IMES DISCUSSION PAPER SERIES

機械学習システムの脆弱性とセキュリティ・リスク：「障害モード」による分類と今後へのインプリケーション

菅 和聖（かん かずとし）

Discussion Paper No. 2020-J-20

IMES

INSTITUTE FOR MONETARY AND ECONOMIC STUDIES

BANK OF JAPAN

日本銀行金融研究所

〒103-8660 東京都中央区日本橋本石町 2-1-1

IMES Discussion Paper Series 2020-J-20
2020 年 12 月

機械学習システムの脆弱性とセキュリティ・リスク：「障害モード」による分類と今後へのインプリケーション

菅 和聖（かん かずとし）*

要　　旨

機械学習は、膨大な入出力データからそれらの関係を自動的に抽出する帰納的な手法であり、これを画像処理などに組み込んだシステム（機械学習システム）の社会実装が進んでいる。一方、機械学習システムには従来の情報システムにはない脆弱性とそれに伴うセキュリティ・リスクが存在するが、その全体像は明確となっておらず、分類方法も確立していない。本稿では、まず、機械学習システムとそのセキュリティ・リスクの特徴について考察する。次に、最近のサーベイ論文を参照しつつ、同システムの主な脆弱性やそれへ攻撃手法を「障害モード（failure mode）」の観点から、(1)外部からの攻撃の有無、(2)脆弱性の所在領域、(3)喪失する機能特性、の 3 つの軸をもとに分類・整理する。最後に、今後の機械学習システムの活用に向けた留意点を述べる。

キーワード：機械学習システム、障害モード、セキュリティ・リスク、脆弱性
JEL classification: L86、L96、M15、Z00

* 日本銀行金融研究所企画役補佐（E-mail: kazutoshi.kan@boj.or.jp）

本稿の作成に当たっては、石川冬樹准教授（国立情報学研究所）から有益なコメントを頂いた。ここに記して感謝したい。ただし、本稿に示されている意見は、筆者個人に属し、日本銀行の公式見解を示すものではない。また、ありうべき誤りはすべて筆者個人に属する。

目　次

１．はじめに 1
２．機械学習システムの特徴 2
（１）機械学習システムとそれ以外のシステムの差異 2
イ．機械学習を用いない情報システムの場合 2
ロ．機械学習システムの場合 2
（２）機械学習システムの留意点 3
イ．機械学習システムへの要求の不明瞭さ 4
ロ．訓練済みモデルの特性の不明瞭さ 4
３．機械学習システムのセキュリティ・リスクの特性 4
（１）情報処理ルールに内在する脆弱性 4
（２）セキュリティ対策が難しい理由 5
イ．脆弱性の洗出し 5
ロ．情報処理ルールの修正 5
ハ．脆弱性と機能の分離 5
４．機械学習の「障害モード」：Kumar らのアプローチ 6
（１）障害モードとは 6
（２）障害モードの分類方法 7
イ．攻撃の有無 7
ロ．脆弱性の所在 8
ハ．喪失する機能特性 8
５．機械学習を組み込んだサービスの提供に関わる主体とその関係 8
６．Kumar らによる障害モードのリスト 9
（１）攻撃による障害 10
イ．機械学習に特有の脆弱性のみに関連する攻撃 10
ロ．一般的なソフトウェアの脆弱性にも関連する攻撃 15
（２）攻撃によらない障害 18
７．結びに代えて：留意点と今後の課題 20
（１）セキュリティ対策に向けての留意点 20
（２）機械学習システムを利活用していくうえでの留意点 21
（３）今後の課題 21

１．はじめに

機械学習は、予め定められたモデルの枠組みの下で、膨大な入出力の組から入出力関係を自動的に抽出する帰納的な手法である。画像処理などの極めて複雑な入出力関係を扱う課題の解決に適しており、機械学習を組み込んだ情報システム（以下、機械学習システム）の社会実装が進んでいる。金融分野では、資産運用や信用スコアの算出などに応用されている。

その一方で、機械学習システムには、さまざまな脆弱性があることが知られている。この一部には、従来の情報システムにおけるセキュリティ対策で対処できない新しい問題も含まれるが、脆弱性やそれに伴うセキュリティ・リスクの全体像が明確となっておらず、分類方法も確立していない。セキュリティ・リスクの影響は、実際のサービスに依存することから一概に評価できないものの、顔認証での不正アクセスや自動運転での誤作動など、システムの用途次第では重大な結果をもたらす怖れがある。

最近では、機械学習システムの脆弱性に関して、情報セキュリティ・インシデントへの対応活動を行う CERT/CC [1]が脆弱性レポート [2]を発信した。個別の機械学習システムの脆弱性を指摘するものではないが、今後、機械学習システムの活用の進展が見込まれる中で、初めて注意喚起を行うものであり、注目を集めている。機械学習システムを利用していく上では、既知の脆弱性について認識しておく必要がある。

本稿では、2 節で機械学習の特徴について説明した後、3 節で機械学習システムに特有のセキュリティ・リスクについて考察する。4、5、6 節では、マイクロソフトとハーバード大学の研究者らによるサーベイ論文（Kumar *et al.* [2019]）をベースに機械学習システムの既知の脆弱性を分類し、個々の脆弱性や攻撃方法について研究事例を交えながら説明する。同論文は、機械学習システムが機能や特性を失う「障害（failure）」の研究事例を収集し、様態（mode）別に分類したものであり、CERT/CC の脆弱性レポートにおいても引用されている。今後、機械学習システムを金融機関が使用していく場合には、この「障害モード（failure modes）」を参照しながら独自にリスクを評価していくことが重要である。

---

[1] CERT Coordination Center (CERT/CC) は、米国のカーネギーメロン大学が運営する非営利組織。情報セキュリティの研究・開発を行うほか、既存システムの脆弱性の情報を収集し、ソフトウェアベンダーやインシデント対応従事者に情報共有し、注意喚起や脆弱性解消を促進するなどの役割も担う。

[2] Vulnerability Note VU#425163、2020 年 3 月 19 日（https://www.kb.cert.org/vuls/id/425163）。邦訳は JPCERT/CC による脆弱性レポート JVNVU#99619336、2020 年 3 月 25 日（https://jvn.jp/vu/JVNVU99619336/）。

