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

相津 一寛
パナソニック株式会社
aizu.kazuhiro@jp.panasonic.com

石川 冬樹
国立情報学研究所
f-ishikawa@nii.ac.jp

小宮山 英明
コニカミノルタ株式会社
hideaki.komiyama@konicaminolta.com

栗田 太郎
ソニー株式会社
taro.kurita@sony.com

柳原 靖司
ブラザー工業株式会社
yasushi.yanagihara@brother.co.jp

徳本 晋
富士通株式会社
tokumoto.susumu@fujitsu.com

要旨

本論文では，AI システムの品質を保証する手段として，IGDM-AIQA 法(Individual Guideline Derivation Method in AI system Quality Assessment)を提案する．現在，AI システムは多くの分野で開発，運用されているにも関わらず，その特性から品質保証の方法が確立されていない．このような問題に対し，AI 開発の知見を集約してガイドライン化することが議論されているが，これらガイドラインは有識者向けで抽象度の高い内容となっており，AI の知見が必ずしも十分でない品質保証担当者が効果的に活用できるとはいえない．IGDM-AIQA 法を用いることで，対象システムの要件に基づいて品質アセスメントに必要な観点を導出し，品質保証の現場担当者が精度良く品質アセスメントを行える．

1. はじめに

機械学習技術の著しい発展により様々な産業分野で AI システムが開発，利用され始めているが，AI システムが内包する特有の性質から従来型の品質保証アプローチが通用しない課題が議論されている．このような課題に対して，AI システムに特化した品質保証のあり方が研究され，様々な AI システムの品質保証ガイドラインが発行されている．しかし，機械学習技術の応用分野は幅広く，既存の品質保証ガイドラインでは AI システムに共通する上位水準の知見，又は個別産業で共通する知見を提示しているのが実際である．これらガイドラインの記述は抽象度が高く，機械学習技術に精通していない品質保証担当者では記述の解釈が難解である．

本研究では，ゴール指向要求分析手法のひとつである AGORA[1](Attributed Goal-Oriented Requirements Analysis method)と，システム安全分析手法のひとつである FRAM[2] (Functional Resonance Analysis Method)を応用して，AI システムの要求から品質保証の点で確認すべき項目（サブガイドライン）を抽出する IGDM-AIQA 法を考案し，実用性を評価した．仮想のクレジットカード与信審査システム（以降，FinTech 与信判定システムと略す）を例とした第三者による実験では，機械学習技術に詳しくない技術者であってもサブガイドラインを活用して有効な指摘ができることが分かり，要求工学やシステム安全性向上の知見に基づいてサブガイドラインを導出する意義を確認した．

以下，本論文の構成を述べる．まず，2 章では現状分析と課題提起を行う．次に 3 章では関連技術について言及し，4 章，5 章で夫々，解決策の提案と評価結果を示す．6 章で評価結果に対する考察を行い，最後に 7 章で成果と将来への発展で結ぶ．

2. 解決すべき課題

2.1. 現状分析

機械学習技術を応用したシステム開発では，外界から収集された学習データに基づきモデルを生成しながら開発を進めて行く帰納的手法を採用しているため，従来の演繹的手法に基づくシステム開発及び品質保証アプローチを適用できない．このような状況に対応するため，開発・運用フェーズにおいて AI システムの品質作り込みや品質評価の指針を与えるものとして，汎用的な AI システム品質のためのガイドライン（以降，汎用ガイドラインと略す）が発行されている．

2.2. 課題提起

既存の汎用ガイドラインは，機械学習技術に対する知識が一定以上ある開発者を想定しており，多岐にわたる AI システムの応用分野における共通事項をまとめているため内容が抽象的で，機械学習技術の知見が十分でない品質保証担当者が活用することが難しい．例えば，機械学習品質マネジメントガイドライン[3]では，「このガイドラインは，できるだけ広範な応用に適用できるような汎用的な記述内容になっており，利用者が具体的な応用に即して，記述内容を取捨選択・具体化して用いることを想定している．実際の利用場面においては，各利用者の状況に応じて，応用分野毎にこのガイドラインの内容を具体化した個別ガイドラインを作ることを想定している．」という記述が本文に記載されている．また，池村・佐藤・松本(2021)は AI 品質マネジメントガイドラインの具体化に関する研究[4]の中で，AI 技術の実務経験有無の違いによりガイドライン解釈の精度に相違がでることを示唆しており， AI 技術初学者がガイドラインを読み解く上で有識者の補助が必要であると述べている．

実際の企業の開発現場に AI 品質の汎用ガイドラインを広く普及させる上では，ガイドラインの具体化と展開のための仕組み作りが必要であると考える．

3. 関連技術の説明

我々の研究の技術的拠り所として用いている，AI システム品質に関する汎用ガイドライン，AGORA 及び FRAM について説明する．関連技術として AGORA と FRAM に着眼した理由であるが，一般に AI システムでは要件の間でトレードオフが発生することが多いため，システムの要求から要件を獲得する過程を俯瞰的に可視化しながら分析できる点で AGORA，AI システムの運用時にプロジェクト関係者だけでなくエンドユーザや社会（コミュニティ）といった多様なステークホルダが関連するため，複雑な機能連関構造を可視化しながら分析できる点で FRAM が夫々適していると考えた．

