一例は，画像分類システムである．そのようなシステムでは，デジタル画像データがシステムに入力される．このようなシステムに対して敵対的入力の印加（injection）を試みる攻撃者は，事前にターゲットとなるデジタル画像データに対して生成した敵対的入力を，新たな画像データとしてシステムに入力することにより，攻撃を実行する．攻撃者は計算した敵対的入力を，ターゲットとなるシステムに対して劣化が一切ない状態で入力できるため，敵対的入力攻撃の成功率が高い．

上記のような条件を満たす，AI システムと敵対的入力攻撃の例は，画像の内容に基づく動的コンテンツフィルタリングシステムへの攻撃である．動的コンテンツフィルタリングは，たとえば Web 等で配布される画像や動画データを解析し，あらかじめ定めたポリシーに従って遮断すべきデータであるか否かを判断する技術である．動的コンテンツフィルタリングシステムの応用例は，広告の遮断（アドブロッキング）や，アダルトコンテンツ等のフィルタリングである．いずれのケースにおいても，コンテンツ提供者がフィルタリングシステムを回避する明快な動機が存在する．すなわち，広告事業者はアドブロッキングを回避することで，アダルトコンテンツ事業者はフィルタリングを回避することで，それぞれ収益の増加を期待できる．

広告を検出する方法の 1 つは，広告画像に対して深層学習（畳み込みニューラルネットワークなど）を適用する方法であり，知覚的アドブロッキング（perceptual adblocking）とも呼ばれる．また，アダルトコンテンツの検出にも深層学習が用いられ，商用化されている．したがって，コンテンツフィルタリングを回避する動機を持つ事業者は，敵対的入力を使うことにより，元のコンテンツに与える影響（ユーザに提供するコンテンツの品質劣化）を極小化しつつも，機械学習モデルによる検出（分類）を誤らせることができる．現時点において，実際にコンテンツ提供事業者が敵対的入力を用いたという明確な事例は報告されていないが，脅威の構図としては普遍的なモデルであることに注意が必要である．実際，スパムフィルタを回避することを目的としたメール文面の構成は古くから知られているが，同様の脅威モデルに基づいている．

## アナログデータへの敵対的入力

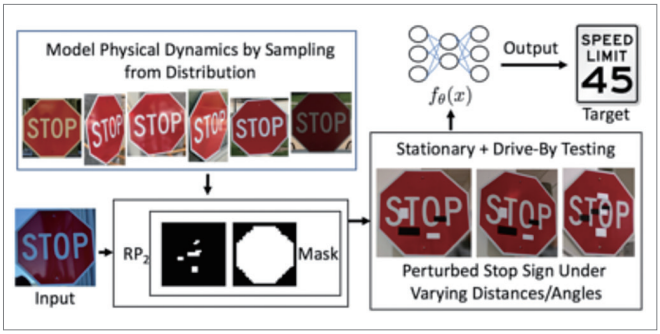
今日広く活用されている AI システムは，元々の入力は光（カメラ）や音声（マイク）などアナログデータであるケースが少なくない．そのようなデータに対する敵対的入力を考えた場合，アナログ領域で敵対的入力を加える攻撃と，デジタル領域で敵対的入力を加える攻撃が考えられる．どちらの領域での敵対的入力が現実的であるかは問題設定によって異なるが，本節では前者のアナログ領域における敵対的サンプル，すなわちアナログデータへの敵対的入力を考える．

はじめに，カメラ画像に基づく物体認識を行う機械学習モデルに対する，アナログ領域での敵対的入力を考える．一般に画像に対する敵対的入力は，文献 1）が示したパンダをテナガザルと誤認識された例のように，デジタル領域で実験・評価されるケースが多い．しかしながら，現実的な脅威モデルを考えると，アナログデータとして敵対的入力を印加する方が自然であることが少なくない．たとえば，自動運転のようにリアルタイムで撮影され続けるカメラ画像に対して物体認識アル

