AI アルゴリズムの透明性について
—AI によるイノベーションを社会が受容し享受するためのガバナンスのあり方の一提案—

結城 東輝[*†]、岡本 昌之[‡]、村山 拓、矢島 桐人[§]、西垣 裕太[**††]

要旨

AI によるイノベーションが生み出す正のインパクトを最大化するためには、当該 AI が内包するリスクをステークホルダーが認識し、その利害調整を適切かつ柔軟に実行できる技術的、組織的、社会的システムの設計と運用、すなわち当該 AI に対するアジャイル・ガバナンスのフレームワークが不可欠である。本稿は、AI アルゴリズムの透明性という最も議論が盛んなガバナンス上の課題について、俯瞰的、体系的な整理に挑戦し、実践的なツールキットを提案する。俯瞰的、体系的な整理がない中で、イノベーションが種々の問題を引き起こす度に規制や規範が形成されることは、設計図のないまま建物の増築が繰り返されるようなものであり、事業者にとって予見可能性を欠くとともに、本来の目的を離れた形式だけの「透明性」が自己目的化していきかねない。統一的な考え方のもと体系的に整理されていることと共に、事業者が実運用しやすい汎用性、明瞭さ、裁量の余地がある柔軟性が追求されなければならない。

本ツールキットは、各国の規制当局の議論のみならず、様々な AI アルゴリズムが生み出す新たなリスク事象を幅広く考慮し、さらには AI アルゴリズムに関する先行研究を踏まえながら、体系的な開示事項・事例集として使えるよう構成した。また、事業者が透明性の程度を工夫することで、社会における信頼性のみならず AI アルゴリズムを組み込んだプロダクトへの満足度など異なる指標にも良い影響が与えられる観点を取り込んでいる。加えて、開示事項をリスト形式にし、それを AI アルゴリズムの提供者、利用者、リスクなどの観点から裁量的な選択性とするなど、自主・共同規制に対応する事業者にとって利便性の高いツールキットとなることを目指し、ユースケースとともに提案する。

[*] SmartNews 株式会社，法律事務所 ZeLo・外国法共同事業
[†] （代表連絡先）tonghwi.soh@gmail.com
[‡] トヨタ自動車株式会社
[§] アフラック生命保険株式会社，アフラックデジタルサービス株式会社
[**] 東京大学大学院法学政治学研究科法曹養成専攻
[††] プログラムアシスタント

目次

1　AI アルゴリズムの透明性が求められる理由

1.1　はじめに

本稿は、第三次人工知能（Artificial Intelligence、以下 AI）ブームによって AI が社会に当然のように浸透している現代において、明らかになりつつあるリスクや弊害を社会がコントロールし、その効用を最大限享受するために、透明性という観点からの解法を提案するものである。イノベーションは、既存のシステムや体制、価値体系、ビジネスモデルなどを何らかの意味で破壊することが多く、社会がそのリスクを過度に恐れた結果、イノベーションが阻害される不幸は後を絶たない。経済産業省が 2021 年に公表した報告書[1]によれば、「社会において生じるリスクをステークホルダーにとって受容可能な水準で管理しつつ、そこからもたらされる正のインパクトを最大化することを目的とする、ステークホルダーによる技術的、組織的、及び社会的システムの設計及び運用」としてイノベーションに対するアジャイル・ガバナンスの必要性が述べられている。AI によるイノベーションと、それが内包するリスクのバランスを取るためには、ステークホルダー間でリスクを受容可能な水準で適切に分配する必要がある。そのガバナンスの一例として「透明性」という観点が求められるのである。そしてそれは、社会や利用者との適切なコミュニケーションにより、社会全体で AI のリスクを受容可能な程度にコントロールし、信頼を構築していくことで、イノベーションによるインパクトを最大化することに繋がっていく。

しかし、すでに AI アルゴリズムの透明性というトピックは国際的に議論されているものの、提案されている内容は百家争鳴の状況である。透明性の担保は重要であるが、本稿で見ていくように、透明性の程度を誤れば開示行為そのものが利用者の信用や満足度を落とすおそれすらある。また、体系的な議論を経ずにイノベーションが引き起こす種々の問題に都度対応する規制が追加されていくようなことは避けなければならない。そうでなければ、ガバナンス・信頼の構築や自己決定権の保護のための透明性という手段が目的化し、その結果、設計図のない増築を繰り返した低強度・低品質の開示項目への対応が至上目的となってしまう。イノベーションのスピードに耐えうる規制のあり方として、自主規制・共同規制が一般化しつつある中、事業者にとって理解しやすく、現実的に自主的対応ができる透明性の実現方法が提案されなければならない。

本稿はこれまでの議論の体系的な整理を試み、実務に耐えうるツールキットバージョン 1.0 を提案することでこれらの問題に対応するものである。本ツールキットを通じ、各国の最新の規制の議論にも対応しながら、AI がもたらす様々なリ

---

[1] 経済産業省「GOVERNANCE INNOVATION Ver.2: アジャイル・ガバナンスのデザインと実装に向けて」95 頁_https://www.meti.go.jp/press/2021/07/20210730005/20210730005-1.pdf（2023 年 1 月 28 日閲覧）

スク事象や AI に対する先行研究を踏まえ、体系的かつ利便性の高い開示事項・事例集を提案する。本稿は、ツールキットを活用するためのガイド又はハンドブックの立ち位置となる。

なお、本稿を通じて、AI とは「データ・情報・知識の学習等により、利活用の過程を通じて自らの出力やプログラムを変化させる機能を有する」[2]ソフトウェアやシステムの総称を指す。本ツールキットは特に機械学習・深層学習を用いた AI アルゴリズムを主眼に記載しているが、ルールベースや知識ベースといった他の AI アルゴリズムにおいても論点は概ね共通する。また、本稿でいう「AI アルゴリズム」とは、AI が問題を処理する一プロセスであり、AI が「数学的問題（解が定まっており計算可能な問題）を解くための有限回数の手続き」[3]を指す。

1.2 AI によるイノベーションが孕むリスク・弊害

本稿が取り組むべき問題を確認するために、AI によるイノベーションが孕むリスク又は弊害の実例を見ていきたい。

あるソーシャル・メディア A は、利用者のエンゲージメント指標を最大化するようなアルゴリズムを発展させた結果、特に若年層の心の健康を阻害している可能性があると指摘されている。別のソーシャル・メディア B では、フィード内の投稿順位が、利用者や社会の与り知れぬ方法で、人為的に、場合によっては何らかの政治的意図を持って操作されている可能性も指摘されている。

求人求職情報をマッチングする人材サービス C では、特定の人種や性別の求職者が優遇されてしまうことが明らかになっている。画像認識サービス D では、黒人をゴリラと判定してしまう問題が発生した。

飲食店のレーティング情報提供サービス E では、レーティングのアルゴリズムが変更されることによって飲食店に多大な損害を及ぼす可能性がある。検索エンジン F のアルゴリズムの変更がメディア事業者や閲覧者に及ぼす影響も同様である。

ある画像生成 AI サービス G は、特定のアーティストの意思に反して、類似する作品を無数に生み出すことができ、別の動画生成 AI サービス H は、政治家のフェイク動画を生み出して選挙を妨害することに成功している。

1.3 AI に関する原則・ガイドライン等で最も言及される「透明性」

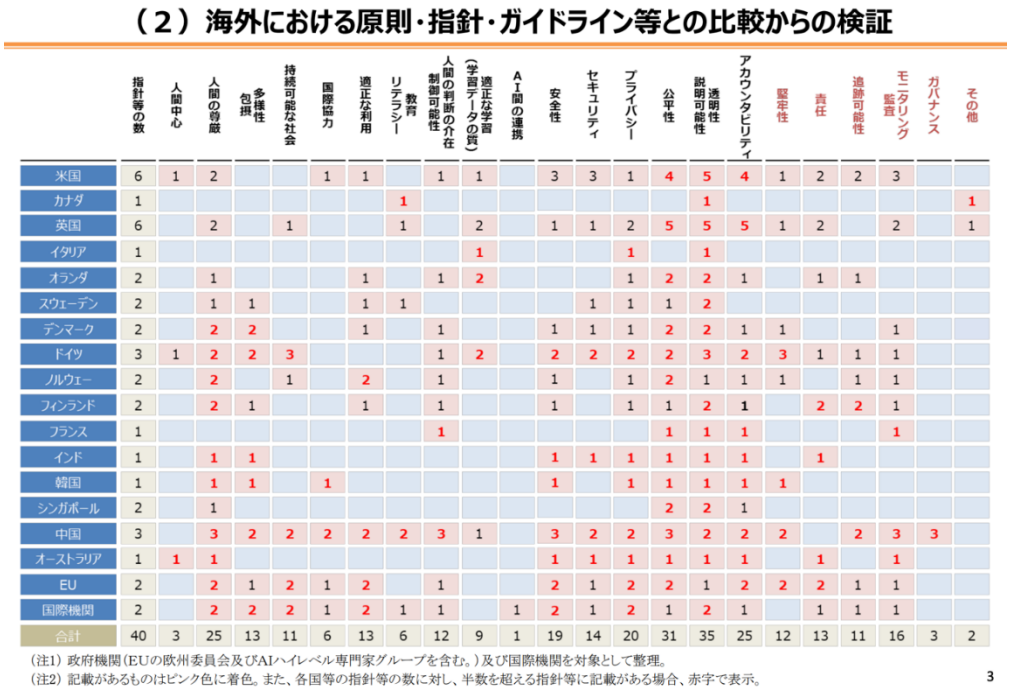
上述したような問題に対し、人間社会はどのように対峙すべきか、すでに議論

---

[2] 総務省 AI ネットワーク社会推進会議「AI 利活用ガイドライン ～AI 利活用のためのプラクティカルリファレンス～」4 頁（https://www.soumu.go.jp/main_content/000637097.pdf（2023 年 1 月 28 日閲覧））.

[3] “Algorithm.” Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/algorithm. Accessed 13 Jan. 2023.