3.1. AI システム品質に関する汎用ガイドライン

機械学習技術を利用したシステムやプロダクトの品質保証に対する共通指針を与えるものとして，機械学習品質マネジメントガイドライン，AI プロダクト品質保証ガイドライン[5]が発行されている．企業・大学等の有識者が知見に基づいて取りまとめたものであり，製品やサービスの開発者，利用者らが参照することを想定して記載されている．

3.2. AGORA

ゴール指向要求分析手法のひとつで，AND-OR ツリーグラフに属性値を付与した上で，主要求から下流に向かって副要求を展開しながら要件を導出する．リーフの各要件について，ステークホルダの満足度行列を調べることで主要求に対するゴール適合度や意見の対立度を計算することができる点が特徴である．図 1 に AGORA のモデル例を示す．

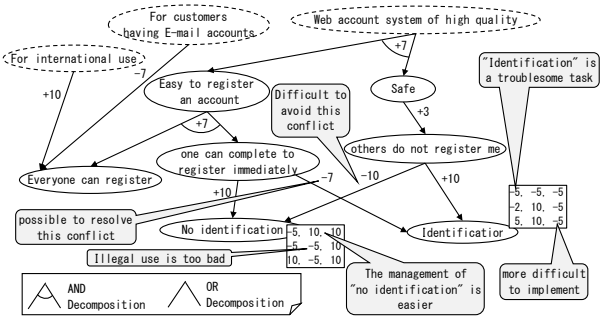

図 1 AGORA による要求分析例

なお，ゴール指向要求工学（GORE: Goal-Oriented RE）の中には先行研究として KAOS[6] (Knowledge Acquisition in Automated Specification), i*[7], NFR Framework[8] (Non-Functional Requirements Framework) 等が存在する．何れもツリー図やネットワーク図を利用しながらゴールを実現するためのサブ要求や要件を分解しながら展開することで，因果関係・AND/OR 関係等を論理的に表現する点が共通している．

3.3. FRAM

システム安全性向上におけるモデリング手法のひとつである. 複数の機能がインタラクションする構造をモデルベースで分析することで, システムの長所や短所を特定できる. システムが正常に働くパターンを増強することを目的として利用する. 図 2 に FRAM のモデル例を示す.

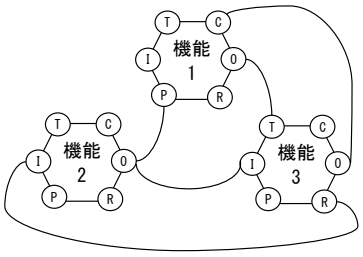

図 2 FRAM による分析例

なお, システム安全分析手法の中には先行研究として STAMP/STPA[9] (Systems-Theoretic Accident Model and Processes/System-Theoretic Process Analysis), MBSE[10] (Model-Based Systems Engineering)等が存在する. 前者はシステム間の連携やシステムと人の関わりにおいて構成要素の相互作用からハザードが発現するプロセスをモデル分析することにより, 創発特性から生じるシステム事故の防止に役立てる手法, 後者は高い抽象度でシステムをモデリングし, 俯瞰的に検証することで全体最適化の実現を支援する手法である. MBSE のモデリング手段として SysML(Systems Modeling Language)を使うことができるが, SysML で記述したモデルを STAMP/STPA で解析した研究報告[11]も存在する.

4. 解決策の提案

4.1. 課題の解決方針

汎用ガイドライン[3]では AI システムの品質観点を, ユーザに対する「利用時の品質」, システムの構成全体に求められる「外部品質」, 構成要素の固有特性として定義される「内部品質」として捉えている. 本研究では, 当該ガイドラインにおける 3 つの品質観点を考慮しながら, 個別 AI システムに対する品質保証上の着眼点をビジネスゴールに適合するように導出する枠組み(IGDM-AIQA 法)を提案する(図 3). また, 現場の品質保証業務に対する IGDM-AIQA 法の有効性を評価するため, FinTech 与信判定システム(図 4)に適用した結果を示す.

[IGDM-AIQA 法の特徴]

・ ゴール指向要求分析手法(AGORA)により利用時の品質と外部品質を考慮しながら AI システムを分析することで, 該当システムに求められる要件群を目的指向で導出する.

・ AI システムの有効性と公平性を定量的に解析する上で, 与えられたデータセットに対する機械学習モデルのシミュレーションを行いながらデータ分析することで, 当該システムの内部品質に求められる特性を把握する.