２．機械学習システムの特徴

（１）機械学習システムとそれ以外のシステムの差異

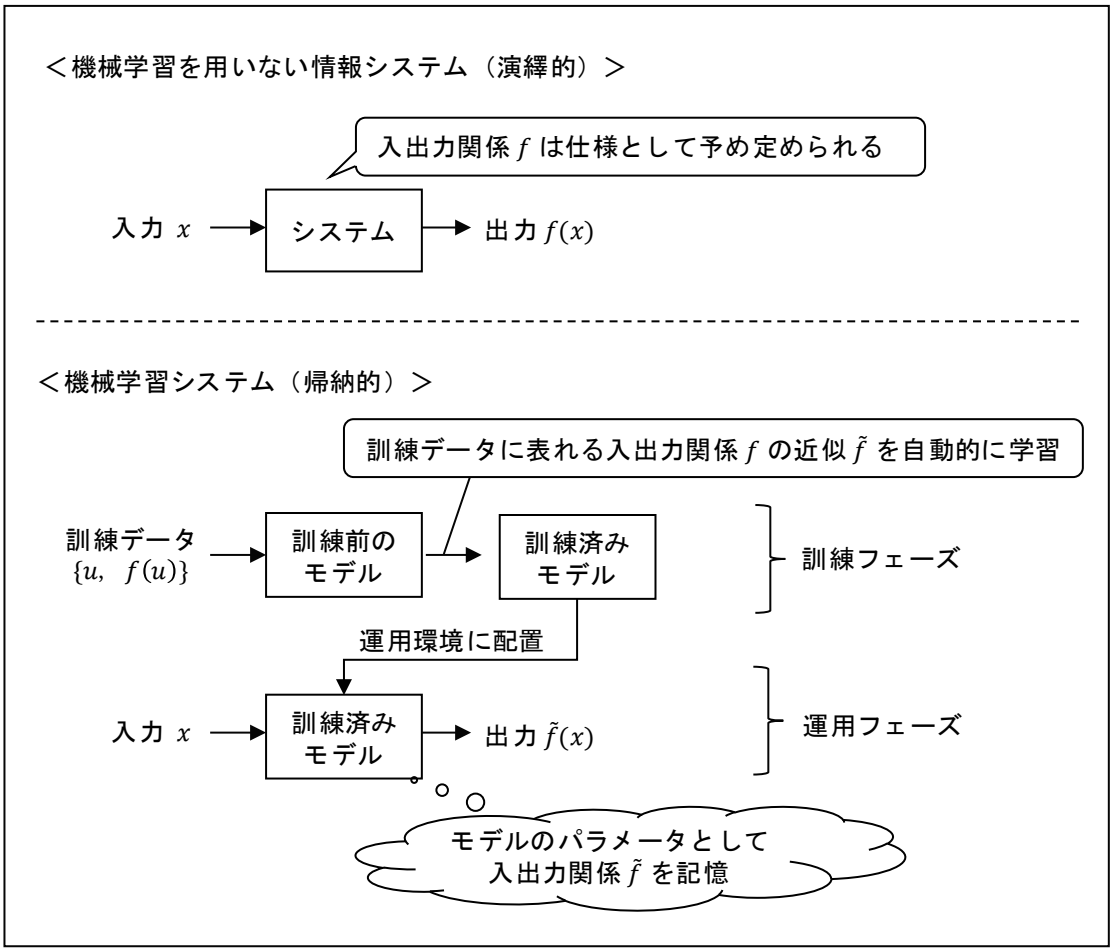
イ．機械学習を用いない情報システムの場合

機械学習を用いない情報システム（図１上を参照）では、入出力関係（f）は開発者によって予め仕様として定められており、システムの情報処理ルールはこの仕様に沿うように実装される。したがって、情報処理ルールはデータに依存しない。機械学習を用いない情報システムは、情報処理ルールを所与として入力から出力を導く演繹的なシステムである。

ロ．機械学習システムの場合

機械学習システム（図１下を参照）では、予め用意した機械学習モデルに対して入出力の組の例示（訓練データ）を読み込ませ、情報処理ルールに相当する

図１ 機械学習システムの入出力の模式図



入出力関係（$\tilde{f}$）を自動的に抽出させる[3]。このとき得られる情報処理ルールは、訓練データに表れる入出力関係（f）の近似になっている。このプロセスは、「訓練」または「学習」と呼ばれる。訓練データを作成することが、機械学習を用いない情報システムにおける仕様策定に相当する。訓練によって得られる情報処理ルールは訓練データと機械学習モデルに依存しており、システムの開発者は直接的には関与できない[4]。このように、機械学習システムは、データから情報処理ルールを導き出す帰納的なシステムである。

情報処理ルールを訓練するための仕組みとして、機械学習では、入出力関係の表現力が高いモデル[5]を用意したうえで、訓練データから訓練済みモデルのパラメータを決定するというアプローチを採る。この訓練を行う情報処理の段階を「訓練フェーズ」と呼ぶ。モデルに求められる表現力が高いほど、より多数のパラメータ（モデルの自由度）が必要になる。訓練された情報処理ルールは、決定されたパラメータの集合のかたちで訓練済みモデルに保持される。訓練済みのモデルをシステムに組み込んで運用する段階を「運用フェーズ」と呼ぶ。

（２）機械学習システムの留意点

機械学習システムは、入出力関係が未知の問題でも、膨大な訓練データを与えることで、一定の答えを出すことができるという点に強みがある。もっとも、こうした強みと表裏の関係にあるものとして、セキュリティ対策や品質管理の難しさにつながる次のような特徴も持つ。

---

[3] 図中の機械学習システムは、入力と出力の組$\{u, f(u)\}$から、これらの関係を学ぶ「教師あり機械学習（supervised learning）」を想定している。このほか、出力$f(u)$に対応する訓練データが存在ない問題を扱う「教師なし機械学習（unsupervised learning）」や、与えられた環境の中で機械学習モデルが自らデータを生み出す「強化学習（reinforcement learning）」がある。いずれの手法でも、機械学習モデルの情報処理ルールがデータに依存する点では共通している。

教師なし機械学習の代表的な応用例としては、未知のデータを自動的に分類するクラスタリングや、データの中から異常値を探し出す異常検知が挙げられる。

強化学習とは、ある環境の中で、エージェントが試行錯誤を通じて、より賢い行動を自律的に習得していく機械学習手法である。エージェントは、「行動（Action）」を選択した結果として環境から「報酬（Reward）」（または「利得」）を受け取ることを繰り返し、より良い行動選択の方法を学習していく。この手法は、行動習得の際に、必ずしも膨大なデータを要しない特殊性から、「教師あり学習」や「教師なし機械学習」とは区別される。強化学習は、ロボットの歩行動作獲得や囲碁を打つ AI ソフトなどに応用される。

[4] 開発者は、訓練データの構築、学習アルゴリズムやモデルの選択を通じて、間接的に機械学習モデルの訓練（機械学習モデルのパラメータ決定）に関与できる。

[5] 回帰や分類、予測、異常検知などの抽象化されたタスクに応用できる、さまざまなモデルが提案されている。例を挙げると、深層学習の基礎となる多層パーセプトロンや、決定木を多数束ねて学習を行うランダム・フォレスト・モデルが頻繁に利用される。

イ．機械学習システムへの要求の不明瞭さ

機械学習システムに解決が期待される課題では、現実世界の複雑かつ自然的な環境への対応を直接的に求められることから、入出力関係がどのような性質を満たすべきか、という情報システムへの要求が明瞭でない場合が多い。例えば、画像処理による自動運転を行う機械学習システムでは、運転で遭遇しうるあらゆる環境を入力したもとで、妥当な操作を出力することが求められる。このとき、入力される環境には、道路交通状況や天候の違いなどの現実世界のさまざまな自然的な要素や人間の作為などを映じて無限のバリエーションがある。これらの環境をすべて列挙することは不可能であり、情報システムの振舞いを定める入力の範囲を定めることができず、出力の妥当性を検証することも困難である。

ロ．訓練済みモデルの特性の不明瞭さ

情報システムへの要求が不明瞭な難しい課題に対処するには、表現力の高い機械学習モデルが必要となる。この場合、情報処理ルールは機械学習モデルの膨大な数のパラメータの集合として表現されるため、機械学習システムは次のような特徴を持つ。第 1 に、機械学習モデルのパラメータを人間が意味のある情報処理ルールとして解釈することが難しい。すなわち、機械学習モデルが上手くいく理由を説明できず、説明可能性が乏しい。第 2 に、機械学習モデルが、訓練データ以外の入力に対してどのように振る舞うかを事前に予想できない。第 3 に、機械学習モデルがもつ情報量が明確でない。機械学習モデルの訓練は、訓練データをモデルのパラメータに変換する操作とみなせるが、変換で保持される情報量は明確ではない[6]。

