AIを護る：複製攻撃とその防御に関する研究動向

東京理科大学 工学部 情報工学科 准教授 中村（なかむら） 和晃（かずあき）

■ はじめに

2010年代半ばより「人工知能（Artificial Intelligence；AI）」という言葉が社会的な注目を集め，ニュースなどで取り上げられる機会も多くなりました．Googleや Amazonといった巨大IT企業のみならず，中小と呼ばれるような企業でもAIの利活用が検討され，それに呼応するように，AI関連スキル教育の強化が叫ばれています．現代のAIは機械学習を基盤としていますが，この「機械学習（Machine Learning；ML）」について解説する講座やWebサイトも乱立しており，いまや，「AI」「機械学習」は専門用語ではなく日常生活用語の域に達した感さえあります．

上記のようなAIの爆発的な普及は，1990年代から2000年代初頭にかけてのインターネットの普及に重なる面があります．当時，ホームページの一般化とともに画像，音声，文書など様々な電子コンテンツがインターネット上に蓄積され，そのようなコンテンツの作成スキルを解説する書籍やWebサイトが多数登場しました．しかし，それと同時に，電子コンテンツの不正複製（海賊版）の問題も浮上し深刻化の一途を辿りました．それが契機となり，不正複製に対する技術的防御策の研究開発や法的対策の整備が進められることともなりました．

歴史は繰り返す……のかどうか，現時点では何とも言えませんが，現在のAIに対しても，その普及とともに不正複製の問題が浮上・深刻化する未来は十分に想定されると言えるでしょう．高度なAIがあれば，その使用権の貸与・販売や，あるいは自社での独占利用に伴う技術的優位性の確保により，大きな利益を上げられる可能性があります．その意味で，AIには莫大な資産価値があり，一種の知的財産と解釈することも可能です．それは，悪意ある者にとっては，不正複製の対象としても魅力的であることを意味します．

本稿では，AIに対する不正複製攻撃の問題と，その対策に向けた取り組みについて，筆者の知る範囲で紹介させて頂きたく思います．

■ AIとは？ 機械学習とは？

AIの不正複製について言及する前に，そもそも「AIとは何か？」という点を論じておきたく思います．

皆様が「AI」と聞いてまずイメージするのはどのようなものでしょうか．1982年生まれの筆者と同世代の方であれば（そうでなくとも？），ドラえもんや「ドラゴンボール」の人造人間，「クロノトリガー」のロボなど，人間と同じように考え自律的に行動するロボットを思い浮かべる方も少なくないものと思われます．しかし，多くの方が実感しておられる通り，現在の，日々のニュースで取り上げられるような「AI」は，そのようなロボットのイメージとは明らかに異なります．

いま実用化が進みつつある医療診断AIや自動翻訳AI，あるいはテキストの読み上げAI（AIアナウンサー）といったAIは，総じて，ある種の「関数」であると言えます．二次関数 $y=x^2$ は，$x=3$ を入力すると $y=9$ を，$x=-4$ を入力すると $y=16$ を，といった具合に，入力に応じた出力を返します．現代のAIも，本質的にはこれと同じです．例えば医療診断AIについて考えてみると，これは，眼底画像や胸部CT画像などを入力として受け取り，そのパターンに応じて病気の有無や具体的な病名を出力する関数と言えます．自動翻訳AIは，入力の日本語テキストを英語テキストへと変換し，それを出力する関数と言えます．テキスト読み上げAIは，同じく日本語テキストを入力し，それを発話音声へと変換して出力する関数となります．昨今大きな話題となったGoogleの画像生成AI「Imagen」なども同様で，ユーザが指定した説明文を入力として受け取り，その内容に適合した合成画像を出力する関数と解釈できます．

こうした「関数」は，具体的にはどのようにして実現されるのでしょうか．実は，そのための手段が機械学習です．機械学習の問題設定は，「二次関数 $y=f(x)$ が3つの点 $(x, y)=(0, 1), (1, 0), (3, 10)$ を通るとき，その式を具体的に求めよ」という高校数学の問題に似ています．この問題の解は，$f(x)=ax^2+bx+c$ とおいた上で3点の座標を実際に代入し，連立方程式

