機械学習システムのトラスト構築に関する課題分析(2)
～期待性能に関する考察～

島 成佳[1]　小川 隆一[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．倫理原則の策定や標準化が試みられているが，筆者ら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に関して検討を進めている．これまで，サプライヤと利用者の共同作業の間でのトラスト構築に関して期待性能に関する課題を分析・整理した．本稿では，これらの課題を事例に基づいてトラスト構築のための合意事項について分析する．また，AI 搭載システムの導入でのトラスト構築について，セキュリティ分野のリスク分析の手法を用いて考察する．

キーワード：AI，トラスト，社会受容，期待性能

Consideration and Analysis of issues on building trust of machine learning systems about expected capability and accuracy

Shigeyoshi Shima[1]　Ryuichi Ogawa[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. We analyzed the trust development problems of machine learning systems from the viewpoint of trustworthy product (or service), regarding capability/accuracy and security. In this paper, we analyze agreement items for trust development of those problems based on usage cases. We consider about trust development of AI-based system with reference to risk analysis method in security domain.

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

1. はじめに

近年，ディープラーニング（深層学習）を代表する機械学習技術の発展・実用化等に伴い，AI（Artificial Intelligence：人工知能）の利活用によるサービスの創造，社会経済の発展，社会課題の解決に対する期待が高まっている．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，これらのリスク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．現在，AI の利活用が進むとともに顕在化するリスクに対して，技術・倫理・制度等の様々な観点から技術者・研究者・法律家等による議論が国内外で行われている．

筆者らは，機械学習を利用した AI に関する議論において代表的な国内外の組織の取り組みを調査し，社会が受け入れるためにコントロールされるべきリスクとして，「不適切な利用・悪用」「責任分担」「説明責任」「性能・品質」「セキュリティ」の 5 項目に整理・分類した。そして，これらの 5 項目から，トラスト構築に関わる問題として，以下の 2 つに整理・分類した[1].

(1) 社会受容のためのトラスト
(2) プロダクト（製品）としてのトラスト

筆者らは，既に AI 機能を搭載したシステム（以下、AI 搭載システム）が導入・利用されている現状から，「(1)社会受容のためのトラスト」よりも，「(2) プロダクト（製品）としてのトラスト」はより短いタイムスパンで実現されるべきであると考え、後者について課題分析を行うこととした．

前回の報告[1]においては，AI 搭載システムを提供する側（サプライヤ）と利用する側（ユーザ）の間のトラスト構築過程を分析し，特に学習性能の維持向上に関する 3 つの課題を挙げた．

① 学習しても不完全である
機械学習による分析は解析的に解けない問題に適用されることが多いが、この場合起こりうるあらゆるケースを網羅的に学習することは難しい（つまり、分析の間違いは必ず起こる）．

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

② 現場で性能がでない
この結果、PoC（Proof of Concept）等の試験運用で期待された分析精度が出ないことがある．

③ 現場で性能が低下していく
さらに実環境の変化に学習が追いつかず、当初出ていた分析精度が低下することもありうる．

このうち③は②の派生的な課題とみることができるので、本稿では③を②に含めて扱うこととする．

筆者らは、機械学習の専門家に対し、実際の AI 搭載システムの導入や品質管理についてインタビューを実施したところ，上記課題が実際に AI 搭載システム導入のネックとなっていることを確認した．

期待される性能が必ずしも実運用で出ないという特性は AI に限らず、セキュリティ対策システムにも当てはまる。例えばテストでマルウェア検知率 98%を達成したとしても、実運用でこれを達成できる保証はない。セキュリティシステムは検知率等の期待性能以外の要因も含めてトラストを担保している可能性がある。本稿ではこれを手掛かりとして、AI 搭載システムのトラスト構築について考察を試みる．以下では、現場で導入が進んでいる AI 搭載システム（製品）、導入に困難が予想される AI 搭載システム、の両極端の事例を比較しながらトラスト構築に関わる分析を行う．また，AI 搭載システムの期待性能は，セキュリティ領域のセキュリティレベルのように定量的に示すのが難しいという類似した特徴を持っており，セキュリティレベルの見積もり方を参考にした分析も行う．

