成果発表会資料

ゴール指向要求分析とシステム安全分析を利用したAIシステム品質の個別ガイドライン導出方法の提案

Individual Guideline Derivation Method in AI System Quality Assessment by use of Goal-Oriented Requirements Analysis and System Safety Analysis

2021年2月26日

日本科学技術連盟 2020年度 ソフトウェア品質管理研究会
研究コース５「人工知能とソフトウェア品質」

研究員 [*AI Quality Fairness*チーム]：
相津 一寛（パナソニック株式会社）
小宮山 英明（コニカミノルタ株式会社）
柳原 靖司（ブラザー工業株式会社）
指導員：
主査　石川 冬樹（国立情報学研究所）
副主査　栗田 太郎（ソニー株式会社）
副主査　徳本 晋　（株式会社富士通研究所）

Agenda

1. 本研究の概要
2. 背景と課題，解決策の提案
3. 提案手法の説明
4. 実験
5. 考察・まとめ

１．本研究の概要

AIの品質保証をする個別ガイドライン*IGDM-AIQA*法の提案

(Individual Guideline Derivation Method in AI system Quality Assessment)

現状

※個々の知見の汎化

A社
プロダクトα

B社
プロダクトβ

C社
プロダクトγ

公になっているガイドラインは
各社AI専門家の知見を集約
抽象度が高くQA担当の活用が困難

提案

開発対象

※システムの要求やステークホルダの利害関係を具体化

要件1

要件2

要件3

個別の品質保証ガイドライン
（サブガイドライン）

AGORAやFRAMで分析した結果
から個別ガイドラインを導出

開発対象であるAIシステムの要求から個別にガイドラインを導出する方式（***IGDM-AIQA法***）を考案，仮想FinTechシステムで有効性を検証

2．背景と課題，解決策の提案

AIの品質を保証するためには個別ガイドラインが必要

既存
ガイドライン

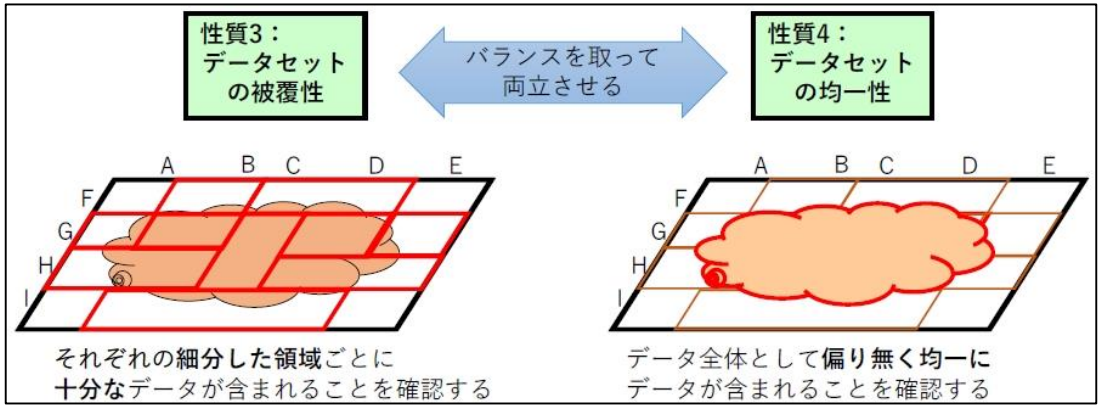
産総研
機械学習品質マネジメント
ガイドライン
(2020.06)

QA4AI
AIプロダクト品質保証
ガイドライン
(2020.08)

既存ガイドラインの問題

- AIの知識がある開発者向け
- 幅広い応用分野の共通事項

➡ 内容が抽象的でQA担当者には難解
個別システムの品質の要諦があいまい



産総研ガイドラインより

目的システム向けに具体化・詳細化した個別ガイドラインが必要

解決策の提案：
要求工学の知見により，AI品質ガイドラインを目的別最適化

システムの要件抽出・モデル解析プロセスの中に，帰納的開発の知見を
取り入れて，サブガイドラインを導出するフレームワーク

⇒ ***IGDM-AIQA法*** *(Individual Guideline Derivation Method in AI system Quality Assessment)*

AI品質の汎用ガイドライン
機械学習品質マネジメントガイドライン[3]，
AIプロダクト品質保証ガイドライン[4]等

AI品質の目的別サブガイドライン
本研究では，サブガイドラインと呼ぶ

現状

機械学習の有識者向け

提案

提案手法
IGDM-AIQA法

サブガイドラインの
導出フレームワーク

個別の品質保証ガイドライン
（サブガイドライン）

品質保証部門のQA担当者向け

3．提案手法の説明

IGDM-AIQA法の特徴

- 汎用ガイドラインの要旨を考慮しながらゴール指向で要件展開 **[STEP1]**
  （リスク回避性，AIパフォーマンス，公平性，その他一般的性質）
- 目的システムの学習データに対する推論特性の分析 **[STEP2]**
- FRAMモデリング技術を利用したステークホルダの機能連関分析 **[STEP3]**