・ AI システムの主要求に関わるステークホルダの機能連関構造(FinTech 与信判定システムでは公平性に配慮した社会受容性の構造)を明らかにする上で, 各機能のインタラクションを FRAM により分析し(図 5), ステークホルダの重要度を特定する.

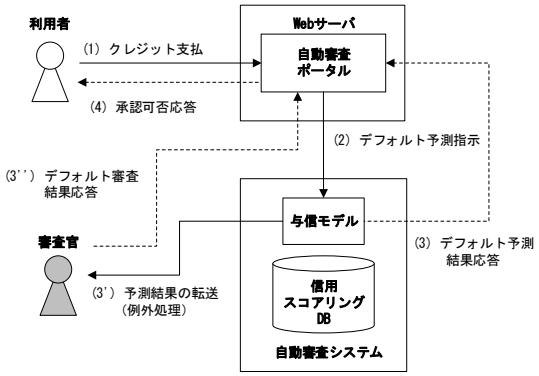

図 4 FinTech 与信判定システム

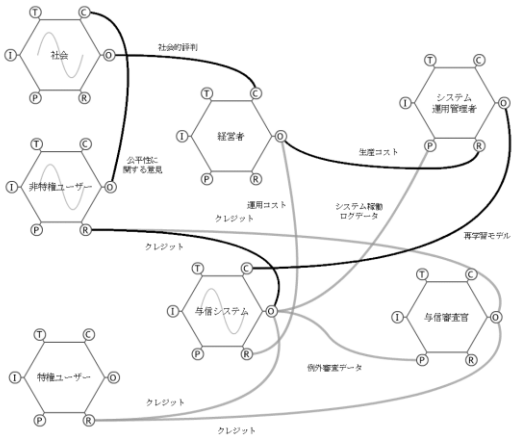

図 5 ステークホルダの機能連関分析 (FRAM 適用例)

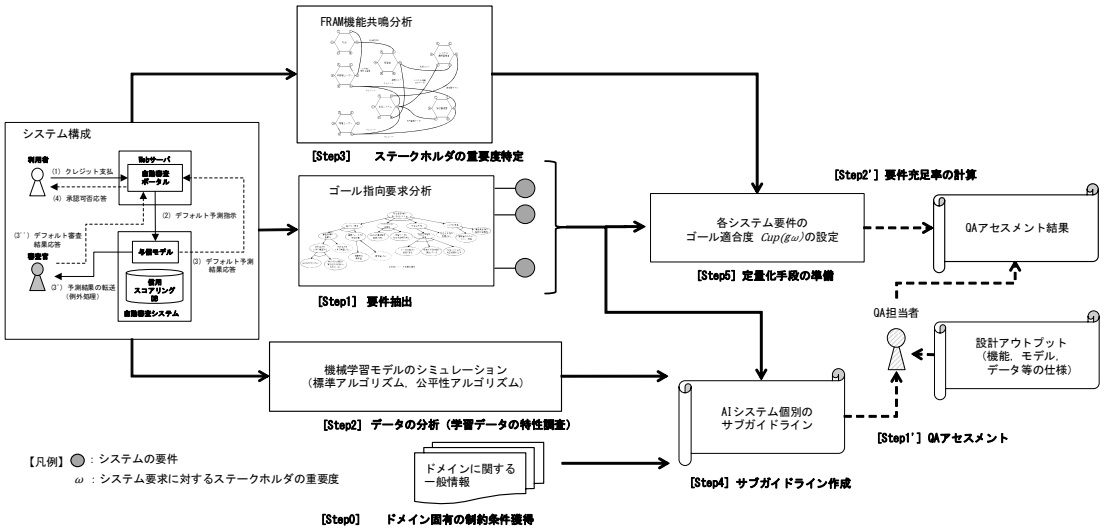

図 3 IGDM-AIQA 法（AI システム品質のサブガイドライン導出の枠組み）

4.2. IGDM-AIQA 法を用いたサブガイドラインの導出

IGDM-AIQA 法は，AGORA，FRAM といったシステム開発で利用される既知の分析手法を活用しながら個別 AI システムに合わせて汎用ガイドラインを解釈するための枠組みである．品質保証業務に携わる QA 担当者は，表 1 に示す手続きに則って進めることで実用的なサブガイドラインを導出することができる．

[ゴール適合度の算出手順]

佐藤・石川・猪原(2011)は，ゴール指向要求分析手法（AGORA）によって導出した各要件に対して，対象システムのゴール適合度（貢献度）を定量化する計算手法を示している[17]．我々の研究では，この定量化方法を参考に，FinTech 与信判定システムに登場するステークホルダ（図 6 凡例参照）のうち主要なステークホルダ（Primary Stakeholder）である，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）のゴール適合度を式(1)に基づき計算することにした．