を解くことにより求まります（ちなみに答えは $f(x)=2x^2-3x+1$）．そして，いちど解が求まると，$x=0, 1, 3$ だけに限らず，任意の x について，それに対応する y を計算できるようになります．AI も同様で，例えば自動翻訳 AI を実現するためには，日本語と英語の対訳文を事例としてまず収集します．これは，上記の二次関数の例で言えば $(x, y)=(0, 1)$ や $(x, y)=(1, 0)$ といった「点」を知ることに相当します．次に，収集した対訳文が再現されるように自動翻訳 AI をチューニングします．これは，二次関数の例で言えばパラメータ a, b, c の値を求めることに相当し，この部分こそがまさに機械学習です．こうして「学習」が完了した後，AI は最初に集めた事例（これを「学習データ」「訓練データ」「教師データ」などと言います）に限らず，任意の日本語文を訳せるようになる，というわけです．正確には「任意」とは言えないのですが，学習データの数や多様性が十分であれば，それに近いものとなり得ます．

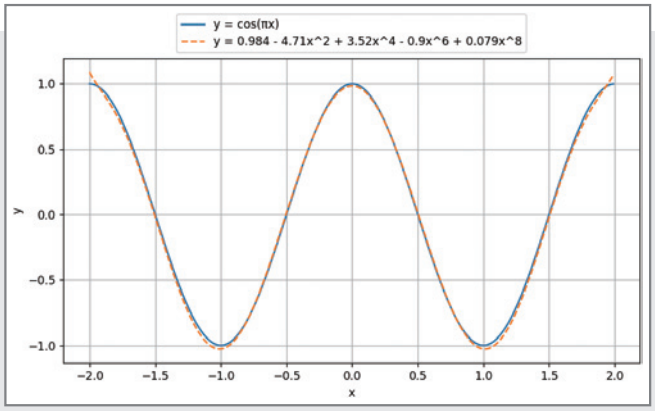
■ AIにおける「複製」の意味

AI と言えど，電子的な実体は有しており，画像や音声などの電子コンテンツと同様，電子データ（0 と 1 からなるビット列）として存在していることに変わりはありません．しかし，上述の通り，AI とは関数なので，その複製は，電子コンテンツの複製とは意味合いが大きく異なります．電子コンテンツにおいて複製とは，その実体である電子データの複製を指します．これに対して，AI の複製とは，関数としての機能の複製を意味します．実体としては全く異なるにも関わらず，関数（機能）としては全く同じ，あるいは極めて類似した AI を設計することが可能なのです．

余弦関数 $y=\cos(\pi x)$ と多項式関数 $y=0.984-4.71x^2+3.52x^4-0.9x^6+0.079x^8$ は，当然ながら別の関数です．しかし，$-2\leq x\leq 2$ において，この両者は極めて類似した挙動を示します【図 1】．これと同様，AI においても，例えば，2 つの医療診断 AI が，電子データとしては（コンピュータプログラムとしての実体は）別であるにも関わらず出力される診断結果は常にほぼ同一となる，といったケースが起こり得ます．

■ AIを不正に複製する攻撃：Model Stealing（MS）

上記のような複製が現実に作成され，それが問題を引き起こす状況は実際にあるのでしょうか．例えば次のような攻撃シナリオが想定されています．



【図 1】実体は別だが機能は類似している関数の例

Google Vision や Amazon Rekognition など，クラウド上で API（Application Programming Interface）を介して利用するタイプの有料の画像認識サービスが既に実用化されています．こうしたサービスでは，利用者から送信されてきた画像を独自の画像認識 AI で処理し，認識結果を返送します．このときに使用される画像認識 AI が複製攻撃を受ける対象となります．攻撃者は，あらかじめ多数の画像を準備しておき，それらを攻撃対象の画像認識サービスへと送信します．すると，サービス提供側から認識結果が返却されます．以上により，「画像」と「認識結果」のペアが多数得られたことになりますので，これらを学習データとする機械学習により新たな AI を作成することが攻撃者には可能となります．機械学習では，学習データとして与えた事例が再現されるように AI がチューニングされますので，新たに得られた AI は，サービス提供元の AI に極めて類似した機能を持つことになります．これはまさに，先述の意味での複製に他なりません．この攻撃を Model Stealing（MS）あるいは Model Extraction Attack（MEA）と呼びます[1,2,3]．また，攻撃の対象となった AI を犠牲者モデル（victim model），攻撃により新たに作成された AI をクローンモデル（clone model）と呼びます【図 2】．