ゴリズムを適用し，衝突回避を行うような応用を考える．このとき，カメラが撮影するデータは不定であり，またカメラで撮影した画像に対して攻撃者がリアルタイムで任意の摂動を印加することは困難であるため，デジタル領域での敵対的入力は現実的ではない．これに対し，アナログ領域での敵対的入力は，カメラが撮影し，機械学習モデルが分類する対象物（たとえば道路標識など）に対して，直接物理的な実体として敵対的入力を加えるアプローチをとる．これにより，前述した問題を解決できる．このとき，対象をカメラが撮影する画像は，距離，角度，光量，ノイズ，周囲の物体などさまざまな条件の影響を受けるため，ユニバーサル敵対的摂動のように，さまざまな入力のバリエーションに対して適用可能な頑健性がある敵対的入力を生成する必要がある．

文献 3）では，上述した背景をもとに各道路標識に対して機械学習モデルが誤認識するような摂動パターンを計算し，印刷したパターンを物理的な標識に貼ることにより，誤認識を引き起こすことが可能なことを実証した．図 -2 は，さまざまな条件下で撮影した一時停止の標識画像を入力とし，機械学習モデルによる認識結果を速度制限（45 mph）とするような敵対的入力パターンを生成する手順を示している．生成にあたり，摂動が標識の外にはみ出さないように，マスクを用いた制約，および印刷可能であるようなパターンとなるような制約を加えている．

次にアナログ領域における，音声認識システム




■図 -2　物理的な道路標識に対する敵対的入力攻撃の例 [3]

をターゲットとした敵対的入力を考える．攻撃者はあらかじめ作成した敵対的入力をスピーカから再生することにより，音声認識システムによる認識結果を意図的に誤らせることができる．敵対的な音声入力は，たとえば通常の人間には音楽のように聴こえるが，音声認識の結果が音声アシスタントシステムへのコマンドとして認識されるものである．攻撃者がテレビやラジオの放送に敵対的入力を忍び込ませることにより，意図せぬ音声認識によるコマンド実行が広範囲で発生するリスクがある．

敵対的な音声入力を物理空間で成功させる際の最大な課題は，カメラのケースと同様，周囲のノイズや環境の変化に対する頑健性である．実際，デジタル領域で生成した敵対的音声入力をそのままスピーカで再生しても，攻撃がうまくいかないことが知られている．

筑波大学の矢倉と佐久間の研究チームは，実空間上での攻撃が可能な，頑健性が高い敵対的音声入力生成方法を提案し，実空間における実験・評価した [4]．鍵となるアイディアは，実空間における反響，バンドパスフィルタ，環境ノイズが与える影響を明示的に最適化問題に組み込んだ上で敵対的入力を生成することである．同論文では，市販のマイクと PC に接続されたスピーカを使った室内実験を行い，生成した敵対的音声入力が Deep Speech によって狙い通りに認識されること，およびユーザスタディにより，ターゲットとなるテキストに対応する音声がまったく認識されないことを報告している．

## AI システムへの敵対的入力に対する防御

敵対的入力攻撃の本質はモデルの出力が不安定になるような（微細な）入力を生成し，それを悪用することにある．機械学習モデルの不安定性を解消する 1 つのアプローチは機械学習モデル自体を敵対的入力に対して頑健にすることであり，前

述した敵対的再訓練が該当する．もう 1 つのアプローチは，機械学習モデルとは別に，チェック機構を設けることである．やはり前述した敵対的入力検出はそのようなアプローチである．敵対的再訓練と敵対的入力検出は機械学習モデルをベースとした対策技術であり，基本的にはデジタル領域で動作する．AI システムの場合は機械学習モデルと独立したチェック機構を設け，さらに対象をアナログ領域に拡張することができる．

AI システムの例として，自動運転の環境認識を考える．自動運転の環境認識では，カメラや LiDAR センサを用いて物体認識を行う．車が走行中に前方に障害物が検出されたら，停止あるいは衝突回避するような運転操作が行われる．カメラを用いた物体認識は 2 次元画像データに畳み込みニューラルネットワークを適用することで，LiDAR を用いた物体認識は 3 次元点群データに畳み込みニューラルネットワークを適用することで，それぞれ物体認識を実現している．これらのニューラルネットワークの目的は同じであるが，異なるデータを入力とする異なるアーキテクチャのモデルであるため，両者は独立している．したがって，カメラ画像に基づく機械学習モデルが認識できないような敵対的入力を生成したとしても，LiDAR センサに基づく機械学習モデルによって，検出することができる．一般に，複数センサを用いることにより，敵対的入力を防ぐことができる場合がある．