表 1 IGDM-AIQA 法を用いたサブガイドラインの導出手順

手順	目的	□入力情報，■出力情報	手続き
STEP0	AI システムの開発に付帯する制約の明確化	□ システム仕様・設計情報，ドメインの関連資料[1] ■ ドメイン固有の制約情報	ドメイン固有の情報から，目的システムの開発や運用に関わる制約情報（機械学習，運用，規制等の知見）を獲得する．
STEP1	AI システムの品質観点を定義する際に参照する要件群の導出	□ STEP0 の出力，汎用ガイドライン ■ 目的システムの要件群（品質観点の定義に適した粒度）	ゴール指向要求分析手法 AGORA を用いて，目的システムの主要求（ゴール）をツリー状にサブゴールへと展開しながら要件群を導出する．品質観点として，汎用ガイドラインに掲載されている内容（リスク回避性，AI パフォーマンス，公平性[2]等）を考慮する．
STEP2	機械学習コンポーネントの学習データに対する推論特性の調査	□ 目的システムが前提とする機械学習のデータセット ■ 機械学習アルゴリズムの推論特性，データセットの被覆性・均一性に関する情報	目的システムに搭載されている機械学習コンポーネントの学習データに対する推論特性をシミュレータ等で解きながら把握する．（機械学習の汎用アルゴリズムに対して Colaboratory(Google)，公平性アルゴリズムに対して AI Fairness 360(IBM)[15]，Fairlearn(Microsoft)[16]等を利用する）

[1] FinTech 与信判定システムの場合は，改正割賦販売法[12]，日本銀行のワークショップ報告書[13]，研究事例[14]等を参照した．

[2] 対象システムよっては公平性に対する要求が小さいため，適宜判断して除外する．（例：株価の予測，交通標識の識別）

STEP3	AI システムのゴールに強い影響を及ぼすステークホルダの特定	□ STEP0, 1, 2 の出力 ■ Primary Stakeholder のリストとゴール(主要求)の貢献度に応じた重みω	システムの運用時に関わるステークホルダのうち，目的システムのゴール適合度に影響する Primary Stakeholder を FRAM 分析[3]により特定する. ステークホルダの機能連関構造を参考にしながら，システムの主要求に対する貢献度に応じて各 Primary Stakeholder の重みωを決める.
STEP4	QA アセスメントで使用するサブガイドラインの導出	□ STEP1, 2, 3 の出力 ■ サブガイドライン	システムの運用時に目的システムの機能が正しく発現されるための品質観点を，要件毎に汎用ガイドライン及び学習データの推論特性に基づいて検討し，ステークホルダの重要度を加味した上でサブガイドラインを導出する.
STEP5	AI システムのゴール適合度$Cup(g_\omega)$の計算	□ STEP1, 3 の出力 ■ ゴール適合度	サブガイドラインを用いた QA アセスメントタスクの達成度を計測するため，図 6 に示す手順に基づき，目的システムのゴール適合度$Cup(g_\omega)$(式 1)を要件毎に求める.

$$Cup(g_\omega) \stackrel{\text{def}}{=} \frac{\sum_{s\in Stakeholder, p\in Primary\ stakeholder} \omega \cdot m(g)_{s,p}}{|\ Stakeholder\ | \cdot |\ Primary\ stakeholder\ |}. \quad (1)\ [17]$$

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

STEP5-1
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(要件に対する重要度)を入力する.

STEP5-2
$Cup(g_\omega)$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
(ω の値は STEP3 参照)
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]

[補足] 本研究では，研究員3名が夫々作成した満足度行列から$Cup(g_\omega)$を求め，これらを平均した.

【凡例】PU: Privileged User(特権ユーザ*1)，UU: Unprivileged User(非特権ユーザ*1)，OW: Owner(経営者)，OM: Operations Manager(システム運用管理者)，DV: Developer(開発者)
(*1) FinTech 与信判定システムでは PU が男性，UU が女性である.(特定の年代で，女性が与信額で不利な扱いを受ける場合がある)

図 6 $Cup(g_\omega)$導出のための満足度行列

4.3. 仮説と研究設問

本研究の仮説と，研究設問 RQ を示す.

[仮説]

IGDM-AIQA 法から導出された AI システム品質評価のためのサブガイドラインを用いれば，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，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.

[3] FinTech 与信判定システムの例では，図 6 の凡例に示すステークホルダのうち，図 5 の機能ループ(太線)を増強する重要なステークホルダとして，非特権ユーザ $_{UU}$，経営者 $_{OW}$，システム運用管理者 $_{OM}$を選定した. なお，AGORA の$Cup(g_\omega)$の計算値に影響を与えるのは UU, OW, OM であるが，各ステークホルダの素点を同列で扱うのでなく，FRAM 機能連関構造内の「社会(六角形のオブジェクト)」に対する影響の仕方(強弱)を加味することにした. 社会に直接作用する非特権ユーザ UU，社会からの出力を受けて指令を出す経営者 OW，間接的に機能ループの強化に作用するシステム運用管理者 OM という解釈を行い，3 名の研究員で合意して ω の重みを決定した. 要件のゴール適合度を$Cup(g_\omega)$と再定義することで，ステークホルダの機能連関構造から生じる影響を要件の重みに反映させた(要件に対する重みは$\omega_{UU} = 1.0$, $\omega_{OW} = 0.9$, $\omega_{OM} = 0.8$を付与した).