以上に述べた MS の攻撃対象は画像認識 AI に限定されるわけではなく，様々な種類の AI が標的となり得ます．しかし，AI の導入が最も進んでいる分野の一つが画像処理であり，具体的な画像認識サービスも既に登場していることから，MS は画像認識 AI を対象に論じられることが一般的です．以下，本稿でもこれに倣うこととします．

改めて，MS の攻撃プロセスを数学的に定式化しておきます．まず，犠牲者モデルを f とすると，f は，画像 x に対し認識結果 $y=f(x)$ を返す関数です．この f に対し，攻撃者は画像集合 $D=\{\tilde{x}_i | i=1, 2, \cdots,$

$N\}$ を送信します．犠牲者モデルはこれらを処理し，認識結果 $\{\tilde{y}_i=f(\tilde{x}_i)\,|\,i=1, 2, \cdots, N\}$ を返送します．これにより得られた事例集合 $\{(\tilde{x}_i, \tilde{y}_i)\,|\,i=1, 2, \cdots, N\}$ を用いて，攻撃者はクローンモデル g を機械学習により作成します．この g もまた関数であり，実体としては f とは別物ですが，多くの x について $f(x)=g(x)$ を満たすものとなっており，機能としては f とほぼ同等となります．以上の攻撃プロセスを先ほどの初等関数の例に当てはめて説明すると，次のようになります．まず，犠牲者モデルは余弦関数 $y=f(x)=\cos(\pi x)$ に相当します．ただし，f が余弦関数であるということは公開されておらず，攻撃者はこの事実を知りません．一方で，攻撃者は $\tilde{x}_1=0, \tilde{x}_2=2, \tilde{x}_3=-1$, といったように有限個の値を犠牲者モデルに送信することは可能であり，それに伴って $f(0), f(2), f(-1)$ などの値を返値 $\tilde{y}_1, \tilde{y}_2, \tilde{y}_3$ として取得できます．これらの返値を参考に，適当な次数（例えば P 次）の多項式関数

$$g(x)=\sum_{k=0}^{P} a_k x^k$$

で f を再現すること，すなわち，可能な限り $f(x)=g(x)$ が満たされるようにパラメータ $a_0, a_1, \cdots, a_P$ を設定することが攻撃者の目標となります．

MS において，攻撃者は特に不正アクセスなどをしているわけではありません．画像を送信して認識結果を受け取る，というのはサービス提供側が想定する正規の利用法です．その範囲内でクローンモデルを作成することができるわけですが，AI の資産価値を考慮すると，MS は知的財産権の侵害に該当すると言えます．また，【図 2】にも示したように，クローンモデルが海賊版サービスのような形態で無料公開されれば，犠牲者モデルを運用する組織が経済的損失を被ったり，海賊版サービスへのアクセスを介して情報流出が発生したり，といったように，より具体的な問題が生ずる恐れもあります．

