［ディジタルエコノミー時代のサイバーセキュリティ―ディジタルトランスフォーメーション促進の基盤確立に向けて―］

⑥ AIをセキュリティリスクから守るために ―AIへのサイバー攻撃とその対策―

古澤一憲 | 三菱総合研究所

AIの可能性とセキュリティ

いわゆる「第3次AIブーム」の到来からもそれなりの時間が経過し，機械学習のアルゴリズムを組み込んだプロダクトが市場に投入されることも一般的になった．医療画像を用いた診断，産業機械の故障予測，利用者へのQ&A対応を行うチャットボット等，さまざまな領域でのタスクを担うAIが次々と開発されている．こうしたAIは，特定の条件下では人間が実施する以上の精度の結果が得られることもあり，大きな可能性を秘めている．

一方で，その有用性から，悪意を持った攻撃を受けた場合には，一転して大きな危険にさらされる可能性もある．また，攻撃者がAI技術を悪用することでより高度なサイバー攻撃を試みる事例も存在する．機械学習を用いてセキュリティ対策ソフトの検知を回避する機能を持ったコンピュータウィルス等がすでに発見されている．

本稿では，はじめにAIのリスクを巡る国内外での議論を紹介し，AIのセキュリティリスクについての考えを述べる．次にAI特有のサイバー脅威について，代表的な研究動向を紹介する．最後にAIシステムを守るために取り得る対策について考察を述べる．

■図-1　アシロマ会議の様子[☆2]

AIのリスクに関する国内外の議論

AIにかかわるリスクに関する議論は国内外で活発に行われている．

米国主導で議論，検討がされている指針としては，IEEEにおける自動システム（AI，ロボット他）設計における倫理的配慮標準化のための議論[☆1]，2017年のアシロマ会議において発表された「**アシロマAI原則**」[☆2]や，米Facebook，Amazon，Alphabet（Google），DeepMind，IBM，Microsoftらを中心とした「Partnership on AI」[☆3]等が代表的である（図-1）．それぞれにおいて，AIの「**倫理性**」や「**透明性**」といった指針が示されている．実際に定められる指針は各々異なるが，「**安全性**」に言及している点は共通である（ただし，これらは，技術開発を抑圧しないことに配慮した非拘束的な指針である）．

一方，欧州ではロボット法を議論する枠組みの中でAIの安全指針に関する言及が見られる．自律ロボットを定義する基準として機械による自己学習が含まれており，民法規則の視点から法的責任も視野に入れた議論を行うアプローチである．

国内では，AIネットワーク社会推進会議（総務省）による**国際的な議論のためのAI開発ガイド**

☆1 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
☆2 https://futureoflife.org/ai-principles/
☆3 https://www.partnershiponai.org/

ライン案（以下，「AI開発ガイドライン案」という）[☆4]が2017年に公表されている．G7やOECD（経済協力開発機構）における国際的議論に供することを念頭に作成された本ガイドラインでは，AIシステムの開発者向けに非拘束型の9原則を提示している．ここで注目したいのは，原則④「**安全の原則**」と明示的に区別する形で，原則⑤「**セキュリティの原則**」を定めている点である（表-1）．

AIシステムのセキュリティ

AI開発ガイドライン案において「安全の原則」は，「**開発者は，AIシステムがアクチュエータ等を通じて利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する**」こととされ，AIを搭載したロボットが危険な動作を行うことを防ぐこと等を目的とした原則であると解説されている．

「**セキュリティの原則**」は，「**開発者は，AIシステムのセキュリティに留意する**」こととされ，情報の機密性・完全性・可用性の確保，必要に応じたAIシステムの信頼性（意図したとおりに動作が行われ，権限を有しない第三者による操作を受けないこと）への留意，セキュリティリスクの評価・抑制等が推奨されている．

セキュリティの概念は，特に安全（セーフティ）との差異に着目する場合，正当な権限を持たない第三者による（悪意ある）不正行為が，その特徴として挙げられることが多い．たとえば，アクセス権限のないユーザからの不正アクセスによる機密性の侵害，データの改ざんによる完全性の侵害，サービス不能攻撃による可用性の侵害といった被害が典型的である．