それでは，カメラと LiDAR，両方のセンサを同時に騙すような敵対的入力を作成することは可能であるだろうか？　実は 2021 年の IEEE S&P にて，そのような入力を生成することが可能であることが報告されている [5]．同論文では，カメラ，LiDAR の 2 つの物体認識モデルを同時に騙すような入力（どちらのモデルにも認識されないようなオブジェクト）を，3D プリンタで印刷できるような制約をつけた上で生成することに成功している☆3．

上記の例のように，攻撃者が複数センサを騙す敵対的入力を生成できるとしたら，防御として打つ手がないかというと，そういうことはない．対策にどこまでコストをかけるかはリスクアセスメントの結果次第であるが，商用の自動運転車では，衝突回避を目的としてミリ波レーダを用いた追突防止機能を実装しているケースがある．このような機械学習モデルを用いないチェック機構を導入することにより，文献 5）のような複数センサを同時に騙すような攻撃を防止することができると期待できる．

同様に，文献 3）のようなカメラ画像を用いた標識認識に対する敵対的入力に対しては，自動運転車が採用しているベクタマップ（地図上に存在する信号や標識などの情報をベクトルデータとして定義したデータ）を用いたチェックができる．すなわち，機械学習モデルが誤って存在しない赤信号を認識したとしても，それが正しい情報であるかはベクタマップを参照することでダブルチェックできる．さらに路車間通信である V2I（Vehicle-to-roadside-Infrastructure）技術では，路側機から車両に向けて直接信号情報が配信される．将来の自動運転車は，そのような信頼性が高いデータを利用できる場面が増えていくと予想される．

以上で示したように，AI システムにおいては AI 以外のモジュールで実装される各種チェック機構を採用することにより，機械学習モデルに依存しない対策が可能である．先に述べたように敵対的入力の本質は機械学習モデルの不安定性にあり，他の機構を利用して二重，三重のチェックをするアプローチが有用である．

☆3 筆者らの研究グループが再現実験を試みた限りでは，3D シミュレーション，3D プリンタの印刷物を用いた実験のどちらにおいても必ずしも敵対的入力は動作せず，不安定な結果を得ている．

## 敵対的入力の現実性

### 敵対的入力をどのように入力するか？

AIシステムに対する敵対的入力を考える場合，入力をデジタル領域で行うのか，アナログ領域で行うかによって，敵対的入力を印加する方法や条件は大きく異なる．デジタル領域であれば，アルゴリズムで計算した敵対的入力をそのままデータに加算すればよい．一方，アナログ領域で敵対的入力を印加する場合，前述したような物理空間におけるアナログ信号の伝達に特有な性質に対し，頑健性を有する敵対的摂動を生成する必要があり，攻撃を成功させる難易度はより高まる．技術的には，それぞれのドメイン知識に基づく制約条件を導入した最適化問題を解くことになる．そのようにして生成したデータを物理空間内におけるオブジェクトに付け加えるか，アナログ信号として出力することにより，攻撃が成立する（可能性がある）．

また，攻撃の対象となるデータ入力がリアルタイムデータであるのか，すでに生成済みのデータであるか（リプレイも含む）によって，敵対的入力を生成すべきタイミングや，生成方法は異なる．前述のカメラの例で見たように，敵対的入力攻撃をリアルタイムで実施する場合，事前に収集した特定のデータに最適化した敵対的入力を用いることは現実的ではない．機械学習モデルに入力されるデータは時々刻々と変化し，またさまざまなバリエーションを持つため，ロバストな敵対的入力の生成が必要である．そのような敵対的入力の生成方法として，ユニバーサル摂動や敵対的パッチなどの手法を使うことができる．これに対し，すでに生成済みのデータに対する敵対的入力の生成はストレートな問題である．

以上を総合すると，組合せとしては，アナログ領域において，リアルタイムで敵対的入力を生成するケースが最も攻撃の難易度が高いと言えるであろう．筆者らは心電図（ECG）データを計測し，種々の不整脈パターンを検出する機械学習モデルに対し，ユニバーサル摂動を用いた敵対的入力攻撃を考案し，実システムを用いてその実現性を評価した[6]．同研究はまさにこのケースに相当している．

