Vol.2019-SPT-32 No.21
2019/3/8

機械学習システムのトラスト構築に関する課題分析

小川隆一[†1]　島成佳[†2]

概要：AI の急速な普及にともない，AI を社会がどのように信頼できるか，の問題に注目が集まっている．これまで倫理原則の策定や標準化が試みられているが，本稿では機械学習機能を備えた AI 搭載システムをプロダクトとみた場合どのような信頼を利用者に提供すべきか，主に性能とセキュリティの観点で課題を分析し，整理した．特に業務向けの AI 応用では，学習データ・学習プロセスの妥当性が信頼に大きく関わり，この点でサプライヤと利用者の共同作業が重要であることを示した．

キーワード：機械学習，AI，トラスト，社会受容，期待性能

Analysis of issues on building trust of machine learning systems

Ryuichi OGAWA[†1]　Shigeyoshi Shima[†2]

Abstract: Along with the rapid spreading of AI-based systems, establishing social trust has become an important issue for such systems to be accepted. So far ethical principles and standards have been under development, but the design of AI trust has not been attempted. In this paper we analyze the trust development problems of machine learning systems from the viewpoint of trustworthy product (or service), regarding capability/accuracy and security. In particular we look at AI application for business that require appropriate learning through the collaboration of AI system supplier and user. This could be an AI-specific feature of trust development.

Keywords: Machine learning, AI, Trust, Social acceptance, Expected capability/accuracy

1. はじめに

近年，深層学習技術の発展・実用化等に伴い，AI（Artificial Intelligence：人工知能）の利活用が新しいサービスの創造，社会経済の発展や社会課題の解決に寄与することが期待されている．同時に，AI の不適切な利用，あるいは AI 利用環境の不備が社会経済活動に新たなリスクが生じさせることが懸念されており，これらを適切にコントロールすることが，AI の利活用には不可欠である．

AI の利活用に関するリスクについては，これまで主に以下の 5 点について，技術・倫理・制度等の観点から技術者・研究者・法律家等により議論されてきたa.

① 不適切な利用・悪用．喫緊の課題としては，AI が公平・公正の原則に反する目的(差別等)に使われるリスクが指摘され，GDPR 等の制度面の対応が始まっている[1]．正規の学習が不公正な結果を生むリスクについては，応用ごとの慎重な検証が必要である．今後さらに，犯罪目的の AI 悪用が課題になると思われる．このほか，IEEE は AI の軍事利用における倫理面・技術面の課題を指摘している[2].

② 責任分担．AI が人間の業務を代替する場合の責任分担があいまいであることは，特にインシデント時のリスクになる．例えば自動走行における AI 利用の責任範囲については，法制，自動走行の社会実装方式，運転者と AI との連携方式等について議論が進められている．

③ 説明責任．AI で高度な分析を行う場合，なぜその分析結果に至ったかの説明が現状ではできない．これが社会の AI 受容を困難にするリスクは技術者・研究者に共有され，エンジンの説明機能を強化する等の試みがなされている[3]．また，DARPA（米国防総省・国防高等研究計画局）が Explainable Artificial Intelligence (XAI)という投資プログラムを発表して研究を促進している[4]．平易な説明に至るにはまだギャップが大きいと思われる．

④ 性能・品質．学習機能を備えた AI は従来のソフトウェアと異なり，学習によって性能・品質が左右されるが，これをどう評価・維持すべきかの方法論が定まっておらず，信頼性のリスクにつながる．このためソフトウェア工学者を中心に，AI 品質工学ともよぶべき技術構築の機運が急激に高まっている．

⑤ セキュリティ．AI のセキュリティ利用が期待されているが，AI 搭載システム自身が脆弱でないことも重要である．近年，アルゴリズムの脆弱性をついて誤判断を起こさせる敵対的学習が研究されているが，