2. 想定する AI の機能とトラスト

2.1 想定する AI の機能

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．このため，実現イメージや実現技術（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「学習によってモデルを作り出す機能」と，「学習モデルを使って予測，分析，計画等の処理を行う機能」を持つ，機械学習技術として捉える．そして，その機械学習技術を使用した機能を含むシステムを AI 搭載システムと定義する．

AI 搭載システムは，「学習によってモデルを作り出す機能」があることで，従来のソフトウェアの演繹的開発に，帰納的開発が加わる．ここで演繹的開発とは，定義された仕様に対して，規則や処理フローに基づいて設計・開発するスタイルをさす．一方，帰納的開発とは，定義された仕様に対して，学習データに基づいて学習モデルの構築を探索的・反復的に進めて設計・開発するスタイルをさす．帰納的開発では，構築される学習モデルをテストや実運用によって，実用可能なレベルであることを実証しなければならず，この実証のための信頼の枠組みが求められている．

2.2 想定するトラストと関連する議論

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，しかし，AI 搭載システムの AI 機能では，「学習によってモデルを作り出す機能」が仕様どおり動作するだけでなく，作り出されたモデルが，定義された仕様において期待される精度で予測や判定ができること（期待性能）が求められる．また AI 搭載システムでは，上記の期待性能に加え，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．ここで学習モデルが妥当であるとは、期待性能実現のために適正に選ばれ、前処理されたデータで学習が行われたことをさす。

一般に AI 搭載システムの期待性能（分析精度）は学習モデル・学習プロセスによって変わりうるため、学習データの選定や前処理には学習アルゴリズムとの適性を考慮する必要がある．また AI 搭載システムが適用される応用分野は解析的なモデル化が困難なケースが多く、期待性能は応用分野依存，業務依存になることが多いと予想される．このため期待性能として、OS のような基盤ソフトウェアの性能指標を作るのと同等のアプローチは難しい、と予想される．

AI 搭載システムのプロダクトとしてのトラストに関連して、ソフトウェアエンジニアリングのコミュニティでは，品質評価や管理に関する活発な議論が行われている[2][3][4]．この中で、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が始まっている．

3. 本研究のこれまでの経緯

筆者らの当初の問題意識は，AI 搭載システム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．

① 学習データは信用できるか
② 学習や分析の精度・妥当性はだれがどう評価するか
③ 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 搭載システムがどういう機能を持つものか，の機能の定義，また機能の何を信頼すればいいのか？の対象範囲の定義が重要であるとの結論に至った．このため，まずは対象範囲を明確化して，その対象範囲におけるトラスト構築の課題を分析することにした．

3.1 対象範囲の明確化

対象範囲を明確化するために，AI に関する議論を行う代表的な国内外の組織の取り組みを調査し，社会が受け入れるためにコントロールされるべきリスクとして「不適切な利用・悪用」「責任分担」「説明責任」「性能・品質」「セキュリティ」の 5 項目に整理し，トラスト構築に関わる問題として，以下の 2 つに分類した．

(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(1)よりも短いタイムスパンで示す必要がある．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．

すでに見たように、ソフトウェアエンジニアリングの分野等で(2) プロダクト（製品）としてのトラストに関わる検討が始まっているが、筆者らは、「期待性能」の指標化が困難な AI 応用分野で(2)をどのように考えるべきか、に注目して分析を行うこととした．

3.2 製品としてのトラスト構築に関わる関係者

筆者らは、AI 搭載システムの開発から利用に至るまでに関わる関係者を図 1 のように表現した。

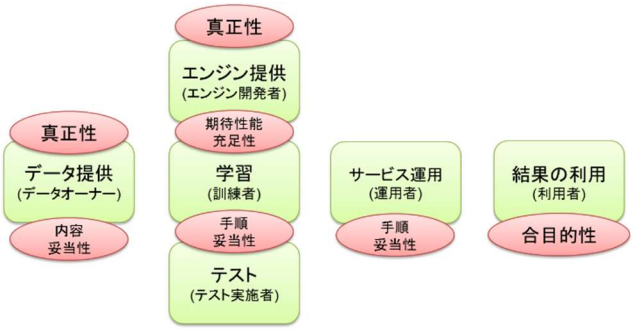

図 1　AI 搭載システムトラスト構築に関わる関係者

Figure 1 Stakeholders for Trust Building AI Software

①　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（品質を含む）が利用目的に合っているかどうか（合目的性）が関わる．