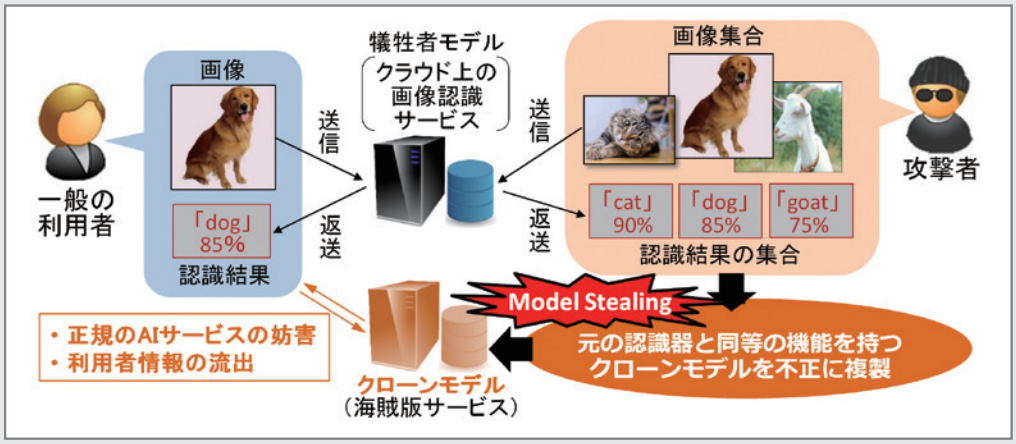
■ MS を目論む攻撃者側の技法

理屈の上では上述の通りなのですが，現実に MS を実行することはどの程度可能なのでしょうか．別の言い方をすれば，クローンモデル g は犠牲者モデル f とどの程度一致するものなのでしょうか．この点について知ることは，MS の脅威度を測る上で，ひいては防御法を考える上で，極めて有益な情報となるはずです．そこで，防御法に先立って，まず攻撃法に関する内容を紹介します．

MS の攻撃性能は一般に，クローンモデルと犠牲者モデルで同じ出力が得られる割合，すなわち $f(x)=g(x)$ となる確率により評価されます．これを以降では「出力一致率（agreement）」と呼びます．攻撃者としては出力一致率を高めることが目標であり，そのためには，攻撃に用いる画像集合 D に工夫が必要となります．具体的には，D に含まれる画像の多様性が重要になります．

対比として，余弦関数を多項式関数で再現したい場合を考えます．このとき，犠牲者モデルである余弦関数に $x=0$ 周辺の値ばかりを送信していれば，$x=0$ 付近の出力一致率（近似精度）は向上しますが，そこから遠ざかるほどに出力一致率は下がるはずです．できる限り広範囲の x について出力一致率を高めるためには，できる限り多様な値を $\tilde{x}_i$ として犠牲者モデルに送信することが重要となります．これと同じことが，AI を対象とした MS にも言えます．すなわち，攻撃者としては，可能な限り様々な画像を用意して画像集合 D の多様性を高めることが重要となります．そのための方法として，次のような合成画像の使用が有効であることが分かっています．なお，画像認識 AI を対象とした MS では，これらの技法の使用が出力一致率の向上のためにほぼ必須であり[4]，その前提の下で防御法が検討されます．

【color perturbation】画像は赤（R），緑（G），青（B）の三つの成分（チャンネル）からなりま



【図 2】Model Stealing の概要と問題点

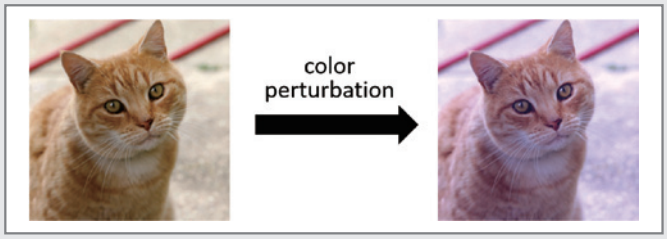
すが，これらの各チャンネルをランダムに強めたり弱めたりすることにより画像集合 D の多様性を高める手法です[5]．例を【図 3】に示します．

【mix-up】ランダムに選択した 2 枚の画像を重ね合わせることにより新たな画像を合成し，それにより画像集合の多様性を高める手法です[6]．例を【図 4】に示します．

■ MS に対する防御

ここからは，以上に述べた MS に対し，これを防御するための手法としてこれまでに研究されてきた内容を紹介したいと思います．

MS に対する防御法には，大きく分けて次の二種類があります．一つは，MS の攻撃プロセスそのものを妨げることによりクローンモデルの作成自体を未然に防ぐ，というものであり，事前防御に相当するものです．もう一つは，クローンモデルが作成されてしまった後にそれを検知して摘発するというものであり，事後防御に相当します．これらは，コンピュータウイルスに対する防御と対比するとイメージしやすいかと思います．コンピュータウイルスに対しては，そもそも感染しないように，アンチウイルスソフトをインストールしておいたり，怪しげな Web サイトには接続しないようにする，といった対策が取られます．これは言わば事前防御です．一方で，感染を 100% 完全に防ぐことは現実的ではないので，定期的にチェックを行い，感染が発覚した場合にはそれを駆除することも推奨されます．こちらが事後防御です．このように，二種類の防御法を組み合わせることにより万全な対策を実現することが理想ですが，これは MS に対する防御においても同様であり，事前・事後両面からの対策が検討され，それぞれ研究が進められています．