†1 独立行政法人情報処理推進機構
Information-technology Promotion Agency, Japan
†2 日本電気株式会社
NEC Corporation

a いわゆる汎用人工知能が人間の知性を凌駕するシンギュラリティがリスクとして議論されることがあるが，本項ではこの課題は扱わない．

学習データの真正性や秘匿，学習・運用の妥当性（悪用しない）等の包括的な対応が必要になる．

これらのリスクはいずれも，AI 搭載システムやサービスを社会が受け入れるためにコントロールされるべきである．本稿ではこのコントロールに向けて，社会が AI を信用する指標となるトラスト構築の問題を検討する．上記のようにトラストに影響する要因は多様であるが，ここでは以下の二つのカテゴリに分類してみる．

１．社会受容のためのトラスト

利用者を含む社会全体が AI を受容し，使ってもいいと実感するため，中長期にわたり分野横断的に構築されるべきトラストである．制度・技術を整備することに加え，利用者の心理的な受容（特に上記項目①～③における合意）が重要である．

２．プロダクト（製品）としてのトラスト

「１」を実現する前段として，AI 搭載システム・サービスが品質・セキュリティ・運用の妥当性等で信頼に足ることを示す必要がある．上記項目では主に③～⑤に関する信頼の枠組みを「１」よりも短いタイムスパンで実現する必要がある．もちろんプロダクトとしてはセキュアで高性能性だが社会が受け入れない，とならないよう「１」のトラストも考慮しなければならない．

本稿では，後者の「製品としてのトラスト」に着目し，次章から課題を分析して整理する．

2. 本論文のトラストと関連議論

筆者らの当初の問題意識は，AI による企業リスク分析が実用化されたことを受け，その結果をどう受容するか，の議論で以下の疑問が出たことから始まっている．

- 学習データは信用できるか
- 学習や分析の精度・妥当性はだれがどう評価するか
- 評価結果を悪用しないことはどう担保されるか

これらを検討する過程で，AI がどういう機能を持つものか，の機能の定義，またそれに関連し，機能の何を信頼すればいいのか？の対象範囲（信頼のスコープと呼ぶ）の定義が非常に重要であることが判明した．次節でこれを整理する．

2.1 想定する AI の機能と信頼のスコープ

AI(Artificial Intelligence：人工知能)は，人間の脳の認知・判断等の機能を人間の脳と異なる仕組みによって実現する技術を総称したものと言えるが，実現イメージや適用する技術（機械学習やエキスパートシステム等）等，人によって定義は異なる．本稿では，AI を近年注目されている機械学習機能を備えたソフトウェアと想定する．学習機能があることで，AI は従来のソフトウェアとは異なる信頼の枠組みが必要になるからである．

信頼のスコープについて述べる．従来のソフトウェアは，各プログラムが仕様どおり，期待される機能や精度で動くことで信頼が担保される．特定の分析を行うよう作成された従来のソフトウェアは，仕様に規定された仕様と精度で分析ができればそれでよいとされる．しかし，学習機能つきの AI エンジンが組み込まれて分析を行う場合，エンジンが仕様どおり動作するだけでなく，事前の学習によってあらかじめ期待される分析機能・精度で予測や判定ができること（期待性能と呼ぶ）が求められる．

すなわち AI 搭載システムは，ソフトウェアに加え，学習データ（訓練データ）や学習モデルの真正性，妥当性が信頼の要件になると考えられる．コンシューマ向けの AI 製品（AI スピーカー等）では厳密な性能を信頼の要件にしなくてよいかもしれないが，業務の一部自動化等の AI 応用では，業務に特化した学習による期待性能が重要になると思われる．

また，学習データや分析結果はプライバシー情報・企業秘密等を含む機微情報である可能性が高い．こうした情報を安全に管理し漏洩させないこと，さらに分析結果を目的外に使わないことは AI 搭載システム信頼の必須要件であり，適切な運用手順，セキュリティ技術，認証技術等が必要になると思われる．

以下では，AI の信頼構築に関連した従来の議論を概観する．

2.2 AI の社会実装に関する倫理原則

AI が社会で信頼され，受容されるために，技術だけでなく，倫理の観点から広く議論されている．これらは 1 章の分類 1（社会受容のためのトラスト）に関わるものとみなせる．国内外の組織による代表的な取り組みを示す．