②　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．

③　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．

④　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．

⑤　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．

⑥　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．

現在，AI 搭載システムの利用は IT 分野が先行している．例えば，Google や Amazon 等の IT 企業は，スマートフォンやホームスピーカー等の AI アシスタントを通して，様々なサービスを利用者（コンシューマ）に提供している．この例の関係者を図 1 に当てはめてみると，利用者がサービスを利用する個人，運用者がサービスを提供する IT 企業となるが、運用者である IT 企業には，エンジン開発者，データオーナー，訓練者，テスト実施者等の関係者も含まれることが多い．

IT 分野以外の利用ケースにおいては，AI 搭載システムの提供を受けてサービスを運用・利用する一般企業と，AI 搭載システムを提供する IT 企業に分けることができる．この場合、図 1 のデータ提供，サービス運用，利用者は一般企業であるが，利用者が個人（コンシューマ）の場合も考えられる．また，IT 企業は，エンジン開発者，訓練者，テスト実施者にあたる．

以下では，一般企業が AI 搭載システムをソフトウェア製品やネットワークサービスの形態で提供され，運用する形態を想定する．ここで、一般企業は製品・サービスを供給されるユーザ，IT 企業は製品・サービスのサプライヤであり，両者の間におけるトラスト構築を分析する．

3.3 製品としてのトラスト構築に関する議論

AI 搭載システムは，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．
従来の AI を搭載していないソフトウェアを利用する場合，図 2 のような流れで，開発のプロセスと動作テストが正しく行われたことが第三者の審査等で保証されていれば，サプライヤとユーザとの間で信頼を構築することができた．また，保証期間中に仕様と異なる動作をして機能が実現できない場合には，ソフトウェア供給する側がバグ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 搭載システムの場合、上記のソフトウェアと同等な信頼を得るためには，動作テストに加え，学習モデルが期待される精度・性能を達成できているかの「学習テスト」が必要になるのではないか（図 3）．本項はこの仮説を立てる．

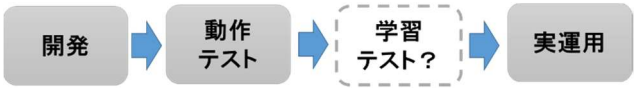

図 3　ソフトウェアの開発から利用の流れ
（AI 搭載システム）
Figure 3　From Development to usage flow of Software
(AI Software)

学習テストでは，学習データを使って学習モデルを作り，その学習モデルをテストデータで評価することになる．このとき，サプライヤが AI 搭載システムを提供し，ユーザが実運用環境を想定した学習データを提供する．すなわち、学習テストではサプライヤとユーザの共同作業によって実施される．
筆者らが機械学習の有識者 6 人に実施したインタビューでは，以下の知見が得られた．

- AI 搭載システムの導入に向けて，サプライヤとユーザで実証実験（PoC）を実施し，サプライヤがコンサルティングをしながら実証性を検証する．
- サプライヤにはデータサイエンティストや機械学習エンジニアが存在し，PoC のノウハウが蓄積されているが、ユーザ側に AI 技術者や PoC のノウハウ存在することは稀である．
- サプライヤは PoC で実現できる性能が実運用でも達成できるかどうかに不安を持ちがちである。またユーザは PoC でどこまで性能が出たら実運用に移るか、の判定方法がわからない。このため、PoC から実運用への移行ができないことがある．

PoC においてユーザが十分な量の学習データ、あるいは利用目的にあった学習データを提供できない場合、実運用で AI 搭載システムが意図しない分析や予測を頻繁に行うことは当然あり得る．学習に関するトラストの構築には、プロバイダだけではなく、ユーザにも求められることがあるのではないか。例えば、期待性能により近い性能を実現するために，ユーザが学習データ提供についてどのような責任を持ち、どのように学習テストに関わるのか、のプロセスやルール作りが必要であるかもしれない．
サプライヤとユーザが，共有できるプロセスやルールに関して，2019 年 5 月に AI プロダクト品質保証コンソーシアムから AI プロダクト品質保証ガイドラインが公開された[6]．ただし，未だ完全ではなく，今後もアップデートしていくとのことである．
このように，サプライヤとユーザで検討すべきプロセス・ルールについて調査・検討した．ただし，現在様々な組織において検討が進んでいることから，十分に網羅できていると考えていない
上記のように、AI 搭載システムの導入において PoC までは進むがその先になかなか行かない、という状況がある．その背景にある課題を整理していく中で、検討すべき重要なものとして，以下の仮説を抽出した．