３．機械学習システムのセキュリティ・リスクの特性

（１）情報処理ルールに内在する脆弱性

機械学習システムの脆弱性のうち、情報処理ルール（機械学習モデルのパラメータ）に内在するものが考えられる[7]。実際、このような脆弱性を悪用する攻撃が研究により多数発見されており、実行可能であることが示されている（Kumar *et al.* [2019]）。

例を挙げると、カメラに映った顔映像で人物の識別を行う際に、ユーザ（攻撃者）が特殊な方法でデザインした眼鏡などの装着物を身に着けると、入力画像にノイズが加わり、機械学習システムが本来の人物であると認識できなくなるこ

---

[6] これらの 3 つの特徴が互いにどのような関係にあるかについても明確になっていない。

[7] 機械学習システム以外の情報システムにも共通して存在する脆弱性もある。例えば、アクセス権限の設定の不備やプログラムの誤りなどが挙げられる。これらについては、既に対処法が確立しているものも多い。

とがある（Sharif *et al.* [2017]）。これは、機械学習モデルが獲得した情報処理ルールの不備によるものであり、これを悪用してカメラによる識別を無効化する攻撃となる。

この例では、攻撃者は、システム設計上は許容される（許容せざるを得ない）権限のみを行使しており、伝統的なソフトウェアのバグなどがないにもかかわらず攻撃が成立してしまう。この点が、機械学習がもたらすセキュリティ・リスクの新しさであり、ソフトウェアのバグの修正、アクセス制御やユーザ権限の適切な管理といった従来のセキュリティ対策では対処が困難なケースが多い。

（２）セキュリティ対策が難しい理由

情報処理ルールに内在する脆弱性への対処は、主に、次の 3 つの要因によって容易でないと考えられる。

イ．脆弱性の洗出し

脆弱性を網羅的に洗い出すことが困難である。これは、2 節（2）イ．で述べたように、機械学習システムへの要求が不明瞭であり、ありうる入力やこれに対応する正しい出力すべてを考慮できないことによる。機械学習の脆弱性とそれに伴うセキュリティ・リスクについての研究は活発に行われているものの、脆弱性やセキュリティ・リスクの全容は明らかではなく、分類の方法も確立されていない。また、最新の研究動向を的確にフォローすること自体にも多大な労力を要する。

ロ．情報処理ルールの修正

情報処理ルールの修正方法の考案や、影響度の予測、妥当性検証が容易でない。機械学習モデルの説明可能性の乏しさから、情報処理ルールの脆弱性とモデルのパラメータとの関係を理解することが難しく、情報処理ルールの修正が機能に及ぼす影響を予測することは簡単でない。また、修正後の情報処理ルールが、システムへの要求を充足していることも保証できない。このため、情報処理ルールの修正は容易ではなく、機械学習システムのセキュリティ対策には不確実性が伴う[8]。

ハ．脆弱性と機能の分離

セキュリティ対策として行う情報処理ルールの修正は、システムの機能にも影響する。機械学習モデルを再訓練する場合には、パラメータ全体が変化する。パラメータの一部を変更する場合にも、すべてのパラメータの組合せによって

[8] 機械学習システムが完璧な精度を達成できず、機能面で不確実性を伴うことと同義である。

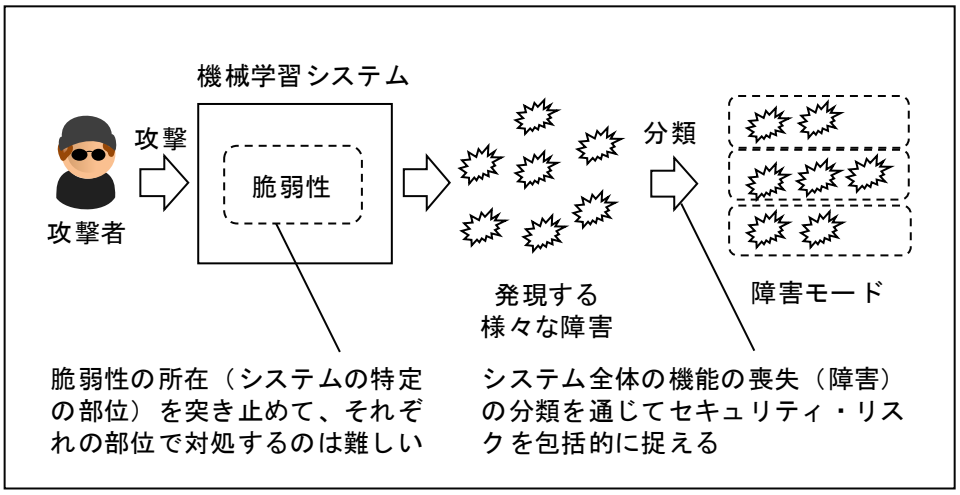
システムの機能が発現しているため、その影響はシステム全体の機能に及ぶ。したがって、機械学習システムでは、情報処理ルールに内在する脆弱性とシステムの機能を分離して考えることができない。セキュリティ対策としての情報処理ルールの修正を、システムの機能を維持しつつどう実施するかが課題となっている[9]。

４．機械学習の「障害モード」: Kumar らのアプローチ

（１）障害モードとは

機械学習システムの情報処理ルールに内在する脆弱性はシステムの機能と不可分である。したがって、セキュリティ・リスクも、個々の脆弱性がシステムのどの部位に存在するかを明らかにして対処する部分分解的なアプローチではなく、機能も含めたシステム全体へのアプローチでなければ捉えられないと考えられる。こうしたアプローチに該当する研究として Kumar *et al.* [2019] が挙げられる。Kumar *et al.* [2019] は、システム全体で発現している機能や保持すべき特

図 2 機械学習システムの障害モード



備考：攻撃がなくても障害は発生しうる。

[9] 機械学習システムのセキュリティ・リスクを捉える難しさは、CERT/CC から公表された脆弱性レポートにも表れている（脚注 2 を参照）。同レポートでは、勾配降下法を利用して学習する機械学習モデルに対して、意図的に誤った識別をさせる入力が作成可能であるとしている。通常では、脆弱性レポートは、既存かつ特定のシステムの脆弱性に関する情報を提供する。しかし、同レポートでは、いまだ開発されていないシステムも対象に含み、特定の機械学習システムのクラス全体に対しての注意喚起となっている点が注目されている。この背景は、セキュリティ対策を施すべきシステムの脆弱性を機能と分離して扱うことができないことから、本質的な対策を行うには、システムの機能の獲得を司る訓練方法自体の修正を要するためと考えられる。

性の喪失である「障害（failure）」に着目し、機械学習システムにおける障害に相当する研究事例などを収集して様態（mode）別に分類・リスト化した（図 2 を参照）。リスト化された「障害モード（failure modes）」は、機械学習のセキュリティ・リスクを、システムの部分（脆弱性）としてではなく、システム全体の現象である障害の分類によって包括的に捉えようとする試みの所産である[10]。

