2021 年 6 月
JDLA 事務局

「AI ガバナンスとその評価」研究会
第 12 回
開催報告

1. はじめに

日本ディープラーニング協会では、人工知能（以下 AI）や Deep Learning（以下 DL）に関連する国内外の政策動向についての知見を深め、議論する場としてテーマごとに研究会を設置している。本研究会「AI ガバナンスとその評価」は、多様なアクターによる管理・評価の体制の在り方を「ガバナンス」と定義し、どのようなガバナンスの形がありうるのか調査し、信頼される AI の構築の一助とする研究会を 2020 年 7 月から立ち上げ、1 年間の検討を行う。

研究会第 12 回（2021 年 4 月 22 日）では、AI ガバナンスに向けた「セキュリティ」というテーマのもと、筑波大学 佐久間 淳氏、株式会社ブレインパッド 韮原 祐介氏より話題提供いただいた。

本レポートは、話題提供の内容と研究会参加者のディスカッションを再構成して作成したものである。

2. AI セキュリティ・プライバシーと信頼性の問題について

佐久間氏より「AI セキュリティ・プライバシーと信頼性の問題について」と題して話題提供いただいた。

AI による判断や意思決定が信頼される条件

近年、深層学習の技術発展により、AI による画像認識の精度が上昇しており、数字の識別等の単純な目的に絞った場合、AI の認識の精度は人間の精度を上回っている。AI 精度の向上から、医療診断や自動運転のような、より複雑な意思決定のタスクに AI を応用する事例が出てきている。

しかしながら、医療診断 AI のように AI をエキスパートとして活用した場合、AI の誤った判断結果が、人間の人生を左右するような重大な意思決定に及ぶ可能性が高くなる。そのため、AI による判断結果が正しいだけではなく、判断結果の根拠やプロセスの説明や裏付けを基に、倫理的に問題がない意思決定をしていく必要があると考える。

なぜなら、人間の場合でも、医師や法律家等、人間の人生に大きな影響を与える意思決定を行う必要がある人たちは、高い職業倫理が求められており、法律や職業のガイドライン、内部的な制約により倫理を守るように定められている。このことから、高い職業倫理が求められる仕事を AI が担う場合にも、AI の判断結果が倫理的に問題ないことを確かめる必要があると考える。

佐久間氏はAIによる判断や意思決定が信頼される条件として、下表1の内容を挙げている。

表1 AIによる判断や意思決定が信頼される条件[1]

最低限の必要要件	エキスパートレベルの品質	
	法令順守	
制約	環境変化・攻撃に対する安定性	
	人権の尊重	自己決定権の尊重
		プライバシー保護
		公平性への配慮
事後検証	判断根拠の説明可能性	
	判断過程の追跡可能性	

- ✓ 制約
  利益最大化の目的を多少制限してでも、品質について考慮すべき条件のことである。
- ✓ 環境変化・攻撃に対する安定性
  後述「AIへの攻撃」参照
- ✓ 自己決定権の尊重
  医療や裁判のように人間の人生に影響を与えるような判断をするAIの場合、自己決定権が奪われていると重大な問題を引き起こす可能性があるため、自己決定権の尊重が条件として必要である。
- ✓ プライバシーの保護
  AIの開発や利用等にあたり、個人や企業の十分な同意なく集められたデータは利用すべきではなく、またAIモデル等から個人情報が搾取されることはあってはならない。技術的な問題にも配慮し、AIを利用していく必要がある。
- ✓ 公平性への配慮
  後述「公平性への配慮」参照
- ✓ 判断根拠の説明可能性
  後述「AIのセキュリティ」参照
- ✓ 判断過程の追跡可能性
  AIが判断を誤った際に、人間がどのようにして誤った判断をしたのかAIの判断結果プロセスを追跡可能なことが条件として必要である。