こうした典型的被害はAIシステムにおいても同様に考え得るものだ．AIシステムの管理者アカウントへのアクセス管理に脆弱性が存在し，ソースコードの改ざんが行われた場合，生じる被害の大きさは想像にかたくない．しかし，この例において改ざんされるソースコードがAIソフトに分類すべきものであるかは本質ではない．元のソースコードの内容を無視し，攻撃者の用意した（AIソフトではない）ソースコードへ全面的な置換えを行うと仮定しても，十分な被害が生じるシナリオであるからだ．

一方で，AI特有の事情に基づく脅威もやはり存在する．機械学習モデルの予測結果はその性質上，ある尤度での確率的推論の結果となる．さらに，予測に至るまでの根拠は必ずしも人間が完全な理解をできる形式ではなく，妥当性を100%の精度で検証することは不可能であることにも留意する必要がある．

こうした特性そのものを悪用することを意図した

■表-1　AI開発ガイドライン案の9原則[☆4]

主にAIネットワーク化の健全な進展及びAIシステムの便益の増進に関する原則		
①	**連携の原則**	開発者は，AIシステムの相互接続性と相互運用性に留意する．
主にAIシステムのリスクの抑制に関する原則		
②	**透明性の原則**	開発者は，AIシステムの入出力の検証可能性および判断結果の説明可能性に留意する．
③	**制御可能性の原則**	開発者は，AIシステムの制御可能性に留意する．
④	**安全の原則**	開発者は，AIシステムがアクチュエータ等を通じて利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する．
⑤	**セキュリティの原則**	開発者は，AIシステムのセキュリティに留意する．
⑥	**プライバシの原則**	開発者は，AIシステムにより利用者及び第三者のプライバシーが侵害されないよう配慮する．
⑦	**倫理の原則**	開発者は，AIシステムの開発において，人間の尊厳と個人の自律を尊重する．
主に利用者等の受容性の向上に関する原則		
⑧	**利用者支援の原則**	開発者は，AIシステムが利用者を支援し，利用者に選択の機会を適切に提供することが可能となるよう配慮する．
⑨	**アカウンタビリティの原則**	開発者は，利用者を含むステークホルダに対しアカウンタビリティを果たすよう努める．

☆4 http://www.soumu.go.jp/menu_news/s-news/01iicp01_02000067.html，AI開発ガイドライン案では，「データ・情報・知識の学習等により，利活用の過程を通じて自らの出力やプログラムを変化させる機能を有するソフトウェア」はAIソフト，AIソフトを構成要素として含むシステムをAIシステムと定義されている．本稿も本定義に従った表記を行っている．

攻撃は，AI に特有の脅威と位置付けるべきだろう．第三者の介入により AI の信頼性が損なわれた場合，開発者の意図しない動作が実行され，さまざまな弊害が生じることとなる．次章では，AI 特有の脅威・攻撃手法に関する具体的な研究や事例を紹介する．

AI 特有のサイバー攻撃手法

ここでは，近年特に普及している「データを基にした機械学習を行った AI」に特有の脅威についての研究，攻撃事例について紹介する．

機械学習を行う AI システムの構成は，図 -2 のような構成が基本形である．訓練系は，訓練データからの学習によってモデルの予測精度向上と評価を実施する環境を意味している．十分な学習を行った訓練済みモデルは予測系に配置され，対象データが分類されるクラス等を予測する．

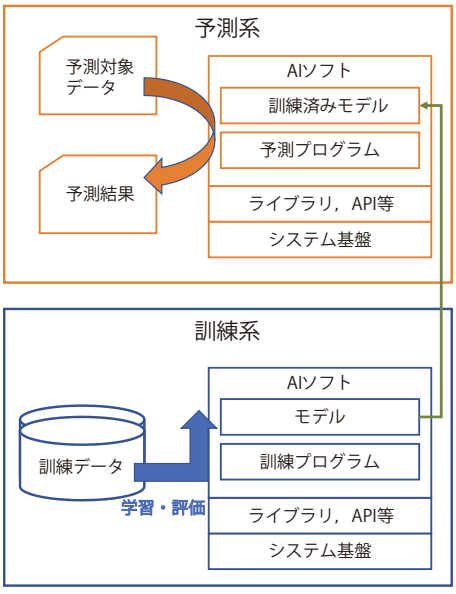
システム構成からも見てとれる通り，AI システムへの脅威は以下の 2 種類に大別して考えられる．