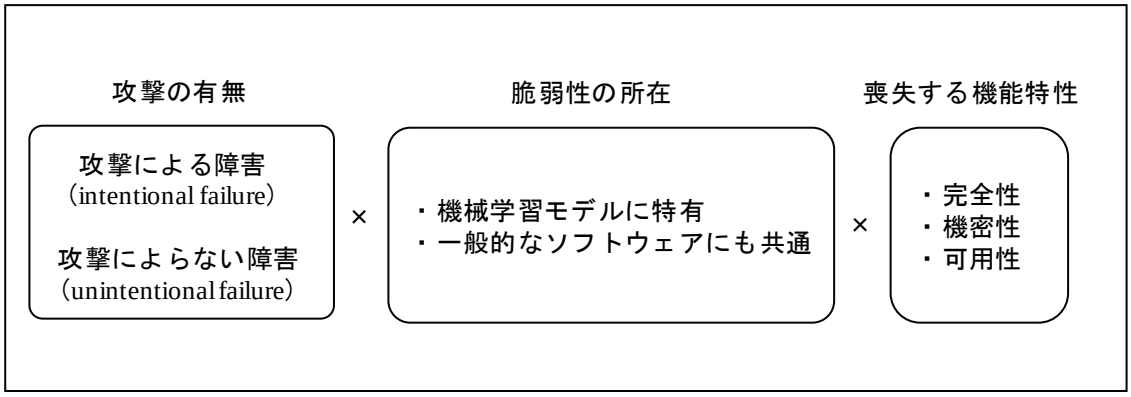
（２）障害モードの分類方法

Kumar *et al.* [2019] では、障害モードを以下の 3 つの軸によって分類している（図 3 を参照）。

イ．攻撃の有無

機械学習システムの障害は、障害を引き起こそうとする攻撃者の意図を原因とする「攻撃による障害（intentional failure）」と、システムの生得的なデザインに起因する「攻撃によらない障害（unintentional failure）」に大きく分けられる。前者の分類では、攻撃者の意図に応じて、（人物の識別誤りなど）モデルから不正な出力を得る攻撃、（特定の条件下で誤作動するなど）情報処理ルールを有害なものに変更する攻撃、（秘匿された訓練データまたはビジネス価値をもつモデル自体の）情報を窃取する攻撃などに分けられる。後者は、攻撃などのシステム外部からの干渉がなくても、訓練の過程で自律的に不安定化するケースに主眼が置かれる。訓練の実行過程の不備や、単純なソフトウェアのバグによる障害は含まない。そのため、外部からの訓練データの入力を必ずしも必要とせず、自律的にデータを生成することで学習を進める強化学習における障害の事例が多い。

図 3　障害モードを分類する 3 つの軸



[10] なお、Kumar *et al.* [2019] の障害モードのリストや研究事例などは、オープンエンドで現在も更新され続けており、最新の研究動向を容易に把握できるように配慮されている。本稿は、執筆時点（2020 年 7 月末）で入手したレポートに基づいている。

ロ．脆弱性の所在

機械学習システムの障害は、脆弱性の所在に応じて便宜的に以下の 2 つに分類できる。1 つ目は、3 節で論じたような、機械学習に特有の脆弱性（情報処理ルールに内在する脆弱性）のみに起因する障害である。2 つ目は、それらに加えて、機械学習を利用しない一般的なソフトウェアと同様の理由で生じた脆弱性にも関連する障害である。近年ではネットワークを介して、機械学習をプラットフォーム上で提供する MLaaS（Machine Learning as a Service）が普及した。このような大規模サービスは、オープン・ソースのプログラムを利用して開発されることから、伝統的なソフトウェアのバグや悪意のあるプログラムが機械学習モデルに混入する可能性がある。これら自体は機械学習に特有の脆弱性ではないが、現実に運用される機械学習システムのセキュリティ・リスクを包括的に捉える際には無視できない。特に、これらと機械学習に特有の脆弱性の組合せにより、機械学習システムに対して新しい様態の攻撃が可能となる点には注意が必要である。こうしたことを踏まえて、脆弱性の所在に応じて障害を分類する。

ハ．喪失する機能特性

情報セキュリティの原則は、CIA の 3 要素、すなわち機密性（Confidentiality）、完全性（Integrity）、可用性（Availability）を保持することである。システムへの攻撃では、これらの特性が喪失する可能性があり、喪失しうる特性に応じて障害を分類することができる。機械学習システムにおいても、この概念的フレームワークは有効である。ここで、機密性は、データが無権限者に対して開示されない特性を表す。完全性は、システムが仕様通りに動作する特性を表す。可用性は、ユーザの要求に応じてシステムが利用可能な状態にある特性を表す。

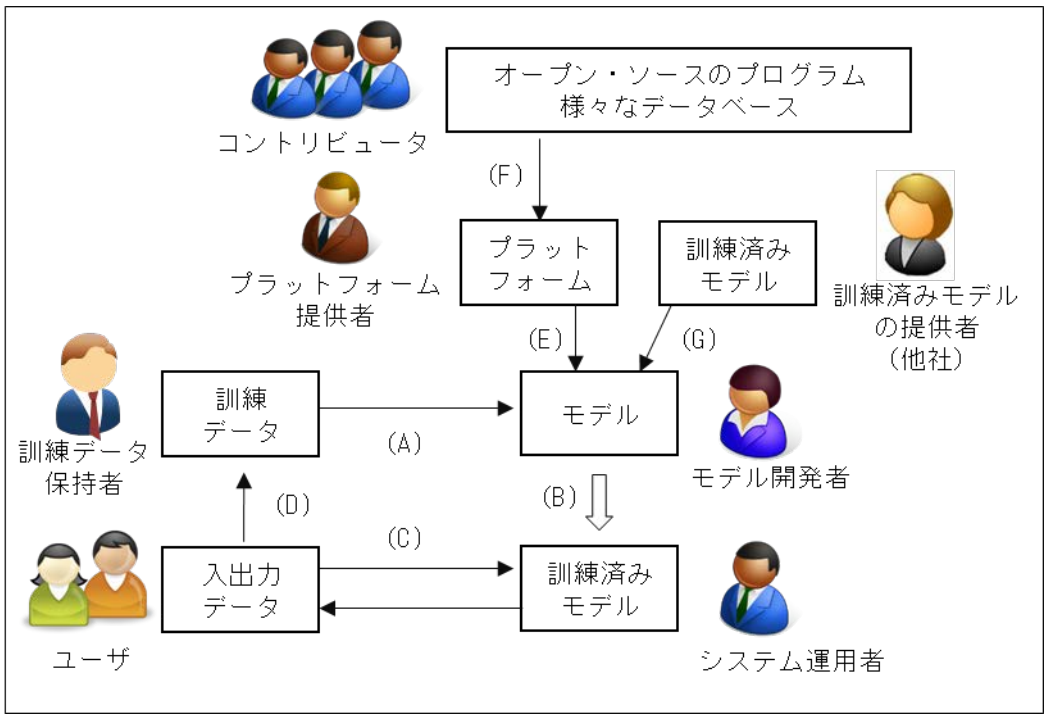
５．機械学習を組み込んだサービスの提供に関わる主体とその関係

次節でみるように、機械学習の障害は攻撃によるものが多い。そのため、攻撃が行われる状況の確認のために、機械学習サービスの開発と提供に関わる主体間の情報の流れを模式図にした（図 4 を参照）[11]。

最も単純な機械学習システムの提供では、（A）モデル開発者が訓練データを訓練データ保持者から受け取り、モデルの訓練を行う。（B）訓練済みのモデルは運用環境に配置され、（C）ユーザからの要求に応じてシステム運用者によっ