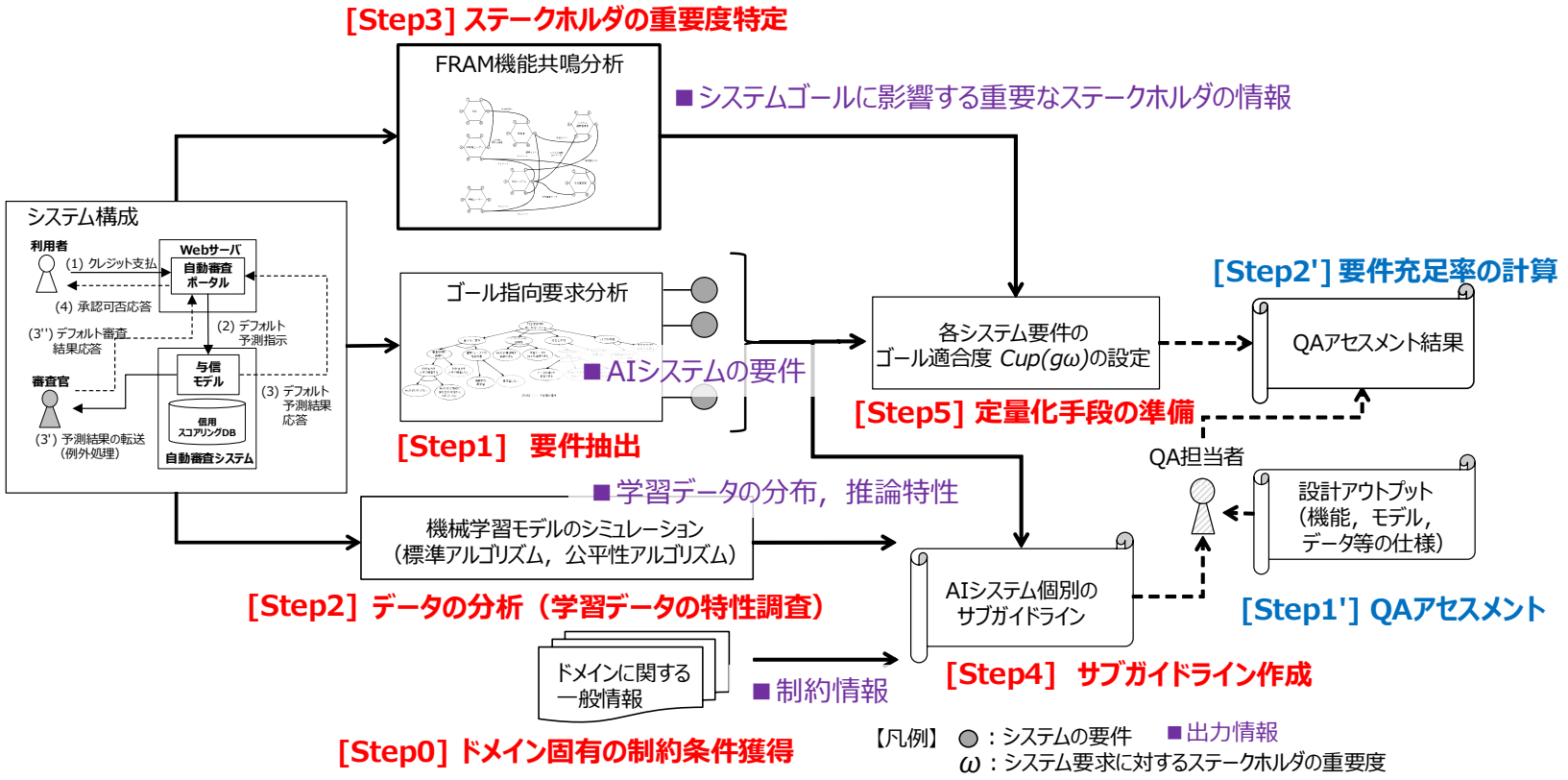


図3-1　AI品質の目的別サブガイドラインの導出フレームワーク（IGDM-AIQA法）

(補足) IGDM-AIQA法の手順

[Step1]
　汎用ガイドラインを参考にしながら，要求分析法のひとつであるAGORAを用いてゴール指向で要件を導出する.
※機械学習応用システムでは要件の間でトレードオフが発生する場合があるので，要件の導出プロセスを俯瞰的に可視化しながら分析できるようにする.

[Step2]
　システムに搭載された機械学習コンポーネントが扱うデータセットに対する推論特性を調べるため，様々な機械学習アルゴリズムで問題を解きながら性能を分析する．機械学習の標準アルゴリズムのほかに，公平性アルゴリズムを適宜利用する.

[Step3]
　システムの運用フェーズを想定して，関連するステークホルダが実社会に及ぼす影響を，機能共鳴分析法（FRAM）によりステークホルダの利害関係を可視化しながら調べる.

[Step4]
　Step1～3で得られた知見をもとに，AIシステムのサブガイドラインを作成する.

[Step5]
　AGORAの満足度行列を利用してサブガイドラインと対になるシステム要件のゴール適合度を計算することで，品質アセスメントの結果を定量的に評価できるようにする.