### 「認知できない」摂動は本質的か？

敵対的入力を生成する問題では，摂動が人間に認知できないほど小さいことを前提としている．そのような前提は，画像や音声など，人間が直接見たり聞いたりすることができ，かつその内容の妥当性や自然さを瞬時に判定できる場合には合理的であるが，AIシステムが扱うデータは必ずしも人間が認知したり，瞬時に妥当性や自然さを判断できるデータとは限らない．たとえば，機械学習モデルを用いてマルウェアを検出・分類するAIシステムを考える．一般に，人間がマルウェアのバイナリを一瞥しただけで，そのファイルに敵対的摂動が加えられているかを瞬時に判定することは困難である．したがって，このようなAIシステムに対する敵対的入力（マルウェアに摂動を加え，良性ファイルと判定させることを狙いとする）は，必ずしも微細な摂動である必要はない．むしろ，元のマルウェアと敵対的入力のバイナリ列が類似していれば，Fuzzy Hashingなど，機械学習と独立した別のロジックによって検出される可能性がある．同様に，自動運転において標識を認識するAIシステムを騙すような敵対的入力を考える．ターゲットとなる自動運転車がクラス5の完全自動運転であれば，人間が標識を見る必要はなくなる．そのようなケースであれば，摂動を小さくする必要はまったくない（極端には，別の標識に置き換えてしまえばよい！）．さらに，一般に機械学習モデルは大量のデータに対して適用されるケースが多い．それらの大量のデータを人間がすべて見たり聞いたりすることは現実的ではないケースが存在し，そのようなケースにおいては摂動が「認

知できない」ことは必ずしも必要ない．

敵対的入力攻撃は，画像認識を例題とした問題設定に端を発する経緯があるため，攻撃者が認知できない摂動を生成する（したい）ことを前提として捉えられがちである（良くも悪くもパンダの例が我々に先入観を与えている）．しかしながら実用的なセキュリティの問題として敵対的入力の脅威を評価する場合は，そもそも入力データを人間が認知可能であるのか，そしてデータの妥当性や自然さを判定でき得るか否かを見定める必要がある．問題によっては，人間の認知は本質的ではなく，機械学習モデルにとって不安定な入力を人為的に生成し，AI システムの出力を任意に制御する脅威のみがセキュリティの問題として重要なケースがある．

## 今後の課題

本稿では，AI システムをターゲットとした敵対的入力攻撃とその防御方法について，いくつかの事例をまじえて解説した．AI システムに特有な点として，アナログ領域の敵対的入力の脅威評価が必要であること，およびリアルタイムの攻撃を実現するためには，環境や入力の多様性に対する頑健性が必要であることを示した．社会実装されているさまざまな重要 AI システムを対象として，敵対的入力の評価を行うことは今後の課題である．本稿では，機械学習モデル単体の評価では，システム全体としてのリスクアセスメントを見誤る恐れがあることを強調した．攻撃の難易度（コスト）と実現性の評価，および機械学習モデルに依存しない対策技術の考案が AI を活用したシステムのセキュリティ向上に資すると考えられる．

**参考文献**

1）Goodfellow, I. J., Shlens, J., and Szegedy, C. : Explaining and Harnessing Adversarial Examples, arXiv:1412.6572v3 [stat.ML] (Mar 2015).
2）Yuan, X. et al. : Adversarial Examples : Attacks and Defenses for Deep Learning, IEEE Transactions on Neural Networks and Learning Systems 30, pp.2805-2824 (2019).
3）Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D. : Robust Physical-World Attacks on Deep Learning Models, arXiv:1707.08945v5, pp.1–11(Apr. 2018).
4）Yakura, H. and Sakuma, J. : Robust Audio Adversarial Example for a Physical Attack, Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp.5334–5341(Aug. 2019).
5）Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., Chen, Q. A., Liu, M. and Li, B. : Invisible for both Camera and LiDAR : Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks, IEEE Symposium on Security and Privacy 2021, pp.176-194.
6）Ono, T., Sugawara, T., Sakuma, J. and Mori, T. : Application of Adversarial Examples to Physical ECG Signals, https://arxiv.org/abs/2108.08972

（2022 年 7 月 17 日受付）

■森　達哉（正会員）　mori@seclab.jp

1999 ～ 2013 年 NTT 研究所．2013 年より早稲田大学．2018 年より理化学研究所革新知能統合研究センター客員研究員兼務．下はハードから上は人間まで，幅広くセキュリティの諸問題に興味を持って研究をしている．