[11] 本稿では、次節でみる攻撃手法と、それらが実行される状況（模式図の場所）との個別の対応付けまでは行っていない。こうした状況は複数あると考えられ、例えば、後述するデータ・ポイズニング攻撃（6 節（１）イ．（ロ）a. を参照）では、訓練データの作成・保管・流通の過程（図 4 中の A、D、E、F など）のすべてにおいて攻撃が成立しうると考えられる。それぞれの攻撃について、攻撃が成立する条件（攻撃者の能力の前提、攻撃者の立場、攻撃が実行される状況）の列挙やそれぞれの条件下でのセキュリティ対策を明らかにすることは今後の課題である。

図 4　機械学習を組み込んだサービス提供の模式図



てサービスが提供される。(D) ユーザがサービスを利用した結果、新たに生み出されたデータは、新たな訓練データとして蓄積されていく。

モデル開発者は、外部リソースを利用して、(E) プラットフォーム上で機械学習システムを開発する場合がある。(F) 巨大なプラットフォームは、オープン・ソースのプログラムなどを利用して開発される。また、(G) 大規模な計算資源を駆使して（他社の）訓練済みモデルを、自社のモデルの一部として利用することもある[12]。

６．Kumar らによる障害モードのリスト

本節では、Kumar *et al.* [2019] の障害モードのリストを紹介する。Kumar らは、障害モードを攻撃による障害（11 項目）と攻撃によらない障害（6 項目）に大別している。このうち、前者については、各障害の関係性を把握しやすくするために、以下のとおり 2 つの階層を分けて整理することとする。すなわち、まず、機

12　転移学習（transfer learning）は、こうしたケースに含まれる。転移学習は、ある領域で学習させたモデルを別の領域に適用する技術である。深層学習などの大量のデータがなければ利用できない手法を、少量のデータのみ利用できる問題にも適用する道を開くことから、注目されている。例を挙げると、他社が訓練し、公開している画像認識を行う深層学習モデルの一部を、自社が開発するモデルに組み込み、そのモデルの一部のパラメータのみ訓練を行う場合が該当する。転移学習の脆弱性については、6 節（1）ロ.（イ）a.「バックドアの設置」を参照。

械学習に特有の脆弱性のみに関連すると考えられるもの（7 項目）と一般的なソフトウェアの脆弱性にも関連すると考えられるもの（4 項目）に分けたうえで、これらをそれぞれ完全性・機密性・可用性の観点からさらに細分化して紹介する。

【障害モードによる分類】
（１）攻撃による障害（11 項目）
　　イ．機械学習に特有の脆弱性のみに関連する攻撃（7 項目）
　　（イ）完全性に対する攻撃１＜入力データの変更＞
　　（ロ）完全性に対する攻撃２＜情報処理ルールの変更＞
　　（ハ）機密性に対する攻撃
　　ロ．一般的なソフトウェアの脆弱性にも関連する攻撃（4 項目）
　　（イ）完全性に対する攻撃
　　（ロ）機密性に対する攻撃
　　（ハ）完全性・機密性・可用性に対する攻撃
（２）攻撃によらない障害（6 項目）

（１）攻撃による障害

攻撃による障害の研究は、機械学習システムの脆弱性の存在を明らかにすることを目的とするものが多く、必ずしも実際の攻撃に悪用される蓋然性の高さを指摘するものではない。それぞれの障害モードについてセキュリティ・リスクの重要度[13]を評価するには、本節で紹介する情報に加えて、機械学習システムの運用目的や運用環境、攻撃手法ごとに異なる攻撃者の能力の前提[14]などを考慮する必要がある点に留意されたい。

イ．機械学習に特有の脆弱性のみに関連する攻撃

機械学習に特有とみられる脆弱性のみに関連する攻撃は、完全性・機密性・可用性の観点から、①訓練済みモデルへの入力データの変更によるもの（完全性に対する攻撃１）と②情報処理ルールの変更を企図するもの（完全性に対する攻撃２）、③機械学習システムで取り扱われるデータの機密性を損なうもの（機密性に対する攻撃）に分けることができる。

13　マイクロソフト社の「bug bar」では、各障害モードについて、重要度の評価の情報も提供されている（https://docs.microsoft.com/en-us/security/engineering/bug-bar-aiml）。

14　本節では攻撃者の立場を述べるにとどめ、攻撃者の能力の前提については詳しく立ち入らない。一般に、攻撃者がモデル内部の情報を利用できる前提を置かなければ成立しない攻撃をホワイトボックス型、そうでないものをブラックボックス型と呼ぶ。攻撃者の能力の前提は、攻撃手法により異なり、例えば同じく 6 節（１）イ．（イ）a. の摂動攻撃に分類される手法の中でも、ブラックボックス型とホワイトボックス型の両方のタイプが存在する。

（イ）完全性に対する攻撃１＜入力データの変更＞

入力データの変更による攻撃として、摂動攻撃（Perturbation Attack）と物理ドメインにおける敵対的サンプルの作成（Adversarial Example in Physical Domain）が挙げられている。これらの概要と研究事例は次のとおりである。

a. 摂動攻撃

障害モード	入力データにノイズ（摂動）が加わることによって、学習済みモデルの出力が所望のものに変更される。
攻撃の形態[15]	攻撃者は、システムを利用する際の入力データにノイズを加える。

【研究事例】

- 深層学習モデルを用いた画像認識において、入力データである皮膚の画像に目視では判別できない微小なノイズを加えることで、正常な部位を異常と判定させることができる。同様の手法により、脳のMRI 画像にノイズを加えることで、脳の領域を正しくセグメンテーションできなくなる（Paschali *et al.* [2018]）。
- 深層学習モデルを用いたテキスト翻訳において、入力データである文字列を僅かに操作することで、翻訳後のテキストから特定の単語を除去または変更するといった高度な操作ができる（Ebrahimi, Lowd and Dou [2018]）。
- 深層学習モデルを用いた音声認識において、入力データである音声に知覚できない小さなノイズを加えることで、任意のテキストを生成させることができる（Carlini and Wagner [2018]）。

b. 物理ドメインにおける敵対的サンプルの作成

障害モード	機械学習モデルを欺くよう設計された物体などに関するデータが入力されると、機械学習システムが誤作動する。
攻撃の形態	攻撃者は、機械学習モデルを欺く物体などを作成し、特定の場所にそれを設置する。

【研究事例】

- 深層学習を使った画像認識システムに対して亀の形の 3D 物体に特

[15] 攻撃の形態は、脆弱性が実際に攻撃に悪用される状況に関する本稿の筆者の想定である（以下同様）。

殊なテクスチャを張り付けることで、さまざまな角度から撮影した亀形の物体画像をライフルであると誤認させることができる（Athalye *et al.* [2018]）。これにより、危険物を検出するシステムを混乱させる。

- 人目を引かない特殊なテクスチャを施したサングラスを装着すると、画像認識システムが人物を正しく判別できなくなる（Sharif *et al.* [2017]）。人間と機械の両方のセキュリティ・チェックをすり抜ける装着物が作成できてしまう。