個別システムのサブガイドライン導出

機械学習モデルを利用したクレジットカードのデフォルト予測を行う，
「FinTech与信判定システム」の事例

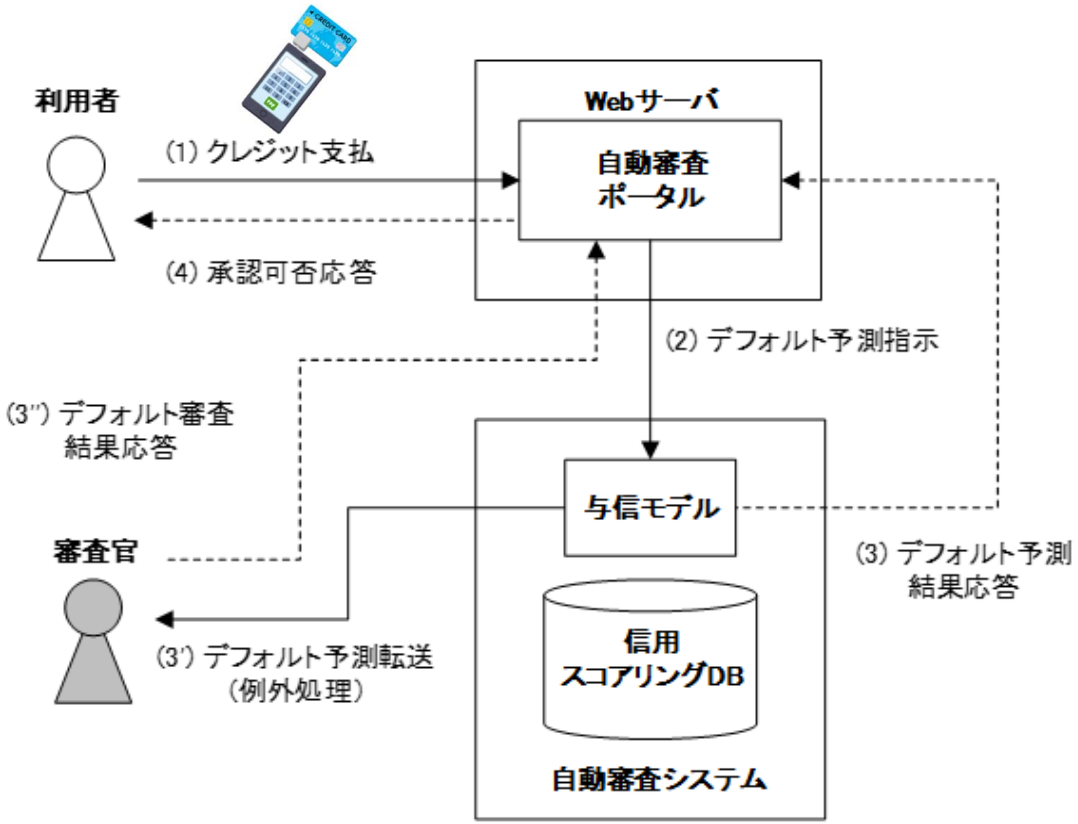


図3-2　FinTech与信判定システム（本研究のケーススタディ対象）

ゴール指向要求による要件抽出

AGORA（Attributed Goal-Oriented Requirements Analysis method）により，
10個の要件群を抽出（ゴール：社会受容性の高い与信システム）

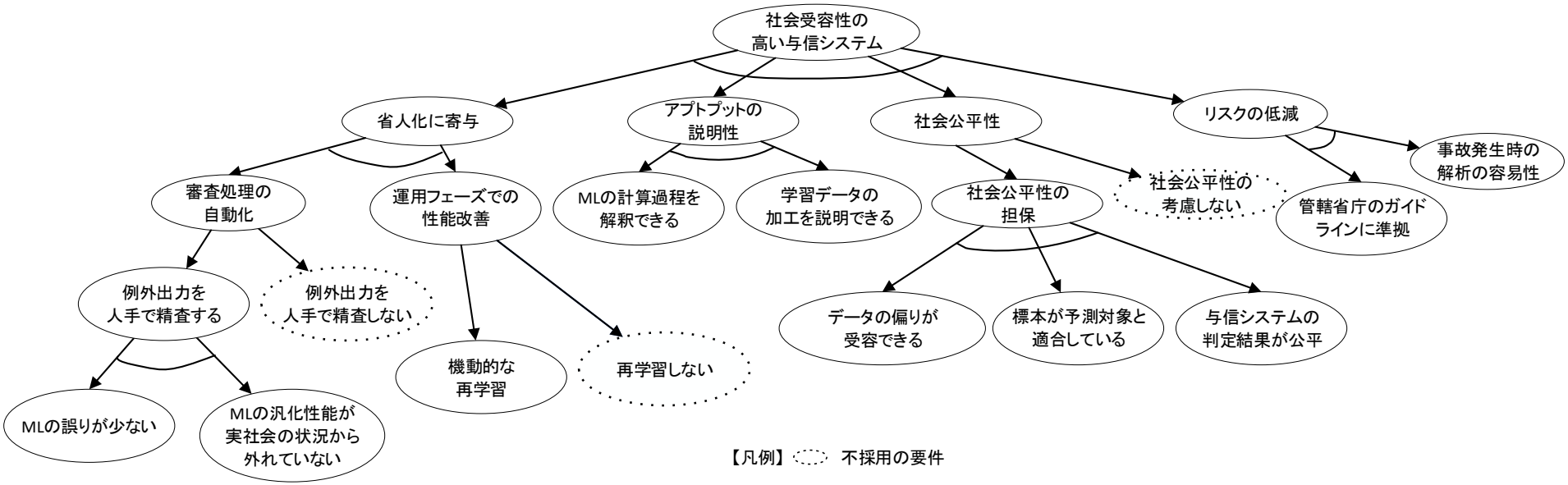


図3-3　FinTech与信判定システムの要求分析結果（AGORAによる展開図）

目的システムの学習データに対する推論特性の分析

与信システムの公平性という視点では，機械学習モデルによる判定結果の不公平性軽減をシミュレーションする

補正無しアルゴリズム（LightGBM）と比べ，Fairlearn[9]に含まれるアルゴリズム（ThresholdOptimizer/GridSearch）では，正答率は下がるが公平性は改善される