AIサービス開発時に、AIによる判断や意思決定が信頼される条件を考慮しないまま設計すると、AIサービス提供者が想定していない第三者を不幸にするおそれがあ

[1] 本研究会公開資料より抜粋

る。前述のような観点があることを念頭に置きながら、AI サービス開発を行うことが肝要である。

AI への攻撃

AI への攻撃として、入力データに微小なノイズを加えることで AI モデルに誤った予測をさせる敵対的サンプルというものがある。例えば、パンダの画像に巧妙にノイズを加えると、AI はパンタではなくテナガザルと認識する。敵対的サンプルは、どのような画像に対しても作成可能なことが判明している。

また敵対的サンプルは、音声に対しても作成することが可能である。音声に微弱なノイズを含ませることで AI には別のものと認識させる。例えば、人間には音楽に聞こえる微弱なノイズを含む音を、音声認識の AI に対しては消防車への通報と認識させることが可能である。さらに音声の敵対的サンプルは、音声サンプルはスピーカーやラジオ等から流すことが可能なため、拡散性が高く、攻撃対象のデバイスに敵対的サンプルを到達させ強制的に操作することが可能である。

なお、敵対的サンプルは説明可能性も攻撃の対象となり、微弱なノイズを含んだ画像を AI に与えると、まったく異なる説明を示す。前述のように AI が不安定な挙動をした場合、AI 自体の性能が高くても AI サービスの利用者には信頼されないおそれがあるため、環境変化・攻撃に対する安定性は、AI による判断や意思決定が信頼される条件として必要である。

公平性への配慮

AI は、人間の人生に影響を与える判断や意思決定(例えば、雇用、信用スコア、入試、保険料率)を担う可能性がある。さらに一部の AI は、既に人間が行う判断や意思決定の一部になっているともいえる。そのため、AI の意思決定では公平性に配慮する必要がある。

AI が差別を引き起こすような意思決定は、人間の属性に依存して発生する。例えば採用 AI が性別に依存した場合、同様の成績や経歴のデータを持つ男女を判断すると、男性では採用と判断し、女性だと不採用と判断される場合がある。これは AI による性差別といわれている。ただし、判断主体が人間か AI というのは関係なく、差別を定義し、公平性に配慮することは AI ではなくても必要である。

アメリカの複数の州で採用されている「COMPAS」という、被告人の再犯可能性を予測するアルゴリズムがあり、保釈金額や判決の決定等において裁判官の決定を支援している。しかし、COMPAS の予測結果は人種の影響を受けているということが、2015 年に Julia Angwin 氏により公表された[2]。同氏は、COMPAS は白人の再犯リスクの可能性を実際よりも低めに判断し、有色人種の再犯リスクの可能性を高めに

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

判断する事例があると指摘した。COMPAS 自体は、AI とは言えないような単純なアルゴリズムではあるが、統計的な判断で差別が引き起こされていることをどのように評価し、公平性を担保していくこと課題があることを示唆している。

公平性の指標として Demographic parity という指標がある。Demographic parity とは、判断の配慮属性(例：性別)に対する条件付き分布の配慮属性値(例：男女)に対する差のことである。公平性を担保する一つの方法として、Demographic parity を小さくするように学習し、公平性の指標を達成することが挙げられる。

AI のセキュリティ

通信のセキュリティは、情報を通信するという単純なタスクに対するものであるが、通信への不正アクセスに対する完璧な防御方法は未だに見出されていない。防御者が攻撃に対して対策を施すと、攻撃者はその対策に対する脆弱性を見つけ攻撃するといった、いたちごっこの状態が長年続いている。そのため、セキュリティ対策の研究を続けることが、高度なセキュリティレベルを維持するための本質となっている。

AI のセキュリティ対策は、通信のセキュリティ対策よりもさらに複雑である。AI のセキュリティでポイントとなるのは、AI に対する攻撃は AI が実施していることが多いため、防御手段として AI による防御策が必要になることである。