- 予測系への攻撃に分類される脅威
- 訓練系への攻撃に分類される脅威

特に AI システムに特徴的な脅威を考える上で，両系におけるデータ（訓練データ・予測対象データ）もしくは AI ソフト（モデル・プログラム）を対象とする攻撃手法に注目したい．

予測系では，モデルは訓練済みであるため，内容は確定済みであるとここでは考える．よって予測対象データを不正に加工する等の手段で，予測結果の妥当性が意図的に低下させられる被害を検討することが中心となる．対して訓練系では，AI ソフトの学習過程自体が悪意ある介入を受けることで，訓練結果が本来の意図と異なるものとなってしまう被害等が主な関心となる．そのほかの要素への攻撃が行われる可能性も当然考えられるが，手法としては一般的な情報システムを対象としたものに近い．

本章ではそれぞれの脅威を対象として，多くの研究が行われている手法，大きな影響を与える可能性がある手法を中心に紹介していく．今回紹介する手法は表 -2 の通りである．



■図 -2　AI システムの構成例

既知のモデルに誤分類を誘発する攻撃（Adversarial Examples）

訓練済みモデルに対して，予測対象データへ悪意ある加工を行うことで，予測結果が本来の結果とは異なる結果へ誘導されてしまう場合がある．Google の Christian Szegedy らは，深層学習等のニューラルネットワークモデルで学習を行った画像分類器に対し，予測対象データに悪性の微小画像（人間の目では違いを認識することができない程度のノイズ画

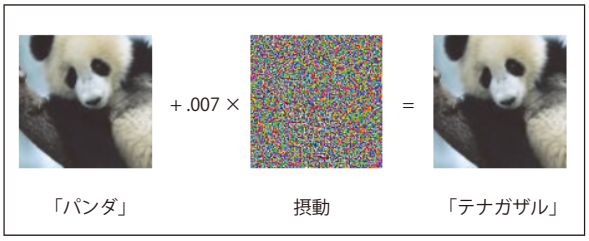
■表 -2　本稿で紹介する AI へのサイバー脅威

予測系への攻撃	訓練済みモデルに誤分類を誘発する攻撃（Evasion Attacks）	
		既知のモデルに誤分類を誘発する攻撃（Adversarial Examples）
		未知のモデルに誤分類を誘発する攻撃（Black-Box Attacks）
訓練系への攻撃	訓練データを汚染する攻撃（Data Poisoning Attacks）	
その他の攻撃	API 経由の情報窃取，機械学習ライブラリの脆弱性を悪用する攻撃等	

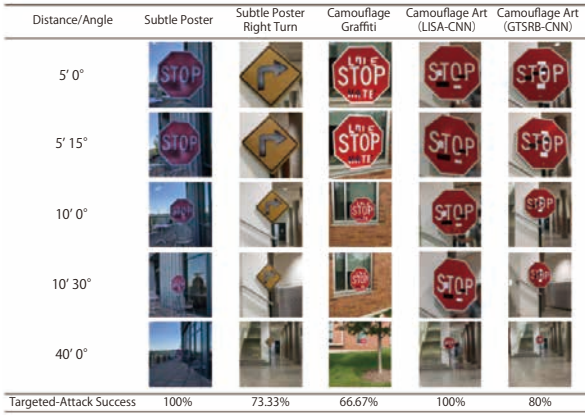
像）を合成することで，予測結果を別のクラスへと変更できることを発見した．

こうした方法で作成された合成画像は**敵対的サンプル（Adversarial Examples）**と呼ばれる．まず，あるモデルに対して，パラメータ（重み）を固定した上で，予測対象データを変化させた場合の勾配を計算する．この際，予測結果に対する摂動が大きくなるデータを選択し，摂動画像とする．最後に，選択した摂動画像に重み付けをした上で，元画像に合成する（図 -3）．不特定の他クラスへの誤分類を試みる手法（Non-targeted）と特定クラスへの誤分類を企図する手法（Targeted）があり，後者の方が難易度は高い．

Ian Goodfellow らは，微小摂動が予測結果に大きな影響を与える原因は，高次元空間が線形性を持つためであるとしている[1]．あるニューラルネットワークモデルで生成した Adversarial Examples が，異なる訓練データセットで学習を行ったモデルや，ロジスティック回帰等の別手法で学習したモデルでも誤分類を誘発しやすいという事実も，高次元空間の線形性によって説明可能としている．また，この線形性を利用して Adversarial Examples を少ない計算量で生成する手法を提唱している．