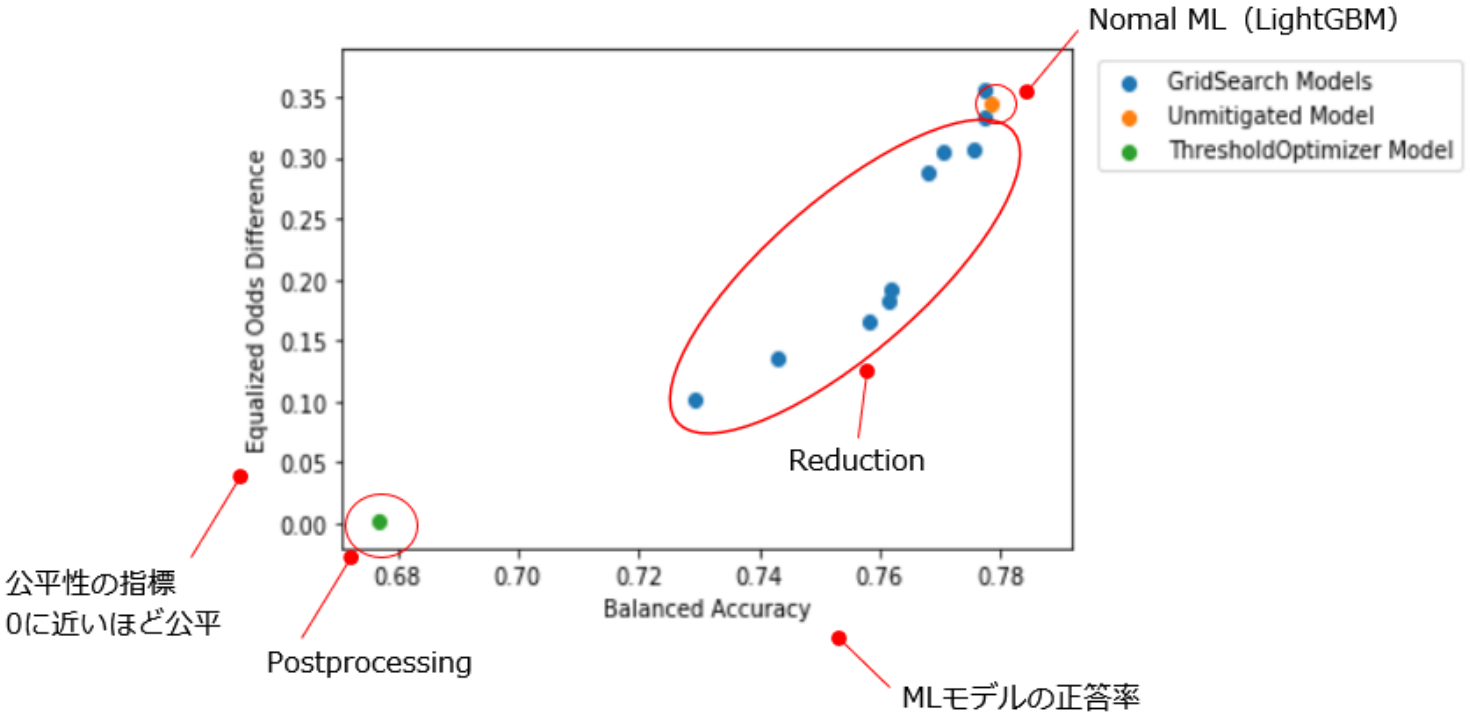


図3-4　Fairlearnを使った公平性改善のシミュレーション

FRAMを利用したステークホルダの機能連関分析

システムのゴールである「社会受容性」を高める上で「鍵」となるステークホルダの特定と，その連関についてモデル化

社会からの評判に基づいて，経営者はシステムの運用改善に対するコスト負担を行い，システム運用管理者が機械学習モデルの公平性を増長するような再学習モデルを作成して与信システムに反映させることで，非特権ユーザの公平性に関する意見が良くなり，社会的評判に還元される.