ただし、通信のセキュリティ攻防と同様に、AI によるセキュリティ攻防もいたちごっこの状態が続いていくと考えられる。

3. AI に係るサイバー脅威と対策の方向性

続いて、韮原氏より「AI に係るサイバー脅威と対策の方向性」と題して話題提供いただいた。

AI システムのセキュリティ

AI システムの攻撃ポイントは 6 カ所ある、学習過程の入力データ、情報処理時のアルゴリズム、学習済みのモデル、推論過程の元データ、モデル、出力結果の 6 か所である。

推論過程の元データの攻撃の例として、AI の画像認識において、元の画像を 1 ピクセルだけ変化させることで AI に誤った判断を行わせることが可能である。出力結果に関する攻撃の一例としては、AI ではないが、Stuxnet[3]を利用したイランの核施設の SCADA システムへのサイバー攻撃を挙げる。当該攻撃では、遠心分離機の回転周期を改竄し、監視モニターには正常と表示させることで数か月～半年にかけて遠心分離

[3] Microsoft Windows で動作するコンピュータワームのことである。

機を破壊した。当該核施設へのサイバー攻撃では、ゼロデイ攻撃[4]が行われた。脆弱性は修正パッチを適用すれば脆弱性は解消されるが、ゼロデイ攻撃は未知の脆弱性を利用したものであるため、攻撃を未然に防ぐことは困難である。また当該事例は、AI システムのセキュリティを考える場合には、AI モデルだけではなく、AI を組み込んだソフトウェア全体のセキュリティを考える必要があることを示唆している。

サイバーレジリエントなソフトウェア

スマートメーターによる電力使用量から住人の在宅状況を判断することで、再配達増加の問題を解決するという取り組みがある。一方で、在宅状況をデジタル化することで誘拐、要人暗殺、盗難、空き巣等の被害に合うリスクが生じる可能性がある。また、セコム警備員による窃盗事件[5]が発生していることから、スマートメーターを利用することで前述の様な事件が起きることを想定した上でサービスの本格実用化が求められる。通信会社、電力会社、配送企業、分析企業等のサービス提供に関わるサプライチェーン上のデータ漏洩のリスクへの対応策を検討する必要がある。利用者はスマートメーターのデータを利用することで再配達が減るというメリットを享受できる一方、自身が前述のリスクにさらされている状況にあることを認識し、合意に基づくサービス運営が必須となるだろう。

しかしながら現実には、サイバー攻撃への対応について、実被害を被ってからでないと本格的な対策が施されにくいことがある。例えば、Google は 2010 年 1 月に中国からのハッキングにより知的財産の窃取被害の公表[6]以降、ゼロトラストネットワーク[7]の構築に取り組んでいる。ゼロトラストネットワークの構築は、防壁を作り守り通すという「サイバーセキュリティ」の考え方から、全てのシステムは必ずハッキングされる、全ての通信を認証しなければならない（ゼロトラスト）という前提の下、サイバー攻撃被害から早期に事業復旧した上で、サイバー攻撃を糧により強固な超回復力を獲得することに焦点を当てる「サイバーレジリエンス」という考え方と一体となって行われるべきである。近年、米国防総省やダボス会議等でも「サイバーレジリエンス」が懸案事項として着目されている。

サイバーレジリエントなソフトウェアを作るためには、「DevSecOps[8]」が有用であ

---

[4] ゼロデイ攻撃とは、OS やアプリケーション等のソフトウェアの誰も気づいていない脆弱性や修正パッチが提供されていない脆弱性を悪用した攻撃である。

[5] https://www.asahi.com/articles/ASMDB5CTPMDBPIHB02C.html

[6] https://googleblog.blogspot.com/2010/01/new-approach-to-china.html

[7] ネットワークのセキュリティの概念。全て信頼できないことを前提として、全てのデバイスのトラフィックの検査やログを取得し、検査するというアプローチである。

[8] DevOps は、開発のチーム(Development)と運用のチーム(Operations)が互いのミッションを補完することにより、ビジネスにおけるシステムの有用性を高め、迅速かつ確実な実装と運用を実現する考え方である。DevSecOps は、この DevOps による開発と運用にセキュリティ(Security)を加えた考え方である。