（ロ）完全性に対する攻撃２＜情報処理ルールの変更＞

情報処理ルールの変更による攻撃として、データ・ポイズニング攻撃（Poisoning Attack）とモデルの転用（Reprogramming ML System）が挙げられている。これらの概要と研究事例は次のとおりである。

a. データ・ポイズニング攻撃

障害モード	不正なデータが訓練データに追加されることで、訓練済みモデルの出力が所望のものに変更される。
攻撃の形態	攻撃者は、訓練データ保持者が管理している訓練データに不正なデータを追加する。または、オンライン学習を採用する機械学習システムでは、攻撃者はユーザの立場で不正なデータを生成し、訓練データに追加させる。

【研究事例】

- 一部のチャットボットにおいて、ユーザとの会話のログをフィードバックに用いるオンライン学習を採用したことにより、複数のユーザとの会話を通じて不適切な表現を出力するようになった（Lee [2016]）。
- 患者に対する抗凝固剤の投与量を予測するシステム（Lasso 回帰モデルを使用）において、患者の属性などに関する訓練データセットに全体の約 8%に相当する分量の不正なデータを追加することで、半数の患者への投与量を 75%も変化させた（Jagielski *et al.* [2018]）。

b. **モデルの転用**

障害モード	ある課題 X のために訓練されたモデルが、再度訓練されることなく、別の課題 Y のためのモデルに転用される。 ―― 通常の入力では課題 X のためのモデルとして機能するが、入力データにノイズを加えた場合のみ、課題 Y のためのモデルとして機能する。
攻撃の形態	攻撃者は、モデル開発者の立場で、訓練実行時に 2 つの課題に対応したモデルを作成する。具体的には、①訓練データに（単一の）特殊なノイズを加えたうえで、ある課題 X のためにモデルを訓練する。②これと同時に、（課題 X のための）モデルの出力を、課題 Y に対する出力に変換する変換器も作成する。 攻撃者は、運用フェーズにおいて、ユーザの立場で入力データにノイズを加える。次に、モデルから得られた出力を変換器に通すことによって、本来の目的とは異なる別のタスクを実行する。

【研究事例】

- 訓練データに単一のノイズを加えて、画像の分類を行う深層学習モデル「ImageNet」を訓練する。その後、入力画像に小さな四角形が複数個含まれる小さな画像を埋め込み、そのモデルに入力すると、画像中の四角形を数え上げるという全く異なる課題をこなすシステムとして機能する（Elsayed, Goodfellow, and Sohl-Dickstein [2018]）[16]。

（ハ）機密性に対する攻撃

機密性を損なう攻撃として、訓練データの逆算（Model Inversion）、メンバーシップ推定（Membership Inference）、モデルの窃取（Model Stealing）が挙げられている。

[16] この研究の手法は、機械学習システムへの攻撃の技術的な可能性を示す研究目的で開発されたものであり、現実的な脅威の程度は不明である。Elsayed, Goodfellow, and Sohl-Dickstein [2018] では、計算資源の窃取などに悪用される例が挙げられている。

a. 訓練データの逆算

障害モード	訓練データの一部（人物名などの項目値）と訓練済みモデルから、秘匿された訓練データや特徴量を逆算する。
攻撃の形態	攻撃者は、訓練データの一部を保有し、ユーザの立場でモデルへのアクセスを繰り返して訓練データや特徴量を推定する。

【研究事例】

- 深層学習モデルを用いた顔認証システムにおいて、個人の名前と学習済みモデルへのアクセスのみから、個人の顔画像をある程度再現した[17]。また、ライフスタイルに関するアンケート調査を学習させた決定木モデルから、訓練データに含まれる「他人を騙したことがあるか」といったセンシティブかつプライバシーに関わる質問への回答を、偽陽性（実際には「いいえ」と回答した人が「はい」と回答したと誤判定されること）なく推定した（Fredrikson, Jha, and Ristenpart [2015]）。

b. メンバーシップ推定

障害モード	あるデータ・レコードが訓練データに含まれているか否かをモデルから推定する。
攻撃の形態	攻撃者は、データ・レコードの候補を保有し、ユーザの立場でモデルへのアクセスを繰り返して、データ・レコードが訓練データに含まれているか否かを推定する。

【研究事例】

- 医療機関の滞在歴のデータベースを学習させたモデルに攻撃を行い、患者の属性情報（性別、年齢など）から、当該患者の治療内容に関する分類項目（訓練データに含まれる）を推定した（Shokri *et al.* [2017]）。

[17] 顔画像の再現度合を検証するための実験として、学習済みモデルから再現された顔画像を被験者に見せたもとで、5 枚の顔写真の中から訓練データに含まれていた人物の顔画像（1 枚）を正しく選択できるかをテストしたところ、95%の正答率で個人を特定することができた。

c. モデルの窃取

障害モード	元のモデルと類似した挙動を示すモデルを作成（複製）する。
攻撃の形態	攻撃者は、ユーザの立場で、モデルへのアクセスを繰り返し、モデルの情報を得る。

【研究事例】

- 一部の信用リスクのデータベースのデータを用いて訓練した決定木モデル（信用スコアを出力）を作成した後、攻撃者の立場でそのモデルを上記手法によって複製できることを示した。こうした攻撃は、ネットワーク経由でアクセス可能な一部の機械学習サービスにおいて有効であることが確認された（Tramèr *et al.* [2016]）。

ロ．一般的なソフトウェアの脆弱性にも関連する攻撃

機械学習だけでなく一般的なソフトウェアの脆弱性にも関連する攻撃については、①バックドアの設置（Backdoor ML。完全性に対する攻撃）とモデル・サプライ・チェーンの悪用（Attacking the ML Supply Chain）、②訓練データの窃取（Malicious ML Provider Recovering Training Data。機密性に対する攻撃）、そして、③完全性・機密性・可用性のいずれも損なう可能性がある攻撃として、伝統的なソフトウェアのバグ（Exploit Software Dependencies）に分けることができる。

（イ）完全性に対する攻撃

a. バックドアの設置[18]

障害モード	バックドアが仕込まれた学習済みモデルが、悪意のあるサード・パーティから提供される。バックドアは、特定の条件を満たすときのみ作動してモデルの誤動作を誘発し、それ以外は、モデルが正常に動作するように実装される。
攻撃の形態	モデルの訓練が他社にアウトソースされた場合[19]などに、攻撃者は、訓練済みモデルの提供者の立場で、バックドアを仕込んだモデルを作成して依頼者（モデル開発者）に提供する。

[18] バックドアの設置に関するサーベイ論文に Gao *et al.* [2020] がある。

[19] 例えば、画像処理による物体認識のモデルでは、訓練に膨大な計算資源を要するため、計算資源に乏しいモデル開発者はサード・パーティに訓練を依頼するインセンティブがある。

【研究事例】

- Gu, Dolan-Gavitt, and Garg [2019] では、道路標識に攻撃者が選んだ特定のステッカーを取り付けた場合にのみ、画像認識システムが標識の意味を誤認するようモデルを訓練することができることを示した。一例として、小さな黄色い四角のステッカーを交通標識に取り付けた場合にだけ、停止（stop）の道路標識を速度制限（speed limit）と誤認させることができる。しかも、特定のステッカーがない場合には、通常の高精度な画像認識システムと区別がつかない。