① 学習しても不完全である
機械学習を利用する AI 搭載システムにおいて、学習が完全にでき、100%の精度が出るということは本質的に難しい．必ず誤った出力は発生する．この点はバグを解消すれば 100%仕様どおり動作するソフトウェアとの大きな違いであり、AI 搭載システムは誤分析のリスクを許容しない限り導入が出来ないことになる．
② 現場で性能がでない
PoC において期待性能を確保した学習モデルを実運用の現場に適用した際に，同等の性能が出る保証はない．学習データはある意味理想化されたデータセットであり、ノイズやデータ欠損、あるいは環境の変化等で性能が出ないことは十分起こり得る．
③ 現場で性能が低下していく
実運用において，環境の変化により学習対象（例えば業務フロー）の挙動が変わる、等により時間経過とともに学習モデルの性能が低下することも起こり得る．性能維持にのためには，運用と並行して再学習を続ける必要がある．

筆者らは、上記仮説が実際の課題となっているかを検証するため，AI 搭載システムの研究者・開発者等の有識者にインタビューを実施した．そこで，上記仮説が AI 搭載システムの品質管理や，PoC から実運用への移行を実際に困難にしていることを確認した．

筆者らは当初，仮説を上記の 3 つとしたが、②「現場で性能が出ない」と③「現場で性能が低下していく」は，類似した課題と考え本稿では，新たな仮説④として「現場での性能低下」を置き，以下で仮説①④について，分析を進めることとする．

4. 期待性能に関わる課題分析

4.1 不完全な学習

本節では，前記①④の課題について，既に AI 搭載システムの導入が進んでいるケースと導入に大きな障壁があるケースを比較し，その影響を分析する．具体的には，導入に大きな障壁があるケースとして自動運転を，AI 導入が進んでいるケースとしてスマートスピーカーやスマートフォン等の AI アシスタントを，両極端の事例として取り上げる．

- 自動運転

自動運転では，2018 年 3 月にウーバーの自動運転車が歩行者の死亡事故を発生させた．この事故の影響によりウーバーが 2019 年半ばに予定していた自動運転を使用したサービス開始を延期するとともに，この死亡事故の影響で自動運転の導入が遅れるだろうと報道された[7]．この記事のとおり，ウーバーは 2019 年 8 月現在自動運転を使ったサービスを行っていない．この死亡事故は，自動運転の学習を強化している際に，現場である公道で，AI の出力に基づく誤った制御によって発生したものである．この死亡事故の発生後も公道での学習を続ければ，死亡事故の発生可能性を低くできると推測できる．しかし，学習を進めても学習は完全にならないため，死亡事故が発生する可能性が常にある．このことは，社会的な影響が大きいことから，現状公道での AI 単独による自動運転は難しいと考える．このため，AI に基づいた制御の誤りを低減するため，必要な場合人が介入して操作するアプローチもとられようとしている．このように，AI 単独ではなく，AI の誤りに起因する事故発生防止手段を組合せることで，AI 搭載システムの信頼性を高めようとしている．ただし，どの程度まで事故発生防止手段が必要であるかは，社会的な議論が必要と考える．

- AI アシスタント

AI アシスタントは，利用者の期待する回答が返ってこなくても普及し続けている．また，これまでに使用中止のような報道は見られない．2018 年の AI アシスタントの各社の回答率をみても，学習が完全とは言えないが[8]，現場であるユーザ宅や，スマートフォンで使われている．さらに，ユーザが使った際の誤りを学習させて，AI アシスタントの学習をさらに進めている．AI アシスタントから期待する回答が返ってこない場合は，ユーザにやり直しコストが発生する．このようにユーザには影響はあるが，自動運転のような社会的な影響まではない．