(1) IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

IEEE はいち早く AI の社会実装について議論を開始し，関連分野の識者により編纂された ETHICALLY ALIGNED DESIGNE Version2（EAD）を公開している[2]．AI の倫理的な設計・開発・実装に関する原則として，以下の 5 項目をあげている．

原則 1 — Human Rights（人権）
原則 2 — Prioritizing Well-being（幸福）
原則 3 — Accountability（アカウンタビリティ）
原則 4 — Transparency.（透明性）
原則 5 — A/IS Technology Misuse and Awareness of It（悪用への警戒）

(2) 倫理原則の標準化

IEEE はさらに上記原則の標準化に着手し，IEEE P7000 Engineering Methodologies for Ethical Life-Cycle Concerns

Working Group [5]にて，設計段階における倫理課題対応のモデルプロセス（P7000），自律型システムの透明性（P7001），データプライバシー（P7002），アルゴリズムのバイアス検討（P7003）等の規格を策定中である．

また，同様な倫理原則の標準化は ISO/IEC SC42 にても行われ[6]，AI 搭載システムや決定のバイアス（TR24027），トラスト（Trustworthiness）（TR24028）の規格策定が始まっている．

(3) AI ネットワーク社会推進会議

日本では AI ネットワーク社会推進会議が，社会・経済の諸分野における AI 搭載システムの利活用のシナリオを検討し，AI の開発及び利活用の促進．AI ネットワーク化の健全の進展等に関する課題を整理した[7]．これらの課題を踏まえ，以下の AI 利活用原則案をまとめている．

同原則案では，「便益の増進」「リスクの抑制」「信頼の醸成」の 3 項目に関連付けられるとしている．「信頼の醸成」に関係する原則としては，⑧公平性の原則．⑨透明性の原則，⑩アカウンタビリティの原則　を挙げている．

(4) AI の開発・利用に関する原則

Google は AI の研究開発や製品開発の方針として，意思決定における 7 原則を公開している[8]．

① 社会的に有益なこと
② 不公正な偏見の創出や強化を避けること
③ 安全性のための構築と検査されること
④ 人に対して責任を負うこと
⑤ プライバシデザインの原則を組み込むこと
⑥ 科学的に卓越した基準を守ること
⑦ これらの原則に一致する用途に利用できるようにすること

他にも Microsoft[9]，EU[10]，IBM[11]等が AI に関する倫理原則を公開している．各原則を見ると，説明責任はほとんどの原則に，透明性やプライバシー保護が多くの原則に含まれている．

以上の中で本稿のトラスト検討に最も近いのは IEEE の P7000, ISO/IEC SC42 のバイアス，透明性，Trustworthiness，に関する議論であると思われるが，これらはまだドラフト段階で詳細は確定していない．

2.3 製品としてのトラストに関する議論

(1) 品質評価

本稿では，AI が製品（あるいはサービス）として，安全で高い精度を備えたものであるか，サプライサイド視点からのトラストに注目する．この「製品としてのトラスト」に関しては，まずソフトウェアエンジニアリングのコミュニティから AI 搭載システムの品質評価や管理に関して議論が立ち上がり，現在活発な討議が行われている[12][13][14]. JST はこれに関し，AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立が必要であるとの戦略プロポーザルを行っている[15]．

しかし前述のとおり，学習機能を備えた AI は学習モデル・学習プロセスによって性能が変わりうるため，期待性能は AI 応用分野依存，業務依存になることが予想される，さらに，アルゴリズムとの適性も考慮した学習データの選定も重要であり，例えば基盤プラットフォームとしてソフトウェア横断的な性能指標を定めるアプローチは容易でない，と予想される．学習データの妥当性（量や質，密度等）や学習モデルの精度，学習のプロセスを誰がどのように評価して性能を達成するかのルール作りはこれからである．

(2) セキュリティ

AI とセキュリティに関しては「AI のセキュリティ分野への応用（AI for cyber security）」と「AI 自身のセキュリティ（Cyber security for AI）」の二つの議論がある．前者については，セキュリティコミュニティにおいてマルウェア解析，異常検知等への適用が検討され[16]，トラフィック監視への深層学習の適用も試行されている．本稿ではスコープ外として触れないこととする．