RQ2[4]: 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．

5. 解決策の評価

5.1. 評価方法

仮説の検証は，FinTech 与信判定システムの概略設計資料（設計アウトプット）に含まれる欠陥事項をアセスメントするような枠組み（図 7）を用いた．独立した 2 つの被験者グループを用意することで RQ の定量的評価と，属性毎の定性的評価を行えるようにした． なお，5.2 節において前者は$\sum_i A_i \cdot Cup(g_{i\omega})$，後者は$\overline{A_i}$及び$\sum_i A_i$に着目して被験者のアセスメント精度を分析した．（$A_i$は要件 i に対する *Assessment accuracy* を意味する）

■ 実験意図の説明
被験者に対し，システム要求（主要求と副要求）及びシステム構成図を提示し，（本実験の）品質アセスメントで問われている期待結果を理解させる．

⇒

■ 検証対象の提示
設計チームが作成した不完全なシステム基本設計情報（表 2）を提示する．当該情報には，モデル，データ，その他の各区分で，設計意図が詳細に記述されている．

⇒

■ 実験の試行
被験者グループ*2 I群とII群は，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の各条件で，基本設計情報に含まれる欠陥を指摘する．
※30 分間で最大 10 個

⇒

■ 実験結果の評価
被験者から提出してもらった指摘事項を，システムの要件（表 3）と照合しながら3段階で採点する*3.
[A_i: *Assessment Accuracy*]
1 点: 指摘と理由が適合
0.5 点: 指摘はできているが理由が不明瞭
0 点: 指摘は要件の範疇外

(*2) 被験者グループ I群：機械学習技術に詳しくない技術者，II群：機械学習技術に詳しい技術者
(*3) AGORA を使って導出された要件群に対する適合性で採点する．

図 7 IGDM-AIQA 法の効果測定の枠組み

表 2 システム基本設計情報（一部抜粋）

区分	項目	説明
モデル	アルゴリズム	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．モデルのハイパーパラメータとして，BATCH_SIZE：25, EPOCH：20とした．

表 3 要件と対応するサブガイドライン

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*4 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．（※国内の与信判定システムでは，一般的に決定木，XGBoost，LightGBM 等が好まれる）
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． －年齢データの比率 －最終学歴データの比率 －学習データ内のデフォルトの比率
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的に含まれているか．
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，SMOTE 等の標準アルゴリズムにより近接データの内挿を行って，データを増やしているか．

(*4) ML は Machine Learning（機械学習）の略

[4] 本研究の構想段階において，IGDM-AIQA 法から導出したサブガイドラインは，機械学習技術に詳しくないが演繹的手法に基づくシステム開発には精通している技術者に対してのみ有効であると仮定した．一方，機械学習技術に詳しい技術者に対してはサブガイドラインの有効性が低く，既存の汎用ガイドラインを参照することで品質の作り込み活動を行えるものと考えた．

6	ML の誤りが少ない	省人化に寄与	4.0	・機械学習コンポーネントの推論結果における正答率, 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	・学習データでデフォルトになっているデータは特定の性別に偏っていないか. ・システムが出力する与信の判定結果が, 特定の属性に対して不公平な結果になっていないか. ・少数の非特権ユーザに対して不利な判定結果になるような, 機械学習アルゴリズムのバイアスに対する補正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	・機械学習コンポーネントの再学習に要する時間は, 運用フェーズにおいて許容できる時間内であるか.
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. (※年代毎の与信のデータ分布)
10	事故発生時の解析の容易性	リスクの低減	3.0	・機械学習モデルの作成時に用いた学習データ・検証データ・モデル学習履歴等を適宜参照できるように記録されているか.

5.2. 評価結果

異なる業種(製造, 情報・通信, 金融)に属する計13社の被験者(機械学習技術に詳しくない技術者 I群: 25 名, 機械学習技術に詳しい技術者 II群: 13 名)に対して, 仮想 FinTech 与信判定システムのシステム概略設計資料をアセスメントしてもらい, 当該資料に含まれる欠陥事項を日本語で指摘してもらった. (回答集計率 95%)

[RQ に関する評価結果]