は始まっている。総務省の報告書[4]によれば、各国政府や研究者らが公表している、40 本に及ぶ AI 関連の開発ガイドライン及び利活用ガイドラインのうち、約 9 割を占める 35 本のガイドラインで「透明性・説明可能性」が掲げられている。また、他の重要視される項目の「プライバシー」や「公平性」などを採用していない国・機関も存在する一方で、唯一「透明性・説明可能性」は全ての国・機関による指針に挙げられている。



	指針等の数	人間中心	人間の尊厳	多様性包摂	持続可能な社会	国際協力	適正な利用	教育リテラシー	人間の判断の介在制御可能性	適正な学習（学習データの質）	AI間の連携	安全性	セキュリティ	プライバシー	公平性	透明性説明可能性	アカウンタビリティ	堅牢性	責任	追跡可能性	モニタリング監査	ガバナンス	その他
米国	6	1	2			1	1		1	1		3	3	1	4	5	4	1	2	2	3		
カナダ	1							1								1							1
英国	6		2		1			1		2		1	1	2	5	5	5	1	2		2		1
イタリア	1									1				1		1							
オランダ	2		1				1		1	2				1	2	2	1		1	1			
スウェーデン	2		1	1			1	1					1	1	1	2							
デンマーク	2		2	2			1		1			1	1	1	2	2	1	1			1		
ドイツ	3	1	2	2	3				1	2		2	2	2	2	3	2	3	1	1	1		
ノルウェー	2		2		1		2		1			1		1	2	1	1	1		1	1		
フィンランド	2		2	1			1		1			1		1	1	2	1		2	2	1		
フランス	1								1						1	1	1				1		
インド	1		1	1								1	1	1	1	1	1		1				
韓国	1		1	1		1						1		1	1	1	1	1					
シンガポール	2		1												2	2	1						
中国	3		3	2	2	2	2	2	3	1		3	2	2	3	2	2	2		2	3	3	
オーストラリア	1	1	1									1	1	1	1	1	1		1		1		
EU	2		2	1	2	1	2		1			2	1	2	2	1	2	2	2	1	1		
国際機関	2		2	2	2	1	2	1	1		1	2	1	2	1	2	1		1	1	1		
合計	40	3	25	13	11	6	13	6	12	9	1	19	14	20	31	35	25	12	13	11	16	3	2

（注1）政府機関（EUの欧州委員会及びAIハイレベル専門家グループを含む。）及び国際機関を対象として整理。
（注2）記載があるものはピンク色に着色。また、各国等の指針等の数に対し、半数を超える指針等に記載がある場合、赤字で表示。

3

表 1 総務省 AI ネットワーク社会推進会議「（２）海外における原則・指針・ガイドライン等との比較からの検証」（「報告書 2022～「安心・安全で信頼性のある AI の社会実装」の更なる推進～」78 頁より引用）

このような指針にとどまらず、すでに EU は一般データ保護規則（GDPR）[5]により AI の透明性を一部実現する規制を開始しており、AI 規則[6]、デジタルサービス法（DSA）、デジタル市場法（DMA）[7]といったさらなる規制の導入へと歩みを

---

[4] 総務省 AI ネットワーク社会推進会議「報告書 2022～「安心・安全で信頼性のある AI の社会実装」の更なる推進～」（2022 年 7 月 25 日）78 頁 https://www.soumu.go.jp/main_content/000826564.pdf （2023 年、1 月 28 日閲覧).

[5] European Commission “EU data protection rules” https://commission.europa.eu/law/law-topic/data-protection_en （2023 年 1 月 28 日閲覧）.

[6] European Commission “A European approach to artificial intelligence」https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence” （2023 年 1 月 28 日閲覧）.

[7] DSA, DMA それぞれにつき European Commission “The Digital Services Act package”

進めている。日本もデジタルプラットフォーム取引透明化法（DPF 取引透明化法）において、透明性の観点からの共同規制を導入した。すでに AI アルゴリズムの透明性は、いつか実現すべき理想の姿ではなく、現実に対応すべき社会からの要請なのである。

1.4 透明性による解決の可能性

なぜ AI アルゴリズムの透明性がこれほど声高に叫ばれるのか。

人間が AI と調和する社会を生み出すには、人間が AI を理解する必要がある。AI に限らず、イノベーションにはリスクがつきものであり、そのリスクは評価されコントロールされることによってイノベーションが社会に根ざす準備が整っていく。リスクを持った AI と共存するためには AI に対する信頼の醸成が不可欠であり、AI との適切な共存方法を人間社会が理解できなければならない。たとえば、我々が口にする食品には栄養成分や原材料、原産国、消費期限などの表示がなされている。これは当該食品を購入するために必要な安心と安全のための透明性である。より講学上の概念としては、自己決定権を保護するための措置である。冒頭に記載した AI によるイノベーションのリスク・弊害事例は、事業者が AI によるリスクやコントロール方法を社会と意識的にコミュニケーションすれば解決の緒が見いだせる。

AI は本来的に価値中立的であり、人間よりも信頼できるという言説が存在するが、問題の本質を見誤っている。AI を開発するのは人間であり、結果的に「体系的なバイアス」[8]を導入したり、「歴史的差別を強化したり」[9]することが指摘されている。人間よりも信頼できるかが問題なのではなく、人間が開発する AI について信頼できるかを個人及び社会が判断する基盤として、AI の透明性が求められるのである。

しかし、AI アルゴリズムの透明性は、利用者個人や社会がメリットを享受するための事業者への足枷ではない。むしろ AI を開発・提供する事業者の観点からも様々な効用が期待できる。事業者は、AI アルゴリズムの透明性を担保することで、競合他社との差別要因を明らかにできる。引き続き食のアナロジーを用いれば、有機栽培や無農薬野菜、産地直送などを差別化要因とすることとパラレルである。また、透明性を実現する取り組みは、事業者のリスクマネジメントにも結

---

https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package （2023 年 1 月 28 日閲覧）.

[8] Kashin, K., King, G., & Soneji, S. (2015). Systematic Bias and Nontransparency in US Social Security Administration Forecasts. Journal of Economic Perspectives, 2(29), 239–258.（https://www.aeaweb.org/articles?id=10.1257/jep.29.2.239 （2023 年 1 月 28 日閲覧））

[9] Kroll, J. A. (2015). Accountable algorithms. (Doctoral Dissertation) Princeton University（https://dataspace.princeton.edu/handle/88435/dsp014b29b837r （2023 年 1 月 28 日閲覧））

びつく。本稿で提案するツールキットは、AI がもたらすリスクを体系的に整理した上でどのような組織的又は技術的措置をとるかを検討するチェックリストの役割を果たすことができる。そのように評価したリスクに対して、事業者がどのようにリスクをコントロールするかを判断し、社会とコミュニケーションを取ることで、いざ顕在化したリスクに対しては社会とともに解決を目指すことが可能になる。結果的に、AI を開発・提供する事業者にとっては、リスクマネジメントのプロセスを俯瞰的に履践するに等しい。

事業者からは、AI アルゴリズムの内部の開示に対して、利用者にハックされるおそれや競合他社に企業秘密を知られてしまう懸念を表明されることがある。しかし、本稿の提案する AI アルゴリズムの透明性とは、つぶさにコードを開示することを意味しない。そのような開示を行ったところでそれが意味するころを理解できるのは限定された研究者や開発者のみであり、大半の利用者に対しては透明性の説明にならない。国際的な議論を踏まえて本稿が提案するのは、特徴量や重み付けに関するより抽象化され、合理的に理解可能な形で言語化された説明である。求められるのは、秘伝のタレのレシピの開示ではなく、そのメニューの具材、調理法、味付けの開示である。

2　ツールキットの概要

2.1　ツールキットの目的・役割

前項で見たように、AI アルゴリズムの透明性は必要性が様々な場所で唱えられているにもかかわらず、具体的な方策は自主的な対応に委ねられていたり、各国の個別具体的な規制が先行したりして、全体的な体系的整理や理解がないままに実務が進み始めているように思われる。本稿が提案するツールキット（以下「本ツールキット」）は、最新の議論を反映しながら AI アルゴリズムの透明性について体系的な整理を行った上で、実践的な内容を備えた対応カタログの役割を果たすことを目指す。翻って言えば、各国の規制や事業者による自主規制が解決しようとする問題は、本ツールキットを参照することで俯瞰して理解することができよう。

透明性に関する原則や指針ではなくツールキットを提案するのは、AI を利用する主体、目的、方法、場面などによって、必要とされる内容や程度が異なるためである。壮大な理想像ではなく、文脈に応じて各ステークホルダーが使い分けることができるカタログを目指す。したがって、本ツールキットは透明性を担保するためのオプションの集合体であって、厳格な開示項目として捉えられるべきではない。AI を開発・提供する事業者と利用者・社会との間の信頼関係をどう構築するか、コミュニケーションはどうあるべきかといった観点から本ツールキットの利用方法を解説していこう。

2.2　ツールキットの射程

1.1 で述べたとおり、本稿が想定する AI とは機械学習モデルや深層学習モデルを用いたソフトウェアやシステムを想定している。一方、本ツールキットの対象となる AI の用途や目的などを制限することはないため、広く AI を用いたソフトウェアやシステム全てに利用できる内容になっている。

ただし、透明性という観点だけでは解決できない AI も存在する。特に生命や身体への危険を伴うプロダクトや利用者にとって代替手段の少ないプロダクトに組み込まれる AI（自動運転やデジタルプラットフォームなど）は、利用者の安全性や利用における公正性を一義的に追求しなければならず、リスクの存在について透明性をもってコミュニケーションするだけでは社会実装は認められない[10]。その意味で、本ツールキットはリスクと共存すべき AI を前提にして構成されている。

2.3　ツールキットの構成

図 1 のように、機械学習モデルのアルゴリズムは、人間が指令や条件を指定してアルゴリズムを開発していく伝統的なプログラミングと異なり、モデルそのものの生成をデータに基づいて機械が担う[11]。本ツールキットの構成は、この点への理解が不可欠であるため、図 1 を用いて具体的に双方の違いを確認したい。

---

[10] 仮に「自動運転車が事故に遭う確率」が開示されたとしても、利用者にできることはその車に乗らないことしかない。

[11] 機械学習における「アルゴリズム」と「モデル」の定義を念のため明確にしておきたい。機械学習アルゴリズムとは、分類、予測といった目的に応じ、入力データから処理の方法を学習し、あるいは入力データに応じ処理結果を返す手順を指す。目的に応じ、ロジスティック回帰、決定木、ニューラルネットワークなどのアルゴリズムが用いられる。機械学習モデルとは、機械学習アルゴリズムが学習データを処理することで得られた結果で、新しい入力データに対し、何らかの識別・判断結果を出力するプログラムを指す。同じアルゴリズムを用いても、学習するデータが異なると出力されるモデルは異なるものになる。本文内の用語もこの定義による。

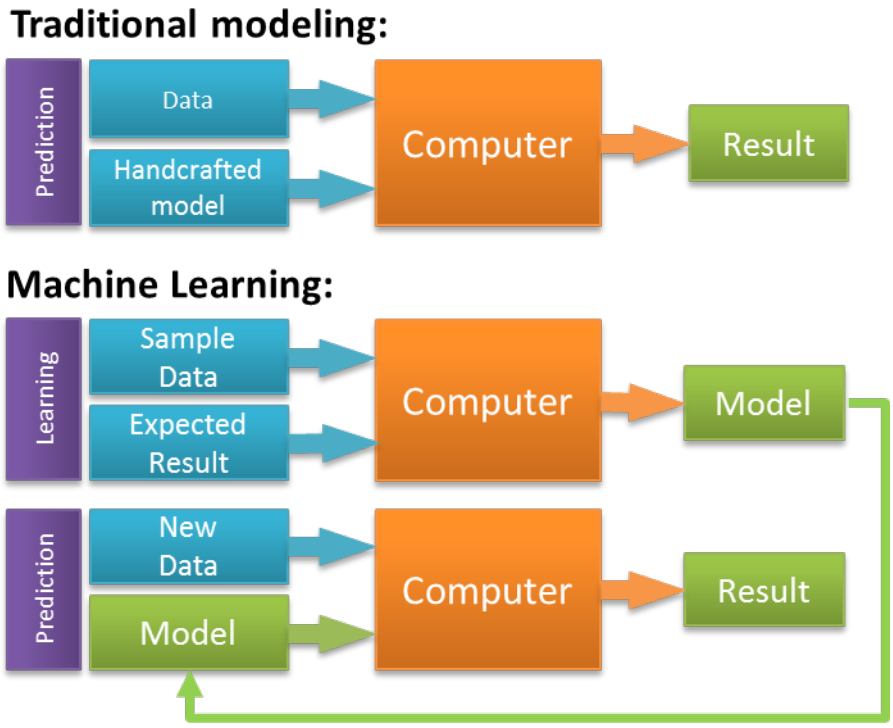


図 1：伝統的なプログラミングと機械学習アルゴリズムの比較（ZEISS[12]より引用）

ルールベースのような、伝統的なソフトウェア開発では、コンピュータ上で専門家が作成したプログラム（図 1 では「Handcrafted model」）がデータを処理することにより出力（認識、推論した結果。図 1 では「Result」）が得られる。