AI 自身のセキュリティについては，AI コミュニティにおいて，アルゴリズムが誤判断を起こす学習データを意図的に与える敵対的学習（Adversarial learning）(例えば[17])が早くから注目されてきた．現在，この攻撃のバリエーションや，分析結果から学習データを再現する等の新しい攻撃について研究が進んでいる．なお，敵対的学習は攻撃対象の AI エンジンの特性を知る必要があるため，容易な手口とはいえない．より包括的な AI 搭載システムの脅威分析については，例えば[18]がある．[18]は学習データ，及び学習プロセスに関わるプレーヤーへの脅威と対策を丁寧に分析しており，本稿にも大きな示唆を与えたものである．

筆者の一人は，CSS2018 の BoF セッションにおいて「AI に対する信頼をどう構築したらよいか」と題して講演した[19]．この講演では，AI エンジンのセキュリティのためには学習プロセスを秘匿すべきだが，信頼構築のための透明性の原則とどう折り合いとつけるか，の問題提起を行った．本稿でこの課題は深耕しないが，引き続き検討していきたい．

以下では，AI 搭載システムの品質の評価や管理に関して，利用者，開発者，運用者等の関係者がどのようにトラストの構築していくのかに注目して，課題の抽出と整理を行う．

3. 課題検討のための関係者アプローチ

AI 搭載のソフトウェアのトラスト構築に関わる関係者を，CSS2018 の BoF で図 1 のように表現した．本論文でも，図 1 を基にトラスト構築に関わる課題を具体化する．

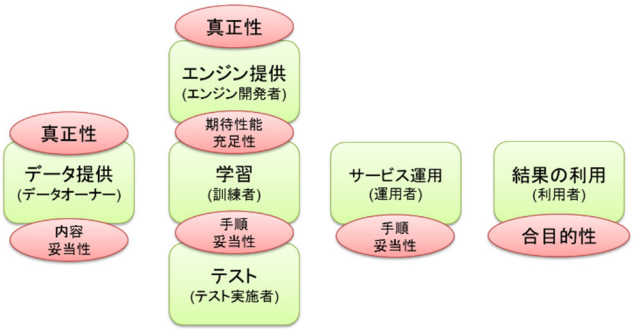

図 1　AI 搭載システムトラスト構築に関わる関係者
Figure 1 Stakeholders for Trust Building AI Software

図 1 の関係者は，AI 搭載システムの開発，運用，利用各フェーズの役割と対応する．1 つの組織に複数の関係者が存在するケースが一般的であり，AI ソフトウェアの開発，運用，利用の形態によって，これらの関係者・組織の組合せが異なると考えられる．各関係者にとってどのようなトラスト要因が重要になるかを以下に示す．

- 利用者
  AI 搭載システムを利用して，その結果を利用する者．トラスト構築には，サービスの内容が利用目的に合っているかどうか（合目的性）が関わる．
- 運用者
  AI 搭載システムを用いたサービスを運用して，利用者にサービスを提供する者．トラスト構築には，サービス運用の手順が正しく，その手順どおりに運用されているかどうか（手順妥当性）が関わる．
- エンジン開発者
  AI 搭載システムを開発する者．トラスト構築には，設計・開発したエンジンが偽物でなく，かつ正しく動作するかどうか（真正性）が関わる．
- 訓練者
  AI 搭載システムの学習モデルをサービス運用可能なレベルまで学習させる者．トラスト構築には，学習モデルの機能や精度を満たしているかどうか（期待性能充足性）が関わる．
- テスト実施者
  AI 搭載システムが，サービス可能な要件であることを確認する者．トラスト構築には，テストの手順が正しく，その手順どおりにテストされたかどうか（手順妥当性）が関わる．
- データオーナー
  AI 搭載システムに学習させるためのデータを保有する者．トラスト構築には，データが改ざんされていないかどうか（真正性）や学習データとして適切なものであるかどうか（内容の妥当性）が関わる．