【図 3】color perturbation による合成画像の例

【図 4】mix-up による合成画像の例

■ 具体的な防御法①：攻撃プロセスの遮断

事前防御を実現するためのポイントは，犠牲者モデルにアクセスしてきた利用者に悪意があるかないかを判定することです．もし，悪意の有無が正確に判定可能であれば，悪意ある利用者に対して以降のアクセスを遮断することにより，MS の攻撃プロセスをそれ以上進行させないようにすることができます．

ここで，上述した攻撃技法の特性が重要な意味を持ちます．MS を目論む攻撃者は，クローンモデルの出力一致率を高めるため，color perturbation や mix-up などによる合成画像を犠牲者モデルに送信する公算が極めて高いと考えられます．これに対し，悪意を持たない一般の利用者は，所望の画像に関する認識結果を得ることが目的なので，合成画像を送信する動機がほとんどありません．このため，攻撃者と一般の利用者では，犠牲者モデルに送信する画像集合の統計的性質に違いが生じることになります．その違いを捉えることができれば，悪意の有無を判定可能となります．

以上の方針で MS への防御を試みた研究はいくつかありますが，その中で特に有名なもの[5]を以下に紹介します．この防御法では，ある利用者について，「その利用者が過去に送信した画像の中からランダムに二枚を選択し，その二枚の類似性（より正確にはユークリッド距離）を評価したときの類似度の値」の統計的傾向に着目します．上記の類似度は，たまたま似た画像が選択されれば高くなり，たまたま正反対の色を持つ画像が選択されれば低くなります．しかし，統計的には正規分布に従うはずです．すなわち，高くもなく低くもない値になる確率が高く，極端に高い値や低い値となる確率は低くなります．ところが，送信画像集合の中に color perturbation や mix-up などによる合成画像が多数混入していると，合成画像とその元となった画像には高い類似性が認められるため，極端に大きい類似度が得られる確率が不自然に高まります．

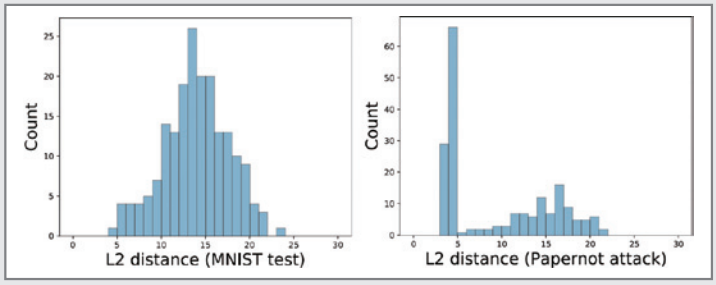
そこで，上記の類似度を多数の画像ペアについて計算した場合の分布を求め，それが正規分布からどの程度外れているかを評価します．その結果，大きく外れていることが判明した場合には，その利用者を悪意ある攻撃者と判定しアクセ

スを遮断します．

上記の防御法について，類似度の分布を実際に解析した結果を【図 5】に示します．これは文献[5]からの引用であり，左の分布が一般の利用者のもの，右の分布が攻撃者のものを示しています．一般の利用者では正規分布に近い結果となっていることが分かります．これに対し，攻撃者では，分布が左側に偏っており，正規分布から大きく外れています（右側ではなく左側に偏っている理由は，文献[5]で計算されている値が実際には画像間の類似度ではなく距離であるためです）．この分析結果は，上述の理論を裏付けるものです．