これら 2 つの事例から，期待性能と影響度の関係を図 4 に整理する．

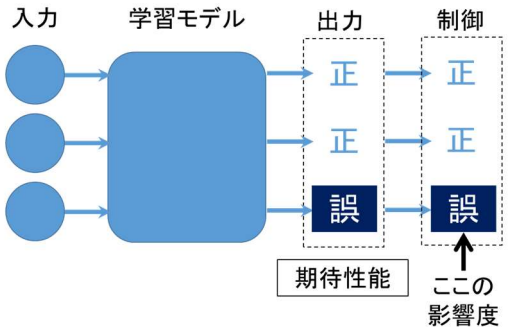

図 4 期待性能と影響度の関係

Figure 4 Relation of expected capability and impact

期待性能は出力結果の精度と捉えると，期待性能が高いほど正しい制御が行われる可能性が高くなる．事例を分析すると，AI 搭載システムの導入に関わるのは，期待性能の精度ではなく，AI の誤った出力によって誤った制御が行われたときの影響度である．期待性能がいくら高くても，誤った制御の影響が大きいと AI 搭載システムの導入が難しい．誤った制御の影響に関しては，事例で自動運転の人命，AI アシスタントのやり直しのコストを示したが，他にも，重要インフラのインシデントによる社会経済活動の停止や，事故発生による金銭的な損失なども考えられる．

サプライヤとユーザは，不完全な学習を前提として，AI の誤った出力によって誤った制御が行われたときの影響について，AI 搭載システム搭載システムの導入で許容できるかどうかを合意（トラスト構築）する必要がある．

4.2 現場での性能の低下

現場での性能が低下する事例として，AI 搭載システムの導入が進んでいるセキュリティ領域のサイバー攻撃対策について分析する．サイバー攻撃対策の AI 搭載システムとしては，IDS やアンチウイルスソフト等がある．

サイバー攻撃は攻撃者の戦略や技術，手順が変化したり，新たなマルウェアや攻撃ツール等が使われるようになったりすることから，これらに対応するために AI 搭載システムは新たな学習が必要となる．例えば，マルウェア検知では新種や亜種のマルウェアが出現すると，IDS やアンチウ

イルスソフトの検知率が低下する．これは攻撃者がマルウェアを検知されないように作るからである．機械学習を使った攻撃検知のアラートに関しては，適合率（攻撃と判断されて，実際に攻撃であった割合）や，再現率（実際の攻撃が，攻撃と判定される割合）が低下することが知られている[9].

自動運転であれば，現場での性能低下は公道での死亡事故につながる可能性が高まることから導入の際の障害となる．しかし，セキュリティ領域のサイバー攻撃対策では，性能低下はサイバー攻撃を侵入させてしまい，セキュリティインシデントを引き起こす可能性を高くするにもかかわらず，AI の導入が進んでいる．

ここで，サイバー攻撃対策で，AI 搭載システムを導入する場合と導入しない場合の対策効果について分析してみる．IDS やアンチウイルスソフトは，システム間の通信やシステム内の処理を監視することでサイバー攻撃を検知する．現状のシステムの通信量や処理量を考えると，通信や処理の監視には，IDS やアンチウイルスソフト等を利用しなければ，人手のみで監視することは不可能である．例えば，セキュリティオペレーションセンター（SOC: Security Operation Center）では，IDS 等のセンサーが不正な通信を検出してアラートとして出力し，そのアラートを分析官が確認して攻撃かどうかを判断する．現状アラートも大量に発生しており，誤検知のアラートを削減にも AI が活用されている．

サイバー攻撃対策では，AI を活用しなければ，通信や処理を監視しきれず，サイバー攻撃を見逃す可能性が高まる．これにより，見逃しによるセキュリティインシデントが発生する可能性も高まる．このため，AI 搭載システムの性能が低下したとしても，導入効果の方が高い．また，誤検知のアラートの削減の性能が低下すると，分析官の確認作業が増加してしまうが，AI 搭載システムを導入しないよりもはるかに確認作業の量は少ない．

サプライヤとユーザは，現場で性能が低下する環境において，AI 搭載システムを導入すると，導入しないときよりもメリットがあることを合意（トラスト構築）する必要がある．