■図 -3　Adversarial Examples の作成例（文献 1）を基に修正）



■図 -4　道路標識を誤認識させる実験[2]

Adversarial Examples に由来する脅威として，実空間上の事故を狙う攻撃が発生する可能性が挙げられている．特に自動運転車向けのセンサ画像認識 AI に関しては具体的な議論が盛んに行われている．

ワシントン大学の Ivan Evtimov らは，道路標識画像に摂動を加えることで，深層学習による訓練済みモデルに対して，停止標識を速度制限標識に誤認識させることが可能であるという研究結果[2]を公表している．この手法では，距離や角度を一定程度変更した場合においても有効であり，自動運転車の脅威になり得るとされている（図 -4）．

MIT の Anish Athalye らの手法も同様に，角度と縮尺を変更した場合にも有効な Adversarial Examples を生成可能である[3]．Google の Tom Brown らは，本手法を基に **Adversarial Patch** と呼ばれるステッカーを作成し，画像認識ソフトウェアが，ステッカーと同時に映った物体の映像を，攻撃者が意図するクラスへと誤認識してしまう様子のデモンストレーションを公開している☆5．

未知のモデルに誤分類を誘発する攻撃（Black-Box Attacks）

Adversarial Examples が訓練済みモデル中のアルゴリズムや重みの情報等を所与のものとする手法であったのに対し，訓練済みモデル内の情報を未知のものとし，予測結果の情報のみを用いた攻撃についての研究も存在する．こうした攻撃手法は**ブラックボックス攻撃（Black-Box Attack）**と呼ばれる．

スコアや尤度の情報まで用いる手法や，予測値／分類結果のみを用いる手法等，さまざまなアプローチの研究が行われているが，今回は二値分類器の分類結果のみを用いて誤分類を誘発した実験事例を紹

☆5　https://youtu.be/i1sp4X57TL4

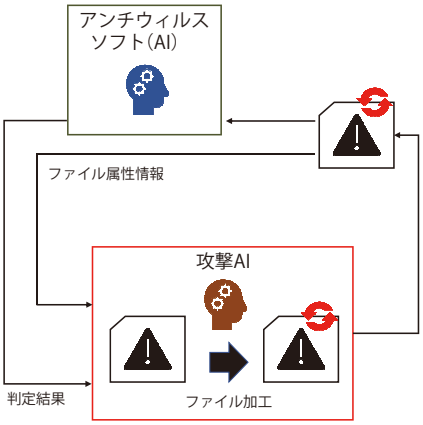
介する．特に情報セキュリティの分野において，ウィルスや迷惑メールの区別等は重要なテーマであり，悪性と良性の二値分類を誤らせる手法については関心が高い．また，分類結果のみを用いる手法は，攻撃者が限られた情報のみで実行であるため，特に注意が必要である．

データサイエンティストの Hyrum Anderson は，Black Hat USA 2017 において，AI による悪性ファイル検出を行うアンチウィルスソフトを対象に，Black-Box Attacks を行った結果を公表した．

この例では，攻撃者側も機械学習を用いる．通常ではアンチウィルスソフトに検知される実行形式の悪性ファイルを基に，ファイルの動作に影響を及ぼさない領域への追記・改変を繰り返す．加工内容は，ファイルの属性情報（サイズ・ヘッダー情報等）を特徴量，アンチウィルスソフトの判定結果を報酬と見なした強化学習によって決定する（図 -5）．

複数の悪性ファイルを用いた検証で，ランダムな改変を加えた場合と比べて悪性と判定される確率が高いという結果が得られている．また，他のアンチウィルスソフトにおける検出率も低下している．

Black-Box Attacks は攻撃対象の情報をほぼ必要としないため，攻撃者目線では利便性が高い手法といえる．物理デバイスに搭載された AI ソフト等，オフラインで複数回の試行が行いやすい対象においては特に注意すべき脅威であると考えられる．



■図 -5　Black-Box Attacks の構成例

訓練データを汚染する攻撃（Data Poisoning Attacks）

訓練系での攻撃は，データの汚染攻撃が第 1 に考えられる．訓練済みモデルの精度は訓練用データの質に依存するため，データのラベルと実態が異なるデータが混入する場合には誤った学習を行うことになる．仮に高い評価結果が得られたとしても，根拠となる評価用データ自体が誤っていれば意味がない．