AI の利用は現状では IT 分野が先行し，Google や Amazon 等の IT 企業がスマートフォンやホームスピーカーを通して，様々なサービスを利用者（コンシューマ）に提供している．図 1 の関係者についてみると，利用者側がサービスを利用する個人で，運用者側がサービスを提供する IT 企業である．運用者側である IT 企業には，エンジン開発者，データオーナー，訓練者，テスト実施者の関係者が含まれる．利用者は，IT 企業との間で信頼を構築できればよい．コンシューマ向けの AI サービスでは，サービスの機能やセキュリティを利用者が厳密にチェックすることはなく（例えば AI スピーカーで収集された音声がどう分析されるかあまり気にせず），いわば企業を信頼することでサービスを受容していると思われる．

今後 AI の利用は IT 分野以外の交通，金融，セキュリティ，法律，医療，軍事等に広がっていく．運用者側は IT 企業にかわり，交通，金融等の業種別サービス提供企業となる．サービスは社会経済活動に直結し，一部は社会基盤インフラ制御にも関わることになるため，トラスト構築には「社会受容のためのトラスト」が大きく関わってくると推測される．

サービス提供企業は，自社が保持する利用者に関するデータを AI 搭載システムに学習させ，既存サービスの利便性や効率性等を向上したいと考える．しかし，サービス提供企業は AI 搭載システムの設計・開発のノウハウを持たないため，IT 企業と連携する必要がある．現在，一部 IT 企業は自社の AI 機能を利用する API を公開しており，AI を利活用したい組織や個人はその API を使ったソフトウェアを作り，運用できる．このケースを図 1 に当てはめると，サービス提供企業が運用者，データオーナー，IT 企業がエンジン開発者と捉えることができる．また，学習やテストは，データオーナーの学習データと，エンジン開発者の AI 搭載システムを用いて行うことから，サービス提供企業と IT 企業の両者が，訓練者とテスト実施者になる．すなわち，サービス提供企業と IT 企業が連携し，AI 搭載システムの期待性能を保証することになる．

本稿では，このようにサービス提供企業と IT 企業の間で，AI 搭載システムの品質の評価や管理を行う上で，関係者間で構築しなければならないトラストに注目して，課題の抽出と整理を行う．

4. 課題の抽出と整理

本節では，サービス提供企業を「ソフトウェアを供給される側」，IT 企業を「ソフトウェアを供給する側」として説明する．ソフトウェアを供給する側は，AI 搭載システムは，クラウド事業者が提供する AI の基盤を利用して，サービス部分を開発するような構成も考えることができる．このため，IT 企業は AI 基盤を提供するクラウド事業者と，そ

の AI 基盤を利用してソフトウェアを開発する SIer に分けることができる．本論文では，簡単なケースから取り組むことを考え，まず 1 つの IT 企業が AI 搭載システムを開発するケースから検討する．

AI 搭載システムは，2.1 節で述べたように，従来のソフトウェアと異なり，期待される機能・精度での予測や判断ができるかどうかの観点も新たに含まれることになる．

これまでの AI を搭載していないソフトウェアであれば，簡易ではあるが図 2 のような流れで，開発のプロセスと動作テストが保証されていれば，ソフトウェア供給する側とソフトウェア供給される側で信頼を構築することができた．また，意図しない動作をした場合には，バグを作りこんだ責任としてソフトウェア供給する側が修正や補償を行ってきた．

図 2　ソフトウェアの開発から利用の流れ
（従来のソフトウェア）
Figure 2　From Development to usage flow of Software
(Current Software)

AI 搭載システムでは，図 3 のように，信頼できるソフトウェアとなるためにこれまでの動作テストに後に，さらに学習モデルが期待される精度・性能がでているかの「学習テスト」も必要になると，本論文では仮説を立てる．

図 3　ソフトウェアの開発から利用の流れ
（AI 搭載システム）
Figure 3　From Development to usage flow of Software
(AI Software)