システムの要求を期待結果とした前記システム概略設計資料のゴール適合度(欠陥指摘精度)を検証するため, 各条件(条件 a: ガイドラインなし, 条件 b: 汎用ガイドライン参照, 条件 c: サブガイドライン参照)における被験者の回答結果(自然言語)を$\sum_i A_i \cdot Cup(g_{i\omega})$で定量化した(表 4, 5). 4.3 節に示す RQ を検証する上で, 被験者回答データが正規分布に従っていると仮定した上で, 条件 a と条件 c, 及び条件 b と条件 c の各母集団に有意な差があるか否かを t 検定(有意水準 5%)によって検証した. t 検定の準備として, F 検定によりサンプルの等分散性を確認したところ, I群については等分散性がなく, II群については等分散性があったので, 夫々Welch の t 検定, Student の t 検定を用いた. I群に関する t 検定の統計量は条件 a と条件 c が 3.01×10^{-8}(＜0.05), 条件 b と条件 c が 6.76×10^{-6}(＜0.05)であり, 両者とも有意な差があることが分かったので RQ1 の妥当性が確認できた. II群に関する t 検定の統計量は, 条件 a と条件 c が 8.27×10^{-5}(＜0.05), 条件 b と条件 c が 1.33×10^{-3}(＜0.05)であり, 両者とも有意な差があることが分かったので RQ2 の妥当性が確認できなかった. 即ち, 保有する機械学習技術の知見に依らずサブガイドラインがあれば, 個別 AI システムの設計アウトプットの欠陥指摘精度が向上するので, 条件付きで仮説を検証できたといえる.

表 4 被験者回答精度(I群 $\sum_i A_i \cdot Cup(g_{i\omega})$)

I群	条件 a	条件 b	条件 c
平均	9.0	13.2	26.1
分散	17.8	37.9	102.9
要件充足率	18.9%	27.6%	54.7%
サンプル数	25	25	23

(補足)一部の被験者から回答データを回収できなかった

表 5 被験者回答精度(II群 $\sum_i A_i \cdot Cup(g_{i\omega})$)

II群	条件 a	条件 b	条件 c
平均	16.8	19.7	30.6
分散	32.3	26.4	73.9
要件充足率	35.2%	41.3%	64.2%
サンプル数	13	12	12

(補足)一部の被験者から回答データを回収できなかった

[定性的な評価結果]

設計資料に含まれる欠陥事項に対する被験者の回答傾向を把握する上で, 回答結果(自然言語)の要件適合性($\overline{A_i}$ 及び$\sum_i A_i$)で可視化した.

<u>要件毎の被験者回答精度</u>

被験者の回答精度をシステムのゴール適合度(主要求への適合性)という全体視点で見ると, FinTech 与信システムの基本設計情報に含まれる欠陥事項の回答指摘精度はサブガイドラインを参照した場合に有意に改善することが分かったが, 要件毎に分解して見た場合には, 各要件の改善幅に差があった(図 8, 9). 表 6 に定性的に分析した結果を示す.

品質保証におけるサブガイドラインの効果

所属企業での役割が QA 担当者(N=6)である被験者の回答結果を抽出して可視化したものを示す(図 10). 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(条件 c), 参照しない場合(条件 a, b)に比べ, 夫々, 4.2 倍(1.8pt→7.5pt), 2.5 倍(3.0pt→7.5pt)の改善効果が認められた. この結果は, 機械学習技術の有識者(II群)の対応する改善効果(1.6～1.7 倍)と比べても顕著である.

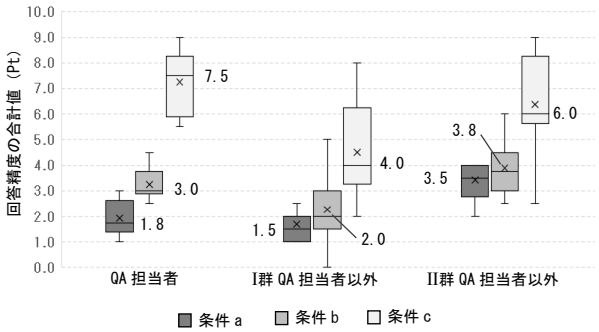

図 10 役割に着目した被験者回答精度$\sum_i A_i$

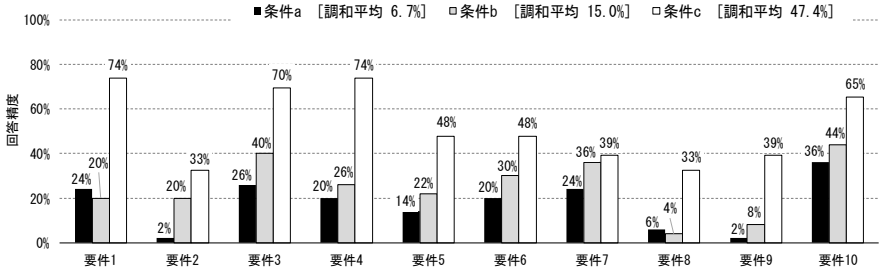

図 8 要件毎の被験者回答精度(I群 $\overline{A_i}$)