5. 期待性能のトラスト構築

本節では，セキュリティ領域において，サプライヤとユーザのトラスト構築について分析するため，具体的な 1 つのセキュリティ業務を前提として期待性能について検討してみた。ただし，そのセキュリティ業務については，AI を適用して実現できるかどうかの議論していない．また，本検討には，AI を活用してセキュリティ業務の効率を向上させたいと考えるセキュリティ担当者の協力を頂いた．

本検討では，トラスト構築のゴールとして，サプライヤ（筆者ら）と，ユーザ（セキュリティ担当者）で，期待性能の合意に至るまでの検討を試みた．検討は，以下の 2 段階で行った．

(1) AI をどのような業務に活用するか（期待する機能）
(2) AI にどの程度の精度を求めるか（期待する性能）

(1)で想定したセキュリティ業務は，セキュア開発での仕様書確認に関する業務で，仕様書がセキュリティ標準やガイドライン等に対応できているか，セキュリティ機能もれをチェックするものである．具体的には，セキュリティ標準やガイドラインに基づいて，以下のような項目ができることを想定した．

- 漏れているセキュリティ機能の抽出
- 抽出理由
- 対応方針

(2)では，上記の項目をどの程度の精度で実現するかを検討した．当初，人手でのチェックともに AI 搭載システムを使うことで効率化するという方針で精度について検討を進めていった．しかし，AI でのチェックが不完全であえば，人手での仕様書全体をチェックしなおすことになる．人がもう一度をチェックしなおすと，AI 搭載システムを導入しても業務の効率化につながらないことため，100%の精度を求められた．本稿で述べているように「学習しても不完全である」ことから，100%の精度は達成できない．このため，残念ながら期待性能の合意が得られなかった．

ここで，同じセキュリティ領域なのに，4.2 節のサイバー攻撃対策の SOC 業務では AI の活用の活用が進んでいるのに，セキュア開発での仕様書確認業務では適用が難しかったのかを分析してみる．SOC 業務では，IDS 等のセンサーによる攻撃検知やアラートの削減等，分析官の確認作業の前処理に活用している（一部の作業を担っている）．このため，精度が 100%でなくても，精度が高ければ高いほど人手の作業の効率化になる．しかし．仕様書確認業務では，人手の作業をそのまま肩代わりしようとした．このため，人手の作業の効率化には，100%の精度でなければ，人がもう一度確認する必要があり，効率化が図れない．このことから，AI 搭載システムが人手の業務をまるごと肩代わりするのではなく，業務の一部を担うような使い方を考えるべきとの知見を得た．

参考として，仕様書確認に機械学習を適用することが困難であるとの結論となったので，議論内容について触れる．仕様書の確認では，セキュリティ標準やガイドラインだけでなく，サービス内容や，構築に用いたシステム，組織の運用ルールなどの背景も確認に用いる．このため，仕様書が様々なルールと整合しているかの確認であり，適切な組

合せ問題として捉えることができる．このため，演繹的な処理の方（最適化や推論等）が合っていると考える．もし，機械学習のように帰納的な処理で実現しようとすると、学習データを少なくともルールの組合せ分を作らなければならず，作りきれない．また，学習データとして，様々なルールに基づいた仕様書を作ったり，集めたりすることは現実的でないと考える．

6. セキュリティ領域から見た考察

6.1 AI とセキュリティとの類似

AI 搭載システムの期待性能は，学習データの質や量に関係することや実運用において性能が低下すること等から，定量的な評価が難しい．この問題は，セキュリティ領域で適切なセキュリティレベルを定量的に示すことが難しいという問題と類似している．また，4 節の分析のように，AI に求められる期待性能は，適用領域で異なる．システムに求められるレベルも，事業分野や適用業務で異なる．この点でも適切なセキュリティレベルとも類似している．

セキュリティ領域では，適切なセキュリティレベルを定性的または半定量的に示すことが一般的である．定量的に表すことが難しいものの，ほとんどの組織では，適切なセキュリティレベルを決めて（ポリシー策定して），ICT システムを運用している．すなわち，ポリシー策定がセキュリティレベル設定のために重要なプロセスである．このことから，筆者らは，組織が適切なセキュリティレベルを決めるプロセスが AI 搭載システムの期待性能を決めるプロセスに参考になるのではないかと考えた．