一方、機械学習アルゴリズムにおける学習フェーズの場合は、「入力データとその入力に期待される出力（正解）」のサンプル（学習データ。図 1 では「Sample Data」と「Expected Result」）から、入出力の関係を示すモデル（図 1 では「Model」）が得られる。このとき、データからモデルを得るための手順が機械学習アルゴリズムにおける学習フェーズである。得られたモデルを用いることで、新たな入力データに対しても認識・推論結果（図 1 では「Result」）を得られるようになる。単純な例として、曲線 $f(x)=w_0+w_1x+w_2x^2 \ldots + w_Mx^M$（$w_0, \ldots, w_M$ は係数、x^n は x の n 乗）を学習することを考えると、与えられた(x,f(x))の組み合わせから係数を推定する手法（例えば最小二乗法）が学習アルゴリズムであり、具体的な係数を当てはめた関数が学習結果のモデルということになる。

さらに、モデルを開発する主体がアウトプット（モデルが抽出する結果）の妥当性や品質をチェックし、また追加のモデル開発を行い、これを繰り返すという点で、人間の関与も引き続き多大である。すなわち、Seaver の言葉を借りれば、

[12] ZEISS「The Relation between Computer Vision and Machine Learning」https://blogs.zeiss.com/digital/page/2/（2023 年 1 月 28 日閲覧）.

「アルゴリズムとは、独立した一つの小さな箱ではなく、何百人もの手が入り込み、調整し、部品を交換し、新しいアレンジメントで実験する、大規模でネットワーク化されたもの」[13]なのである。

したがって、単純にモデルのコードを開示したり、モデルについてのみ説明したりするだけでは何ら問題は解決されない。

そこで、本ツールキットは、AI アルゴリズムをデータ、モデル、アウトプット、人間の関与、そして全体のシステム設計の五つが有機的に連携することで成立していると捉え、別表 1 のとおり、これらの観点を踏まえた AI 全体の透明性を担保しようとする。次項より各項目の詳細について述べる。

3　ツールキットの解説

AI アルゴリズムの透明性を議論するにあたって、まずはこの開発・提供のプロセスについて図 2 をもとに大まかに認識を整えたい。

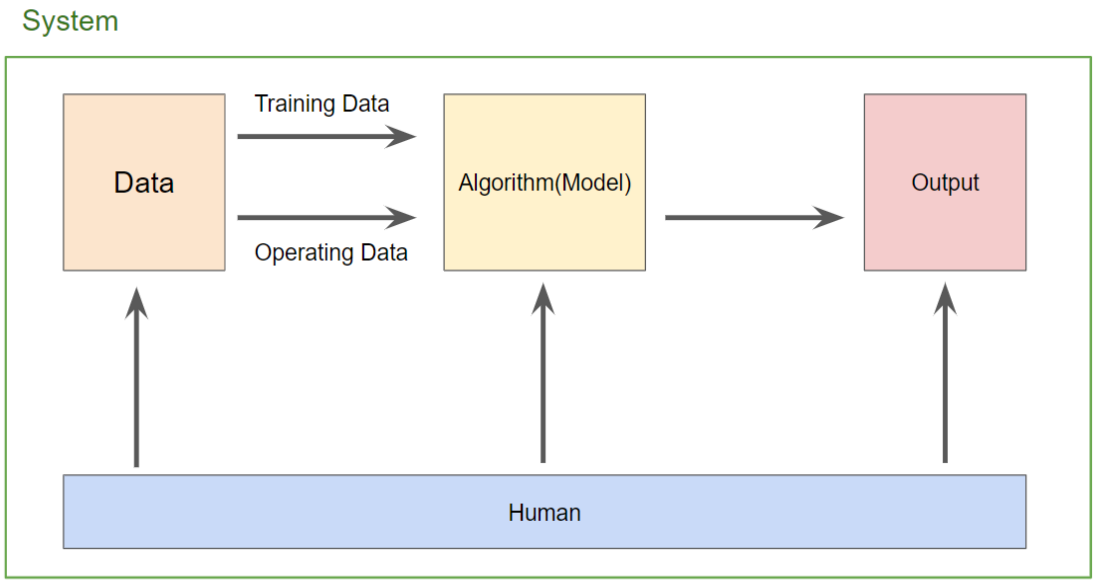


図 2：AI の開発プロセスの概要（著者作成）

まず、AI のモデルを開発するために、学習データ（Training Data）が必要となる。これをもとに AI はモデルを構築し、そのモデルが実運用に耐えうるかをテストデータで確認を行う。そして実運用にあたっては、実際に当該モデルに入力するインプットとしてのデータ（Operating Data）が入力され、それに基づいて当該モデルが出力（Output）を生み出す。この全ての過程に人間による関与（Human）が含まれる。これらの各要素を横断する思想やデザインを本ツールキットでは全体のシステム設計（System）と呼

[13] Seaver, N. (2019). Knowing algorithms. DigitalSTS, 412-422. https://digitalsts.net/wp-content/uploads/2019/03/26_Knowing-Algorithms.pdf 2023 年 1 月 28 日閲覧）

んでいる。

別表 1「AI アルゴリズムの透明性開示項目リスト」は、AI 開発の大まかなライフサイクルに基づいた開示事項を昇順に列挙し、各開示事項が上述の五つの要素のいずれに関するものであるかを明示している。以下では、五つの要素毎で実現すべき透明性のあり方を具体的に見ていくことにする。なお、説明の便宜上、データ、モデル、アウトプット、全体のシステム設計、人間による関与の順に見ていくものとするが、基本的には順不同である。

3.1 データ（Data）

3.1.1 学習データ（Training Data）の概要

ある AI を開発するにあたって、用いられている学習データ（テストに用いるためのデータを含む。以下同様）の概要を開示する。たとえば、テキストの生成 AI であれば利用する大規模言語モデルによって用いられている学習データは異なりうる。画像生成 AI であれば、用いている画像のデータセットの特徴を利用者や一般社会が理解しやすい形で言語化する。

学習方法によって異なる開示を工夫することも重要である。たとえば、教師あり学習の場合にはどのようなデータを教師データとしているのか、何を教えているのか。教師なし学習の場合にはモデルに何をさせようとしているのか。強化学習の場合にはモデルに与えるインセンティブなどを明確にすることが考えられる。

また、データセットを構築する全ての変数や項目を開示することはできずとも、何らかの特殊な事項やセンシティブな事項が含まれる場合にはその旨を明示することも必要となる可能性がある。

3.1.2 学習データのソース

利用者や研究者がアクセス可能な形式で公開されているデータセットへのアクセスリンクを提供する。

オープンデータであればこれはそれほど困難なものではないが、そうでない場合、全体のデータセットと内容に齟齬のないサンプル化したデータセットを提供することも考えられる。

3.1.3 学習データの入手経路

学習データがどのような方法で入手されたのかを説明する。

すでに個人情報に関しては、各国の個人情報保護規制によって少なくとも取得段階で取得方法や利用目的に関する開示が求められていることがほとんどであるが、AI の利用段階でもこの開示は求められるべきである。また、個人情報以外のデータ（たとえば統計データやテキストデータ等）についても同じく開示されるべきケースはあるだろう。

3.1.4 学習データの DEIB（Diversity, Equity, Inclusion, and Belonging）ポリシー

AI のモデルを開発するにあたって、学習データが考慮している多様性や公平性、網羅性、代表性（当該 AI が利用対象にする母集団を適切に代表しているか）、包摂性といったいわゆる「Diversity, Equity, Inclusion, and Belonging」に関する事業者のポリシーや見解を記述する。

AI がバイアスを持つとき、その一つの要因は学習データそのものが偏ってしまっていることが挙げられる。したがって、実運用に耐えうるように偏りをいかになくしているか、あるいは偏りをあえて残す場合はその理由を説明することが、利用者の便宜に資するのみならず、事業者の開発時の意識強化にも繋がる。たとえば、日本を市場として AI を提供する場合、米国社会にとっては明らかに偏りがあるとされるデータセットであっても日本社会における多様性や包摂性の観点からは問題が少ない場合がある。その場合、どの社会からも受け入れられるような多様性のあるデータセットの利用を追求するのではなく、当該市場における DEIB 観点からの十分性の検討結果を事業者が明らかにすることが重要である。

また、学習データから意図的に除外されたデータを説明することで、多様性や網羅性、代表性に関する説明が可能となる場合がある。

3.1.5 実運用データ（Operating Data）

AI が実際に活用される時点で用いられるデータの概要を記述する。

すでに述べた学習データに関する項目が実運用データにも全て適用できる場合、「データ」という大きな項目に整理可能である。一方、学習データとは異なり、たとえば利用者が直接入力するデータが実運用データになる場合には、実際に用いられるデータについて、学習データとは別途説明が必要となる場合がある。

3.2 モデル・Model

3.2.1 利用フェーズ

AI で用いられているモデルが製品・サービス開発のどのようなフェーズで使われているかを説明する。たとえばアイデア段階、研究段階、ベータ版、実運用段階といったフェーズを明示することで、利用者は当該 AI に対する期待値や利用方法を調整することが可能になる。

3.2.2 学習手法

当該モデルがどのような学習手法によって開発されているかを記述する。機械学習、深層学習、エキスパートシステム、ルールベースなど、その説明方法や粒度は利用者や研究者の理解に資する観点からケースバイケースで選択すべきである。

3.2.3 説明可能性

当該モデルを事業者が合理的に説明することのできる程度に限界がある

場合、事業者の理解を記述する。

たとえば深層学習モデルについては、どうしてもそのブラックボックス性ゆえに、抽象化したパラメータ・特徴量等による言語化にも限界があるケースが存在する。しかし、根本に立ち返って、透明性が実現しようとするのは社会や利用者による AI プロダクトへの信頼、あるいは自らの自己決定権を含む権利保護であると考えれば、「人間が理解可能な粒度のある程度の説明」という透明性で解決できる事項もまた多い。昨今導入が進んでいる Explainable AI（XAI）などにもまさにそういった観点からの取り組みが存在している。そこで、事業者による過度な説明責任を軽減し、代替できる説得的な説明方法を案内することが考えられる。