残念ながら，上述の事前防御には，いくつかの欠点もあります．第一に，ある利用者からの送信画像の統計的傾向（画像間類似度の分布など）を分析するためには，その利用者から十分な数の画像を受信していなければならない，という点です．これは，ある程度の枚数の画像を受け付けた後でなければ，その利用者に悪意があるか否かを判定できない，ということに他なりません．従って，攻撃者による AI の複製プロセスを完全には遮断できません．続いて，第二の欠点は，結託攻撃に脆弱であるという点です．結託攻撃とは，複数の攻撃者が結託して行う攻撃の総称であり，MS の場合では，攻撃に用いる画像集合 D を複数の攻撃者が分担して犠牲者モデルに送信する方法のことを言います．第一の欠点から，個々の攻撃者がアクセス遮断を受けるのは一定程度の枚数の画像を送信した後となりますので，複数の攻撃者が結託すれば，最終的には十分な量の画像を犠牲者モデルに送信することも可能となります．この欠点はかなり本質的であり，事前防御のみで解決することは困難です．このため，次に述べる事後防御との組み合わせが重要となります．

そのほかの欠点として，悪意の有無の誤判定という問題もありますが，この点については，攻撃者と一般の利用者のふるまいの違いをより詳細に解析するなどして，判定精度の向上を図ることにより解決することが望まれます．



【図 5】利用者ごとの画像間距離の分布
（左：一般の利用者，右：攻撃者．参考文献[5]の Fig. 5 より引用）

■ 具体的な防御法②：クローンモデルの検知

事後防御において重要なのは，二つの AI が互いに互いの複製でないにも関わらず似たような挙動を示すケースもあり得る（例えば，自動運転車の開発を目指す二つの企業が各々独立に道路標識画像の認識 AI を実用化した場合など），という点です．このことを踏まえた上で，ある AI が別の何らかの AI のクローンモデルであるか否かをどのように判定するか，を考えなければなりません．ここで鍵となるのが「電子透かし」と呼ばれる技術です．

画像などの電子コンテンツにコンテンツ自体とは別の情報を埋め込む技術を電子透かし技術と呼び，それにより埋め込まれた情報のことを電子透かしと呼びます．例えば画像では，少数の画素の値を数ビット程度書き換えたとしても，人間にはその変化を知覚することはできません．このような人間の知覚特性を利用することにより，電子コンテンツを視聴する人々に違和感を与えることなく情報を埋め込むことが可能となります．埋め込まれる情報として最も典型的なのは，その電子コンテンツの作者の署名です．署名を埋め込んでおけば，何らかの理由で電子コンテンツが他者によって複製され一般公開されたとしても，それが不正に複製されたものであることを事後的に突き止めることが可能となり，作者は著作権を正当に主張することができます．また，そのことが複製を目論む人物に対する抑止力として働くことも期待できます．現代社会において，電子透かしは著作権保護のための基盤技術となっています．

これと同様の電子透かし技術を AI に対しても確立することができれば，クローンモデルの検知が可能となり，MS を目論む攻撃者に対する抑止力となることが期待されます．しかし，先述したように，AI とは関数なので，その実体である電子データに透かし情報（署名）を埋め込んだとしても，それはクローンモデルには引き継がれず，署名としての役割を果たさないことになります．従って，あくまでも「関数」の観点から透かし情報を埋め込む必要があります．

これを実現する方法の一つとして，特定の画像に対してのみ故意に誤った認識結果を返すように画像認識 AI をあらかじめ設計しておく，という方法が提案されています[7]．例えば，野生動物のモニタリング調査などの目的で設計された動物画像認識 AI（画像中に写っている動物の種類を認識する AI）において，ある特定の 1 枚のウサギの画

像に対してのみ「タヌキ」という認識結果を返すよう，その AI を予め設計しておく，ということです．このように設計しておいた AI がのちに MS の標的となり被害者モデルとなった場合，故意に設定しておいた上記の誤認識がクローンモデル側でも再現されるはずです．これに対し，被害者モデルとは独立に設計された全く別物の動物画像認識 AI の場合，仮にその AI が被害者モデルにたまたま似た挙動を示すとしても，故意に設定した誤認識までもが同じように発生することはまずあり得ません．従って，特定の画像に対する「故意の誤認識」は AI における電子透かしとして活用できる可能性があります．