る。DevSecOps は、ソフトウェアのプランニング、サービス設計、開発、テスト、運用の各段階でセキュリティについて考慮している。例えば、プランニングでは、どのような攻撃脅威があるのかを検討する。サービス設計では、AI サービス利用者が危険にさらされる可能性を考慮する。開発時には、コーディングをセキュアにし、ソースコードの脆弱性解析をする。テスト時には、自動化した脆弱性のテストや、ホワイトハッカーから受けた攻撃により識別した脆弱性への対策、運用時は、定期的なセキュリティスキャンやセキュリティ監査の実施、オープンソースの中に脆弱性が発見されればオープンソースのバージョンアップを実施する。

最近は DevSecOps と MLOps[9]を組み合わせた、「MLSecOpes」というセキュリティを考慮した AI 開発の考え方も北米を中心に提唱されるようになってきた。また、機械学習システムに対する攻撃を検出・対応・修復を可能にすることを目的としたオープンソースフレームワーク「Adversarial ML Threat Matrix（敵対的機械学習に対する脅威マトリクス）」が、Microsoft や IBM、カーネギーメロン大学等の有志コミュニティから公表[10]されている。

ハイブリッド脅威

今後の国際的な安全保障における主要な脅威は、物理空間、情報空間、及び宇宙空間（陸海空、サイバー、認知空間、及び宇宙空間）への攻撃手法を組み合わせたハイブリッドな脅威（以下ハイブリッド脅威）である。

ハイブリット脅威が注目されたのは、2014 年のロシアによるクリミア併合の時に行われた作戦がきっかけで、ドメイン横断での戦争が、ハイブリッド戦争と呼ばれるようになった。ハイブリッド戦争は軍事戦略の一つであり、正規戦[11]、非正規戦[12]、サイバー戦、情報戦等を組み合わせて行われる。ロシアによるクリミア併合は、ウクライナのクリミア自治共和国がクリミア議会による住民投票で 95%以上の賛成票が投じられ、ウクライナから独立し、ロシア連邦に併合した[13]。

＜クリミア併合時に見られたハイブリッド脅威＞

- ✓ ウクライナ・クリミア間の通信回線の遮断
- ✓ SNS によるウクライナのデモ・暴動やクリミア域内のロシア系住民の不安を煽る情報拡散

---

[9] DevOps をベースとしており、機械学習モデルの開発と運用を一体化し、開発から運用まで一連のライフサイクルを管理する手法のことである。

[10] https://github.com/mitre/advmlthreatmatrix

[11] 国家によって組織された訓練されている軍隊が行う戦争形態のこと。

[12] 民衆が自発的あるいは外国の援助を得て武装する紛争等の正規戦に分類できない形態のこと。

[13] クリミア併合は、国際的には承認を得られていないと考えられている。

- ✓ ウェブサイトへの DDoS 攻撃
- ✓ リトル・グリーン・マンと呼ばれる自警団による空港や政府施設、クリミア州議会を占拠

なお宇宙空間に着目すると、測位衛星（GPS など）、通信衛星、気象衛星、早期警戒衛星などが宇宙空間と地球間で通信を行っているため、宇宙空間もサイバーセキュリティの対象となる。そのため、衛星の内部機構、地球-衛星間、衛星-衛星間のセキュリティリスク（衛星への攻撃[14]）についても検討する必要がある。さらに、例えば小惑星探査機「はやぶさ」のように AI 技術による自律制御機能が宇宙開発における探査機やロケットに必須であることからすると、これらのセキュリティリスクについても対策しながら、今後の宇宙開発を行っていく必要がある。

ハイブリッド脅威下でのセキュリティレジリエンスの考え方として、マルチドメインでのセキュリティレジリエンスのマインドが必要である。AI システムは重要だが、あくまでソフトウェア全体の一部に過ぎない（陸海空、サイバー空間、認知世界、宇宙もある中の一部に AI がある）。ハイブリッド脅威への対応は、国の安全保障政策に基づき検討していく必要があると考える。