3.2.4 パラメータ・特徴量とその目的・理由

当該モデルを構成するパラメータ・特徴量について説明する。その目的や理由について事業者の理解が記載することが望ましい。

パラメータまたは特徴量とは、データセットが持つ特徴を定量的に表現したものを指すが、ここではその厳密な定義を求めるものではなく、AI アルゴリズムがある出力結果を求める際に考慮している数値化可能な特徴・設定値などといった認識で揃えたい。また、どのパラメータをどの程度重視するかを定量化したものを重みと呼ぶ（説明は後述する）。AI プロダクト開発においてはパラメータをどのように設計しまたはデータから獲得するか、重みをどのように設定したり最適化したりするかが主な論点となる。よく挙げられる事例として、アイスクリームの売上を予測する AI アルゴリズムを考えると、天気や気温、商品数、立地（最寄駅からの距離など）など様々なデータと売上データの関連性を学習しながら AI のモデルが構築され、それら一つひとつの要素がパラメータを構成することになる。深層学習に代表されるように、人手で個々のパラメータを設計するのではなく、データから学習する中で、入力データの中で目的に有効なパラメータが「獲得」される場合もある。

しかし、細かなパラメータなどを全て開示することが求められるわけではない。詳細なパラメータの開示は利用者の理解を困難にさせ、事業者にとっても秘伝のたれとしての企業秘密を開示することにも繋がりかねない。むしろ、利用者が理解可能な程度に抽象化された、モデルが考慮する重要なパラメータについて、次項に述べる重み付けと併せて言語化することが最も実務的である。先の事例でいえば、「天気を以下の厳格な定義に基づいて 18 種類に分類し、…」といった説明が求められるのではなく、「当日の天気」というパラメータの説明で利用者は適切な理解を得ることが多いだろう。また、深層学習のようにパラメータ自体を学習する場合は、学習の手法についての言

語化・開示となることも考えられる。また、モデルに影響を与えないが、影響を与えるのではないかと誤解されやすい要素を明示的に説明することで、AI アルゴリズムの適正な利用を促すことも考えられる。

この観点で 2023 年 1 月 28 日現在最も参考になる先行規制は EU のプラットフォーム透明化規則（P2B 規則）[14]だろう。同規則は、オンラインプラットフォーム上において中小企業を含む事業者の取引の公正性と透明性を実現すべく、オンラインプラットフォーム事業者に対して主要なアルゴリズムやパラメータの透明性を実現するよう求めている。具体的に同規則では、ランキングを決定する主要なパラメータを決定し、説明することが求められている。AI アルゴリズム内に多数のパラメータが存在している場合、それらをいくつかのカテゴリに分類し、決定的な役割を果たす大きなカテゴリを残していく作業を履践する。そして、主要なパラメータを決定するまでの判断プロセス自体も記述することがベストプラクティスとされている。さらに、一般的に用いられるパラメータ一覧も公表[15]されており、どういったものがいわゆる「パラメータ」の記述候補となるのかがわかりやすい。たとえば「コンテンツの品質（ウェブサイトへの相互リンク、豊富度、言語の質・数）」や「地理的な近接性」、「商品・サービスの人気」、「在庫の有無」などが記載されている。

また、日本の DPF 取引透明化法でも主要なパラメータの開示は求められているが、P2B 規則ではさらにその理由の説明までを求めている点に特徴がある。AI アルゴリズムの類型にもよるが、目的や理由、意図するところまでを開示できれば、利用者や社会にとって、どのように当該 AI アルゴリズムを利活用すればよいかを一歩踏み込んで理解できるだろう。

実際の開示事例をいくつか抜粋する。

たとえば Google LLC は、公式サイト[16]において、Google Play におけるアプリの表示順位について、「ユーザとの関連性」、「アプリのエクスペリエンスの品質」、「エディターによる評価」、「広告」、「ユーザーエクスペリエンス」を「主な要因」とし、ただし「ランキングに与える影響の大きさは、ユーザ

---

[14] European Commission “Platform-to-business trading practices」https://digital-strategy.ec.europa.eu/en/policies/platform-business-trading-practices”（2023 年 1 月 28 日閲覧）.

[15] EUR-Lex “Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council”(Annex1) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020XC1208%2801%29（2023 年 1 月 28 日閲覧）.

[16] Google LLC 「アプリの検出とランキング」 https://support.google.com/googleplay/android-developer/answer/9958766?hl=ja（2023 年 1 月 28 日閲覧）

ーが見ている Google Play 上の場所、ユーザーが使用しているデバイス、ユーザーの設定によって異な」るとしている。さらにそれぞれの項目について、より深い理解を促すために、詳しい説明を続けており、その中でそのような要素を「主な要因」とした理由や経緯も記載されている。

アマゾンジャパン合同会社[17]は、オンラインモール「Amazon 出品サービス」の商品の表示順位を決定する主要なパラメータとして、商品の提供事業者向けには「商品情報にあるテキストその他の情報との一致率、価格、ポイント、在庫状況、品揃え、販売履歴などの要素」によって決まると開示しており、一般利用者向けには、「お客様の行動（商品の購入頻度など）、商品の情報（商品名、価格、商品の説明など）、在庫の有無、お届けまでにかかる時間、各種料金（配送料など）、その商品に興味を持たれるかどうか（新商品など）」といった内容で開示している。出店者向けと利用者向けでそれぞれ異なる説明を置いている点が特徴的である。

ヤフー株式会社は、公式サイト[18]において、「検索順位やランキングの決定の仕組み、各表示枠」について、「優良配送アイコンが付された商品」を「主に」「考慮」し、さらに、「その他の考慮要素」として、検索ワードとの関連性、商品の購入件数、購入顧客数、販売個数、商品レビュー数、ストア評価数、ストア評価の平均値、ストア評価の合計値等」を挙げている。同時に、「出店者様からの販促費等の支払が、ランキングの順位に直接的に影響を及ぼすことはありません」と明記することで、誤解を生みやすい内容にも対応している。また、報告書[19]において、このような主要なパラメータの抽出に至った理由や観点を丁寧に説明している。すなわち、「当社オンラインモールの『おすすめ順』の表示順位を決定する主要な事項は、表示順位 を改善したいと考える出店者に参考にしていただける要素を公開すべきという方針のもとで、約 120 の要素の中から、順位を左右する度合いの大きい要素を抽出しました。ストアツールにおいて閲覧可能な数値（購入顧客数、注文件数、レビュー数）や、「ストアパフォーマンス」として開示されているスコア（注文件数、ストア評価平均点等）とできる限り一致する形で項目化することで、出店者が改善に繋げやすいようにしています」とのことである。さら

---

17 アマゾンジャパン合同会社「令和 3 年度特定デジタルプラットフォーム提供者による定期報告書概要」50 頁,66 頁
https://www.meti.go.jp/press/2022/12/20221222005/20221222005-c.pdf （2023 年 1 月 28 日閲覧）

18 ヤフー株式会社「透明性向上のための取組みのご紹介」 https://business-ec.yahoo.co.jp/shopping/digitalplatformer/ （2023 年 1 月 28 日閲覧）

19 ヤフー株式会社「令和 3 年度特定デジタルプラットフォーム提供者による定期報告書概要」43 頁

に機械学習アルゴリズムの特質に基づく説明として、「表示順位を決定する要素の内容や重みづけは、機械学習により日々変動するため、その全てを常に正確に開示することはほぼ不可能ですが、順位を左右する度合いが大きい主要な要素を新たに追加する場合は、必ず開示内容を更新することとしています」と明記している。

3.2.5 重み付けとその目的・理由

各パラメータに対する重み付けや係数について説明する。重みとは、どのパラメータを重視するかを定量化した係数である。先のアイスクリームの売上予測の例で言えば、気温を重視するのであれば気温を表すパラメータに対する重みを大きくすることになり、立地を重視する場合は対応する重みが大きくなる。重みは、解決したい問題に対する専門家が設定する場合もあれば、データからの学習により調整・最適化される場合もある。重みについては、パラメータ・特徴量と同じく、その目的や理由について事業者の理解が記載することが望ましい。

パラメータはモデル内で相互に対等ではないことが多く、相互関係や相対的な重要度を考慮し、利用者がパラメータ間の重要性の違いを理解できるような説明が求められることがある。ただし、重み付けは日常的に変化しうるし、細かな調整こそ事業者内部に秘匿すべきことが少なくないことから、個別の既述よりも相対的な関係性の記述が好まれることが多いだろう。

EU の P2B 規則を例に取ってみても、正確な重み付けの指標の開示やパラメータ間の係数の開示まで求められているわけではない。利用者にとって有益なのは、どのような挙動をすべきか想定できることであるから、相対的に重要なパラメータに限定して開示することで十分に目的は達成できるだろう。したがって、本項目を独立して記載せず、前項のパラメータ・特徴量とともに「主要なパラメータ」や「重要な要素」のような形で、パラメータ間の相対的な比較から影響の大きいパラメータを抽出する方法が考えられる。パラメータの重要性の考え方としては、モデルの抽出するアウトプット（たとえば表示順位やレーティング結果、審査結果、判定結果など）に影響を及ぼすもののうち、特に事業者が重視している事項（モデルを通じて、AI アルゴリズムへの事業者の思想・哲学が反映される）、又は特に利用者が AI アルゴリズムを最適に利用するために考慮すべき要素（利用者の最適化行動を通じて、AI アルゴリズムへの事業者の思想・哲学が反映される）が一つの尺度になるだろう。

前項ですでに挙げたいくつかの開示事例も基本的に主要なパラメータを説明する方針を採っており、今後一般化されるように思われる。一方、公的機関が利用する AI アルゴリズムや、利用に対するリスクが大きいと考えら

れる AI アルゴリズムについては、主要なパラメータに限らず、相互の関係性などが記述されるべき場合も考えられるため、そのような場合には個別に独立した説明がなされるべきである。

3.2.6 課金を含む利用者区分毎の異なる取扱い

パラメータや重み付けが利用者の課金行為その他の行為又は属性によって変化しうる場合、どのような区分で異なる取扱いがありうるのかを説明する。たとえば課金の有無による区分、事業者による区分、年齢による区分、性別による区分、総利用期間・頻度による区分などで異なる取扱いを行う場合などが考えられる。ここには、事業者（グループ企業を含む）が自ら運営するサービスとその他のサービスとを AI アルゴリズム上で異なる取扱いをする場合にはそれを説明することも含まれる。

実際に、EU の P2B 規則では、直接又は間接の対価の支払いによってランキングアルゴリズムにどのような影響を及ぼすかを開示するよう求める規制がある。また、同じく自社優遇に関する規制も P2B 規則や DMA などに含まれている。

たとえば、プラットフォーム事業者が AI を用いてレコメンデーションエンジンやレーティングを調整する場合を想定してみよう。あるソーシャル・メディアでは、課金した利用者の投稿を表示順位の観点で優先させることや、返信や拡散投稿を表示されやすくすることを開示している。また、特定の年齢の利用者に対しては一定の AI アルゴリズムを使用しないポリシーを有する事業者なども存在する。

3.2.7 変更の適正手続

AI アルゴリズムが変更される場合に事業者が履践する適正手続を記述する。

AI アルゴリズムの変更の内容が実質的に利用者に影響を及ぼす場合、たとえばレーティングに影響を及ぼしたり、レコメンデーションエンジンの結果が変更されたりする場合、利用者の予測可能性や防御の観点から、事前の通知、変更の目的、変更によって予想される影響などを説明することが考えられる。当然ながら、利用者による悪用（エンジンのハック）は予防されるべきであるため、事前の通知内容を一定程度抽象化することはありうる。