セキュリティ領域では，事業分野や組織で，適切なセキュリティレベルが異なる．例えば，重要インフラの ICT システムは，一般企業の ICT システムに比べて，高いセキュリティレベルが必要である．

このセキュリティレベルは，リスクアセスメントにて，対象脅威を明確にして，その脅威が組織に与えるリスクの大きさを算出することで決まる．リスクアセスメントのプロセスや手法について，NIST SP-800-30[10]を参考にすると，リスクアセスメントでは，大きく「リスク因子の定義」と「リスク分析」の 2 つに分けることができる[11].「リスク因子の定義」では，組織に影響を及ぼす可能性のある脅威を洗い出す.「リスク分析」では，影響度や発生確率から見積もる．

リスク分析は，定量的なリスク分析や定性的なリスク分析の手法があるが，今回は簡単化のため，図 5 のリスクポートフォリオを利用する．リスクポートフォリオは，影響度と発生頻度で，図 5 のように，4 つの象限に分かれる．

- リスク保有：リスクの発生確率が低く，影響度が小さいことから，リスクを許容できる．新たな対策を講じる必要はない．
- リスク低減：リスクの発生頻度が高く，影響度が小さいことから，対策を講じて発生頻度を低減することで，リスクを許容できる．
- リスク回避：リスクの発生度が高く，影響度が大きいことから，リスクを許容できない．リスクが許容できないため，利用しないことで，脅威の発生を取り除く．
- リスク移転：リスクの発生度が低いが，影響度が大きいことから，保険等にリスクを移転できれば，リスクを許容できる．ただし，リスクをすべて移転できるとは限らない．

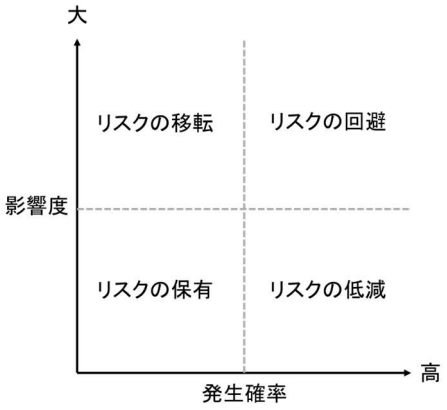

図 5　リスクポートフォリオ
Figure 5　Risk Portfolio

6.2 期待性能と誤分析リスクの統制

AI 搭載システムの誤制御のリスクにおいて，期待性能はリスク低減の 1 要素であると考え，リスク分析の概念を参考にして見積もることを述べる．

AI 搭載システムにおいて，リスクアセスメントの「リスク因子の定義」は，4.1 節で述べた不完全な学習を前提として，トラスト構築の際に AI からの誤った出力結果によって，誤った制御が行われたときの悪影響の洗い出しがと捉える．

また，リスクアセスメントの「リスク分析」は，AI 搭載システムにおいて，発生確率を期待性能，影響度を誤った制御の際の影響，と捉えることができる．このため，AI 搭載システムも，リスクを見積もることができるではなかと考える．ここで，期待性能は，誤った制御の発生確率に関連する 1 つの項目として捉える．リスク低減では，他の対策も講じることで，発生確率を低くする．他の対策としては，セキュリティ領域で言われるように，技術・制度・人によって低減できると考える．4.1 節の自動運転でも，事故に備えて人も制御できることで，導入のアプローチをとろうとしている事例から，そうであると言える．

また，影響度として，セキュリティでは多層防御の概念があり，インシデントの影響を小さくする仕組みが実装さ

れている．AI 搭載システムでも，多段階の影響度の低減の仕組みが有効である．自動運転のケースは、安全工学の知見が参考になる．

このことから，AI 搭載システムの導入する際に，図 5 のように，AI 搭載システムの期待性能とその他の施策を組み合わせた発生確率と，事故発生防止等を考慮した影響度からリスクを見積もる検討を行うことによって，サプライヤとユーザの間のトラスト構築に貢献すると，筆者らは仮説を立てる．

今後，この仮説が，AI 搭載システムの導入に向けた検討に貢献できるかどうかを検証していきたい．この際，この仮説を具体化したのちに有識者等に確認したいと感がる．さらに，貢献することが確認できれば，リスクポートフォリオ以外のツールも活用できないかを検討していきたい．