Cyber Grand Challenge

アメリカ国防高等研究計画局(DARPA[15])主催で、Cyber Grand Challenge（以下 CGC）が 2016 年 8 月に初めて開催された。CGC では用意されたプログラムの脆弱性を探し、修正パッチの作成、敵チームのプログラムを攻撃するという行動を自動化されたプログラムによって行われた。2016 年開催の CGC では、カーネギーメロン大学スピンオフ企業の ForAllSecure チームの「Mayhem」というシステムが優勝した。Mayhem にはファジング(fuzzing)というテストの自動化技術が利用されており、プログラムの脆弱性を自動でテストできる。ファジングはプログラムへの入力を自動生成してプログラムに与えると脆弱性のある個所を特定する。

＜ファジングの種類＞

- ✓ ランダム(random fuzzing)：ランダムな入力値をプログラムに与える。
- ✓ テンプレート(template fuzzing)：マニュアルで入力値テンプレートを準備してプログラムに与える。
- ✓ 生成的(generational fuzzing)：入力値を自動で生成してプログラムに与える。

---

[14] 衛星への攻撃の種類として、①軍事衛星からのアーム、科学スプレー、ネット、高出力マイクロ波、ジャミングによる攻撃、②地上からの光学衛星へのレーザー照射、③対衛星攻撃ミサイルによる攻撃、④地上衛星間通信のスプーフィング・ジャミング攻撃、④地上管制施設へのサイバー攻撃がある。

[15] Defense Advanced Research Projects Agency の略称

✓ ガイド(guided fuzzing)：入力値を自動生成した上で、結果を学習しながら、次の入力値を自動生成してプログラムに与える。各テストケースの効果をスコアリングし、強化学習していく。

ファジングは、グーグルやマイクロソフトで脆弱性発見のために日常的に使われる技術である。このファジングなどのテスト技術を組み合わせた Mayhem はその後、米国防総省にも採用されており、精度の高い脆弱性発見 AI システムとして利用が進んでいる。こうしたサイバー防衛のための AI 開発をいかに構築できるかも、今後のサイバー世界の安全保障においての重大なポイントになってきている。

4. 質疑応答での論点

第 12 回研究会では AI ガバナンスに向けた「セキュリティ」の検討内容について議論が行われた。話題提供を踏まえて以下のような質疑応答が行われた。

AI システムのセキュリティの考え方

✓ AI モデル単体のリスク評価結果と、AI システム全体のリスク評価結果は異なる可能性があることを意識する必要がある。

✓ AI といってもシステムの一部であるため、一般的なシステムセキュリティのフレームワークや標準的な手順を活用し、システムセキュリティを設定することは必要である。ただし、AI のセキュリティリスクに対して全ての対応を行うことは難しいと考えている。そのため、考えうる脅威について想定する知識を持ち、脅威が起こった際に対応できるように想定して、AI システムを構築する必要がある。

✓ AI サービスが攻撃される可能性を認識し、AI サービスが攻撃される目的や攻撃対象を想定することが必要である。なお、サイバー攻撃の目的は愉快犯だけではなく、軍事目的、金銭目的等の目的が存在することを認識すべきである。

✓ AI サービスへの攻撃のシナリオに応じて、セキュリティ対策は変える必要がある。AI サービスの中で一番のリスクとなる問題は何か、その問題が起こる可能性について検討する必要がある。

✓ セキュリティ対策は、ワーストケースを想定して考えることが多い。利用形態から想定して対策を講じるというアプローチが必ずしも完全なセキュリティを達成するとはいえない。しかしながら、サイバーセキュリティリスクの全てに対して対応することは現実的ではないため、サービスの利用形態に鑑み、脆弱性が外部環境に影響する可能性が高い箇所から対応することはセキュリティ戦略としてあり得る。