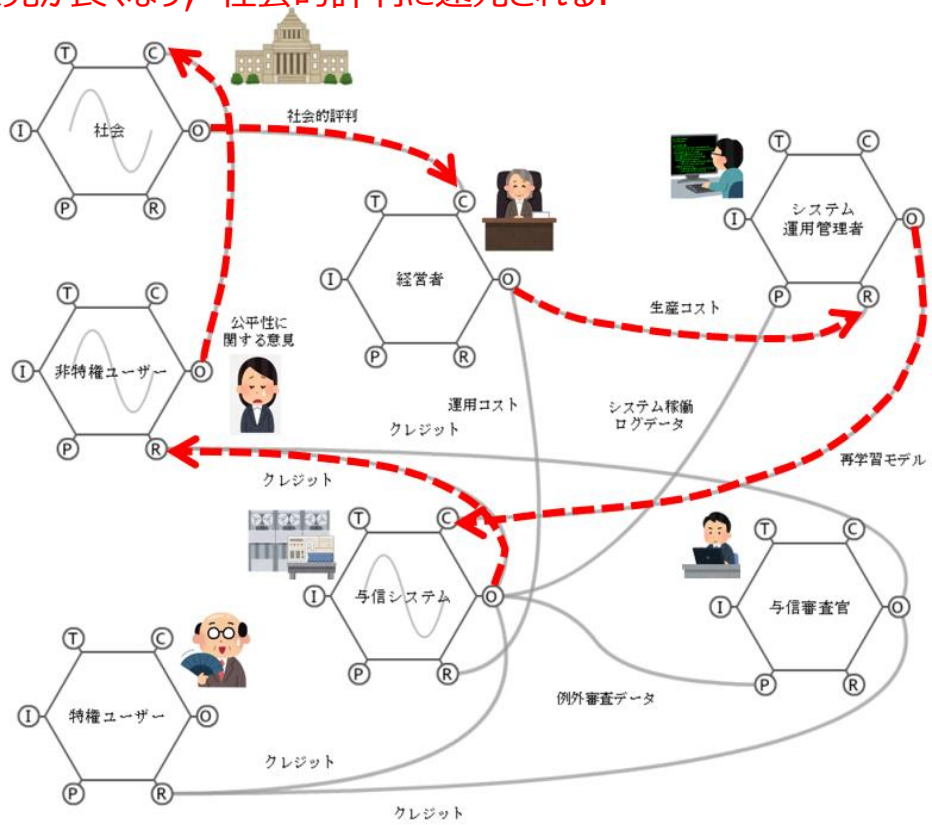


図3-5　FRAMによるステークホルダの機能連関構造

FinTech与信判定システムのサブガイドライン

主要求： 省人化に寄与する性能だけでなく，公平性に配慮した社会受容性の高い与信システム				
#i	要件	副要求	$Cup(g_{i\omega})*$	サブガイドライン（要約）
1	MLの計算過程を解釈できる	アウトプットの透明性	5.4	モデルのアルゴリズムは，説明性の高いアルゴリズムを使用しているか.
2	MLの汎化性能が実社会の状況から外れていない	省人化に寄与	4.7	モデルのアルゴリズムに含まれる汎化のために採用している制約によって，少数の重要なデータが無視されていないか.
3	データの偏りが受容できる	社会公平性	7.7	学習データの内容の分布が，偏っていないか.
4	標本が予測対象と適合している	社会公平性	6.5	学習データにクレジットカードのデフォルト予測で取り扱うすべての審査対象者のデータが網羅されているか.
5	学習データの加工を説明できる	アウトプットの透明性	3.5	正例（デフォルト），負例（非デフォルト）の不均衡を解消するため，近接データの内挿を行って，データを増やしているか.
6	MLの誤りが少ない	省人化に寄与	4.0	機械学習の推論結果に関する正答率，F1値，AUCが十分であるか.
7	与信システムの判定結果が公平	社会公平性	7.1	・学習データの目的変数の値が性別で偏っていないか. ・推論結果が不公平な結果になっていないか. ・機械学習のバイアスを補正する処理が実施されているか.
8	機動的なモデルの再学習	省人化に寄与	2.9	再学習の時間は，運用で許容できる時間以内であるか.
9	管轄省庁のガイドラインに準拠	リスクの低減	2.9	収入が低い世代の人に対してのクレジット額が高くなっていないか.
10	事故発生時の解析の容易性	リスクの低減	3.0	学習，検証データと，モデルの学習履歴が必要な時に，確認することができるか.

(*) 要件のゴール適合度： AGORA満足度行列を利用した定量化（次ページで補足）

(参考) 各要件に対するゴール適合度の計算方法

AGORAの満足度行列[10]を利用した自然言語記述要件の定量化

$$Cup(g_\omega) \stackrel{\text{def}}{=} \frac{\sum_{s \in Stakeholder, p \in Primary\ stakeholder} \omega \cdot m(g)_{s,p}}{|\ Stakeholder\ | \cdot |\ Primary\ stakeholder\ |}.$$

評価の視点（被評価者の立場）

評価者（役割）	PU	UU	OW	OM	DV	役割毎の評価基準
要件名: 与信システムの判定結果が公平						
PU	5	8	8	7	6	社会における機会均等性は理解できる
UU	8	10	9	8	6	より良い社会に向けて弱者に対する配慮が欲しい
OW	5	8	8	7	7	AI システムの性能と公平性はバランスが必要
OM	5	8	8	7	6	運用時に継続的にモデルを改善する必要がある
DV	4	7	7	6	6	技術によって不公平性を緩和することは難しい

① 満足度行列（左図）に各評価者の役割で，被評価者の視点に着目して-10～+10の素点（要件重要度）を入力する.

② $Cup(g_\omega)$を計算する. [左図例 7]
分子: 左図のグレー部分の和
[左図例 105] ※役割毎に重みωを付与

分母: 左図の実線部と破線部に含まれる要素の集合濃度の積の平方根
[左図例 15]

【凡例】PU: Privileged User（特権ユーザ），UU: Unprivileged User（非特権ユーザ），OW: Owner（経営者），OM: Operations Manager（システム運用管理者），DV: Developer（開発者）

４．実験

本研究の仮説を検証するための実験を実施

［仮説］

IGDM-AIQA法から導出されたサブガイドラインを用いれば，社会受容性を含むビジネスゴールを持ったAIシステムの品質保証のアセスメント精度が向上する．

［研究設問］

RQ1：機械学習技術に詳しくない技術者がサブガイドラインを参照すると，ガイドラインがない場合，及び汎用ガイドラインを参照した場合に比べ，システム要件に関わる欠陥指摘の精度が改善する．

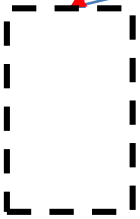
RQ2：機械学習技術に詳しい技術者がサブガイドラインを使っても，ガイドラインがない場合，及び汎用ガイドラインを参照した場合に比べ，システム要件に関わる欠陥指摘の精度は改善しない．