  こうしたバックドアは、転移学習を経ても有効に機能しうる。上記と同様の条件で発動するバックドアを埋め込んだモデルをベースに転移学習を行うと、特定のステッカーがある場合にのみ道路標識の認識精度が落ちる。
- 機械学習モデルのパラメータを一部変更することにより、入力データが特定の条件（トリガー）を満たすときのみ誤作動するモデルを作成できる。顔認証システムに適用すると、入力画像に小さな半透明の四角形のスタンプが付された場合のみ、特定の人物について判定を誤るモデルが作成できる。この攻撃は、音声認識や自動運転などのタスクを行う幅広いモデルに対して適用できる（Liu *et al.* [2017]）[20]。

b. モデル・サプライ・チェーンの悪用

障害モード	訓練済みモデルを提供するリポジトリなどが汚染される。不正なモデルがダウンロードされた場合、これを（そのまま、あるいは転移学習などに）利用したモデルが適切に動作しないなどの影響を受ける。
攻撃の形態	攻撃者は、リポジトリに不正なモデルを追加するか、既存のモデルを改竄する。または、モデルを提供するサーバーを侵害して、不正なモデルをダウンロードさせる。

【研究事例】

- 訓練済みモデルを提供するリポジトリ[21]（例えば、Caffe Model Zoo）からバックドアが設置されたモデルが提供され、オープン・ソース

[20] この文献は、Kumar *et al.* [2019] には引用されていない。

[21] データやプログラムなどを集積して一元的に保管する場所や、そうした場所を提供するサービスを指す。

で提供される機械学習フレームワーク（例えば、TensorFlow、Keras、Core ML、Theano、MXNet）に組み込まれうる。このようにモデル間の依存関係が悪用されると、広範なモデルが汚染される恐れがある。例えば、Caffe Model Zoo には、掲載されている訓練済みモデルに対するハッシュ値が当該モデルの作成時に付与されたものと異なるものや、ハッシュ値が付与されていないものが存在している。その場合、モデルのユーザは改竄を検出できない場合がある（Gu, Dolan-Gavitt, and Garg [2019]）。

（ロ）機密性に対する攻撃：訓練データの窃取

障害モード	悪意あるプラットフォーム提供者などが、訓練に利用された非公開のデータを窃取する。
攻撃の形態	機械学習システムの提供が他社（プラットフォーム提供者）などにアウトソースされた状況で、プラットフォーム提供者などが、訓練済みモデルから訓練データを窃取できる機能を埋め込む。

（ハ）完全性・機密性・可用性に対する攻撃：伝統的なソフトウェアのバグ

障害モード	機械学習システムが依存しているオープン・ソースのプログラムを介して、システムに伝統的なバグが混入する。
攻撃の形態	攻撃者は、数値計算ライブラリなど、機械学習システムの基礎を支えるプログラムのバグを発見または放置する、あるいは、バグを意図的に埋め込むことで、このプログラムを利用する機械学習システムでバグを悪用する。

【研究事例】

- 無限ループを引き起こすバグを悪用して、機械学習モデルによるサービスの稼働を妨げる攻撃（Denial-of-Service）が可能になる。訓練済みモデルを動作させるソフトウェアにバッファ・オーバーフローの脆弱性が存在した場合、メモリ管理領域を超えるサイズのデータを入力するなどして脆弱性を悪用し、権限昇格などを誘発してモデルの出力の上書きやソフトウェア制御の乗っ取りを行うことができる可能性がある。一例として、機械学習モデルによる正常な

クラス分類を妨げる攻撃（evasion attack）が可能になる（Xiao *et al.* [2018]）。

（２）攻撃によらない障害

機械学習システムにおける攻撃によらない障害として、①利得の誤謬（Reward Hacking）、②副次的効果（Side Effect）、③入力データの分布の変化（Distributional Shifts）、④敵対的サンプルの自動収集（Natural Adversarial Examples）、⑤普遍的な変化（Common Corruption）、⑥現実的な環境に対する不十分なテスト（Incomplete Testing）が挙げられている。

イ．利得の誤謬

障害モード	開発者の意図に整合的な利得と実装された利得とのミスマッチから、強化学習システムが意図せざる挙動を示す[22]。

ロ．副次的効果

障害モード	強化学習システムが開発者により設定された目的を達成する際に、外部環境に副作用や負の影響を与える。

【研究事例】

- あるロボットが物体を移動させる課題を与えられたときに、移動経路上にある水の入った花瓶を倒してしまうといったシナリオが考えられる（Amodei *et al.* [2016]）。

ハ．入力データの分布の変化

障害モード	訓練された環境と異なるテスト環境において、機械学習システムが入力データの確率分布の違いに対応できないことから安定的に動作しない。

【研究事例】

- 強化学習手法である Rainbow DQN と A2C を利用して、スタート地点からゴール地点まで溶岩の領域を避けて到達するように訓練されたエージェントが、溶岩の位置を変化させたテスト環境では不安定な挙

[22] 強化学習システムの利用例として、強化学習などの AI が組み込まれたゲームが収集され、リスト化されている（https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml）。

動を示し、溶岩を避けることができなかった（Leike *et al.* [2017]）。

二．敵対的サンプルの自動収集

障害モード	機械学習システムが判別を誤るような難しいサンプルを自動的に収集する。こうしたサンプルを単純に入力するだけで、システムが不安定化する。

【研究事例】

- 機械学習モデルの訓練の際に、判別が難しいサンプルを自動抽出するHEM（Hard Example Mining）[23]を使って精度を上げる手法がある。もっとも、HEM の設定が不適切である場合、機械学習システムが判別を誤るような難しい例を過剰に自動収集してしまうことがある（Gilmer *et al.* [2018]）。

ホ．普遍的な変化

障害モード	画像における物体の傾きや縮尺、ノイズといった普遍的にみられる入力データの変化に機械学習システムがうまく対応できない。

【研究事例】

- 画像に生じる普遍的な変化（明暗、コントラスト、霧、ノイズ、縮尺、傾きなど）によって、機械学習システムによる画像識別の精度が大幅に低下する（Hendrycks and Dietterich [2019]）。

ヘ．現実的な環境に対する不十分なテスト

障害モード	現実的な環境のもとでの可用性が十分に検証されておらず、本番運用で機械学習システムがうまく動作しない。

【研究事例】

- 機械学習アルゴリズムの頑健性を高めても、現実的な環境のもとでう

[23] HEM は、モデルによる判別が簡単なデータが大量に存在する一方、判別が難しいデータ（hard examples）が少量しかない場合に、後者を収集して重点的に学習することにより、判別の精度を高める目的で利用される。HEM は、小さな訓練データからスタートして機械学習モデルを訓練し、判別を誤ったサンプルを訓練データに追加していくことで行われる。

まく機能しない場合がある（Gilmer *et al.* [2018]）。一例として、強風で倒れた「止まれ（stop）」の道路標識が認識されない場合がある。

７．結びに代えて：留意点と今後の課題

本節では、セキュリティ対策を含め機械学習システムを社会で利活用していく際の留意点や障害モードの有用性について考察する。

（１）セキュリティ対策に向けての留意点

第 1 に、機械学習システムの脆弱性やそれに伴うセキュリティ・リスクには解明されていない点も多いため、これらへの対策手法について最新の知見を収集することが重要である。障害モードのリストは脆弱性に関する情報を整理する際に有用であり、Kumar *et al.* [2019] で指摘されているように、脆弱性や障害に関する研究をタイムリーにフォローし障害モードのリストをアップデートしていくことが求められる。