実際には，故意の誤認識を引き起こす画像が一枚だけでは頑健性に欠けるため，そのような画像を複数枚用意しておくことになります．その上で，クローンモデルか否かを判定したい別 AI が現れた際には，あらかじめ用意しておいた画像のうちの何枚で犠牲者モデルと全く同じ誤認識が引き起こされるかを調べます．その結果，高い割合で同じ誤認識が生じる場合，判定対象の AI はクローンモデルであると判定します．

上記の方法には，出力一致率の高い（＝危険度の高い）クローンモデルであればあるほど，より正確な判定を下すことができる，という利点があります．しかし，逆に言えば，攻撃者があえて出力一致率をやや低下させるようにクローンモデルを機械学習した場合には正確な判定が困難になる，という欠点を抱えているとも言えます．また，現代の AI の特長として，既に機械学習済みの AI に対し新たな学習データを与えて追加で機械学習を行い，それにより更なる性能の向上を図る，といった運用の仕方が可能である点が挙げられますが，このような追加学習時に透かし情報が失われる（予め設定しておいた誤認識が追加学習後の AI では再現されない），という欠点もあります．

以上のように，事後防御手法も今のところ完全ではありません．従って，より高度な透かし埋め込み手法の研究を進めるとともに，事前防御との組み合わせも検討し，少しでも万全な対策へと近づけていくことが重要であると言えます．

■ おわりに

本稿では，主に画像認識 AI を題材として，AI を不正に複製しようとする攻撃「Model Stealing」とその防御技術に関する研究動向を紹介しました．

AI が普及するにつれ，その負の側面が語られる機会も増えつつあるように思われます．AI を利用して捏造されたフェイクニュースの問題や，AI を悪用したプライバシーの暴露といった問題は，その代表的な例と言えます．これらの例のように，AI を攻撃側とする視点から AI の負の側面が語られることは珍しくないのですが，実際には，本稿で紹介したように，AI が攻撃を受ける側となる状況も起こり得ます．今後 AI の利活用がますます進み，その資産価値が認識されるにつれ，AI が攻撃を受けるケースも目立って増えてくるものと危惧されます．本稿では紹介しませんでしたが，AI を機械学習する際に用いられた学習データを AI そのものから復元しようとする Model Inversion Attack や，特定のデータに対し人間の目には知覚できないような微細な変化を与えることにより AI の判断を誤らせる Adversarial Attack など，AI を標的とした攻撃が MS のほかにも，いろいろと存在します．そのような攻撃の存在をできる限り多くの方に知って頂くことにより，技術的・法的な対策に関する議論を活性化し，来るべき AI 社会の安全性向上に少しでも寄与することができれば，筆者としては大変幸いに思います．

参考文献

1）F. Tramer, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing Machine Learning Models via Prediction APIs”, Proc. USENIX Security Symposium, pp. 601–618（2016）.

2）Y. Shi, Y. Sagduyu, and A. Grushin, “How to Steal a Machine Learning Classifier with Deep Learning”, Proc. IEEE International Symposium on Technologies for Homeland Security, pp. 1–5（2017）.

3）B. Wang and N. Z. Gong, “Stealing Hyperparameters in Machine Learning”, Proc. IEEE Symposium on Security and Privacy, pp. 36–52（2018）.

4）J. Lee, S. Han, and S. Lee, “Model Stealing Defense against Exploiting Information Leak through the Interpretation of Deep Neural Nets”, Proc. International Joint Conference on Artificial Intelligence, pp. 710–716（2022）.

5）M. Juuti, S. Szyller, and S. Marchal, “PRADA: Protecting against DNN Model Stealing Attacks”, Proc. IEEE European Symposium on Security and Privacy, pp. 512–527（2019）.

6）K. Nakamura, Y. Mori, N. Nitta, and N. Babaguchi, “Recognizer Cloning Attack on Image Recognition Services and Its Defending Method”, Frontiers in Fake Media Generation and Detection, Chapter 10, Springer, pp. 235–247, June（2022）.

7）E. L. Merrer, P. Perez, and G. Tredan, “Adversarial Frontier Stitching for Remote Neural Network Watermarking”, Neural Computing and Applications, Vol. 32, No. 13, pp. 9233–9244（2020）.