FinTech AI与信システムをケーススタディとして，各実験条件の品質保証のアセスメント精度を検証

FinTech AI与信システムの設計情報
（機械学習モデル・学習データ・
運用等に関する設計方針を記述）

被験者Ⅰ群
（機械学習に詳しくない）

被験者Ⅱ群
（機械学習に詳しい）

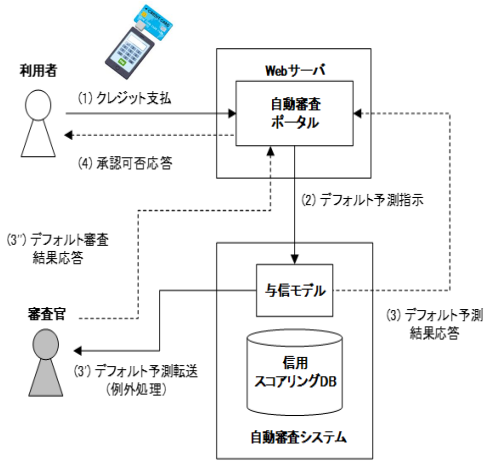


ガイドライン
参照無し

汎用ガイドライン
参照

サブガイドライン
参照

図4-1　実験の概要



業種	製造，情報・通信，金融
会社数	13社
業務上の役割	QA，開発，研究
被験者数	Ⅰ群：25名 Ⅱ群：13名

被験者が出した欠陥指摘をシステムの各要件と比較し，要件の適合度を採点する

被験者には，FinTech与信判定システムに適していない* 設計情報を提示

(*) 意図的に欠陥情報が混入されている

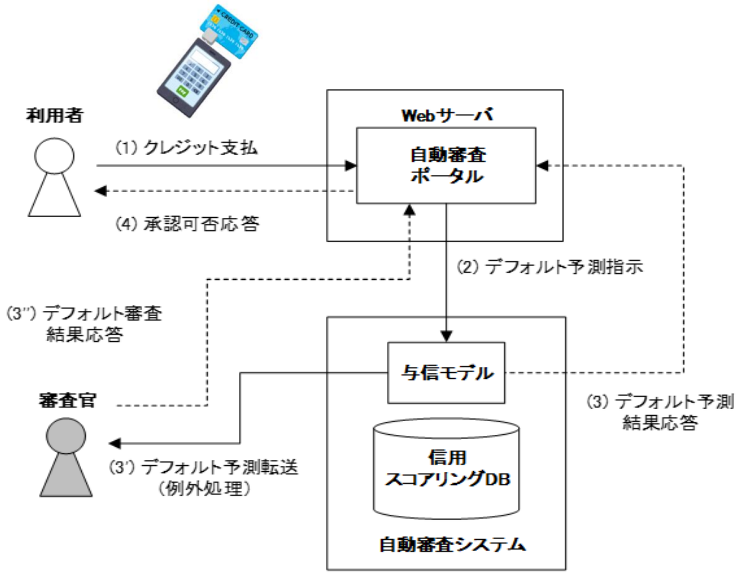


図4-2　FinTech AI与信システム（再掲）

区分	項目	説明
モデル	アルゴリズム	2項分類問題を解くためのアルゴリズムとして，表現力が高く，正答率を高めやすいDNN（Deep Neural Network）を選定した.

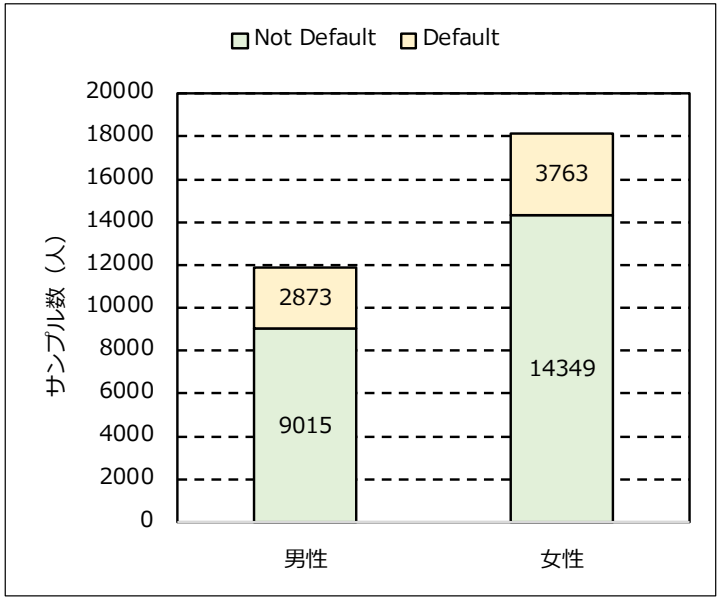


図4-3　被験者に提示した性別毎のデフォルト情報

機械学習の知識に依らず，サブガイドラインを参照すると，要件に合致した指摘ができることがt検定から言えた

各条件における被験者の回答に有意な差があるかをゴール適合度$\sum_i A_i \cdot Cup(g_{i\omega})$のサンプルをt検定（有意水準5%)によって検証した.

$$Cup(g_\omega) \stackrel{\text{def}}{=} \frac{\sum_{s \in Stakeholder, p \in Primary\ stakeholder} \omega \cdot m(g)_{s,p}}{|Stakeholder| \cdot |Primary\ stakeholder|}.$$

Ⅰ群

[ゴール適合度]	ガイドライン参照なし	汎用ガイドライン参照	サブガイドライン参照
平均	9.0	13.2	26.1
分散	17.8	37.9	102.9

Ⅱ群

[ゴール適合度]	ガイドラインなし	汎用ガイドライン参照	サブガイドライン参照
平均	16.8	19.7	30.6
分散	32.3	26.4	73.9

RQ1：
Ⅰ群はサブガイドラインを参照することで，欠陥指摘の精度が改善する.

t検定

RQ２：
Ⅱ群はサブガイドラインを参照しても，欠陥指摘の精度が改善しない.

ガイドラインなしとの差
3.01×10^{-8}（<0.05)
汎用ガイドライン参照との差
6.76×10^{-6}（<0.05)

有意な差あり

ガイドラインなしとの差
8.27×10^{-5}（<0.05)
汎用ガイドライン参照との差
1.33×10^{-3}（<0.05)

有意な差あり

RQ1の妥当性は確認できたが，RQ2の妥当性は確認できなかった

要件により，サブガイドライン参照による改善度が異なった

要件毎の回答精度の平均値（$\overline{A_i}$）で比較.