第 2 に、セキュリティ対策を実施するにあたって、機械学習モデルの脆弱性の特性を理解しておくことが重要である。情報処理ルールに脆弱性が内在する場合、不用意にユーザに対して機械学習モデルへのアクセスを提供することが、完全性や機密性への攻撃につながる怖れがある。また、訓練データを汚染（データ・ポイズニング攻撃）されても、攻撃を検知することが難しい。こうした問題への対応策は今のところ確立されていない [24]が、少なくとも障害モードのリストに学び、機械学習システムのリスク特性を理解することにより、セキュリティ対策へのさまざまな示唆を得ることができる。

第 3 に、機械学習システムにサプライ・チェーンを介して脆弱性が入り込むリスクに留意する必要がある。近年の機械学習システムは巨大であり、システムのすべてを独自に開発することは困難である。ネットワーク依存性に起因するセキュリティ・リスクは新しいものではないが、機械学習システムに特有の脆弱性やセキュリティ・リスクは、これを悪用する攻撃の検知が難しいだけに、これらがシステムに流入しないよう注意を払うことが重要である。機械学習システムを他社と共同で開発する場合や、機械学習サービスのプラットフォームやオープン・アクセスのデータを利用する場合には、協業相手やデータ・ソースをどの程度信頼できるかを確認することが重要である。

第 4 に、機械学習システムのセキュリティ対策はシステムの他の機能にも影

[24] 攻撃手法によっては対策が提案されているものもある。例えば、画像認識を行う機械学習モデルにおける摂動攻撃に対しては、敵対的サンプルを生成して、これらを訓練データに含める敵対的学習（adversarial training）が提案されている。これにより、モデルの精度と敵対的サンプルへの耐性を同時に高めることができる（Xie *et al.* [2020]）。

響するため、情報セキュリティの専門家のみでは完結できない。セキュリティ対策では、影響を受けうる他の機能の開発者の協力が不可欠である。

第 5 に、機械学習システムのセキュリティ対策を手順を踏んで実施することが重要である。具体的には、機械学習システムに関して、既知の脆弱性とこれを悪用する攻撃をまず洗い出し、次に対策手法を検討することが重要である。その際、Kumar *et al.* [2019] も指摘するように、障害モードのリストは脆弱性と攻撃手法の洗い出しに有用である[25]。

（２）機械学習システムを利活用していくうえでの留意点

機械学習システムの提供者に加えて、利用者や規制当局にとっても、機械学習システムに特有のセキュリティ・リスクを理解することは重要である。機械学習システムを利用する際に、システムが予期しない動作をすることや、これによって損害が発生する可能性がある。その際、システムの提供者と利用者の双方が、機械学習システムの特性を理解しておくことは、リスク・コミュニケーションを円滑にする。この点、Kumar *et al.* [2019] が指摘するように、障害モードは、さまざまな当事者が機械学習システムの脆弱性やそれに伴うセキュリティ・リスクの特性を理解し、これについて対話するための基礎的な概念と共通言語を提供する点で有用である。機械学習システムの社会実装がより深化した場合には、当局による規制が望まれる可能性もある。こうした政策オプションを検討した例としては、Kumar *et al.* [2018] や Calo *et al.* [2018] がある。

（３）今後の課題

機械学習は本来的に帰納的なアプローチであり、これまでの演繹的な情報システムとは学習のパラダイムが異なる。今後、機械学習システムの脆弱性やそれに伴うセキュリティ・リスクに関する研究を深めるとともに、発見されたセキュリティ・リスクを軽減する手法の確立も求められる。金融機関を含め、機械学習を利用したサービスを提供する主体においては、機械学習システムの脆弱性に関する情報収集を継続的に行い、各社における同システムの運用目的・運用環境に照らしてリスクの重要度を評価し、セキュリティ対策を実施していく取り組みが重要になる。

以　上

---

[25] 例として、マイクロソフト社は、脆弱性の分類をもとに、情報システムのセキュリティ・リスクを洗い出すための「脅威モデル（threat model）」化手法を機械学習にも拡張し、本番稼働前に行う安全性の検証に利用している（https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml）。

【参考文献】

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, “Concrete Problems in AI Safety,” arXiv: 1606.06565, 2016.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, “Synthesizing Robust Adversarial Examples,” *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 284-293.

Calo, Ryan, Ivan Evtimov, Earlence Fernandes, Tadayoshi Kohno, and David O’Hair, “Is Tricking a Robot Hacking?” University of Washington School of Law Research Paper 2018-05, 2018 (available at SSRN: https://ssrn.com/abstract=3150530).

Carlini, Nicholas, and David Wagner, “Audio Adversarial Examples: Targeted Attacks on Speech-to-Text,” arXiv: 1801.01944, 2018.

Ebrahimi, Javid, Daniel Lowd, and Dejing Dou, “On Adversarial Examples for Character-Level Neural Machine Translation,” arXiv: 1806.09030, 2018.

Elsayed, F. Gamaleldin, Ian Goodfellow, and Jascha Sohl-Dickstein, “Adversarial Reprogramming of Neural Networks,” arXiv: 1806.11146, 2018.

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart, “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures,” *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS) 2015*, Association for Computing Machinery, 2015, pp. 1322-1333.

Gao, Yansong, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim, “Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review,” arXiv: 2007.10760, 2020.

Gilmer, Justin, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl, “Motivating the Rules of the Game for Adversarial Example Research,” arXiv: 1807.06732, 2018.

Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg, “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,” arXiv: 1708.06733, 2019.

Hendrycks, Dan, and Thomas Dietterich, “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations,” arXiv: 1903.12261, 2019.

Jagielski, Matthew, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li, “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning,” arXiv: 1804.00308, 2018.

Kumar, Ram Shankar Siva, David O’Brien, Kendra Albert, and Salome Vijoen, “Law

and Adversarial Machine Learning," arXiv: 1810.10731, 2018.

———, ———, ———, ———, and Jeffrey Snover, "Failure Modes in Machine Learning," arXiv: 1911.11034, 2019.

Lee, Peter, "Learning from Tay's Introduction," Official Microsoft Blog, March 25 2016 (available at https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction、2020 年 11 月 27 日).

Leike, Jan, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg, "AI Safety Gridworlds," arXiv: 1711.09883, 2017.

Liu, Yingqi, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang, "Trojaning Attack on Neural Networks," Department of Computer Science Technical Reports, Paper 1781, Purdue University, 2017 (available at https://docs.lib.purdue.edu/cstech/1781、2020 年 9 月 23 日).

Paschali, Magdalini, Sailesh Conjeti, Fernando Navarro, and Nassir Navab, "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging," arXiv: 1804.00504, 2018.

Sharif, Mahmood, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter, "Adversarial Generative Nets: Neural Network Attacks on State-of-the Art Face Recognition," arXiv: 1801.00349v1, 2017.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, "Membership Inference Attacks Against Machine Learning Models," *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, IEEE, 2017, pp. 3-18.

Tramèr, Florian, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart, "Stealing Machine Learning Models via Prediction APIs," *Proceedings of USENIX Security Symposium*, Advanced Computing Systems Association, 2016, pp. 601-618.

Xiao, Qixue, Kang Li, Deyue Zhang, and Weilin Xu, "Security Risks in Deep Learning Implementations," *Proceedings of 2018 IEEE Security and Privacy Workshops*, IEEE, 2018, pp. 123-128.

Xie, Cihang, Mingxing Tan, Boqing Gong, Jiang Wang, Alan Yuille, and Quoc V. Le, "Adversarial Examples Improve Image Recognition," arXiv: 1911.09665, 2020.