学習テストでは，AI 搭載システムと，学習データを使って学習モデルを作り，その学習モデルを評価することになる．このとき，ソフトウェア供給する側が AI 搭載システムを提供し，ソフトウェア供給される側が学習データを提供する．このため，学習テストではソフトウェア供給する側とソフトウェア供給される側の両者が実施する．
このように両者で学習テストを行うことから，実運用で AI 搭載システムが学習テスト不足によって意図しない動きをすると，ソフトウェア供給する側の責任のみでなく，ソフトウェアを供給される側の責任も考えられる．さらに，学習データに問題があった場合には，ソフトウェアを供給された側に責任があることになるかもしれない．このため，ソフトウェアを供給する側と供給される側で，信頼できる AI 搭載システムとするために，どのような責任を持って学習テストを行うべきかのプロセスやルール作りが必要と考えられる．

ソフトウェアを提供する側には，AI に関わる技術者（データサイエンティストや機械学習エンジニア等）がいることから，組織内にプロセスやルールが存在するかもしれない．ソフトウェアを供給される側は，AI に関わる技術者がいることが稀であり，ほとんどの組織にプロセスやルールは存在しないと推測される．また，ソフトウェアを提供する側とソフトウェアを提供される側が，共有できるプロセスやルールに関する公的なドキュメントは現状見当らない．

以下に，ソフトウェアを供給する側とソフトウェアを供給される側が協同でトラストを構築するプロセス・ルールについて，これまでの調査・検討から抽出した項目を挙げる．ただし，現在も調査・検討中であり，以下の項目で十分に網羅できていると考えていない．

- ソフトウェア供給する側
  - 学習モデルの適切な評価
    学習モデルの構築に用いた AI アルゴリズムに応じた適切な評価手法をとっているか．
  - 学習モデルに用いた学習データの適性
    学習モデル構築に役立つ適切な学習データが選択されているか（適切な特徴量の選択）．
  - 学習モデル構築に必要なデータ量の見積もり
    学習テストにおいて，学習モデルを構築するために必要な学習データの量はどれくらいか．
  - 学習モデル検証に必要なデータ量の見積もり
    学習テストにおいて，学習モデルを検証するために必要な検証データの量はどれくらいか．
  - 学習モデルへの攻撃への評価
    学習モデルに対して考えられる攻撃に対策しているか．
- ソフトウェア供給される側
  - 期待される精度・性能の明確化
    提供サービスに求められる信頼性から AI 搭載システムで期待される精度や性能を明確にしているか．
  - 学習テスト完了（本番環境適用）の要件の明確化
    本番環境に適用するために，学習テストで何を何処まで行えばよいかの要件を明確にしているか．
  - 偏りのない網羅性のあるデータ
    学習データとして，偏りのない網羅性のあるデータであるか．
  - 欠損値のないデータ
    学習データに欠損値がないように，適切な欠損値の処理を行ったか．

5. 具体的な課題

本節では，4 節で課題を抽出・整理していく中で得られた，ソフトウェア供給する側とソフトウェア供給される側の両者で協力して検討していくべき具体的な課題について，以下に例に基づいた仮説を挙げる．

- （例１）現場で性能がでない

テストデータやシミュレーションで十分性能を確保した学習モデルを，サービス運用の現場に適用した際に，必ずしも同等の性能が出るとは限らない．

- （例２）現場で性能が低下していく

サービス運用において，時間が経つにつれて学習モデルの性能が低下して，期待できる精度や性能を満たせなくなるかもしれない．そのため，再学習しつづけないといけない．

- （例３）学習しても不完全である

不完全ということを前提として，AI を利用するということ供給される側が理解して受け入れる必要がある．従来のソフトウェアの利用とは違う意識が必要になる．技術面だけでは捉えきれない課題である．

6. おわりに

本稿では「製品としての AI のトラストをどう構築するか？」の問いをたて，特に機械学習機能を備えた AI 搭載システムについて，AI エンジン，学習データ，学習プロセスに関して考慮すべきトラストの課題を具体化した．また，具体化した課題に関して，仮説となる検討していくべき項目をいくつか示した．

特に AI 搭載システムの業務への適用を行う場合，提供者と利用者が共同で期待性能を定義し，協力しながらシステムの信頼を高めていく必要がある，という仮説を提示した．従来の CC (Common Criteria) 等に依拠するサプライサイドのトラスト構築とは異なるアプローチが求められる，という点を重視したものである．