Ⅰ群

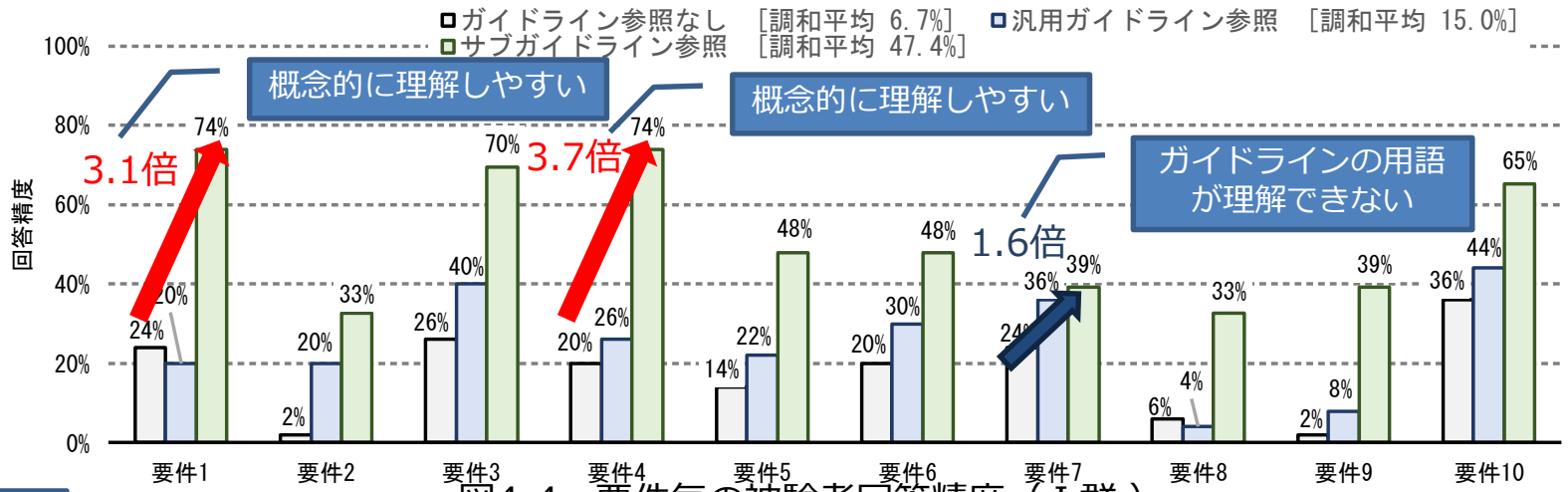


図4-4　要件毎の被験者回答精度（Ⅰ群）

Ⅱ群

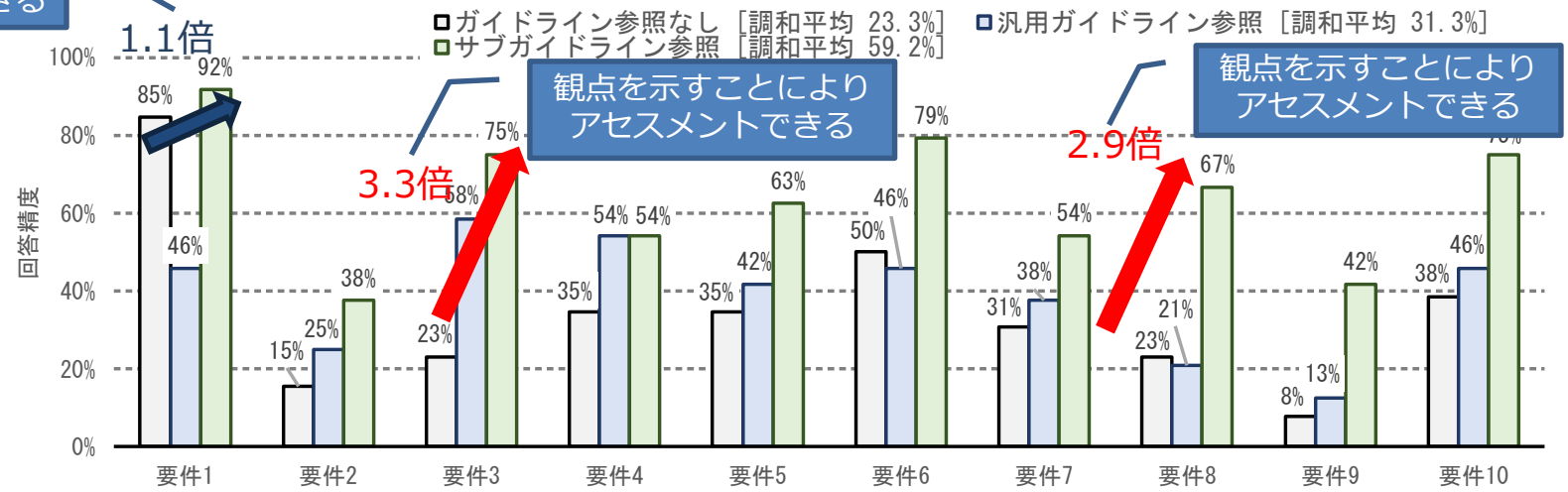


図4-5　要件毎の被験者回答精度（Ⅱ群）

機械学習技術の知識が十分でないQA担当者でも，実用レベルの品質保証アセスメント精度を出せた

QA担当者とそれ以外の役割で，被験者の回答精度の合計値（ $\sum_i A_i$ ）を比較.