3.3 アウトプット（Output）

3.3.1 パフォーマンス精度・限界

AI のモデルが出力した結果の精度や限界を説明する。

たとえばその結果がどれほど信頼できるのかを有意確率やエラー率、正解率などによって記述することで、利用者は当該 AI に対して適切な期待値や利用方法を調整することができる。

また、よく起こる誤りの記述や不正解を起こしやすいケースの説明なども利用者にとって有意義となるだろう。

3.4 全体のシステム設計（System）

AI は、用いる学習手法によっては、一般の利用者の理解の範疇を超えるロジックをもってアウトプットが出力される。その AI プロダクトを信頼してよいのかどうかは、事前事後の透明性や説明責任、あるいはその AI プロダクトを開発する主体やガバナンス体制への信頼へと委ねられる。そこで、全体のシステム設計と人間の関与に関する項目は、AI を用いる主体としての人間の関与方法や AI の設計思想、ガバナンス体制について記述することで、当該 AI への理解や信頼獲得を実現しようとするものである。

3.4.1 利用目的

提供するサービスやシステムにおいて、AI を利用する目的を記載する。

利用目的は、開発の経緯や将来像などを含むことで、より具体的に事業者の利用意図や思想を理解することに繋がる。

3.4.2 ベネフィット又はインパクト

当該 AI を用いることで得られる便益（ベネフィット）又は影響（インパクト）を説明する。

この項目は利用者の目線、事業者の目線、社会全体の目線など様々な観点からの記述が可能である。リスクとは別に、想定するポジティブな影響を中心に記述することで、利用者や社会全体の理解へと繋げることが可能となる。

3.4.3 利用方法

提供するサービスやシステムにおいて、どのように AI を用いるのかを説明する。

サービス全体の中で AI がどのように利用されているのかは、一見すると利用者にはわからないことがしばしば発生する。具体的には、どのプロセスで AI を用いているのか、AI が出力したアウトプットは人間の意思決定をサポートする役割にすぎないのか、それとも AI によるアウトプットそのものが意思決定となるのかを具体的に説明することが望ましい。

たとえば、クレジットカードの決済履歴情報や購買履歴などに基づく与信審査で AI が用いられる際、AI による審査結果そのものが最終的な審査判断となるのか、それとも人間がその審査結果を参考にしながら審査判断を下すのかは意味合いが全く異なり、かつ開示されない限り利用者にはわからない。

3.4.4 リスクマネジメント

当該 AI の利用によって生じうるリスクについて、評価（アセスメント）、リスクの最小化方法、又はコントロール方法を説明する。

AI が生み出すリスクを秘匿するのではなく、むしろリスクと共存してイ

ノベーションを生み出すために、リスクの存在を利用者と共有し、事業者がどのような努力にとってこれを最小化しようとしているか、あるいはコントロールすべき項目として対応しているかを記述する。このようなスタンスによって、利用者もリスクを認識した利用が可能となるし、利用者自身のリスクマネジメントにも繋がる。

また、一般的に同様の AI が包含するリスクに対して、自社の AI がどのような優位性を持つのかをリスクマネジメントの観点から説明することで差別化を図ることも可能かもしれない。

なお、リスクマネジメントについては、データ、モデル、アウトプットなどそれぞれの観点から整理することでより理解しやすい体系となるだろう。

3.4.5 教育体制

AI の利用によって利用者や第三者に対して損害を生じさせることや、社会に対して新たな問題を生じさせることを可能な限り避けるため、事業者内で AI の開発や提供に関して求められる倫理意識や留意点などの教育体制をどのように整えているか、トレーニングの実施内容や頻度、認証制度などを説明する。これにより、事業者が目指す取扱いの品質水準が可視化される。

3.4.6 監査体制

データやモデル、アウトプットなどについて、どのような監査体制を整備しているかを説明する。監査主体の別（内部か外部か）、監査対象の別（開発プロセスの監査か、技術そのものの監査か）などの観点から、事後的にどのような監査を行い、フィードバックがなされているかを明らかにする。

3.4.7 利用者によるコントロール

3.4.7.1 AI アルゴリズム利用の選択権

AI アルゴリズムを利用するか否か、又は複数の選択肢がある場合にどの AI アルゴリズムを利用するかを利用者が選択できることができるかを説明する。

たとえば、あるソーシャル・メディアではレコメンデーションエンジンに基づくフィードと投稿時間のみに基づくフィードを区別し、利用者が選択できるようにしている。しかし、一瞥してその区別や利用者による選択の可否がわからないことがある。利用者による主体的な AI アルゴリズムの選択を可能にするのであれば、選択権が存在すること自体が適切に伝達されなければならない。

3.4.7.2 フィードバック

AI アルゴリズムやデータへの利用者によるフィードバック方法の有無、また存在する場合はその利用方法を説明する。

たとえば、あるアグリゲーション・プラットフォームは、利用者が

関心を持っている項目を選択することができ、それによって AI アルゴリズムにフィードバックできる仕組みとしている。さらに、事業者によっては、事業者が推測したプロファイリング結果について、利用者がそれを訂正したり追加したりすることで、より最適な AI アルゴリズムに調整することを可能にしている。

3.4.7.3 データ・オプトアウト

利用者が、AI のモデルのために用いられている学習データにおいて自らの情報や自身の知的財産等の利用を停止するよう申請する方法の有無、また存在する場合はその申請方法を説明する。

特に生成 AI で問題になることが多いが、学習データにデータを用いられたくないと考える利用者が一定数存在することを想定するサービスやシステムは、このようなオプトアウト権の実装を検討すべきである。個人データの利用に関するオプトアウト権については、EU や米国の一部の州で認められているが、個人データにとどまらないデータ利用については事業者の判断が求められている。

3.5 人間の関与（Human）

3.5.1 開発責任者・管理責任者

AI を開発・提供する責任者及び管理責任者を記載する。本記載によって、当該 AI によって生じた責任の帰属先を明確にする。個人情報の取扱責任者の明示は国際的にも必須となり始めているが、その流れにも即するものである。

3.5.2 開発チームの DEIB ポリシー

AI を開発するチームが考慮している多様性や公平性、網羅性、代表性、包摂性といったいわゆる「Diversity, Equity, Inclusion, and Belonging」に関する事業者のポリシーや見解を記述する。

学習データの取扱いや AI のアウトプットへの評価を行うチームがアンコンシャス・バイアスを持つことは避けられないため、これに対する事業者の措置を説明する項目である。AI が持つ偏りを最小化する方法の一つとして、これを開発・提供するチームの多様性を担保すること、あるいは想定利用者の母集団との整合性を取ることが考えられる。しかし、開発チームがこれら全ての観点で必要十分な属性を持つことができるのは稀である。そこで、どのようにアンコンシャス・バイアスを排除しようとしているか、複数の観点からチェックを行う体制は整備されているかなどについて、事業者のポリシーを説明することも考えられる。

3.5.3 委託先管理

自社以外の委託先とともに AI を開発・提供する場合、その管理方法や体

制について説明する。委託先についても各事業者が自社のガイドラインによる開発を求めるなど、適切な開発を担保する措置を記述することが求められる。

3.5.4 事業者によるモニタリング及びメンテナンス

特に AI が出力したアウトプットに対する品質評価又は安全性評価について、事業者による能動的なモニタリングやメンテナンス体制を記述する。

実運用段階では、利用者に誤ったアウトプットや違法有害な情報が出る影響を最小化するためにトラストアンドセーフティの観点から能動的にアウトプットをモニタリングする体制が必要となるケースがある。またそれらの具体的な運用状況（検出件数等）を利用者や一般社会に共有し、どのようにそれらを AI にフィードバックするかなどを説明することでさらなる信頼の構築へと繋がるだろう。

3.5.5 問い合わせ対応

利用者や研究者による問い合わせ窓口やアウトプットに対する不服申立て窓口の案内、カスタマーエクスペリエンス体制及びその運用結果（問い合わせの内容、頻度等）などを記述する。

たとえば、日本の DPF 取引透明化法に基づく開示では、Amazon や Google らが苦情の類型、苦情に対する処理期間の平均期間、処理の結果に関する概要（処理結果の割合等）を年次報告書の形式で提出している[20]。各事業者がどのようにユーザからの問い合わせに向き合っているかがわかる良質な情報開示である。

3.6 留意点

なお、特に Human や System に関する項目については、複数のサービスやプロダクトを提供しているような事業者は、サービスやプロダクト単位ではなく事業体単位として横断的に統一した体制を整備している場合もあるだろう。また、事業者全体としてみれば更に透明性を高められる事項や求められる開示事項もありうる可能性がある。しかし、本ツールキットは ver1.0 としてひとまず一つのサービスやシステムにフォーカスした内容としている。

4 ツールキットの利用のための視点

以上のように、本ツールキットの各項目に対して説明を加えてきたが、繰り返し述べたとおり、全ての事業者がこれら全ての項目について常に詳細に開示することが求め

---

[20] 経済産業省「『特定デジタルプラットフォームの透明性及び公正性についての評価』を取りまとめました」
https://www.meti.go.jp/press/2022/12/20221222005/20221222005.html （2023 年 1 月 28 日閲覧）.

られる訳ではない。どの項目についてどのような開示を行うかを決定するために重要な視点として、利用される AI に内在するリスク、AI を利用する主体、情報が必要なステークホルダーの別がある。

4.1 AI 利用によるリスク・影響による分類

EU が AI 規則において、AI をリスクベースで 4 つに分類[21]したように、全ての AI について一律の規制や透明性を求めるのは妥当ではない。当然ながら与えるインパクトの大きさによって問われる透明性は異なるだろう。

一つの参考に過ぎないが、たとえば、以下のような分類をした場合に、Tier 1 ほど高度な透明性が求められると考えられる。ただし、Tier1 が用いる AI の種類によっては、透明性では解決できない問題（リスクが顕在化した時点で多大な被害が生じる場合など）もあるため、その点は留意が必要である。

	Tier 1	Tier 2	Tier3
リスク	利用者の生命・身体や、社会が共有する基本的人権・自由民主主義の価値観に危険を及ぼすもの	利用者に経済的影響を及ぼすもの	その他の AI（利用者の意思決定の補助や業務サポートを行うもの）
具体例	医療的判断を下す AI、大規模な映像監視 AI、投票先マッチング等	求人求職マッチング AI、契約書審査 AI、（クリエイターにとっての）画像生成 AI、飲食店やホテルにとっての事業者レコメンデーションエンジン等	勤務シフト調整 AI、検索エンジン等

4.2 利用主体による分類

AI を用いる主体によっても、求められる透明性は高まる。特に、公的機関がその業務で AI を用いる場合、影響を受けるのは行政サービスの対象としての市民であり、当該 AI の利用については高度な情報の透明性が求められる。また、ある市場において支配的な地位を有する巨大なプラットフォーム事業者についても、その取引の公正性を担保するために高い透明性が求められるといえよう。

---

[21] The European Union ”Regulatory framework proposal on artificial intelligence” https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai（2023 年 1 月 28 日閲覧）.