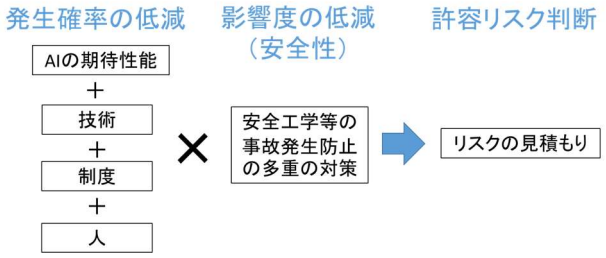

図 5　トラスト構築のためのリスクの見積もり

Figure 5　Risk estimation of trust development

また，セキュリティ領域では，事業分野ごとで，求められるセキュリティレベルをガイドライン等で決めていることから，AI の期待性能に関しても，事業分野ごとにリスクアセスメントに関するガイドラインを作ることにより，事業分野での AI 搭載システムの議論がしやすくなると考えられる．

セキュリティ領域では，セキュリティインシデントの発生する可能性が低いが影響が大きい場合は，リスクの移転として，サイバーセキュリティ保険がある．実際，サイバーセキュリティ保険が販売されており，利用が広がっている．このことから，同様に AI 搭載システムに普及に向けて，事故発生を前提として，リスク移転のための保険を用意する必要でてくるかもしれない．また，セキュリティインシデントが発生した場合には，説明責任が求められる．同様に，AI 搭載システムでも事故が発生したときに説明責任が求められるのではないかと考える．

7. おわりに

本稿では，前回の論文で示した 3 つの課題について，有識者へのインタビューを通して，AI 搭載システムを現場に適用する際の課題であることを確認した．この際，「学習しても不完全である」と「現場で性能が低下していく」2 つが本質的な課題であると整理した．

この 2 つの課題に基づいて，AI 搭載システムの導入が進む領域と導入が進んでいない分野を比較して，トラスト構築に関わる項目について分析した．

また，トラスト構築に関わる項目をセキュリティ領域のセキュリティレベルが，AI 搭載システムの性能問題と類似していることを発見し，セキュリティレベルを決めるときのリスク分析を AI 搭載システムの導入のリスク分析に適用した．そして，サプライヤとユーザで AI 搭載システムの導入のリスクを見積もることが，トラスト構築に貢献すると仮説を立てた。

今後，この仮説を具体化していくとともに，実際に使うことができるかどうかを有識者や現場の方と検証していきたい．

参考文献

[1]小川隆一，島成佳，“機械学習システムのトラスト構築に関する課題分析” 研究報告セキュリティ心理学とトラスト（SPT）,2019-SPT-32(21),1-6 (2019-02-28) , 2188-8671(参照 2019-08-20).

[2] 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 2019-08-20).

[3] 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 2019-080-20).

[4] 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 2019-08-20).

[5] 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 2019-08-20).

[6] AI プロダクト品質保証コンソーシアム，“AI プロダクト品質保証ガイドライン 2019.05 版,” http://www.qa4ai.jp/QA4AI.Guideline.201905.pdf (参照 2019-08-20).

[7] BUSINESS INSIDER, “Self-driving cars could face a 'huge setback' after the tragic death of a woman struck by an autonomous Uber,” https://www.businessinsider.com/uber-self-driving-car-death-could-hurt-adoption-2018-3 (参照 2019-08-20).

[8] Perficientdigital, “Rating the Smarts of the Digital Personal Assistants in 2018,” https://www.perficientdigital.com/insights/our-research/digital-personal-assistants-study (参照 2019-08-20).

[9] Michael Bierma, Justin E. Doak, Corey Hudson, “Learning to Rank for Alert Triage、” https://www.osti.gov/servlets/purl/1365098 (参照 2019-08-20).

[10] National Institute of Standards and Technology, “NIST Special Publication 800-30 Revision 1, Guide for Conducting Risk Assessments”

[11] 島成佳，芦野佑樹，佐々木良一，“リスクポートフォリオに基づいたサイバーセキュリティ対策選定プロセスの提案,” Computer Security Symposium 2014 (参照 2019-08-20).