✓ 自社サービスのサイバーセキュリティリスク対策レベルの程度は、GAFAM[16]の担当者ですら頭を悩ませていることである。セキュリティ対策の優先順位付けを実施し、ある程度のリスクを許容することもビジネス運営上は妥協が必要となる。
✓ システムの期待から外れた挙動をバグと捉えると、AI の場合、頑健性や公平性の問題も含めて、期待する意思決定から外れた挙動はバグと捉えることもできるため、このような AI の仕組みに起因する問題も「セキュリティ」の範囲の問題と捉えることができるだろう。

AI に関するセキュリティ教育

✓ 大学の研究者を増やす教育をしていくことが重要である。現在、コンピュータサイエンス領域の学部を設けている大学の数は多くはない。
✓ 情報技術の発展に伴い、セキュリティを学問とする領域も発展している。ML／DL がさらに発展していけば、従来のセキュリティに関する研究所でも、一定の割合の研究者が AI セキュリティの研究（ML／DL のセキュリティの研究）に従事していくことが予想される。
✓ AI 単体のセキュリティだけではなく、サイバーや軍事戦略と絡めて AI セキュリティの研究を行うことが必要である。

AI への攻撃

✓ AI は攻撃されていなくても、AI の誤判定はゼロにはならず、また AI の判断結果は 100%当たるものではないため、前提として AI は 100%当たらないと考えるべきである。
✓ 音声の敵対的サンプルでは、意図的に「消防署へ通報」させるようなノイズを開発することは可能である。
✓ 音声の敵対的サンプルでは、人間に聞こえない超音波の帯域を利用しても攻撃可能である。
✓ 専門教育を受けた学生は、敵対的サンプルを作成することが可能である。一定レベルの ML/DL の知識は必要となるが、敵対的サンプルの作成自体は難しいことではない。しかし、敵対的サンプルを外部環境へ拡散する条件を整えることは難しく、音声の敵対的サンプルの場合、ただちに他人のスマートフォンに敵対的サンプルを拡散させることができるわけではない。音声認識モデルにアクセスすることを前提として敵対的サンプルは作成されるが、スマートフォンの音声認識モデルには簡単にアクセスすることができない仕様になっている。
✓ AI サービスへの攻撃は、AI モデルへの技術的な攻撃だけではなく、人の手によ

[16] GAFAM とは、米国の大手 IT 関連企業 5 社（Google, Apple, Facebook, Amazon, Microsoft）の頭文字をとって名付けられた造語である。

る攻撃も存在するため、技術で対応しようとしても、人の手で脅かされる可能性がある。

セキュリティレジリエンスを高める方法

- ✓ 常にセキュリティリスクにさらされているという事実を知ることが重要である。例えば、Google Workspace や Microsoft Office365 を導入することで、（スパムメールが自動で除外されている状況や、システムへの不正ログイン時にデバイス認証ではじかれている状態等の）セキュリティリスクにさらされている状態の可視化が可能となる。
- ✓ システムが動作しなくなる可能性等の影響を恐れて、OS バージョンアップをしないことがあるが、既知の脆弱性を解決するパッチを適用していくことが重要。またそれが行える運用体制の整備が重要となる。
- ✓ セキュリティリスクを可視化し、組織的な対策、判断を積み重ねていく必要がある。

ビジネスにおけるセキュリティの位置づけ

- ✓ DX 改革においては利益を創出できるかが優先されており、セキュリティは後回しになっている。
- ✓ システムの機能改善対応と、セキュリティに脆弱性への対応のどちらを優先して対応すべきは、ビジネス的な判断となる。CRM やサービスのブランディングの一環としてセキュリティを考え、ビジネスの土俵に載せて判断していくことが重要である。

次回以降も引き続き、本研究会を通じて、日本国内外の AI ガバナンスに係る検討を続ける。

（文責：清見友来）

＜第 12 回開催概要＞
日時 ： 4 月 22 日（木）14:30-16:30（Zoom 開催）
内容 ： 話題提供：「AI セキュリティ・プライバシーと信頼性の問題について」
佐久間 淳氏（筑波大学）
話題提供：「AI に係るサイバー脅威と対策の方向性」
韮原 祐介氏（株式会社ブレインパッド）
質疑応答・ディスカッション