	Tier 1	Tier 2
主体	政府・公的機関 巨大プラットフォーム	その他

実際に英国政府が 2022 年末に公表したアルゴリズムの透明性記録基準（Algorithmic Transparency Recording Standard）[22]は、政府を含むパブリックセクターがアルゴリズムツールを用いる際の情報の開示方法を網羅的に整理している。本稿と同じく、モデルのみに集中せず、データやアウトプット、人間による関与に至るまで、様々な視点でパブリックセクターが果たすべき説明責任を言語化している。パブリックセクターという主体の重要性に鑑みた透明性の基準として今後様々な場所で参照されるものになるだろう。

4.3 情報の受け手の視点

本ツールキットを用いて AI アルゴリズムの透明性の観点から開示する項目とその記載方法は、開示された結果を確認するステークホルダーの別を考慮した内容でなければならない。いわゆる読み手の視点である。なぜなら、ステークホルダー別に透明性の理解や必要性、用途は全く異なるからであり、それに応じた説明方法が個別に選択されるべきだからである。表 2 は、ロボティクスや AI に関する透明性の理解がステークホルダー毎に異なることを指摘した先行研究の抜粋である。

ステークホルダー	透明性の類型・目的
開発者	・デバッグや改良のために、システムがどのように動作しているかを理解すること ・安全基準のためのモニタリングとテストを容易にすること
利用者	・システムが何をしているのか、なぜしているのかを理解し、不測の事態に何が起こるかを予測できるようにし、技術に対する信頼感を構築すること ・ある特定の予測や決定がなぜ下されたかを理解し、システムが適切に機能したことを確認し、意味のある挑戦を可能にすること（例：審査の承認や刑事判決）
社会	広くシステムの長所と限界を理解し、慣れ親しみ、未知

[22] GOV.UK “Algorithmic Transparency Recording Standard” https://www.gov.uk/government/publications/algorithmic-transparency-template （2023 年 1 月 28 日閲覧）.

	のものに対する無理もない不安を克服すること
専門家・規制当局	特に何か問題が発生した場合（自律走行車による衝突など）、予測や決定の痕跡を詳細に監査する能力を提供すること。これには、主要なデータストリームを保存し、各論理的ステップを追跡することが必要となる場合があり、説明責任と法的責任の割り当てを容易にする。
デプロイヤー（deployer）	・利用者が予測や判断に納得し、システムを使い続けられるようにすること ・利用者を何らかの行動や振る舞いに導くこと。例えば、アマゾンがある商品を勧め、その説明をすることで、クリックして購入するように仕向けること。

表 2 ロボティクス及び AI に関する透明性についてステークホルダー毎の理解の差異（Weller, 2017 より筆者翻訳による引用[23]）

この表が明らかにするとおり、透明性を実現する相手が開発者なのか、利用者なのか、社会一般なのか、専門家や規制当局なのかによって、それぞれのステークホルダーは異なる目的を持っており、同時に理解度も異なる。本ツールキットの項目選択や説明の記載方法もそれに応じた使い分けがなされるべきである。

またこれらのステークホルダーが所在する国・地域の観点も重要である。国・地域によって法規制や社会規範、価値意識は大きく異なる。それに伴い、社会が問題意識を持つ項目にも濃淡があり、それに応じて担保すべき透明性は異なっているだろう。

なお、本ツールキットは、別のステークホルダーとして社内の監査部門やリスク・コンプライアンス部門による利用も想定している。本ツールキットをチェックリストのように用いることで、AI の社内理解や当該 AI のリスク評価やマネジメントの促進に繋がるだろう。

4.4 透明性の粒度・程度について

透明性といっても、本ツールキットの各項目について、ただ闇雲に開示すれば良いというわけではない。研究者の中には、透明性を高めることに意義がない、あるいは有害でさえあるという批判も存在する[24]が、その強い根拠の一つは開示事項に対する利用者の理解に限界があることである。曖昧すぎる記載やあえて詳

[23] Weller, A. (2017). Transparency: Motivations and Challenges. arXiv:1708.01870. https://arxiv.org/pdf/1708.01870.pdf（2023 年 1 月 17 日閲覧）.

[24] Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645（2023 年 1 月 17 日閲覧）

細をはぐらかしたような説明は、透明性が目指す信頼獲得や理解の向上には全く資さない。一方、詳細に開示したからこれらが実現するかといえば利用者の理解に限界がある以上、これも首肯できない。しかし、AI アルゴリズムの透明性の粒度・程度に関する複数の先行研究を整理した表 3 には、現実的な可能性への示唆が存在する。

研究トピック	透明性の結果指標	発見事項
映画推薦システム	・システムの受容度 ・パフォーマンス	混合的な結果：パフォーマンスには効果はないが、受容度には正の効果あり
音楽推薦システム	・システムの満足度 ・信頼度	満足度と信頼度に対する正の効果
社会的ロボット	・責任の帰属 ・信用の帰属	弱い効果
文化財レコメンダーシステム	・システムの受容度 ・信頼度 ・コンピテンス	弱い効果：信頼度とコンピテンスには効果はなく、受容度には部分的に効果あり
行動認識システム	・理解度 ・信頼度 ・パフォーマンス	なぜそうなのか、そうでないのかの説明が、理解度、信頼度、パフォーマンスを向上させる
音楽推薦システム	・メンタルモデル ・利用者の信頼度	理解や信頼を深めるために健全で完全な説明が最適である
Facebook のニュースフィードアルゴリズム	・最初の驚き、怒り、不満の度合い ・段階的な満足度	アルゴリズムの存在を知ることで正負の効果が生まれる
MOOC におけるピア評価	信頼度	透明性の効果は利用者の期待とそれに背くか次第。結果が利用者の期待に反した場合、透明性が高くとも信頼度は低くなる。
環境に配慮したモバイルアプリ	・信頼度 ・認識されたコントロール	信頼度に対する正の効果はあるが、認識されたコントロールへの影響はない
オンライン広告	・信頼度 ・気味悪さ	具体的すぎる説明と一般的な説明は気味悪さに繋がる。中

	・満足度	間的な説明は信頼度と満足度を高める。アルゴリズムの透明性は幻滅に繋がる。
Facebook のニュースフィードアルゴリズム	・認知度 ・正しさ ・解釈可能性 ・説明責任	認知度への効果が最も高く、ついで説明責任への効果が高い

表 3 様々な AI の透明性に対する利用者個人への影響に関する先行研究（Felzmann et al.,2019 より筆者翻訳による引用[25]）

先行研究が示すのは、AI の種類や役割によって、透明性の粒度・程度に工夫しなければ、場合によっては利用者の信頼度や満足度、パフォーマンスをむしろ下げてしまうおそれがあるということである。特に、具体的に説明しようとして情報を過剰に提供し、結果的に利用者の理解を得られないようではむしろ AI への気味悪さを助長させたり、一見透明性が高いようでいて AI の挙動によってはむしろ信頼度を落としたりする可能性がある点には十分注意する必要がある。一方で、一般的すぎる説明をアリバイ作りのように置くだけでは、PR 効果を含む一切の期待する効果は得られない。利用者を始めとするステークホルダーの目線から、効果を最大化する開示の程度を AI の種類や役割に応じて使い分ける必要がある。

5 ユースケース

5.1 ユースケースの目的

本ツールキットを具体的にどのように利用できるか、簡単なユースケースを 2 つほど記しておきたい。これらのユースケースは決して理想的な開示事例を示すものではなく、意図的に議論の対象となる事項を含んでいる。ユースケースを通じて、開示項目の選択や記載方法の検討の一助になれば幸いである。なお、全てのユースケースは架空の事案であり、特に念頭に置いたシステムやサービスがあるわけではない。

5.2 政府機関が補助金制度の申請審査に AI を用いる場合

このユースケースでは、AI アルゴリズムの透明性を実現するにあたって、以下のような点が考慮要素となり、開示の項目、方法、程度はそれらに鑑みた対応で

[25] Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1). https://doi.org/10.1177/2053951719860542（2023 年 1 月 17 日閲覧）.

なければならない。

（1）主体：

XX 庁（政府機関）であり、透明性が強く求められる。

（2）AI の利用によるリスク・影響：

利用者に経済的影響（補助金申請の可否）を及ぼす。

（3）情報の受け手：

広く国民一般であり、そのリテラシーには幅広い差異がある。

このようなケースでは、強い透明性が求められるため開示事項は多岐にわたる一方、影響を受ける国民広く一般が理解可能な粒度に情報の開示方法を工夫する必要がある。

別表 2 として本ユースケースの開示例を記載している。決して理想的な事例を記載しているわけではなく、説明の便宜のためあえて注意が必要な記載をしている箇所もある。特に留意すべき事項について、以下簡単に検討する。

「ベネフィット・インパクト」は、利用者の理解を得るためにもなるべく具体的に、そしてステークホルダー毎の観点で記載することが望ましい。特に政府機関が利用する場合は、税金を用いることへの費用対効果の観点も含まれるべきである。

「利用方法」では、AI が意思決定を下すのではなく、人間のサポートに過ぎないことが強調されている。最終的に人間が責任をもって審査することが明示されている点が重要である。

「開発チームの DEIB ポリシー」や「学習データの DEIB ポリシー」は、広く国民一般を利用対象とするサービスであればあるほど、高度なポリシーが求められる。特に何らかの属性に対する差別（合理的説明ができない異なる取り扱い）が生じないための措置を現在の記載以上に詳細に記載することが検討されるべきである。

「パラメータ・特徴量」では、主要なパラメータが 5 つ挙げられている。そして、それぞれの項目について簡単な記載が置かれている。また、特に行政による非合理な差別との誤解を生みやすい「年齢及び性別」というパラメータについては、詳細について別のページに誘導する記載がある。利用者間のリテラシーに差異が想定されるケースでは、このような階層的な記載が好まれる。

「パフォーマンス精度・限界」及び「事業者によるモニタリング・メンテナンス」については、政府機関が利用する以上、利用者が審査結果に疑義を抱かれないよう、事前に AI アルゴリズムの精度や誤った判断の可能性への対応方法を説明することが求められる。

政府機関により利用され、結果によって経済的影響が生じる AI アルゴリズムについては、性質上開示が許されないもの（個人データを含む学習データ等）を

除いて基本的には全ての事項をわかりやすく開示することが考えられる。先述したとおり、英国が先んじて政府機関による透明性の取組みを始めており、今後の実務対応を注意深く追っていきたい。

5.3 ソーシャル・メディアのフィードに AI によるレコメンデーション機能を用いる場合

次に民間事業者によるソーシャル・メディア上のフィードに用いる AI アルゴリズムの透明性を検討する。別表 3 の開示例を見ると、前項のユースケースとは異なり、開示していない事項も存在する一方、ユーザビリティの観点から利用者にコントロールを委ねているものも少なくない。

「学習データのソース」は、サンプルデータを格納するレポジトリが指定されている。事業者に競争優位性を持たせる学習データを全て開示することはできずとも、ランダムサンプリングなどによって抽出された一定のサンプルデータを公表することは現実的な対応オプションとして検討されるべきである。