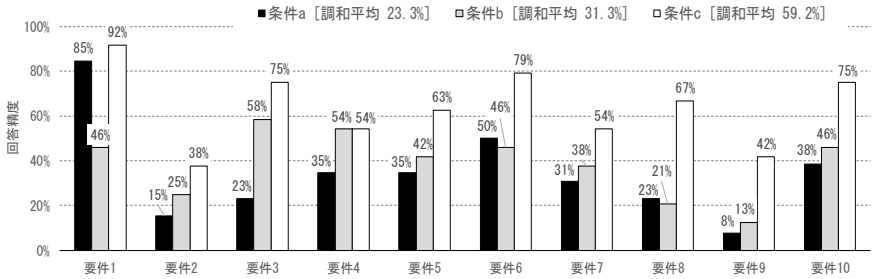

図 9 要件毎の被験者回答精度(II群 $\overline{A_i}$)

表 6 要件別被験者回答精度の定性的評価

I群 $\overline{A_i}$	推定要因
改善幅が大きい要件:要件 1, 4 (条件 a と c の差 50pt, 54pt)	サブガイドラインの対応箇所では機械学習アルゴリズムの種類, 学習データの網羅性について言及しているが, 機械学習技術に対する豊富な知識がなくても概念として理解しやすい.
改善幅が小さい要件:要件 7 (条件 a と c の差 15pt)	サブガイドラインの該当箇所では機械学習の公平性について言及しているが, 推論結果の偏りやそのバイアスを補正する考え方に対する理解が難しい. (後者は, 研究領域として発展中)

II群 $\overline{A_i}$	推定要因
改善幅が大きい要件:要件 3, 8 (条件 a と c の差 52pt, 44pt)	機械学習技術に詳しい技術者は, 機械学習アルゴリズムの種類と性能(要件 1, 6)を深く掘り下げて欠陥事項を指摘する傾向があり, 学習データの偏りと機械学習の再学習時間(要件 3, 8)については関心が低かった. サブガイドラインの中で着目すべき視点を明示することで該当箇所にも意識を向けられる.
改善幅が小さい要件:要件 1 (条件 a と c の差 7pt)	サブガイドラインの対応箇所では機械学習アルゴリズムの説明性について言及しているが, 既に被験者が保有している機械学習技術の知見でも, 基本設計情報に含まれる欠陥を容易に指摘できる.

共通	推定要因
条件 b(汎用ガイドライン参照)の精度が条件 a, c に比べて低い要件:要件 1, 8	要件 1:「説明性が高い機械学習アルゴリズム」という点について, 個別システムに応じた解釈方法が汎用ガイドラインでは解説されていないので精度が下がった. 要件 8:「機械学習の再学習」は当該分野の初歩の知識領域であるため, 敢えてガイドラインには掲載されておらず, それ故, 視点として着眼されにくかった.

6. 考察

6.1. 得られた知見

[仮説に対する整合性]

(関連: 5.2 節 [RQ に関する評価結果])

AI システムの品質保証活動において, 基本設計情報の欠陥を要件群に照らして指摘するタスクについては, 機械学習技術に関する知識や業務経験に依らず, 要件群から導出した QA アセスメントのためのサブガイドラインに基づいて実施した方が精度を改善できる. 記述の抽象度が高い, 既存の汎用ガイドライン[3]を参照するよりも有意に改善する.

[サブガイドラインの記述]

(関連: 5.2 節 要件毎の被験者回答精度)

個別 AI システムの要件群に適合するようなかたちでサブガイドラインを導出する際, 機械学習技術に関する背景知識に依存して理解度にばらつきが生じないようにするため, IGDM-AIQA 法のサブガイドライン導出部において言語化の工夫が必要である.

[品質保証活動の現場での有効性]

(関連: 5.2 節 品質保証におけるサブガイドラインの効果)

現場の実務でシステムやソフトウェアの品質保証活動を行っている担当者が, IGDM-AIQA 法から導出されたサブガイドラインを参照すると欠陥指摘の精度を 4 倍近く改善できるので, QA 担当者の本業の知見を補完しながら実務を遂行する上で合理性が高い.

6.2. IGDM-AIQA 法の実用性

[サブガイドライン導出の再現性]

IGDM-AIQA 法では機械学習, 要求工学(AGORA), 及びシステム安全性向上(FRAM)の学際的知見を活用している. 品質保証部門の現場で必要となる QA アセスメントタスクでは, 各分野の緻密な理論を習得するよりは寧ろ欠陥検出やレビューに必要な視点を保有する方が重要であるため, 初学者レベルの知識を保有すればよい. 本研究のケーススタディに関わった 3 名の研究者は, 計半年間の各分野の自習によりサブガイドラインの導出を行えた. なお, 複数人の視点で抜け漏れを検証すれば, 網羅性や十分性を担保できると考える.

また, 5.2 節「要件毎の被験者回答精度」で示したサブガイドライン毎の QA アセスメント精度のばらつきを考慮すると, サブガイドラインの項目によっては十分な精度を得られていないものがあるため, 振り返りを行って得られた知見を初学者向けにパターン化して反映する等, サブライドラインの記述内容へのフィードバックを繰り返すことで精度向上が見込まれる.