AI を有効に利活用したい企業は，AI 搭載システムを導入したら終わりではなく，環境変化等に合わせ，継続的な期待性能の再定義・学習が必要になるだろう．これを「社会的受容のトラスト」の観点からみると，利用者は学習機能つきの AI を完成品とみるべきではなく，人材と同じように継続的なチューニング（育成）を行うもの，と認識すべきであるかもしれない．今後，本稿の課題や解決のアプローチに関し，ケーススタディや有識者ヒアリング等を通じて検証していきたい．

参考文献

[1] European Union, “General Data Protection Regulation,” https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC（参照 2019-2-1).
[2]IEEE, “ETHICALLY ALIGNED DESIGN, v2”, https://ethicsinaction.ieee.org/（参照 2019-01-26）.
[3]Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, Fosca Giannotti, "A Survey of Methods For Explaining Black Box Models," ACM Computing Surveys (CSUR) Surveys Homepage archive. Volume 51 Issue 5, January 2019, Article No. 93.
[4]DARPA, “Explainable Artificial Intelligence (XAI),” https://www.darpa.mil/attachments/XAIProgramUpdate.pdf(参照 2019-2-1).
[5]IEEE, “IEEE P7000 Working Group,” http://sites.ieee.org/sagroups-7000/ (参照 2019-2-1).
[6]ISO, ”ISO/IEC JTC 1/SC 42 Artificial intelligence,” https://www.iso.org/committee/6794475.html (参照 2019-2-1)
[7]AI 社会推進会議，“報告書 2018－AI の利活用の促進及び AI ネットワーク化の健全な進展に向けて－,” http://www.soumu.go.jp/main_content/000564147.pdf (参照 2019-01-26).
[8] Google，“AI at Google: our principles,” https://www.blog.google/technology/ai/ai-principles/ (参照 2019-01-26).
[9] Microsoft，“Microsoft AI principles,” https://www.microsoft.com/en-us/ai/our-approach-to-ai (参照 2019-01-26).
[10] IBM，“Everyday Ethics for Artificial Intelligence,” https://medium.com/design-ibm/everyday-ethics-for-artificial-intelligence-75e173a9d8e8 (参照 2019-01-26).
[11] The European Commission’s HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE，“Draft Ethics Guidelines for Trustworthy AI,” https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56433 (参照 2019-01-26).
[12] 石川冬樹，“AI 時代における品質保証のチャレンジ～機械学習の難しさと（AI による）テスティング,” http://research.nii.ac.jp/~f-ishikawa/work/1807-ESTIC18-AI+Testing.pdf (参照 2019-01-26).
[13] 明神智之，“AI 搭載システムの品質保証,” http://jasst.jp/symposium/jasst18tokyo/pdf/C5-1.pdf (参照 2019-01026).
[14] 桑島洋，安岡宏俊，中江俊博，“機械学習モデルを搭載したセーフティクリティカルなシステムの品質保証,” https://swest.toppers.jp/SWEST20/program/pdfs/s2b_public.pdf (参照 2019-01026).
[15]科学技術振興機構,“（戦略プロポーザル）AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立,” https://www.jst.go.jp/crds/report/report01/CRDS-FY2018-SP-03.html (参照 2019-2-1).
[16]小澤誠一,“セキュリティ分野における AI 活用の現状と期待,” 第 30 回 AI セミナー 人の能力を拡張するサービスインテリジェンス IoT と機械学習を専門知識の構造化技術で融合, https://www.airc.aist.go.jp/seminar_detail/docs/09bd252401da9f9b6d92fb09812932c9226746c3.pdf (参照 2019-02-03).
[17]Daniel Lord, Christopher Meek, “Adversarial Learning,” KDD2005: proceedings of the eleventh ACM SIGKDD, 2005.
[18]宇根正志,“機械学習システムのセキュリティに関する研究動向と課題,” http://www.imes.boj.or.jp/research/papers/japanese/18-J-16.pdf (参照 2019-02-03).
[19]情報処理学会,“BoF (Birds of Feather)セッション in CSS2018,” https://www.iwsec.org/css/2018/bof.html (参照 2019-02-03).