「パラメータ・特徴量」は、主要なパラメータについてある程度抽象化したカテゴリで表記し、「重み付け」で相対的に重視されるものを列挙している。「2023 年 X 月時点」と明記しているのも特徴的である。また、特徴的なのは、個別のポストを選択することで表示された理由がパラメータと共に説明される設計になっていることである。大手のソーシャル・メディアなどでは徐々にこのような実装も始まっているが、ソーシャル・メディアにとどまらず、レーティングサービスやマッチングサービスなどでも積極的な活用が期待される。

「利用者によるコントロール」について、様々な記載が置かれている。本来はこのような項目を確認せずとも利用者がフィードを利用している最中に気づき、コントロールできる UI・UX を設計することが理想的である。

6 各国の規制状況と本ツールキットの対比

AI アルゴリズムをめぐる各国の規制は、特に GDPR、P2B 規則、AI 規則、DSA、DMA といった規則等で EU が先んじて議論や導入を進めている。AI と人間社会の共存を図る上で、EU がいかに人権や民主主義といった価値観を重要視しているかが見て取れる。GDPR はデータに関する権利を人権の議論に引き上げ、AI 規則は AI の内包するリスクに応じて異なる規制を当てはめる、いわゆるリスクベースアプローチを採用した点で特徴的である。本稿も、AI を開発・提供する主体、AI を利用する者、利用方法、AI が及ぼすリスクに応じて透明性のレベルを区分する点で同じ立場を取っている。P2B 規則や DSA、DMA は、特にプラットフォーム事業者やオンラインサービス事業者が社会に果たす役割の大きさに着目し、規模や業態に応じて、透明性の観点から社会との適切なコミュニケーションの実施を求める規制が含まれている。

一方、日本でも DPF 取引透明化法において、指定された特定の巨大プラットフォーム事業者を対象に AI アルゴリズムの透明性を求める共同規制が導入されている。本稿

でもすでに触れたように、プラットフォーム事業者による自主的な開示が実現しており、プラットフォームの利用者による理解も進んできている。

なお、アメリカは依然として AI アルゴリズムの透明性に関する具体的な規制は整備されていないが、すでに議会での議論は開始している。最も直接的な法案としてはアルゴリズム説明責任法（Algorithmic Accountability Act）[26]が挙げられ、こちらもやはり重要な意思決定を担う AI のようなリスクの高いケースに対する高度な透明性・説明責任などを求める内容になっている。ただし、当該法案は未成立であり、今後も米国議会はこの観点について当面議論を続けるだろう。

それぞれの規制について本稿でその詳細な説明は避けるが、別表 4 として各国の規制状況と本ツールキットの対比表を作成した。もちろん法令毎に規制の対象も異なれば規制の内容も様々であるため、対比表は個別の法律・法案の解釈を説明するものではなく、大雑把にそれぞれの法律・法案がどういった項目について関心を有するかを示したに過ぎない。この対比表からも、本ツールキットが透明性というトピックで俯瞰的、体系的な議論に挑戦していることがご理解いただけるかと思う。

7　結語

AI によるイノベーションが生み出す正のインパクトを最大化するためには、当該 AI が内包するリスクをステークホルダーが認識し、その利害調整を適切かつ柔軟に実行できる技術的、組織的、社会的システムの設計と運用、すなわち当該 AI に対するアジャイルなガバナンスのフレームワークが不可欠である。本稿は、AI アルゴリズムの透明性という最も議論が盛んなガバナンスの手法について、俯瞰的、体系的な整理に挑戦し、実務に耐えうるツールキットを提案した。イノベーティブな AI が不完全なコミュニケーションに起因する誤解によって社会実装を阻害される不幸は回避されるべきであり、そのために透明性が果たすことのできる役割は小さくない。ただし、体系的な整理がない中で、イノベーションが種々の問題を引き起こす度に規制や規範が形成されることは、設計図のない建物の増築が繰り返されるようなものであり、予見可能性を欠くとともに、本来の目的を離れた形式だけの「透明性」が自己目的化していきかねない。そのような状況では、イノベーションは阻害されていくだろう。このような状況を避けるためには、アルゴリズムの透明性に関するルールについて、統一的な考え方のもとに体系的に整理されていることと、事業者が実運用しやすい汎用性、明瞭さ、裁量の余地がある柔軟性が追求されなければならない。

本稿では、各国当局による規制案のみならず、様々な AI によって生じる新たなリスク事象を幅広く考慮し、さらには AI アルゴリズムに関する先行研究を踏まえながら、

---

[26] CONGRESS.GOV “H.R.6580 - Algorithmic Accountability Act of 2022” https://www.congress.gov/bill/117th-congress/house-bill/6580 （2023 年 1 月 28 日閲覧）.

体系的な開示事項・事例集としてのツールキットを構築した。また、事業者が透明性の程度を工夫することで、社会における信頼性のみならず満足度など異なる指標にも良い影響が与えられる観点を取り込んでいる。加えて、開示事項をリスト形式にし、それを AI アルゴリズムの提供者、利用者、リスクなどの観点から裁量的な選択性とするなど、自主・共同規制に対応する事業者にとって利便性の高いツールキットとなることを目指した。事業者が、利用者や社会、当局、専門家らとのコミュニケーションのツールとして、あるいは社内のリスクマネジメントツールとして、本ツールキットを利用していただくことができれば、これほど嬉しいことはない。それが AI によるイノベーションのインパクト最大化に繋がる最適なアジャイル・ガバナンスの一つだと信じている。

本ツールキットは ver1.0 とし、今後の指摘や議論の発展を踏まえて、アップデートを重ねていきたいと考えている。読者の皆様におかれては、ぜひ過不足の指摘などご指導いただきたい。

8 謝辞

最後に、本稿の提案の場を提供してくださり、様々な観点から有益な指摘を続けてくださった宗像直子先生、羽深宏樹先生、東京大学公共政策大学院、同「イノベーションガバナンスエキスパート養成プログラム」受講生の皆様に心から感謝申し上げる。

別表１　AI アルゴリズムの透明性開示項目リスト

Lv1	Lv2	本文項目	System	Data	Model	Output	Human	開示項目
プロダクト思想	利用目的	3.4.1	○					AI を利用する目的
	ベネフィット・インパクト	3.4.2	○					AI を利用することによって得られる便益（ベネフィット）、影響（インパクト）のアセスメント
	利用方法	3.4.3	○					AI が活用されているプロセス、AI が担う役割（意思決定そのものを行うのか、人間の意思決定をサポートするのか等）
開発チーム	開発責任者・管理責任者	3.5.1					○	管理責任部門・責任者の所在
	開発チームの DEIB ポリシー	3.5.2					○	開発チームの構成・多様性、想定利用者との整合性
	委託先管理	3.5.3					○	委託先の管理体制・方法
データ	学習データの概要	3.1.1		○				学習データ・テストデータの概要
	学習データのソース	3.1.2		○				学習データ・テストデータの所在
	学習データの入手経路	3.1.3		○				学習データ・テストデータの入手方法・入手経路
	学習データの DEIB ポリシー	3.1.4		○				学習データ・テストデータの多様性、想定利用者との整合性
	実運用データ	3.1.5		○				実際に用いられるデータの概要（AI は実際に何を Data として Output を抽出するのか）
AI アルゴリズム	利用フェーズ	3.2.1	○					利用されるアルゴリズムのフェーズ（研究、テスト、実運用等）

別表１　AIアルゴリズムの透明性開示項目リスト

	学習手法	3.2.2			○			学習手法AI（エキスパートシステム、ルールベース、機械学習、深層学習等）
	説明可能性	3.2.3			○			アルゴリズム自体の説明可能性
	パラメータ・特徴量とその目的・理由	3.2.4			○			合理的に理解可能な程度に抽象化・言語化されたパラメータ
	重み付けとその目的・理由	3.2.5			○			重要なパラメータに対する重み付け・係数、あるいは他のパラメータに比して相対的に重要である理由
	課金を含む利用者区分毎の異なる取扱い	3.2.6			○			パラメータや重み付けと課金を含む行為・属性等利用者区分の関係
	変更の適正手続	3.2.7	○					アルゴリズムの変更に係る適正手続、変更による影響、変更の目的等
アウトプット	パフォーマンス精度・限界	3.3.1				○		アウトプットの有意確率、誤謬率、偽陰性・偽陽性等
	事業者によるモニタリング・メンテナンス	3.4.4					○	アウトプットの品質に対するモニタリング・メンテナンス体制（頻度・方法・体制）、検出件数
利用者によるコントロール	アルゴリズム利用の選択権	3.4.4.1	○					アルゴリズムの採否の選択権
	フィードバック	3.4.4.2	○					利用者によるアルゴリズムやデータセットへのフィードバックの有無、可否及び方法
	データ・オプトアウト	3.4.4.3	○					学習データからのオプトアウト方法の有無

別表１　AIアルゴリズムの透明性開示項目リスト

マネジメント体制	リスクマネジメント	3.5.5	○					AIを用いることのリスクに対するアセスメント、最小化又はマネジメント対応
	教育体制	3.5.6	○					AIの開発・提供に係る教育状況
	監査体制	3.5.7	○					AIの開発・提供に対する監査体制及び実施状況
	問い合わせ対応	3.5.8					○	問い合わせ先・不服申立窓口、カスタマーエクスペリエンス体制（頻度・方法・体制）、一定期間経過後の苦情件数

別表２：ユースケース１　政府機関が補助金制度の申請審査に AI を用いる場合

Lv1	Lv2	開示例
プロダクト思想	利用目的	補助金申請の審査担当課が、多数の審査書類を精査し、審査基準に満たない申請を抽出するために AI アルゴリズム（以下「本件 AI」）を利用します。
	ベネフィット・インパクト	・審査担当課にとっての便益：月次 250 万件の審査を実施する予定ですが、本件 AI の利用により、審査工数が 70%削減されます。そのため、全ての審査を人間が実施した場合に比べ、審査日数約 1 週間の縮減、人件費月額約 4.2 億円の削減が望めます。 ・申請者に対する便益：本件 AI の利用により補助金を受け取るまでの時間が約 1 週間早まります。
	利用方法	本件 AI は以下の審査において用いられます。ただし、全ての審査において、最終的な判断は審査担当官が行い、あくまで本件 AI の役割はその判断のための補助的な仕分けです。 ・書面の形式審査：誤記や不明瞭な記載などの形式面を審査します。 ・対象基準の形式審査：補助金の給付対象者に該当する者かを、書面上に記載された事業所得、業種などから審査します。 ・対象基準の実質審査：予算上限がある中で分配される補助金の給付対象者の優先順位を様々なデータ（後述）から審査します。
開発チーム	開発責任者・管理責任者	XX 庁 CTO（情報システム部門長兼務）：XX
	開発チームの DEIB ポリシー	本件 AI を開発する XX 社は、申請者の属性（特に年齢、性別及び業種）が多様であることを認識した上で、テスト段階において、それぞれの属性の被験者（試験的に申請を実施する方）が差別的結果の発生の有無を確認するプロセスを設けています。
	委託先管理	本件 AI を開発する XX 社との間で、我が国が定める開発ガイドラインの遵守を求め、特に以下の項目について本庁に対する報告を求めています。 ・説明不可能又は極めて困難な結果が出た場合 ・本件 AI の判断の正確性に疑義が生じた場合