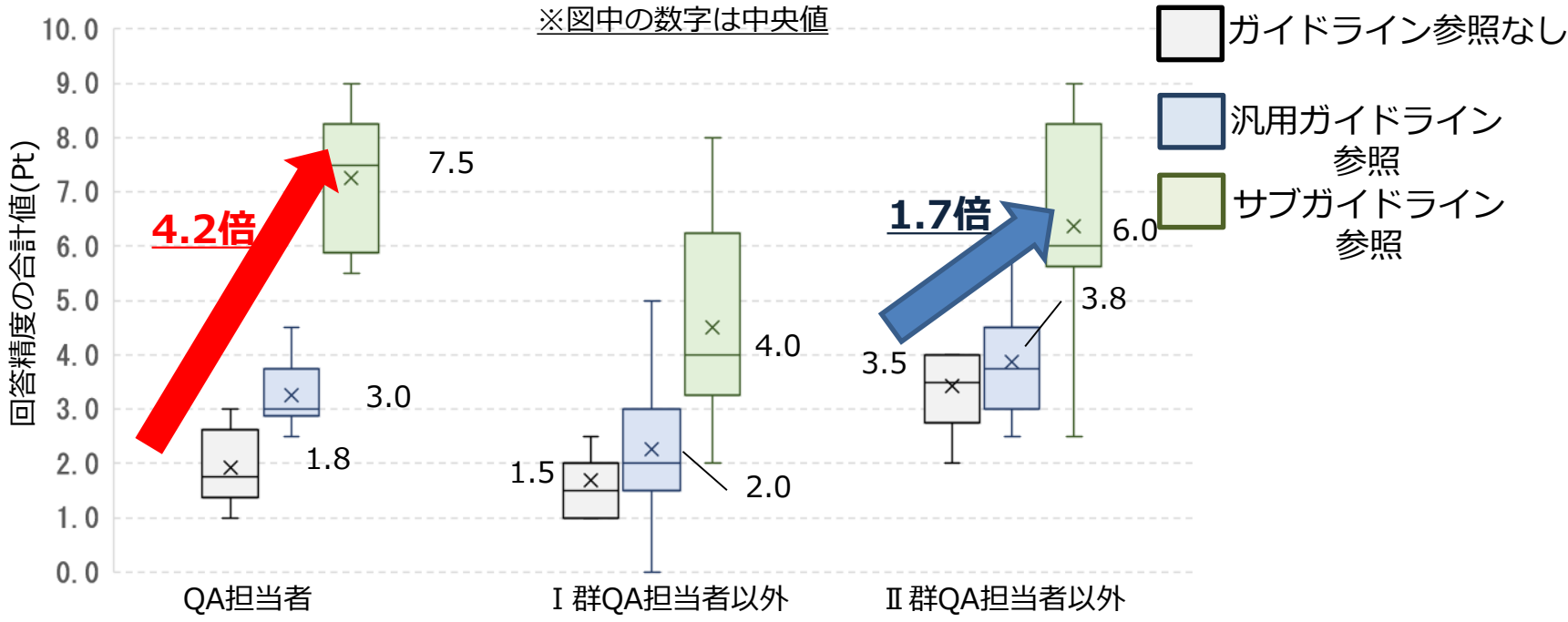


図4-6　役割に着目した被験者回答精度

QA担当者の場合，サブガイドラインを参照しながら欠陥指摘を行うと，
サブガイドラインを参照しない場合に比べて欠陥指摘精度が大幅に改善された

5．考察・まとめ

研究の結果に対する考察を示す

■仮説に対する整合性

・機械学習の知識や実務経験に関わらず，サブガイドラインに基づいて実施した方が品質保証アセスメントの精度が高まる.

■サブガイドラインの記述

・サブガイドラインを導出する際，読み手の背景知識に依存した理解度にばらつきが生じないよう，言語化の工夫が必要である.

■現場での有効性

・IGDM-AIQA法から導出されたサブガイドラインを参照するとQA担当者の本業の知見を補完しながら実務を遂行できる点で合理性が高い.

■サブガイドライン導出の再現性

・IGDM-AIQA法でのサブガイドライン導出は，機械学習，AGORA，FRAMの初学者レベルの知識を保有すれば可能である.

本研究の結論を示す

・研究では，品質保証の現場の実務で活用しやすいAIシステムの品質アセスメントのためのサブガイドラインを導出する枠組みとしてIGDM-AIQA法を提案した.

・FinTech与信判定システムを事例に本手法から導出したサブガイドラインを品質保証ケーススタディに適用し，品質保証アセスメントの精度が向上することを確認した.

・本研究では，FinTech与信判定システムを対象にIGDM-AIQA法の有効性を評価したが，特定のドメインに関係なく汎用的な手法であるため，他ドメインのシステムについても適用が期待される.

ご清聴ありがとうございました

参考文献

[1] H. Kaiya *et al*. (2002), AGORA: attributed goal-oriented requirements analysis method, 10th Anniversary IEEE Joint International Requirements Engineering Conference, pp.13-22.
[2] Erik Hollnagel, Örjan Goteman (2004), The Functional Resonance Accident Model, Cognitive System Engineering in Process Control 2004.
[3] 産業技術総合研究所, 機械学習品質マネジメントガイドライン第1版, https://www.cpsec.aist.go.jp/achievements/aiqm/ (閲覧2020-12-27).
[4] 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2020.08版, http://www.qa4ai.jp/download/ (閲覧2020-12-27).
[5] 経済産業省商務情報政策局, 割賦販売法, https://www.meti.go.jp/policy/economy /consumer/credit/11kappuhanbaihou.html (閲覧2020-12-20).
[6] 日本銀行 金融機構局, AI を活用した金融の高度化に関するワークショップ 第3回, https://www.boj.or.jp/announcements/release_2019/rel190215d.htm/ (閲覧2020-12-20).
[7] 小野潔 (2016)，インテックの与信モデルの特徴と今後の展開, ITJ2016.9 第17号.
[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind *et al*. (2019), AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, IBM Journal of Research and Development, Vol.63, Issue: 4/5, July-Sept. 2019.
[9] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík *et al*. (2018), A Reductions Approach to Fair Classification, In Proceedings of the 35th International. Conference on Machine Learning
[10] 佐藤慎一，石川冬樹，猪原健弘 (2011)，貢献度と顧客のニーズに関する妥当性の間のコンフリクト検出指標, ソフトウェアエンジニアリングシンポジウム2011, pp.1-6.