[投資対効果]

品質保証の現場に IGDM-AIQA 法を展開する上で, 従来の開発手法に慣れた QA 担当者に対して, 前記 3 つの分野の初学者レベルの知識獲得を目的とした教育投資を行えばよい. これにより, ガイドラインを使わない場合(図 10 条件 a)の約 4 倍の QA アセスメント精度を見込める. 他方, 汎用ガイドラインを活用するためには, 機械学習有識者の指導の下, プロジェクトでの活用を経験し, ガイドラインの咀嚼方法を学ぶ必要がある. 経営的視点で機械学習有識者の資源を教育活動に振り向けることは, 投資対効果の点で課題がある.

6.3. 妥当性への脅威

・ 被験者の在籍している企業の多くは製造業や情報・通信に属しており, FinTech に特有の機械学習技術の知見は必ずしも持ち合わせていない. 5.2 節の実験(条件 a, b)の試行において, ドメイン固有の知識を十分に持っていなかったことが不利に働き, 条件 c の改善効果を見かけ上増長させた可能性がある.

・ 今回の実験では, 夫々の被験者が条件 a, b, c の下で QA アセスメントを行うような内容とした. 各条件での試行は 1～2 週間間隔で取り組んでもらったが, 過去の実験で得られた知識が次の実験に影響するような経験効果が働いた可能性を否定できない. そこで, 被験者が回答したデータを精査し, 各回での重複回答の比率を分析したところ, 下記に示すとおり調和平均で 14.8～20.9%の範囲に留まっていた. また, 条件 a→b→c という連続的な実験の試行による経験累積効果は確認されなかった. これらのことから, 今回の実験では, 過去の実験で得られた経験知よりも, 各実験回で被験者へ提示したインプット情報の方が実験結果に与える影響が大きかったと考える.

＜各条件間の被験者回答データの重複率＞

I群: [a, b] 19.3%, [a, c] 14.8%, [b, c] 20.9%

II群: [a, b] 17.9%, [a, c] 20.6%, [b, c] 20.5%

・ 本研究では, FinTech(クレジットカードの与信審査)という, 特定領域に対して IGDM-AIQA 法を適用したので, 手法の汎用性を示す上ではさらなる実験が必要である.

7. まとめ

7.1. 成果

本研究では，品質保証の現場で活用しやすい 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た結果，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と比較してアセスメントの精度が向上することを確認した．特に，現場の QA 担当者がサブガイドラインを活用した場合に 4 倍程度の精度改善を見込めることが分かった．

7.2. 将来への発展

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，IGDM-AIQA 法によるサブガイドラインの導出方法は，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．

8. 謝辞

本研究は，日本科学技術連盟 2020 年度ソフトウェア品質管理研究会（SQiP）研究コース 5「人工知能とソフトウェア品質」の中で取り組んだ内容です．研究活動の場をご提供頂きました SQiP 研究会，並びに本研究の実験にご協力頂きました研究コース 5，一般企業の有志の方々に厚く御礼申し上げます．

付録

本研究のケーススタディとして用いた FinTech 与信判定システムに対する IGDM-AIQA 法の適用結果の詳細は，本論文の著者らが SQiP 研究会で執筆した成果報告書（付録）[18]に記載がある．

参考文献

[1] H. Kaiya, H. Horai, M. Saeki (2002), AGORA: attributed goal-oriented requirements analysis method, 10th Anniversary IEEE Joint International Requirements Engineering Conference (RE'02), 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] 池村，佐藤，松本 (2021), AI 品質マネジメントガイドライン具体化における AI 経験有無の影響明確化, 第 36 年度ソフトウェア品質管理研究会 分科会報告書 pp.107-131.

[5] 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).

[6] A. van Lamsweerde (2001), Goal-Oriented Requirements Engineering: A Guided Tour, Symposium on Requirements Engineering, 2001, pp. 249-263.

[7] E. Yu and J. Mylopoulos (1994), Understanding “Why” in Software Process Modelling, Analysis, and Design, Proc. 16th Int. Conf. Software Engineering.

[8] L. Chung, B.A. Nixon, E. Yu, J. Mylopoulos (1999), Non-Functional Requirements in Software Engineering, Kluwer Academic Publishers.

[9] Nancy G. Leveson (2011), Engineering a Safer World – Systems Thinking Applied to Safety, The MIT Press.

[10] INCOSE (2006), Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities, INCOSE-TP-2003-002-03.

[11] 三原，十山，石井，松田，三縄，八嶋 (2016), システムの安全性・信頼性分析手法，SEC journal Vol.12 No.1 Jul. 2016.

[12] 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).

[13] 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).

[14] 小野 (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.

[15] Rachel K. E. Bellamy *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, Numbers 4/5, July-Sept. 2019.

[16] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, Hanna Wallach (2018), A Reductions Approach to Fair Classification, In Proceedings of the 35th International. Conference on Machine Learning

[17] 佐藤，石川，猪原 (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.

[18] 相津，小宮山，柳原(2021)，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，第 36 年度ソフトウェア品質管理研究会 分科会報告書 pp.132-155.