別表２：ユースケース１　政府機関が補助金制度の申請審査に AI を用いる場合

		・XX
データ	学習データの概要	本件 AI を開発するために用いたのは、本庁が保有する過去の補助金申請書類、過去の審査結果及び理由、業種毎の年度別収益状況を含む政府統計データ、確定申告に関するデータ、納税に関するデータです。
	学習データのソース	－（機微な個人情報であり、データ自体を開示することはできません）
	学習データの入手経路	本庁又はデータを保有する他の省庁に対して、本人から直接提出されたデータです。
	学習データの DEIB ポリシー	上述のとおり、学習データとして過去の同趣旨の補助金申請にかかるデータを用いていますが、今回の母集団を構成しないと思料される外れ値となるデータ（明確に対象外となった業種等）は事前にデータセットから除外しております。
	実運用データ	審査に際して用いられるのは、今回の補助金の申請のために本人から提出されたデータです。
AI アルゴリズム	利用フェーズ	本件 AI はすでに実運用段階（本番環境）にあります。
	学習手法	機械学習が中心ですが、形式審査については一部ルールベースのものが存在します。
	説明可能性	本件 AI が出力する結果は基本的に以下のパラメータ・特徴量や重み付けによって説明可能です。ただし、極めて稀に説明が難しいケースが存在し、その場合には本件 AI による出力結果は利用せず、審査担当官が個別に判断を行います。
	パラメータ・特徴量とその目的・理由	特に対象基準の実質審査では、主要なパラメータ・特徴量として以下を有しています。 ・業種（複数の政府統計等により今年度収益が悪化したと考えられる業種に対してポジティブな結果が生まれやすくなります） ・年齢及び性別（過去の審査結果に鑑み、業種や収益状況との関係性から特に保護すべき可能性がある属性としてポジティブな結果が生じる場合があります。詳細については、こちらをご覧ください） ・本年度の収益状況（過去 5 事業年度との平均に比べて減収減益となっていればポジティブな結果が生まれやすくなります）

別表２：ユースケース１　政府機関が補助金制度の申請審査に AI を用いる場合

		・過去 3 事業年度における修正申告履歴 ・過去 10 事業年度における納税状況（滞納や追徴課税がある場合のみネガティブな結果が生まれやすくなります）
	重み付けとその目的・理由	業種と本年度の収益状況が最も重視されています。
	課金を含む利用者区分毎の異なる取扱い	特にございません。
	変更の適正手続	本件 AI のアルゴリズムを事業年度内に変更する予定はございません。
アウトプット	パフォーマンス精度・限界	本件 AI を、過去の審査資料と審査結果と照らし合わせたところ、97%の精度でした。残りの 3%については、外れ値が含まれる申請であるか、過去の審査結果の妥当性に疑義があるものが多いと言えますが、なお本件 AI の精度は 100%ではありません。したがって、複数の審査担当官が本件 AI の結果をレビューし、最終的には審査担当官が審査判断を下すものとしています。
	事業者によるモニタリング・メンテナンス	本件 AI の結果は、2 名の審査担当官によって交互にレビューされます。2 名の審査担当官がともに本件 AI と異なる判断を下した場合、その結果は本件 AI を開発する XX 社にフィードバックされ、本件 AI の改善に利用されます。
利用者によるコントロール	アルゴリズム利用の選択権	本件 AI の利用方法の性質上、特にございません。
	フィードバック	本件 AI の利用方法の性質上、特にございません。
	データ・オプトアウト	本件 AI の利用方法の性質上、特にございません。
マネジメント体制	リスクマネジメント	本件 AI は、出力する審査結果が不適切であるリスクがあり、これは以下のように対応しています。 ・複数の審査担当官によるレビューとフィードバック ・過去の審査結果との整合性の確認

別表２：ユースケース１　政府機関が補助金制度の申請審査に AI を用いる場合

		・申請者による不服申立てとそれによって本件 AI の誤りが明らかになった場合のフィードバック
	教育体制	審査担当官は、本件 AI の仕様及び利用方法について合計 15 時間の訓練を受け、本庁が求める基準試験に合格しています。
	監査体制	本庁の監査部門が本年度の補助金申請の判断について 1 年以内に監査を実施する予定であり、本件 AI や審査担当官の判断に疑義があった場合は公表することがあるとともに、本件 AI を開発する XX 社にフィードバックされます。
	問い合わせ対応	本件 AI に関する問い合わせ又は補助金申請に関する問い合わせ若しくは不服申立ては以下までお願いいたします。 XX 庁 XX 補助金申請問い合わせ窓口 電話番号：XX メールアドレス：XX

別表３：ユースケース２　ソーシャル・メディアのフィードに AI によるレコメンデーション機能を用いる場合

Lv1	Lv2	開示例
プロダクト思想	利用目的	ソーシャル・メディア「YY」の投稿フィード内で、速報順のコンテンツ表示ではなくユーザにとって利便性や関連性の高いコンテンツが提供されるように、レコメンデーション AI（以下「本件 AI」）を利用します。
	ベネフィット・インパクト	ユーザは満足度の高いコンテンツをより多く受け取ることができ、限られた時間を最適に過ごすことができます。
	利用方法	－
開発チーム	開発責任者・管理責任者	－
	開発チームの DEIB ポリシー	開発チームは、人種、性別、宗教、教育などの観点で多様性が担保されており、本件 AI の開発にあたって最低一人はマイノリティの観点からレビューすることが求められています。
	委託先管理	－
データ	学習データの概要	YY 内への過去 3 年間のポスト、ポストへのユーザのリアクション（リアクションボタン、コメント、拡散、閲覧時間等）、ポストフォーマット（テキスト、画像、動画）などを学習データとしています。
	学習データのソース	サンプルデータセットは以下のレポジトリから入手可能です。 YY.com/repository/feed/sample-dataset/
	学習データの入手経路	YY 内フィード
	学習データの DEIB ポリシー	YY 内フィードから取得されたデータであるため、YY ユーザの代表性が一定程度担保されています。たとえば、初回起動後利用を停止したユーザや bot のような挙動を行うユーザは外れ値としてデータセットの母集団から外しています。
	実運用データ	実運用に用いられるのはユーザによる実際のポストデータです。
AI アルゴリズム	利用フェーズ	－

別表３：ユースケース２　ソーシャル・メディアのフィードにAIによるレコメンデーション機能を用いる場合

	学習手法	－
	説明可能性	－
	パラメータ・特徴量とその目的・理由	本件AIは以下のパラメータを主に保有しています。また、個別のポストの詳細メニュー内「なぜこのポストが表示されるのですか」を選択していただけば、個別のポストの表示理由が、用いられているパラメータとともに説明されます。 ・　ユーザ間の関係性 　・　ユーザ間の共通の友人の分量 　・　ユーザ同士の交流内容（相互の反応等） 　・　ユーザ間のメッセージの分量 　・　一方のユーザの他方ユーザのポスト上の滞在時間 　・　ユーザ区分 ・　ポストの内容 　・　フォーマット（テキスト、画像、動画） 　・　時事性 　・　関連性 　・　地域性 ・　ポストへのリアクション 　・　リアクションの分量 　・　リアクションの頻度 　・　リアクションのクロスコミュニティ度合い
	重み付けとその目的・理由	2023年X月時点で相対的に重要なパラメータは、ユーザ間の関係性とリアクションの分量及びクロスコミュニティ度合いです。

別表３：ユースケース２　ソーシャル・メディアのフィードに AI によるレコメンデーション機能を用いる場合

	課金を含む利用者区分毎の異なる取扱い	課金ユーザのポストは、優先的に非課金ユーザのポストよりも多く、高頻度で表示されます。
	変更の適正手続	本件 AI の重要な変更については、ユーザへの 7 日以上前の予告を行います。また課金ユーザに関する不利益な変更については定期支払いを解除できるための合理的な期間をおいて実施します。
アウトプット	パフォーマンス精度・限界	－
	事業者によるモニタリング・メンテナンス	本件 AI が違法有害な情報を多く取り扱うことのないよう、定期的に当社 Trust&Safety チームがフィード内をモニタリングし、該当ポストを削除したり、ダウンランク（表示の優先順位を落とす行為）したりしています。これらの内容と件数は毎事業年度の年次レポートで公表されています。
利用者によるコントロール	アルゴリズム利用の選択権	YY のフィードを、本件 AI による推薦順ではなく、速報順で楽しみたい場合、「設定画面」→「表示順位の変更」→「速報順に見る」を選択してください。同様に、同設定よりいつでも本件 AI による推薦順をお楽しみいただけます。
	フィードバック	YY のフィード内で満足度の低いポストがあった場合、ポストを長押しするか、ポスト下の「…」ボタンから、「このようなポストに興味がない」ボタンを押して本件 AI にフィードバックしてください。またその先の画面で理由を選択することで、よりあなた個人に適した推薦が可能になります。ただし、本件 AI がフィードバックをうまく活用できるまで時間的なラグがある点、ご認識ください。
	データ・オプトアウト	個人データに関する権利行使については、「設定画面」→「個人データについて」にお進みください。
マネジメント体制	リスクマネジメント	－
	教育体制	－
	監査体制	－
	問い合わせ対応	本件 AI に関する問い合わせは以下まで support-recommendation@example.com

別表４　各国の規制状況と本ツールキットの対比表

Lv1	Lv2	GDPR (EU)	AI 規則 (EU)	P2B 規則 (EU)	DSA (EU)	DMA (EU)	ATRS (UK)	AAA (US)	DPF 透明化法(JP)
プロダクト思想	利用目的	○	○	○	○		○		
	ベネフィット・インパクト		○				○		
	利用方法	○	○				○		
開発チーム	開発責任者・管理責任者		○		○		○		
	開発チームの DEIB ポリシー								
	委託先管理						○		
データ	学習データの概要	○			○		○		
	学習データのソース				○	○	○		
	学習データの入手経路	○					○		
	学習データの DEIB ポリシー								
	実運用データ	○				○	○		
AI アルゴリズム	利用フェーズ						○		
	学習手法						○		
	説明可能性			○	○				○

別表４　各国の規制状況と本ツールキットの対比表

	パラメータ・特徴量とその目的・理由			○	○				○
	重み付けとその目的・理由			○	○				○
	課金を含む利用者区分毎の異なる取扱い			○					
	変更の適正手続			○			○	○	○
アウトプット	パフォーマンス精度・限界						○		
	事業者によるモニタリング・メンテナンス		○		○		○		○
利用者によるコントロール	アルゴリズム利用の選択権	○		○	○	○	○	○	
	フィードバック			○			○		
	データ・オプトアウト	○				○			
マネジメント体制	リスクマネジメント		○		○		○	○	○
	教育体制		○		○			○	○
	監査体制		○		○			○	○
	問い合わせ対応	○			○		○		○

*GDPR: General Data Protection Regulation *AI 規則: AI Regulation *P2B 規則: The EU Regulation on platform-to-business relations

*DSA: Digital Services Act *DMA: Digital Market Act *ATRS: Algorithmic Transparency Recording Standard

別表４　各国の規制状況と本ツールキットの対比表

*AAA: Algorithmic Accountability Act *DPF 透明化法：特定デジタルプラットフォームの透明性及び公正性の向上に関する法律