Microsoft が 2016 年に Twitter 上で公開したチャットボット「**Tay**」は，ユーザとの対話を学習することで応答精度を向上させる AI として制作された．しかし，公開から 1 日とたたないうちに問題発言を繰り返すようになり，即座に公開停止となった．詳細な原因は公開されていないが，悪意をもったユーザが協力して，Tay が問題のある学習を行うよう作為的な対話を行ったと見られている．

Tay のケースでは，ユーザとの対話ログが学習用データとして明示されていた．開発者のデータ収集元が特定可能な場合には，収集元へのデータ汚染リスクが存在することになる．インターネット上のオープンなデータプールが訓練用データとして用いられる場合は多くあるが，こうしたデータベースに，訓練結果を悪化させるデータを意図的に混入された場合，これらのデータセットで訓練をすると，モデル自体が汚染されてしまうことにもつながる．

AI システムへのサイバー攻撃にどう対策すべきか

これまで，AI システムへの脅威となるサイバー攻撃の概要とその具体的な手法について見てきた．ここからは，こうした脅威から AI システムを守るための対策について考える．

第 1 に考えるべきは通常の IT システムと同様にシステムの機密性・完全性・可用性を守る対策となるだろう．攻撃者にシステムへ侵入され，管理者権限を取得された場合には，あらゆる被害が生じる可

能性がある．こうした根本的な脆弱性対策を講じることは必須である．AIシステムでは外部のAPIやライブラリを用いる場合も多いため，開発プロセスやサプライチェーンリスクにも注意を払いたい．

その上で，AIシステム特有の脅威への対策を考える必要がある．本稿では，予測系における誤分類の誘発とデータの汚染と分類される脅威を紹介した．それぞれの脅威への対策を順に考えていく．

予測系における誤分類の誘発は，特定の予測対象データに対して望ましい結果を返せないという視点からは，予測精度の問題とも解釈できる．そこで考慮すべき対策が，ロバスト（誤分類を起こしにくい）なモデルの訓練である．

画像識別器の訓練において，反転画像や色違いの画像等を学習させることで予測精度を高める手法は一般に用いられている．同様に，意図的に誤分類を誘発させる画像を作成した上で，正しいラベルをつけて学習させることで，ロバストなモデルを訓練することができる．さらに，データを分類するモデルと，誤分類を起こしやすいデータを自動的に大量生成するモデルを相互に競わせることで，効率的にロバストなモデルの訓練を行う手法（**Generative Adversarial Network**）も提案されている．

Black-Box Attacksへの対策も同様に，攻撃AIに対抗した学習を行う対策が考えられる．高度な攻撃へ対応可能なAIを開発するためには，高度な攻撃のデータを学習させる必要がある．米DARPA（国防高等研究計画局）が主催する**Cyber Grand Challenge**は，攻撃AIと防御AIを競い合わせるコンテストであり，3年間で50億円以上の予算が投入された．

データの汚染対策としては，まずは学習データの品質担保の対策を考えたい．意図的な汚染の有無によらず，データの品質確保は重要である．また，モデルの評価プロセスにおいて，学習用データと評価用データの分割を工夫するなど，汚染データを検知，除外するための対策も推奨される．ほかのデータで学習を行ったモデルとの比較検証を行うことも有効であると考えられる．

いずれの場合においても，AIシステムへの脅威を把握し，分析をすることが重要である．新しい脅威は次々と出現してくるが，自らの開発するシステムへの脅威を正しく理解し，実効性のあるセキュリティ対策を選択したい．

参考文献

1) Goodfellow, I. J. et al. : Explaining and Harnessing Adversarial Examples, arXiv:1412.6572 (2014).
2) Evtimov, I. et al. : Robust Physical-World Attacks on Deep Learning Visual Classification, arXiv:1707.08945 (2017).
3) Athalye, A. et al. : Synthesizing Robust Adversarial Examples, arXiv:1707.07397 (2017).

(2018年9月14日受付)

古澤一憲　kazunori_furusawa@mri.co.jp

（株）三菱総合研究所 サイバーセキュリティ戦略グループ研究員．専門はセキュリティ政策，セキュリティ技術検証等．